Long Non-Coding RNA in Controlling Chromatin Modification and Transcription

by

M. Jordan Rowley

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Molecular, Cellular, and Developmental Biology)
in the University of Michigan
2014

Doctoral Committee:

Assistant Professor Andrzej Wierzbicki
Professor Kenneth M. Cadigan
Professor Steven E. Clark
Associate Professor Györgyi Csankovski
Professor David R. Engelke

DEDICATION

To my children, Sophie and Grayson, and
especially to Chelsee, whose smile brightens
my life.

ACKNOWLEDGMENTS

My mentor, Andrzej Wierzbicki, has provided tremendous support and encouragement over the years. He has always made time for me, even though I'm sure that I popped into his office far too often because I was excited about a result or idea. At the beginning of this work he urged me to pursue bioinformatics and provided avenues and resources to develop these skills. His foresight and flexibility has provided the guidance and freedom to develop as a scientist. I believe the training in the Wierzbicki lab is beyond compare and I am glad to have had a chance to work there.

I would also like to thank all those in the lab with whom I've worked. An enormous amount of help has come from undergraduate students; Ivan Amies, Maria Avrutsky, Natalie Blackwood, Lilia Bouzit, Elaine Chang, Lois Dodson, Paula Gajewski, and Alex Stuart. They have greatly facilitated this work and should be applauded for the time they have spent on this research. Other past members of the lab have been great resources; in particular Yongyou Zhu, who is a master of cloning and has given me great advice numerous times. I would particularly like to thank Gudrun Böhmdorfer and Jan Kuciński for all their tremendous help and support, for the great discussions, and for reading this dissertation.

Most of all, I am grateful to my family for supporting me. My children, Sophie and Grayson, make life fun and enjoyable. Their excitement when I come home immediately washes away any frustration I may have had. My wife Chelsee is the most wonderful, patient, and supportive person I know. She backs me in all my endeavors and always knows how to cheer me up. She also puts up with all my faults and pushes me forward to always be better. I count myself lucky, as each day I am excited for work and excited for home. Thank you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1
Introduction

Gene Expression Control

Living organisms comprise an amalgam of molecular interactions, each component contributing to the complexity of life. Some of the most basic, yet essential processes and components of life hinge on nucleic acids and proteins. Indeed, the central dogma of biology is based on interactions between nucleic acids (DNA and RNA) and proteins. Decades have been spent studying these processes producing great insights into both simple and complex organisms.

Past research on DNA allows speculation that a conserved and basic role of nucleic acids is to encode information. DNA is comprised of nucleotides strung together in sequences forming a code to direct functions much like how computer codes can be used to direct actions. In the cell, the nucleotide sequence of DNA in genes is used for protein synthesis. In this process, cellular machinery reads the DNA sequence code and produces a transcript, or copy of the sequence, in the form of RNA. The transcript can then be modified, transported, and read by machinery responsible for producing proteins. Through this process, proteins are made according to the distinct information encoded in each gene. Additionally, by controlling the amount of transcript available of a gene, the cell can increase or reduce the levels of any particular protein.

The amount of RNA produced from a gene is referred to as gene expression and varies from gene to gene. In order to perform both simple and complex processes, the cell controls the amount and timing of specific protein production. A major way of doing this is by controlling the levels of gene expression which is essential for a myriad of biological processes including proliferation, differentiation, and just plain cell survival. How the cell controls gene expression is a basic and fundamental question in biology with no one particular answer. An impressive range of mechanisms have been explored in this regard; however, the more that is understood of this process, the more the

depths of our ignorance are known. This leaves a vast and exciting frontier of exploration into such a fundamental aspect of biology.

Controlling the expression of genes is a complex task which can involve communication between two or more regions of DNA, between DNA and proteins, and between proteins interacting with each other. Regions of DNA which are involved in controlling gene expression are often referred to as regulatory regions. These are sections of DNA which interact with proteins to influence transcription. Proteins, known as transcription factors (TFs), are targeted to and bind to regulatory regions and from there are able to control the expression of a gene. Some ways in which proteins binding to regulatory regions can control gene expression are discussed below.

In order to obtain expression of a gene, protein machinery responsible for transcription must be recruited to that site. Research into this machinery has revealed that RNA POLYMERASE II (Pol II) is an enzyme responsible for transcribing most protein coding genes[1]. By modulating Pol II binding and/or activity, specific transcription factors can enhance or inhibit transcription[2]. This is often seen by changes in gene expression upon mutation of a transcription factor. In eukaryotes, transcription factors can have combinatorial effects in that multiple proteins, or protein complexes, at one site are required to modulate expression. Furthermore, transcription factors influence each other; for example, a protein binding by itself may cause increases in transcription, but when other proteins bind it may repress transcription. These variations in transcription factor activity depend on many factors including the local nuclear environment of the DNA and the type of regulatory region present as well as regulatory regions nearby.

One type of regulatory region, called a promoter, often occurs near the transcriptional start site (TSS) of genes. Core promoters, most often found directly at the TSS, are regions bound by transcription factors necessary for recruiting and activating Pol II[1,3]. By controlling TF binding to this region, the cell is able to control gene expression[3,4]. Also near the TSS, proteins can bind and may cause Pol II to pause or stall, providing another way to control expression[5]. Pausing polymerase is thought to allow hair-trigger transcription by keeping polymerase and other components ready at the site of synthesis. Similar to, and in addition to the core promoter, are proximal

promoters where transcription factors often bind[4]. This region represents an even greater dynamic influence on gene expression where specific transcription factor interactions at proximal and core promoters can alter transcription at a range of levels. Transcription factors influencing gene expression in this way are often called activators or repressors because they can either cause increases or decreases to gene expression depending on the type of protein, the DNA region targeted, and the conditions in the cell.

While promoters and regions near transcriptional start sites are important regions of regulatory DNA, other regions play a significant role in influencing gene expression. These regions, known as enhancers, are also bound by transcription factors which can interact with transcriptional machinery at promoters[4,6]. While promoters are generally directly upstream of protein coding genes, enhancers can be quite distant or downstream of genes, or even on separate chromosomes[6]. The mechanism by which enhancers exert long range control of transcription is an exciting and relatively new field and will be discussed in detail later in this chapter. Enhancer / promoter activity can vary by tissue type and can respond to environmental stimuli[4], allowing specific and precise control of gene expression under various conditions. The term enhancer suggests that TF binding to these elements positively controls, or enhances, transcription; however, silencers are also distant regulatory elements where TFs can bind, but ones which have a repressive effect on transcription[4]. Silencers are often difficult to delineate from enhancers as they both are TF binding sites and the expression output can depend on cellular conditions, thus enhancer is often used as a general term for distant regulatory regions and will be used as such in this text. In either case, distant regions of DNA and their associated proteins are able to dynamically control the expression of genes.

In addition to enhancers, insulators are also regions of DNA which are bound by proteins and control gene expression. Insulator proteins are thought to function mainly by influencing enhancer-promoter interactions[7,8]. This function can take the form of blocking enhancers from influencing a nearby promoter and / or can cause long-range enhancer – promoter interactions[8,9]. Insulator proteins have also been found to be involved in the three dimensional nuclear organization of DNA[10] and these proteins in particular will be discussed further on. The role of insulators as regulatory regions for

genes is gaining increasing attention and their relationship to enhancers and promoters seems to be of vast importance[6,8].

Genes and regulatory regions comprise important aspects of transcriptional control, but other regions of DNA can also be transcribed. Some of these regions can cause drastic changes to genes by even altering the genetic sequence itself. Transposable elements (transposons) are DNA sequences which have the ability to move, or “jump”, to other areas of DNA[11,12]. Movement of transposons can occur through transcription followed by reverse transcription of these regions to make DNA copies which can be inserted into the genome (nuclear DNA content)[12]. Other types of transposons use a method involving DNA excision followed by insertion elsewhere in the genome[12]. In either case insertion of DNA sequence can cause problems for gene expression. If insertion occurs inside the gene’s transcribed region, the sequence is disrupted and the correct protein will not be formed[11]. Insertion into regulatory regions can also cause problems by disrupting protein recognition sites and thereby alter how gene expression is controlled[11,13].

Transposons are present in abundance in the human genome taking up an estimated 45% of the human DNA content which is more than thirty times that of protein coding genes[14]. In some organisms (i.e. maize), transposons are even estimated to take up approximately 85% of the total DNA content (genome)[14]. Such a large amount of potentially harmful DNA sequence is tightly controlled by mechanisms which inhibit protein binding and transcription specifically at these regions[12]. If transposon movement is left unchecked, insertion into gene sequence can disrupt genes or regulatory regions and can lead to disease or cell death[15]. Mechanistic understanding of transposon movement and inhibition has allowed insights into several diseases like cancer and has generated new tools in bioengineering[15,16].

Although transposons pose potential threats, advantages to the cell in keeping these elements present have been proposed. Features and mechanisms controlling transposons are also used by the cell in viral defense or in repairing DNA[12]. These elements may also be a way in which organisms can rapidly increase mutagenesis and thus may allow the species to adapt to stress[12,15]. By activation of transposons, organisms can randomly alter their genetic sequence, increasing mutagenesis, and thus

increasing the likely hood of an advantageous mutation in a population. Transposons can also jump imperfectly and carry other genomic (DNA) sequences with them, allowing the copying or reshuffling of useful DNA sequences[12]. Additionally, transposons themselves have been proposed to work similar to enhancers, in that transcription factors target these regions and can control expression of protein coding genes[11,17]. Very little is known about the regulatory activity of transposons as these are most often maintained in a repressed state by the cell; however, repression of transposons may also contribute to gene expression. If a silenced transposon is close enough to a gene or to other regulatory regions, the mechanism silencing the transposon may also control gene expression[12].

Although reports suggest that about 1.5% of the human genome is used as genes for protein synthesis[18], it is unknown how much of the genome is devoted to potential transcription or transcriptional control. It is known, however, that non-coding regions (regions not directly containing information for protein synthesis) are essential to the cell. For example, by controlling TF binding to enhancers, organisms allow activation or repression in an environmental or tissue specific manner[4]. In addition, transposons are loci which can affect gene expression, are abundant in the genome, and can be transcribed[11,13]. In addition to all of these genomic features described, an increasing amount of evidence indicates that non-coding RNA transcription is prevalent throughout the genome. Non-coding RNA comes from transcription of loci which do not encode protein, but the RNA molecules themselves are used for a variety of functions. Roles of non-coding RNA in controlling gene expression and transposons will be discussed further in this work[19–21].

Overall, genomes are complex organizations providing dynamic and responsive gene expression control. Understanding mechanisms which influence regulatory regions, and thereby subsequent gene expression, will be valuable in understanding the basics of life and will naturally provide insights into organismal development and disease.

Chromatin

The environment in which regulatory regions are located is an important way that genes expression can be controlled[1,22]. One way of influencing this environment is by controlling chromatin, which is composed of DNA and associated proteins and is utilized by eukaryotes to package genetic information. Organisms generally have large amounts of DNA in each cell. Stretched linearly, the DNA from one human cell would be approximately two meters long, all of which must fit into a ~5-10 micrometer nucleus[23]. For further perspective, the total amount of DNA within one human straightened and lined end to end could stretch far past the edges of the solar system[24]. The packed organization of DNA in the nucleus is not random and the ability to effectively organize and package DNA is essential to complex organisms. Changes to the packaging at specific points can affect the accessibility of genes or regulatory regions to transcription machinery, which in turn can affect gene expression[22].

One way in which the cell packages DNA is by wrapping short regions (~147 base pairs [bp]) around core histone proteins to make up nucleosomes[25]. This core is often composed of two of each H2A, H2B, H3, and H4 proteins with an additional histone, H1, which binds to DNA not wrapped around the histone core[25,26]. Tails of these histone proteins extrude from the core nucleosome area and have amino acid residues which are prone to chemical modification[27]. The specific placement and / or modification of nucleosomes are fundamental contributors to gene expression control[26,27] and will be explored further in the work presented here.

Several lines of evidence support a role of nucleosome positioning in gene expression control. Promoters for genes with active transcription are often devoid of nucleosomes, a trend which indicates that the presence of these units on regulatory regions may influence expression[26,28,29]. Indeed removal of nucleosomes from promoters results in increased transcription[30]. The effects of nucleosomes on gene expression are thought to occur by making TF bindings sites inaccessible[26,31]. Some TFs require specific bends or conformation in the DNA in order to bind, thus it is possible that nucleosomes may inhibit TFs by altering the conformation of DNA[26]. Specific nucleosome positioning is also important at other regulatory regions such as

enhancers and insulators, and is thought to influence transcription factor binding at these sites[26,32].

Additionally, nucleosome occupancy in a given population or at different times can vary at individual loci[26,30,33]. Some nucleosomes can appear to be stable and precise, while others regions have more limited or dynamic binding, and others appear sloppy without a distinct target in the cellular population[26]. Nucleosome positioning can also change as part of normal cellular processes or in response to environmental conditions[26,30]. This can take the form of insertion or eviction, or by sliding nucleosomes to nearby regions[26]. Changes in nucleosome positioning correspond to changes to gene expression and mutants for nucleosome remodeler enzymes can cause dramatic phenotypic effects, illustrating the importance of nucleosome remodeling for proper development[34].

While nucleosomes at regulatory regions can control gene expression, so can well placed nucleosomes often found immediately downstream of the transcriptional start site[5,26]. This nucleosome may control expression by causing polymerase to pause or stall[5]. TFs such as SPT4 and SPT5 (Supressor of Ty insertion 4 / 5) (see Appendix A Table 1.1) work as elongation factors to overcome polymerase pausing[5,35]. This may occur through displacement of nucleosomes during polymerase pausing, followed by repositioning after transcription has passed through[5,35]. This is an example of the multifaceted relationship between chromatin and transcription factors; chromatin can affect TF binding and TF binding can affect chromatin.

The relationship between chromatin and transcription factors takes on even further depth by modification of nucleosome components which stick out from the core, called histone tails. Extensive evidence supports the idea that changes to histone modifications can influence gene expression[27,36,37]. This is thought to occur by altering the accessibility of DNA through compaction (into heterochromatin) or by decompaction / release (into euchromatin)[27]. Histone modifications may also work directly as protein binding sites, and changing the modification can create or eliminate recognition sites for transcription factors[27].

A myriad of histone modifications at dozens of residues have been characterized which have varying effects on gene expression[27]. General rules for histone

modifications have been proposed, such that acetylation of H3 (H3Ac) and methylation of H3 lysine 4 (H3K4me) often define euchromatin or active transcription, while methylation of H3 lysine 9 (H3K9me) often characterizes heterochromatin or inactive transcription[27,36,38]. These are only a few of the known histone modifications and interplay between others must also be considered. Additionally, variable effects to gene expression between organisms or within the same organism at different genomic locations occur and histone modifications must often be viewed as part of a bigger picture[39,40]. These variable effects are likely dependent on the interplay between the type of modification present and the specific TF recruited / inhibited.

Modification of chromatin by nucleosome positioning and histone modification have a large role in gene expression control, but another feature of chromatin which can affect gene expression is DNA methylation at cytosines[41,42]. In mammals, DNA methylation mainly occurs in the context of CG dinucleotide sequences [41]. Unmethylated CG islands often occur on regulatory regions associated with active transcription, and methylation of these regions is associated with transcriptional repression[41,43]. It is also suggested that more widespread methylation occurs in stem cells and that dynamic DNA methylation of regulatory regions is important for tissue differentiation and embryogenesis[43]. In particular, methylation of cytosines in other contexts such as CHG or CHH (H = A, C, or T) are thought to occur more frequently in embryonic stem cells to help control transcription[44].

CG and CHG methylation are symmetric sequences in that there are cytosines on both DNA strands which are generally methylated together[41]; thus after DNA replication, one strand in each copy retains the methylation information (Figure 1.1). Maintenance methyltransferases (DNMT1 [DNA METHYLTRANSFERASE 1] in humans and MET1 [METHYLTRANSFERASE 1] or CMT3 [CHROMOMETHYLASE 3] in plants – see Appendix A Table 1.1) can use single stranded methylation marks in CG or CHG sequence contexts to place methylation on the opposite strand[41]. In fact DNMT1 has been shown to associate with the replication machinery and relies on the protein UHRF1 for targeting[45]. UHRF1 recognizes hemimethylated DNA and H3K9me2 for targeting and can flip out methylated cytosines which are likely passed to maintenance

methyltransferases[46,47]. It is likely that this flip-out mechanism allows recognition of un-methylated cytosines on the opposite strand[46,47].

In contrast to CG or CHG methylation, CHH sites are asymmetric in nature and after replication of CHH methylated sites only one copy of the DNA retains the methylation (Figure 1.1). In this sequence context the second copy must undergo *de novo* methylation after each round of replication. Even in CG and CHG methylation *de novo* methyltransferases are necessary for the original placement of these marks[41]; thus studying *de novo* methylation is an effective means to study pathways responsible for placing methylation marks regardless of sequence context.

In plants CHH methylation sites occur much more frequently than mammals and are commonly used to study an important pathway for *de novo* methylation called Transcriptional Gene Silencing (TGS)[41,48]. Studies of this pathway show that targeting of *de novo* DNA methylation to specific loci can alter gene expression[34,49]. Others have shown evidence that misregulation of DNA methylation is involved in cancer and many other diseases[50]. Research also supports an essential role for DNA methylation in transposon silencing[41,51].

DNA methylation, like histone modification and nucleosome positioning, could exert its function by directly inhibiting transcription factor binding or by recruitment of methyl binding proteins. The interplay between DNA methylation and histone modifications is strong, such that alterations in DNA methylation can change histone modifications and thus affect the accessibility of chromatin[52]. The inverse is also true, where histone modifications can alter the placement of DNA methylation and thereby influence transcription factor binding[52]. It is apparent that both are closely entwined as chromatin modulators of the DNA environment.

Chromatin modifications in the form of DNA methylation, histone modifications, and nucleosome positioning are important for gene expression control; however, the three dimensional spatial organization of chromatin can also play a role in controlling genes. It is often tempting to think of DNA as linear, and even when adding nucleosomes to the picture we think of beads on a string. That string, however, is more like a tangled ball of yarn. DNA is packed into the nucleus and can loop around to linearly distant regions or even to other chromosomes[8,53,54]. The three dimensional

organization of DNA has a function; it allows linearly distant regions of chromatin to come into close proximity and influence each other[6,54–56]. Recent advances in technology involving chromosome conformation capture and high throughput sequencing have rapidly deepened our understanding of chromosome conformation and allow detection of linearly distant enhancer – promoter interactions[57]. Although studying the three-dimensional genome is a relatively new field, general principles of nuclear organization have emerged. Data shows the clustering of regions with similar chromatin features including centromeres, telomeres, and inactive regions into distinct domains, as well as long range interactions between more active regions[54,58].

Chromatin organization also includes the clumping of large genomic sections into Topologically Associated Domains (TADs)[54,59]. These are defined as regions with topological similarities which interact at high frequency. These TADs generally consist of regions having similar histone modifications and have borders at active chromatin marks[60]. Little is known of their function, but it seems that TADs may play a vital role since they have been found in many organisms and are relatively static throughout the cell cycle, quickly reforming after mitosis[53]. However, TADs have not been discovered in organisms such as yeast or plants, but this may be due to differences in the developmental stage used or in other experimental conditions[61,62].

The organization of chromatin into specific domains is mediated by insulator proteins[10]. These proteins bind preferentially to nucleosome free regions and form complexes which bring long distance sections of chromatin into close proximity[7,10]. Insulator proteins cluster at TAD boundaries which may be one way in which these are defined, however insulators can also be found throughout the domain[60]. Several insulator proteins have been found in mammals as well as in flies (*Drosophila melanogaster)* with CTCF being the most well known[60,63,64]. CTCF and other known insulators are not found in yeast or plants, which may also explain a lack of evidence supporting TAD formation in these systems. However, the lack of TADs does not mean that chromatin is not organized specifically. Long distance chromosome looping exists in both of these systems and several other proteins might display insulator activity[61,62,65].

In addition to organization of chromatin into TADs, chromatin looping mediated by insulators allows for enhancer-promoter interactions[6,9]. Insulators were traditionally thought to “insulate” particular genes from the effects of a nearby enhancer. This is generally thought to be at least partially caused by specific chromosome looping to a different gene. As mentioned, these interactions can occur at a diverse range of linear distances and even between different chromosomes[6,54,55]; however, the mechanisms for precise pairing between specific promoters and enhancers remains a mystery. Additionally, it is thought that chromatin modifications influence these interactions, a subject about which relatively little is known but which merits further exploration[22,55]. It has been shown that silenced genes and chromatin with repressive marks are much less likely to participate in chromosome looping[54,58]. Similarly promoters with active chromatin marks are more likely to form loops than ones with inactive marks[54,58]. This suggests that mechanisms affecting local chromatin landscapes may also affect chromosome looping thereby contributing to gene expression control.

One intriguing idea is that transcription machinery is localized at specific points within the nucleus, and that DNA is brought to these sites[66]. As part of this model of transcription factories, active enhancers cluster together and are available for RNA Polymerase binding. Evidence for this exists as labeling of nascent RNA forms distinct spots within the nucleus[67]. Other researchers used polymerase fixed in place to show that DNA can be moved through the Polymerase[68]. This model is still under debate and others have shown the inverse: that fixing DNA in place causes polymerase to move[69]. It is unsure which is happening *in vivo* and either mechanism is possible, while an alternate scenario could occur where neither is fixed in location and both types of molecules move to associate. In any case, chromatin looping can occur and may affect transcription.

The effects of chromatin modifications and looping on the transcription of genes and transposons are fast growing fields of study. There are several mechanisms affecting chromatin modification, some of which are highlighted here. Nucleosome remodelers such as SWI/SNF proteins may slide, place, or evict nucleosomes which in turn may block or release potential transcription factor binding sites. Histone acetyltransferases and methyltransferases may alter modifications to histone tails, and

thus again affect accessibility to chromatin. Similarly, DNA methyltransferases can affect accessibility and thus recruitment or inhibition of proteins. The actual three dimensional arrangement of DNA can allow long distant regions of chromatin to interact and affect gene expression by enhancer-promoter pairing. Each of these aspects of chromatin may influence another, creating complex feedback systems between features. An additional aspect of chromatin which can control gene expression is non-coding RNA (ncRNA). Pathways involving ncRNA are involved in gene activation, post-transcriptional silencing, and in altering chromatin modifications for transcriptional silencing[70].

Non-Coding RNA

As mentioned, cells contain a huge amount of information in the form of DNA packed into the nucleus, but approximately 98.5% of this information does not encode protein[18]. The functional role of much of this DNA content remains unknown; however, increasing evidence indicates that non-coding regions of chromatin are extremely important. Although transcription of DNA encoding proteins occurs at very few sites in the genome, many other regions are transcribed into non-coding RNA (ncRNA)[56]. While it is unknown how much of the genome is transcribed to form ncRNA, numbers of discovered ncRNAs are constantly increasing, to the extent that low levels of transcription may occur throughout the entire genome[71]. These ncRNAs affect gene expression in a variety of ways and have been shown to be involved in the RNA splicing machinery, translation machinery, post-transcriptional silencing pathways, transcriptional regulation, and in chromatin modification[70]. Although several classes of ncRNA exist, the ones relevant to this work are generally classified as small RNA (sRNA) and long non-coding RNA (lncRNA).

Much of what is known of sRNA mechanisms suggests that they play a large role in silencing gene expression. One class of sRNA, termed microRNA (miRNA), participates in post-transcriptional silencing[72]. miRNA is generated from hairpin RNA formed by base-pairing with itself[73,74]. This hairpin form is exported out of the nucleus and recognized as double-stranded RNA for cleavage by DICER protein into 21 nucleotide (nt) fragments[73]. The 21 nt products are miRNAs which load onto an

ARGONAUTE (AGO) protein as single stranded fragments. The sequence specificity of the miRNA is able to then guide AGO to mRNA targets with similar sequences[72]. Once bound, AGO and associated small RNAs can silence transcripts in a few ways. If the miRNA sequence perfectly base pairs a region of mRNA, AGO can then directly cleave the transcript[72]. However if a few nucleotides do not match up with the mRNA, binding is still achieved, but in lieu of cleavage by AGO, this process causes rapid degradation of the transcript, and / or blocking of the translational machinery on this mRNA[72]. In this way, mRNA levels are controlled post-transcriptionally in the cytoplasm to regulate gene expression, a pathway also referred to as Post Transcriptional Gene Silencing.

Another class of small RNA is piwi-interacting RNAs (piRNAs) which are found mainly in the germ line and are important in gonad formation[74]. piRNAs are slightly longer, approximating 25-33 nucleotides in length and differ from miRNAs in that they do not require DICER for processing and that they are loaded onto PIWI proteins which are part of the AGO family[74]. This class of sRNA mainly targets transcribed transposons for cleavage, the fragments of which can be processed into additional piRNAs[74]. Another function of piRNAs is to target *de novo* DNA methylation to transposons in order to silence transcription, thus exhibiting a role in chromatin modification[74].

A third class of small RNA is small interfering RNA (siRNA) processed from double-stranded RNA (dsRNA) which can act in a broad range of tissue type and are found in most organisms[74]. These are loaded onto AGO proteins and can work similarly to miRNA in Post Transcriptional Gene Silencing[72]. Depending on the biogenesis and processing of double stranded RNA these siRNA-AGO complexes can instead target chromatin where they direct DNA methylation and chromatin modifications in a process known as Transcriptional Gene Silencing (TGS) or RNA-directed DNA Methylation (RdDM)[41,74]. Traditionally, siRNAs affecting chromatin were mainly thought to target and silence transposons, but work from several groups have shown that they also target regulatory regions of DNA (chapter 2) [49,75,76].

In addition to small RNA, long non-coding RNA (lncRNA) can affect gene expression[70]. These are loosely defined as transcripts which form no known protein product. Some lncRNAs are known to act at enhancers and are important for enhancing transcription of genes[77,78]. This may be through the recruitment of TFs to enhancers /

promoters, by causing conformational changes in TFs, or by acting as competitors for repressor binding so that DNA regulatory regions are not targeted[70]. Activating lncRNAs such as these are generally found in active chromatin and are implicated as important features of chromosome looping[79].

Perhaps a more prominent role of lncRNA lies in transcriptional silencing. One widely known lncRNA, termed HOTAIR, guides repressive chromatin modifications to target loci and acts as a scaffold to help proteins associate with chromatin and influence expression of genes[70,80]. Another well known lncRNA, Xist, is essential for dosage compensation of the X chromosome[81]. Xist accumulates along the X chromosome marked for inactivation, where it acts as a scaffold for chromatin modifiers to place repressive marks[81]. These lncRNAs illustrate general principles of lncRNA in which they act as a scaffold for proteins to guide chromatin modification and thus represent ways in which cells may target repressive chromatin marks to specific regions.

The Transcriptional Gene Silencing Pathway

In most organisms, lncRNAs involved in transcriptional silencing are thought to be transcribed by the same polymerase necessary for protein coding genes, Polymerase II (Pol II). This dual function makes genetic studies difficult in many organisms, as disruption of Pol II would cause widespread changes in gene expression. However, in *Arabidopsis thaliana* a specialized RNA polymerase, Pol V, is responsible for transcribing much of the lncRNA in the cell[82,83]. Pol V recently diverged from Pol II and these complexes share several subunits, however a few subunits are specific for lncRNA production by Pol V; most notably the largest subunit, NRPE1[83]. This allows knockout of Pol V (*nrpe1* mutant) without disruption of the Pol II protein coding transcriptional machinery.

In Arabidopsis, the *nrpe1* mutant displays no visible phenotype except under specific conditions, indicating the maintained functionality of Pol II[83]. However, knockout of Pol V results in activation of a number of transposons as seen by increased Pol II transcription at these loci[82] and the minor visible phenotype in Arabidopsis is thought to be due to the low transposon content of the genome. Indeed in Maize mutations of this pathway show more severe phenotypes[51].

Very little is known of how Pol V is recruited to chromatin since no discernible core promoter or motif regions have been found. However, proteins such as RDM1 (RNA-DIRECTED DNA METHYLATION 1), DMS3 (DEFECTIVE IN MERISTEM SILENCING 3), and DRD1 (DEFECTIVE IN RNA-DIRECTED DNA METHYLATION 1) form complexes which are necessary for Pol V to bind chromatin (see Appendix A Table 1.1)[51,84]. Additionally, there is some indication that chromatin modifications can play a role in Pol V recruitment to chromatin. Mutation in a maintenance DNA methyltransferase causes loss of *de novo* DNA methylation[85]. It is also suggested that the DNA methylation binding proteins, SUVH2 (SU[VAR] 3-9 HOMOLOG 2) and SUVH9 (SU[VAR] 3-9 HOMOLOG 9), are necessary for proper Pol V chromatin binding [86]. This represents an example in which chromatin structure may affect not only transcription of genes, but also transcription of lncRNA.

In addition to Pol V, *Arabidopsis* has another specialized RNA polymerase, Pol IV which is involved in creating siRNA[48]. Pol IV shares several subunits with Pol II and Pol V, but the largest subunit, NRPD1, is distinct, of which knockout lines are used in genetic studies[83]. Transcription by this polymerase also contributes to TGS in a different role; lncRNA produced by Pol IV is processed into siRNA[48,83]. Most of the siRNA in the nucleus is thought to be derived from Pol IV transcripts, although some can depend on Pol V[76]. Feedback between the two pathways can also influence siRNA levels, to an effect that much of the Pol V dependent siRNA may also depend on Pol IV[76].

Pol IV is closely associated with RNA DEPENDENT RNA POLYMERASE 2 (RDR2) which uses Pol IV transcripts as a template to make double stranded RNA (dsRNA) (Figure 1.2)[82,87]. It is thought that this happens very quickly and that RDR2 forms a complex with Pol IV so that dsRNA may be synthesized directly as the locus is transcribed[87].

Very little is known of how Pol IV and RDR2 are targeted, but some evidence suggests that it is dependent on SAWADEE HOMEODOMAIN HOMOLOGUE 1 (SHH1)[88]. SHH1 contains lysine binding pockets which may recognize histone tail modifications[88]. These pockets recognize methylated H3K9 and unmethylated H3K4, the combination of which is a mark of heterochromatin[88]. In this way SHH1 may bind already present chromatin marks and recruit Pol IV for siRNA production.

Once Pol IV has transcribed and double stranded RNA is produced, the RNA is then processed by a DICER-LIKE (DCL) protein into 24 nucleotide (nt) siRNA (Figure 1.2)[51]. In *Arabidopsis* there are four different DCL proteins involved in small RNA biogenesis[89]. While some overlap between pathways occurs, DCL3 generally produces 24 nt siRNA and is the DCL mainly responsible for processing Pol IV / RDR2 produced dsRNA for TGS (DCL2 and DCL4 also contribute significantly to the pool of siRNA and may have some redundancy with DCL3)[51,89].

An important next step in small RNA biogenesis is methylation on the 3' end by HEN1 (HUA ENHANCER 1)[90], but the purpose of methylating small RNA remains mysterious. The processed and finalized small RNA products are part of a pool of sequences which direct components of TGS to their targets in a sequence specific manner.

As previously discussed, small RNAs can direct silencing by binding to AGO or AGO-like proteins[41,51,72]. In the plant TGS pathway, siRNAs are loaded onto AGO4 which is then targeted to chromatin with sequences matching the small RNA[51]. However, this targeting of AGO4 seems to be dependent on both siRNA and Pol V transcription (Figure 1.2)[49]. It is likely that siRNA targets AGO4 to loci by base pairing with lncRNA produced by Pol V. An alternative also exists where base pairing of siRNA to DNA occurs at the region transcribed and opened by Pol V. In any case, the sequence specificity is likely important for AGO4 targeting to diverse loci in a precise manner.

Evidence also indicates that AGO4 binds to reiterated WG/GW motifs located on the Pol V C-terminal domain (CTD)[91]. These motifs are found on the NRPE1 subunit and may help AGO4 distinguish between various polymerases[91]. By making AGO4 dependent on Pol V and on siRNA, a dual check system is in place. In this scenario two conditions must be met for AGO4 to recognize a locus for silencing: 1. siRNA matching the sequence must be present; 2. Pol V must also be present. This would help hone the precision of silencing and eliminate unwanted silencing of other regions matched by siRNA or transcribed by Pol V.

Another component of the TGS pathway, SPT5L (also known as KTF1) (SUPRESSOR OF TY INSERTION 5-LIKE / KOW DOMAIN CONTAINING

TRANSCRIPTION FACTOR 1), contains WG/GW motifs and is thought to interact with AGO4[92]. Like AGO4, SPT5L associates with Pol V transcripts and is dependent on Pol V to bind chromatin (Figure 1.2)[38,92]. SPT5L is structurally similar to the transcription elongation factor SPT5, which is part of a nucleosome remodeling complex and is thought to help guide Pol II through nucleosome bound DNA[35]. However, unlike SPT5, SPT5L binds chromatin dependent on Pol V and is thought to have little effect on lncRNA transcription[38]. The specific roles of AGO4 and SPT5L in TGS were previously unknown, but are investigated in chapters 4 and 5.

INVOLVED IN DE NOVO 2 (IDN2) also binds to Pol V transcripts and is important for TGS (Figure 1.2)[34,93,94]. There are several IDN or IDN-like proteins which are thought to be able to form complexes with IDN2 or to act redundantly in this pathway[95]. It was shown that IDN2 interacts with SWI3B (SWITCH3B) a member of the SWI/SNF family of nucleosome remodelers[34]. This interaction and the roles these proteins play in the TGS pathway are discussed in chapter 3 and provide novel insights into nucleosome positioning.

The TGS pathway, otherwise known as RNA-directed DNA Methylation (RdDM), results in modification to chromatin, one form of which is the placement of DNA methylation[41]. DNA methylation is targeted through the coordinated action of TGS / RdDM pathway components to allow a *de novo* DNA methyltransferase to bind, called DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) (Figure 1.2)[41]. DRM2 is a homolog of the mammalian *de novo* DNA methyltransferase and is thought to function similarly. Evidence suggests that DRM2 binds to both lncRNA and DNA, where it places cytosines methylation[94]. As stated earlier, in Arabidopsis this is usually studied in the context of CHH methylation; these marks must be placed after each round of replication and thus represent *de novo* DNA methylation (Figure 1.1).

Genome-wide maps of DNA methylation have revealed that although an abundance of CHH methylation occurs near centromeres (where transposons are abundant), much of the TGS dependent methylation occurs in pericentromeric regions and chromosome arms (in more gene dense regions)[75,76,85]. Presence of CHH methylation on chromosome arms near genes is intriguing and could suggest a role in controlling expression of protein coding genes.

In addition to DNA methylation and nucleosome positioning, TGS is responsible for directing histone modification at specific loci. This generally includes removal of active marks such as H3 acetylation and addition of repressive marks such as H3K9me2[41,51]. These effects usually are present in conjunction with DNA methylation and a great amount of feedback between them may occur[51]. Despite this feedback, these modifications represent another way in which the TGS / RdDM pathway targets discrete loci for silencing by alterations to chromatin.

Functional Roles of Transcriptional Gene Silencing

Traditionally, the role of TGS / RdDM is thought to lie in transposon silencing. Transposons in many organisms, including humans, are more abundant than genes throughout the genome, and, as discussed, can pose a threat to genomic stability[14]. Much of the transposon content in the genome is centered in centromeric regions where silencing is maintained[49]. However, evolutionarily "young" transposons (those which have recently moved) are present throughout the genome. RdDM components were found to be enriched on these transposons and mutations in the RdDM pathway result in loss of repressive chromatin marks and transposon activation[49,75].

In addition to transposon silencing, this work shows that RdDM also targets regulatory regions and functions in controlling the expression of protein coding genes[49] (Chapter 2). Due to the presence of RdDM components on regulatory regions, and due to the great impact this pathway has on chromatin, the RdDM pathway may influence long range chromosome interactions. Findings from genome wide chromosome looping data are discussed in chapter 6. Here I examine looping in the context of RdDM to study if the effects this pathway has on gene expression may be explained by an ability to influence enhancer – promoter interactions.

RdDM's control of gene expression may represent a dynamic way in which chromatin responds to environmental stimuli. Genes controlled by RdDM often encode proteins involved in stress response pathways and evidence shows involvement in heat tolerance, pathogen defense, and DNA repair[96–98]. I propose that under stressful conditions, RdDM components are turned off which not only allows the upregulation of stress response genes, but also increases mutagenesis from transposon activation.

This may allow the cell to not only respond to the change in environment but also may provide means for adaptation.

A unifying current underlies the diverse functions of TGS / RdDM in that chromatin is controlled in order to manipulate genomic features. Whether these features are transposable elements or protein coding genes, TGS / RdDM can modify chromatin in order to control expression.

Overview of Dissertation

The work presented here focuses on how the TGS / RdDM pathway can exert control of transcription in *Arabidopsis thaliana.* I seek to understand the relationship between central components of TGS / RdDM (SPT5L and AGO4) in order to better understand each of their specific roles. I investigate various ways in which TGS / RdDM components are involved in chromatin modification, including nucleosome positioning. The main purpose of chromatin modifications placed by TGS / RdDM was thought to be in silencing transposons , but although transposons are found in abundance near the centromere, they can also be found throughout the genome where regulatory regions occur. By using genome wide maps of AGO4 binding sites, I investigate a possible role in gene expression control through modification of regulatory regions. Due to the myriad of chromatin modifications and the role in controlling gene expression, I also examine whether TGS / RdDM is able to influence enhancer – promoter interactions in the form of chromosome looping.

References

1. Kornberg, R. D. The molecular basis of eukaryotic transcription. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 12955–12961 (2007).
2. *Molecular cell biology*. (W.H. Freeman, 2000).
3. Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72,** 449–479 (2003).
4. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7,** 29–59 (2006).
5. Kwak, H. & Lis, J. T. Control of transcriptional elongation. *Annu. Rev. Genet.* **47,** 483–508 (2013).
6. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14,** 288–295 (2013).

7. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4,** e1000138 (2008).
8. Phillips-Cremins, J. E. & Corces, V. G. Chromatin insulators: linking genome organization to cellular function. *Mol. Cell* **50,** 461–474 (2013).
9. Mohan, M. *et al.* The Drosophila insulator proteins CTCF and CP190 link enhancer blocking to body patterning. *EMBO J.* **26,** 4203–4214 (2007).
10. Van Bortle, K. & Corces, V. G. The role of chromatin insulators in nuclear architecture and genome function. *Curr. Opin. Genet. Dev.* **23,** 212–218 (2013).
11. Rebollo, R., Romanish, M. T. & Mager, D. L. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu. Rev. Genet.* **46,** 21–42 (2012).
12. Fedoroff, N. V. Transposable Elements, Epigenetics, and Genome Evolution. *Science* **338,** 758–767 (2012).
13. Lockton, S. & Gaut, B. S. The contribution of transposable elements to expressed coding sequence in Arabidopsis thaliana. *J. Mol. Evol.* **68,** 80–89 (2009).
14. Huang, C. R. L., Burns, K. H. & Boeke, J. D. Active Transposition in Genomes. *Annual Review of Genetics* **46,** 651–675 (2012).
15. O'Donnell, K. A. & Burns, K. H. Mobilizing diversity: transposable element insertions in genetic variation and disease. *Mobile DNA* **1,** 21 (2010).
16. Munoz-Lopez, M. & Garcia-Perez, J. DNA Transposons: Nature and Applications in Genomics. *Current Genomics* **11,** 115–128 (2010).
17. Polak, P. & Domany, E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* **7,** 133 (2006).
18. Gregory, T. R. Synergy between sequence and size in large-scale genomics. *Nat. Rev. Genet.* **6,** 699–708 (2005).
19. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81,** 145–166 (2012).
20. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22,** 1775–1789 (2012).
21. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136,** 629–641 (2009).
22. Harmston, N. & Lenhard, B. Chromatin and epigenetic features of long-range gene regulation. *Nucleic Acids Res.* **41,** 7185–7199 (2013).
23. BioNumbers - The Database of Useful Biological Numbers. at <http://bionumbers.hms.harvard.edu/>
24. Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Ann. Hum. Biol.* **40,** 463–471 (2013).
25. Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423,** 145–150 (2003).
26. Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.* **10,** 161–172 (2009).
27. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21,** 381–395 (2011).

28. Sadeh, R. & Allis, C. D. Genome-wide 're'-modeling of nucleosome positions. *Cell* **147,** 263–266 (2011).
29. Weiner, A., Hughes, A., Yassour, M., Rando, O. J. & Friedman, N. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res.* **20,** 90–100 (2010).
30. Hogan, G. J., Lee, C.-K. & Lieb, J. D. Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. *PLoS Genet.* **2,** e158 (2006).
31. Szerlong, H. J. & Hansen, J. C. Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure. *Biochem. Cell Biol.* **89,** 24–34 (2011).
32. He, H. H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.* **42,** 343–347 (2010).
33. Teif, V. B. *et al.* Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol.* **19,** 1185–1192 (2012).
34. Zhu, Y., Rowley, M. J., Böhmdorfer, G. & Wierzbicki, A. T. A SWI/SNF Chromatin-Remodeling Complex Acts in Noncoding RNA-Mediated Transcriptional Silencing. *Mol. Cell* **49,** 1–12 (2013).
35. Hartzog, G. A. & Fu, J. The Spt4-Spt5 complex: a multi-faceted regulator of transcription elongation. *Biochim. Biophys. Acta* **1829,** 105–115 (2013).
36. Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 2926–2931 (2010).
37. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* **12,** 7–18 (2011).
38. Rowley, M. J., Avrutsky, M. I., Sifuentes, C. J., Pereira, L. & Wierzbicki, A. T. Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. *PLoS Genet.* **7,** e1002120 (2011).
39. Cantone, I. & Fisher, A. G. Epigenetic programming and reprogramming during development. *Nat. Struct. Mol. Biol.* **20,** 282–289 (2013).
40. Pfluger, J. & Wagner, D. Histone modifications and dynamic regulation of genome accessibility in plants. *Curr. Opin. Plant Biol.* **10,** 645–652 (2007).
41. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11,** 204–220 (2010).
42. Dowen, R. H. *et al.* Widespread dynamic DNA methylation in response to biotic stress. *Proc. Natl. Acad. Sci. U.S.A.* **109,** E2183–2191 (2012).
43. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14,** 204–220 (2013).
44. Dyachenko, O. V., Schevchuk, T. V., Kretzner, L., Buryanov, Y. I. & Smith, S. S. Human non-CG methylation: are human stem cells plant-like? *Epigenetics* **5,** 569–572 (2010).
45. Bostick, M. *et al.* UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science* **317,** 1760–1764 (2007).
46. Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y. & Shirakawa, M. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* **455,** 818–821 (2008).

47. Liu, X. *et al.* UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. *Nat Commun* **4,** 1563 (2013).
48. Wierzbicki, A. T. The role of long non-coding RNA in transcriptional gene silencing. *Curr. Opin. Plant Biol.* **15,** 517–522 (2012).
49. Zheng, Q. *et al.* RNA polymerase V targets transcriptional silencing components to promoters of protein-coding genes. *Plant J.* (2012). doi:10.1111/tpj.12034
50. Robertson, K. D. DNA methylation and human disease. *Nature Reviews Genetics* **6,** 597–610 (2005).
51. Matzke, M. A. & Mosher, R. A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15,** 394–408 (2014).
52. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* **10,** 295–304 (2009).
53. Naumova, N. *et al.* Organization of the mitotic chromosome. *Science* **342,** 948–953 (2013).
54. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148,** 458–472 (2012).
55. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489,** 109–113 (2012).
56. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
57. Belton, J.-M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58,** 268–276 (2012).
58. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503,** 290–294 (2013).
59. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).
60. Van Bortle, K. *et al.* Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. *Genome Res.* **22,** 2176–2187 (2012).
61. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465,** 363–367 (2010).
62. Feng, S. *et al.* Genome-wide Hi-C Analyses in Wild-Type and Mutants Reveal High-Resolution Chromatin Interactions in Arabidopsis. *Molecular Cell* **55,** 694–707 (2014).
63. Yang, J., Ramos, E. & Corces, V. G. The BEAF-32 insulator coordinates genome organization and function during the evolution of Drosophila species. *Genome Res.* **22,** 2199–2207 (2012).
64. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137,** 1194–1211 (2009).
65. Grob, S., Schmid, M. W., Luedtke, N. W., Wicker, T. & Grossniklaus, U. Characterization of chromosomal architecture in Arabidopsis by chromosome conformation capture. *Genome Biol.* **14,** R129 (2013).
66. Rieder, D., Trajanoski, Z. & McNally, J. G. Transcription factories. *Frontiers in Genetics* **3,** (2012).
67. Jackson, D. A., Hassan, A. B., Errington, R. J. & Cook, P. R. Visualization of focal sites of transcription within human nuclei. *EMBO J.* **12,** 1059–1065 (1993).

68. Guthold, M. *et al.* Direct observation of one-dimensional diffusion and transcription by Escherichia coli RNA polymerase. *Biophys. J.* **77,** 2284–2294 (1999).
69. Kabata, H. *et al.* Visualization of single molecules of RNA polymerase sliding along DNA. *Science* **262,** 1561–1563 (1993).
70. Geisler, S. & Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.* **14,** 699–712 (2013).
71. Berretta, J. & Morillon, A. Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO reports* **10,** 973–982 (2009).
72. Filipowicz, W., Jaskiewicz, L., Kolb, F. A. & Pillai, R. S. Post-transcriptional gene silencing by siRNAs and miRNAs. *Curr. Opin. Struct. Biol.* **15,** 331–341 (2005).
73. Kim, V. N., Han, J. & Siomi, M. C. Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology* **10,** 126–139 (2009).
74. Moazed, D. Small RNAs in transcriptional gene silencing and genome defence. *Nature* **457,** 413–420 (2009).
75. Zhong, X. *et al.* DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. *Nat. Struct. Mol. Biol.* **19,** 870–875 (2012).
76. Wierzbicki, A. T. *et al.* Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. *Genes Dev.* **26,** 1825–1836 (2012).
77. Ørom, U. A. & Shiekhattar, R. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell* **154,** 1190–1193 (2013).
78. Jiao, A. L. & Slack, F. J. RNA-mediated gene activation. *Epigenetics* **9,** (2013).
79. Lai, F. *et al.* Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* **494,** 497–501 (2013).
80. Sabin, L. R., Delás, M. J. & Hannon, G. J. Dogma derailed: the many influences of RNA on the genome. *Mol. Cell* **49,** 783–794 (2013).
81. Jeon, Y., Sarma, K. & Lee, J. T. New and Xisting regulatory mechanisms of X chromosome inactivation. *Curr. Opin. Genet. Dev.* **22,** 62–71 (2012).
82. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135,** 635–648 (2008).
83. Haag, J. R. & Pikaard, C. S. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nat. Rev. Mol. Cell Biol.* **12,** 483–492 (2011).
84. Law, J. A. *et al.* A protein complex required for polymerase V transcripts and RNA-directed DNA methylation in Arabidopsis. *Curr. Biol.* **20,** 951–956 (2010).
85. Stroud, H., Greenberg, M. V. C., Feng, S., Bernatavichute, Y. V. & Jacobsen, S. E. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. *Cell* **152,** 352–364 (2013).
86. Kuhlmann, M. & Mette, M. F. Developmentally non-redundant SET domain proteins SUVH2 and SUVH9 are required for transcriptional gene silencing in Arabidopsis thaliana. *Plant Mol. Biol.* **79,** 623–633 (2012).
87. Haag, J. R. *et al.* In vitro transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. *Mol. Cell* **48,** 811–818 (2012).

88. Law, J. A., Vashisht, A. A., Wohlschlegel, J. A. & Jacobsen, S. E. SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. *PLoS Genet.* **7,** e1002195 (2011).
89. Henderson, I. R. *et al.* Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nature Genetics* **38,** 721–725 (2006).
90. Yang, Z. HEN1 recognizes 21-24 nt small RNA duplexes and deposits a methyl group onto the 2' OH of the 3' terminal nucleotide. *Nucleic Acids Research* **34,** 667–675 (2006).
91. El-Shami, M. *et al.* Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components. *Genes & Development* **21,** 2539–2544 (2007).
92. He, X.-J. *et al.* An effector of RNA-directed DNA methylation in arabidopsis is an ARGONAUTE 4- and RNA-binding protein. *Cell* **137,** 498–508 (2009).
93. Ausin, I., Mockler, T. C., Chory, J. & Jacobsen, S. E. IDN1 and IDN2 are required for de novo DNA methylation in Arabidopsis thaliana. *Nat. Struct. Mol. Biol.* **16,** 1325–1327 (2009).
94. Böhmdorfer, G. *et al.* RNA-directed DNA methylation requires stepwise binding of silencing factors to long non-coding RNA. *Plant J.* **79,** 181–191 (2014).
95. Zhang, C.-J. *et al.* IDN2 and its paralogs form a complex required for RNA-directed DNA methylation. *PLoS Genet.* **8,** e1002693 (2012).
96. Wei, W. *et al.* A role for small RNAs in DNA double-strand break repair. *Cell* **149,** 101–112 (2012).
97. Popova, O. V., Dinh, H. Q., Aufsatz, W. & Jonak, C. The RdDM pathway is required for basal heat tolerance in Arabidopsis. *Mol Plant* **6,** 396–410 (2013).
98. López, A., Ramírez, V., García-Andrade, J., Flors, V. & Vera, P. The RNA silencing enzyme RNA polymerase v is required for plant immunity. *PLoS Genet.* **7,** e1002434 (2011).

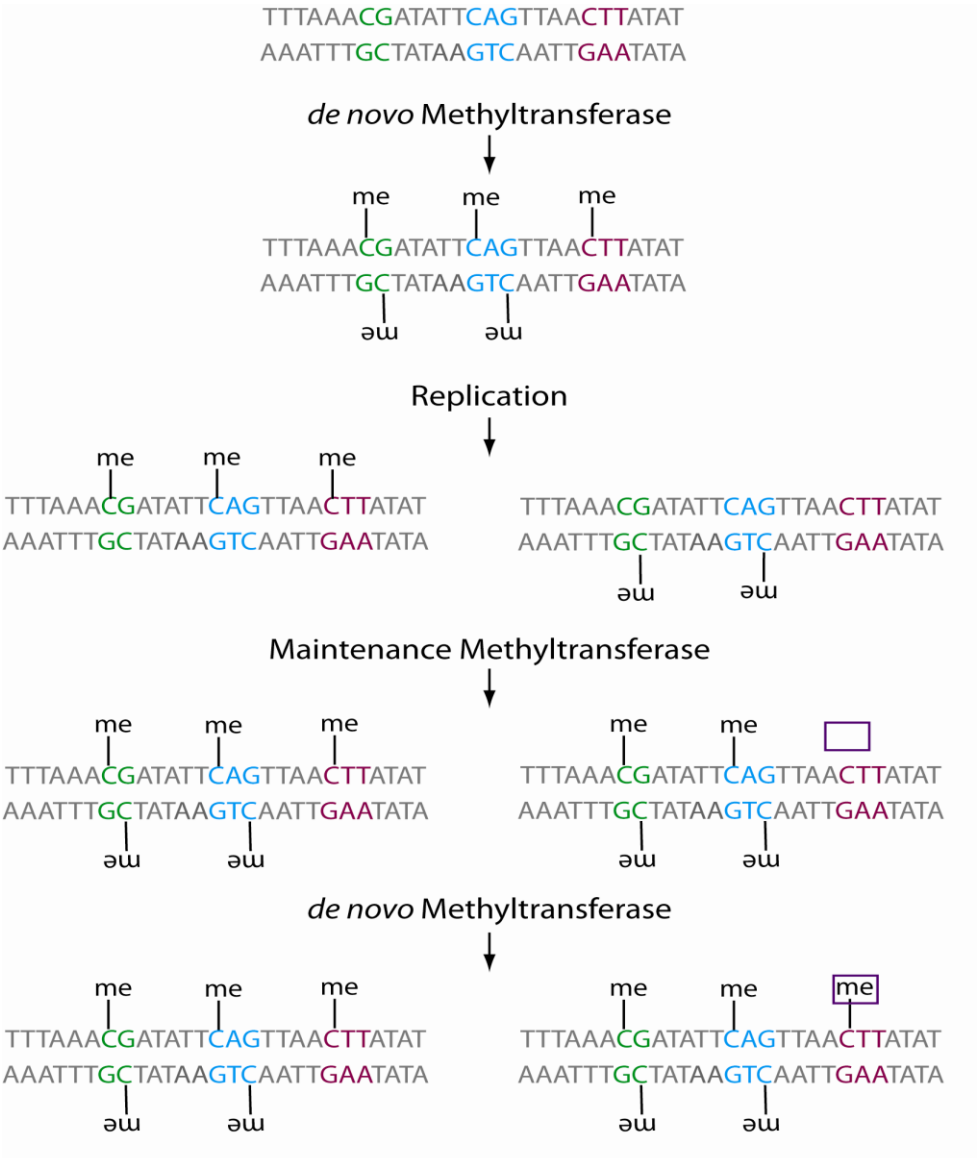


Figure 1.1 DNA methylation in different contexts
Cytosines can be found in three different sequence contexts: CG (green), CHG (blue), or CHH (red) (H= A, T, or C). Once DNA methylation is established CG and CHG methylation can be maintained based on the symmetry of the sequence. CHH methylation is asymmetric and thus requires *de novo* methylation after each round of replication.

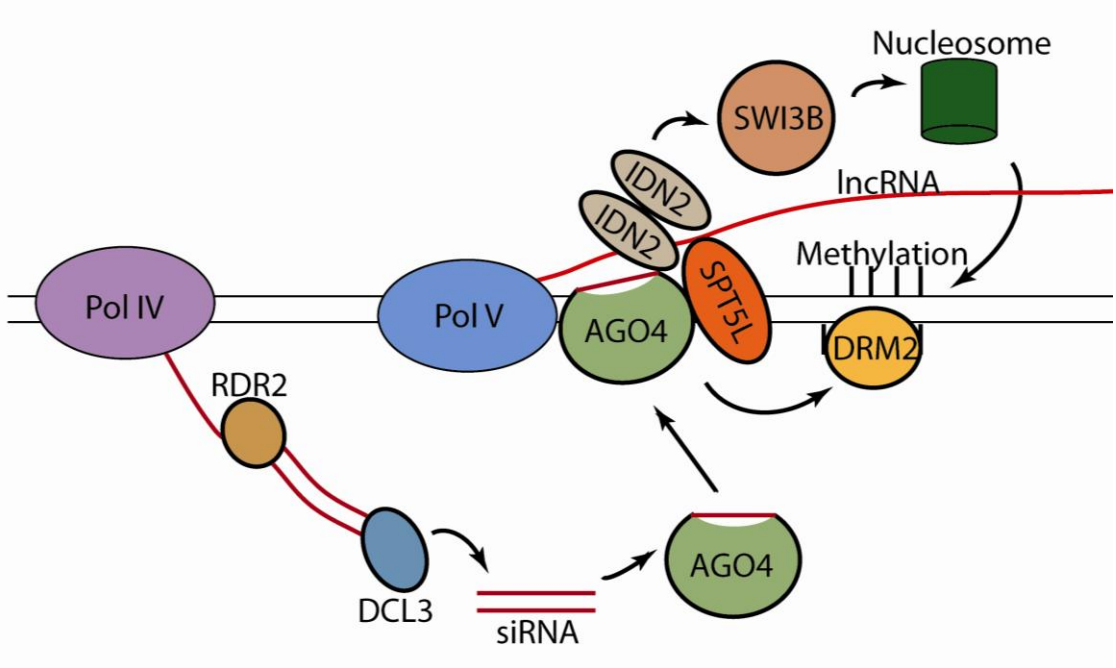


Figure 1.2 Model of Transcriptional Gene Silencing (TGS) in *Arabidopsis thaliana*

Transcriptional Gene Silencing (TGS) has many components whose roles have been discovered. Pol IV (purple) produces transcripts which are made double-stranded by RDR2 (gold). Double stranded RNA is then cleaved by DCL3 (pale blue) into 24 nt siRNA. Pol V (light blue) produces long non coding RNA (lncRNA) which acts as a scaffold for several proteins. AGO4 (pale green) binds siRNA and lncRNA and may interact directly with Pol V. SPT5L (orange) is also dependent on Pol V and binds chromatin independent of AGO4. IDN2 (beige) forms homodimers and binds lncRNA dependent on AGO4. IDN2 interacts with SWI3B (pale orange), a chromatin remodeling enzyme which positions nucleosomes at these loci. Many of these factors such as AGO4 and SPT5L are also important for *de novo* DNA methylation placed by DRM2 (yellow).

CHAPTER 2

RNA Polymerase V Targets Transcriptional Silencing Components to Promoters of Protein-Coding Genes

The contents of this chapter were published and featured on the cover of The Plant Journal in 2012. This work was done in partnership with collaborators (Brian Gregory) from the University of Pennsylvania. Gudrun Böhmdorfer performed work shown in Figure 2.3G-I. Davinder Sandhu constructed libraries and handled the initial data processing. Qi Zheng performed work shown in Figure 2.1G, Figure 2.2A,C,D, Figure 2.3A-C, Figure 2.4A-D, Figure 2.5, Figure 2.8, Figure 2.9B, Figure 2.10, Figure 2.11A, Figure 2.12A,B, and Figure 2.13. I generated samples for sequencing and performed all other experiments and data analysis shown in this chapter.

Abstract

Transcriptional gene silencing controls transposons and other repetitive elements through RNA-directed DNA methylation (RdDM) and heterochromatin formation. A key component of the *Arabidopsis* RdDM pathway is ARGONAUTE4 (AGO4), which associates with siRNAs to mediate DNA methylation. Here, we show that AGO4 preferentially targets transposable elements embedded within promoters of protein-coding genes. We find that this pattern of AGO4 binding cannot be simply explained by the sequences of AGO4-bound siRNAs, but instead AGO4 binding to specific gene promoters is also mediated by long non-coding RNAs (lncRNAs) produced by RNA Polymerase V. lncRNA-mediated AGO4 binding to gene promoters directs asymmetric DNA methylation to these genomic regions and is involved in regulating the expression of targeted genes. Finally, AGO4-binding overlaps sites of DNA methylation affected by biotic stress response. Based on these findings, we propose that targets of AGO4-directed RdDM are regulatory units responsible for controlling gene expression under specific environmental conditions.

Introduction

Transcriptional gene silencing is mediated by repressive chromatin modifications directed to transposable elements and other repetitive sequences to prevent their expression, which if uncontrolled could result in detrimental effects to the cell. In eukaryotic organisms, the primary factors driving the functional mechanism of silencing are the conserved Argonaute proteins [1]. In *Arabidopsis thaliana*, the RNA-mediated transcriptional gene silencing pathway (also known as RNA-mediated DNA methylation; RdDM) is mediated by ARGONAUTE4 (AGO4) [2]. Specific genomic localization of AGO4 has been hypothesized to require the joint activity of two classes of non-coding RNAs. The first is small interfering RNAs (siRNAs), which are produced by the activities of RNA Polymerase IV (Pol IV), RNA-DEPENDENT RNA POLYMERASE 2 (RDR2) and DICER-LIKE 3 (DCL3) [3]. siRNAs bind AGO4 and provide sequence-specificity [4] through direct base-pairing interactions with complementary loci. The other class of non-coding RNAs involved in targeting AGO4 to specific genomic loci is likely long non-coding RNA (lncRNA) produced by a plant-specific RNA Polymerase V (Pol V) [5] with some involvement of RNA Polymerase II [6]. Pol V-produced lncRNAs have been proposed to act as binding scaffolds for AGO4-siRNA complexes [7,8]. Upon binding to chromatin, AGO4 is believed to work with at least one more RNA-associated protein (SPT5L/KTF1) [9], guide *de novo* DNA methyltransferase DRM2 and thereby mediate DNA methylation primarily in CHH contexts [10].

Little is known about the genome-wide distribution of AGO4 or other RdDM components or the mechanisms that direct them to specific loci. It is also unknown to what extent the RdDM pathway controls expression of protein-coding genes involved in specific biological processes. To answer these questions we characterized the genome-wide distribution of AGO4 binding to chromatin. We found that AGO4 preferentially targets promoters of protein-coding genes. This specific binding pattern cannot be explained by the sequences of AGO4-associated small RNAs and seems to be primarily mediated by Pol V-produced lncRNAs. AGO4 binding to gene promoters mediates CHH methylation and in some cases affects expression levels of genes controlled by these promoters. Moreover, AGO4 binding overlaps DNA methylation affected by biotic stress response. This combination of results leads to the intriguing hypothesis that AGO4

binding sites are regulatory units controlling gene expression under specific environmental conditions.

Results

AGO4 has no preference towards TE-rich pericentromeric regions

As the first step towards explaining the mechanism by which AGO4 directs RdDM-mediated silencing to specific loci, we assayed the genome-wide distribution of AGO4 binding targets using chromatin immunoprecipitation with an anti-AGO4 polyclonal antibody followed by high-throughput sequencing (ChIP-seq). Using a combinatorial comparison approach, where ChIP-seq samples from Col-0 wild type were compared to those from *ago4* mutant as well as to input sample controls, we identified 820 AGO4 binding regions (referred to also as peaks; Figs. 2.1A – C and 2.7). We used ChIP followed by real-time PCR (ChIP-qPCR) to validate 24 AGO4 binding regions. Binding was confirmed at all 11 tested regions ranked in top 20% by the peak calling algorithm (Figs. 2.1D and 2.7B), all 3 tested regions ranked in the middle 60% (Figs. 2.1E and 2.7D) and 10 out of 13 tested regions ranked in bottom 20% (Figs. 2.1F and 2.7F). Additionally, most previously known AGO4 targets [7,9] displayed evidence of strong AGO4 binding in our ChIP-seq, however only *IGN25* met the stringent criteria to be included on the list of significant AGO4 chromatin binding sites. In total, these results indicate that our analysis has a high stringency with low proportion of false positives even among the lowest ranking AGO4 binding sites.

The *Arabidopsis* RNA-mediated transcriptional gene silencing pathway targets mostly transposable elements and other repetitive sequences [3]. Because most TEs in the *Arabidopsis* genome cluster within pericentromeric regions [11], silencing components may also be expected to be enriched around the centromeres. To test whether the genome-wide AGO4 chromatin binding data confirm this prediction, we mapped AGO4 peaks onto the five nuclear chromosomes of *Arabidopsis.* Surprisingly, we found AGO4 peaks to be distributed evenly across all five chromosomes and their density to be comparable in both TE-rich pericentromeric regions as well as gene-rich chromosome arms (Figs. 2.1G and 2.8). Therefore, genome-wide identification of significant AGO4

binding sites indicates that this protein is not preferentially targeted to large heterochromatic and repetitive genomic domains.

AGO4 binds TEs within gene promoters

Widespread AGO4 binding within gene-rich chromosome arms often overlapped protein coding genes (Figs. 2.1G and 2.8), suggesting that AGO4 binding may be enriched on genes. To test this possibility, we classified AGO4 peaks based on overlaps with annotated genomic features. AGO4 was not enriched on the transcribed regions of protein-coding genes (Fig. 2.2A), instead we observed a significant enrichment on gene promoters defined as 1 kb regions upstream of transcription start sites ($p < 0.001$; Fig. 2.2A) with 64% of all AGO4 peaks mapping to promoters of protein-coding genes. This pattern was confirmed by profiling ChIP-seq signal around transcription start sites, which revealed preferential AGO4 binding in the approximate region between -500 and -200 upstream of target gene transcription start sites (TSSs) (Fig. 2.2B). Moreover, AGO4 peaks were also depleted in nucleosomes (Fig. 2.9A-B), the absence of nucleosomes being a characteristic feature of gene promoters [12]. These results demonstrate that AGO4 is preferentially binding promoters of protein-coding genes.

AGO4 binding was also enriched on transposable elements and tandem repeats ($p \rightarrow 0$; Fig. 2.2A). This enrichment was significant on most Class I and Class II transposable elements (Fig. 2.2C). Interestingly, AGO4 binding was significantly depleted on En-Spm DNA transposons as well as Copia long terminal repeat (LTR) retrotransposons (Fig. 2.2C), both of which are enriched within coding sequences of protein-coding genes [13]. AGO4 binding was also depleted on Gypsy LTR retrotransposons (Fig. 2.2C). These results reveal that AGO4 has a preference towards specific families of transposable elements in the *Arabidopsis* genome.

Further analysis of AGO4 binding to gene promoters revealed that out of 528 AGO4 binding regions identified within gene promoters, 362 (69%) overlapped transposable elements (Fig. 2.2D). In contrast, in a comparable set of random control regions 436 mapped to gene promoters and only 163 (37%) of them overlapped transposable elements (Fig. 2.2E, $p < 7x10^{-22}$), indicating that AGO4 binding to gene promoters does not reflect preferential insertions of TEs into promoter regions. These results demonstrate that AGO4 binding shows a significant preference for both gene

promoters and transposable elements. Together, our findings reveal that AGO4 preferentially binds transposons embedded within promoters of protein-coding genes.

AGO4 binding pattern is mediated by lncRNA

Sequence specificity of AGO4 binding to chromatin has been proposed to be directed by the sequences of incorporated 24 nt siRNAs [4]. To test whether 24nt siRNAs have a function in directing AGO4 to TEs within promoters of protein coding genes, we mapped AGO4-bound siRNAs [14] to AGO4 peaks. We found that AGO4-associated 24 nt siRNAs are enriched on AGO4 peaks (Fig. 2.3A). As controls, similar analyses with AGO1-bound 24 nt siRNAs [14] demonstrated only minimal enrichment, and 21 nt siRNAs bound by either AGO protein revealed negligible enrichment on AGO4 peaks (Fig. 2.3A). Consistent with these findings, we observed that the total population of 24 nt but not 21 nt siRNAs, which are implicated in posttranscriptional silencing [1], was enriched on AGO4 peaks (Fig. 2.10A). Furthermore, only 10% of AGO4 peaks were found to have little or no association with AGO4-bound siRNAs. These results suggest that AGO4 binding to chromatin is correlated with the presence of 24nt siRNAs, which likely have a function in guiding AGO4 to specific genomic loci.

To further test whether the sequences of siRNAs are able to explain the specific pattern of AGO4 binding to chromatin, we mapped AGO4-associated 24 nt siRNAs onto the five nuclear chromosomes of *Arabidopsis.* Surprisingly, we found these siRNAs to be strongly enriched within TE-rich pericentromeric regions and much less abundant within gene-rich chromosome arms (Figs. 2.3B and 2.10B). Therefore, AGO4-associated siRNAs are not solely responsible for targeting AGO4 to its DNA interaction sites. This is consistent with the model where 24 nt siRNAs are necessary but not sufficient for mediating AGO4 binding to specific loci.

Another factor previously implicated in AGO4 binding to specific genomic loci is transcription by Pol V, which has been proposed to provide lncRNA scaffolds for AGO4 binding to chromatin [7]. To test if Pol V is required for genome-wide targeting of AGO4, we performed ChIP-seq using anti-AGO4 antibody on *nrpe1* mutant plants, which are deficient for the largest subunit of Pol V. By comparing the ChIP-seq datasets from *nrpe1* mutant plants to Col-0 wild type and *ago4* we tested if AGO4 binding to specific loci requires Pol V. Surprisingly, we identified only 7 (0.85%) Pol V-independent AGO4

peaks and 41 (4.96%) binding sites that demonstrated intermediate levels of AGO4 binding in *nrpe1* mutant plants (Fig. 2.3C). We also analyzed Pol V-dependence by comparing normalized read counts of AGO4 binding sites in Col-0 wild type to the *nrpe1* mutant, which confirmed that the vast majority of sites have ChIP signal strongly reduced in the *nrpe1* mutant (Fig. 2.11A). These results indicate that Pol V is generally required for AGO4 binding to chromatin. The small fraction of Pol V-independent peaks or differences in AGO4 chromatin interaction strength may reflect a minor Pol V-independent mechanism of AGO4 binding or alternatively that this type of targeting is not actually biologically significant. Importance of Pol V for AGO4 binding to chromatin was further supported by our ChIP-qPCR validation, which demonstrated that AGO4 binding to all validated loci is dependent on Pol V (Figs. 2.1D – F and 2.7B, D, and F). Furthermore, all 11 tested high ranking loci, 3 middle ranking loci, and 9 low ranking loci show detectable Pol V binding by ChIP-qPCR with anti-NRPE1 antibody (Figs. 2.3D – F and 2.11B – D). Importantly, hitherto undetected AGO4 binding sites showed evidence of Pol V-dependent transcription (Figs. 2.3G – I), indicating that Pol V produces lncRNA at these loci.

In total, these results demonstrate that Pol V is required for AGO4 binding to most if not all of its target loci. Furthermore, our observations of 1) a strong preference for gene promoter binding by AGO4, 2) the lack of concordance between AGO4 interaction sites and siRNA sequences bound by this protein, and 3) Pol V transcription within AGO4 promoter bound regions, suggest that lncRNAs produced by Pol V are also a critical factor in mediating the interaction of AGO4 with promoters of specific protein-coding genes.

lncRNA-mediated AGO4 binding controls gene activity

Our observation of lncRNA-mediated AGO4 binding to promoters of protein-coding genes suggests that non-coding transcription and AGO4 binding may control the expression levels of the targeted genes by mediating DNA methylation. To test this possibility, we first examined whether AGO4 binding was correlated with DNA methylation [15]. AGO4 peaks were significantly enriched (p -> 0) in CHH methylation relative to the genome-wide level and also demonstrated a less pronounced enrichment in CG and CHG methylation (Fig. 2.4A). A similar pattern of DNA methylation coincident

with AGO4 binding regions was also present in *ros1 dml2 dml3* triple mutant plants that are deficient in three DNA demethylases [15] (Fig. 2.12A). Significant enrichment in CHH methylation within AGO4-bound regions was also present in *met1*, a mutant of the major CG methyltransferase of *Arabidopsis* [15] (Fig. 2.12B). However, in *drm1 drm2 cmt3* triple mutant plants [15] CHH methylation of AGO4-bound sites was strongly reduced relative to Col-0 wild type (Fig. 2.4B), suggesting that this methylation is established by the *de novo* methyltransferase DRM2, although involvement of CMT3 cannot be excluded. We also found that DNA methylation within AGO4 binding sites was most prominent on transposable elements embedded within promoters of protein-coding genes (Fig. 2.4C). Additionally, CHH methylation within AGO4 peaks was significantly reduced (p -> 0) in the *nrpe1* mutant relative to wild type [16] (Fig. 2.4D). Together, these results demonstrate that AGO4 binding is correlated primarily with CHH methylation and predicts that AGO4 recruitment to specific genomic loci, including TEs in gene promoters, likely mediates their CHH methylation.

To test this prediction, we probed DNA methylation levels on 26 AGO4-bound promoter regions in Col-0 wild type as well as *ago4* and *nrpe1* mutants. Digestion with methylation-sensitive restriction endonucleases followed by PCR revealed that tested AGO4-bound promoter regions contain CHH methylation, which was strongly reduced in both *nrpe1* and *ago4* mutants (Figs. 2.4E and 2.12C). Taken together, these results indicate that lncRNA-mediated AGO4 binding in gene promoters is directing CHH methylation, and suggest this might control transcription of those genes.

To test whether lncRNA-mediated AGO4 binding within gene promoters affects expression of proximal genes, we screened 41 genes with AGO4 peaks in their promoter regions for significant expression changes in *nrpe1* and *ago4* mutants. Real-time RT-PCR identified three genes upregulated in *nrpe1* and *ago4* mutants and two downregulated in *nrpe1* and *ago4* mutants (Figs. 2.4F – J). These results demonstrate that expression of at least a subset of AGO4-bound genes is affected by AGO4 and Pol V under standard growth conditions, revealing that lncRNA-mediated AGO4 binding within gene promoters is capable of affecting gene expression. One of the genes where RNA accumulation was reduced in *nrpe1* and *ago4* mutants under standard growth conditions is *ROS1* (*AT2G36490*; Fig. 2.4I), which encodes a DNA demethylase, which

has previously been shown to be positively regulated by CG DNA methylation [17]. This suggests the presence of a compensatory mechanism, where a reduction in CG methylation or RNA-directed CHH methylation results in reduction of DNA demethylase production to prevent excessive loss of DNA methylation. In total, these results demonstrate that AGO4 binding within promoter regions is capable of controlling the expression of targeted genes.

AGO4 binding is correlated with DNA methylation affected by biotic stress responses

Our observation that only five out of 41 tested AGO4-associated genes are affected in *ago4* and *nrpe1* mutants is consistent with the lack of morphological phenotypes associated with *Arabidopsis ago4* and *nrpe1* mutant plants grown under optimal conditions [2,18,19]. To test if AGO4 target genes may be controlled in response to stress, we performed Gene Ontology (GO) analysis, which revealed significant enrichment of genes responsive to biotic and abiotic stimuli (Fig. 2.5A). To test if DNA methylation levels at AGO4 binding sites may be affected by stress we calculated the average changes in DNA methylation levels at differentially methylated regions (DMRs) identified in plants subjected to biotic stressors [20]. Differential methylation was significantly enriched on AGO4 binding sites relative to the genome overall (Fig. 2.5B). In fact, stress-responsive differential CHH methylation was 8-fold more pronounced on AGO4 binding sites than on genome overall ($p \to 0$). This was much higher than the 3.5-fold enrichment of total CHH methylation on AGO4 binding sites (Fig. 2.4A). These results suggest that enrichment of stress-induced differential methylation on AGO4 interaction regions is not merely a byproduct of overall higher levels of DNA methylation in these genomic sites. This was further confirmed by the observation that AGO4 binding sites significantly overlapped salicylic acid-induced CHH DMRs compared to 1000 random genomic permutations and vice versa (Figs. 2.13A and 2.13B; $p < 0.001$ for both comparisons). These results demonstrate that a significant fraction of AGO4 binding sites contains DNA methylation that can be dynamically regulated during the plant's response to biotic stresses. Taken together, these findings suggest that changes in DNA methylation patterns at AGO4 target genes are a part of a natural gene regulatory mechanism during plant biotic stress responses.

Discussion

Argonaute proteins have been shown to recognize the sequences of specific target RNAs and genomic loci using incorporated small RNAs [4]. Our findings are consistent with 24 nt siRNAs being required for AGO4 binding to chromatin, but also demonstrate that they are not sufficient. Instead, lncRNAs produced by Pol V mediate the specific binding of AGO4 to its genomic targets, many of which are transposons embedded within the promoters of protein-coding genes. Intriguingly, these results suggest that widespread AGO4-bound transposons within gene promoters may be controlling elements as proposed by Barbara McClintock [21], and identify Pol V-produced lncRNAs as the primary determinant of their status as regulatory modules.

Once the overlapping action of 24 nt siRNAs and lncRNAs guides AGO4 to specific genomic regions, chromatin modifying enzymes are recruited and repressive DNA and histone modifications are established. These modifications in turn affect gene expression. A possible mechanism by which RdDM controls gene expression is by affecting the binding of transcription factors or other DNA-binding proteins to *cis*-elements within promoters (Fig. 2.6). This possibility is consistent with our data showing both up- and down-regulation of AGO4-controlled genes in *ago4* mutant plants reflecting either repressive or activating transcription factors being affected by DNA methylation, respectively. It is however also possible that AGO4 binding and RdDM may affect the spreading of chromatin modifications [22] or RNA processing. In addition to serving as switchable regulatory elements controlled by DNA methylation status, AGO4-targeted transposable elements may also insert into novel locations providing an additional level of transcription regulation relative to the pre-insertion promoter sequence. Our model predicts that pericentromeric silenced genomic regions which are not bound by AGO4 but give rise to siRNAs, are not transcribed by Pol V. Instead they are likely targeted by a different transcriptional silencing pathway.

Our work provides direct evidence of preferential binding of an RdDM component to promoters of protein coding genes. It is consistent with immunostaining data showing the presence of AGO4 outside of chromocenters [23,24], with the preferential upregulation of euchromatic genes in *drm1 drm2 cmt3* triple mutant [25], and with the presence of some well characterized RdDM targets in euchromatin [26,27]. Targeting of AGO4 towards

promoters of protein-coding genes also reveals an additional level of gene expression control, likely conserved between plants and animals [22,28]. It is interesting that only minimal morphological phenotypes are observed in *Arabidopsis* RdDM mutant plants grown under standard conditions [2,18], which suggests that this mechanism may be more prevalent in organisms with higher transposon content, like maize where disruption of RdDM results in more dramatic phenotypes [29,30]. This mechanism may also have much greater impact in plants like tomato, where the majority of 24nt siRNAs map to gene-rich chromosomal regions [31]. AGO4-mediated control of gene expression may also work in certain developmental stages as suggested in early embryonic development [32,33] or provide a common response to environmental stimuli [34–36] (Fig. 2.5).

A potential involvement of RdDM-targeted TEs in response to environmental stimuli is supported by our observations that AGO4 binding sites significantly overlap genomic regions, where biotic stresses have been demonstrated to affect DNA methylation levels [20] (Fig. 2.5B). Thus, our findings likely provide an explanation for earlier reports showing the involvement of AGO4 and Pol V in pathogen responses [36,37]. We propose that pathogen infection affects siRNA production and/or Pol V transcription, which in turn causes changes in promoter DNA methylation and affects gene expression levels. In conclusion, our findings establish further exploration of the regulatory functions of AGO4 and the entire RdDM pathway in normal plant development and stress responses as an important goal for future research.

Materials and Methods

Plant material

Arabidopsis thaliana nrpe1 (*nrpd1b-11*) and *ago4* (*ago4-1* [2]) introgressed into the Col-0 background) were described previously [7,38]. Plants were cultivated at 22 ˚C under long day conditions (16 h day, 8 h night).

RNA Analysis

For assays of mRNA accumulation, total RNA was extracted from 2-3 week old plants using RNeasy Plant Mini Kit (Qiagen) and three biological replicates were amplified using SuperScript III Platinum SYBR Green One-Step qRT-PCR Kit (Invitrogen) in an Applied Biosystems 7500 real time PCR machine. For assays of Pol V transcript

accumulation, total RNA was extracted from 2-3 week old plants using RNeasy Plant Mini Kit (Qiagen) and assayed as described [5], except that random primers were used and cDNA was amplified in a BioRad CFX Connect real time PCR machine. Two independent biological repeats were performed.

DNA Methylation Analysis

Genomic DNA was extracted from above-ground tissue of 2-week-old plants using DNeasy Plant Mini Kit (Qiagen). 100 ng of genomic DNA was digested with 10 units of *Alu*I, *Dde*I, or *Sau*3AI restriction enzymes (NEB) for 20 min. After heat-inactivation of the enzyme, DNA was amplified using 0.75 units Platinum Taq (Invitrogen).

Antibodies

Affinity-purified rabbit polyclonal anti-AGO4 and anti-NRPE1 antibodies were described previously [7,39].

Chromatin Immunoprecipitation

ChIP was performed as described [9] with slight modifications. A detailed protocol is provided in the Appendix B.

ChIP-seq library preparation and sequencing

All ChIP-seq and input libraries were prepared according to the Illumina ChIP-seq library preparation protocol, and subjected to sequencing on a Genome Analyzer IIx as per manufacturer's instructions (Illumina, La Jolla, CA).

Sequence processing

Raw reads were pre-processed and mapped to the *Arabidopsis* genome using a pipeline as previously described [40] with slight modifications. Specifically, we used the Bowtie program [41] instead of the original cross_match aligner. All valid alignments were reported so as to tolerate non-uniquely mapping reads, since AGO4 is thought to target heterochromatin and repetitive elements in *Arabidopsis*. A detailed procedure is provided in the Appendix B.

AGO4 binding site identification

AGO4-bound peaks (AGO4 binding regions) were called using the CSAR R package [42]. To do this, all mapped reads were extended up to 250 nt and merged from both strands. Then, peaks were required to reach a significant fold-enrichment between test and control with an FDR < 0.05. To identify high-quality peaks with minimum false positives,

5 sets of peaks were called either between ChIP and input samples ("traditional calls") or between Col-0 and *ago4* or *nrpe1* mutants (“direct-comparison”) as our basis for defining substantial peaks. Then Pol V-dependent and Pol V-independent peaks were determined by “peak-arithmetic” manipulations, which reliably identify peaks enriched for both ChIP vs. input and wild-type vs. mutant comparisons. See Appendix B for descriptions of these manipulations.

An additional filtering step was also implemented to exclude peaks with a potential ecotype bias, because the *ago4* mutant plants used in this study were originally identified [2] in the Landsberg (Ler-1) ecotype of *Arabidopsis* and subsequently back-crossed to Col-0 plants three successive times. To do this, any peak that either 1) cannot be mapped to the Ler-1 draft genome (Ler-1 unmappable) or 2) can be better mapped to the Ler-1 draft genome (Ler-1 better mapped) were discarded from further analysis.

To distinguish the AGO4 peaks that are completely dependent from those that are partially dependent on Pol V activity, we determined if the clone-abundance of AGO4 binding sites is comparable (less than 2 fold difference) between *nrpe1* ChIP and *ago4* ChIP samples (Pol V-dependent) or not (Pol V partially dependent). The vast majority of Pol V-dependent peaks were completely dependent, and therefore we didn’t separate these peaks in further analyses.

AGO4 binding site classification

To classify and annotate AGO4 peaks, their genomic coordinates were compared to various classes of known genetic elements annotated by TAIR (TAIR9 release) on the *Arabidopsis* genome, including protein-coding genes (exons and introns), rRNAs, tRNAs, miRNAs, snoRNAs, snRNAs, ncRNAs, pseudogenes, and transposable elements (TEs). To supplement this analysis, additional repetitive elements were defined using the RepeatMasker program (http://www.repeatmasker.org/). We also defined gene promoters as the regions 1 kb upstream from the transcription start sites (TSS) of protein-coding genes. As a negative control, 1000 sets of random peaks (NC-peaks) were sampled from the genome, classified, and annotated similarly, and the p-values of enrichment or depletion in specific categories were estimated using a bootstrapping method based on these NC-peaks. To comprehensively characterize the

classes and families of transposable elements in AGO4 peaks, we used the RepeatMasker identified TEs and their corresponding annotation information.

To characterize small RNA profiles near AGO4 peaks, published small RNA-IP and total small RNA datasets [14] for both AGO4 and AGO1 from *Arabidopsis* seedlings were used for our analysis; the smRNA-IP or total smRNA reads were searched within the AGO4 peaks as well as their flanking regions (2 kb up- and downstream).

To characterize the cytosine methylation (mC) in AGO4 peaks, we used the published single nucleotide mC datasets datasets including genome-wide mC profiles from Col-0, *met1*, *ddc* and *rdd* mutant plants [15] kindly provided by Dr. Ryan Lister. The mC sites were searched within all AGO4 peaks as well as NC-peaks, and the mC density was calculated and compared between AGO4 peaks and NC-peaks for CG, CHG and CHH methylation or as a whole. The mC density was also directly compared between Col-0 and *nrpe1* mutant plants using recently published mC datasets [16].

To characterize AGO4 binding profiles around TSSs, the log fold change profile of ChIP-seq reads between Col-0 and *ago4* samples was generated using the CEAS program [43] relative to positions in the TSS of all protein-coding genes. Similarly, to characterize the nucleosome profile around AGO4 peaks, we used published MNase-seq datasets [12] and calculated the log fold change of MNase-seq reads between Col-0 and *ago4* samples using the CEAS program [43] relative to positions in the TSS. We also called the well-positioned nucleosomes as previously described [44], and determined the nucleosome density profiles for all or promoter overlapping AGO4 peaks.

GO analysis

To identify significantly enriched biological processes for the genes corresponding to AGO4 bound promoters, the gene IDs of these loci were analyzed using the GOEAST online analysis tool [45] with an FDR < 0.05.

Accession number and genome-browser link

All six ChIP-seq library datasets were deposited into NCBI GEO under the accession GSE35381. The AnnoJ genome-browser for all of the ChIP-seq libraries as well as external datasets (smRNAs and DNA methylation) presented in this paper is publicly available at: http://gregorylab.bio.upenn.edu/annoj_atAGO4/.

Acknowledgements

We thank Eran Pichersky, David Engelke, and Eric Richards for critical reading of the manuscript. This work was supported by National Science Foundation grants MCB 1120271 to Andrzej Wierzbicki and CAREER Award MCB 1053846 to Brian Gregory., Austrian Science Fund (FWF) fellowship J3199-B09 to Gudrun Böhmdorfer and the NIH National Research Service Award #5-T32-GM07544. Andrzej Wierzbicki, Brian Gregory, Jordan Rowley, and Qi Zheng designed the experiments. Gudrun Böhmdorfer performed RT-PCR assays for Pol V transcripts. Jordan Rowley performed all remaining ChIP, RT-PCR and DNA methylation assays and generated profiles shown in Fig. 2.2B and Fig. 2.9A. Davinder Sandhu constructed all high-throughput sequencing libraries and was responsible for initial data management. Qi Zheng performed all remaining bioinformatic analysis.

References

1. Hutvagner, G. & Simard, M. J. Argonaute proteins: key players in RNA silencing. *Nat. Rev. Mol. Cell Biol.* **9,** 22–32 (2008).
2. Zilberman, D., Cao, X. & Jacobsen, S. E. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299,** 716–719 (2003).
3. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11,** 204–220 (2010).
4. Qi, Y. *et al.* Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature* **443,** 1008–1012 (2006).
5. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135,** 635–648 (2008).
6. Zheng, B. *et al.* Intergenic transcription by RNA polymerase II coordinates Pol IV and Pol V in siRNA-directed transcriptional gene silencing in Arabidopsis. *Genes Dev.* **23,** 2850–2860 (2009).
7. Wierzbicki, A. T., Ream, T. S., Haag, J. R. & Pikaard, C. S. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat. Genet.* **41,** 630–634 (2009).
8. Wierzbicki, A. T. The role of long non-coding RNA in transcriptional gene silencing. *Curr. Opin. Plant Biol.* **15,** 517–522 (2012).
9. Rowley, M. J., Avrutsky, M. I., Sifuentes, C. J., Pereira, L. & Wierzbicki, A. T. Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. *PLoS Genet.* **7,** e1002120 (2011).
10. Wierzbicki, A. T. The role of long non-coding RNA in transcriptional gene silencing. *Curr. Opin. Plant Biol.* **15,** 517–522 (2012).

11. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408,** 796–815 (2000).
12. Chodavarapu, R. K. *et al.* Relationship between nucleosome positioning and DNA methylation. *Nature* **466,** 388–392 (2010).
13. Lockton, S. & Gaut, B. S. The contribution of transposable elements to expressed coding sequence in Arabidopsis thaliana. *J. Mol. Evol.* **68,** 80–89 (2009).
14. Wang, H. *et al.* Deep sequencing of small RNAs specifically associated with Arabidopsis AGO1 and AGO4 uncovers new AGO functions. *Plant J.* **67,** 292–304 (2011).
15. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133,** 523–536 (2008).
16. Wierzbicki, A. T. *et al.* Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. *Genes Dev.* **26,** 1825–1836 (2012).
17. Mathieu, O., Reinders, J., Caikovski, M., Smathajitt, C. & Paszkowski, J. Transgenerational stability of the Arabidopsis epigenome is coordinated by CG methylation. *Cell* **130,** 851–862 (2007).
18. Pontier, D. *et al.* Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. *Genes Dev.* **19,** 2030–2040 (2005).
19. Kanno, T. *et al.* Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nat. Genet.* **37,** 761–765 (2005).
20. Dowen, R. H. *et al.* Widespread dynamic DNA methylation in response to biotic stress. *Proc. Natl. Acad. Sci. U.S.A.* **109,** E2183–2191 (2012).
21. MCCLINTOCK, B. Controlling elements and the gene. *Cold Spring Harb. Symp. Quant. Biol.* **21,** 197–216 (1956).
22. Moshkovich, N. *et al.* RNAi-independent role for Argonaute2 in CTCF/CP190 chromatin insulator function. *Genes Dev.* **25,** 1686–1701 (2011).
23. Pontes, O. *et al.* The Arabidopsis chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell* **126,** 79–92 (2006).
24. Li, C. F. *et al.* An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in Arabidopsis thaliana. *Cell* **126,** 93–106 (2006).
25. Zhang, X. *et al.* Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* **126,** 1189–1201 (2006).
26. Henderson, I. R. & Jacobsen, S. E. Tandem repeats upstream of the Arabidopsis endogene SDC recruit non-CG DNA methylation and initiate siRNA spreading. *Genes Dev.* **22,** 1597–1606 (2008).
27. Huettel, B. *et al.* Endogenous targets of RNA-directed DNA methylation and Pol IV in Arabidopsis. *EMBO J.* **25,** 2828–2836 (2006).
28. Cernilogar, F. M. *et al.* Chromatin-associated RNA interference components contribute to transcriptional regulation in Drosophila. *Nature* **480,** 391–395 (2011).
29. Alleman, M. *et al.* An RNA-dependent RNA polymerase is required for paramutation in maize. *Nature* **442,** 295–298 (2006).
30. Erhard, K. F., Jr *et al.* RNA polymerase IV functions in paramutation in Zea mays. *Science* **323,** 1201–1205 (2009).

31. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485,** 635–641 (2012).
32. Autran, D. *et al.* Maternal epigenetic pathways control parental contributions to Arabidopsis early embryogenesis. *Cell* **145,** 707–719 (2011).
33. Mosher, R. A., Schwach, F., Studholme, D. & Baulcombe, D. C. PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 3145–3150 (2008).
34. Pecinka, A. *et al.* Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in Arabidopsis. *Plant Cell* **22,** 3118–3129 (2010).
35. Tittel-Elmer, M. *et al.* Stress-induced activation of heterochromatic transcription. *PLoS Genet.* **6,** e1001175 (2010).
36. Agorio, A. & Vera, P. ARGONAUTE4 is required for resistance to Pseudomonas syringae in Arabidopsis. *Plant Cell* **19,** 3778–3790 (2007).
37. López, A., Ramírez, V., García-Andrade, J., Flors, V. & Vera, P. The RNA silencing enzyme RNA polymerase v is required for plant immunity. *PLoS Genet.* **7,** e1002434 (2011).
38. Onodera, Y. *et al.* Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120,** 613–622 (2005).
39. Ream, T. S. *et al.* Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. *Mol. Cell* **33,** 192–203 (2009).
40. Zheng, Q. *et al.* Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS Genet.* **6,** e1001141 (2010).
41. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
42. Muiño, J. M., Kaufmann, K., van Ham, R. C., Angenent, G. C. & Krajewski, P. ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. *Plant Methods* **7,** 11 (2011).
43. Shin, H., Liu, T., Manrai, A. K. & Liu, X. S. CEAS: cis-regulatory element annotation system. *Bioinformatics* **25,** 2605–2606 (2009).
44. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458,** 362–366 (2009).
45. Zheng, Q. & Wang, X.-J. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.* **36,** W358–363 (2008).

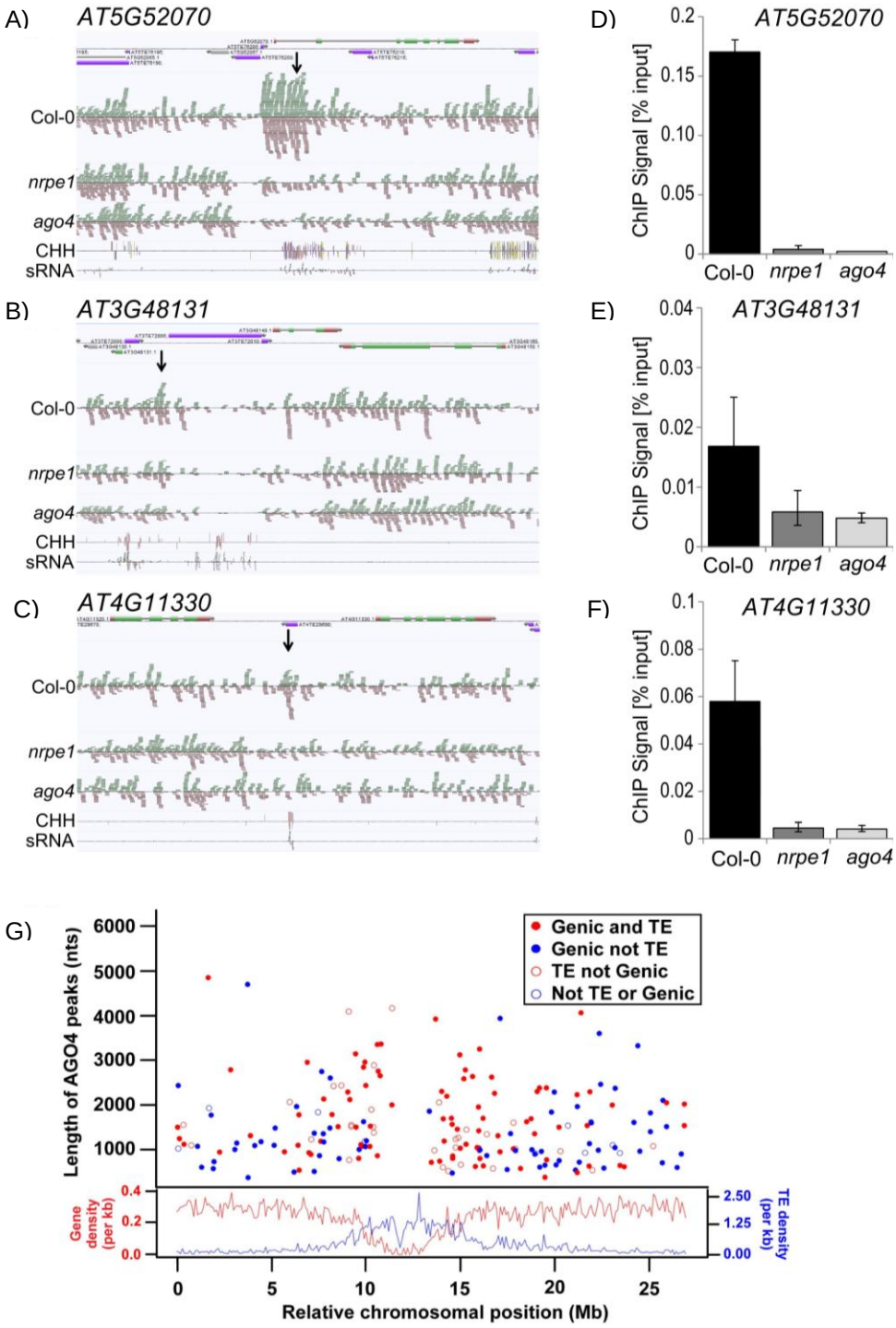


Figure 2.1 Identification of AGO4-bound loci

A – C) Graphical representation of sample AGO4-bound loci identified using ChIP-seq. The genome browser screenshots show from top genome annotation, ChIP-seq sequencing reads from Col-0 wild type, *nrpe1* and *ago4* strains, CHH DNA methylation and total small RNA reads. More loci are shown in Fig. 2.7.

D – F) ChIP-real time PCR validation of AGO4 binding to chromatin on AGO4 peaks identified using ChIP-seq in Col-0 wild type, *nrpe1* and *ago4* mutants. Bars represent

averages from three independent amplifications. Error bars represent standard deviations. More loci are shown in Fig. 2.7.
G) AGO4 binding shows no preference towards transposon-rich pericentromeric regions. The top graph shows the distribution of AGO4 peaks along the length of chromosome 5 vs. peak length in nucleotides (nts). The corresponding category for each colored dot can be found in the legend. The lower graph provides the density of genes (red line and y-axis label) and transposable elements (blue line and y-axis label) along the length of chromosome 5.

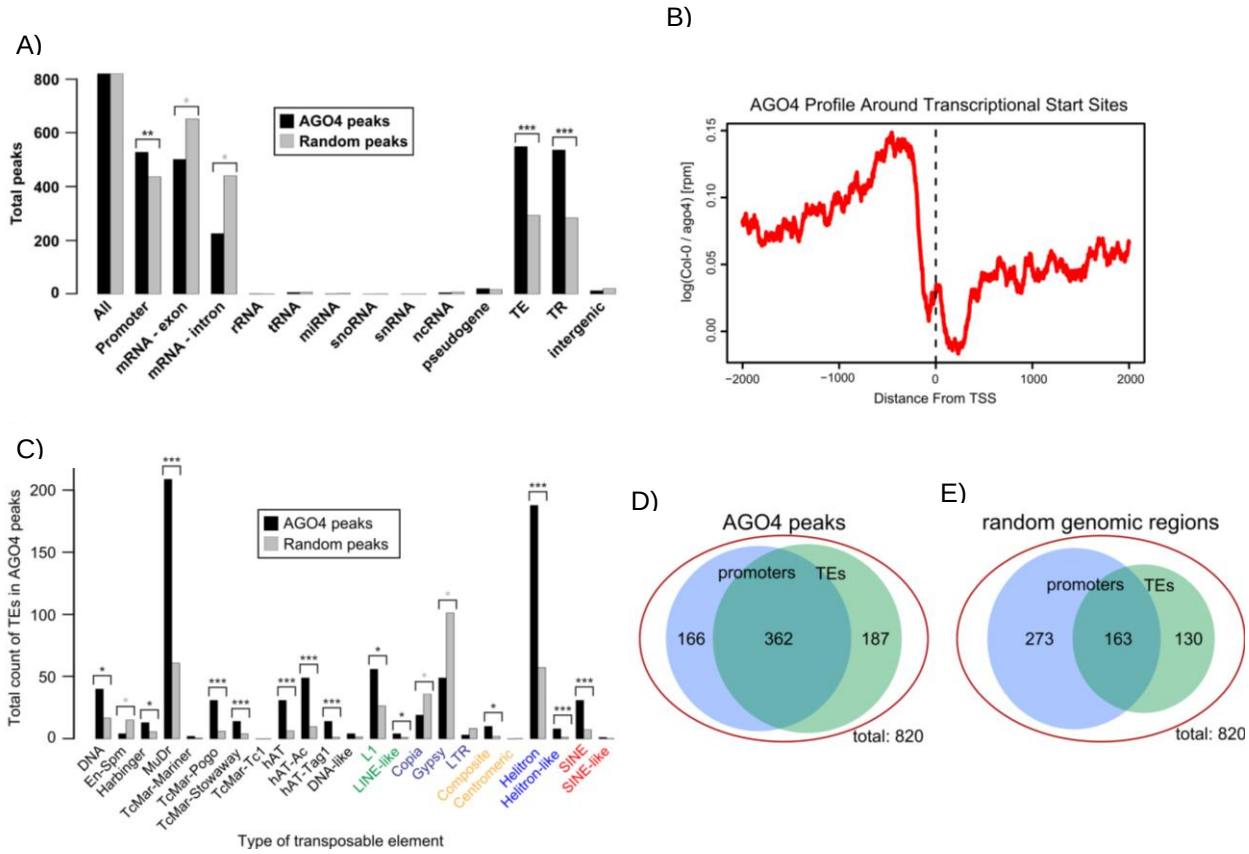


Figure 2.2 AGO4 is enriched on transposable elements within promoters of protein-coding genes

A) AGO4 binding is significantly enriched at promoters, transposable elements, and tandem repeats, but deficient in gene bodies. Classification of all AGO4 (black bars) or a set of randomly generated (grey bars) peaks. TE and TR are transposable elements and tandem repeats, respectively. *, **, and *** denote p-values < .05, < .001, and → 0, respectively.

B) AGO4 binding is enriched on regions upstream of transcription start sites. Profile of AGO4 binding around transcription start sites of all annotated *Arabidopsis* genes showing proportion of ChIP-seq reads in Col-0 wild type to the *ago4* mutant plants. rpm – reads per million.

C) AGO4 binding sites are significantly enriched in DNA transposons, but are devoid of LTR transposable elements. Classification of specific transposable element levels in AGO4 (black bars) or a set of randomly generated (grey bars) peaks. Transposable elements are as specified on the x-axis. *, **, and *** denote p-values < .05, < .001, and → 0, respectively.

D – E) AGO4 preferentially binds transposable elements within gene promoters. D) Venn diagram showing AGO4 peaks mapping to regions 1kb upstream of transcription start sites, transposable elements and both. E) Venn diagram showing random genomic regions 1kb upstream of transcription start sites, transposable elements and both.

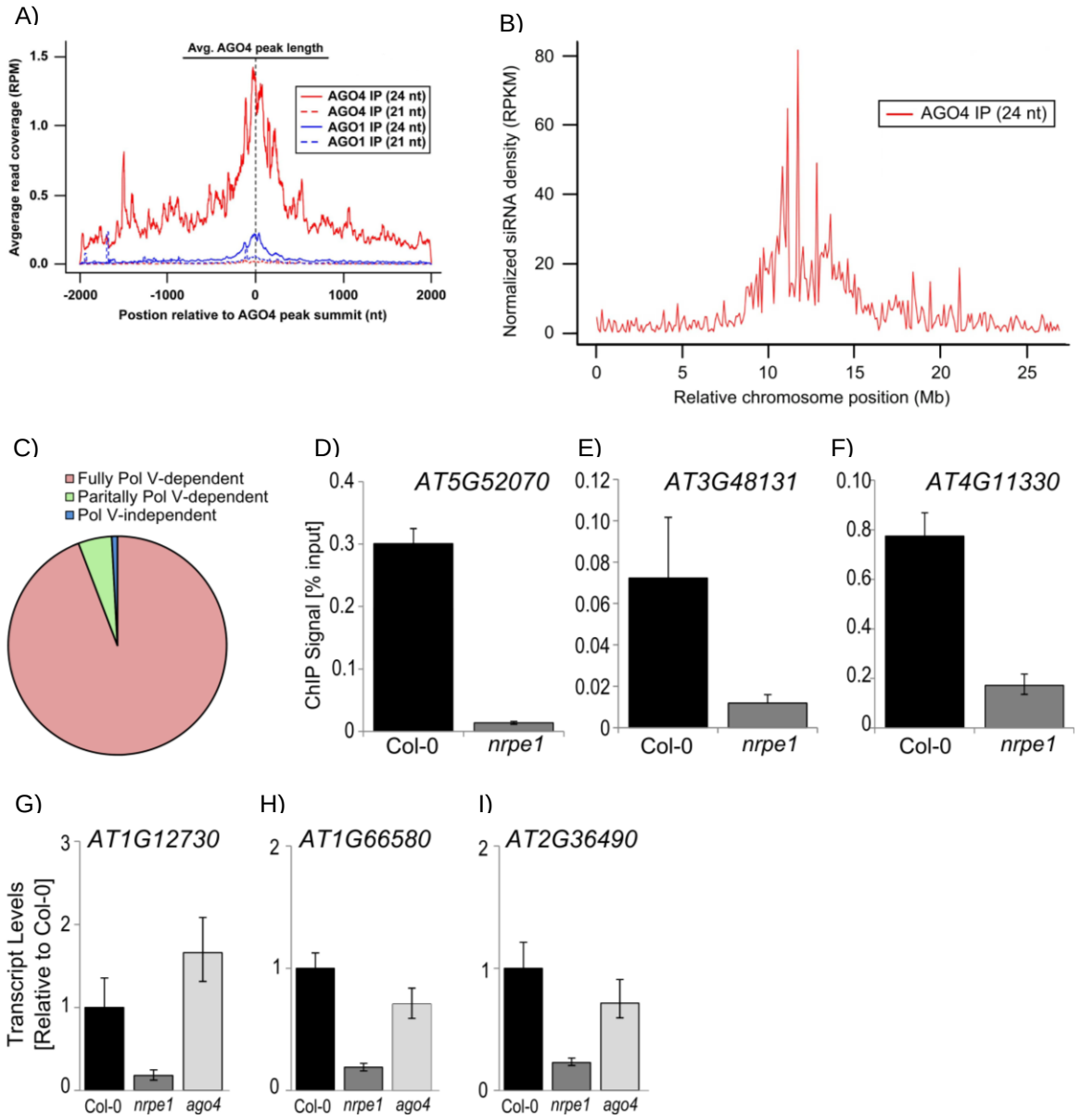


Figure 2.3 AGO4 binding specificity towards TEs in gene promoters is mediated by lncRNA

A) AGO4 binding sites have a significant overlap with AGO4-bound 24 nt but not 21 nt smRNAs. The plot demonstrates the levels of previously identified 24 nt and 21 nt AGO4- (red lines as indicated) and AGO1-bound (blue lines as indicated) smRNAs. The solid black line at the top of the graph indicates the average AGO4 peak size. The dashed black line denotes the position of AGO4 peak summits.

B) Density of AGO4-bound 24 nt siRNAs along the length of chromosome 5. RPKM – reads per kilobase per million.

C) AGO4 binding is dependent on Pol V. Proportions of AGO4 binding sites identified as fully Pol V-dependent, partially Pol V-dependent and Pol V-independent are shown.

D – F) ChIP-real time PCR showing Pol V binding to AGO4 peaks. Bars represent averages from three independent amplifications. Error bars represent standard

deviations. More loci are shown in Figs. 2.11B – D.
G – I) Pol V-dependent transcripts are present on AGO4 binding sites. Graphs show ncRNA accumulation assayed using real-time RT-PCR in Col-0 wild type, *nrpe1* and *ago4* mutant plants normalized to Actin as a control. Bars represent averages from three independent amplifications. Error bars represent standard deviations.

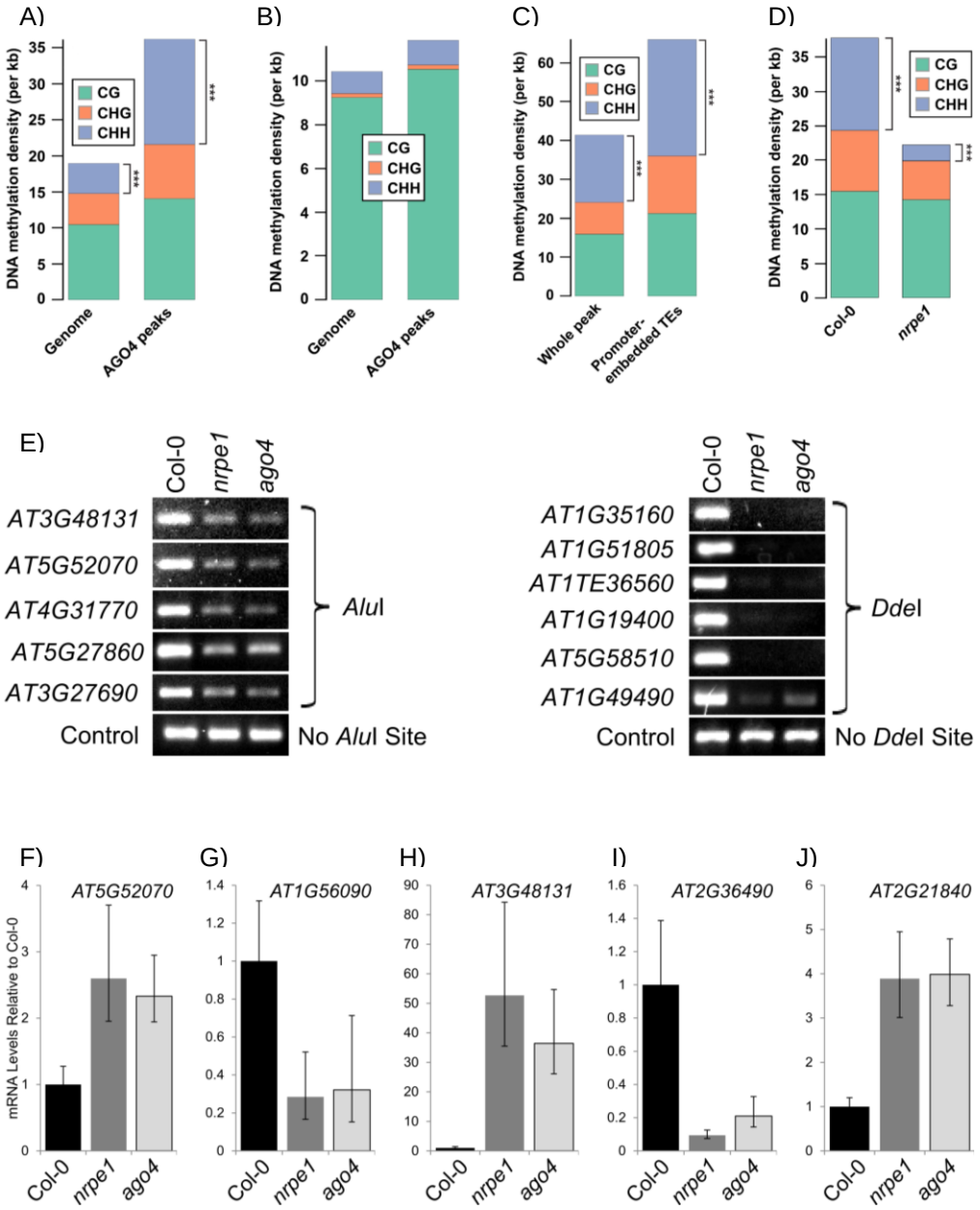


Figure 2.4 AGO4 binding mediates DNA methylation and controls gene activity
A) AGO4 binding sites are enriched in DNA methylation, especially at CHH sites. The graph shows the average levels of CG (green portion of bars), CHG (red portion of bars), and CHH (purple portion of bars) per kilobase (kb) of the *Arabidopsis* genome overall (left bar) or within AGO4 peaks (right bar). Methylation data is previously published for Col-0 wild-type plants. *** denotes p-value → 0.
B) CHH methylation enrichment of AGO4 binding sites requires non-CG DNA methyltransferases. The graph shows the average levels of CG (green portion of bars), CHG (red portion of bars), and CHH (purple portion of bars) per kilobase (kb) of the *Arabidopsis* genome overall (left bar) or within AGO4 peaks (right bar). Methylation data is previously published for *drm1 drm2 cmt3* triple mutant plants.
C) DNA methylation within AGO4 binding sites is significantly higher on transposable

elements embedded within gene promoters. The graph shows the average levels of CG (green portion of bars), CHG (red portion of bars), and CHH (purple portion of bars) per kilobase (kb) of all promoter-associated AGO4 binding sites (left bar) or within TEs embedded in promoter-associated AGO4 binding sites (right bar). Methylation data is previously published for Col-0 wild-type plants. *** denotes p-value → 0.

D) CHH methylation within AGO4 binding sites is dependent on Pol V. The graph shows the average levels of CG (green portion of bars), CHG (red portion of bars), and CHH (purple portion of bars) per kilobase (kb) of AGO4 binding sites in *nrpe1* mutant (right bar) and Col-0 wild type derived from a corresponding dataset (left bar). Methylation data is previously published (Wierzbicki *et al.*, 2012). *** denotes p-value → 0.

E) AGO4 and Pol V are required for CHH DNA methylation on AGO4 binding sites. DNA methylation analysis using *Alu*I DNA methylation-sensitive restriction endonuclease. Digested genomic DNA was amplified using PCR. A sequence lacking *Alu*I sites (IGN5) was used as a loading control. More loci are shown in Fig. 2.12C.

F – J) AGO4 affects the expression levels of certain protein-coding genes whose promoters are binding sites. Graphs show mRNA accumulation assayed using real-time RT-PCR in Col-0 wild type, *nrpe1* and *ago4* mutant plants normalized to Actin as a control. Bars represent averages from three biological replicates. Error bars represent standard deviations.

A)

Biological Process	p-value
Response to carbohydrate stimulus	5.6E-4
Regulation of microtubules	6.6E-4
Response to stress	7.0E-4
Response to biotic stimulus	1.0E-3
Photosynthesis	1.8E-3
Response to other organism	2.3E-3
Response to organic substance	2.4E-3
Response to endogenous stimulus	2.9E-3
Cellular response to red light	3.0E-3
Response to herbivore	3.0E-3
Response to jasmonic acid	3.3E-3

B)

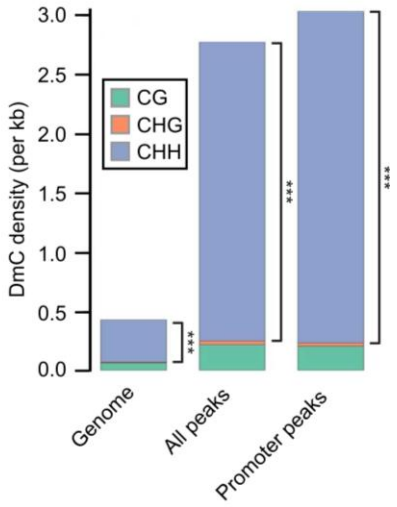


Figure 2.5 AGO4 binding is enriched on genes affected by stress
A) AGO4 promoter binding may regulate genes encoding proteins in stress, environmental, and hormone responses. The 11 most significantly enriched biological processes (and corresponding p-value) for all the genes whose promoters are bound by AGO4.
B) Biotic stress-mediated differential DNA methylation is enriched on AGO4 binding sites. The graph shows the average levels of CG (green portion of bars), CHG (red portion of bars), and CHH (purple portion of bars) per kilobase (kb) of the *Arabidopsis* genome overall (left bar), within all AGO4 peaks (middle bar) or within promoter-associated AGO4 peaks (right bar). Methylation data is previously published (Dowen *et al.*, 2012). *** denotes p-value → 0.

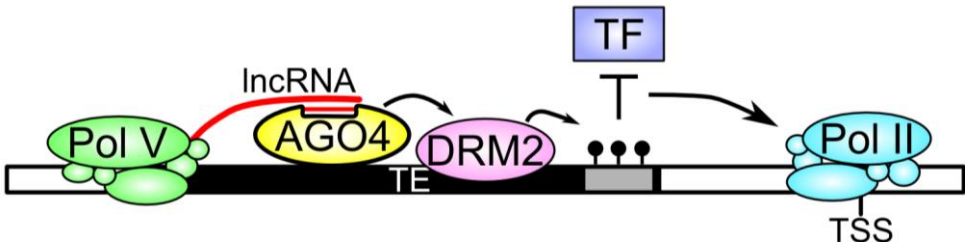


Figure 2.6 A model for AGO4 function on gene promoters
Pol V produces long non-coding RNA which is a scaffold for AGO4-siRNA binding. AGO4 recruits DRM2 *de novo* DNA methyltransferase. CHH methylation affects transcription factor binding within gene promoters which in turn positively or negatively affects Pol II transcription and gene expression.

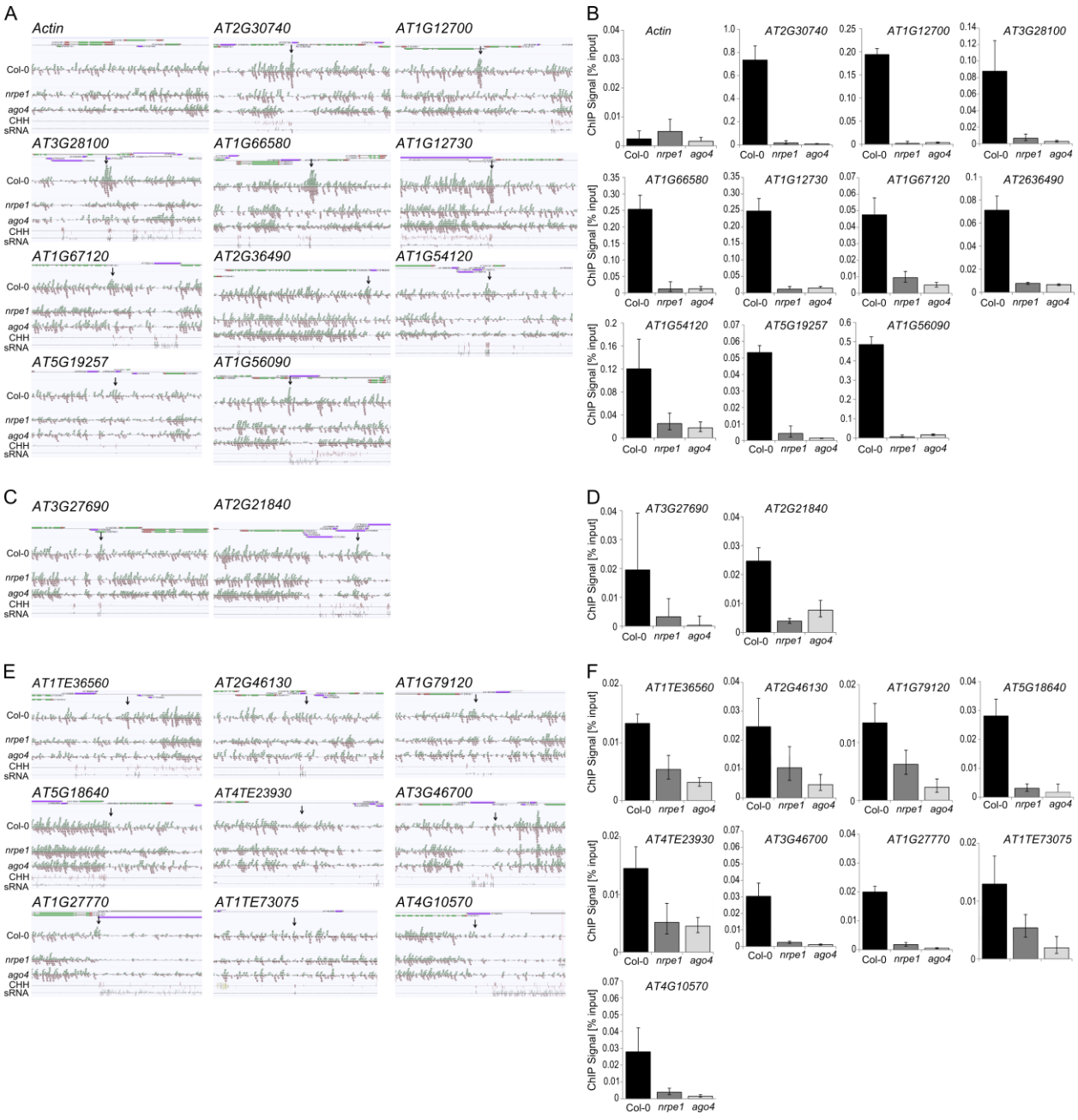


Figure 2.7 Identification of AGO4-bound loci (supplementary to Figure 2.1)
A, C, E) Graphical representation of AGO4-bound loci identified using ChIP-seq. A) Actin control and loci with the score of AGO4 binding ranking in the top 20%. C) Loci with the score of AGO4 binding ranking in the middle 60%. E) Loci with the score of AGO4 binding ranking in the bottom 20%. The genome browser screenshots show from top genome annotation, ChIP-seq sequencing reads from Col-0 wild type, *nrpe1* and *ago4* strains, CHH DNA methylation and total small RNA reads
B, D, F) ChIP-real time PCR validation of Pol V-dependent AGO4 binding to chromatin on AGO4 peaks identified using ChIP-seq. B) Actin control and loci with the score of AGO4 binding ranking in the top 20%. D) Loci with the score of AGO4 binding ranking in

the middle 60%. F) Loci with the score of AGO4 binding ranking in the bottom 20%. Bars represent averages from three independent amplifications. Error bars represent standard deviations.

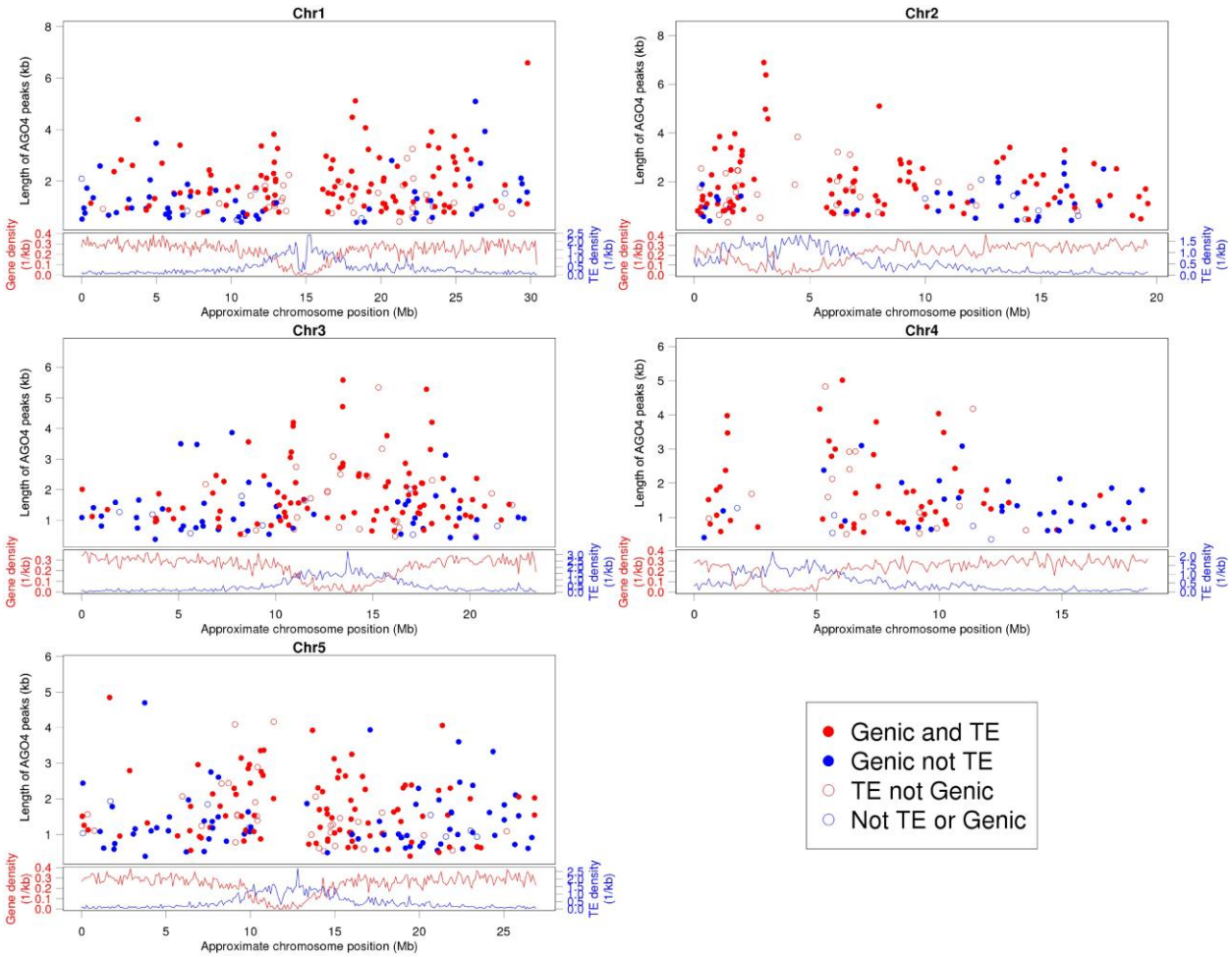


Figure 2.8 AGO4 binding shows no preference towards transposon-rich pericentromeric regions (supplementary to Figure 2.2)

The graphs show the distribution of AGO4 peaks along the length of the five nuclear chromosomes of *Arabidopsis* (as indicated in the Figure). The corresponding category for each colored dot can be found in the legend. The lower graphs for each chromosome provide the density of genes (red line and y-axis label) and transposable elements (blue line and y-axis label).

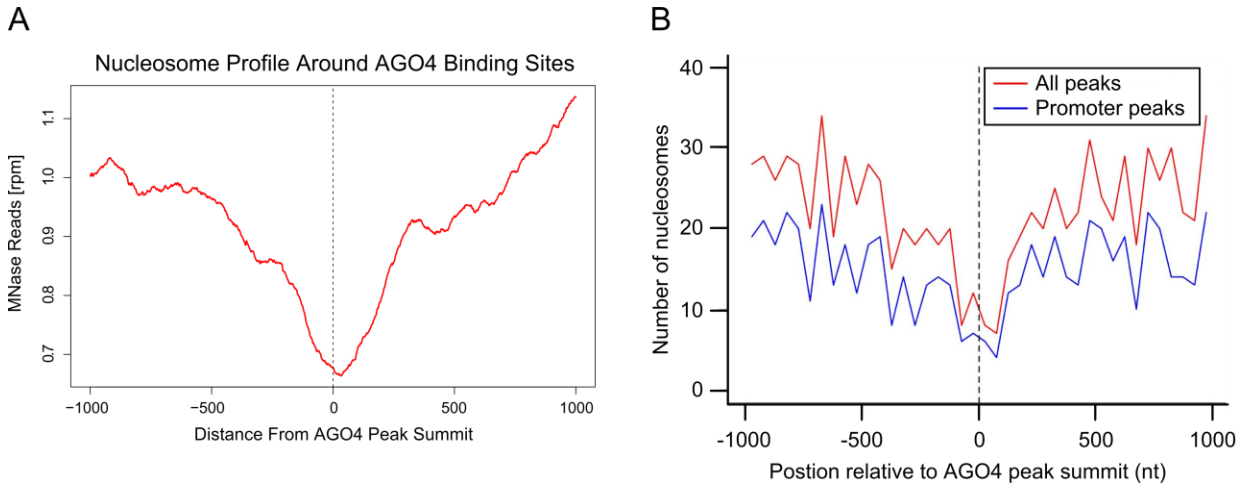


Figure 2.9 The summit of AGO4 peaks is devoid of nucleosomes (supplementary to Figure 2.2)
The average number of mononucleosome-sequencing reads (A) and nucleosomes (B) is plotted over the average length of all (red line) or only promoter-localized (blue line) AGO4 peaks.

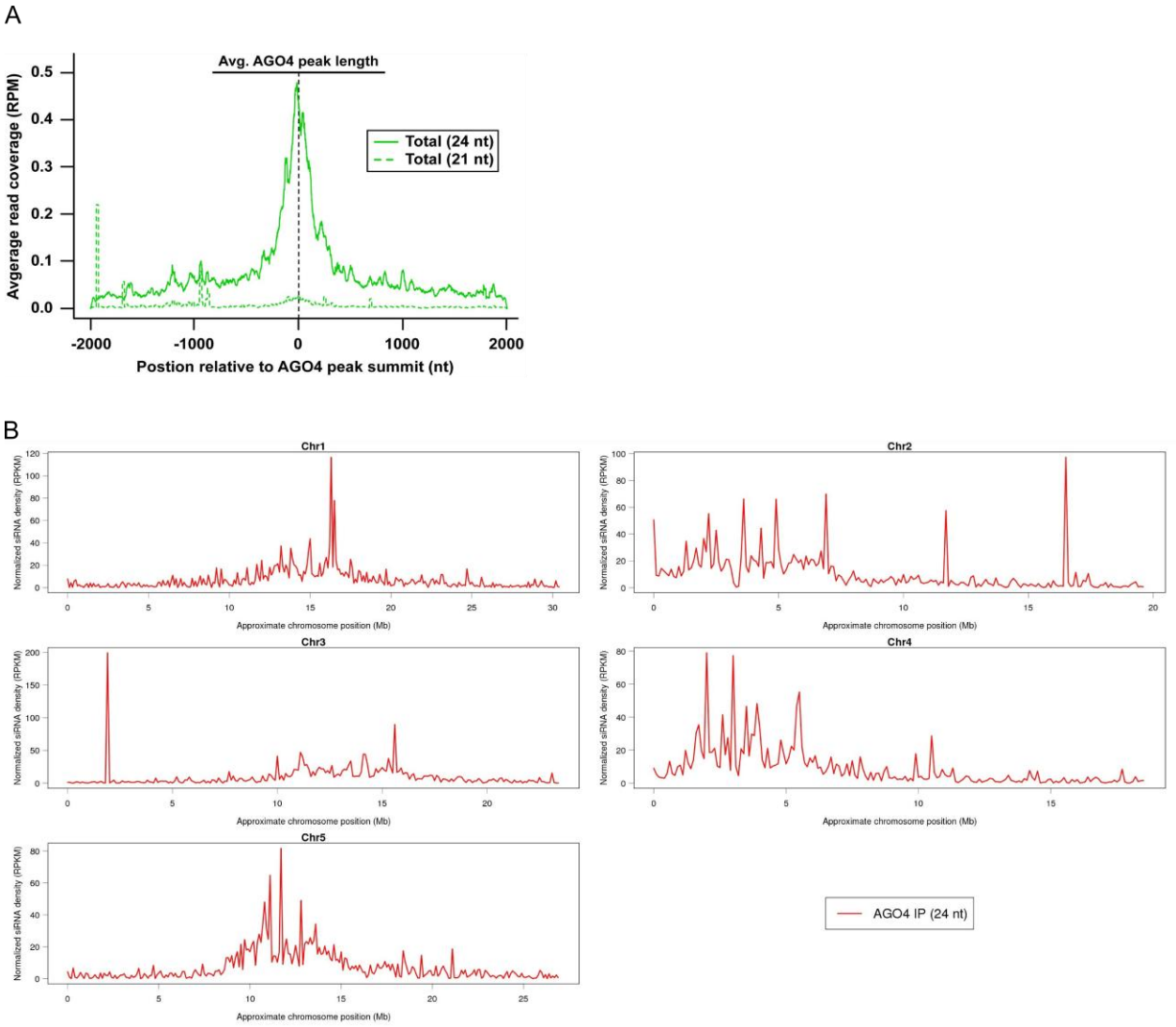


Figure 2.10 Pattern of AGO4 binding cannot be explained by AGO4-bound siRNAs (supplementary to Figure 2.3)

A) AGO4 binding sites have a significant overlap with previously identified 24 nt but not 21 nt smRNAs. The plot demonstrates the levels of previously identified 24 nt (solid green line) and 21 nt (dashed green line) smRNAs. The solid black line at the top of the graph indicates the average AGO4 peak size. The dashed black line denotes the position of AGO4 peak summits.

B) Density of AGO4-bound 24 nt siRNAs along the length of the five nuclear chromosomes of *Arabidopsis thaliana*.

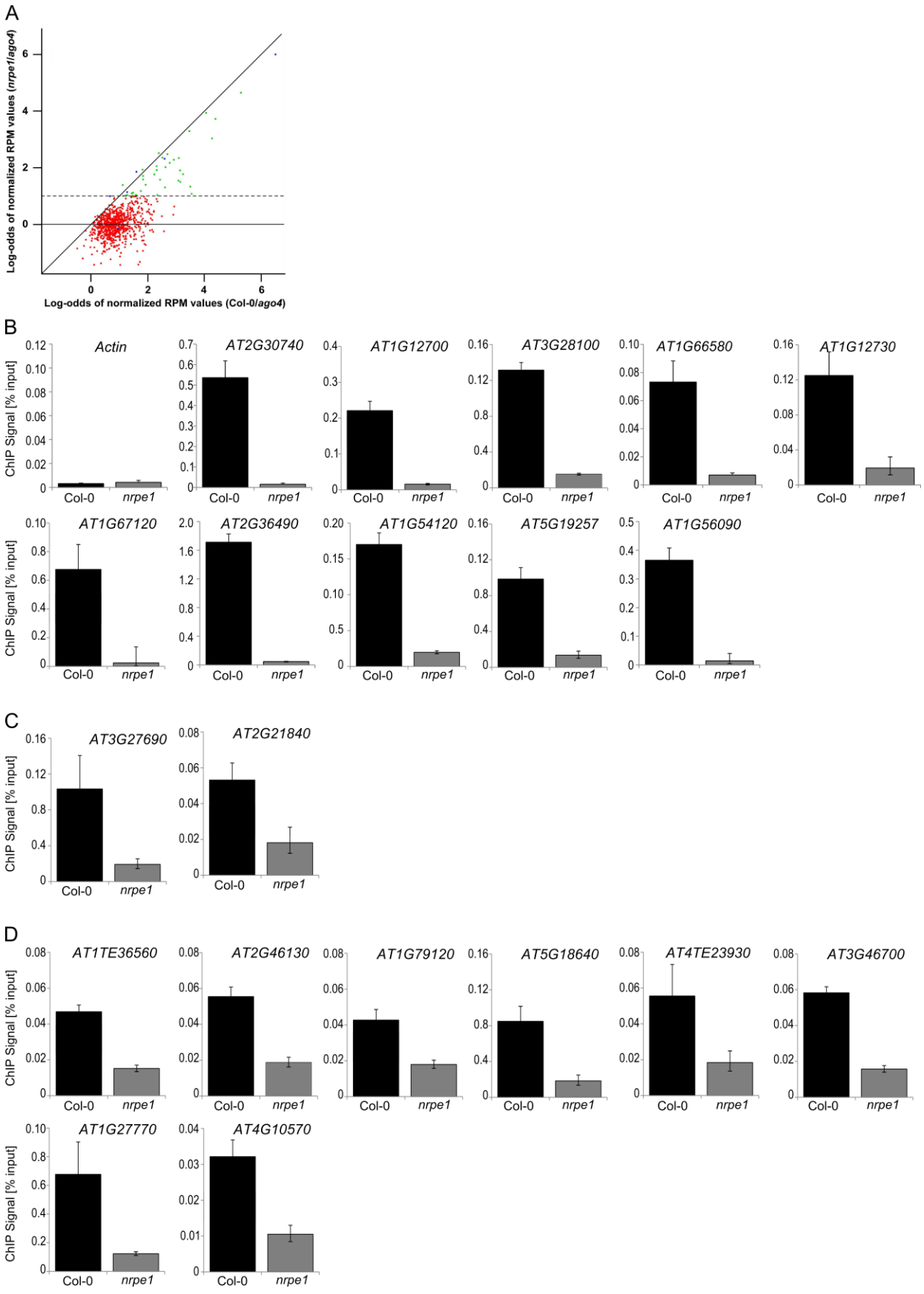

A
Log-odds of normalized RPM values (nrpe1/ago4)
Log-odds of normalized RPM values (Col-0/ago4)
B
ChIP Signal [% input]
Actin
AT2G30740
AT1G12700
AT3G28100
AT1G66580
AT1G12730
AT1G67120
AT2G36490
AT1G54120
AT5G19257
AT1G56090
Col-0
nrpe1
C
AT3G27690
AT2G21840
D
AT1TE36560
AT2G46130
AT1G79120
AT5G18640
AT4TE23930
AT3G46700
AT1G27770
AT4G10570

Figure 2.11 AGO4 binding specificity towards TEs in gene promoters is mediated by Pol V-produced lncRNAs (supplementary to Figure 2.3)
A) AGO4 binding is dependent on Pol V. A scatterplot showing the log-odds of normalized ChIP-seq reads per million (RPM) values for Col-0 wild type (x-axis) and *nrpe1* (y-axis) versus *ago4* mutant plants. The dashed line indicates the value at which AGO4 binding can be described as Pol V partially dependent. The red, green, and blue dots indicate AGO4 peaks that are completely, partially, and not dependent on Pol V, respectively.
B-D) ChIP-real time PCR showing Pol V binding to AGO4 peaks. B) Actin control and loci with the score of AGO4 binding ranking in the top 20%. C) Loci with the score of AGO4 binding ranking in the middle 60%. D) Loci with the score of AGO4 binding ranking in the bottom 20%. Bars represent averages from three independent amplifications. Error bars represent standard deviations.

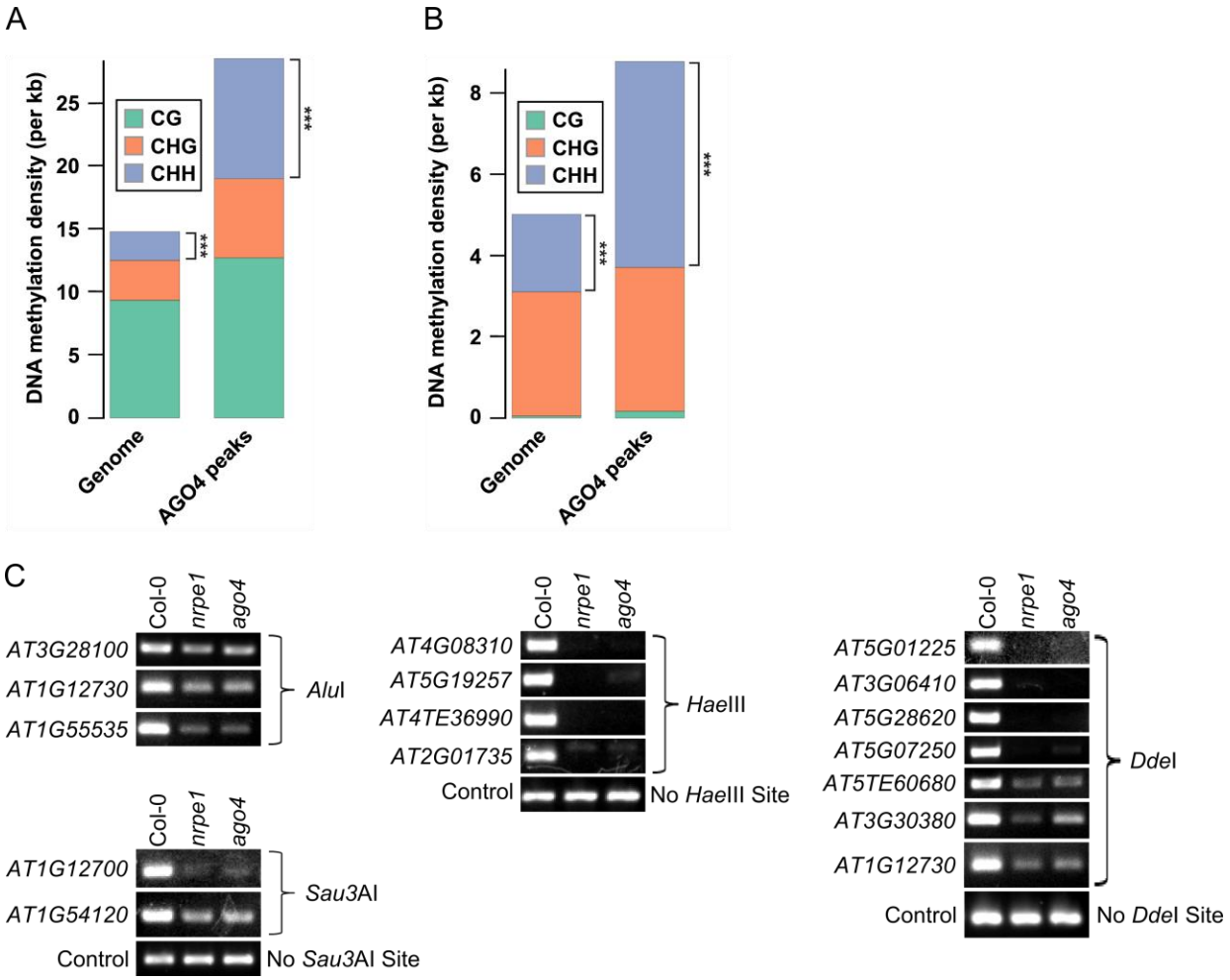


Figure 2.12 AGO4 binding mediates DNA methylation (supplementary to Figure 2.4)

A-B) AGO4 binding sites are enriched in DNA methylation, especially at CHH sites. The graphs show the average levels of CG (green portion of bars), CHG (red portion of bars), and CHH (purple portion of bars) per kilobase (kb) of the *Arabidopsis* genome overall (left bar) or within AGO4 peaks (right bar). Methylation data is previously published (Lister *et al.*, 2008) for *ros1 dml2 dml3 (*A) and *met1* (B) mutant plants. *** denotes p-value → 0.

C) AGO4 and Pol V are required for CHH DNA methylation on AGO4 binding sites. DNA methylation analysis using *Alu*I, *Sau*3AI, *Hae*III and *Dde*I DNA methylation-sensitive restriction endonucleases. Digested genomic DNA was amplified using PCR. Sequence lacking *Sau*3AI and *Dde*I sites (AT2G36490) or *Hae*III (AT2G27860) were used as loading controls. More loci and a control for *Alu*I are shown in Fig. 2.4E.

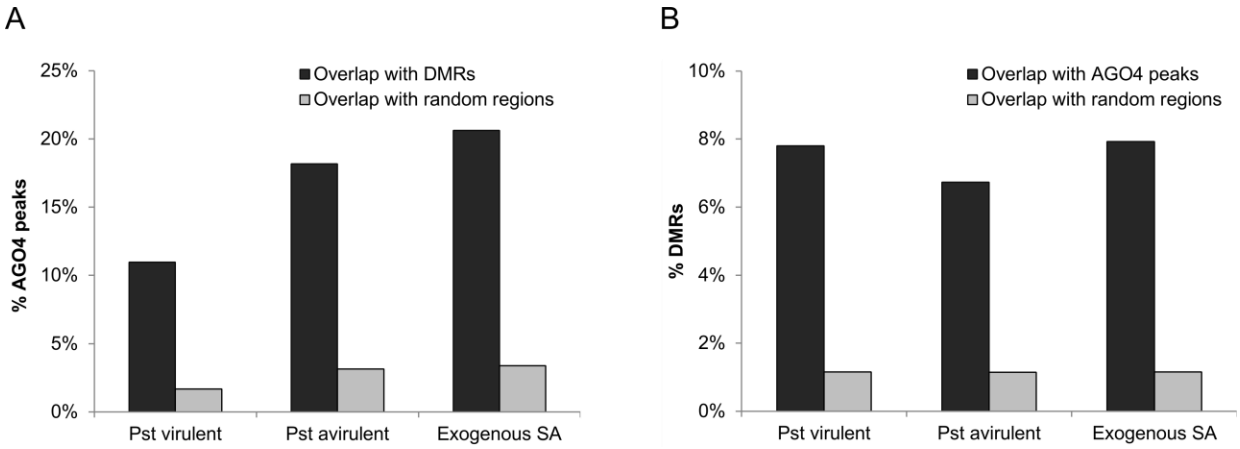


Figure 2.13 AGO4 binding overlaps genomic regions containing stress-induced differential CHH methylation (supplementary to Figure 2.5)
A) Percentages of AGO4 peaks overlapping differentially methylated regions (DMRs) identified upon treatment with biotic stress. Average overlaps of AGO4 peaks with 1000 random genomic regions are shown as controls (light grey bars). All enrichments are statistically significant ($p < 0.001$).
B) Percentages of differentially methylated regions (DMRs) identified upon treatment with biotic stress overlapping AGO4 peaks. Average overlaps of DMRs with 1000 random genomic regions are shown as controls (light grey bars). All enrichments are statistically significant ($p < 0.001$).

CHAPTER 3

A SWI/SNF Nucleosome Remodeling Complex Acts in Non-Coding RNA-Mediated Transcriptional Silencing

The contents of this chapter were published in Molecular Cell in 2013. Gudrun Böhmdorfer performed work shown in Figure 3.1A,B, Figure 3.4D-H, Figure 3.8A, and Figure 3.11 E-I. Yongyou Zhu performed work shown in Figure 3.1C-F, Figure 3.2, Figure 3.3A-E, Figure 3.4A-C, Figure 3.6, Figure 3.8 B-F, Figure 3.9A-D, Figure 3.10A-E, and Figure 3.11 A-D,K. Andrzej Wierzbicki performed experiment in Figure 3.12 A. I performed all other experiments and data analysis shown in this chapter.

Abstract

RNA-mediated transcriptional silencing prevents deleterious effects of transposon activity and controls the expression of protein-coding genes. It involves long non-coding RNAs (lncRNAs). In *Arabidopsis thaliana* some of those lncRNAs are produced by a specialized RNA Polymerase V (Pol V). The mechanism by which lncRNAs affect chromatin structure and mRNA production remains mostly unknown. Here we identify the SWI/SNF ATP-dependent nucleosome remodeling complex as a component of the RNA-mediated transcriptional silencing pathway. We found that SWI3B, an essential subunit of the SWI/SNF complex, physically interacts with a lncRNA-binding protein, IDN2. SWI/SNF subunits contribute to lncRNA-mediated transcriptional silencing. Moreover, Pol V mediates stabilization of nucleosomes on silenced regions. We propose that Pol V-produced lncRNAs mediate transcriptional silencing by guiding the SWI/SNF complex and establishing positioned nucleosomes on specific genomic loci. We further propose that guiding ATP-dependent chromatin remodeling complexes may be a more general function of lncRNAs.

Introduction

Transposable elements and other classes of repetitive genomic elements are controlled by RNA-mediated transcriptional silencing [1], in plants also known as RNA-dependent DNA methylation (RdDM) [2]. This process not only prevents deleterious activities of transposons but is also believed to regulate the expression of protein-coding genes [3,4].

RNA-mediated transcriptional silencing pathway involves two independent classes of non-coding RNA: small interfering RNA (siRNA) and long non-coding RNA (lncRNA) [2]. In *Arabidopsis thaliana* siRNA is produced by the activities of RNA Polymerase IV (Pol IV), RNA-DEPENDENT RNA POLYMERASE 2 (RDR2) and DICER-LIKE 3 (DCL3). It then incorporates into the ARGONAUTE4 (AGO4) protein and gives it specificity towards genomic loci with the same sequence as siRNA [2]. Despite being necessary for AGO4 to bind chromatin and mediate repressive chromatin modifications, siRNA is not sufficient for those events. Another required component is lncRNA produced by RNA Polymerases II and V (Pol II and Pol V, respectively), which is necessary for siRNA to recognize its genomic target loci [5–7].

Pol V is a DNA-dependent RNA polymerase with subunit composition similar to Pol II [8,9]. Pol V-produced lncRNAs originating from silencing targets have been shown to be required for AGO4 association with chromatin, CHH DNA methylation and transcriptional silencing [4,6,10]. These RNAs are believed to work as scaffolds for the assembly of silencing complexes, which establish DNA methylation and repressive histone modifications [5,11]. They have also been proposed to be the primary factors guiding repressive chromatin modifications to gene regulatory regions [4,12]. The mechanism connecting lncRNAs to the activities of chromatin modifying enzymes is not well understood and only two lncRNA-binding proteins have been identified so far: AGO4 and SUPPRESSOR OF TY INSERTION 5 – LIKE (SPT5L).

SPT5L is a homolog of the SPT5 Pol II elongation factor [8,13,14]. It binds chromatin at silenced loci independently of AGO4 and has been proposed to work together with AGO4 in the recruitment of chromatin modifying enzymes [15]. lncRNAs were also hypothesized to interact with IDN2, an RNA-binding protein required for RNA-mediated

transcriptional silencing, which has been shown to bind double-stranded RNA *in vitro* [16–19]. However, the *in vivo* function of IDN2 remains mysterious.

Besides DNA methylation and posttranslational histone modifications, chromatin status may also be affected by active nucleosome positioning by ATP-dependent chromatin remodeling complexes, which control DNA accessibility by positioning, moving or exchanging nucleosomes [20,21]. Although active nucleosome remodeling has been shown to control nucleosome positioning and affect transcription throughout genomes [22], it remains unknown if nucleosomes are controlled by the transcriptional silencing pathways. One of the families of ATP-dependent chromatin remodeling complexes is known as SWI/SNF [21]. In *Arabidopsis thaliana* SWI/SNF complexes are known to contain at least five core subunits, including an Snf2-family ATPase and two SWI3 proteins [23,24].

In this work we discovered a hitherto unidentified component of the RNA-mediated transcriptional silencing pathway. Looking for lncRNA-associated proteins we identified SWI3B, a subunit of the SWI/SNF ATP-dependent chromatin remodeling complex. SWI3B interacts with Pol V-produced lncRNAs indirectly with IDN2 dimers serving as adaptors. SWI3B is required for RNA-mediated transcriptional silencing as indicated by derepression of several known silenced loci and a significant overlap of genes misregulated in mutants defective in lncRNA production or SWI3B activity. Defects in silencing were also observed in mutants of other SWI/SNF subunits indicating that this process involves the whole ATP-dependent chromatin remodeling complex. Consistently, elimination of Pol V-produced lncRNAs resulted in widespread changes in nucleosomes positioning on silencing targets. Moreover, DNA methylation levels were partially reduced in the *swi3b* mutant. These results support a model, where lncRNA recruits IDN2 dimers, which interact with the SWI/SNF complex by its SWI3B subunit. The SWI/SNF complex positions nucleosomes, which affect Pol II transcription by facilitating DNA methylation and/or restricting protein access to DNA.

Results

Pol V-produced lncRNAs associate with IDN2

To test if IDN2 physically interacts with Pol V-produced lncRNAs we performed RNA immunoprecipitation (RNA IP) with affinity-purified anti-IDN2 antibody and assayed the obtained samples using real time RT-PCR with primers specific for Pol V transcripts identified by Pol V ChIP-seq [15,25]. IDN2 pulled down RNA from Col-0 wild type plants at much higher levels than from the *idn2-1* mutant (Figures 3.1A-B and 3.8A). RNA recovery was also strongly reduced in *nrpe1*, a mutant defective in the largest subunit of Pol V and unable to produce lncRNA [6], which suggests that pulled down RNAs are Pol V-produced lncRNAs. Western blot demonstrated that the anti-IDN2 antibody was specific and that IDN2 stability was not affected in the *nrpe1* mutant (Figure 3.8B). This demonstrates that IDN2 interacts at least with a subset of Pol V-produced lncRNAs. The *idn2-1* mutant has a substitution followed by an eight aminoacid deletion in its XS domain, which adopts an RNA recognition motif fold and interacts with RNA [16,19,26]. In contrast to the knock-out T-DNA mutant allele *idn2-2*, *idn2-1* line still accumulates IDN2, although at a lower level than wild type (Figure 3.8B). Because the *idn2-1* mutant is defective in transcriptional silencing [16] and unable to bind lncRNA (Figure 3.1A-B) this suggests that interaction of IDN2 with lncRNA may be important for its function in RdDM.

IDN2 interacts with SWI3B

Having established that IDN2 physically interacts with lncRNA we used a yeast two hybrid screen with IDN2 as a bait to identify proteins that might be indirectly associated with lncRNA. Among positive clones we found several corresponding to SWI3B, a subunit of the SWI/SNF ATP-dependent chromatin remodeling complex [23] (Figure 3.1C). To validate this finding we used a targeted yeast two hybrid test, which confirmed that IDN2 interacts with full length SWI3B but not with other SWI3 homologs (SWI3A, SWI3C or SWI3D; Fig 3.1D). This interaction was then confirmed *in vivo* by reciprocal co-immunoprecipitations of FLAG- and GFP-tagged IDN2 and SWI3B transiently overexpressed in tobacco leaves (Figure 3.1E and Figure 3.8C) and by co-immunoprecipitation of SWI3B-GFP and IDN2-FLAG driven by their respective native

promoters in stable *Arabidopsis* transformants (Figure 3.1F). This interaction was lost in IDN2 truncations eliminating predicted coiled-coil regions (Figure 3.8D) but not in a deletion mutant in the RNA-binding XS domain (Figure 3.8E-F) [16], suggesting that IDN2 binds SWI3B using its coiled-coil domain. Together, these results indicate that SWI3B physically interacts with IDN2 and therefore SWI3B might be involved in lncRNA-mediated transcriptional silencing.

IDN2 dimerization is required for silencing but not for interaction with SWI3B

To check if recently reported IDN2 dimerization [19,26] (Figure 3.2A) plays a role in the interaction with SWI3B we identified a region within the IDN2 coiled-coil domain responsible for homodimerization (Figure 3.2B). We created the IDN2 M8 mutant with eight point mutations introduced within this region (Figure 3.2C) to disrupt homodimerization by specific interactions of the coiled-coil domain. The IDN2 M8 mutant was unable to dimerize in the yeast two hybrid assay (Figure 3.9A) and *in vivo* co-immunoprecipitation in tobacco leaves (Figure 3.2D). We tested DNA methylation levels in *idn2-2* knock out mutants transformed with wild-type *IDN2* and *IDN2 M8* by digesting genomic DNA with methylation-sensitive restriction endonucleases followed by PCR. DNA methylation levels on silencing targets *AtSN1*, *IGN5* and *MEA-ISR* were reduced in the *idn2-2* knock out mutant (Figure 3.2E, Figure 3.9B). Transformation of the *idn2-2* mutant with wild type *IDN2* restored DNA methylation to wild-type levels, however *IDN2 M8* was unable to restore DNA methylation (Figure 3.2E, Figure 3.9B). This indicates that dimerization of IDN2 is required for the biological function of IDN2 in the transcriptional gene silencing pathway. We further tested if IDN2 dimerization was disrupted by deletion in the XS domain present in the *idn2-1* mutant. Yeast two hybrid and co-immunoprecipitation in tobacco leaves revealed that deletion in the XS domain did not affect IDN2 dimerization (Figure 3.9C-D), suggesting that the XS domain is not needed for IDN2 dimerization. Together with our observation that deletion in the XS domain does not disrupt IDN2 interaction with SWI3B (Figure 3.8E-F), this suggests that deletion within the XS domain only disrupts IDN2 binding to RNA. This provides additional support for interaction with lncRNA being important for IDN2 function in RdDM. We further tested if dimerization of IDN2 is required for its interaction with

SWI3B. Yeast two hybrid assay revealed that IDN2 M8 was still able to interact with SWI3B (Figure 3.2F). Although we cannot exclude that the M8 mutation affects other aspects of IDN2 function, these results suggest that IDN2 dimerization is required for RdDM but not for the interaction of IDN2 with SWI3B.

SWI3B contributes to transcriptional silencing

To test if SWI3B is involved in RNA-mediated transcriptional silencing we assayed accumulation of RNA produced from silencing targets in a *swi3b* mutant. Because homozygous knock-out mutants of *SWI3B* are embryo lethal [23], we used plants heterozygous for the *swi3b-2* mutation (*swi3b*/+), which have been shown to have the expression level of *SWI3B* reduced to about 50% and display phenotypes attributed to SWI3B deficiency [27] (Figure 3.10A). Real time RT-PCR revealed that two transposon-originating transcripts within the *solo LTR* region, a well characterized target of RNA-mediated transcriptional silencing [6,28], were derepressed in the *swi3b*/+ mutant as well as in the *idn2* and *nrpe1* mutants (Figure 3.3A-B). Derepression of *solo LTR* was more pronounced in *nrpe1* mutant than in *swi3b*/+ and *idn2* mutants, which is consistent with partial reduction of *SWI3B* in the *swi3b*/+ line [27] (Figure 3.10A) and with IDN2 having several potentially redundant homologs [16,19]. Similar partial reactivation of RdDM targets in the *swi3b*/+ mutant was observed on several other RdDM targets (Figures 3.3C-E and 3.10B-D). These results show that SWI3B contributes to RNA-mediated transcriptional silencing of at least a subset of silencing targets.

SWI3B controls expression of genes affected by silencing

Interaction of SWI3B with lncRNA-binding protein IDN2 and requirement of SWI3B for transcriptional silencing suggest a functional relationship between lncRNA and SWI3B. To test this possibility we used a genome-wide approach to identify genes whose expression is affected in the *swi3b*/+, *nrpe1* and *idn2-1* mutants. RNA-seq from three biological repeats revealed that out of 280 genes significantly upregulated in the *nrpe1* mutant, 137 (49%) were also upregulated in the *swi3b*/+ line (Figures 3.3F and 3.10G). This is significantly more than 1.8% expected by chance (p -> 0, chi-square test). Similarly, out of 343 genes significantly upregulated in the *idn2-1* mutant, 122 (36%) were also upregulated in the *swi3b*/+ line (Figures 3.3F and 3.10G). This is

significantly more than 1.7% expected by chance (p -> 0, chi-square test). Importantly, 55 genes were upregulated in all three mutants (Figure 3.3F), which is significantly more than less than one gene expected by chance (p -> 0, chi-square test). A slightly smaller, although still highly significant overlap was observed among genes downregulated in *nrpe1*, *idn2* and *swi3b*/+ mutants (Figure 3.10F). In contrast, genes upregulated in one genotype and downregulated in another genotype were found at rates not significantly higher than expected by chance (Figure 3.10G). These results suggest that Pol V, IDN2 and SWI3B affect overlapping groups of genes. Although it is unknown which of those genes are direct targets of Pol V, IDN2 and SWI3B, these results are consistent with SWI3B being involved in the same gene regulatory pathway as Pol V and IDN2. Together with IDN2-lncRNA interaction, IDN2-SWI3B interaction and derepression of silencing targets in the *swi3b*/+ mutant, these data demonstrate that SWI3B is involved in the RNA-mediated transcriptional silencing pathway.

SWI/SNF complex contributes to transcriptional silencing

SWI3B is a subunit of a putative SWI/SNF ATP-dependent chromatin remodeling complex [23]. Therefore, the involvement of SWI3B in RNA-mediated transcriptional silencing suggests that the SWI/SNF complex may be involved in this process. To test this possibility, we assayed transcriptional silencing in mutants defective in several known components of the SWI/SNF complexes, including four SWI3 proteins [23,29] and two Snf2-family ATPases SPLAYED (SYD) [30] and BRAHMA (BRM) [24,31]. *Solo LTR* was significantly derepressed in the *swi3b*/+, *swi3d* and *syd* mutants (Figure 3.4A), which shows that this locus is controlled by a SWI/SNF complex including SWI3B, SWI3D and SYD. Another silencing target *At2TE78930* was derepressed in *swi3b*/+, *swi3c* and *brm* mutants (Figure 3.4B) indicating that this locus is also controlled by a SWI/SNF complex, however the subunit composition of the complex acting on *At2TE78930* is different than on *solo LTR*. Other tested loci also displayed locus-specific contributions of specific SWI/SNF subunits for transcriptional silencing (Figures 3.4C and 3.11A-C). These results indicate that SWI3B is not the only subunit of SWI/SNF required for transcriptional silencing. They further suggest that the SWI/SNF complex as a whole

contributes to RNA-mediated transcriptional silencing, but subunit contributions may be locus-specific.

SWI/SNF functions downstream of lncRNA production

The physical interaction of a SWI/SNF subunit with lncRNA-binding protein IDN2 suggests that SWI/SNF functions downstream of Pol V and might be recruited by Pol V and IDN2 to specific targets in the genome in a similar fashion as AGO4 [10] and SPT5L [15]. Alternatively, SWI/SNF could work with the RDD complex [32] and mediate Pol V binding to chromatin and/or Pol V transcription [6,10,33]. To distinguish between these possibilities we assayed the accumulation of known Pol V transcripts [6,15,25] in *idn2* and *swi3b*/+ mutants. Because of low abundance, these transcripts were not detectable in our RNA-seq datasets and could only be assayed using targeted RT-PCR. We found levels of all tested Pol V transcripts to be unchanged in both *idn2* and *swi3b*/+ mutants (Figure 3.4E-H and 3.11F-I). This shows that IDN2 and SWI/SNF work either in parallel to Pol V transcription or downstream of Pol V-produced lncRNA. This was further confirmed by chromatin immunoprecipitation (ChIP) with anti-SWI3B antibody, which indicated that at least on the tested loci SWI3B binding to chromatin was reduced in the *nrpe1* mutant (Figure 3.11J). Together with the IDN2-SWI3B interaction and the requirement of SWI/SNF for silencing these results suggest that Pol V-produced lncRNAs guide the SWI/SNF complex to specific genomic loci with IDN2 being an intermediate adaptor protein.

Pol V affects nucleosome positioning

Because SWI/SNF is a putative ATP-dependent chromatin remodeling complex, if it is guided to chromatin by lncRNA, it is expected to affect nucleosome positioning in an lncRNA-dependent manner. To test this prediction we digested nuclei from Col-0 wild type and *nrpe1* mutant plants with micrococcal nuclease (MNase), which specifically cuts genomic DNA not protected by nucleosomes [34] (Figure 3.12A). We sequenced mononucleosomal DNA and identified 2544 nucleosomes which were significantly weaker in *nrpe1* than Col-0 wild type (“Pol V-stabilized”) and 2362 nucleosomes which were significantly stronger in *nrpe1* than Col-0 wild type (“Pol V-destabilized”). We further narrowed down the list of differential nucleosomes by

performing two biological repeats of ChIP-seq with anti-H3 antibody in Col-0 wild type and *nrpe1* mutant, although limited resolution of ChIP affects the ability to detect differential nucleosomes located in close proximity to well stabilized unaffected nucleosomes. This stringent approach yielded 108 Pol V-stabilized and 655 Pol V-destabilized high confidence nucleosomes. We validated several identified nucleosomes (Figure 3.12B) using a locus-specific assay, where MNase-digested chromatin is subject to anti-H3 ChIP followed by real time PCR (Figure 3.5A). These nucleosomes were also destabilized in the *idn2-1* mutant (Figure 3.5A), which is consistent with IDN2 being and adaptor protein connecting Pol V-produced lncRNA to SWI/SNF. A subset of assayed nucleosomes was also affected in a knock-out *brm* mutant (Figure 3.5A), which further suggests that at least a fraction of lncRNA-mediated nucleosome positioning is mediated by the SWI/SNF complex. Together, these data allow speculation that Pol V-produced lncRNAs mediate nucleosome positioning by guiding the SWI/SNF complex to specific genomic loci with IDN2 being an intermediate adaptor protein.

Pol V mediates nucleosome stabilization on silencing targets

Nucleosomes have been shown to be generally correlated with DNA methylation [35]. To test if Pol V may contribute to this phenomenon, we compared published DNA methylation data [12] to the nucleosomes affected by Pol V. We found that in Col-0 wild type, Pol V-stabilized nucleosomes were significantly enriched in CHH methylation when compared to nucleosomes with no significant changes in *nrpe1* ($p < 10^{-9}$; t-test; Figures 3.5B and 3.12C). Pol V-stabilized nucleosomes were also significantly, yet to a lesser extent enriched in CHG methylation ($p < 10^{-4}$; t-test; Figures 3.5C and 3.12D). We further compared DNA methylation on nucleosomes in Col-0 wild type to the *nrpe1* mutant. We found that DNA methylation on Pol V-stabilized nucleosomes was reduced in the *nrpe1* mutant to levels comparable to observed on nucleosomes unaffected in *nrpe1* or throughout the entire genome (Figure 3.5B-D). Observed reduction in DNA methylation was especially apparent in CHH ($p < 10^{-9}$; t-test) and CHG ($p < 0.002$; t-test) contexts. This indicates that Pol V-stabilized nucleosomes overlap Pol V-mediated DNA methylation. In contrast, Pol V-destabilized nucleosomes showed no enrichment in

wild type DNA methylation compared to nucleosomes unaffected in *nrpe1* or to the entire genome (Figures 3.5B-D and 3.12C-E). In the *nrpe1* mutant DNA methylation on Pol V-destabilized nucleosomes was slightly increased (CHH $p < 0.002$; CHG $p < 0.03$; CG $p < 0.002$; t-test), which reflects similar effects on nucleosomes unaffected in *nrpe1* or throughout the entire genome [25] (Figure 3.5B-D). These findings indicate that Pol V-stabilized but not Pol V-destabilized nucleosomes are preferentially present on regions of Pol V-dependent DNA methylation. This is consistent with Pol V-produced lncRNAs mediating both DNA methylation and nucleosome positioning.

To further test if Pol V-stabilized nucleosomes overlap direct Pol V targets, we used a published Pol V ChIP-seq dataset [25] to calculate enrichment of Pol V binding on the three categories of nucleosomes. We found that Pol V-stabilized nucleosomes were enriched in Pol V binding compared to Pol V-destabilized nucleosomes or nucleosomes unaffected in *nrpe1* (Figure 3.5E). Pol V binding and Pol V-dependent CHH methylation are also clearly visible on validated Pol V-stabilized nucleosomes (Figure 3.12B). These results show that Pol V-stabilized nucleosomes are present on Pol V targets. This suggests that, in addition to facilitating DNA methylation, Pol V mediates nucleosome positioning and further supports the model that Pol V-produced lncRNAs control nucleosome positioning on silencing targets.

Pol V-stabilized nucleosomes were also enriched on gene promoters and depleted on transcribed sequences (Figure 3.12F) suggesting that nucleosome remodeling may be involved in the control of gene expression. To test this possibility we overlapped the long list of differential nucleosomes generated using only MNase-seq with genes affected in *swi3b*/+, *nrpe1* and *idn2-1* mutants. Pol V-stabilized nucleosomes were slightly, yet significantly enriched on genes upregulated or downregulated in all three mutants compared to corresponding permutations of genes (Figure 3.5F). Consistently, when we mapped nucleosomes affected in the *nrpe1* mutant to chromosomes, Pol V-stabilized nucleosomes were present throughout gene-rich chromosome arms (Figure 3.12G). In contrast, Pol V-destabilized nucleosomes displayed a preference towards pericentromeric regions (Figure 3.12G), which is consistent with redistribution of silencing towards pericentromeric regions in the *nrpe1*

mutant [25]. This shows that Pol V-stabilized nucleosomes can be correlated with Pol V- and SWI/SNF-controlled genes. This is further consistent with lncRNA-guided nucleosome positioning being an important factor in transcriptional silencing.

SWI/SNF contributes to DNA methylation

Correlation between Pol V-stabilized nucleosomes and Pol V-mediated DNA methylation may be explained by both nucleosome stabilization and DNA methylation being independently guided by lncRNA to the same genomic regions. Alternatively, preexisting DNA methylation may promote nucleosome stabilization or nucleosomes may be preferred targets for DNA methyltransferases. We tested these possibilities by assaying positioning of selected Pol V-stabilized nucleosomes in *drm2*, a mutant in *de novo* DNA methyltransferase. Five of the tested nucleosomes were destabilized in the *drm2* mutant, while one nucleosome was not affected (Figure 3.5A). This is consistent with DNA methylation promoting nucleosome positioning but only on a subset of loci. To test if nucleosome positioning affects DNA methylation we assayed DNA methylation on silenced loci in the *swi3b*/+ mutant. *Solo LTR*, *siR02* and *At2TE78930* all had CHH methylation levels reduced to around 50% in *swi3b*/+ (Figure 3.6A-C). Less pronounced reduction was observed at *IGN6*, *IGN22* and *LTRCO3* (Figure 3.6D-F). This suggests that nucleosome positioning affects DNA methylation possibly by well positioned nucleosomes being preferential targets for DNA methyltransferases, however indirect effects cannot be excluded. Together, these results are consistent with a model where transcriptional silencing is established by Pol V-produced lncRNA guiding both positioned nucleosomes and DNA methylation, while maintenance of silencing is additionally facilitated by a mutual feedback loop between well positioned nucleosomes and DNA methylation.

Discussion

Our results uncover an additional pathway connecting Pol V-produced lncRNA to the establishment of silent chromatin status. The initial step in this pathway is the association of lncRNA with IDN2 (Figure 3.1A), a process which may involve an entire IDN2-containing complex composed of an IDN2 dimer (Figure 3.2A) and two IDN2-

related proteins [18,19,26]. IDN2 has been shown *in vitro* to bind double-stranded RNA using its XS domain [16,19,26], therefore IDN2 may bind lncRNA-siRNA duplexes or hairpin regions within lncRNA. It also implies that the IDN2-lncRNA association may be directed either by AGO4-siRNA complexes or IDN2 may bind lncRNA without the requirement for siRNA in a manner similar to SPT5L [15].

The next step in this pathway is the physical interaction between IDN2 and SWI3B, a subunit of the ATP-dependent chromatin remodeling complex (Figure 3.1C-F). Consistently, transcriptional silencing was partially lost on several known loci in the *swi3b*/+ mutant (Figure 3.3A-E). Although it remains possible that IDN2-lncRNA complexes are distinct from the ones formed by IDN2 and SWI/SNF, our transcriptome analysis shows a significant overlap between genes affected by the *nrpe1* mutant, *idn2-1* mutant, where IDN2 is unable to interact with lncRNA and the *swi3b*/+ mutant (Figure 3.3F). This is consistent with lncRNA, IDN2 and SWI/SNF working together and suggests that the silencing signal is transmitted from Pol V-produced lncRNA to SWI3B by IDN2 (Figure 3.7).

Defects in transcriptional silencing observed in plant lines defective for the other tested subunits of the SWI/SNF complex (Figures 3.4A-C) suggest that the involvement of SWI3B in transcriptional silencing reflects the involvement of the entire SWI/SNF complex. Interestingly, however, subunit composition of the SWI/SNF complex seems to be locus specific (Figures 3.4A-C). This is consistent with the subunit variants being responsible for functional diversification of the SWI/SNF complexes [21,23,36]. High level of functional diversification of the *Arabidopsis* SWI/SNF complexes is further supported by the observation that mutants in SWI/SNF subunits have additional phenotypes, which are likely not associated with RNA-mediated transcriptional silencing [23,36–39]. This indicates that SWI/SNF complexes also have other biological functions beyond RNA-mediated transcriptional silencing.

The following step in this pathway is nucleosome positioning, as demonstrated by the changes in nucleosome patterns in the *nrpe1* mutant. Our findings are consistent with a model (Figure 3.7), where these effects are mediated by the ATP-dependent chromatin remodeling activity of SWI/SNF, however alternative mechanisms cannot be

excluded at this moment. Pol V-produced lncRNA does not seem to be involved in the establishment of conserved nucleosome patterns [35] on most protein-coding genes. Instead, it specifically mediates stabilization of nucleosomes on sequences enriched in Pol V-dependent non-CG DNA methylation (Figure 3.5B-D), a hallmark of RNA-mediated transcriptional silencing [32]. This is consistent with previously observed genome-wide correlation between positioned nucleosomes and DNA methylation in all sequence contexts [35]. The Pol V-stabilized nucleosomes were also significantly enriched on genes upregulated in the *nrpe1, idn2-1* and *swi3b*/+ mutants (Figure 3.5F), indicating that these nucleosomes may affect Pol II transcription of at least a subset of silencing targets.

The final step of the pathway is repression of Pol II transcription on silenced regions, demonstrated by reduction of transcriptional silencing in *nrpe1*, *idn2* and *swi3b*/+ mutants (Figure 3.3A-E). Locus-specific destabilization of nucleosomes in the *drm2* mutant and partial reduction of DNA methylation in the *swi3b*/+ mutant (Figures 3.5A and 3.6A-F) support a model, where transcriptional silencing is established by Pol V-produced lncRNA guiding both positioned nucleosomes and DNA methylation. On the other hand maintenance of silencing is mediated by continuous action of lncRNA further reinforced by a mutual feedback loop between well positioned nucleosomes and DNA methylation. Pol II is then repressed by DNA methylation of *cis*-regulatory regions and/or well positioned nucleosomes directly affecting the ability of transcriptional machinery to bind DNA.

Our data show the role of nucleosome positioning in the final steps of the RNA-mediated transcriptional silencing pathway. It is however possible that active changes in nucleosome occupancy are also critical at other steps of the pathway and in other silencing pathways. Production of siRNA and Pol V-produced lncRNA has been shown to require two distinct putative chromatin remodelers CLASSY1 and DRD1 [6,40–42]. Additionally, maintenance of DNA methylation requires the DDM1 protein, which has been shown to have a nucleosome remodeling activity *in vitro* [43,44]. This indicates that nucleosome remodeling may play a multitude of roles in transcriptional silencing.

Generally, Involvement of lncRNAs is a common theme in transcriptional regulation in various groups of organisms [45] and guiding protein factors to specific genomic loci is a function of lncRNA in several regulatory processes, including RNA-mediated transcriptional silencing [46]. It is therefore possible that the recruitment of ATP-dependent chromatin remodeling complexes and nucleosome positioning may be a general and conserved feature of lncRNA.

Overall, our results are consistent with a model where lncRNAs produced by Pol V affect gene expression by mediating nucleosome positioning (Figure 3.7). Nascent Pol V transcripts physically interact with IDN2 dimers, which then recruit the SWI/SNF ATP-dependent chromatin remodeling complex by the physical interaction with SWI3B. SWI/SNF positions nucleosomes, which affect transcription machinery. Additionally, maintenance of silencing is reinforced by a feedback loop between DNA methylation and nucleosome positioning.

Materials and Methods

Plant lines

Arabidopsis thaliana nrpe1 (*nrpd1b-11*) mutant was described previously [47]. *swi3a-1* (SALK_035320), *swi3c-2* (Koncz_3737) and *swi3d-2* (Koncz_14259) were kindly provided by Tomasz Sarnowski [23]. *swi3b-2* [23] (GABI_302G08), *syd-4* (SALK_149549) and *brm-4* [48] (WiscDsLox 436E9) were obtained from ABRC. *idn2-1* [16] was kindly provided by Steve Jacobsen. *idn2-2* (FLAG_550B05) was obtained from INRA.

Protein-protein interaction assays

Protein-protein interactions were assayed using yeast two hybrid as well as co-immunoprecipitation of proteins expressed in tobacco leaves or in *Arabidopsis* [49]. Details are provided in the Appendix C.

Site-directed mutagenesis

Entry plasmids containing full length genomic IDN2 or cDNA clones were used to introduce deletions or mutations in *IDN2* using the Quickchange Site-Directed Mutagenesis Kit (Stratagene).

RNA-immunoprecipitation (RNA IP)

RNA IP was performed as described [6] except that IP was performed with 40 µl of Dynabeads protein A, an anti-IDN2 antibody at 4°C over-night and analyzed by real-time RT-PCR. Amplified cDNA was generated with the Ovation RNA-Seq System V2 (Nugen) according to the manufacturer's protocol. Rabbit polyclonal anti-IDN2 antibody was raised against an N-terminal portion of the IDN2 protein (aa 4-201) expressed in bacteria and affinity purified.

Chromatin-immunoprecipitation (ChIP)

ChIP-seq was performed as described [4]. Library generation and Illumina sequencing were performed by the University of Michigan Sequencing Core. ChIP-real time PCR protocol was based on [6] with an additional Micrococcal Nuclease (MNase) digestion prior to IP. We used anti-histone H3 antibody (ab1791, Abcam) or affinity-purified rabbit polyclonal anti-SWI3B antibody raised against a C-terminal portion of the SWI3B protein (aa 248-469). Details are provided in the Appendix C.

MNase-seq

Nuclei were extracted from two weeks old *Arabidopsis* seedlings as described [6] and digested with Micrococcal Nuclease (MNase; NEB). Mononucleosomal DNA was gel-purified and used for library generation and Illumina sequencing. Details are provided in the Appendix C.

Bioinformatic analysis

RNA-seq reads from three independent biological repeats were aligned and processed to the Arabidopsis TAIR10 genome using the BOWTIE suite (BOWTIE, TOPHAT, and CUFFLINKS) [50]. Overlaps between sets of differentially expressed genes were determined based on gene ID while expected values were derived by the formula: set1 (set2/total genes); p values were generated from these values using chi-square test.

MNase-seq and ChIP-seq reads were aligned using BOWTIE default settings and nucleosomes were called using MNase-seq data as described [51]. Nucleosome lists were generated in Col-0 as well as in *nrpe1* and the lists from both genotypes were combined. Reads corresponding to nucleosomes were counted in both genotypes,

quantile normalization was applied and the highest 5% nucleosomes according to read depth were removed. The remaining nucleosomes were classified as either decreased in *nrpe1* (Pol V-stabilized) or increased in *nrpe1* (Pol V-destabilized) via a combination of 2 fold change cutoff and Poisson significance of $p<0.001$; unchanged nucleosomes were classified as having less than 1.5 fold difference between genotypes. Nucleosomes were further filtered using enrichment scores from H3 ChIP-seq with a Poisson significance of $p<0.05$. Published DNA methylation data [12] were overlapped with the nucleosome list and methylation profiles were generated in 10bp windows with a 5bp sliding window for smoothing. Published Pol V ChIP-seq data [25] was similarly overlapped with nucleosomes and profiles were generated with Sitepro from the CEAS suite [52]. Overlaps with differential transcripts were performed using 1000 permutated gene sets to obtain expected numbers and p values.

Accession Numbers

Next generation sequencing data reported in this manuscript have been deposited in GEO (Accession number GSE38464).

Acknowledgements

We thank Ken Cadigan and David Engelke for critical reading of the manuscript, Eva Czarnecka-Verner, Steve Jacobsen and Tomasz Sarnowski for reagents and David Akey for helpful suggestions. We also thank Robert Lyons and Brendan Tarrier from the University of Michigan DNA Sequencing Core for performing next generation sequencing. This work was supported by National Science Foundation grant MCB 1120271 to Andrzej Wierzbicki and Austrian Science Fund (FWF) fellowship J3199-B09 to Gudrun Böhmdorfer. Jordan Rowley was supported by the NIH National Research Service Award #5-T32-GM07544.

Yongyou Zhu initiated the Y2H screen, performed all protein-protein interaction experiments, obtained and analyzed IDN2 M8 mutant and transgenic plants, performed RT-PCRs shown in Figs. 3.3A-E, 3.4A-C, performed all DNA methylation assays and generated anti-IDN2 and anti-SWI3B antibodies. Jordan Rowley performed all ChIP and ChIP-seq experiments as well as all bioinformatic analyses. Gudrun Böhmdorfer

perfomed RNA IP shown in Fig. 3.1A, RT-PCR shown in Fig. 3.4D-H and prepared samples for RNA-seq and MNase-seq.

References

1. Girard, A. & Hannon, G. J. Conserved themes in small-RNA-mediated transposon control. *Trends Cell Biol.* **18,** 136–148 (2008).
2. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11,** 204–220 (2010).
3. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **41,** 563–571 (2009).
4. Zheng, Q. *et al.* RNA polymerase V targets transcriptional silencing components to promoters of protein-coding genes. *Plant J.* (2012). doi:10.1111/tpj.12034
5. Wierzbicki, A. T. The role of long non-coding RNA in transcriptional gene silencing. *Curr. Opin. Plant Biol.* **15,** 517–522 (2012).
6. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135,** 635–648 (2008).
7. Zheng, B. *et al.* Intergenic transcription by RNA polymerase II coordinates Pol IV and Pol V in siRNA-directed transcriptional gene silencing in Arabidopsis. *Genes Dev.* **23,** 2850–2860 (2009).
8. Huang, L. *et al.* An atypical RNA polymerase involved in RNA silencing shares small subunits with RNA polymerase II. *Nat. Struct. Mol. Biol.* **16,** 91–93 (2009).
9. Ream, T. S. *et al.* Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. *Mol. Cell* **33,** 192–203 (2009).
10. Wierzbicki, A. T., Ream, T. S., Haag, J. R. & Pikaard, C. S. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat. Genet.* **41,** 630–634 (2009).
11. Haag, J. R. & Pikaard, C. S. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nat. Rev. Mol. Cell Biol.* **12,** 483–492 (2011).
12. Zhong, X. *et al.* DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. *Nat. Struct. Mol. Biol.* **19,** 870–875 (2012).
13. Bies-Etheve, N. *et al.* RNA-directed DNA methylation requires an AGO4-interacting member of the SPT5 elongation factor family. *EMBO Rep.* **10,** 649–654 (2009).
14. He, X.-J. *et al.* An effector of RNA-directed DNA methylation in arabidopsis is an ARGONAUTE 4- and RNA-binding protein. *Cell* **137,** 498–508 (2009).
15. Rowley, M. J., Avrutsky, M. I., Sifuentes, C. J., Pereira, L. & Wierzbicki, A. T. Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. *PLoS Genet.* **7,** e1002120 (2011).
16. Ausin, I., Mockler, T. C., Chory, J. & Jacobsen, S. E. IDN1 and IDN2 are required for de novo DNA methylation in Arabidopsis thaliana. *Nat. Struct. Mol. Biol.* **16,** 1325–1327 (2009).

17. Zheng, Q. *et al.* Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS Genet.* **6,** e1001141 (2010).
18. Xie, M., Ren, G., Costa-Nunes, P., Pontes, O. & Yu, B. A subgroup of SGS3-like proteins act redundantly in RNA-directed DNA methylation. *Nucleic Acids Res.* **40,** 4422–4431 (2012).
19. Zhang, C.-J. *et al.* IDN2 and its paralogs form a complex required for RNA-directed DNA methylation. *PLoS Genet.* **8,** e1002693 (2012).
20. Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.* **10,** 161–172 (2009).
21. Hargreaves, D. C. & Crabtree, G. R. ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. *Cell Res.* **21,** 396–420 (2011).
22. Sadeh, R. & Allis, C. D. Genome-wide 're'-modeling of nucleosome positions. *Cell* **147,** 263–266 (2011).
23. Sarnowski, T. J. *et al.* SWI3 subunits of putative SWI/SNF chromatin-remodeling complexes play distinct roles during Arabidopsis development. *Plant Cell* **17,** 2454–2472 (2005).
24. Knizewski, L., Ginalski, K. & Jerzmanowski, A. Snf2 proteins in plants: gene silencing and beyond. *Trends Plant Sci.* **13,** 557–565 (2008).
25. Wierzbicki, A. T. *et al.* Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. *Genes Dev.* **26,** 1825–1836 (2012).
26. Ausin, I. *et al.* INVOLVED IN DE NOVO 2-containing complex involved in RNA-directed DNA methylation in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 8374–8381 (2012).
27. Saez, A., Rodrigues, A., Santiago, J., Rubio, S. & Rodriguez, P. L. HAB1-SWI3B interaction reveals a link between abscisic acid signaling and putative SWI/SNF chromatin-remodeling complexes in Arabidopsis. *Plant Cell* **20,** 2972–2988 (2008).
28. Huettel, B. *et al.* Endogenous targets of RNA-directed DNA methylation and Pol IV in Arabidopsis. *EMBO J.* **25,** 2828–2836 (2006).
29. Sarnowski, T. J., Swiezewski, S., Pawlikowska, K., Kaczanowski, S. & Jerzmanowski, A. 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 Res.* **30,** 3412–3421 (2002).
30. Wagner, D. & Meyerowitz, E. M. SPLAYED, a novel SWI/SNF ATPase homolog, controls reproductive development in Arabidopsis. *Curr. Biol.* **12,** 85–94 (2002).
31. Farrona, S., Hurtado, L., Bowman, J. L. & Reyes, J. C. The Arabidopsis thaliana SNF2 homolog AtBRM controls shoot development and flowering. *Development* **131,** 4965–4975 (2004).
32. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11,** 204–220 (2010).
33. Law, J. A. *et al.* A protein complex required for polymerase V transcripts and RNA-directed DNA methylation in Arabidopsis. *Curr. Biol.* **20,** 951–956 (2010).
34. Reeves, R. & Jones, A. Genomic transcriptional activity and the structure of chromatin. *Nature* **260,** 495–500 (1976).

35. Chodavarapu, R. K. *et al.* Relationship between nucleosome positioning and DNA methylation. *Nature* **466,** 388–392 (2010).
36. Bezhani, S. *et al.* Unique, shared, and redundant roles for the Arabidopsis SWI/SNF chromatin remodeling ATPases BRAHMA and SPLAYED. *Plant Cell* **19,** 403–416 (2007).
37. Walley, J. W. *et al.* The chromatin remodeler SPLAYED regulates specific stress signaling pathways. *PLoS Pathog.* **4,** e1000237 (2008).
38. Farrona, S. *et al.* Brahma is required for proper expression of the floral repressor FLC in Arabidopsis. *PLoS ONE* **6,** e17997 (2011).
39. Wu, M.-F. *et al.* SWI2/SNF2 chromatin remodeling ATPases overcome polycomb repression and control floral organ identity with the LEAFY and SEPALLATA3 transcription factors. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 3576–3581 (2012).
40. Kanno, T. *et al.* Involvement of putative SNF2 chromatin remodeling protein DRD1 in RNA-directed DNA methylation. *Curr. Biol.* **14,** 801–805 (2004).
41. Smith, L. M. *et al.* An SNF2 protein associated with nuclear RNA silencing and the spread of a silencing signal between cells in Arabidopsis. *Plant Cell* **19,** 1507–1521 (2007).
42. Law, J. A., Vashisht, A. A., Wohlschlegel, J. A. & Jacobsen, S. E. SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. *PLoS Genet.* **7,** e1002195 (2011).
43. Jeddeloh, J. A., Stokes, T. L. & Richards, E. J. Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nat. Genet.* **22,** 94–97 (1999).
44. Brzeski, J. & Jerzmanowski, A. Deficient in DNA methylation 1 (DDM1) defines a novel family of chromatin-remodeling factors. *J. Biol. Chem.* **278,** 823–828 (2003).
45. Wang, K. C. & Chang, H. Y. Molecular mechanisms of long noncoding RNAs. *Mol. Cell* **43,** 904–914 (2011).
46. Cam, H. P., Chen, E. S. & Grewal, S. I. S. Transcriptional scaffolds for heterochromatin assembly. *Cell* **136,** 610–614 (2009).
47. Pontes, O. *et al.* The Arabidopsis chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell* **126,** 79–92 (2006).
48. Tang, X. *et al.* The Arabidopsis BRAHMA chromatin-remodeling ATPase is involved in repression of seed maturation genes in leaves. *Plant Physiol.* **147,** 1143–1157 (2008).
49. Su, W., Liu, Y., Xia, Y., Hong, Z. & Li, J. Conserved endoplasmic reticulum-associated degradation system to eliminate mutated receptor-like kinases in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 870–875 (2011).
50. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
51. Weiner, A., Hughes, A., Yassour, M., Rando, O. J. & Friedman, N. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res.* **20,** 90–100 (2010).
52. Shin, H., Liu, T., Manrai, A. K. & Liu, X. S. CEAS: cis-regulatory element annotation system. *Bioinformatics* **25,** 2605–2606 (2009).

53. Ausin, I., Mockler, T. C., Chory, J. & Jacobsen, S. E. IDN1 and IDN2 are required for de novo DNA methylation in Arabidopsis thaliana. *Nat. Struct. Mol. Biol.* **16,** 1325–1327 (2009).

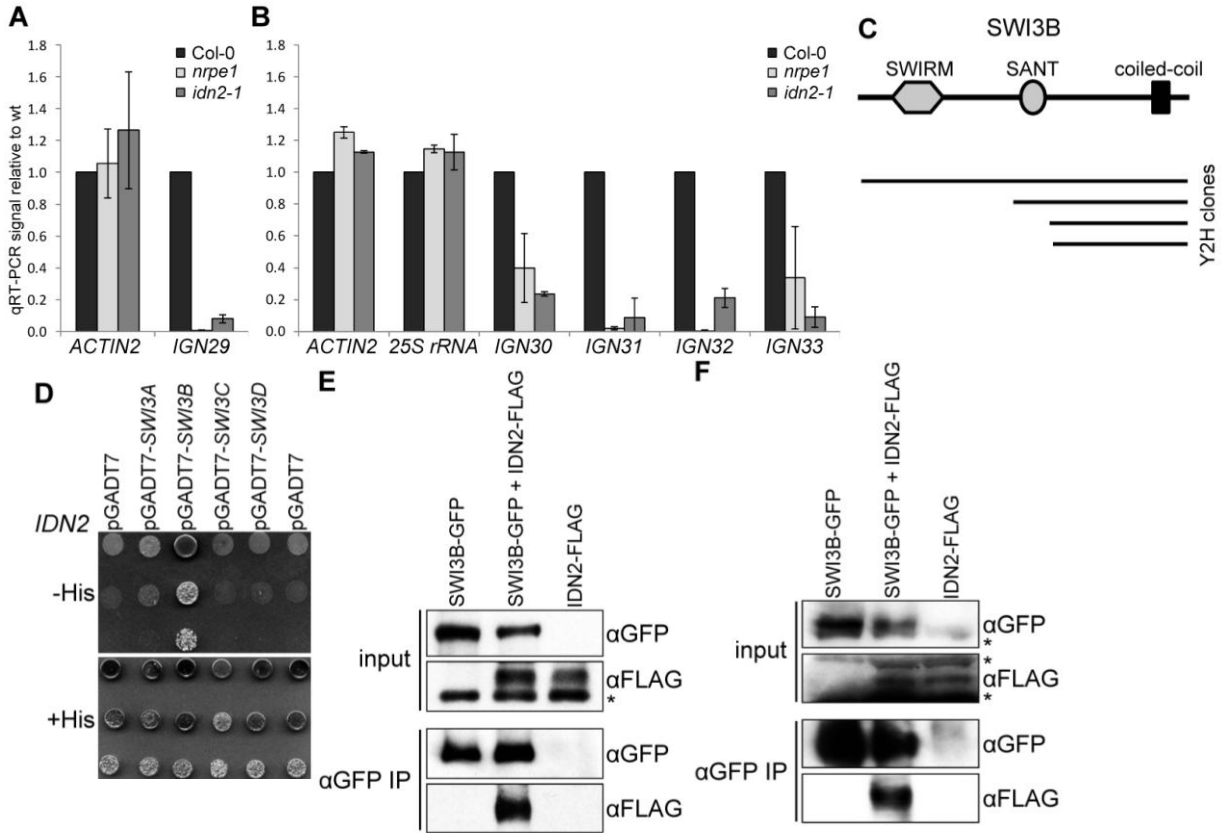


Figure 3.1 IDN2 interacts with lncRNA and SWI3B

A-B) IDN2 interacts with Pol V-produced lncRNA. RNA immunoprecipitation was performed using anti-IDN2 antibody in Col-0 wild type, *nrpe1* mutant and *idn2-1* mutant with a deletion in the RNA-binding XS domain. Recovered RNA was digested with DNase I and assayed using real time RT-PCR A) or reverse transcribed and amplified followed by real time PCR B). *ACTIN2* signal serves as a loading control. Graphs show averages normalized to wild type and SD from four (A) or two (B) biological repeats. Input, no antibody and no RT controls as well as RNA IP results normalized to wild type inputs are shown in Figure 3.8A.

C) Domain structure of SWI3B and SWI3B clones identified using the yeast two-hybrid screen with IDN2 as a bait.

D) IDN2 interacts with SWI3B but not with its homologs SWI3A, SWI3C or SWI3D. Interaction of full length SWI3A, SWI3B, SWI3C and SWI3D with IDN2 was tested using yeast two-hybrid. A series of three 10x dilutions is shown. Yeast growth on a plate with His is shown as a loading control.

E) IDN2 interacts with SWI3B in tobacco. GFP-tagged SWI3B was co-expressed in tobacco leaves with FLAG-tagged IDN2. After immunoprecipitation with anti-GFP antibody the sample was analyzed using western blot with anti-FLAG antibody. Plants expressing only single construct were used as controls. Total protein extracts (inputs) were assayed using western blot to demonstrate comparable protein expression levels. Asterisk indicates a non-specific band. Reciprocal co-immunoprecipitation is shown in Figure 3.8C.

F) IDN2 interacts with SWI3B in *Arabidopsis*. GFP tagged SWI3B and FLAG-tagged IDN2 under the control of their respective native promoters were transformed into *Arabidopsis*. Obtained transgenic lines were crossed and analyzed using co-immunoprecipitation with anti-GFP antibody and western blot with anti-FLAG antibody. Asterisks indicate non-specific bands.

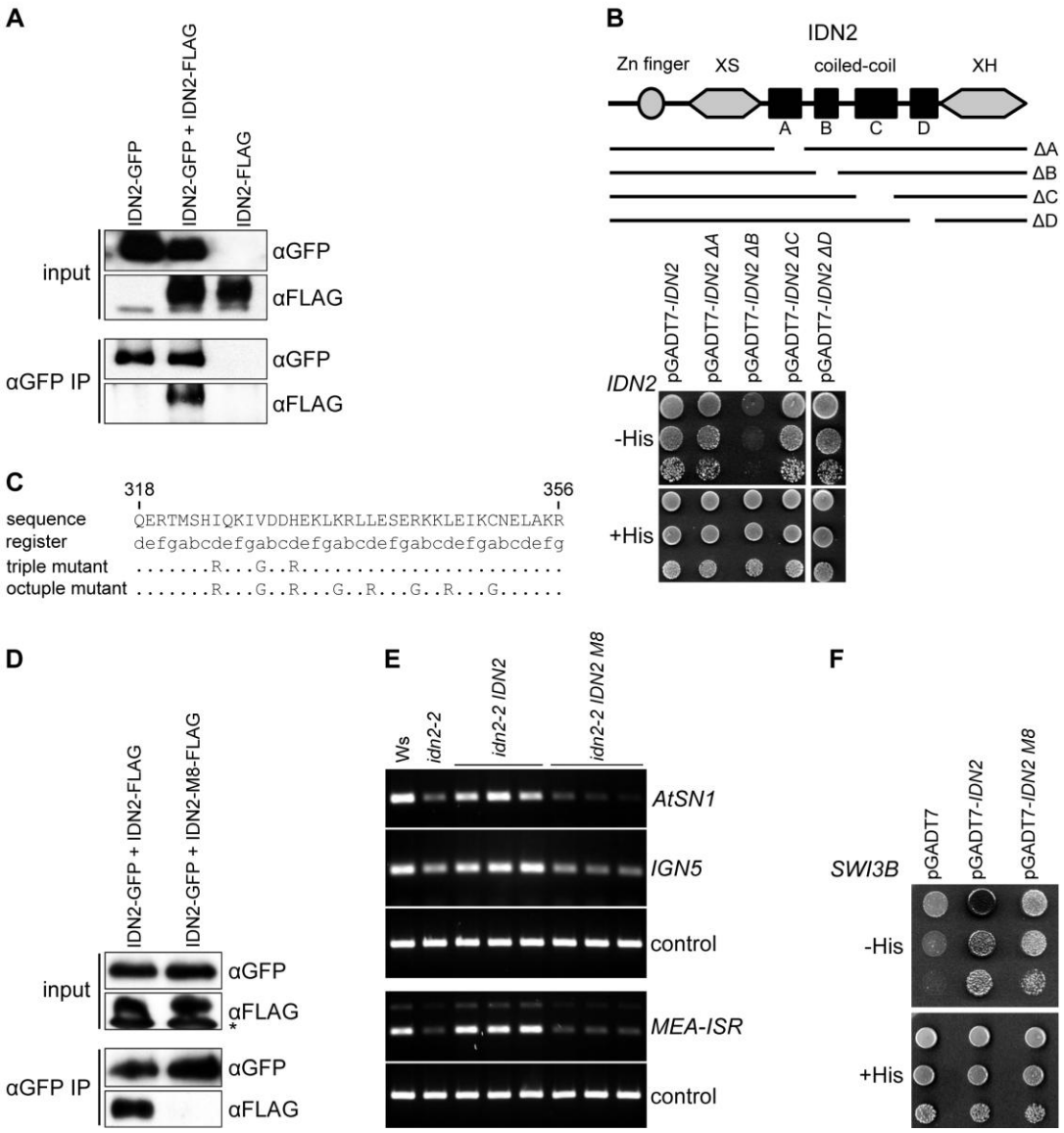


Figure 3.2 IDN2 homodimerization is required for silencing but not for the interaction with SWI3B

A) IDN2 dimerizes. GFP- and FLAG-tagged IDN2 were coexpressed in tobacco leaves. After immunoprecipitation with anti-GFP antibody the sample was analyzed using western blot with anti-FLAG antibody.

B) Subdomain B within coiled-coil region of IDN2 is responsible for IDN2 dimerization. Four coiled-coil sub-domains were identified within IDN2 using Parcoil2 and corresponding deletion mutants were tested using yeast two hybrid for interaction with wild type IDN2.

C) Point mutants within subdomain B of the coiled-coil region of IDN2. Point mutants were designed to contain three (triple) or eight (octuple, M8) aminoacids within registers A and D of the coiled-coil alpha helix changed to arginines or glycines to disrupt interactions mediated by the coiled-coil with minimal impact on the alpha-helix of the coiled-coil domain.

D) Dimerization is lost in IDN2 with a mutated coiled-coil domain. GFP-tagged wild type IDN2 was coexpressed in tobacco leaves with FLAG-tagged IDN2 M8. After immunoprecipitation with anti-GFP antibody the sample was analyzed using western blot with anti-FLAG antibody. FLAG-tagged wild type IDN2 was used as a control. Total protein extracts (inputs) were assayed using western blot to demonstrate comparable protein expression levels. Asterisk indicates a non-specific band.
E) IDN2 dimerization is required for DNA methylation. *idn2-2* knock out mutant *Arabidopsis* plants were transformed with wild type *IDN2* and *IDN2 M8*. Flowers of obtained transgenic plants were assayed for changes in DNA methylation by digesting with DNA methylation-sensitive restriction endonucleases (*Hae*III for *AtSN1* and *IGN5*, *Sau3*AI for *MEA-ISR*) followed by PCR. Sequences with no restriction sites were used as controls (*ACTIN2* for *Hae*III and JA35/JA36 for *Sau3*AI). More independent transgenic lines are shown in Figure 3.9B.
F) IDN2 dimerization is not required for interaction with SWI3B. IDN2 M8 mutant within the subdomain B of the coiled-coil region was tested for interaction with SWI3B using yeast two hybrid.

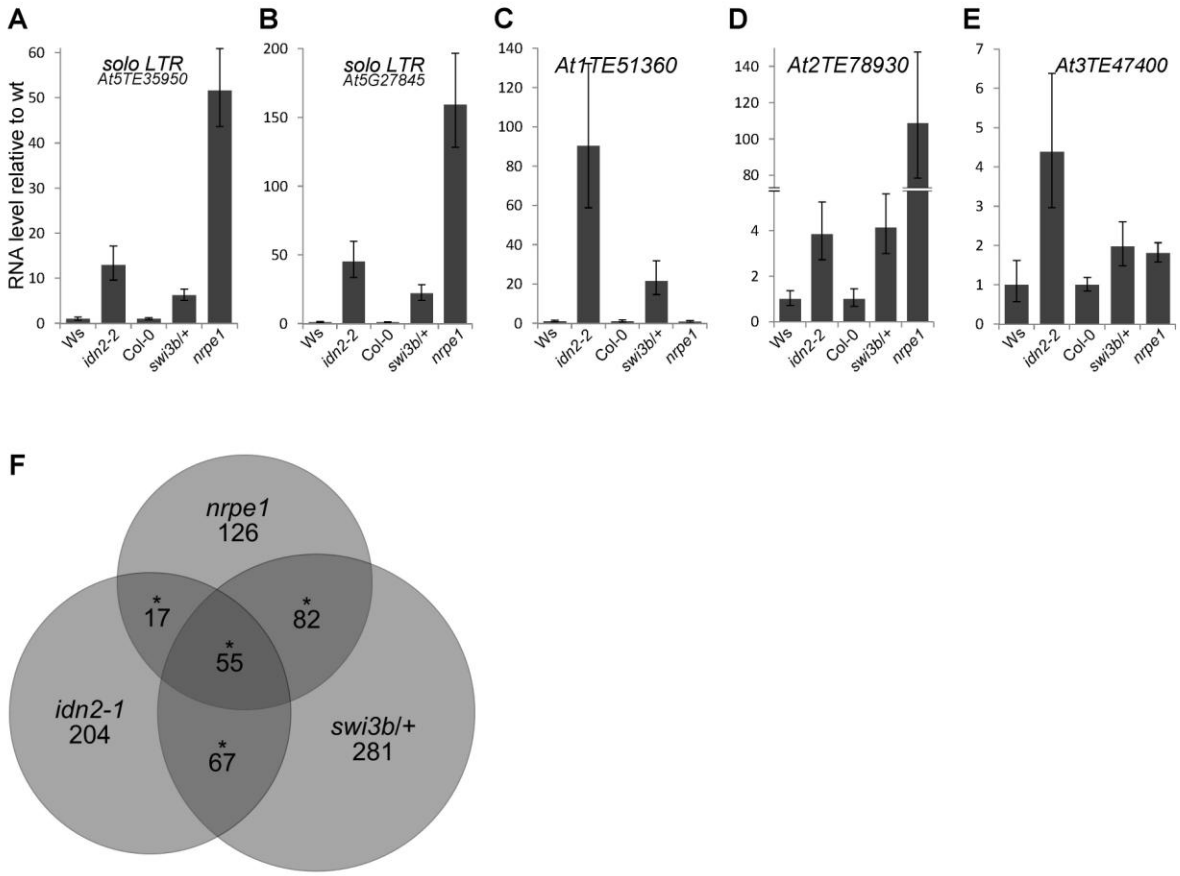


Figure 3.3 SWI3B contributes to RNA-mediated transcriptional silencing
A-E) Silencing targets are derepressed in the *swi3b*/+ mutant. RNA accumulation in flowers from *solo LTR* (A-B), *At1TE51360* (C), *At2TE78930* (D) and *At3TE47400* (E) were assayed using real time RT-PCR in *idn2-2* mutant compared to Ws wild type and in *swi3b*/+ and *nrpe1* mutants compared to Col-0 wild type. Graphs show averages normalized to *ACTIN2* and SD from three biological repeats. More loci are shown in Figure 3.10B-E.
F) SWI3B controls the expression levels of a significant subset of Pol V and IDN2 targets. Venn diagram showing genes identified using RNA-seq to be upregulated in *nrpe1, idn2-1* and/or *swi3b*/+ mutants. RNA-seq was performed in three independent biological repeats in seedlings. * denotes statistically significant enrichment of overlaps (see text and Figure 3.10G for details).

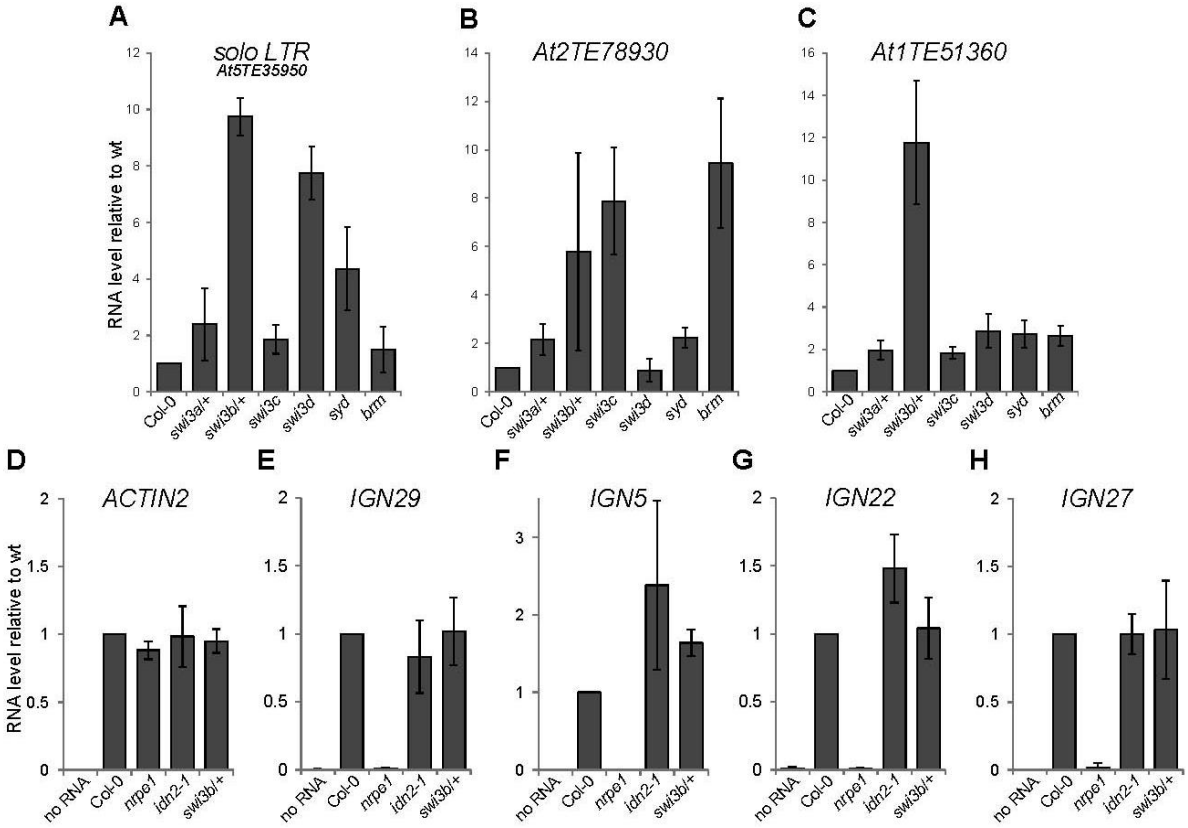


Figure 3.4 RNA-mediated transcriptional silencing involves the SWI/SNF complex, which works downstream of lncRNA production

A-C) Silencing targets are derepressed in mutants defective in SWI/SNF subunits. RNA accumulation in flowers from *solo LTR* (A), *At2TE78930* (B) or At1TE51360 (C) was assayed using real time RT-PCR in *swi3a/+, swi3b/+, swi3c, swi3d, syd* and *brm* mutants compared to Col-0 wild type. Graphs show averages normalized to *ACTIN2* and SD from three biological repeats, normalized to Col-0 wild type.

D-H) IDN2 and SWI3B are not required for lncRNA production. Pol V-produced lncRNAs *IGN29* (E), *IGN5* (F), *IGN22* (G) and *IGN27* (H) were assayed in seedlings using real time RT-PCR in *idn2-1* and *swi3b/+* mutants compared to Col-0 wild type. *nrpe1* mutant was used as a negative control. To check for potential DNA contaminations no RT control was performed on *ACTIN2* (Figure 3.11E) and additionally no RNA controls were performed for all primer pairs tested. *ACTIN2* (D) is a loading control. More loci are shown in Figures 3.11F-I.

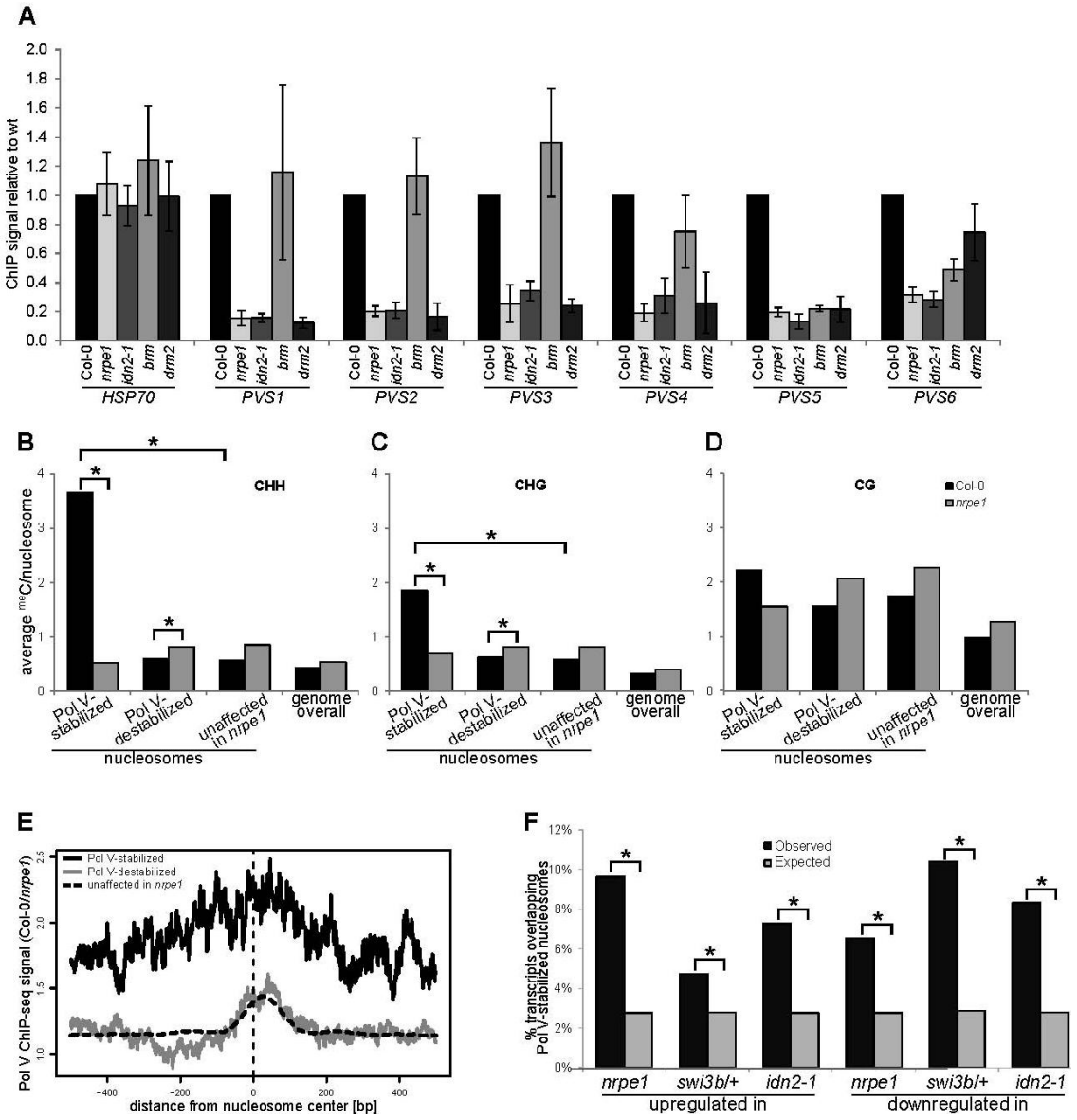


Figure 3.5 Pol V mediates nucleosome positioning

A) Validation of Pol V-stabilized nucleosomes. Nucleosomes identified using genome-wide assays (Figure 3.12B) were assayed in seedlings and in case of *brm* in mature leaves with MNase digestion followed by H3 ChIP and real time PCR. ChIP signal values were normalized to *ACTIN2* and to wild type. *HSP70* is a negative control. Bars show averages and SD from three independent biological repeats.

B-D) Nucleosomes stabilized by Pol V are enriched in Pol V-dependent DNA methylation. Published genome-wide DNA methylation datasets from Col-0 wild type and *nrpe1* mutant were used to calculate average DNA methylation levels on nucleosomes identified using MNase-seq and H3 ChIP-seq. Pol V-stabilized and Pol V-destabilized nucleosomes were compared to nucleosomes

unaffected in *nrpe1* and to the entire genome (genome overall). DNA methylation levels were independently calculated in CHH (B), CHG (C) and CG (D) contexts. Asterisks indicate significant enrichment (see text).
E) Pol V-stabilized nucleosomes are enriched in Pol V binding. Published Pol V ChIP-seq dataset was used to calculate profiles of Pol V binding around the centers of nucleosomes identified using MNase-seq and H3 ChIP-seq, which are Pol V-stabilized, Pol V-destabilized or unaffected in *nrpe1*.
F) Pol V-stabilized nucleosomes are enriched on genes upregulated or downregulated in *nrpe1* and *swi3b*/+ mutants. Nucleosomes identified using MNase-seq as Pol V-stabilized were overlapped with genes identified using RNA-seq as upregulated or downregulated in *nrpe1* or *swi3b.* Permutations of gene sets were overlapped in parallel to calculate enrichment. Asterisks indicate significant enrichment ($p<0.02$).

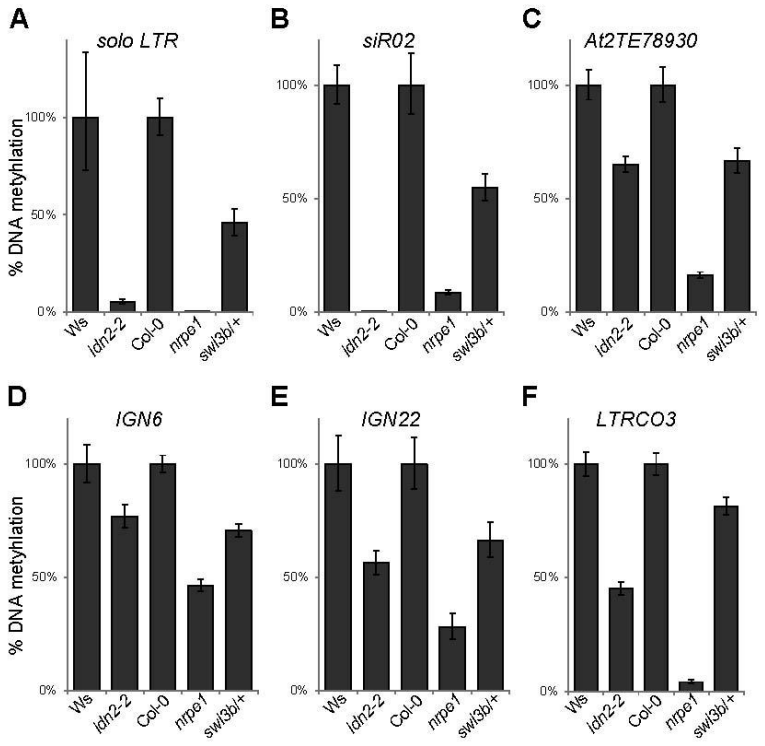


Figure 3.6 SWI/SNF is required for wild type levels of CHH methylation
A-F) DNA methylation was assayed in flowers using digestion with methylation sensitive restriction endonucleases followed by real time PCR amplification of *solo LTR* (A), *siR02* (B), *At2TE78930* (C), *IGN6* (D), *IGN22* (E) and *LTRCO3* (F). Graphs show average DNA methylation levels normalized to *ACTIN2* and to wild type. Error bars are SD from three biological repeats.

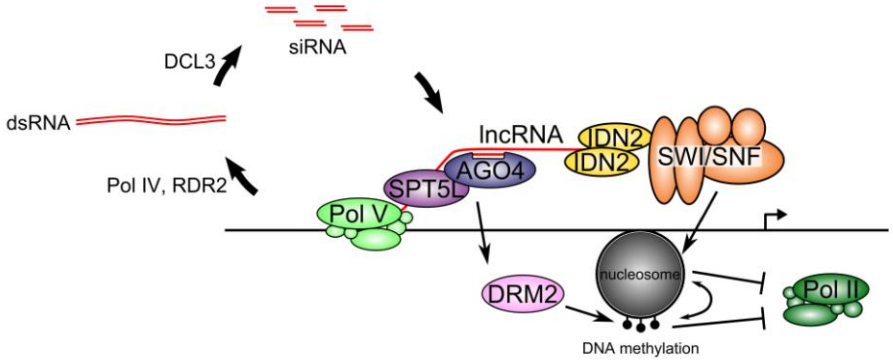


Figure 3.7 A model of the involvement of nucleosome positioning in lncRNA-mediated transcriptional silencing

siRNA is produced by the activities of Pol IV, RDR2 and DCL3 to give AGO4 sequence-specificity. Pol V produces lncRNAs, which are bound by IDN2 dimers. IDN2 recruits the SWI/SNF ATP-dependent chromatin remodeling complex by physical interaction with SWI3B. The SWI/SNF complex positions nucleosomes. Positioned nucleosomes silence transcription directly or by facilitating DNA methylation by the *de novo* methyltransferase, DRM2. Maintenance of silencing is reinforced by a positive feedback between DNA methylation and nucleosome positioning.

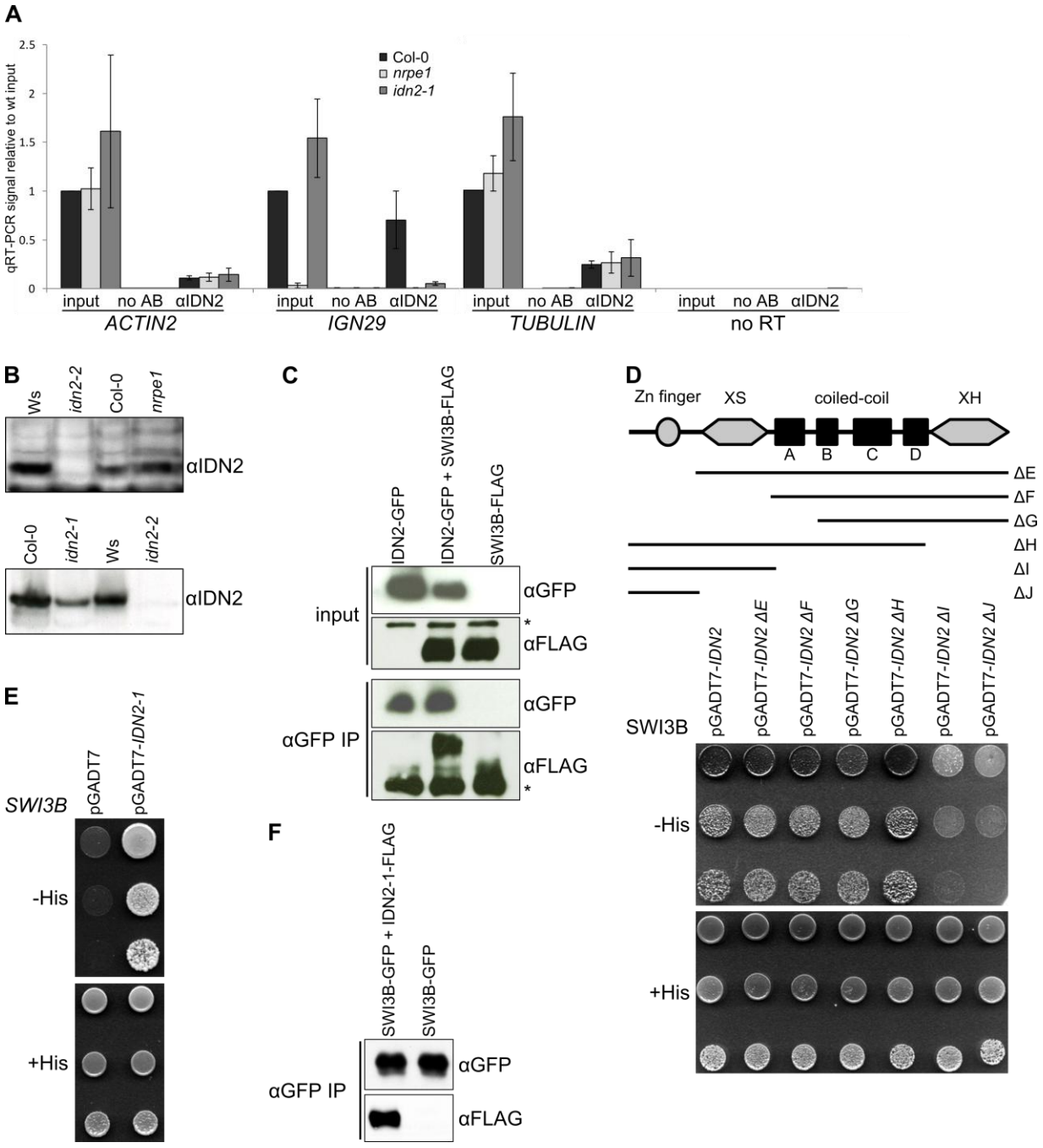


Figure 3.8. IDN2 interacts with SWI3B (supplementary to Figure 3.1)
A) IDN2 interacts with Pol V-produced lncRNA. Input, no antibody and no RT controls as well as RNA IP results shown in Figure 3.1A were normalized to wild type input. No RT control was performed using *ACTIN2* primers. Bars show averages normalized to wild type and standard deviations from four biological repeats.
B) IDN2 protein levels. Equal amounts of total protein extracts from *idn2* mutants, *nrpe1* mutant and corresponding wild type controls were assayed using western blot with affinity purified rabbit polyclonal anti-IDN2 antibody.

C) IDN2 interacts with SWI3B in tobacco. FLAG-tagged SWI3B was coexpressed in tobacco leaves with GFP-tagged IDN2. After immunoprecipitation with anti-GFP antibody the sample was analyzed using western blot with anti-FLAG antibody. Plants expressing only single construct were used as controls. Total protein extracts (inputs) were assayed using western blot to demonstrate comparable protein expression levels. Asterisks indicate non-specific bands. Reciprocal co-immunoprecipitation is shown in Figure 3.1E.
D) IDN2 interacts with SWI3B using its coiled-coil domain. Truncated IDN2 was assayed for interactions with SWI3B using yeast two hybrid. A series of three 10x dilutions is shown. Yeast growth on a plate with His is shown as a loading control.
E) The XS domain within IDN2 is not required for interaction with SWI3B in yeast two hybrid assay. A deletion mutant in the XS domain of IDN2 corresponding to the *idn2-1* mutant was tested for interaction with SWI3B using yeast two hybrid. A series of three 10x dilutions is shown. Yeast growth on a plate with His is shown as a loading control.
F) The XS domain within IDN2 is not required for interaction with SWI3B in tobacco leaves. A FLAG-tagged deletion mutant in the XS domain of IDN2 corresponding to the *idn2-1* mutant was coexpressed with GFP-tagged SWI3B in tobacco leaves. After immunoprecipitation with anti-GFP antibody the sample was analyzed using western blot with anti-FLAG antibody.

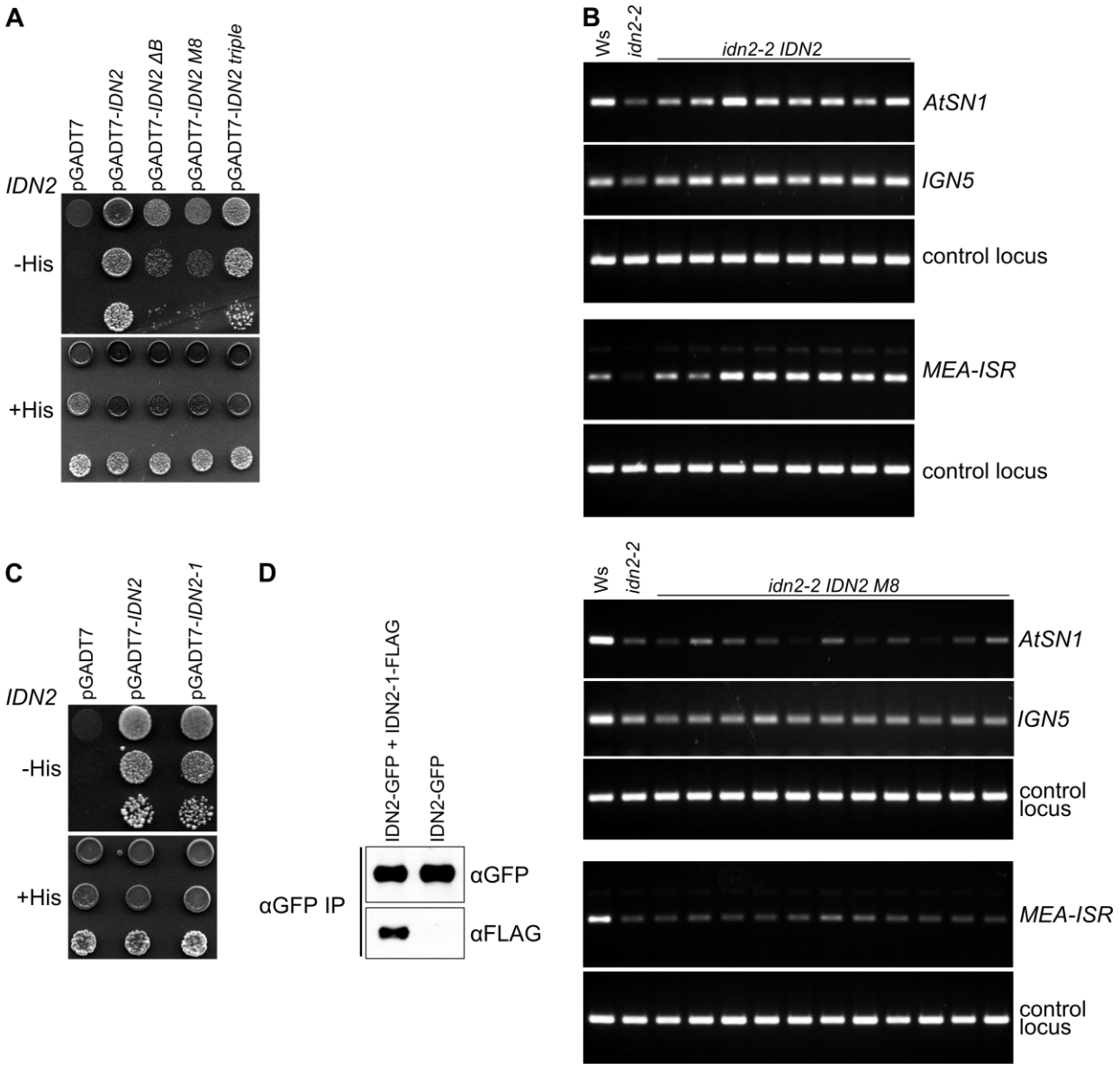
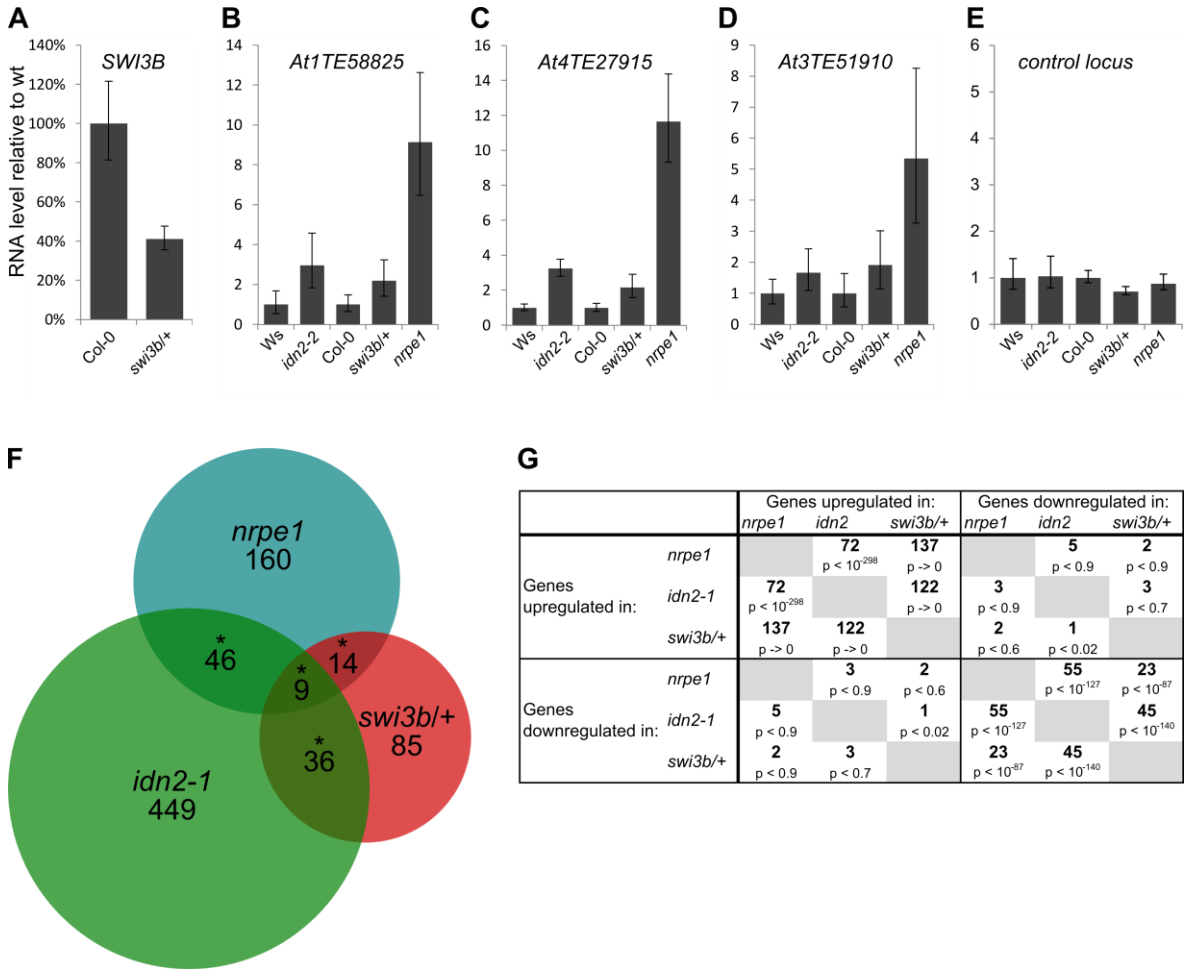


Figure 3.9 Characterization of IDN2 dimerization domain and its functional significance (supplementary to Figure 3.2)

A) Octuple mutations in subdomain B of the coiled-coil region disrupt IDN2 dimerization. Interaction of wild type IDN2, IDN2 with subdomain B deleted, IDN2 with octuple mutations (M8) and IDN2 with triple mutations were assayed for interaction with wild type IDN2 using yeast two-hybrid. A series of three 10x dilutions is shown. Yeast growth on a plate with His is shown as a loading control.

B) IDN2 dimerization is required for its function – additional independent transgenic lines extending the result shown in Figure 3.2E. *idn2-2* knock out mutant *Arabidopsis* plants were transformed with wild type *IDN2* or *IDN2 M8*. Obtained transgenic plants were assayed for changes in DNA methylation by digesting with DNA methylation-sensitive restriction endonucleases (*Hae*III for *AtSN1* and *IGN5*, *Sau3*AI for *MEA-ISR*) followed by PCR. Transformation of the *idn2-2* knock out mutant with wild type *IDN2*

restored DNA methylation to wild type levels at all tested loci. *IDN2 M8* was unable to restore DNA methylation at any of the tested loci. Sequences with no restriction sites were used as controls (*ACTIN2* for *Hae*III and JA35/JA36 for *Sau3*AI).
C) The XS domain of IDN2 is not required for dimerization in yeast two hybrid. A mutated IDN2 corresponding to the *idn2-1* mutant was tested for interaction with wild type IDN2 using yeast two hybrid. A series of three 10x dilutions is shown. Yeast growth on a plate with His is shown as a loading control.
D) The XS domain of IDN2 is not required for dimerization in tobacco leaves. A FLAG-tagged deletion mutant in the XS domain of IDN2 corresponding to the *idn2-1* mutant [53] was coexpressed with GFP-tagged wild type IDN2 in tobacco leaves. After immunoprecipitation with anti-GFP antibody the sample was analyzed using western blot with anti-FLAG antibody.



G

		Genes upregulated in:			Genes downregulated in:		
		nrpe1	*idn2*	*swi3b/+*	*nrpe1*	*idn2*	*swi3b/+*
Genes upregulated in:	*nrpe1*		**72** $p < 10^{-298}$	**137** p -> 0		**5** p < 0.9	**2** p < 0.9
	idn2-1	**72** $p < 10^{-298}$		**122** p -> 0	**3** p < 0.9		**3** p < 0.7
	swi3b/+	**137** p -> 0	**122** p -> 0		**2** p < 0.6	**1** p < 0.02	
Genes downregulated in:	*nrpe1*		**3** p < 0.9	**2** p < 0.6		**55** $p < 10^{-127}$	**23** $p < 10^{-87}$
	idn2-1	**5** p < 0.9		**1** p < 0.02	**55** $p < 10^{-127}$		**45** $p < 10^{-140}$
	swi3b/+	**2** p < 0.9	**3** p < 0.7		**23** $p < 10^{-87}$	**45** $p < 10^{-140}$	

Figure 3.10 SWI3B contributes to RNA-mediated transcriptional silencing (supplementary to Figure 3.3)

A) *SWI3B* expression is reduced in the *swi3b*/+ line. RNA accumulation of *SWI3B* was assayed using real time RT-PCR in *swi3b*/+ mutant compared to Col-0 wild type. Graphs show averages and standard deviations from three biological repeats.

B-E) Silencing targets are derepressed in *swi3b*/+ mutant. RNA accumulation from *At1TE58825*, *At4TE27915* and *At3TE51910* was assayed using real time RT-PCR in *idn2-2* mutant compared to Ws wild type and in *swi3b*/+ and *nrpe1* mutants compared to Col-0 wild type. *UBQ10* was tested as a control (E). Graphs show averages normalized to *ACTIN2* and wild type and standard deviations from three biological repeats.

F) SWI3B controls the expression levels of a significant subset of Pol V and IDN2 targets. Venn diagram showing genes identified using RNA-seq to be downregulated in *nrpe1, idn2-1* and/or *swi3b*/+ mutants. RNA-seq was performed in three independent biological repeats. * denotes statistically significant enrichment of overlaps (see text and Figure 3.10G for details).

G) Overlaps between genes identified by RNA-seq to be upregulated or downregulated in the analyzed mutants. p-values correspond to the observed overlap compared to overlap expected by chance and were obtained using chi-square test.

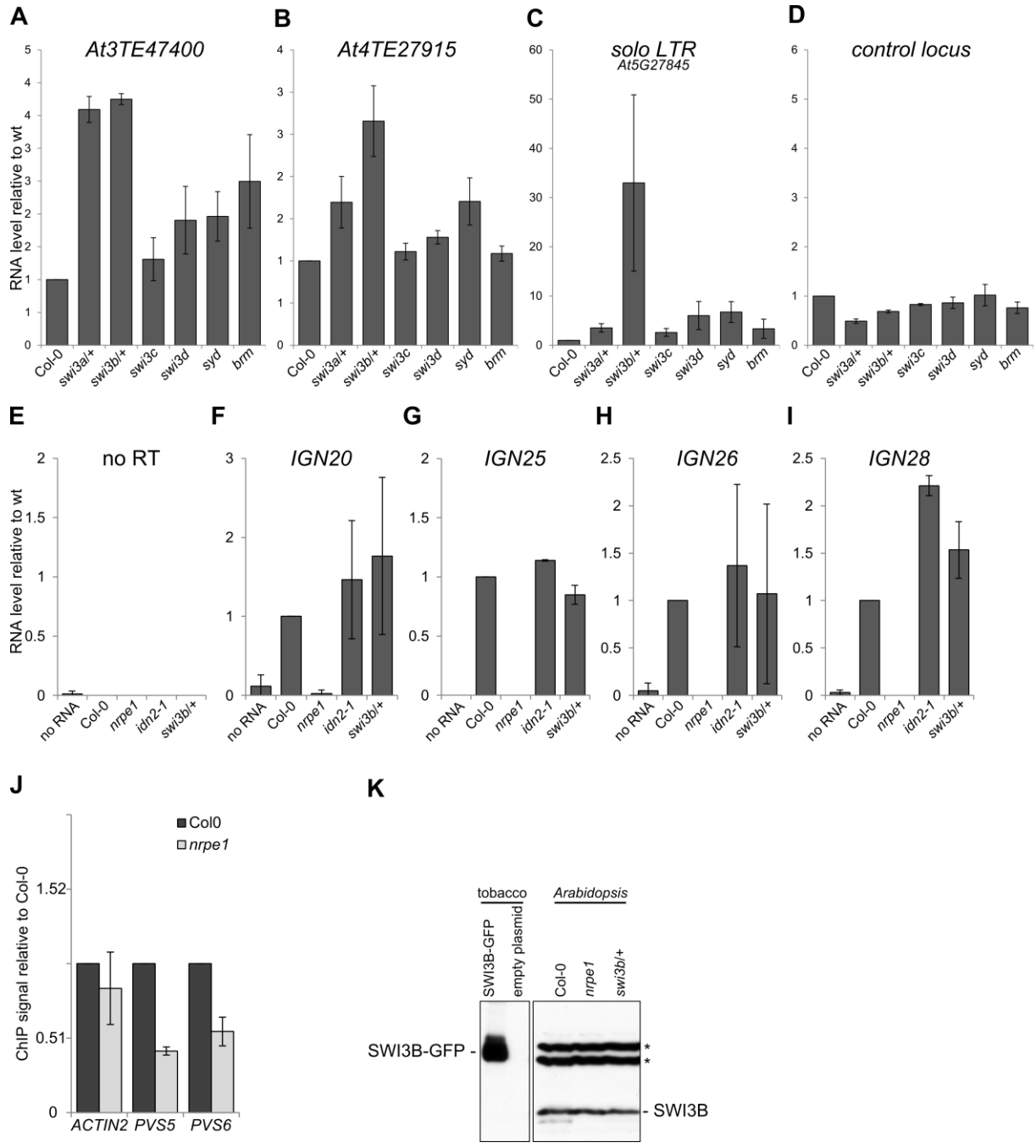


Figure 3.11 RNA-mediated transcriptional silencing involves the SWI/SNF complex, which works downstream of lncRNA production (supplementary to Figure 3.4)

A-D) Silencing targets are derepressed in mutants defective in SWI/SNF subunits. RNA accumulation from *At3TE47400* (A), *At4TE27915* (B) or *At5G27845* (C) was assayed using real time RT-PCR in *swi3a/+, swi3b/+, swi3c, swi3d, syd* and *brm* mutants compared to Col-0 wild type. *ROC3* was tested as a control (D). Graphs show averages

normalized to *ACTIN2* and wild type and standard deviations from three biological repeats.
E-I) IDN2 and SWI3B function downstream of lncRNA production. Pol V-produced lncRNAs *IGN20* (F), *IGN25* (G), *IGN26* (H) and *IGN28* (I) were assayed using real time RT-PCR in *idn2-1* and *swi3b*/+ mutants compared to Col-0 wild type. *nrpe1* mutant was used as a negative control. To check for potential DNA contaminations no RT control was performed on *ACTIN2* (E) and additionally no RNA controls were performed for all primer pairs tested. Graphs show averages normalized to wild type and standard deviations from three biological repeats.
J) SWI3B binding to chromatin is reduced in the *nrpe1* mutant. ChIP with anti-SWI3B antibody was performed in Col-0 wild type and *nrpe1* mutant. Bars show averages and standard deviations from three biological repeats, normalized to inputs and Col-0. Western blot showing antibody specificity is in (K).
(K) Western blot showing specificity of anti-SWI3B antibody. Total proteins from tobacco leaves expressing epitope-tagged SWI3B and from Col-0, *nrpe1* and *swi3b*/+ *Arabidopsis* plants were assayed using affinity-purified anti-SWI3B antibody. Asterisks indicate non-specific bands detectable only in *Arabidopsis* extracts.

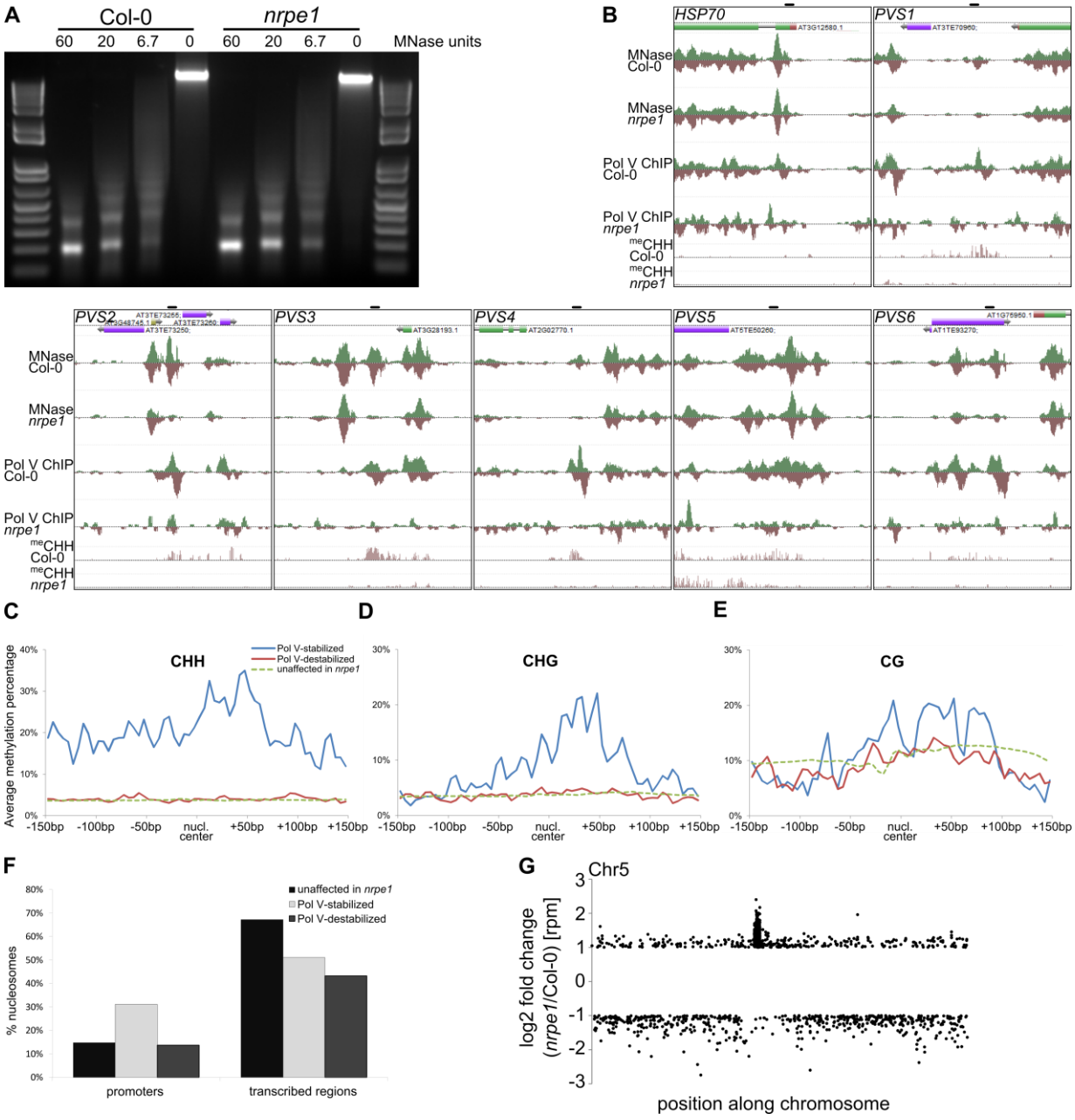


Figure 3.12 Pol V mediates nucleosome positioning (supplementary to Figure 3.5)
A) Micrococcal Nuclease (MNase) digestion of nuclei from Col-0 wild type and *nrpe1* mutant. MNase activity is shown in Kunitz Units. Mononucleosomal DNA was later sequenced using Illumina sequencing.
B) Genome browser screenshots showing regions of Pol V-stabilized nucleosomes selected for validation (Figure 3.5A). Shown data include from top: annotation, MNase-seq in Col-0 wild type and *nrpe1* mutant, Pol V ChIP-seq in Col-0 wild type and *nrpe1* mutant, CHH methylation in Col-0 wild type and *nrpe1* mutant.
C-E) Nucleosomes stabilized by Pol V are enriched in Pol V-dependent DNA methylation. Profiles of CHH methylation (C), CHG methylation (D) and CG methylation

(E) were calculated and plotted on nucleosomes identified using MNase-seq and H3 ChIP-seq with nucleosome center in the middle of each graph. Nucleosomes unaffected in *nrpe1* were tested as controls. Published DNA methylation data were used.
F) Nucleosomes stabilized by Pol V are enriched on gene promoters. Nucleosomes identified using MNase-seq were overlapped with gene promoters and transcribed regions.
G) Nucleosomes stabilized by Pol V are distributed throughout the chromosomes but nucleosomes destabilized by Pol V are enriched at the centromere. Differential nucleosomes identified using MNase-seq were plotted on the chromosome 5 with the fold value of the change in *nrpe1* mutant compared to Col-0 wild type. Pol V-stabilized nucleosomes have negative and Pol V-destabilized nucleosomes have positive enrichment values.

CHAPTER 4

Independent Chromatin Binding of ARGONAUTE4 and SPT5L/KTF1 Mediates Transcriptional Gene Silencing

The contents of this chapter were published in PLOS Genetics in 2011. Maria Avrutsky performed experiments shown in Figure 4.4J. Christopher Sifuentes' work is shown in Figure 4.1K. Ligia Pereira's work is shown in Figure 4.1J and 4.3J. I performed all other experiments shown in this chapter.

Abstract

Eukaryotic genomes contain significant amounts of transposons and repetitive DNA elements, which, if transcribed, can be detrimental to the organism. Expression of these elements is suppressed by establishment of repressive chromatin modifications. In *Arabidopsis thaliana*, they are silenced by the siRNA-mediated transcriptional gene silencing pathway where long non-coding RNAs (lncRNAs) produced by RNA Polymerase V (Pol V) guide ARGONAUTE4 (AGO4) to chromatin and attract enzymes that establish repressive chromatin modifications. It is unknown how chromatin modifying enzymes are recruited to chromatin. We show through chromatin immunoprecipitation (ChIP) that SPT5L/KTF1, a silencing factor and a homolog of SPT5 elongation factors, binds chromatin at loci subject to transcriptional silencing. Chromatin binding of SPT5L/KTF1 occurs downstream of RNA Polymerase V, but independently from the presence of 24-nt siRNA. We also show that SPT5L/KTF1 and AGO4 are recruited to chromatin in parallel and independently of each other. As shown using methylation sensitive restriction enzymes, binding of both AGO4 and SPT5L/KTF1 is required for DNA methylation and repressive histone modifications of several loci. We propose that the coordinate binding of SPT5L and AGO4 creates a platform for direct or indirect recruitment of chromatin modifying enzymes.

Author Summary

Transposons and other repetitive elements occupy vast areas of the eukaryotic genomes. They pose a threat to genome integrity but at the same time regulate expression of many genes and have been proposed to be a major factor contributing to genome evolution. One of the processes responsible for controlling activity of transposons and other repetitive elements is transcriptional gene silencing. This process uses small interfering RNA and long non-coding RNA to recruit enzymes that establish repressive chromatin modifications. Several proteins have been identified to be needed for siRNA-mediated transcriptional silencing in *Arabidopsis thaliana*, however for many of them their position in the silencing pathway is unknown. One of those proteins is SPT5L/KTF1, a homolog of an elongation factor associated with RNA Polymerase II. Here we establish the position of SPT5L in the silencing pathway and propose the molecular mechanism of its function. This gives further knowledge of the mechanism of transcriptional gene silencing and is important to understand how transposons are controlled.

Introduction

Eukaryotic genomes contain significant amounts of transposons and other repetitive DNA elements, which usually remain transcriptionally inactive. Efficient silencing of transposon transcription is essential for preventing their mobility and for maintaining genome integrity [1]. Transposon silencing has also been hypothesized to regulate expression of genes that contain transposable elements in their promoters and to facilitate the evolution of genomes [2].

Transposons are silenced at both transcriptional and post-transcriptional levels by mechanisms that involve small interfering RNAs (siRNAs) [3]. These 20-25-nt RNA molecules are generated by the RNase III enzyme Dicer and provide sequence specificity for effector complexes mediating RNA cleavage and/or the establishment of chromatin modifications that silence transcriptional activity [3]. In *Arabidopsis thaliana,* single-stranded RNA precursors for siRNA biogenesis are produced by RNA Polymerase II (Pol II) or RNA Polymerase IV (Pol IV), while the second strand is synthesized by RDR2 (RNA-Dependent RNA Polymerase 2). DCL3 (Dicer-like 3)

cleaves double-stranded RNA into siRNAs that are then incorporated into ARGONAUTE4 (AGO4) [4,5]. This mechanism seems to be similar in maize where homologs of RDR2 and Pol IV have been shown to be involved in transcriptional gene silencing [6–8].

Recognition of target loci by AGO4-siRNA complexes requires sequence identity between siRNAs and the genomic loci. These loci, however, are often actively transcribed, and it is not clear if siRNAs base-pair interact with DNA or nascent RNA transcripts [3,9]. The latter possibility is well supported in *Schizosaccharomyces pombe* where loci subject to siRNA-mediated transcriptional silencing are actively transcribed by RNA Polymerase II [10–12]. The central role of nascent transcripts in recognition of siRNA targets in *S. pombe* was observed by the ability of Argonaute proteins to cleave RNA. This ability is required for the establishment of repressive chromatin modifications [13]. Moreover, tethering Argonaute and siRNA-containing RITS (RNA-induced initiation of transcriptional gene silencing) complex to nascent transcripts is sufficient for the initiation of repressive chromatin modifications and transcriptional silencing [14].

This mechanism may be similar in *Arabidopsis* where transcriptional silencing requires a specialized RNA Polymerase complex known as RNA Polymerase V (Pol V) [15–17]. Pol V produces non-coding transcripts in otherwise silent chromatin, and its activity is required for the establishment and maintenance of repressive chromatin modifications [18]. Pol V-produced non-coding transcripts physically interact with AGO4 and recruit siRNA-AGO4 complexes to their targets [19]. Additionally, transcriptional silencing of several loci needs AGO4 slicer activity [20], suggesting that in plants siRNAs may recognize their targets by base-pair interactions with Pol V transcripts [19].

RNA Polymerases and AGO4 are assisted in their functions by several other known protein components of the plant silencing system, all of which are required for efficient establishment and maintenance of transcriptional silencing [5]. One of them is SPT5L (Suppressor of Ty insertion 5 - like; also known as SPT5-like or KTF1), a homolog of SPT5 Pol II-associated elongation factor. It was shown to contain a domain rich in WG/GW repeats that facilitate physical interaction with AGO4 [21–23]. Because SPT5L interacts with RNA but is not required for the accumulation of Pol V-dependent transcripts, it was hypothesized to work downstream of Pol V and recruit AGO4 to Pol V-

transcribed loci [22,24].

Despite the recent progress in understanding the mechanisms of transcriptional gene silencing, it is not known how siRNAs work with Pol V transcripts, AGO4 and other proteins to recruit chromatin modifying enzymes to their target loci in chromatin. It is unknown how chromatin-bound AGO4 recruits enzymes that establish repressive chromatin modifications. It is unknown if other protein components of the silencing system help AGO4 recruit chromatin modifying enzymes. It is also unknown in what order proteins involved in silencing are recruited to chromatin. Here we try to resolve the mechanism of siRNA-mediated recruitment of chromatin modifying enzymes to chromatin and the function of SPT5L in this process. We show that SPT5L physically interacts with chromatin and that SPT5L works downstream of Pol V but does not require 24-nt siRNA. SPT5L and AGO4 are recruited to chromatin in parallel and at least partially independently of each other and both are needed for DNA methylation and repressive histone modifications at several loci. We propose that the coordinate binding of SPT5L and AGO4 creates a platform for direct or indirect recruitment of chromatin modifying enzymes.

Results

SPT5L interacts with chromatin

The interaction of SPT5L with AGO4 [21,22] suggested that like AGO4 [19], SPT5L may bind loci targeted by siRNA-mediated transcriptional gene silencing. We first used chromatin immunoprecipitation (ChIP) with anti-SPT5L antibody to test if SPT5L binds chromatin. Subsequent real-time PCR demonstrated recovery of *IGN5* and *solo LTR* DNA from Col-0 wild type at much higher levels than from *spt5l* mutant which represents the background level (Fig. 4.1CD). This shows that SPT5L physically interacts with *IGN5* and *solo LTR* loci which are known to be transcribed by Pol V and silenced by the siRNA-mediated transcriptional gene silencing pathway [18,19,25,26]. There was, however, no enrichment on the control *Actin 2* and *Tubulin 8* (*TUB 8*) loci (Fig. 4.1AB), which are transcribed by Pol II and not occupied by components of the silencing pathway [18,27]. This suggests that SPT5L is present at the loci undergoing transcriptional silencing and that its function in silencing is most likely direct.

Interaction of SPT5L with chromatin was also demonstrated at *IGN20*, *IGN22*, *IGN23*, *IGN25* and *IGN26* (Fig. 4.1E-I), which have been identified in a genome-wide screen of Pol V occupancy (A. Wierzbicki, R. Lister, B. Gregory, J. Ecker and C. Pikaard, unpublished data), suggesting that SPT5L binding may be a general feature of Pol V-transcribed loci.

SPT5L works downstream of Pol V

SPT5L interacts with chromatin (Fig. 4.1C-I) as well as with AGO4, Pol V complex and Pol V transcripts [21–23]. SPT5L is also not required for the accumulation of Pol V-dependent transcripts at *IGN5*, *IGN6* or *AtSN1* [22]. This suggests that SPT5L should work downstream of Pol V. To test this prediction we assayed Pol V binding to chromatin by ChIP with antibody against NRPE1, the largest subunit of Pol V. Subsequent real-time PCR demonstrated recovery of DNA from Col-0 wild type at much higher level than from the *nrpe1* mutant at *IGN5*, *solo LTR* and *AtSN1* loci but not at *Actin 2* or *Tubulin 8* loci (Fig. 4.2A-E) demonstrating that Pol V binds chromatin at *IGN5*, *solo LTR* and *AtSN1* loci. DNA recovery from *spt5l* mutant was comparable to Col-0 wild type (Fig. 4.2A-E) showing that SPT5L is not needed for Pol V binding to chromatin. Interestingly, Pol V binding to chromatin was reproducibly increased at *solo LTR* locus in *ago4* mutant (Fig. 4.2C), indicating that AGO4 may inhibit Pol V binding to chromatin possibly by affecting initiation and/or elongation of Pol V transcription. We conclude that SPT5L does not work upstream of Pol V in siRNA-mediated transcriptional gene silencing pathway.

Because both Pol V and SPT5L are required for DNA methylation at several silenced loci [21–23], SPT5L may be functionally dependent on Pol V and/or Pol V transcription. To test this possibility we performed western blot with anti-SPT5L antibody in *nrpe1* mutant background. Accumulation of SPT5L was strongly reduced in the *nrpe1* mutant (Fig. 4.1J). To test if *nrpe1* mutation affects accumulation of *SPT5L* mRNA or SPT5L protein stability, we assayed *SPT5L* RNA using real time RT-PCR. Accumulation of *SPT5L* RNA was not reduced in the *nrpe1* mutant (Fig. 4.1K) indicating that Pol V is needed for SPT5L protein stability. This behavior of SPT5L in *nrpe1* mutant is reminiscent of reduced AGO4 protein stability in mutants that reduce siRNA production [28]. Interestingly, we observed a slight increase in *SPT5L* RNA level in the *nrpe1* mutant

which may be explained by the presence of an AtMU10 transposon in *SPT5L* coding region. Overall, these results suggest that SPT5L is functionally dependent on Pol V.

We further tested the functional relationship between Pol V and SPT5L by performing ChIP with anti-SPT5L antibody in *nrpe1* mutant background. Consistent with the reduced stability of SPT5L in *nrpe1* (Fig. 4.1J), DNA recovery from Pol V-transcribed loci was reduced to the level observed in the *spt5l* mutant (Fig. 4.1C-I). This result may be explained by the overall reduction in the amount of SPT5L. However, a similar reduction in the SPT5L protein accumulation in *rdr2* mutant (Fig. 4.1J) did not affect SPT5L binding to chromatin (see below). This suggests that *nrpe1* may affect the ChIP signal not only by destabilizing SPT5L, but also by affecting its ability to bind chromatin. Because SPT5L does not work upstream of Pol V and is functionally dependent on Pol V, we conclude that SPT5L works downstream of Pol V and/or Pol V transcription and may be recruited to chromatin by Pol V.

SPT5L binds chromatin independently of AGO4

The recruitment of SPT5L to chromatin by Pol V (Fig. 4.1, Fig. 4.2) is consistent with the interaction of SPT5L with Pol V transcripts and AGO4 [21,22]. There are at least two explanations of SPT5L function in the establishment of siRNA-mediated transcriptional gene silencing. SPT5L may be recruited by Pol V and then help recruit AGO4-siRNA complexes. Alternatively, AGO4-siRNA may recognize target loci and then recruit SPT5L which further recruits chromatin modifying enzymes. To test the latter possibility we performed ChIP with αSPT5L antibody in the *ago4* mutant. DNA recovery of all tested Pol V-transcribed loci was comparable from Col-0 wild type and the *ago4* mutant (Fig. 4.1C-I). This shows that binding of SPT5L to chromatin was not affected in the *ago4* mutant, and suggests that SPT5L is not recruited to its target loci by AGO4-siRNA complexes. We conclude that SPT5L does not work downstream of AGO4 in the siRNA-mediated transcriptional gene silencing pathway.

AGO4 binds chromatin partially independently of SPT5L

Having concluded that SPT5L does not work downstream of AGO4, we tested the alternative hypothesis that SPT5L may work upstream of AGO4 by binding Pol V and/or Pol V transcripts and recruiting AGO4 to chromatin. To test this possibility we performed ChIP with anti-AGO4 antibody. As demonstrated by real-time PCR we

recovered DNA from wild type plants above the background level observed in the *ago4* mutant at *IGN5* and *solo LTR* (Fig. 4.3CD) as well as at *IGN20*, *IGN22*, *IGN23*, *IGN25* and *IGN26* loci (Fig. 4.3E-I). This indicates that AGO4 binds chromatin at all tested Pol V-transcribed loci. In the *spt5l* mutant total accumulation of AGO4 protein was not affected (Fig. 4.3J). At all assayed Pol V-transcribed loci AGO4 binding to chromatin in the *spt5l* mutant was reproducibly above the background level observed in the *ago4* mutant indicating that AGO4 is able to bind chromatin in the absence of SPT5L (Fig. 4.3C-I). Interestingly, we observed that the intensity of AGO4 binding to chromatin is slightly reduced in the *spt5l* mutant at *solo LTR*, *IGN20*, *IGN22*, *IGN23*, *IGN25* and *IGN26* (Fig. 4.3C-I). This indicates that although SPT5L is not required for AGO4 recruitment to chromatin, it enhances AGO4 chromatin binding. Alternatively, most loci may be occupied by two pools of AGO4. One being SPT5L-dependent and other recruited to chromatin independently of SPT5L.

We conclude that SPT5L is not required for recruitment of a pool of AGO4 to specific loci in chromatin and therefore does not work upstream of AGO4 in the siRNA-mediated transcriptional gene silencing pathway. Since SPT5L also does not work downstream of AGO4, they are most likely recruited in parallel and at least partially independently of each other.

SPT5L binds chromatin independently of 24-nt siRNA

The parallel and independent recruitment of SPT5L and AGO4 to chromatin suggests that they are both guided by the interactions with Pol V complex and/or Pol V transcripts. To test if SPT5L is also guided by siRNA we used ChIP to assay SPT5L binding to chromatin in *rdr2*, a mutant in an RNA-dependent RNA polymerase responsible for production of the majority of 24-nt siRNA [29]. The *rdr2* mutation reduced the stability of SPT5L protein (Fig. 4.1JK) but did not cause reduction in DNA recovery of the tested loci after ChIP (Fig. 4.4C-I). This suggests that although RDR2 increases the amount of SPT5L protein, the chromatin-bound fraction of SPT5L is not affected by the *rdr2* mutation. This also suggests the presence of siRNA-dependent pool of SPT5L that does not physically interact with assayed Pol V-transcribed loci.

These results demonstrate that binding of SPT5L to chromatin is not affected in the *rdr2* mutant and suggest that RDR2-dependent siRNA is not required for SPT5L

binding to chromatin. In contrast, RDR2 is necessary for proper establishment of DNA methylation at *AtSN1*, *IGN5*, *IGN25*, *IGN23*, *IGN26*, *solo LTR* and *IGN22* (Fig. 4.4J); demonstrating that all assayed loci are in fact targets of the siRNA-mediated transcriptional gene silencing pathway. We conclude that SPT5L is recruited to chromatin in a manner independent of 24-nt siRNA.

Both AGO4 and SPT5L are needed for repressive chromatin modifications

Parallel and at least partially independent recruitment of SPT5L and AGO4 by Pol V suggests that at Pol V-transcribed loci none of them is sufficient for the establishment and maintenance of silent chromatin modifications. To further test this possibility we assayed several Pol V-transcribed loci for DNA methylation side-by-side in *nrpe1*, *ago4* and *spt5l* mutants using DNA methylation-sensitive restriction endonucleases. Methylation of cytosines in *Hae*III, *Alu*I or *Ava*II restriction sites blocks the enzymes from cutting and allows amplification of the genomic region by PCR. However, unmethylated sites are cleaved and PCR amplification fails. All three enzymes recognize asymmetric (CHH) methylation at tested loci. Consistently with previous reports, DNA methylation was strongly reduced at *AtSN1* locus in both *ago4* [19,24,30] and *spt5l* mutants [21–23] and at *IGN5* locus in *ago4* mutant [19] (Fig. 4.5A). DNA methylation was also reduced at *IGN5* locus in *spt5l* mutant and at *IGN23*, *IGN25* and *IGN26* loci in both *ago4* and *spt5l* mutants (Fig. 4.5AB). Importantly, in all these cases reduction of DNA methylation was comparable in *ago4* and *spt5l* mutants (Fig. 4.5AB) suggesting that neither AGO4 nor SPT5L is sufficient for the establishment of asymmetric DNA methylation at Pol V-transcribed loci.

We also tested the effect of *nrpe1*, *ago4* and *spt5l* mutations on dimethylation of lysine 9 of histone H3 (H3K9me2). At *IGN5* and *IGN26* loci, H3K9me2 was reduced in all three mutants (Fig. 4.5DF) showing that both AGO4 and SPT5L are required not only for the establishment and/or maintenance of DNA methylation but also H3K9me2. We conclude that at least at a subset of loci SPT5L and AGO4 work together to recruit repressive chromatin modifications. We propose that it is the coordinate action of SPT5L and AGO4 that directly or indirectly recruits *de novo* DNA methyltransferase DRM2 and H3K9 methyltransferases.

SPT5L contributes to repressive chromatin modifications in a locus-specific manner

While *AtSN1*, *IGN5*, *IGN23*, *IGN25* and *IGN26* loci require both AGO4 and SPT5L for repressive chromatin modifications (Fig. 4.5), *soloLTR* has been shown to be methylated independently of SPT5L [21,23]. We confirm this result and further show that *solo LTR* and *IGN22* which, like other Pol V-transcribed loci, are methylated in a Pol V and AGO4-dependent manner (Fig. 4.6A) did not show reduction of DNA methylation on *Alu*I or *Ava*II sites in the *spt5l* mutant (Fig. 4.6A). This suggests that there is some significant locus specificity in SPT5L contributions to DNA methylation. Furthermore, H3K9me2 was reduced at both *soloLTR* and *IGN22* in *nrpe1* and *ago4* mutants but not in the *spt5l* mutant (Fig. 4.6BC). Also acetylation of histone H3 (H3Ac) at *solo LTR* was increased in *nrpe1* and *ago4* but not in *spt5l* (Fig. 4.6D). This demonstrates that the locus-specific function of SPT5L affects not only DNA methylation but also H3K9me2 and H3Ac.

The requirement of SPT5L for repressive chromatin modifications (Fig. 4.5,2.6) does not correlate with the extent of partial SPT5L-dependency of AGO4 binding to chromatin (Fig. 4.3). It suggests that the pool of AGO4 that is bound to chromatin in an SPT5L-dependent manner is not required for silencing. This is consistent with our interpretation that AGO4 and SPT5L are recruited to chromatin in parallel and independently of each other.

Discussion

Order of events in siRNA-mediated silencing

Our findings establish the order of events leading to siRNA-mediated establishment of transcriptional silencing. This process is initiated by recognition of silencing targets and production of two classes of non-coding RNA. The first class is siRNA which is produced from double-stranded RDR2 products by DCL3 and becomes incorporated into AGO4 and possibly also AGO6 and AGO9 [4,5,24]. The second class is long non-coding RNA produced by Pol V and/or Pol II [18,26]. Pol V transcription is initiated independently of siRNA and Pol V transcripts most likely are not precursors for siRNA biogenesis [18,31]. Pol V recruitment to chromatin and transcription requires the presence

of DMS3, DRD1 and RDM1, which either help initiate Pol V transcription or assist elongation of Pol V transcripts [18,19,32].

Pol V transcription is followed by association of two RNA-binding proteins with chromatin (Fig. 4.7). First is AGO4 which is recruited to chromatin by Pol V transcripts and uses the incorporated siRNA to provide sequence-specificity of silencing [19]. The second is SPT5L (Fig. 4.1,2.2), which is recruited to chromatin by an unknown mechanism, possibly involving interactions between SPT5L and Pol V complex and/or with Pol V transcripts [22,23]. SPT5L binds chromatin independently of 24-nt siRNA (Fig. 4.4) and is likely a general factor associated with transcribing Pol V and its transcripts [21–23]. Since SPT5L binds chromatin in the absence of AGO4 (Fig. 4.1), and the functional pool of AGO4 is able to bind chromatin in the absence of SPT5L (Fig. 4.3), we concluded that they are recruited to chromatin in parallel and independently of each other. Both AGO4 and SPT5L are required for the establishment and/or maintenance of DNA methylation and repressive histone modifications at the majority of tested loci (Fig. 4.5). This suggests that both are needed for the recruitment of enzymes establishing repressive chromatin modifications.

Mechanism recruiting chromatin modifying enzymes

Because AGO4 and SPT5L bind chromatin independently of each other, and, at the majority of tested loci both are required for establishment and maintenance of silencing, we propose that AGO4 and SPT5L create a binding platform for the recruitment of chromatin modifying proteins. One possibility is that both weakly interact with a downstream protein but the interaction becomes strong enough to recruit chromatin modifying enzymes only when both are present. Alternatively, AGO4 may be a sole interacting partner of downstream proteins but SPT5L, which has a C-terminal domain rich in WG/GW motifs, interacts with AGO4 and alters its conformation to facilitate the recruitment of chromatin modifying enzymes.

Our results show that there are loci where DNA methylation is established in a Pol V, AGO4 and SPT5L-dependent manner (Fig. 4.5A), but these loci have an overall low level of H3K9me2 and no change in the histone modifications in tested mutants (*IGN23* in Fig. 4.5E). It suggests that the *de novo* DNA methyltransferase DRM2 is likely the chromatin modifying enzyme directly recruited by the AGO4-SPT5L platform. It is

also possible that DRM2 may be recruited indirectly by another protein that binds the AGO4-SPT5L platform.

Assembly of the silencing complexes

Binding of AGO4 and SPT5L to chromatin is mediated by multiple protein-protein and protein-RNA interactions. These interactions may mediate recruitment of proteins to specific genomic regions and/or stabilize binding after recruitment by an independent mechanism.

SPT5L binding to chromatin occurs downstream of Pol V and is most likely mediated by protein-RNA interaction between SPT5L and Pol V transcripts [22]. Like canonical SPT5, SPT5L may also form a heterodimer with SPT4 [33]. Alternatively, SPT5L may be recruited to chromatin by protein-protein interaction with Pol V complex as suggested by identification of SPT5L in Pol V holoenzyme [23] and interactions between yeast SPT5 as well as bacterial homolog of SPT5, nusG, with RNA polymerases [34,35]. It is also possible that SPT5L is recruited to chromatin by interacting with both Pol V transcripts and Pol V complex. All these mechanisms explain the AGO4-independent binding of SPT5L to Pol V-transcribed loci.

Interaction with Pol V transcripts seems to be the major factor recruiting AGO4-siRNA to chromatin [19]. AGO4 also interacts with WG/GW-rich C-terminal domains of Pol V and SPT5L [21,22,36]. Because Argonautes contain only one WG/GW binding pocket [37] these interactions may be employed sequentially. First, they help recruit AGO4 to chromatin by interaction with Pol V and then they stabilize the binding of AGO4 to chromatin on its target loci by interaction with SPT5L. It is consistent with our observation that AGO4 binding to chromatin is slightly reduced in the *spt5l* mutant (Fig. 4.3).

Locus-specific regulation of silencing

We show that SPT5L contributes to regulation of siRNA-mediated transcriptional silencing in a highly locus-specific manner. This is demonstrated by the observation that two of the tested loci require Pol V and AGO4 but not SPT5L for establishment and/or maintenance of repressive chromatin modifications (Fig. 4.6). It could be explained by presence of the canonical SPT5 at a subset of silenced loci. However, both loci are occupied by SPT5L in wild type plants (Fig. 4.1) suggesting that SPT5L is in fact

involved in their silencing. Only when SPT5L is mutated, the canonical SPT5 is able to compensate the deficiency at these particular loci. Alternatively, it is possible that the observed locus-specificity of SPT5L is caused by the presence of both Pol V and Pol II at a subset of loci [26]. Pol II-bound canonical SPT5 may be able to compensate the lack of Pol V-bound SPT5L. The mechanism deciding locus specificity of the SPT5L function remains unknown.

Our results also suggest the presence of two pools of AGO4: SPT5L-dependent and SPT5L-independent. Because both pools are detectable at loci that are silenced in a SPT5L-independent manner, the SPT5L-dependent pool of AGO4 is likely not required for silencing. It may be recruited independently of siRNA by direct interaction with SPT5L and may have some other, yet unknown and locus-specific functions.

Materials and Methods

Plant lines and antibodies

Arabiodopsis thaliana nrpe1 (*nrpd1b-11*), *dms3-4,* and *ago4-1* introgressed into Col-0 background were described previously [19,38]. *rdr2-1* mutant was obtained from J. Carrington. *spt5l-1* (*rdm3-3*; SALK_001254) mutant line, affinity-purified anti-SPT5L (anti-KTF1), affinity-purified anti-Pol V (anti-NRPE1) and affinity-purified anti-AGO4 antibodies were described previously [19,22,39]. Mouse monoclonal anti-H3K9me2 antibody (cat. #ab1220) was obtained from Abcam, rabbit polyclonal anti-H3Ac antibody (cat. #06-599) was obtained from Millipore.

Chromatin Immunoprecipitation

ChIP was performed essentially as described [18,19]. Detailed ChIP protocol is included in Appendix A. ChIP samples were amplified in triplicate in Applied Biosystems 7500 real time PCR machine and obtained data were analyzed using comparative C_T method relative to inputs [40]. All ChIP experiments were performed in three independent biological replicates. Results from every biological replicate were normalized to Col-0 wild type and normalized data were used to obtain averages and standard deviations that show fold difference between analyzed strains. Normalized data were subsequently multiplied by average ChIP signal level of Col-0 wild type. This way data are corrected for variability in overall signal strength between independent experiments, the unit is

%input and presented data reflect the relative signal strength observed at particular loci. Standard deviations for Col-0 wild type are not available because Col-0 wild type was used to normalize data.

DNA and RNA analysis

For DNA methylation analysis genomic DNA was extracted from above-ground tissue of 2-week old plants using DNeasy Plant Mini Kit (Qiagen). 100 ng of genomic DNA was digested with 10u of *Hae*III, *Alu*I or *Ava*II restriction enzymes (NEB) for 20 min. After heat-inactivation of the enzyme DNA was amplified using 0.75u Platinum Taq (Invitrogen).

Total RNA was extracted from 2-week old plants using RNeasy Plant Mini Kit (Qiagen) and amplified using SuperScript III Platinum SYBR Green One-Step qRT-PCR Kit (Invitrogen) in Applied Biosystems 7500 real time PCR machine.

Oligonucleotide primers used in this study are shown in a table in Appendix D.

Acknowledgements

We thank Eran Pichersky and David Engelke for critical reading of the manuscript.

References

1. Girard, A. & Hannon, G. J. Conserved themes in small-RNA-mediated transposon control. *Trends Cell Biol.* **18,** 136–148 (2008).
2. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **41,** 563–571 (2009).
3. Moazed, D. Small RNAs in transcriptional gene silencing and genome defence. *Nature* **457,** 413–420 (2009).
4. Matzke, M., Kanno, T., Daxinger, L., Huettel, B. & Matzke, A. J. M. RNA-mediated chromatin-based silencing in plants. *Curr. Opin. Cell Biol.* **21,** 367–376 (2009).
5. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11,** 204–220 (2010).
6. Alleman, M. *et al.* An RNA-dependent RNA polymerase is required for paramutation in maize. *Nature* **442,** 295–298 (2006).
7. Erhard, K. F., Jr *et al.* RNA polymerase IV functions in paramutation in Zea mays. *Science* **323,** 1201–1205 (2009).
8. Sidorenko, L. *et al.* A dominant mutation in mediator of paramutation2, one of three second-largest subunits of a plant-specific RNA polymerase, disrupts multiple siRNA silencing processes. *PLoS Genet.* **5,** e1000725 (2009).

9. Cam, H. P., Chen, E. S. & Grewal, S. I. S. Transcriptional scaffolds for heterochromatin assembly. *Cell* **136,** 610–614 (2009).
10. Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297,** 1833–1837 (2002).
11. Cam, H. P. *et al.* Comprehensive analysis of heterochromatin- and RNAi-mediated epigenetic control of the fission yeast genome. *Nat. Genet.* **37,** 809–819 (2005).
12. Djupedal, I. *et al.* RNA Pol II subunit Rpb7 promotes centromeric transcription and RNAi-directed chromatin silencing. *Genes Dev.* **19,** 2301–2306 (2005).
13. Irvine, D. V. *et al.* Argonaute slicing is required for heterochromatic silencing and spreading. *Science* **313,** 1134–1137 (2006).
14. Bühler, M., Verdel, A. & Moazed, D. Tethering RITS to a nascent transcript initiates RNAi- and heterochromatin-dependent gene silencing. *Cell* **125,** 873–886 (2006).
15. Pontier, D. *et al.* Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. *Genes Dev.* **19,** 2030–2040 (2005).
16. Kanno, T. *et al.* Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nat. Genet.* **37,** 761–765 (2005).
17. Ream, T. S. *et al.* Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. *Mol. Cell* **33,** 192–203 (2009).
18. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135,** 635–648 (2008).
19. Wierzbicki, A. T., Ream, T. S., Haag, J. R. & Pikaard, C. S. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat. Genet.* **41,** 630–634 (2009).
20. Qi, Y. *et al.* Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature* **443,** 1008–1012 (2006).
21. Bies-Etheve, N. *et al.* RNA-directed DNA methylation requires an AGO4-interacting member of the SPT5 elongation factor family. *EMBO Rep.* **10,** 649–654 (2009).
22. He, X.-J. *et al.* An effector of RNA-directed DNA methylation in arabidopsis is an ARGONAUTE 4- and RNA-binding protein. *Cell* **137,** 498–508 (2009).
23. Huang, L. *et al.* An atypical RNA polymerase involved in RNA silencing shares small subunits with RNA polymerase II. *Nat. Struct. Mol. Biol.* **16,** 91–93 (2009).
24. Havecker, E. R. *et al.* The Arabidopsis RNA-directed DNA methylation argonautes functionally diverge based on their expression and interaction with target loci. *Plant Cell* **22,** 321–334 (2010).
25. Huettel, B. *et al.* Endogenous targets of RNA-directed DNA methylation and Pol IV in Arabidopsis. *EMBO J.* **25,** 2828–2836 (2006).
26. Zheng, B. *et al.* Intergenic transcription by RNA polymerase II coordinates Pol IV and Pol V in siRNA-directed transcriptional gene silencing in Arabidopsis. *Genes Dev.* **23,** 2850–2860 (2009).
27. Numa, H. *et al.* Transduction of RNA-directed DNA methylation signals to repressive histone marks in Arabidopsis thaliana. *EMBO J.* **29,** 352–362 (2010).
28. Li, C. F. *et al.* An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in Arabidopsis thaliana. *Cell* **126,** 93–106 (2006).

29. Kasschau, K. D. *et al.* Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biol.* **5,** e57 (2007).
30. Zilberman, D., Cao, X. & Jacobsen, S. E. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299,** 716–719 (2003).
31. Mosher, R. A., Schwach, F., Studholme, D. & Baulcombe, D. C. PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 3145–3150 (2008).
32. Law, J. A. *et al.* A protein complex required for polymerase V transcripts and RNA-directed DNA methylation in Arabidopsis. *Curr. Biol.* **20,** 951–956 (2010).
33. Schwer, B., Schneider, S., Pei, Y., Aronova, A. & Shuman, S. Characterization of the Schizosaccharomyces pombe Spt5-Spt4 complex. *RNA* **15,** 1241–1250 (2009).
34. Hartzog, G. A., Wada, T., Handa, H. & Winston, F. Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in Saccharomyces cerevisiae. *Genes Dev.* **12,** 357–369 (1998).
35. Li, J., Horwitz, R., McCracken, S. & Greenblatt, J. NusG, a new Escherichia coli elongation factor involved in transcriptional antitermination by the N protein of phage lambda. *J. Biol. Chem.* **267,** 6012–6019 (1992).
36. El-Shami, M. *et al.* Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components. *Genes & Development* **21,** 2539–2544 (2007).
37. Eulalio, A., Huntzinger, E. & Izaurralde, E. GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. *Nat. Struct. Mol. Biol.* **15,** 346–353 (2008).
38. Pontes, O. *et al.* The Arabidopsis chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell* **126,** 79–92 (2006).
39. Onodera, Y. *et al.* Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120,** 613–622 (2005).
40. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25,** 402–408 (2001).

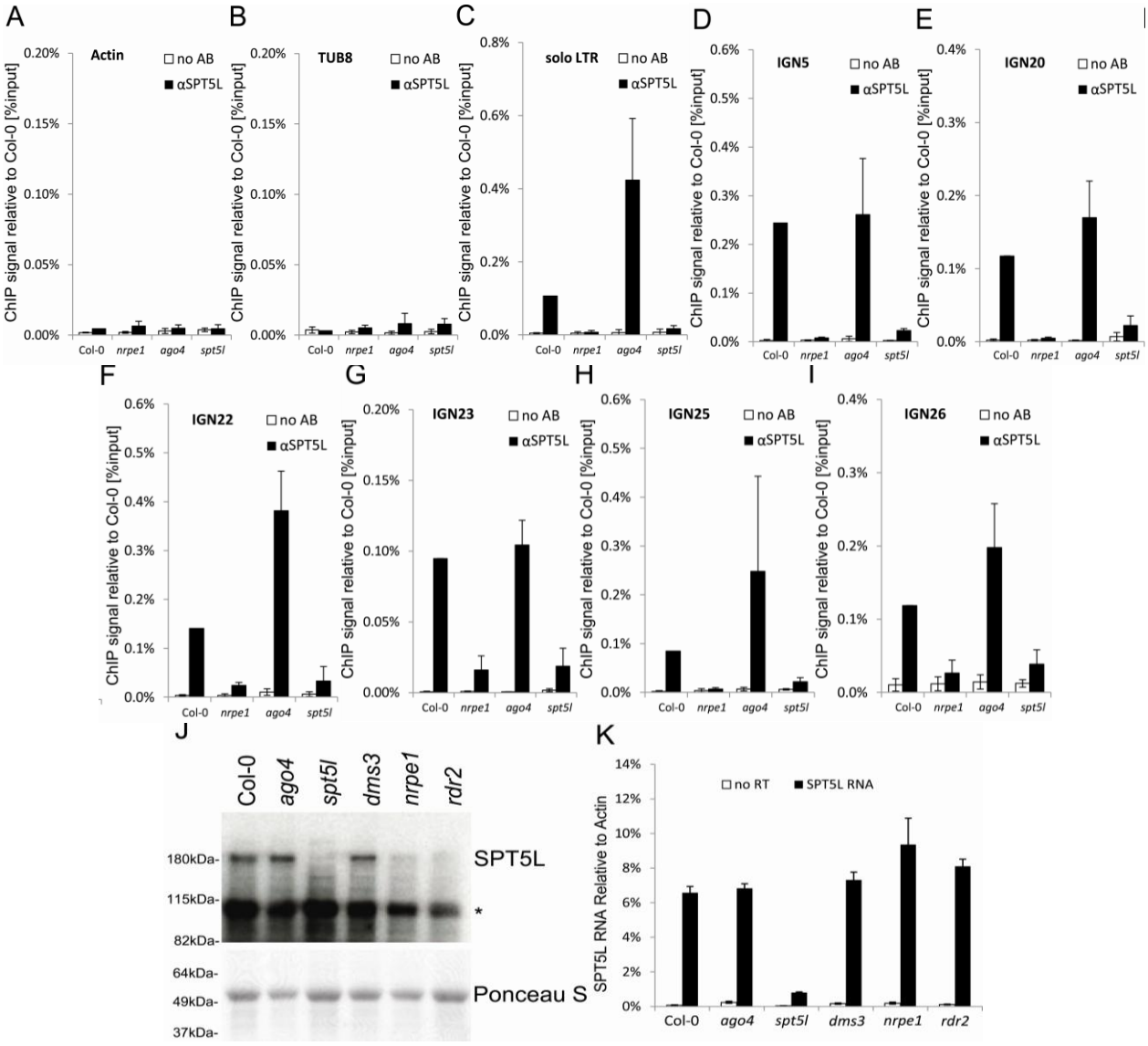


Figure 4.1 SPT5L interacts with chromatin in a Pol V-dependent and AGO4 independent manner

A-I) ChIP data showing SPT5L binding to chromatin in Col-0 wild type, *nrpe1, ago4* and *spt5l* mutants at loci transcribed by Pol V and silenced by siRNA-mediated transcriptional silencing: *solo LTR*(C), *IGN5* (D), *IGN20* (E), *IGN22* (F), *IGN23* (G), *IGN25* (H) and *IGN26* (I). Two loci transcribed by Pol II are shown as controls: *Actin 2* (A) and *Tubulin 8* (B). No antibody controls (white bars) provide background level for ChIP samples (black bars). Bars represent mean value of ChIP signals normalized to Col-0 wild type. Error bars are standard deviations of three independent biological replicates.

J) Immunoblot detection of SPT5L in whole-cell protein extracts from Col-0 wild type, *ago4, spt5l, dms3, nrpe1* and *rdr2* mutants. Ponceau S staining of the membrane is a loading control. Asterisk denotes nonspecific bands.

K) Real time RT-PCR detection of *SPT5L* RNA in Col-0 wild type, *ago4, spt5l, dms3, nrpe1* and *rdr2* mutants. Bars represent average *SPT5L* mRNA accumulation relative to *Actin 2* from three biological replicates. Error bars represent standard deviation.

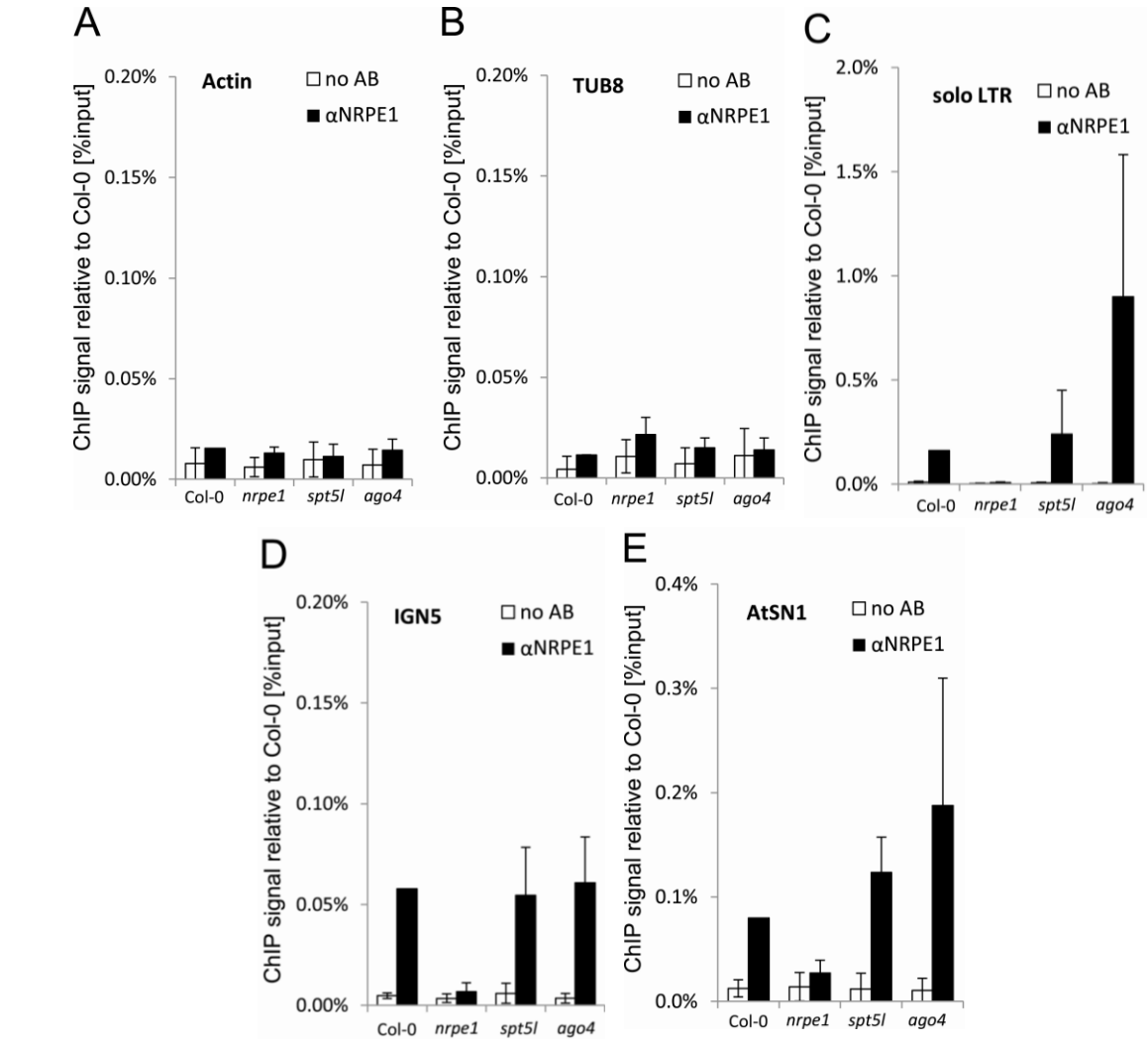


Figure 4.2 SPT5L and AGO4 are not required for Pol V binding to chromatin
A-E) Pol V occupancy of *Actin 2* (A) and *Tubulin 8* (B) control loci, *solo LTR* (C), *IGN5* (D) and *AtSN1* (E) assayed by ChIP in Col-0 wild type, *nrpe1*, *spt5l* and *ago4*. No antibody controls (white bars) provide background level for ChIP samples (black bars). Bars represent mean value of ChIP signals normalized to Col-0 wild type. Error bars are standard deviations of three independent biological replicates.

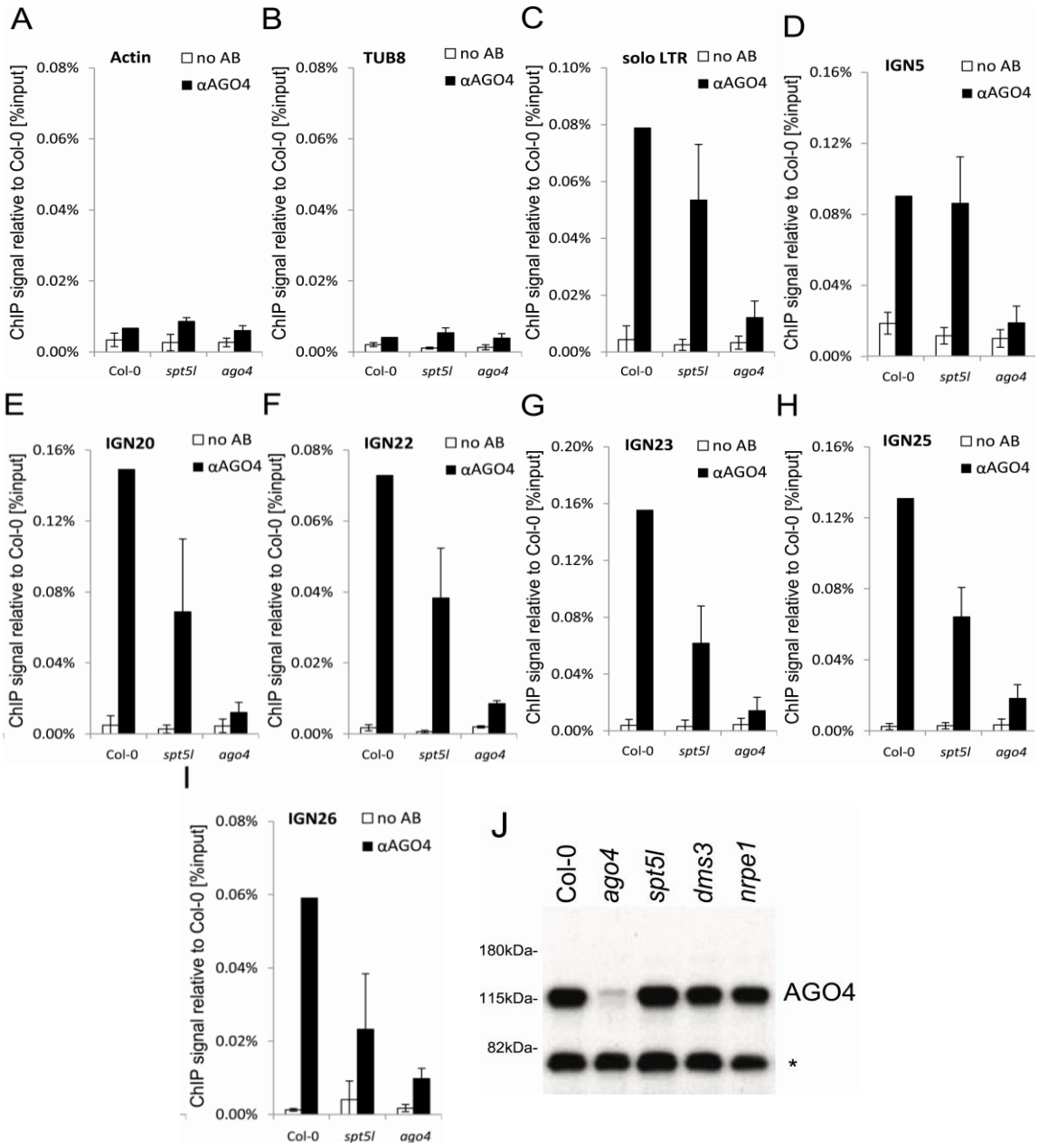


Figure 4.3 AGO4 can bind chromatin independently of SPT5L

A-I) ChIP data showing AGO4 binding to chromatin in Col-0 wild type, *spt5l* and *ago4* mutants at *Actin 2* (A) and *Tubulin 8* (B) control loci, *solo LTR* (C), *IGN5* (D), *IGN20* (E), *IGN22* (F), *IGN23* (G), *IGN25* (H) and *IGN26* (I). No antibody controls (white bars) provide background level for ChIP samples (black bars). Bars represent mean value of ChIP signals normalized to Col-0 wild type. Error bars are standard deviations of three independent biological replicates.

J) Immunoblot detection of AGO4 in whole-cell protein extracts from Col-0 wild type, *ago4, spt5l, dms3* and *nrpe1* mutants using anti-AGO4 antibody. Asterisk denotes a nonspecific band. Ponceau S staining of the membrane shown in Fig. 4.1J is a loading control.

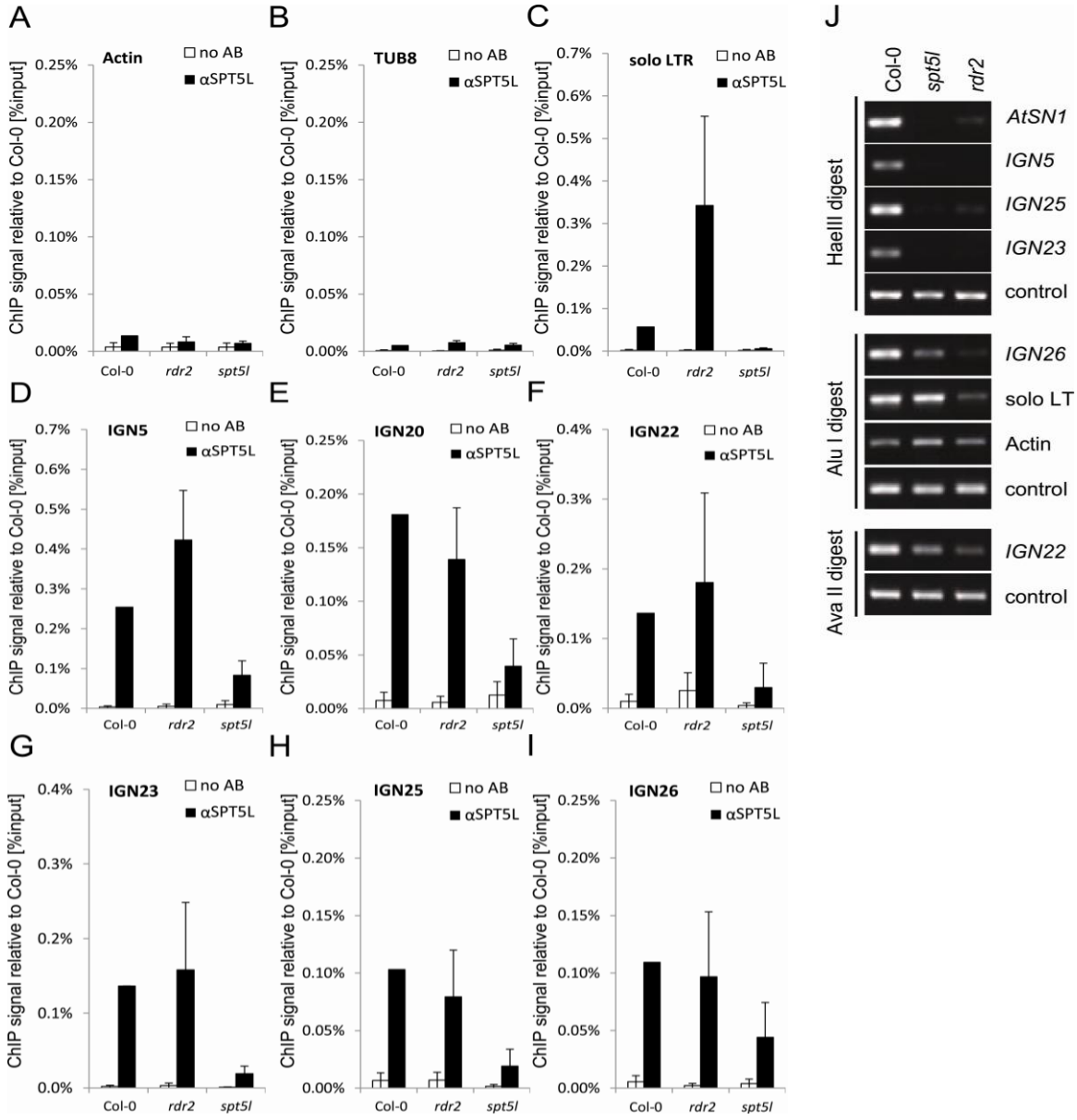


Figure 4.4 SPT5L interacts with chromatin in an siRNA-independent manner
A-I) ChIP data showing SPT5L binding to chromatin in Col-0 wild type, *rdr2* and *spt5l* mutants at loci transcribed by Pol V and silenced by siRNA-mediated transcriptional silencing: *solo LTR*(C), *IGN5* (D), *IGN20* (E), *IGN22* (F), *IGN23* (G), *IGN25* (H) and *IGN26* (I). Two loci transcribed by Pol II are shown as controls: *Actin 2* (A) and *Tubulin 8* (B). No antibody controls (white bars) provide background level for ChIP samples (black bars). Bars represent mean value of ChIP signals normalized to Col-0 wild type. Error bars are standard deviations of three independent biological replicates.
J) DNA methylation analysis of *AtSN1*, *IGN5*, *IGN23* and *IGN25* performed by digestion with *Hae*III restriction endonuclease, *IGN26* and *solo LTR* performed by digestion with *Alu*I restriction endonuclease and *IGN22* performed by digestion with *Ava*II restriction endonuclease. Digested genomic DNA was amplified by PCR. Sequences lacking *Hae*III (*Actin 2*), *Alu*I (*IGN5*) or *Ava*II (*Actin 2*) were used as loading controls.

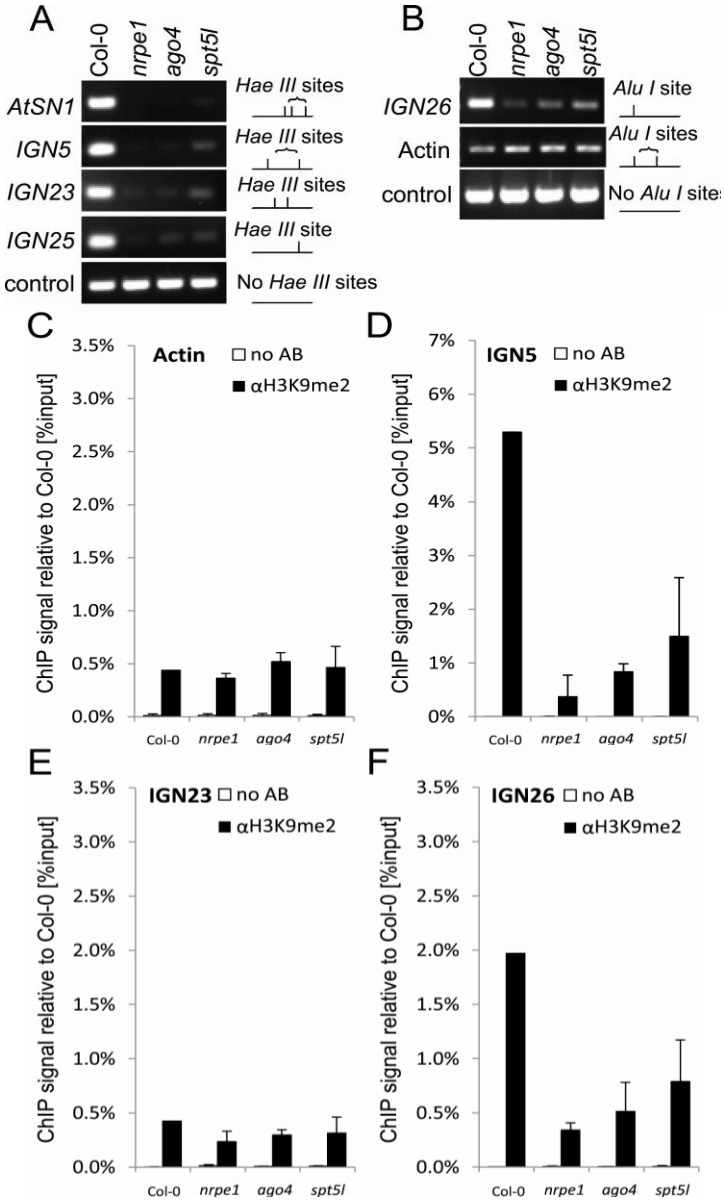


Figure 4.5 Both SPT5L and AGO4 are required for silencing at certain loci
A) DNA methylation analysis of *AtSN1*, *IGN5*, *IGN23* and *IGN25* performed by digestion with *Hae*III restriction endonuclease. Digested genomic DNA was amplified by PCR. Sequence lacking *Hae*III (*Actin 2*) sites was used as a loading control.
B) DNA methylation analysis of *IGN26* performed by digestion with *Alu*I restriction endonuclease. Digested genomic DNA was amplified by PCR. Sequence lacking *Alu*I (*IGN5*) sites was used as a loading control.
C-F) Analysis of H3K9me2 at *actin 2* (C), *IGN5* (D), *IGN23* (E) and *IGN26* (F) loci performed by ChIP with anti-H3K9me2 antibody in Col-0 wild type, *nrpe1*, *ago4* and *spt5l*. No antibody controls (white bars) provide background level for ChIP samples (black bars). Bars represent mean value of ChIP signals normalized to Col-0 wild type. Error bars are standard deviations of three independent biological replicates.

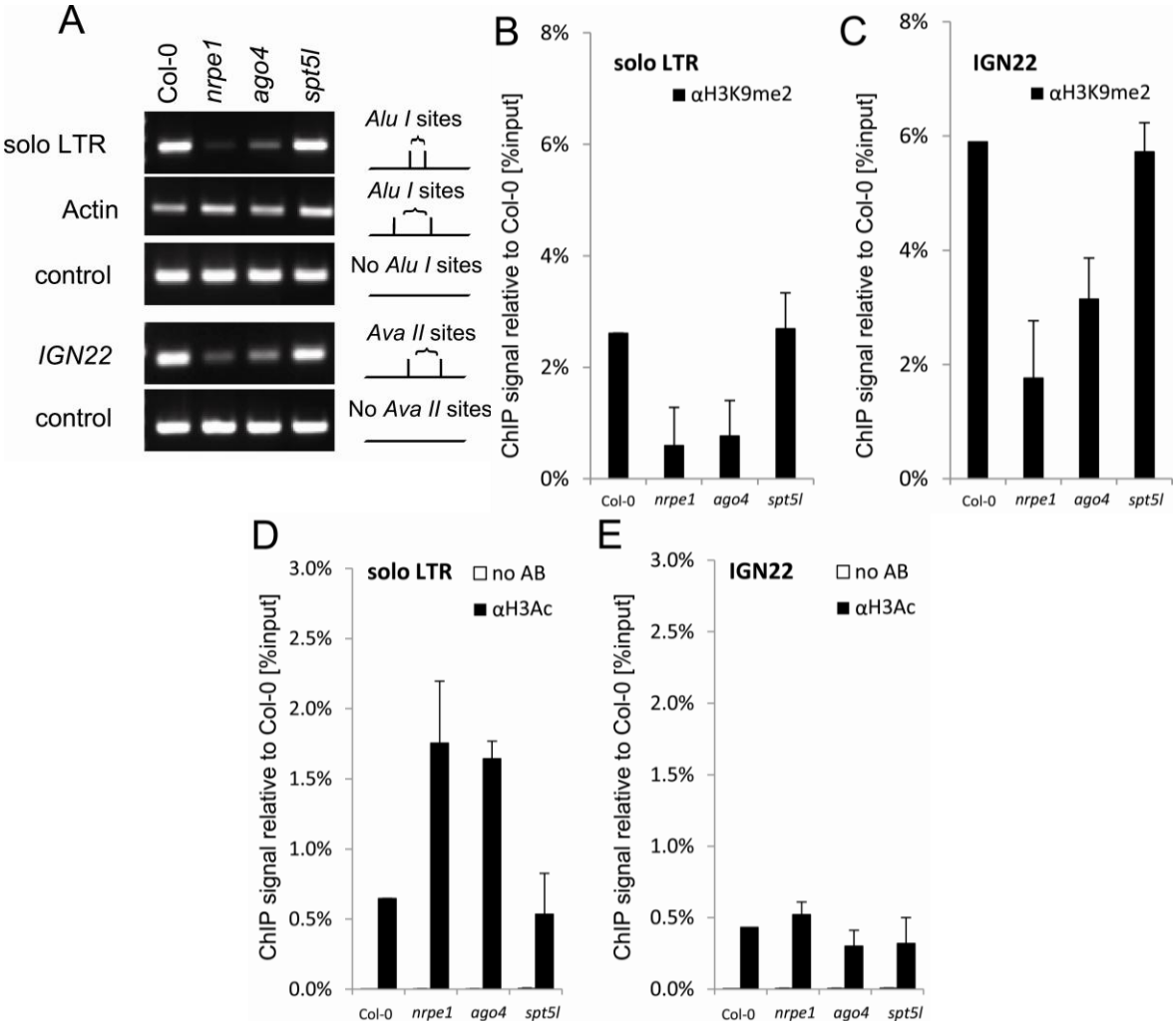


Figure 4.6 Locus-specific effects of SPT5L on silencing

A) DNA methylation analysis of *solo LTR* performed by digestion with *Alu*I restriction endonuclease and *IGN22* performed by digestion with *Ava*II restriction endonuclease. Digested genomic DNA was amplified by PCR. Sequences lacking *Alu*I (*IGN5*) or *Ava*II (*Actin 2*) sites were used as loading controls.

B-C) ChIP analysis of H3K9me2 at *solo LTR* (B) and *IGN22* (C) in Col-0 wild type, *nrpe1*, *ago4* and *spt5l*. Corresponding no antibody controls are shown in panels D and E. Bars represent mean value of ChIP signals normalized to Col-0 wild type. Error bars are standard deviations of three independent biological replicates.

D-E) ChIP analysis of H3Ac at *solo LTR* (D) and *IGN 22* (E) in Col-0 wild type, *nrpe1*, *ago4* and *spt5l*. No antibody controls (white bars) provide background level for ChIP samples (black bars). Bars represent mean value of ChIP signals normalized to Col-0 wild type. Error bars are standard deviations of three independent biological replicates.

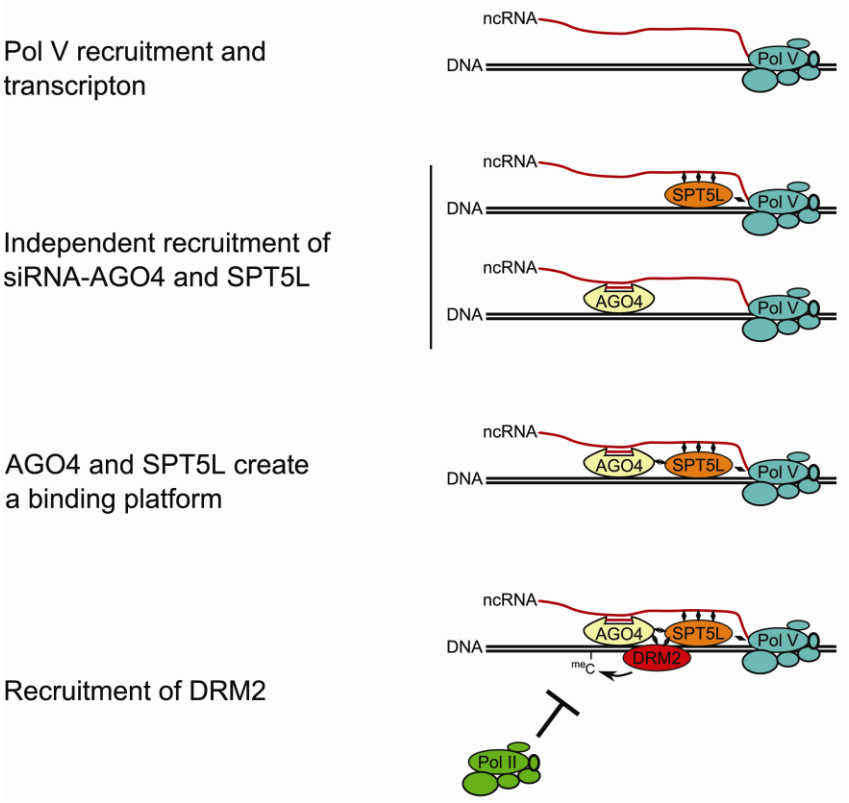


Figure 4.7 Model of SPT5L involvement in the recruitment of chromatin modifying enzymes

Pol V produces intergenic non-coding transcripts which are the binding points for SPT5L and AGO4-siRNA complex. Both AGO4 and SPT5L may interact with both Pol V transcripts and Pol V complex itself. SPT5L and AGO4 are recruited to Pol V transcripts in parallel and independently of each other. When both AGO4 and SPT5L are present they create a binding platform for direct or indirect recruitment of DRM2 *de novo* DNA methyltransferase and other chromatin modifying enzymes. Establishment of chromatin modifications represses Pol II transcription.

CHAPTER 5

Distinct Roles of SPT5L and AGO4 in Transcriptional Gene Silencing

The contents of this chapter will be submitted for publication in the near future. Jan Kuciński prepared AGO4 ChIP samples for sequencing. Lilia Bouzit generated profiles shown in Figure 5.1A,C,D. Natalie Blackwood performed experiments shown in Figure 5.9C. I performed all other experiments and data analysis shown in this chapter.

Abstract

Transcriptional Gene Silencing (TGS) protects genomes from harmful transposons and controls gene expression. This is accomplished through chromatin modification directed by small RNA and long non-coding RNA (lncRNA) bound proteins. In *Arabidopsis thaliana*, lncRNA is produced by RNA Polymerase V (Pol V) which helps guide ARGONAUTE4 (AGO4) to chromatin and directs *de novo* DNA methylation. Other proteins such as SPT5L (Supressor of Ty insertion 5 – Like) and IDN2 (INVOLVED IN DE NOVO 2) also interact with lncRNA and are important for TGS. However, the role of SPT5L in directing DNA methylation varies at individual loci. We show through genome wide maps of SPT5L and AGO4 binding sites (ChIP-seq) that these proteins bind chromatin mostly independent of each other, but that feedback through Pol V transcription can occur. We also show that SPT5L has a more limited role in directing *de novo* DNA methylation than does AGO4, and that this limited role is similar to that of IDN2. Furthermore, by examining genome-wide maps of H3 occupancy, we show that both SPT5L and AGO4 help direct nucleosome positioning. We propose that coordinated binding of SPT5L and AGO4 creates an additional level of silencing by increasing the nucleosomal density at TGS targets.

Author Summary

Transposable elements are regions of DNA which are silenced to decrease the potential risks to genomic integrity. This silencing is often in the form of modification such as DNA methylation and chromatin compaction. One pathway responsible for establishing these modifications is Transcriptional Gene Silencing, otherwise known as RNA-directed DNA Methylation. As the name implies, chromatin modification is directed by long non-coding RNAs which act as scaffolds for RNA binding proteins to interact with chromatin. AGO4 and SPT5L are two such proteins which bind RNA and are thought to direct chromatin modifiers. While the role of AGO4 in directing *de novo* DNA methylation is established, the role of SPT5L is unknown. Here we show that while AGO4 is more important for DNA methylation, both AGO4 and SPT5L are necessary for nucleosome positioning.

Introduction

The Transcriptional Gene Silencing (TGS) pathway is responsible for silencing transposons by directing chromatin modification to discrete loci [1]. Central to this pathway is long non-coding RNA (lncRNA) which acts as a scaffold for proteins to bind chromatin [2,3]. While most features of TGS are conserved in eukaryotes [1,4], in *Arabidopsis thaliana* lncRNA is produced by RNA Polymerase V (Pol V) [3], an RNA Polymerase separate from Pol II. This distinction allows effective genetic studies in this organism.

TGS, also known as RNA-directed DNA methylation (RdDM), directs chromatin modification by both lncRNA and siRNA to which ARGONAUTE (AGO) proteins bind[1,2,5,6]. AGO4 is directed to transposons and gene promoters by the presence of Pol V / lncRNA and by siRNA [5,7]. These components are then necessary for the *de novo* DNA methyltransferase DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) to place DNA methylation [1].

In addition to DNA methylation, RdDM directs other types of chromatin modification, one of which is nucleosome positioning [8]. Another protein, INVOLVED IN DE NOVO 2 (IDN2), interacts with Pol V transcripts and is thought to guide SWI/SNF chromatin remodelers to RdDM targets [8,9]. It was found that knockout of Pol V (*nrpe1*

mutant) and thus loss of IDN2 binding, leads to decreases in nucleosome occupancy [8]. AGO4 is also necessary for nucleosome positioning at several loci, but whether SPT5L participates in this function is unknown [8].

A less well understood component of RdDM is SPT5L (Supressor of Ty insertion 5 - Like) which is similar to SPT5 a Pol II elongation factor [10]. While it has been proposed that SPT5 works to guide stalled Pol II transcription through nucleosomes [11], SPT5L on the other hand interacts with lncRNA and is dependent on Pol V to bind chromatin [10,12]. Also unlike canonical SPT5, SPT5L contains a GW/WG sequence repeat domain, also known as an AGO hook, and interacts with AGO4 [10]. Locus specific assays have shown that SPT5L binds chromatin independent of AGO4 and that DNA methylation defects occur in the *spt5l* mutant only at some loci [12]. The role of SPT5L in RdDM remains mostly unknown as the limited effects to DNA methylation suggest a lesser or different role than that of AGO4.

Here we investigate locus variability of RdDM components and the role of SPT5L in RdDM. By using genome-wide binding site data we show that SPT5L and AGO4 bind chromatin independent of each other at a majority of loci, but that they can influence each other through effects to Pol V transcripts. We further show that AGO4 binding to chromatin plays a major role in directing *de novo* DNA methylation, while SPT5L binding does not. We also show that *spt5l* and *idn2* defects in DNA methylation correlate well with each other and that both AGO4 and SPT5L are necessary for nucleosome positioning.

Results

Spt5l binding sites reflect RdDM activity

While it is known that SPT5L and AGO4 are both important components of RdDM [1,10,12], their individual roles in this process remain mysterious. Previously, we found that mutation of *spt5l* leads to DNA methylation defects at some RdDM targets, but not at others. To investigate this locus variability in SPT5L dependent chromatin modification, we first asked whether SPT5L works mainly in RdDM or elsewhere.

To investigate whether SPT5L is important in RdDM genome wide, we performed ChIP-seq using SPT5L antibody in Col-0 wild-type, *ago4*, and *spt5l* mutant. We found

5011 loci showing significant SPT5L binding (enrichment compared to the *spt5l* mutant), with a false discovery rate (FDR) < 0.001. These loci display prominent peaks in Col-0 indicating that binding site discovery was accurate (Figure 5.1A). To further verify ChIP-seq data we tested several loci by ChIP-qPCR with SPT5L antibody in three biological replicates, and found that all tested loci showed enrichment in Col-0 compared to *spt5l* (Figure 5.7A). These loci included sites with low (JR693), middle (soloLTR, JR341), or high (JR213) peak calling scores, as well as a few sites that did not meet our stringent enrichment criteria for peak calling but showed at least minor read enrichment (JA19, JR587, JR687), illustrating the stringency of our SPT5L peak calls. We also tested whether SPT5L binding is dependent on Pol V at these loci and found that at each locus tested, SPT5L signal is reduced to background levels in *nrpe1* (Figure 5.7A). This is consistent with previous findings that SPT5L targets chromatin through interaction with POL V transcripts [10,12].

Due to the similarities between SPT5L and SPT5, we asked whether SPT5L acts outside of RdDM [11,13]. To test whether detected SPT5L binding sites represent RdDM targets or function like SPT5 at Pol II transcripts, we first examined SPT5L binding sites in relation to transcriptional start sites (TSS) of protein coding genes. We see enrichment of SPT5L directly upstream of genes, matching that which was found for AGO4 [7] (Figure 5.7B). To further investigate the possibility of SPT5L working at genes we examined binding on genes, promoters, and transposons. Protein coding genes actually show depletion of detectable SPT5L compared to random peaks (random genomic regions; $p < .001$); however, promoters and transposons show enrichment of SPT5L (Figure 5.1B, $p < .001$). This localization to promoters and transposons is similar to what has been shown for both Pol V and AGO4 [7,14,15] and further implicates SPT5L's function as part of RdDM.

To test whether SPT5L bound loci have other components of RdDM, we examined previously published Pol V ChIP-seq data [14]. Indeed we see Pol V signal enrichment present at loci where SPT5L is bound (Figure 5.1C). We also examined previously published AGO4 ChIP-seq data [7] and found enrichment of AGO4 signal at SPT5L binding sites (Figure 5.1D). Furthermore, SPT5L binding sites represent functionally significant RdDM loci as seen by enrichment of CHH methylation at these

loci compared to random genomic regions (Figure 5.1E, p<.001). Due to the presence of several hallmarks of RdDM at SPT5L binding sites, we conclude that SPT5L mainly functions as part of RdDM.

SPT5L and AGO4 independent chromatin binding is a genome-wide trend

SPT5L interacts with AGO4 through a reiterated GW/WG repeat doman and it was originally proposed that SPT5L functions as a bridge between AGO4 and DRM2. However, we have shown that SPT5L and AGO4 bind chromatin independent of each other at individual loci. I is unknown whether these loci represent the majority of RdDM targets in the genome. We decided to test whether SPT5L and AGO4 binding chromatin independently represents RdDM sites genome-wide.

In order to explore the relationship between SPT5L and AGO4 we performed ChIP-seq with AGO4 antibody in Col-0, *spt5l*, and *ago4* mutant. Using our same set of stringent peak calling criteria, we found 3988 significantly enriched AGO4 binding sites (FDR < .001). Median profiles around these sites display enrichment of AGO4 in Col-0 vs *ago4* indicating accurate binding site discovery (Figure 5.2A). We also tested several loci by ChIP-qPCR with AGO4 antibody, and found that all tested loci showed enrichment in Col-0 compared to *ago4* (Figure 5.8A). These loci also included ones not present in the list of called peaks (JR341, JR587), illustrating the stringent requirements of peak calling. These loci were also reduced to background levels in *nrpe1* which is consistent with previous findings that AGO4 depends on Pol V to bind chromatin (Figure 5.1D) [5,7].

To test if detected AGO4 binding sites represent RdDM targets we examined Pol V ChIP-seq data [14] at these loci. Pol V ChIP-seq signal is indeed enriched on AGO4 peaks (Figure 5.2B) as it was on SPT5L peaks (Figure 5.1C) and is consistent with a previous genome-wide study of AGO4 [7]. We then checked whether SPT5L is enriched at AGO4 binding sites and see enrichment of SPT5L (Figure 5.2C). To further verify the accuracy of this data we also examined AGO4 signal on SPT5L peaks and see strong enrichment there (Figure 5.8B). Furthermore, detected AGO4 binding sites display an enrichment of CHH methylation (Figure 5.2D, p<.001) consistent with what was previously shown [7]. From this we conclude that detected AGO4 binding sites represent

targets of RdDM, and that Pol V, SPT5L, and AGO4 bind chromatin at the same loci consistent with a model where these components work in complex with each other.

In order to test whether SPT5L binds chromatin independent of AGO4, we performed SPT5L ChIP-seq in the *ago4* mutant. Overall levels of SPT5L are unaffected by *ago4* (Figure 5.1A), indicating that at a majority of loci SPT5L does not depend of AGO4 to bind chromatin. We also tested whether AGO4 binds chromatin independent of SPT5L by performing AGO4 ChIP-seq in the *spt5l* mutant. Overall levels of AGO4 are unaffected by *spt5l*, indicating that AGO4 does not depend on SPT5L to bind chromatin (Figure 5.2A). We conclude that AGO4 and SPT5L binding to chromatin independent of each other is indeed a genome-wide phenomenon.

Pol V transcript availability determines AGO4 and SPT5L binding to chromatin

While it is true generally that SPT5L and AGO4 bind to chromatin independent of each other, variations of RdDM occur. This is seen in variations in feedback between chromatin marks, variations in small RNA production, and variations in RdDM components necessary for silencing [12,14,16]. We decided to ask whether variations in SPT5L and AGO4 binding dependencies occur.

To check whether SPT5L is influenced by *ago4* at some loci, we categorized SPT5L binding sites into those reduced in the *ago4* mutant (*ago4*/Col-0 < 0 .5), those unchanged in *ago4* (0.5 < *ago4*/Col-0 < 2), and those that increase in *ago4* (*ago4*/Col-0 > 2). While the majority of SPT5L peaks are independent of AGO4 (63%), some are dependent on (21%) or are suppressed by AGO4 (16%) (Figure 5.3A, Figure 5.9A). This indicates that while the general trend confirms that SPT5L binds chromatin independently of AGO4, locus specific variation can occur.

To determine if AGO4 is influenced by *spt5l* we divided AGO4 peaks into those reduced, unchanged, or increased in *spt5l*. While AGO4 mostly binds chromatin independent of SPT5L (73%), loci where AGO4 is reduced (9%) or increased (18%) in *spt5l* mutant are present (Figure 5.3B, Figure 5.9B). This indicates that AGO4 is indeed influenced by SPT5L at some loci.

SPT5L and AGO4 mostly bind chromatin independent of each other, but dependent on Pol V (Figure 5.1A, Figure 5.1B) [7,12]. Furthermore, the signal strength of both SPT5L and AGO4 corresponded to that of Pol V (Figure 1C, Figure 2B). We

sought to determine whether variations of SPT5L and AGO4 binding to chromatin are explained by variations to Pol V transcripts upon which they depend. We first tested whether changes to SPT5L binding in the *ago4* mutant can be explained by changes to Pol V transcripts in the *ago4* mutant. Pol V transcripts identified by RT-PCR were tested in the *ago4* mutant and we found that the *ago4* mutant can result in Pol V transcript levels to decrease, increase, or remain the same (Figure 5.9C). When comparing the effects of *ago4* on Pol V transcripts to the effects of *ago4* on SPT5L ChIP-seq data we see a correlation between changes in SPT5L signal and changes in Pol V transcript signal in the *ago4* mutant (Figure 5.3C, $R^2 = .6356$). In essence Pol V transcripts that decrease in the *ago4* mutant correspond to SPT5L signal decreases in *ago4*. Likewise, Pol V transcript increases occur at the same loci where SPT5L signal increases in *ago4*. This is consistent with a model where Pol V transcripts act as scaffolds for SPT5L binding to chromatin and can explain the locus specific effects that *ago4* has on SPT5L.

Next we tested whether the *spt5l* mutant can affect Pol V transcript levels; we tested the aforementioned Pol V transcripts and found that *spt5l* can indeed cause changes to transcript levels (Figure 5.9C). We compared these changes in transcript levels to the changes in AGO4 binding signal in *spt5l* and found correlation between transcript level changes and AGO4 signal changes (Figure 5.3D – $R^2 = .5425$). Loci with increased Pol V transcripts in *spt5l* corresponded to AGO4 signal increased in *spt5l*. Likewise, decreased Pol V transcripts in *spt5l* corresponded to AGO4 signal decreased in *spt5l*. We propose that locus specific effects to AGO4 by *spt5l* can be explained at least in part by the effects of *spt5l* on Pol V transcripts.

Since variations of chromatin binding of both SPT5L and AGO4 may be explained by Pol V transcript levels, we sought to determine whether these effects on the two proteins occur at the same loci. We combined both lists of binding sites (AGO4 and SPT5L) and examined the overlap between SPT5L reduced in *ago4* and AGO4 reduced in *spt5l*. Although SPT5L sites reduced in *ago4* coincided relatively well with AGO4 sites reduced in *spt5l*, they most often coincide with sites where AGO4 is unchanged in *spt5l* (Figure 5.3E, Figure 5.9D). This is true for all sites tested whether examining binding reduction, or binding increases. This indicates that locus variability of SPT5L and AGO4 chromatin binding do not occur at the same loci. We propose that the

effects exerted by SPT5L and AGO4 on each other likely result from the altered availability of Pol V transcripts, but that they affect transcription in different ways. Despite these locus specific effects, however, it should be noted that the majority of loci reflect a mechanism where SPT5L and AGO4 bind chromatin independent of each other.

SPT5L plays a more limited role than AGO4 in directing DNA methylation

Although both SPT5L and AGO4 bind at RdDM targets (Figure 1E, Figure 2D), the effect of SPT5L on DNA methylation varies at individual loci [12]. We sought to determine whether the variable role of SPT5L on DNA methylation can be explained by variations in chromatin binding.

Utilizing previously published bisulfite sequencing data [17], we identified regions with Pol V dependent CHH methylation (*nrpe1* DMRs) and found that the majority of DMRs overlapped those of *ago4* (Figure 5.4A, Figure 5.10A). In contrast, only about half of *nrpe1* DMRs were also *spt5l* DMRs (Figure 5.4A). This confirms that, while AGO4 is necessary for DNA methylation at most RdDM targets, SPT5L's role is more variable.

We next tested if *spt5l* causes severe DNA methylation reduction at some loci and leaves others completely unchanged. In fact, most *nrpe1* DMRs display at least some reduction in *spt5l,* although not nearly as much as in *ago4* (Figure 5.4BC, Figure 5.10BC). The contrast between the effects of *ago4* and *spt5l* on DNA methylation indicates that SPT5L's role in directing DNA methylation is not an "on" or "off" scenario, but a gradient of effects.

The limited role of SPT5L in directing DNA methylation may result from limited binding of SPT5L at these sites. To test this possibility we tested if SPT5L binding is stronger at *nrpe1* DMRs with severely reduced DNA methylation in *spt5l* than at *nrpe1* DMRs with minor or unchanged methylation in *spt5l*. Interestingly, SPT5L binding is enriched for these sites to the same degree (Figure 5.4D). This suggests that SPT5L binding to chromatin is not directly linked to directing *de novo* DNA methylation.

Although, AGO4 is necessary at a majority of nrpe1 DMRs, variations in DNA methylation defects occur in this mutant also (Figure 5.10A). We tested whether variations in methylation dependency on AGO4 can be explained by variations in AGO4 binding. AGO4 binding signal is seen at *nrpe1* DMRs affected in *ago4*, but not at those

unaffected in *ago4* (Figure 5.4E). This is in stark contrast to SPT5L where SPT5L binding was seen at sites unaffected in *spt5l* (Figure 5.4D). This suggests that AGO4, more than SPT5L, corresponds to directing DNA methylation at a majority of loci.

We found loci where AGO4 is reduced in *spt5l* (Figure 5.3C); we also found loci where DNA methylation depends heavily on *spt5l* (Figure 5.4AB, Figure 5.10AB)*;* we sought to determine whether *spt5l* affects DNA methylation by influencing AGO4 binding to chromatin. We took categorized AGO4 binding sites that are reduced, unchanged, or increased in *spt5l* and tested whether these corresponded to the effects of *spt5l* on CHH methylation. Loci with AGO4 signal reduced in *spt5l* correspond to more severe reductions to DNA methylation in *spt5l* mutant (Figure 5.4F). This means that the most dramatic methylation changes in *spt5l* are in fact due to reduction in AGO4 binding. In contrast AGO4 binding sites that are unchanged or increased in *spt5l* have less severe DNA methylation reductions in *spt5l* (Figure 5.4F). This suggests that spt5l has a limited effect on DNA methylation at most loci, and that locus specific severe effects of this mutant may be explained by feedback to AGO4 chromatin binding. Overall, we propose that AGO4 is important for directing DNA methylation, and that SPT5L plays a more limited role in this aspect.

Both SPT5L and AGO4 are necessary for RdDM specific nucleosome positioning

SPT5L's limited role in directing CHH methylation could possibly indicate that it has a distinct role from that of AGO4 in the RdDM pathway. To find components with similar effects as *spt5l*, we checked whether other mutants in the RdDM pathway display limited methylation defects. Using a principle component analysis of methylation at *nrpe1* DMRs we examined the effects of 11 different mutants on CHH methylation. Interestingly *spt5l*, *idn2*, and the *idn2/idnl1/idnl2* triple mutant clustered together indicating that these have similar effects on *de novo* DNA methylation (Figure 5.5A). Additionally, while these mutants cluster together, they all were closer to mutants for proteins involved in DNA methylation than to Col-0 (WT) demonstrating their role, however limited, in directing methylation (Figure 5.5A). This is in contrast to that of *ago1* which is not connected to POLV dependent DNA methylation and has CHH methylation most similar to Col-0 wild-type (Figure 5.5A).

In order to explore the relationship between SPT5L and IDN2 we further examined methylation levels at *nrpe1* DMRs. We first compared CHH methylation changes in *idn2/idnl1/idnl2* and in *spt5l* and found that the methylation level reductions in these mutants correlate well with each other (Figure 5.5B, Pearson Correlation = 0.758). We next compared the *idn2* triple mutant to *ago4* and found that CHH methylation is generally reduced more in *ago4* (Figure 5.5C). This is similar to *spt5l* and *ago4* (Figure 5.4BC) in that *ago4* has a much larger effect on methylation than *spt5l* (Figure 5.5D). In contrast *ago4* and *drm1/drm2* show more similar methylation defects to each other (Figure 5.11A, Pearson Correlation = 0.43). Overall this suggests that AGO4 is mainly responsible for directing DRM2 dependent methylation, and that SPT5L and IDN2 may work together or similarly in RdDM.

Previously we have shown that IDN2 plays an important role in nucleosome positioning [8]. Since *idn2* and *spt5l* methylation patterns suggest that a connection may exist between the two, we tested whether SPT5L is important for nucleosome positioning. We performed H3 ChIP-seq on MNase digested chromatin to obtain genome-wide nucleosome maps in Col-0, *nrpe1*, *ago4*, and *spt5l*. We then took nucleosomes depleted in *nrpe1* and examined their occupancy in *ago4* and *spt5l*. Interestingly, most nucleosomes reduced in *nrpe1* were also reduced in both *ago4* and *spt5l* (Figure 5.5E). This indicates that while SPT5L plays a limited role in directing DNA methylation, both AGO4 and SPT5L are necessary for nucleosome positioning.

Due to the strong effects of the *ago4* mutant on CHH methylation (Figure 5.4C) we checked whether the methylation defects can explain nucleosome dependency. We took nucleosomes reduced in *nrpe1* and examined changes to DNA methylation in *ago4* and *spt5l*. Many of these nucleosomes actually displayed no change in DNA methylation in either *ago4* or *spt5l,* suggesting that nucleosome positioning functions independent of *de novo* methylation (Figure 5.11B). We further filtered these nucleosomes for those that represent *nrpe1* DMRs. These nucleosomes display the same pattern of DNA methylation changes as total *nrpe1* DMRs (Figure 5.5F, Figure 5.5D). This indicates that AGO4 and SPT5L are important for nucleosome positioning independent of their roles in directing *de novo* DNA methylation. We also compared nucleosome changes in *ago4* to CHH methylation changes in *ago4* and see no

correlation (Figure 5.11C). Similarly comparison of nucleosome changes in *spt5l* to CHH methylation changes in *spt5l* displays no correlation (Figure 5.11D). We propose that AGO4 and SPT5L direct nucleosome positioning independent of their functions in directing *de novo* DNA methylation.

Discussion

Despite the finding that SPT5L is involved in RdDM [10], it was found that the *spt5l* mutant has locus specific effects to DNA methylation and histone modifications [12]. Since then the exact function of SPT5L in RdDM has been in question. Here we show that AGO4 and SPT5L bind chromatin independent of each other, but that one may affect the other indirectly by feedback to Pol V transcripts. Furthermore, while it was supposed that SPT5L was only important for DNA methylation at some RdDM sites, *spt5l* causes at least minor changes in DNA methylation at nearly all *nrpe1* DMRs (Figure 5.4B). This methylation pattern is similar to that of *idn2* mutants (Figure 5.5D), which have been identified as proteins necessary for nucleosome positioning [8]. Indeed SPT5L is also important for nucleosome positioning to the same degree (if not more than) AGO4 (Figure 5.5E).

Therefore, despite the limited effects to DNA methylation in *spt5l*, we propose a model where SPT5L and AGO4 are necessary for directing nucleosome remodeler binding or function at RdDM loci (Figure 5.6). In this model SPT5L and AGO4 bind chromatin independent of each other, but dependent on Pol V transcripts. AGO4 is then responsible for directing *de novo* DNA methylation and both SPT5L and AGO4 are essential to direct nucleosome positioning.

A role in nucleosome positioning is intriguing due to the similarities between SPT5L and the Pol II elongation factor, SPT5 [10,11]. Canonical SPT5 has been proposed to guide Pol II through nucleosomes and reposition the nucleosome after transcription has passed [11,13]. SPT5L may function at Pol V transcripts similar to SPT5, however it is likely that Pol V transcribes mostly independent of *spt5l* (as seen by limited effects to AGO4 binding in *spt5l*). It is more likely that Pol V transcription occurs before placement of nucleosomes and that SPT5L and AGO4 help direct chromatin modifiers to RdDM targets.

The similarity between SPT5L and SPT5 could also allow for compensation between the two. This might explain locus variability of the effects of *spt5l* on DNA methylation in that SPT5 may be able to at least partially compensate for SPT5L. However, changes in nucleosome positioning were as dramatic in *spt5l* as they were in *ago4* suggesting that full compensation does not occur. Additionally, SPT5L is likely specific for RdDM as it was enriched upstream of genes and was conspicuously absent inside Pol II genes (Figure 5.1B, Figure 5.7B). In comparison, SPT5 in mouse was also shown to bind near genes, however in this system two profile summits are present: one directly upstream of the TSS and one slightly downstream or at the TSS [18], suggesting that two separate functions of SPT5 may exist in other systems. We propose that, like other components of RdDM in Arabidopsis, SPT5L's function has been specialized for RdDM and is separate from that of SPT5.

Overall, despite the limited role of SPT5L in directing DNA methylation, this component of RdDM is still vital for effective chromatin silencing. In conclusion, our findings indicate that while AGO4 is necessary for directing *de novo* DNA methylation, SPT5L and AGO4 together help direct nucleosome occupancy at RdDM targets.

Materials and methods

Plant Material and Antibodies

nrpe1 (*nrpd1b-11*), *spt5l* (*rdm3-3*), and *ago4* [(*ago4-1*) introgressed into the Col-0 background] have been described previously [5,10,19]. α-SPT5L, α-AGO4, and α-IDN2 antibodies were also described previously [5,10,20].

RT-PCR

Total RNA was isolated from above ground tissue of three week old seedlings using the SV total RNA Isolation kit from Promega. After isolation, an additional incubation with Turbo DNase was performed as described [21]. cDNA synthesis was performed using 3µg RNA with random primers (Invitrogen). Mean values relative to Col-0 and standard deviations were calculated from three biological replicates. Pol V transcripts were identified as loci with signal reduction in *nrpe1* and were then tested in *ago4* and *spt5l*. ChIP-seq reads covering these regions were counted and plotted relative to Col-0. Primers used are included in Appendix E.

Chromatin Immunoprecipitation

3g above ground tissue of three week old seedlings were crosslinked and used in ChIP mostly as described [21]. In lieu of phenol : chloroform extraction, immunoprecipitated samples were purified using the QIAquick PCR Purification Kit from Qiagen, after which libraries were prepared using the Diagenode Microplex Library Preparation Kit followed by Illumina sequencing. For H3 ChIP-seq, nuclei were resuspended in MNase buffer (NEB) after final wash in Honda buffer. Chromatin was digested in the presence of 6,000 Gel Units MNase (NEB) at 37°C for ten minutes. Nuclei were then broken by sonicating on ice three times for ten seconds each on the lowest setting. Immunoprecipitation and sample preparation were then performed as described for SPT5L and AGO4.

Bioinformatics analysis

Mapping to the TAIR10 genome and peak calling were performed using SOAP2 and CSAR as previously described [21,23,24]. SPT5L peaks found in Col-0 vs *spt5l* were combined with those found in *ago4* vs *spt5l* for downstream analysis. AGO4 peaks found in Col-0 vs *ago4* were combined with those found in *spt5l* vs *ago4* for downstream analysis. Overlap between peaks and genomic features were done using TAIR10 annotations. Genome-wide ChIP-seq for Pol V and AGO4, and for bisulfite sequencing data used was taken from the NCBI data repository SRP013929 [14], GSE35381 [7], and GSE39901 [17] respectively. The *ago4* mutant was backcrossed from the Landsberg ecotype and could still contain some genomic sequence similarity, thus AGO4 peaks matching these regions were filtered as described [7]. CHH methylated regions were identified using a 2 step process: 1. Regions were identified as having more than one methylated base in a CHH context (> 10% methylation score) within 100bp of each other; 2. Regions less than 50bp in length and with less than 10% methylation per 10 bp were filtered to keep only highly methylated regions. DMRs were then called as having less than 25% methylation in the mutant vs Col-0. Weighted Venn diagrams of peaks and of DMRs were created using the Venneuler package in R [25]. PCA analysis of *nrpe1* DMRs was performed using the prcomp function in R and the first two principle components were plotted. Pol V dependent nucleosomal regions were

identified as sites were H3 ChIP signal in Col-0 was at least two fold greater than in *nrpe1* and reads in *ago4* and *spt5l* on these regions were counted.

References

1. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11,** 204–220 (2010).
2. Wierzbicki, A. T. The role of long non-coding RNA in transcriptional gene silencing. *Curr. Opin. Plant Biol.* **15,** 517–522 (2012).
3. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135,** 635–648 (2008).
4. Girard, A. & Hannon, G. J. Conserved themes in small-RNA-mediated transposon control. *Trends Cell Biol.* **18,** 136–148 (2008).
5. Wierzbicki, A. T., Ream, T. S., Haag, J. R. & Pikaard, C. S. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat. Genet.* **41,** 630–634 (2009).
6. Hutvagner, G. & Simard, M. J. Argonaute proteins: key players in RNA silencing. *Nat. Rev. Mol. Cell Biol.* **9,** 22–32 (2008).
7. Zheng, Q. *et al.* RNA polymerase V targets transcriptional silencing components to promoters of protein-coding genes. *Plant J.* (2012). doi:10.1111/tpj.12034
8. Zhu, Y., Rowley, M. J., Böhmdorfer, G. & Wierzbicki, A. T. A SWI/SNF Chromatin-Remodeling Complex Acts in Noncoding RNA-Mediated Transcriptional Silencing. *Mol. Cell* **49,** 1–12 (2013).
9. Ausin, I., Mockler, T. C., Chory, J. & Jacobsen, S. E. IDN1 and IDN2 are required for de novo DNA methylation in Arabidopsis thaliana. *Nat. Struct. Mol. Biol.* **16,** 1325–1327 (2009).
10. He, X.-J. *et al.* An effector of RNA-directed DNA methylation in arabidopsis is an ARGONAUTE 4- and RNA-binding protein. *Cell* **137,** 498–508 (2009).
11. Hartzog, G. A. & Fu, J. The Spt4-Spt5 complex: a multi-faceted regulator of transcription elongation. *Biochim. Biophys. Acta* **1829,** 105–115 (2013).
12. Rowley, M. J., Avrutsky, M. I., Sifuentes, C. J., Pereira, L. & Wierzbicki, A. T. Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. *PLoS Genet.* **7,** e1002120 (2011).
13. Hartzog, G. A., Wada, T., Handa, H. & Winston, F. Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in Saccharomyces cerevisiae. *Genes Dev.* **12,** 357–369 (1998).
14. Wierzbicki, A. T. *et al.* Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. *Genes Dev.* **26,** 1825–1836 (2012).
15. Zhong, X. *et al.* DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. *Nat. Struct. Mol. Biol.* **19,** 870–875 (2012).
16. Matzke, M. A. & Mosher, R. A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15,** 394–408 (2014).

17. Stroud, H., Greenberg, M. V. C., Feng, S., Bernatavichute, Y. V. & Jacobsen, S. E. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. *Cell* **152,** 352–364 (2013).
18. Pavri, R. *et al.* Activation-Induced Cytidine Deaminase Targets DNA at Sites of RNA Polymerase II Stalling by Interaction with Spt5. *Cell* **143,** 122–133 (2010).
19. Onodera, Y. *et al.* Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120,** 613–622 (2005).
20. Zhu, Y., Rowley, M. J., Böhmdorfer, G. & Wierzbicki, A. T. A SWI/SNF chromatin-remodeling complex acts in noncoding RNA-mediated transcriptional silencing. *Mol. Cell* **49,** 298–309 (2013).
21. Rowley, M. J., Böhmdorfer, G. & Wierzbicki, A. T. Analysis of long non-coding RNAs produced by a specialized RNA polymerase in Arabidopsis thaliana. *Methods* **63,** 160–169 (2013).
22. Rowley, M. J., Böhmdorfer, G. & Wierzbicki, A. T. Analysis of long non-coding RNAs produced by a specialized RNA polymerase in Arabidopsis thaliana. *Methods* **63,** 160–169 (2013).
23. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).
24. Muiño, J. M., Kaufmann, K., van Ham, R. C., Angenent, G. C. & Krajewski, P. ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. *Plant Methods* **7,** 11 (2011).
25. Wilkinson, L. Exact and approximate area-proportional circular Venn and Euler diagrams. *IEEE Trans Vis Comput Graph* **18,** 321–331 (2012).

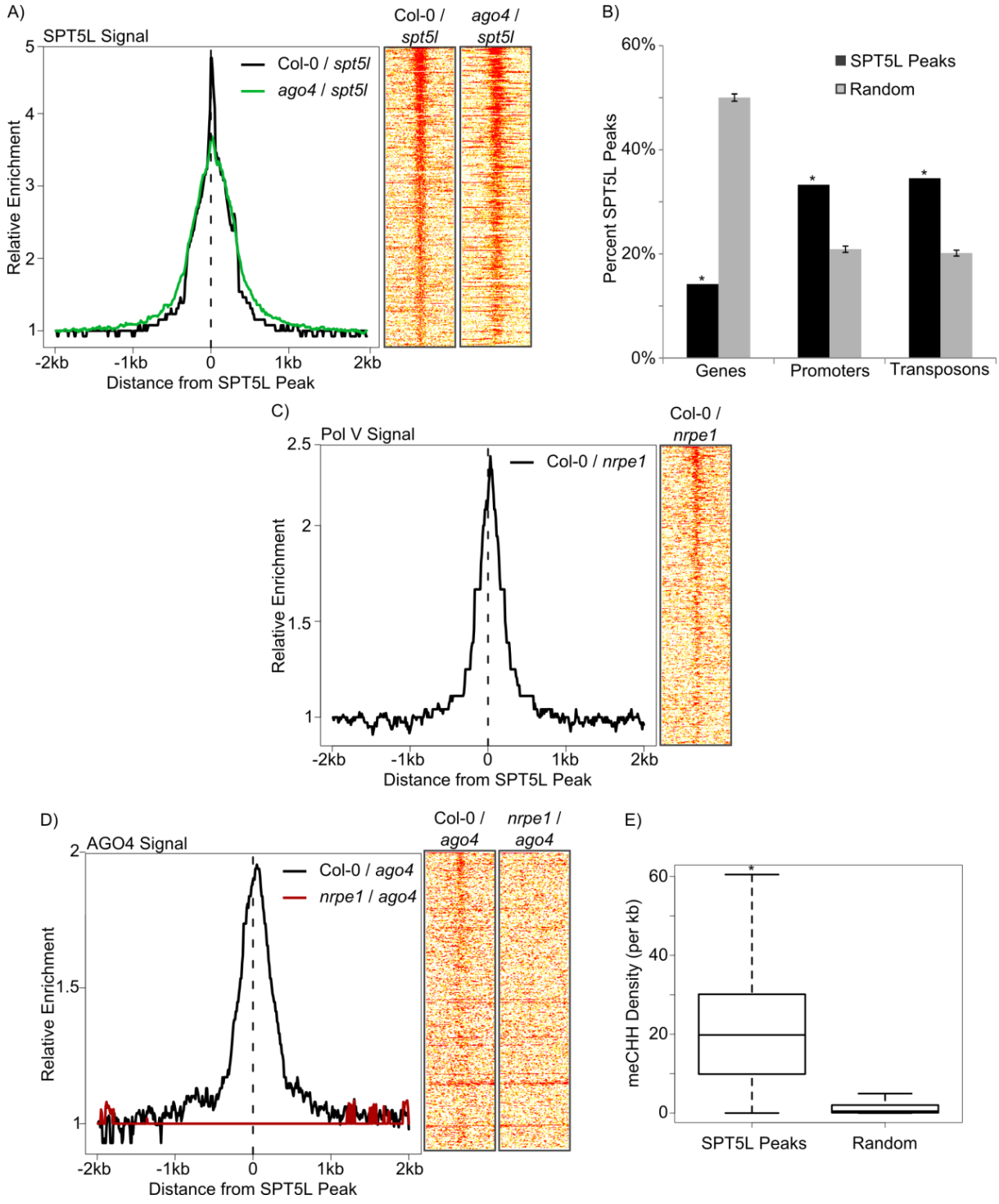


Figure 5.1 SPT5L binding sites reflect RdDM activity

A) Profile and heatmap of SPT5L ChIP-seq signal on called SPT5L binding sites. Graph shows median SPT5L ChIP-seq signal ratios in Col-0 vs *spt5l* (black) or *ago4* vs *spt5l* (green) relative to background around SPT5L peaks. Dashed line indicates SPT5L peak summit around which the plot is oriented. Heatmaps display this same region ordered from highest to lowest SPT5L peak score with red indicating strong Col-0/*spt5l* or *ago4*/*spt5l* SPT5L ChIP-seq signal ratios and white/yellow indicating

low signal ratios. All signal ratios were taken as reads per million (RPM).
B) SPT5L is enriched on Promoters and Transposons. Barplot indicating the percentage of SPT5L peaks (black) overlapping genes, promoters, and transposons. Random permutations matching the size distribution of SPT5L peaks were overlapped with genes, promoters, and transposons for comparison. Permutations were run 1000 times and the median value was plotted (grey). Error bars indicate standard deviation between permutations. * indicates $p < .001$ as calculated by permutation.
C) Profile and heatmap of Pol V ChIP-seq signal on called SPT5L binding sites. Graph shows median Pol V ChIP-seq signal ratios in Col-0 vs *nrpe1* (black) relative to background around SPT5L peaks. Dashed line indicates SPT5L peak summit around which the plot is oriented. Heatmaps display this same region ordered from highest to lowest SPT5L peak score with red indicating strong Col-0/*nrpe1* Pol V ChIP-seq signal ratios and white/yellow indicating low signal ratios. All signal ratios were taken as reads per million (RPM).
D) Profile and heatmap of AGO4 ChIP-seq signal on called SPT5L binding sites. Graph shows median AGO4 ChIP-seq signal ratios in Col-0 vs *ago4* (black) or *nrpe1*/*ago4* (red) relative to background around SPT5L peaks. Dashed line indicates SPT5L peak summit around which the plot is oriented. Heatmaps display this same region ordered from highest to lowest SPT5L peak score with red indicating strong Col-0/*ago4* or *nrpe1*/*ago4* AGO4 ChIP-seq signal ratios and white/yellow indicating low signal ratios. All signal ratios were taken as reads per million (RPM).
E) *De novo* DNA methylation enrichment on SPT5L binding sites. meCHH signal on SPT5L peaks was counted and normalized by region size (per kb). Random genomic regions were taken and plotted for comparison. * indicates $p < .001$ calculated by t-test.

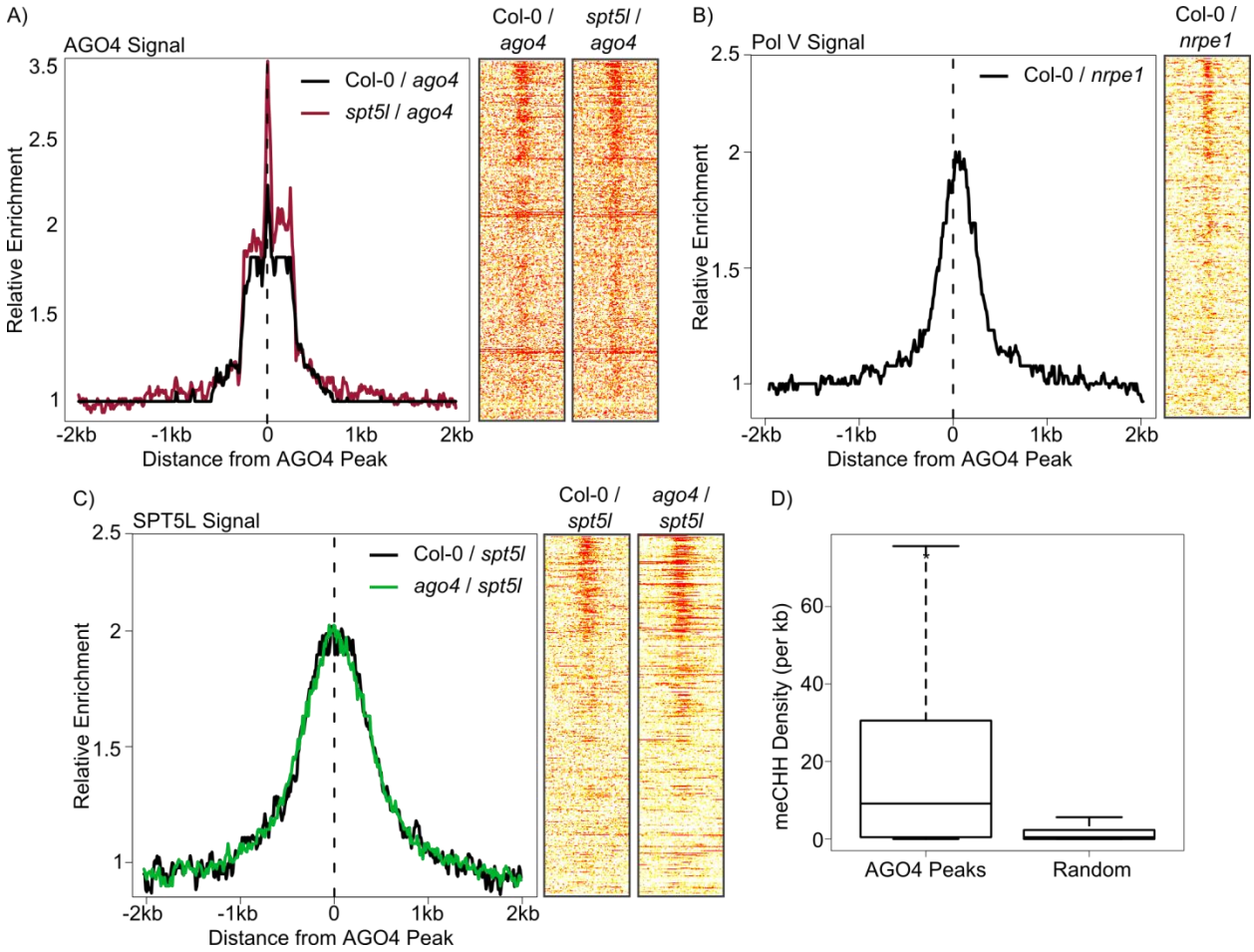


Figure 5.2 AGO4 binding sites reflect RdDM activity

A) Profile and heatmap of AGO4 ChIP-seq signal on called AGO4 binding sites. Graph shows median AGO4 ChIP-seq signal ratios in Col-0 vs *ago4* (black) or *spt5l* vs *ago4* (red) relative to background around AGO4 peaks. Dashed line indicates AGO4 peak summit around which the plot is oriented. Heatmaps display this same region ordered from highest to lowest AGO4 peak score with red indicating strong Col-0/*ago4* or *spt5l*/*ago4* AGO4 ChIP-seq signal ratios and white/yellow indicating low signal ratios. All signal ratios were taken as reads per million (RPM).

B) Profile and heatmap of Pol V ChIP-seq signal on called AGO4 binding sites. Graph shows median Pol V ChIP-seq signal ratios in Col-0 vs *nrpe1* (black) relative to background around AGO4 peaks. Dashed line indicates AGO4 peak summit around which the plot is oriented. Heatmaps display this same region ordered from highest to lowest AGO4 peak score with red indicating strong Col-0/*nrpe1* Pol V ChIP-seq signal ratios and white/yellow indicating low signal ratios. All signal ratios were taken as reads per million (RPM).

C) Profile and heatmap of SPT5L ChIP-seq signal on called AGO4 binding sites. Graph shows median SPT5L ChIP-seq signal ratios in Col-0 vs *spt5l* (black) or *ago4/spt5l* (green) relative to background around AGO4 peaks. Dashed line indicates AGO4 peak summit around which the plot is oriented. Heatmaps display this same region ordered from highest to lowest AGO4 peak score with red indicating strong Col-0/*spt5l* or *ago4/spt5l* SPT5L ChIP-seq signal ratios and white/yellow indicating

low signal ratios. All signal ratios were taken as reads per million (RPM).
D) *De novo* DNA methylation enrichment on AGO4 binding sites. meCHH signal on AGO4 peaks was counted and normalized by region size (per kb). Random genomic regions were taken and plotted for comparison. * indicates $p < .001$ calculated by t-test.

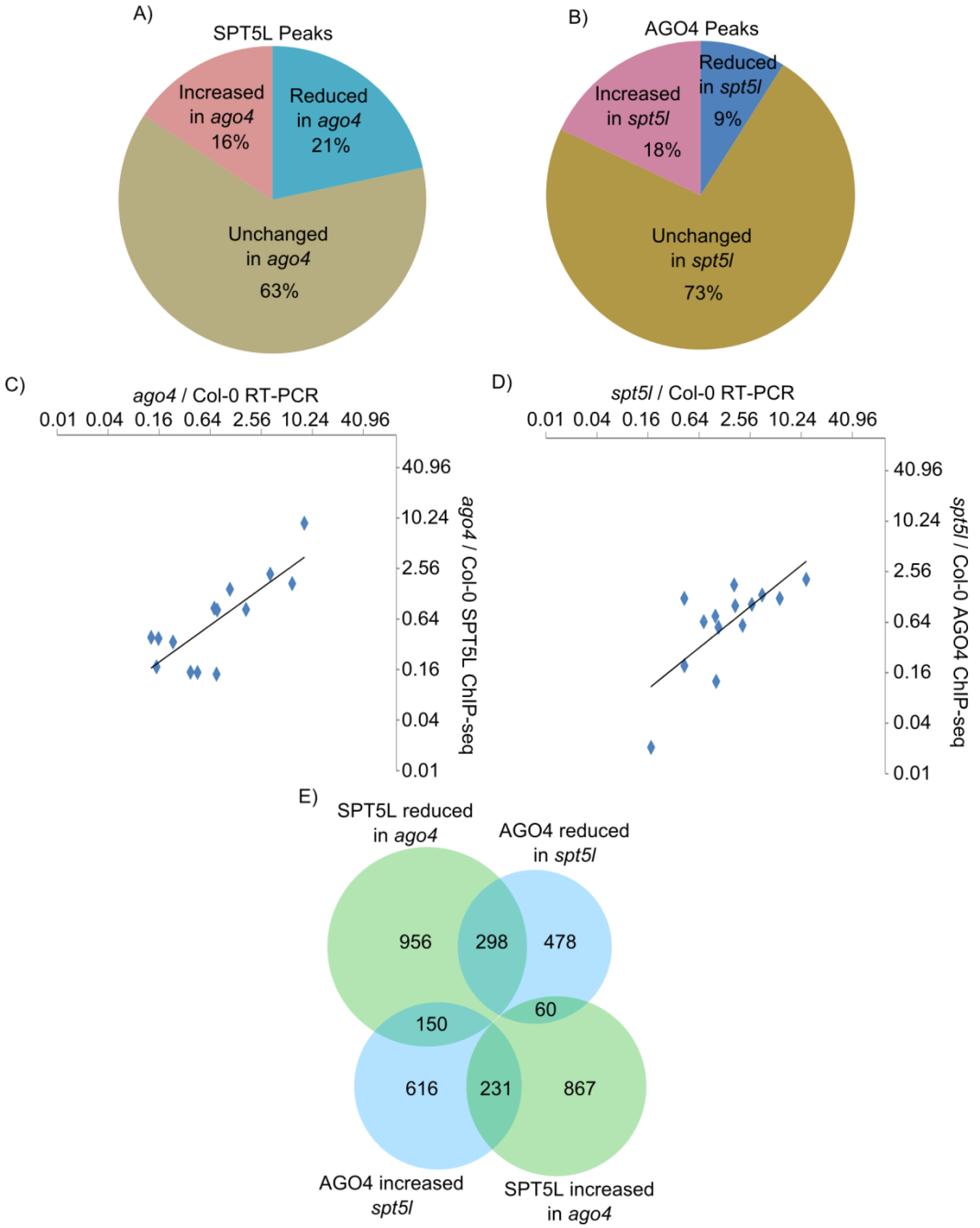


Figure 5.3 Pol V transcript availability determines AGO4 and SPT5L binding to chromatin

A) Categories of SPT5L chromatin binding. SPT5L ChIP-seq signal was grouped into those reduced, increased, or unchanged in *ago4*. Reductions were defined as peaks where *ago4*/Col-0 < 0.5 (21% of called peaks), increases where *ago4*/Col-0 > 2 (16% of called paks), and unchanged where 0.5 < *ago4*/Col-0 < 2 (63% of called peaks). B) Categories of AGO4 chromatin binding. AGO4 ChIP-seq signal was grouped into those reduced, increased, or unchanged in *spt5l*. Reductions were defined as peaks where *spt5l*/Col-0 < 0.5 (9% of called peaks), increases where *spt5l*/Col-0 > 2 (18% of called paks), and unchanged where 0.5 < *spt5l*/Col-0 < 2 (73% of called peaks). C) Relationship between SPT5L binding and Pol V transcript availability. Pol V

transcript changes in *ago4* (*ago4*/Col-0; x-axis) as measured by RT-PCR are plotted in relation to SPT5L ChIP-seq signal changes in ago4 (*ago4*/Col-0; y-axis). A best fit linear correlation line is shown (R^2=0.6356).
D) Relationship between AGO4 binding and Pol V transcript availability. Pol V transcript changes in *spt5l* (*spt5l*/Col-0 RT-PCR; x-axis) as measured by RT-PCR are plotted in relation to AGO4 ChIP-seq signal changes in *spt5l* (*spt5l*/Col-0 AGO4 ChIP-seq; y-axis). A best fit linear correlation line is shown (R^2=0.5245).
E) Overlap between categories of SPT5L and AGO4 bound loci. AGO4 peak and SPT5L peak lists were combined and categorized as in Figure 5.3A. Loci found in different categories are shown with the size of the circle and the amount of overlap relative to the number of peaks in each category.

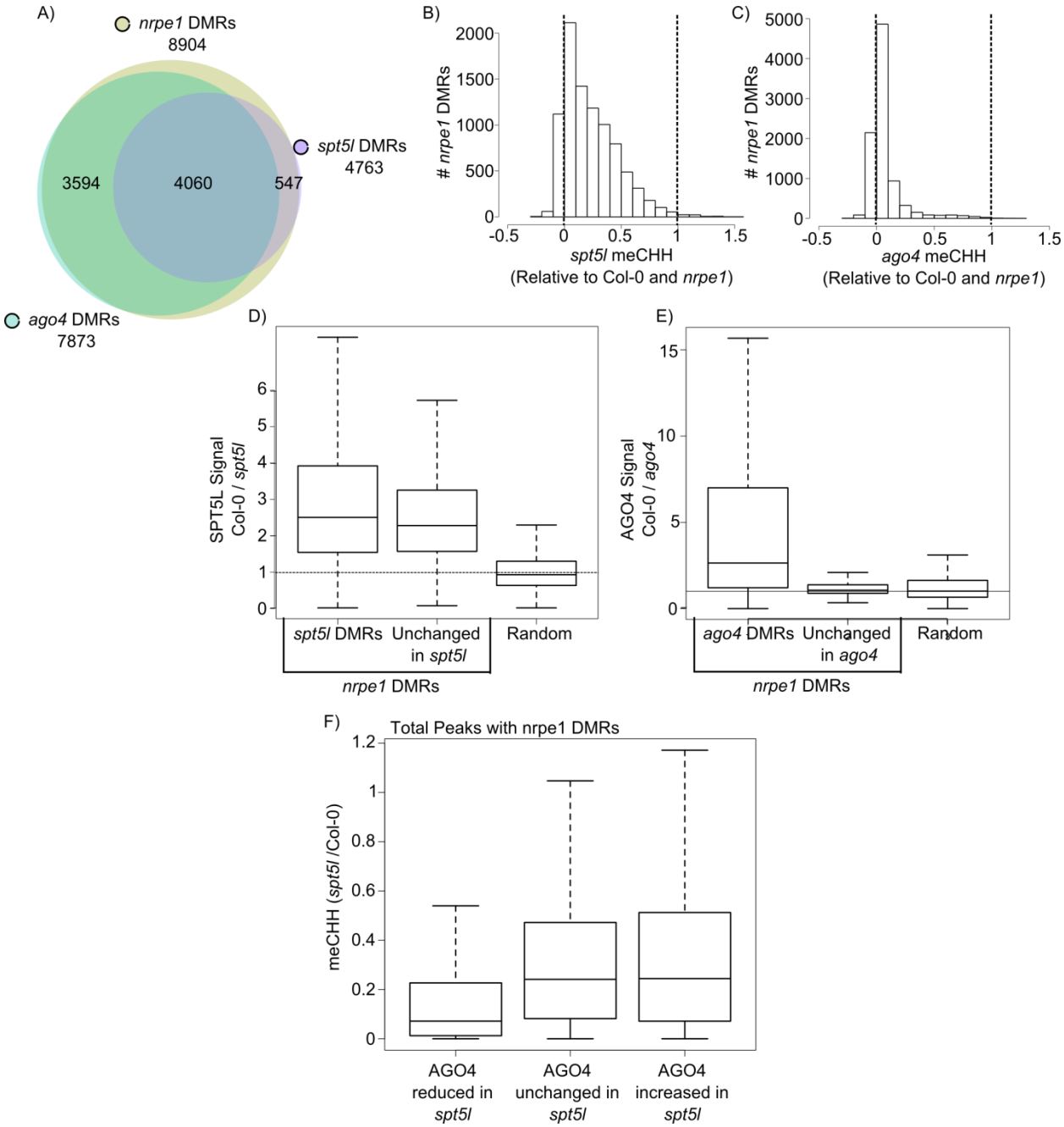


Figure 5.4 SPT5L plays a more limited role than AGO4 in directing DNA methylation

A) Overlap of differentially methylated regions (DMRs). Venn Diagram showing overlap between *nrpe1*, *ago4*, and *spt5l* DMRs identified as regions with at least four fold decreases in CHH methylation compared to Col-0. The size of the circle and the amount of overlap is relative to the number of regions in each category.

B) *spt5l* mutant has limited effects on CHH methylation. Histogram of CHH methylation on *nrpe1* DMRs in the *spt5l* mutant relative to Col-0 and *nrpe1*. Dashed lines at 0 and 1 indicate *spt5l* methylation level similarities to *nrpe1* and Col-0 respectively ([*spt5l* – *nrpe1*] / [Col-0 - *nrpe1*]).

C) *ago4* mutant has dramatic effects on CHH methylation. Histogram of CHH

methylation on *nrpe1* DMRs in the *ago4* mutant relative to Col-0 and *nrpe1*. Dashed lines at 0 and 1 indicate *ago4* methylation level similarities to *nrpe1* and Col-0 respectively ([*ago4* - *nrpe1*] / [Col-0 - *nrpe1*]).
D) SPT5L binds *nrpe1*, *spt5l* DMRs and *nrpe1* DMRs unchanged in *spt5l*. *nrpe1* DMRs were divided into those overlapping *spt5l* DMRs (*spt5l*/Col-0 <= .25) and those with little change in *spt5l* (*spt5l*/Col-0 >= .75) and SPT5L ChIP-seq signal was plotted. SPT5L ChIP-seq signal on random genomic regions is plotted as an estimate of background signal. Dashed line indicates a ChIP-seq signal ratio expected if no enrichment of SPT5L is present.
E) AGO4 binds *nrpe1*, *ago4* DMRs, but not *nrpe1* DMRs unchanged in *ago4*. *nrpe1* DMRs were divided into those overlapping *ago4* DMRs (*ago4*/Col-0 <= .25) and those with little change in *ago4* (*ago4*/Col-0 >= .75) and AGO4 ChIP-seq signal was plotted. AGO4 ChIP-seq signal on random genomic regions is plotted as an estimate of background signal. Dashed line indicates a ChIP-seq signal ratio expected if no enrichment of AGO4 is present.
F) Dramatic meCHH reductions in *spt5l* are explained by reductions in AGO4 binding. AGO4 and SPT5L peaks were pooled and those overlapping *nrpe1* DMRs were kept. These were further categorized as in Figure 5.3C into AGO4 sites reduced, unchanged, or increased in *spt5l*. The level of CHH methylation in *spt5l* relative to Col-0 was plotted in each category.

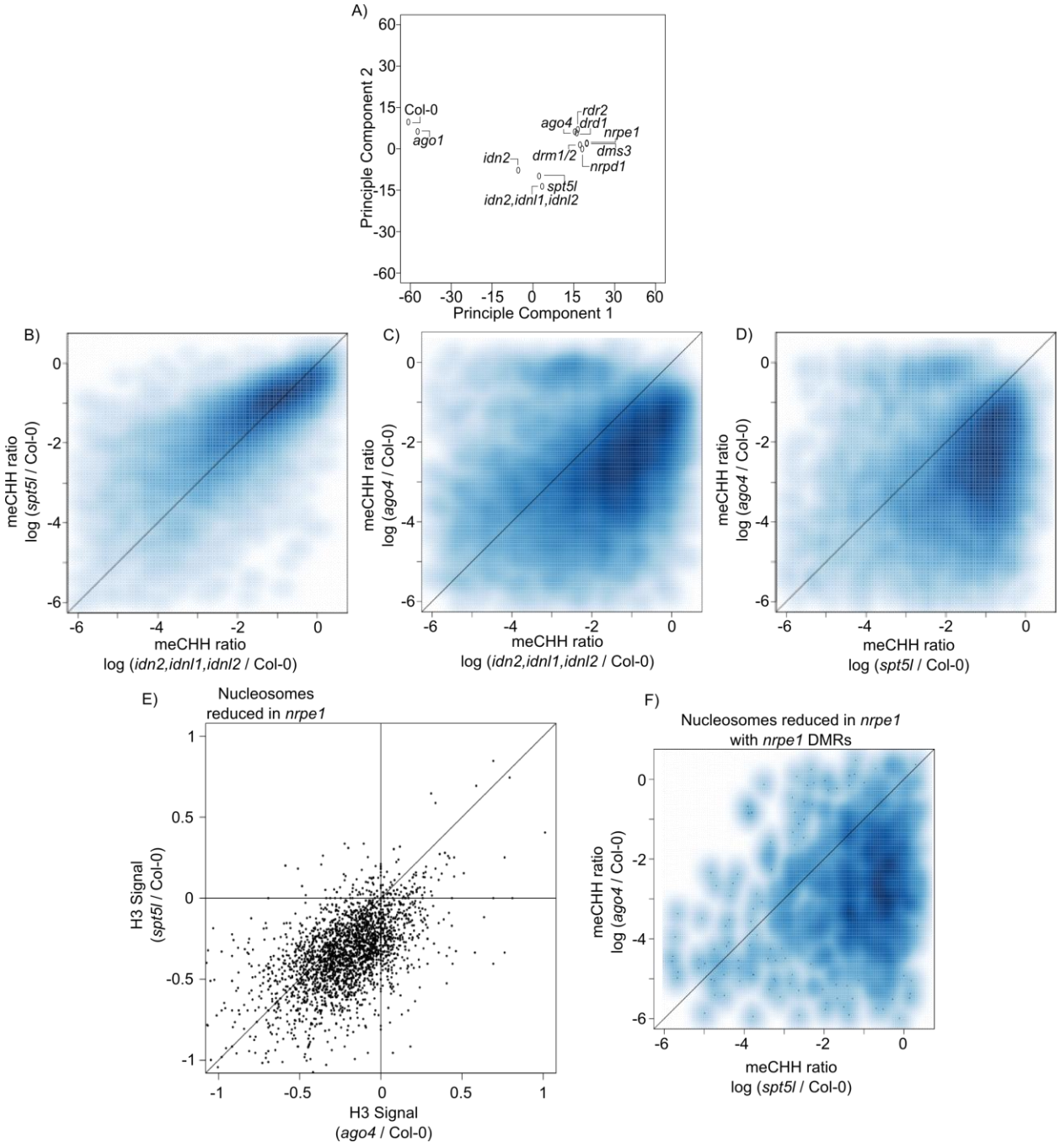


Figure 5.5 Both SPT5L and AGO4 are necessary for RdDM specific nucleosome positioning

A) Principle component analysis of CHH methylation in mutants of RdDM. Ten mutants known to be involved in RdDM were chosen and compared to each other and to Col-0 (wild-type) and *ago1* (not involved in RdDM). The first two principle components are plotted against each other.

B) spt5l mutant affects CHH methylation similar to *idn2/idnl1/idnl2* triple mutant. *nrpe1* DMRs were taken and meCHH levels for *spt5l* and *idn2,idnl1,idnl2* relative to Col-0 were calculated and plotted. The gradient from white to dark blue indicates low to high density points. Solid diagonal line indicates a slope of 1 (i.e. *spt5l* =

idn2/idnl1/idnl2).
C) *ago4* mutant affects CHH methylation more than does *idn2,idnl1,idnl2* triple mutant*. nrpe1* DMRs were taken and meCHH levels for *ago4* and *idn2,idnl1,idnl2* relative to Col-0 were calculated and plotted. The gradient from white to dark blue indicates low to high density points. Solid diagonal line indicates a slope of 1 (i.e. *ago4 = idn2,idnl1,idnl2*).
D) *ago4* mutant affects CHH methylation more than does *spt5l. nrpe1* DMRs were taken and meCHH levels for *ago4* and *spt5l* relative to Col-0 were calculated and plotted. The gradient from white to dark blue indicates low to high density points. Solid diagonal line indicates a slope of 1 (i.e. *ago4 = spt5l*).
E) Both SPT5L and AGO4 are necessary for RdDM specific nucleosome positioning. Nucleosomes showing at least two fold signal reduction in *nrpe1* (*nrpe1*/Col-0 <= 0.5) were selected. H3 ChIP-seq signal in *spt5l* and *ago4* relative to Col-0 were plotted.
F) DNA methylation and nucleosome positioning are separate functions of RdDM. Nucleosomes reduced in *nrpe1* with *nrpe1* DMRs were taken and CHH methylation changes in *ago4* (*ago4*/Col-0) and *spt5l* (*spt5l*/Col-0) were plotted. The gradient from white to dark blue indicates low to high density points. Solid diagonal line indicates a slope of 1 (i.e. *ago4 = spt5l*).

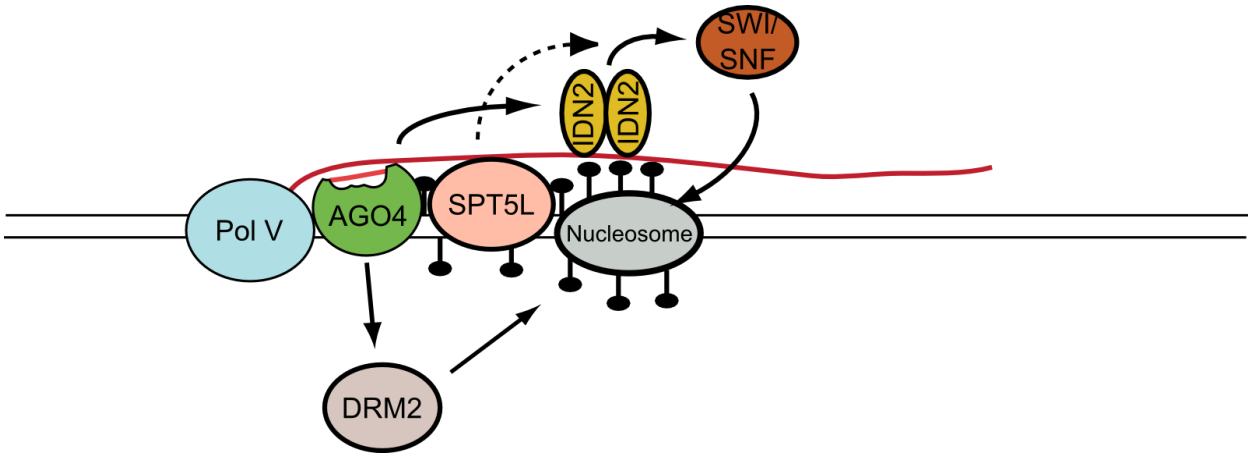


Figure 5.6 Model of SPT5L and AGO4 in RdDM
Pol V (blue) produces lncRNA scaffolds (red line) to which proteins bind. siRNA bound AGO4 (green) and SPT5L (peach) are guided to chromatin by Pol V transcripts where they bind independent of each other. AGO4 directs *de novo* DNA methylation placed by DRM2 (light brown). AGO4 and SPT5L interact with each other and are both important for nucleosome positioning. AGO4 is important for IDN2 (yellow) to bind lncRNA. IDN2 dimerizes and helps SWI/SNF (orange) deposit nucleosomes (grey) at RdDM targets. Mechanism of SPT5L activity in nucleosome positioning is yet to be determined (dashed line).

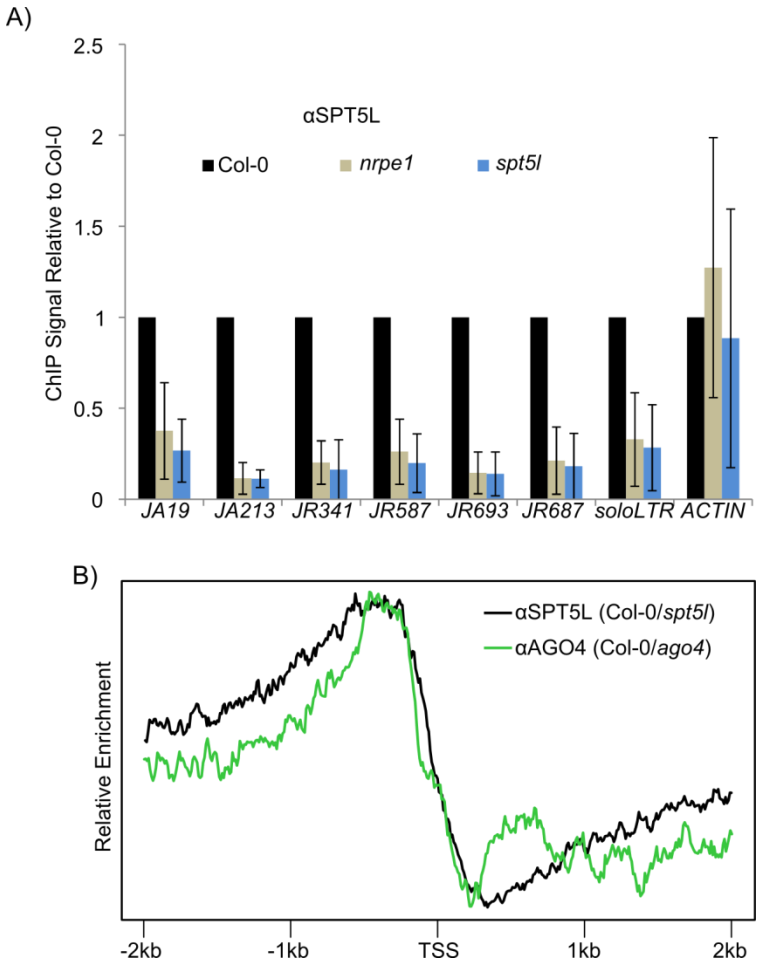


Figure 5.7 SPT5L binding sites reflect RdDM activity (supplementary to Figure 5.1)

A) Locus specific validation of SPT5L ChIP-seq peaks. ChIP-qPCR showing enrichment of SPT5L in Col-0 (black) compared to *nrpe1* (tan) and *spt5l* (blue). Error bars represent standard deviation from three biological replicates.

B) SPT5L and AGO4 bind promoters of protein coding genes. Profile of SPT5L (black) and AGO4 (green) relative to the transcriptional start site (TSS) of protein coding genes. Profiles were scaled to the minimum and maximum value.

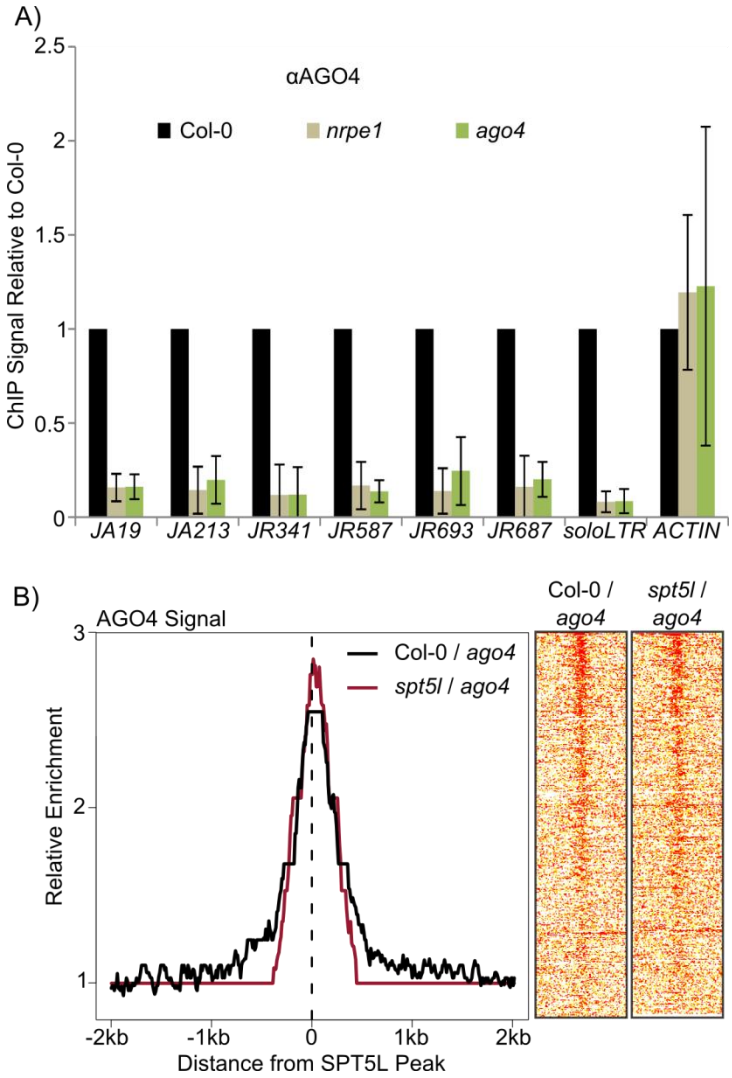


Figure 5.8 AGO4 binding sites reflect RdDM activity (supplementary to Figure 5.2)

A) Locus specific validation of AGO4 ChIP-seq peaks. ChIP-qPCR showing enrichment of AGO4 in Col-0 (black) compared to *nrpe1* (tan) and *ago4* (green). Error bars represent standard deviation from three biological replicates.

B) Profile and heatmap of AGO4 ChIP-seq signal on called SPT5L binding sites. Graph shows median AGO4 ChIP-seq signal ratios in Col-0 *vs ago4* (black) or *spt5l vs ago4* (red) relative to background around SPT5L peaks. Dashed line indicates SPT5L peak summit around which the plot is oriented. Heatmaps display this same region ordered from highest to lowest SPT5L peak score with red indicating strong Col-0/*ago4* or *spt5l*/*ago4* AGO4 ChIP-seq signal ratios and white/yellow indicating low signal ratios. All signal ratios were taken as reads per million (RPM).

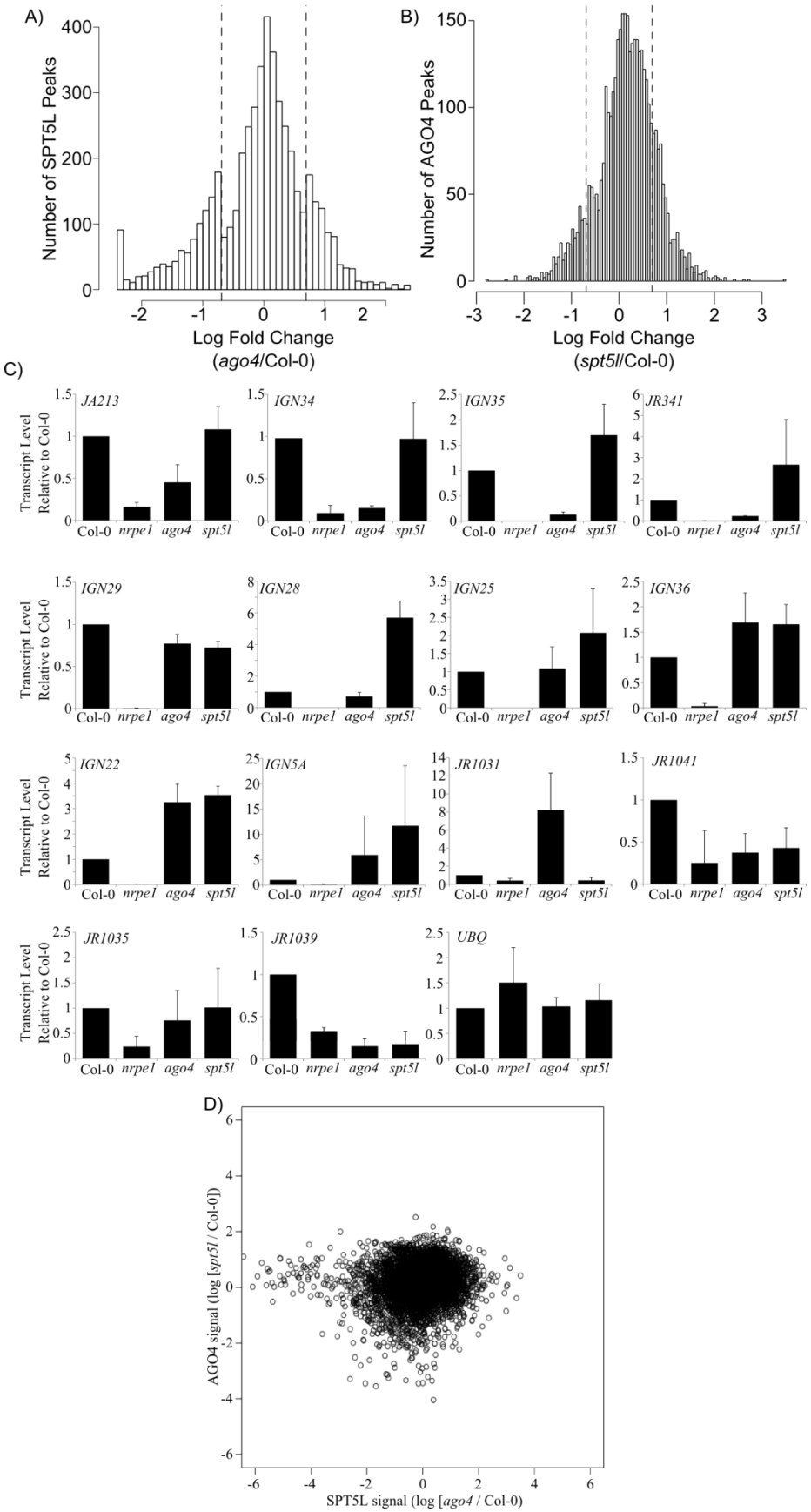

A)
Number of SPT5L Peaks
Log Fold Change
(ago4/Col-0)
B)
Number of AGO4 Peaks
Log Fold Change
(spt5l/Col-0)
C)
Transcript Level
Relative to Col-0
JA213
IGN34
IGN35
JR341
IGN29
IGN28
IGN25
IGN36
IGN22
IGN5A
JR1031
JR1041
JR1035
JR1039
UBQ
Col-0
nrpe1
ago4
spt5l
D)
AGO4 signal (log [spt5l / Col-0])
SPT5L signal (log [ago4 / Col-0)

Figure 5.9 Pol V transcript availability determines AGO4 and SPT5L binding to chromatin (supplementary to Figure 5.3)
A) Histogram of *ago4* effects on SPT5L chromatin binding. The log fold change in *ago4* compared to Col-0 is plotted as a histogram. Vertical dashed lines indicate two-fold changes.
B) Histogram of *spt5l* effects on AGO4 chromatin binding. The log fold change in *spt5l* compared to Col-0 is plotted as a histogram. Vertical dashed lines indicate two-fold changes.
C) RT-PCR of Pol V transcripts shown in figure 5.3CD. Q-PCR signal was normalized to *ACTIN* and plotted relative to Col-0. UBQ is a control locus (not transcribed by Pol V). Error bars indicate standard deviation from three biological replicates.
D) Effects of *spt5l* on AGO4 do not correlate to the effects of *ago4* on SPT5L. Scatterplot showing changes in SPT5L ChIP-seq signal in *ago4* (log[*ago4*/Col-0], x-axis) compare to the changes to AGO4 ChIP-seq signal in *spt5l* (log[*spt5l*/Col-0], y-axis).

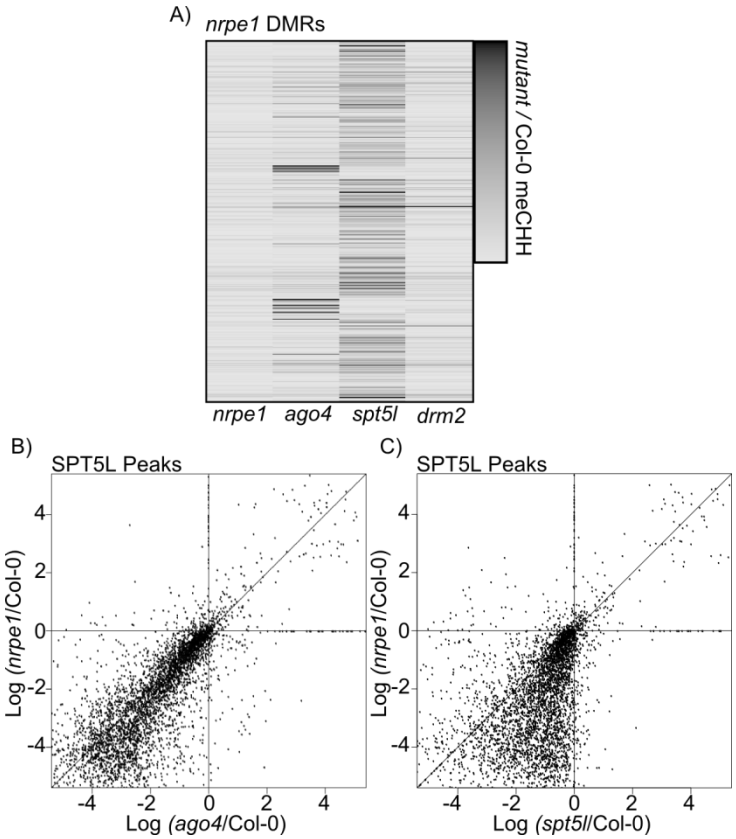


Figure 5.10 SPT5L plays a more limited role than AGO4 in directing DNA methylation (supplementary to Figure 5.4)

A) Methylation reduction in *spt5l* is limited. Heatmap of *nrpe1* DMRs showing CHH methylation changes in *nrpe1*, *ago4*, *spt5l*, and *drm2* relative to Col-0.

B) AGO4 is important for DNA methylation at SPT5L binding sites. Scatterplot showing CHH methylation changes in *ago4* (log[*ago4*/Col-0], x-axis) compared to *nrpe1* (log[*nrpe1*/Col-0], y-axis). Vertical and horizontal lines indicate no change from Col-0. Diagonal line indicates a slope of 1 (i.e. *ago4* = *nrpe1*).

C) SPT5L is semi-important for DNA methylation at SPT5L binding sites. Scatterplot showing CHH methylation changes in *spt5l* (log[*spt5l*/Col-0], x-axis) compared to *nrpe1* (log[*nrpe1*/Col-0], y-axis). Vertical and horizontal lines indicate no change from Col-0. Diagonal line indicates a slope of 1 (i.e. *spt5l* = *nrpe1*).

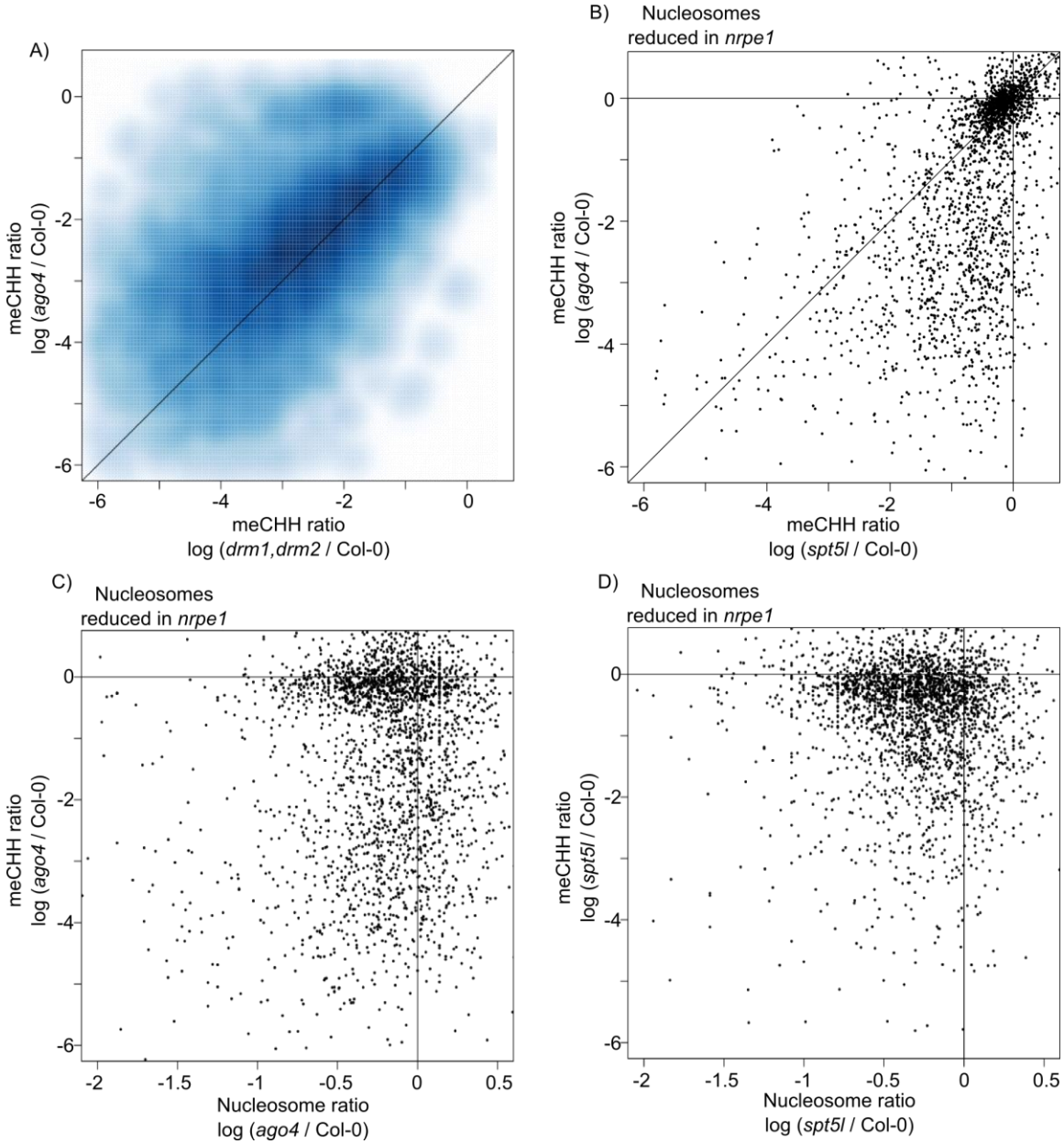


Figure 5.11 Both SPT5L and AGO4 are necessary for RdDM specific nucleosome positioning (supplementary to Figure 5.5)

A) *ago4* mutant affects CHH methylation similar to the *drm1/drm2* double mutant. *nrpe1* DMRs were taken and meCHH levels for *ago4* and *drm1,drm2* relative to Col-0 were calculated and plotted. The gradient from white to dark blue indicates low to high density points. Solid diagonal line indicates a slope of 1 (i.e. *ago4* = *drm1/drm2*).

B) DNA methylation and nucleosome positioning are separate functions of RdDM. Nucleosomes reduced in *nrpe1* were taken and CHH methylation changes in *ago4* (*ago4*/Col-0) and *spt5l* (*spt5l*/Col-0) were plotted. Solid diagonal line indicates a slope of 1 (i.e. *ago4* = *spt5l*).

C) DNA methylation changes and nucleosome positioning in *ago4* are not correlated. Nucleosomes reduced in *nrpe1* were taken and CHH methylation changes in *ago4* (*ago4*/Col-0) were plotted against nucleosome changes in *ago4* (*ago4*/Col-0). Vertical

and horizontal lines indicate no change from Col-0.
D) DNA methylation changes and nucleosome positioning in *spt5l* are not correlated. Nucleosomes reduced in *nrpe1* were taken and CHH methylation changes in *spt5l* (*spt5l*/Col-0) were plotted against nucleosome changes in *spt5l* (*spt5l*/Col-0). Vertical and horizontal lines indicate no change from Col-0.

CHAPTER 6

RNA-directed DNA Methylation Controls Gene Expression via Chromosome Looping

The contents of this chapter will be submitted for publication in the near future. Gudrun Böhmdorfer prepared RNA-seq samples. I performed all other experiments and data analysis shown in this chapter.

Abstract

RNA-mediated transcriptional silencing, in plants known as RNA-directed DNA methylation (RdDM), is a conserved process where small interfering RNA (siRNA) and long non-coding RNA (lncRNA) work together to establish repressive chromatin modifications [1,2]. This process not only represses transposons but also affects expression of protein-coding genes [3,4]. Mechanisms used by RdDM to control transcription remain mostly unknown. In this work we show that RdDM controls chromosome looping between distant genomic regions, specifically at RdDM target loci in *Arabidopsis thaliana*. Presence of chromosome loops is correlated with high levels of gene expression and the absence of repressive chromatin modifications on corresponding distant regulatory regions. Direct RdDM targets have the potential to engage in chromosome looping, which is prevented by the presence of repressive chromatin modifications. In mutants defective in RdDM, looping at RdDM targets is strongly increased. This includes increased looping between genes repressed by RdDM and distant regions, which are direct RdDM targets and are enriched in transcription factor binding sites. This suggests that RdDM may repress looping between genes and potential regulatory regions like enhancers. We propose a model where RdDM contributes to the regulation of gene expression by controlling looping between genes and their distant enhancers.

Significance Statement

RNA-directed DNA methylation (RdDM) is a process, where small interfering RNA (siRNA) and long non-coding RNA (lncRNA) work together to repress transposons. In addition to controlling repetitive sequences, it contributes to the regulation of gene expression. Molecular mechanisms of transcription regulation by RdDM remain mostly unknown. We show that RdDM is involved in repression of long range chromosomal interactions. We propose a model where RdDM contributes to the regulation of gene expression by controlling looping between genes and their distant enhancers.

Introduction

RNA-mediated transcriptional silencing is a process, which directs repressive chromatin modifications to transposons and other repetitive loci. In *Arabidopsis thaliana* it is known as RNA-directed DNA methylation (RdDM) and uses two specialized RNA polymerases, RNA polymerase IV (Pol IV) and RNA polymerase V (Pol V) [5]. While Pol IV is believed to produce siRNA precursors, Pol V produces lncRNA scaffolds which help guide proteins to chromatin [5,6]. ARGONAUTE 4 (AGO4) interacts with both siRNA and lncRNA to target specific loci for silencing [7,8]. At these loci AGO4 and other RNA binding proteins, such as SPT5L and IDN2, are important for chromatin modifiers to bind and/or function [9–12]. Components of RdDM are important for *de novo* DNA methylation, nucleosome positioning, and repressive histone modifications [9,10,12–14].

The RdDM pathway not only represses transposons but also contributes to the regulation of gene expression. We used genome-wide chromosome conformation capture assay (Hi-C) to test the hypothesis that RdDM affects gene expression by controlling long range chromosomal interactions. We found that chromosome looping is less likely to occur at silenced genes and at regions with repressive chromatin marks. Our data also show that RdDM inhibits chromosome loops specifically at RdDM targets. We also show that gene expression may be controlled by RdDM via inhibiting interactions with distant transcription factor binding sites. These findings lead us to propose a model where RdDM affects gene expression by inhibiting long distance chromosomal interactions.

Results

Gene expression is correlated with looping to euchromatin

Components of the RNA-dependent DNA methylation pathway (AGO4, Pol V) have been shown to bind chromatin in putative promoter regions upstream of protein-coding genes [8,15]. To test if AGO4 controls the expression of proximally located genes, we performed RNA-seq in the *ago4* mutant and compared changes in RNA levels to our ChIP-seq data of AGO4 binding to chromatin [8]. Only 11% of genes with expression levels changed in the *ago4* mutant had evidence of AGO4 binding within their proximal regulatory regions defined as 2.5kb upstream of the transcription start site. Although this overlap is significant, the non-overlapping features suggest that the majority of genes differentially expressed in the *ago4* mutant do not have evidence of detectable AGO4 binding within their proximal regulatory regions (Figure 6.1A). One explanation of this disparity may be posttranscriptional or indirect effects. Alternatively, AGO4 may affect transcription over long distances by controlling chromosome looping.

To test this possibility, we performed chromosome conformation capture experiments followed by high throughput sequencing (Hi-C) [16,17]. To facilitate comparative analysis of Hi-C and ChIP-seq datasets, we used the same tissue type, similar crosslinking conditions, and the *Dpn*II restriction enzyme, which is insensitive to DNA methylation and produces a median fragment size of 166 bp in the *Arabidopsis* genome. We first analyzed Hi-C data from Col-0 wild type *Arabidopsis* seedlings and tested if gene expression is correlated with looping. Consistent with reports from other organisms [18–20], genes with lower expression (approximated by RNA-seq signal) are less likely to engage in chromosome looping than more highly expressing genes (Figure 6.1B, Figure 6.5D). This is consistent with a model where looping brings regulatory elements into close proximity to genes thereby activating transcription.

To test if chromatin modifications present at distant regions may affect looping and thereby gene expression, we identified loops that connect genes to other genomic regions and compared the presence of specific histone modifications with gene expression levels. High levels of an active/euchromatin mark (H3K4me2) at distant interacting regions were associated with higher levels of gene expression (Figure 6.1C). Similarly, high levels of an inactive/heterochromatin mark (H3K9me2) at distant

interacting sites were associated with lower levels of gene expression (Figure 6.1D). Although exceptions to this trend clearly exist (Figure 6.1C, 6.1D), it is consistent with a proposed model [21] where chromatin modifications at distant regions affect the formation of loops to genes and/or control the expression of genes distally located. Because RdDM establishes DNA methylation and other repressive chromatin modifications, it is possible that RdDM may also affect chromosome looping.

We then tested if the frequency of looping corresponds to particular chromatin modifications. Presence of activating chromatin modifications (higher H3K4me2 and lower H3K9me2 levels) was associated with higher probability of looping (Figure 6.1E). Conversely, the presence of repressive chromatin modifications (lower H3K4me2 and higher H3K9me2 levels) was associated with lower probability of looping (Figure 6.1E). These results were also supported by a second biological repeat of Hi-C, which was sequenced at lower coverage (Figure 6.5E-H). This is consistent with chromatin modifications being important for chromosome looping [22–24]. Together, these results indicate a correlation between activating chromatin modifications, gene expression, and chromosome looping.

RdDM inhibits chromosome looping at RdDM sites

Negative correlation between repressive chromatin modifications and looping (Figure 6.1E) indicates that RdDM may prevent the formation of chromosome loops. To test this possibility we performed the Hi-C experiment in *nrpe1* and *ago4* mutants, both of which are defective in RdDM (*NRPE1* encodes the largest subunit of RNA Polymerase V, which produces long non-coding RNA required for RdDM) [25–28]. Examining the Hi-C data, we identified chromosome looping at RdDM chromatin targets, defined as regions where NRPE1 and AGO4 bind to chromatin (ChIP-seq peaks) and mediate CHH DNA methylation (loss of methylation in *nrpe1* mutant). We found that in Col-0 wild type plants, RdDM targets were significantly less likely to engage in looping than random genomic regions (Figure 6.2A). In contrast, this repression of looping was alleviated in *nrpe1* and *ago4* mutants (Figure 6.2A). Moreover, loops to *nrpe1* differentially methylated regions (DMRs) were also supported by higher numbers of Hi-C sequencing reads in *nrpe1* and *ago4* mutants compared to Col-0 wild type (Figure 6.2B, Figure 6.6C). Enhanced looping in RdDM mutants was

consistently observed in the second low coverage repeat of Hi-C (Figure 6.6AB) and validated using a locus-specific 3C assay on a few selected loci (Figure 6.2C). This repression of looping can also be seen at non-centromeric transposons, which are generally thought to be targets of RdDM [8] (Figure 6.6D). Additionally, these effects were specific to RdDM targets and most likely not caused by a widespread disruption of chromosome looping since there was no difference in the overall relationship between total gene expression levels in Col-0 wild type and looping between Col-0 wild type and *nrpe1* or *ago4* mutants (Figure 6.5D). These results indicate that RdDM prevents formation of chromosome loops specifically at RdDM targeted chromatin.

Gene repression correlates with repression of chromosome looping by RdDM

Chromatin modifications on distant looping regions are correlated with the frequency of looping (Figure 6.1E) and gene expression levels (Figure 6.1CD). Moreover, site specific chromosome looping is repressed by RdDM (Figure 6.2A). This suggests that RdDM may affect gene expression by repressing gene looping to heterochromatin. To test this possibility we analyzed genes repressed by RdDM. For this analysis we define genes repressed by RdDM as those increased in both *nrpe1* and *ago4* mutants (Figure 6.3A) since genes increased in a single mutant may be due to indirect effects or limitations of RNA-seq; however, overlap of differentially expressed genes between the two mutants is significant (Figure 6.3A) and supports the model of coordinated action by Pol V and AGO4. Having obtained high confidence differentially expressed genes, we tested what fraction of those genes loop to regions bound by AGO4. When we analyzed loops identified in Col-0 wild type, we found that genes upregulated in RdDM mutants were less likely to loop to AGO4 binding sites in the wild type than expected by chance (Figure 6.3B, Figure 6.7). However, when we analyzed loops identified in *nrpe1* and *ago4* mutants, we found that genes upregulated in both RdDM mutants were more likely to loop to AGO4 binding sites in the mutants than expected by chance (Figure 6.3B, Figure 6.7). Similar results were obtained for larger groups of genes upregulated in *nrpe1* or in *ago4* (Figure 6.7). This indicates that genes repressed by RdDM have the potential to loop to regions bound by AGO4 but in Col-0 wild type this looping is prevented.

Changes in gene expression observed in RdDM mutants may be explained by two non-mutually exclusive mechanisms. First, RdDM targets on gene promoters or promoter proximal elements may affect transcription without the involvement of looping. Second, RdDM targets on distant regulatory regions (potential enhancers) may affect gene expression via chromosome looping. For these purposes we define distal regulatory regions (enhancers) as located at least three *Dpn*II restriction sites away from the gene or its proximal regulatory region and connected by detectable chromosome looping from Hi-C. To test if RdDM influences gene expression long-range we examined genes with differential expression in *nrpe1* for AGO4 binding to chromatin on the promoters or distant regulatory regions. AGO4 has evidence of binding in three different scenarios: on the distal regulatory region (Figure 6.3C - Q1), on the proximal region (Figure 6.3C – Q3), or on both (Figure 6.3C – Q2). We also calculated how many genes upregulated in the *ago4* mutant show evidence of AGO4 binding to their proximal regulatory regions, distant regulatory regions, or both. We found evidence of detectable AGO4 binding within only the proximal regulatory regions, within only the distant regulatory regions (enhancers), or within both proximal and distant regulatory regions (Figure 6.3D). This indicates that long and short distance effects of RdDM are likely to coexist. Together these results suggest that RdDM binds regulatory regions either proximally or distally from genes where gene expression is controlled through the inhibition of chromosome looping.

RdDM represses looping between genes and their enhancers

Connecting genes to distant enhancers is thought to be an important function of chromosome looping [21]. To determine if this is the case in *Arabidopsis* we tested whether genomic regions that loop to protein-coding genes are enriched in transcription factor binding. By reanalyzing seven independent genome-wide transcription factor binding datasets from seedlings (see Methods), we found that in general, genes are more likely to loop to transcription factor binding sites than expected by chance (random regions in the genome) (Figure 6.4A). Since the presence of transcription factor binding sites is a hallmark of enhancers [29], this suggests that *Arabidopsis* genes preferentially loop to enhancers.

To examine whether RdDM inhibits looping to enhancers, we further tested if distant regulatory regions which loop to genes upregulated in the *nrpe1* mutant are also enriched in transcription factor binding sites. We found that these distant regions are enriched in the presence of transcription factor binding sites (Figure 6.4B). This suggests that distant regions looping to genes upregulated in the *nrpe1* mutant are possible regulatory elements like enhancers. This is consistent with our hypothesis that RdDM affects looping between genes and their distant regulatory regions / enhancers.

Discussion

Based on presented results we propose a model, where RdDM affects gene expression by repressing looping between genes and their enhancers. According to this model, at least some genes which are repressed by the RdDM pathway have direct RdDM targets on their proximal promoters and/or on their distant enhancers (Figure 6.4C – wild type). In wild type plants, regions directly targeted by RdDM have high levels of DNA methylation and repressive histone modifications. These repressive chromatin marks prevent the formation of chromosome loops between genes and enhancers (Figure 6.4C – wild type). Lack of gene-enhancer looping contributes to gene repression. When RdDM is not active in *nrpe1* or *ago4* mutants or under specific environmental conditions [30], repressive chromatin marks are eliminated. When repressive chromatin modifications are removed or lost, chromosome looping is no longer repressed. Loops between genes and their distant enhancers may form and transcription factors binding to enhancers activate gene expression (Figure 6.4C – No RdDM).

Our model leads to an important question of what feature of RdDM directly affects the formation of chromosome loops. One possibility is that chromatin modifications directly affect chromosomal interactions. Another possibility is that looping is directly mediated by transcription factors, which are unable to bind to chromatin with repressive modifications. In both cases the effects on looping are not expected to be limited to RdDM. Repressive chromatin modifications established using other mechanisms are likely to have similar effects on chromosome looping.

Our results are consistent with a recent report studying higher order organization of *Arabidopsis* chromosomes using chromosome conformation capture, where a correlation in chromatin modifications between both ends of loops has been shown [31]. However, the mechanism we observe appears to be different than looping reported on the *FLC* gene, which was shown not to be affected by ncRNA or chromatin modifications [32] but may be affected by other factors altering chromatin modifications at this locus.

Looping between specific chromosomal regions is a conserved process found by chromosome conformation capture from bacteria to mammals [19,20,33–35] and is believed to contribute to regulation of gene expression [36–38]. Chromosome looping is often mediated by insulator proteins which preferentially bind to open chromatin [19,39]. Activating histone modifications and gene activity have also been shown to be correlated with chromosome looping [19,20]. This is consistent with our model where RdDM prevents looping by establishing repressive chromatin marks. These similarities suggest a conserved mechanism where transcriptional silencing pathways affect gene expression by inhibiting chromosome looping.

Chromosome conformation capture (Hi-C), relies on *in vivo* crosslinking and ligation, in which the effects of biological variability and conditions are not well understood. We attempt to mitigate some of the limitations of Hi-C by performing biological repeats and locus-specific validation. We also perform comparative analysis with ChIP-seq, which involves a similar crosslinking step. Our observations of both correlations and anti-correlations between Hi-C and ChIP-seq datasets (Figures 6.1E and 6.3B) suggest that crosslinking or sequencing biases have a limited impact on our datasets. Also our findings are at least in part based on a comparison of genetic backgrounds. Although currently, there is very limited availability of tools allowing independent verification of Hi-C data over a broad range of chromosomal distances, obtaining such an independent evidence confirming our model is an important goal for future research.

We have recently shown that RdDM mediates nucleosome positioning [12] by the SWI/SNF ATP-dependent chromatin remodeling complex. Together with the results presented here, this indicates that RNA-mediated transcriptional silencing may affect

several aspects of chromatin structure. This further identifies lncRNA as a master regulator of chromatin structure.

Materials and Methods

Plant Material

Arabidopsis thaliana mutant lines *nrpe1 (nrpd1b-11)* and *ago4 (ago4-1* introgressed into Col-0) were described previously [25,40,41]. Seedling tissue was used in all experiments and only datasets from the same tissue type were used in analysis.

3C/Hi-C sample preparation

Arabidopsis seedlings were grown for 2 weeks under long day conditions after which above ground tissue was harvested and cross-linked in 0.5% formaldehyde as previously described [42]. Nuclei were extracted using the same protocol as ChIP [42], with the exception that the final wash was performed in 1.2x *Dpn*II buffer (NEB). Nuclei were then incubated in *Dpn*II reaction buffer supplemented with 0.3% SDS at 65°C for 40 minutes and 37°C for 20 minutes. Triton-X was added to a final concentration of 1.8% and samples were incubated for 1 hour at 37°C after which 600 units of *Dpn*II were added and samples were incubated overnight at 37°C with mixing. To stop digestion, SDS was added to 1.25% and samples were incubated at 65°C for 25 minutes. They were then diluted in 7ml 1x Ligation Buffer (NEB) supplemented with 1% Triton-X and incubated for 1 hour at 37°C. Ligation was performed with 600 units T4 DNA Ligase (NEB) for 4 hours at 16°C followed by 1 hour at 25°C. 600 µg of Proteinase K (Invitrogen) was added and de-crosslinking was performed at 65°C overnight. RNA was eliminated by incubation with 300 µg RNAse A (Invitrogen) at 37°C for 30 minutes. DNA was purified using phenol:chloroform extraction and ethanol precipitation as described[42]. Library preparation and Illumina Paired-End Sequencing were performed by the University of Michigan Sequencing Core. Primers for 3C can be found in Appendix F.

Hi-C data analysis

Each end of paired-end reads with unique alignment was mapped to the *Arabidopsis* genome (TAIR10) using Bowtie and then paired. The genome was divided into 250bp windows and interactions between windows were counted. Interaction

events more than three *Dpn*II sites apart were kept (Figure 6.5A) and plotted as a function of window distance (Figure 6.5B). A polynomial for each curve was generated and multiplied by 0.1 to provide distance to read based cutoffs keeping only the highest confidence short range interactions while maintaining long-range interaction events with a minimum of two reads supporting independent ligation events (Figure 6.5BC). Genome annotations were obtained from the TAIR website (www.arabidopsis.org) and gene proximal elements were defined as 2.5kb upstream of the transcription start site. H3K9me2 and H3K4me2 datasets were obtained from [43,44] (GSE49090 and GSE37644 respectively). AGO4 and Pol V ChIP-seq datasets were obtained from [3,28] (GSE35381 and SRA054962 respectively). Genome-wide bisulfite sequencing datasets were obtained from [45] (GSE39901). *nrpe1* DMRs were identified using a 3 step process: 1. Regions were identified as having more than one methylated base in a CHH context (> 10% methylation score) within 100bp of each other; 2. Regions less than 50bp in length and with less than 10% methylation level per 10 bp were filtered to keep only highly methylated regions; 3. DMRs were then called as having less than 25% methylation in *nrpe1* vs. Col-0.

RNA-seq

RNA from *ago4* seedlings was prepared in three biological repeats as described [12] and libraries were prepared by the University of Michigan Sequencing Core. Reads were mapped to the TAIR10 genome assembly and differential expression was called using the Tophat / Cufflinks suite. Published RNA-seq datasets from seedlings (Col-0 wild type and the *nrpe1* mutant) [12] (GSE38464) were grown, harvested, isolated, and sequenced in parallel to the *ago4* dataset. Overlaps in differential expression were calculated and plotted as a weighted Venn diagram using the Venneuler package in R.

Transcription factor datasets

Raw reads for transcription factor ChIP-seq data was downloaded for AL5 (GSE56706), IBH1 (GSE51120) [46], LFY (GSE24568) [47], PIF4 (GSE35315) [48], PRR7 (GSE49282) [49], SPL7 (GSE45213), and SVP (GSE33120) [50]. Reads were then mapped to the TAIR10 genome and immunoprecipitation vs. control were counted in 250bp windows for comparison to Hi-C. Transcription factor binding was defined as the

presence of at least 10 sequencing reads and signal to background ratio of at least 2 in a 250 bp window.

References

1. Haag, J. R. & Pikaard, C. S. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nat. Rev. Mol. Cell Biol.* **12,** 483–492 (2011).
2. Wierzbicki, A. T. The role of long non-coding RNA in transcriptional gene silencing. *Curr. Opin. Plant Biol.* **15,** 517–522 (2012).
3. Zheng, Q. *et al.* RNA polymerase V targets transcriptional silencing components to promoters of protein-coding genes. *Plant J.* (2012). doi:10.1111/tpj.12034
4. Zhong, X. *et al.* DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. *Nat. Struct. Mol. Biol.* **19,** 870–875 (2012).
5. Haag, J. R. & Pikaard, C. S. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nat. Rev. Mol. Cell Biol.* **12,** 483–492 (2011).
6. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135,** 635–648 (2008).
7. Wierzbicki, A. T., Ream, T. S., Haag, J. R. & Pikaard, C. S. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat. Genet.* **41,** 630–634 (2009).
8. Zheng, Q. *et al.* RNA polymerase V targets transcriptional silencing components to promoters of protein-coding genes. *Plant J.* (2012). doi:10.1111/tpj.12034
9. Rowley, M. J., Avrutsky, M. I., Sifuentes, C. J., Pereira, L. & Wierzbicki, A. T. Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. *PLoS Genet.* **7,** e1002120 (2011).
10. Ausin, I., Mockler, T. C., Chory, J. & Jacobsen, S. E. IDN1 and IDN2 are required for de novo DNA methylation in Arabidopsis thaliana. *Nat. Struct. Mol. Biol.* **16,** 1325–1327 (2009).
11. Bies-Etheve, N. *et al.* RNA-directed DNA methylation requires an AGO4-interacting member of the SPT5 elongation factor family. *EMBO Rep.* **10,** 649–654 (2009).
12. Zhu, Y., Rowley, M. J., Böhmdorfer, G. & Wierzbicki, A. T. A SWI/SNF Chromatin-Remodeling Complex Acts in Noncoding RNA-Mediated Transcriptional Silencing. *Mol. Cell* **49,** 1–12 (2013).
13. Matzke, M., Kanno, T., Daxinger, L., Huettel, B. & Matzke, A. J. M. RNA-mediated chromatin-based silencing in plants. *Curr. Opin. Cell Biol.* **21,** 367–376 (2009).
14. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11,** 204–220 (2010).
15. Zhong, X. *et al.* DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. *Nat. Struct. Mol. Biol.* **19,** 870–875 (2012).
16. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).

17. Belton, J.-M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58,** 268–276 (2012).
18. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30,** 90–98 (2012).
19. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148,** 458–472 (2012).
20. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503,** 290–294 (2013).
21. Harmston, N. & Lenhard, B. Chromatin and epigenetic features of long-range gene regulation. *Nucleic Acids Res.* **41,** 7185–7199 (2013).
22. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503,** 290–294 (2013).
23. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148,** 458–472 (2012).
24. Van Bortle, K. & Corces, V. G. The role of chromatin insulators in nuclear architecture and genome function. *Curr. Opin. Genet. Dev.* **23,** 212–218 (2013).
25. Onodera, Y. *et al.* Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120,** 613–622 (2005).
26. Kanno, T. *et al.* Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nat. Genet.* **37,** 761–765 (2005).
27. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135,** 635–648 (2008).
28. Wierzbicki, A. T. *et al.* Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. *Genes Dev.* **26,** 1825–1836 (2012).
29. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15,** 272–286 (2014).
30. Matzke, M. A. & Mosher, R. A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15,** 394–408 (2014).
31. Grob, S., Schmid, M. W., Luedtke, N. W., Wicker, T. & Grossniklaus, U. Characterization of chromosomal architecture in Arabidopsis by chromosome conformation capture. *Genome Biol.* **14,** R129 (2013).
32. Crevillén, P., Sonmez, C., Wu, Z. & Dean, C. A gene loop containing the floral repressor FLC is disrupted in the early phase of vernalization. *EMBO J.* **32,** 140–148 (2013).
33. Cagliero, C., Grand, R. S., Jones, M. B., Jin, D. J. & O'Sullivan, J. M. Genome conformation capture reveals that the Escherichia coli chromosome is organized by replication and transcription. *Nucleic Acids Res.* **41,** 6058–6071 (2013).
34. Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol. Cell* **48,** 471–484 (2012).
35. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465,** 363–367 (2010).

36. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14,** 288–295 (2013).
37. Harmston, N. & Lenhard, B. Chromatin and epigenetic features of long-range gene regulation. *Nucleic Acids Res.* **41,** 7185–7199 (2013).
38. Van Bortle, K. & Corces, V. G. The role of chromatin insulators in nuclear architecture and genome function. *Curr. Opin. Genet. Dev.* **23,** 212–218 (2013).
39. Van Bortle, K. *et al.* Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. *Genome Res.* **22,** 2176–2187 (2012).
40. Zilberman, D., Cao, X. & Jacobsen, S. E. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299,** 716–719 (2003).
41. Wierzbicki, A. T., Ream, T. S., Haag, J. R. & Pikaard, C. S. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat. Genet.* **41,** 630–634 (2009).
42. Rowley, M. J., Böhmdorfer, G. & Wierzbicki, A. T. Analysis of long non-coding RNAs produced by a specialized RNA polymerase in Arabidopsis thaliana. *Methods* **63,** 160–169 (2013).
43. Greenberg, M. V. C. *et al.* Interplay between active chromatin marks and RNA-directed DNA methylation in Arabidopsis thaliana. *PLoS Genet.* **9,** e1003946 (2013).
44. Moissiard, G. *et al.* MORC family ATPases required for heterochromatin condensation and gene silencing. *Science* **336,** 1448–1451 (2012).
45. Stroud, H., Greenberg, M. V. C., Feng, S., Bernatavichute, Y. V. & Jacobsen, S. E. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. *Cell* **152,** 352–364 (2013).
46. Zhiponova, M. K. *et al.* Helix-loop-helix/basic helix-loop-helix transcription factor network represses cell elongation in Arabidopsis through an apparent incoherent feed-forward loop. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 2824–2829 (2014).
47. Moyroud, E. *et al.* Prediction of regulatory interactions from genome sequences using a biophysical model for the Arabidopsis LEAFY transcription factor. *Plant Cell* **23,** 1293–1306 (2011).
48. Oh, E., Zhu, J.-Y. & Wang, Z.-Y. Interaction between BZR1 and PIF4 integrates brassinosteroid and environmental responses. *Nat. Cell Biol.* **14,** 802–809 (2012).
49. Liu, T., Carlsson, J., Takeuchi, T., Newton, L. & Farré, E. M. Direct regulation of abiotic responses by the Arabidopsis circadian clock component PRR7. *Plant J.* **76,** 101–114 (2013).
50. Gregis, V. *et al.* Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in Arabidopsis. *Genome Biol.* **14,** R56 (2013).

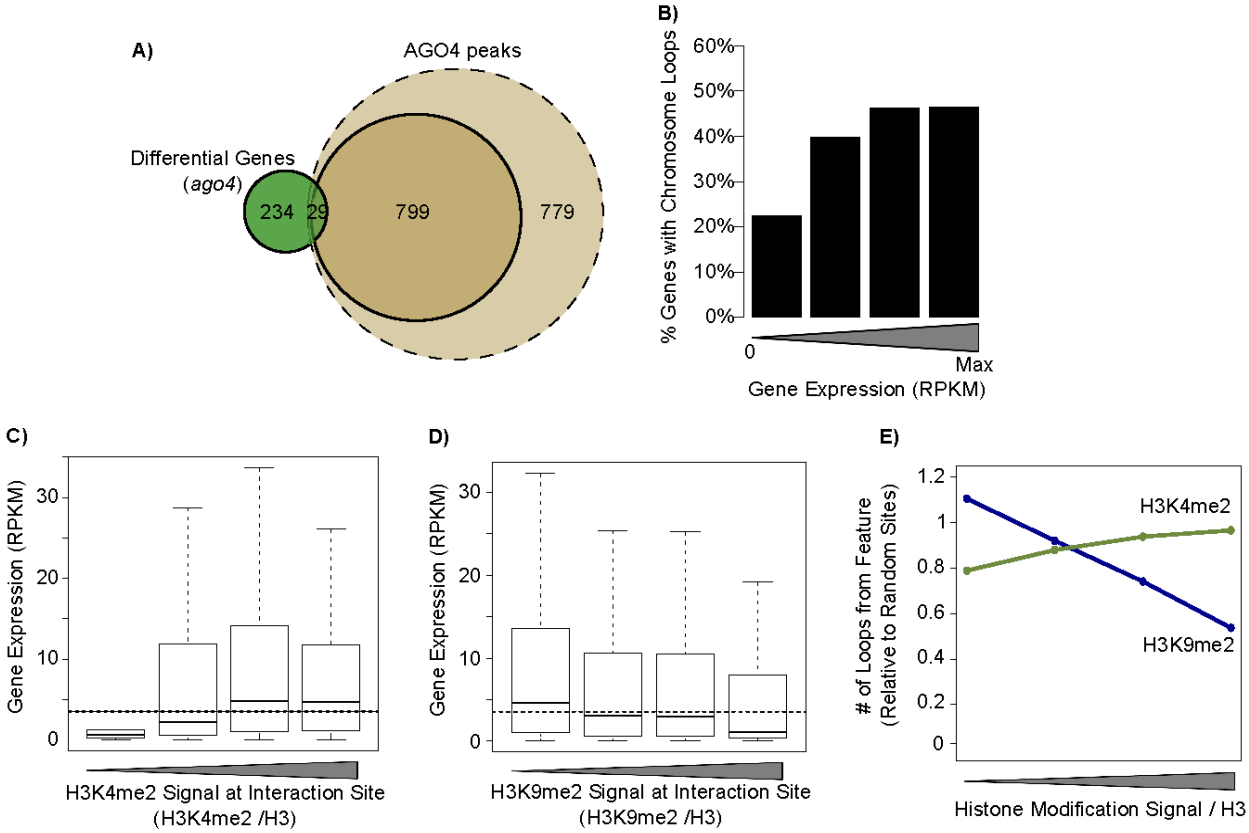


Figure 6.1 Gene expression corresponds to looping to euchromatin

A) Overlap between genes with differential expression in *ago4* mutant (green) and genes with AGO4 peaks defined using high confidence (brown) or lower confidence (light brown) score cutoffs. AGO4 binding to a gene is defined as binding within a region from 2.5kb upstream of the TSS to the 3' end of the transcribed region.

B) Correlation between RNA accumulation and chromosome looping. Genes were divided into inactive (0 RPKM) and a range of categories from low to high expression (RPKM) in Col-0. The number of genes in each expression category that had evidence of long range chromatin interactions in Col-0 were plotted as a percentage of each category. RPKM – reads per kilobase per million.

C) Correlation between gene expression and the level of H3K4me2 at distant regulatory regions. Genes with evidence of looping in Col-0 were divided into categories based on the levels of wild type H3K4me2 ChIP-seq signal normalized to H3 ChIP-seq at distant regulatory regions, and RNA-seq signal from Col-0 was plotted. Dashed horizontal line indicates median gene expression in the genome.

D) Anti-correlation between gene expression and the level of H3K9me2 at distant regulatory regions. Analysis was performed like for (C) except that a H3K9me2 dataset was used.

E) Correlation between histone modifications and the probability of chromosome looping. Genomic regions were divided into categories based on H3K4me2 ChIP-seq signal or H3K9me2 ChIP-seq signal. The numbers of regions in each category with detectable long range chromatin interactions in Col-0 were plotted relative to permutated random genomic regions. ChIP-seq signal for each genomic region was normalized to H3 ChIP-seq.

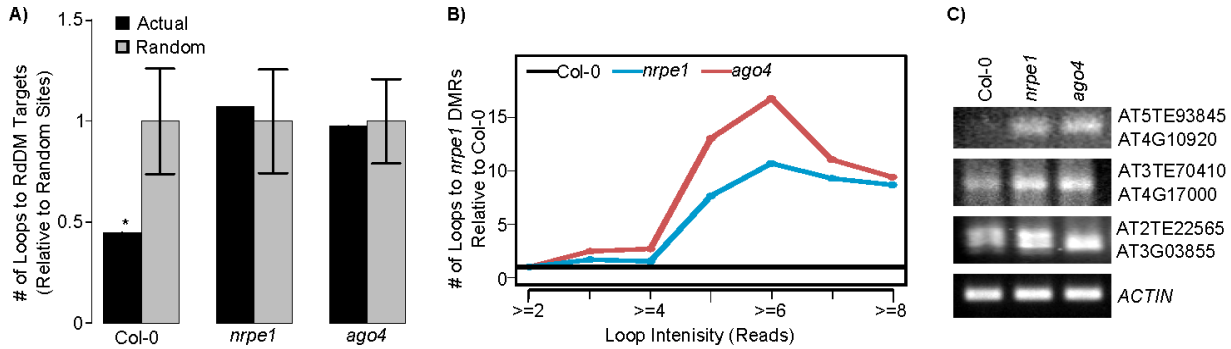


Figure 6.2 RdDM inhibits chromosome looping at RdDM sites

A) Comparison of looping from RdDM target loci in Col-0, *nrpe1* and *ago4*. The numbers of RdDM targets with detectable chromosome looping in Col-0, *nrpe1,* or *ago4* (Actual) relative to the average numbers of permutated random regions (Random). Error bars indicate standard deviations between 1000 permutations, * indicates $p<0.005$.

B) Numbers of sequencing reads supporting chromosome looping to *nrpe1* DMRs. Numbers in each signal intensity cutoff were normalized to those from total identified interactions in each genotype and to Col-0 wild type.

C) Locus-specific validation of chromosome looping from RdDM targets using 3C. Both ends of loops are indicated with the closest annotated features. *ACTIN2* locus serves as a loading control. A representative result from at least two biological replicates is shown.

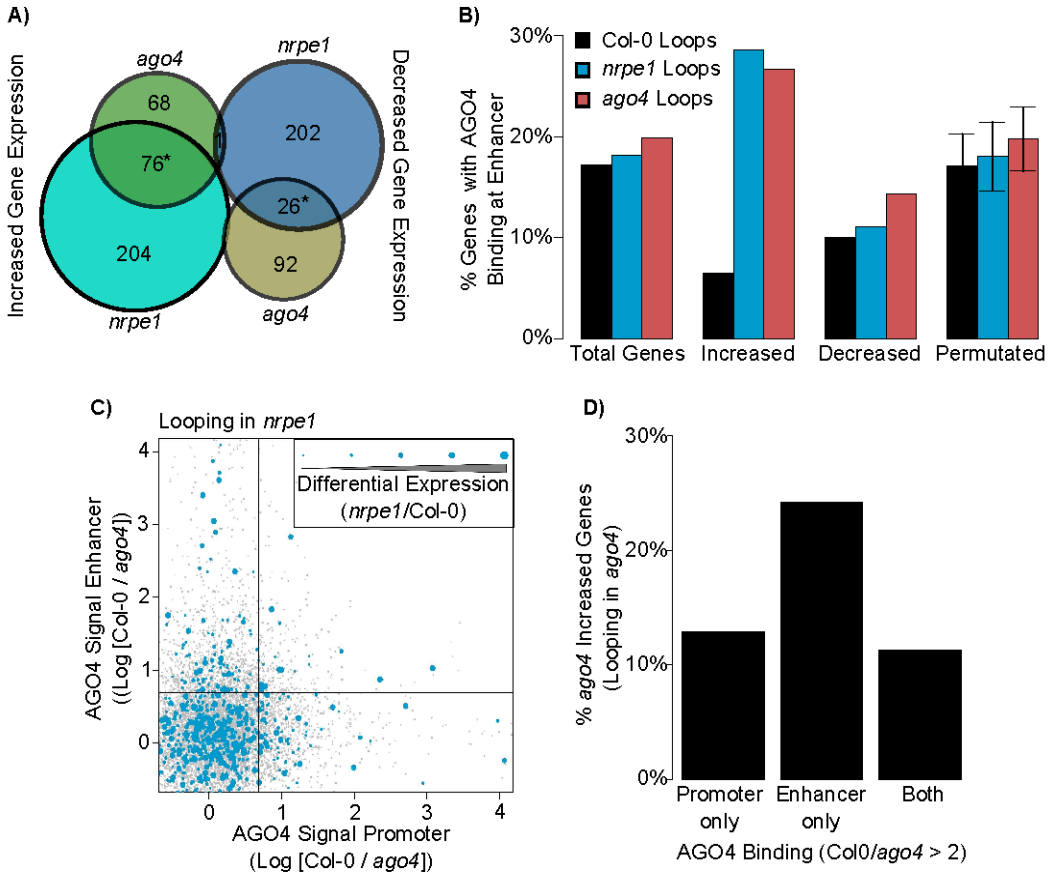


Figure 6.3 Repression of chromosome looping to RdDM targets correlates with gene silencing

A) Overlaps between genes with increases or decreases in RNA accumulation in *nrpe1* and *ago4* mutants. * indicates p <0.05. Overlapping genes are used in Figure 6.3B.

B) Fraction of genes, which form chromosome loops with AGO4-bound direct RdDM targets. Total genes, genes affected in both *ago4* and *nrpe1* (Figure 6.3A) and random permutations of genes are shown. Only genes with detectable looping were included. AGO4 binding is defined by AGO4 ChIP-seq signal (Col-0 > 10 reads and Col-0 / *ago4*> 2) at the distant regulatory region. Error bars indicate standard deviation between 1000 permutations.

C) AGO4 binding intensity to proximal and distant regulatory regions. Genes with increased expression in *nrpe1* (blue) or total genes (grey) were taken and AGO4 signal was counted on the proximal regulatory regions and distant regulatory regions identified by looping detected in the *nrpe1* mutant. Dot sizes represent the intensity of change in gene expression in *nrpe1*. AGO4 binding is presented as log enrichment (Col-0/*ago4*) in 250 bp windows. Lines indicate 2 fold cutoffs.

D) Fractions of genes with RNA levels increased in the *ago4* mutant, which show AGO4 binding to proximal regulatory regions, distant regulatory regions identified in the *ago4* mutant or both. AGO4 binding is defined as Col-0 / *ago4* AGO4 ChIP-seq signal > 2 with more than 10 reads in Col-0 in a 250 bp window.

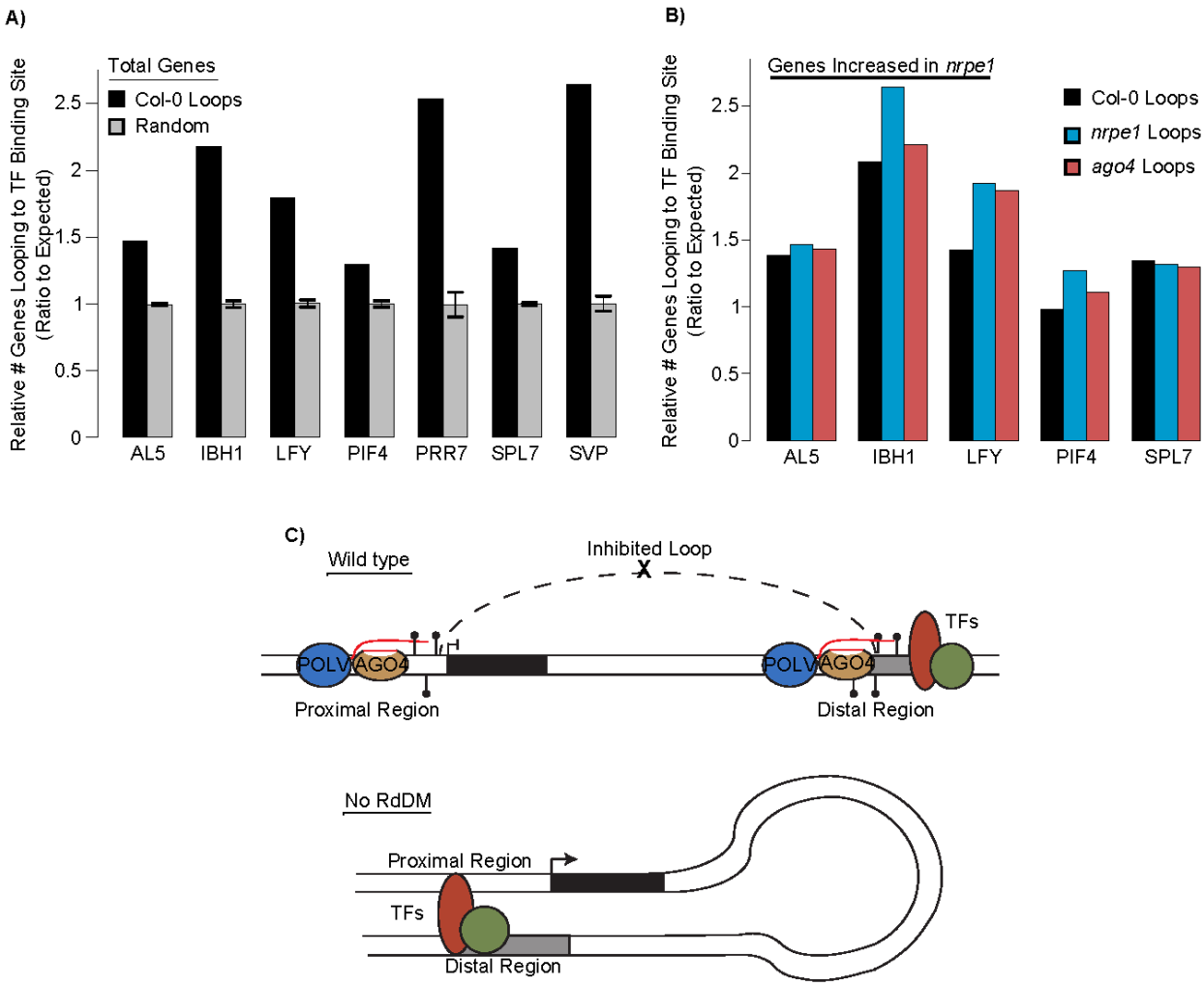


Figure 6.4 RdDM represses looping between genes and their enhancers

A) Transcription factor (TF) binding to distant regulatory regions of total genes. Bars represent numbers of total genes, which create detectable loops in Col-0 to previously published binding sites of transcription factors AL5, IBH1, LFY, PIF4, PRR7, SPL7 and SVP. The total number of 250bp windows with evidence of TF binding was used to calculate the expected numbers. Black bars indicate enrichment of genes which loop to TF binding sites relative to the expected number derived from 1000 permutations of random regions (grey). Error bars indicate standard deviations between permutations.

B) Transcription factor binding to distant regulatory regions of genes upregulated in the *nrpe1* mutant. Values were calculated like on Figure 6.4A except distant regulatory regions were identified based on loops deteted in Col-0, *nrpe1* or *ago4* and only transcription factors with evidence of binding to more than 2000 250bp windows were included.

C) Model of gene regulation by RdDM. In wild type plants (top) Pol V (blue) and AGO4 (yellow) bind to chromatin and help establish repressive chromatin marks at regulatory regions. These chromatin modifications then inhibit looping between genes and distant regulatory regions. When RdDM is absent (bottom), chromatin looping is able to occur between promoters and distant regulatory regions bound by transcription factors (orange and green) and thereby increases gene expression.

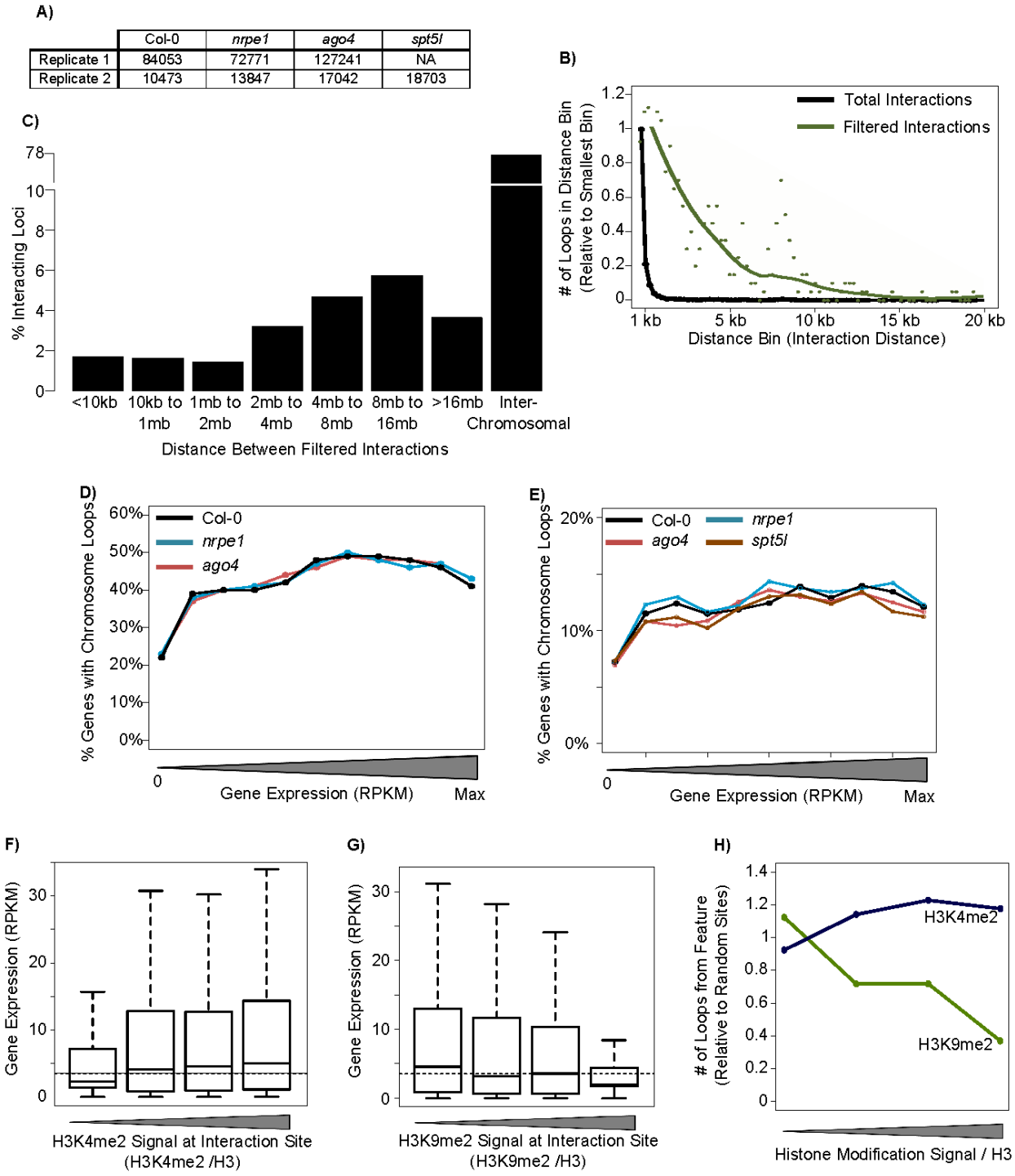


Figure 6.5 Features of the Hi-C datasets and analysis of a second biological repeat (supplementary to Figure 6.1)

A) Numbers of long range ligation events identified in two biological repeats of Hi-C. B) Numbers of identified long range ligation events as a function of the interaction distance. The *Arabidopsis* genome was divided into 250bp bins and numbers of interactions supported by at least two separate reads and at least more than 3 bins apart were plotted as function of the distance between bins (black). Data was then

filtered using polynomial cutoffs and plotted (green).
C) Distances between identified filtered chromosome loops.
D) Correlation between RNA accumulation and chromosome looping. Genes were divided into inactive (0 RPKM) and a range of categories from low to high expression (RPKM) in Col-0 (black), *nrpe1* (blue), and *ago4* (red). The number of genes in each expression category that had evidence of long range chromatin interactions were plotted as a percentage of each category. RPKM – reads per kilobase per million.
E) Correlation between RNA accumulation and chromosome looping in a second biological repeat. Genes were divided into inactive (0 RPKM) and a range of categories from low to high expression (RPKM) in Col-0 (black), *nrpe1* (blue), *ago4* (red), and *spt5l* (brown). The number of genes in each expression category that had evidence of long range chromatin interactions were plotted as a percentage of each category. RPKM – reads per kilobase per million. Second biological repeat corresponding to Figure 6.1D.
F) Correlation between gene expression and the level of H3K4me2 at distant regulatory regions in a second biological repeat. Genes with evidence of looping in Col-0 were divided into categories based on the levels of wild type H3K4me2 ChIP-seq signal normalized to H3 ChIP-seq at distant regulatory regions, and RNA-seq signal from Col-0 was plotted. Dashed horizontal line indicates median gene expression in the genome.
G) Anti-correlation between gene expression and the level of H3K9me2 at distant regulatory regions in a second biological repeat. Analysis was performed like for (C) except that a H3K9me2 dataset was used.
H) Correlation between histone modifications and the probability of chromosome looping in a second biological repeat. Genomic regions were divided into categories based on H3K4me2 ChIP-seq signal or H3K9me2 ChIP-seq signal. The numbers of regions in each category with detectable long range chromatin interactions in Col-0 were plotted relative to permutated random genomic regions. ChIP-seq signal for each genomic region was normalized to H3 ChIP-seq.

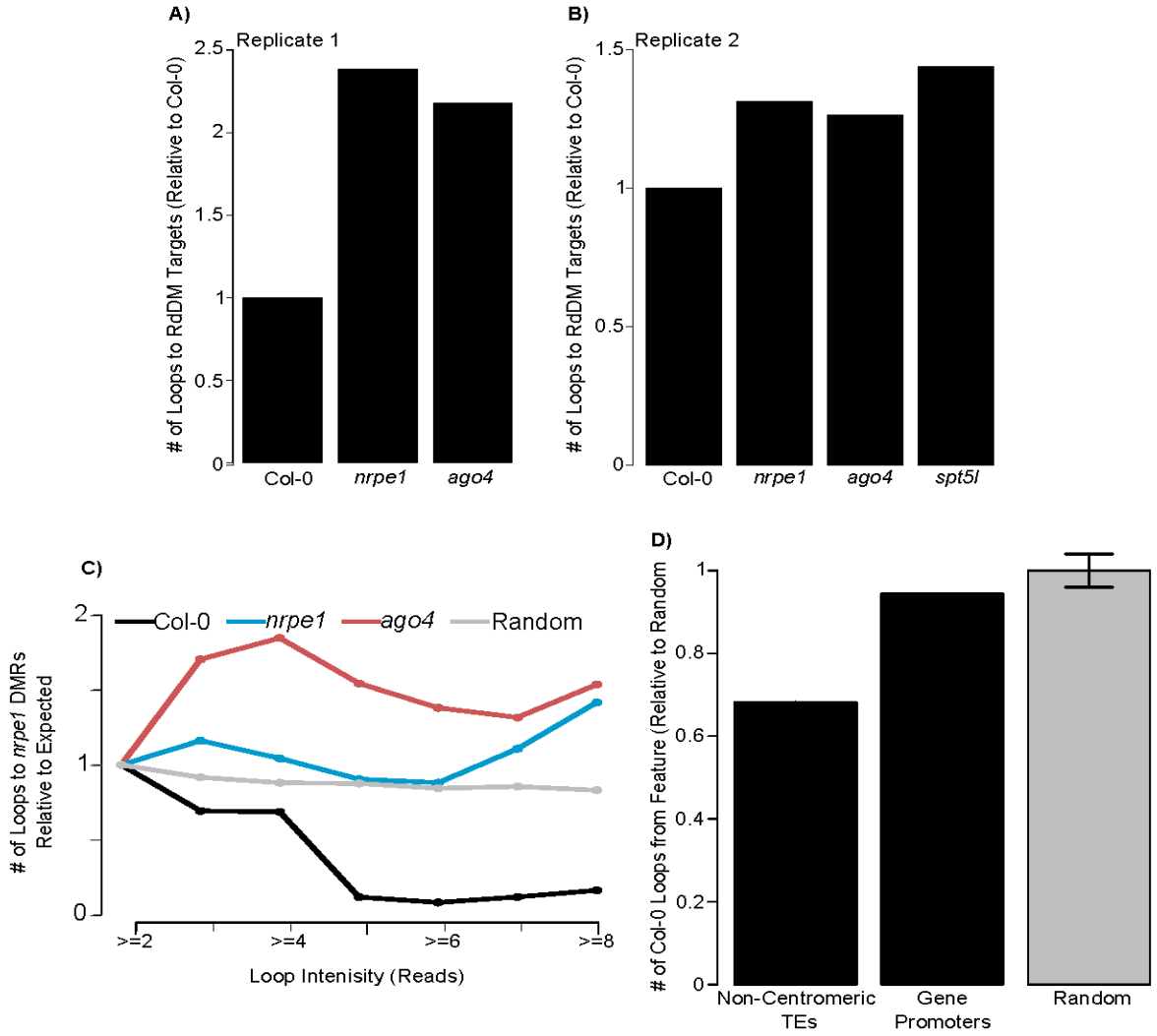


Figure 6.6 RdDM inhibits chromosome looping at RdDM sites (supplementary to Figure 6.2)

A) Comparison of chromosome looping from RdDM target loci in Col-0, *nrpe1* and *ago4* in biological repeat 1. Bars show numbers of RdDM targets with detectable chromosome looping in Col-0, *nrpe1,* or *ago4* relative to the wild type level.

B) Comparison of chromosome looping from RdDM target loci in Col-0, *nrpe1* and *ago4* in biological repeat 2. Bars show numbers of RdDM targets with detectable chromosome looping in Col-0, *nrpe1,* or *ago4* relative to the wild type level.

C) Numbers of sequencing reads supporting chromosome looping to *nrpe1* DMRs. Numbers in each signal intensity cutoff were normalized to those from total identified interactions in each genotype. Loop signal intensity in Col-0, *nrpe1*, *ago4* and permutated genomic regions (Random) were plotted as a ratio to values expected over the entire genome.

D) Depletion in chromosome looping to non-centromeric transposable elements (TEs) in Col-0 wild type. Chromatin interactions were counted if one side overlapped non-centromeric TEs or gene promoters. Each category was normalized by the ability to map sequencing reads to those regions and calculated as a ratio to permutated random regions. Error bars indicate standard deviation of permutations.

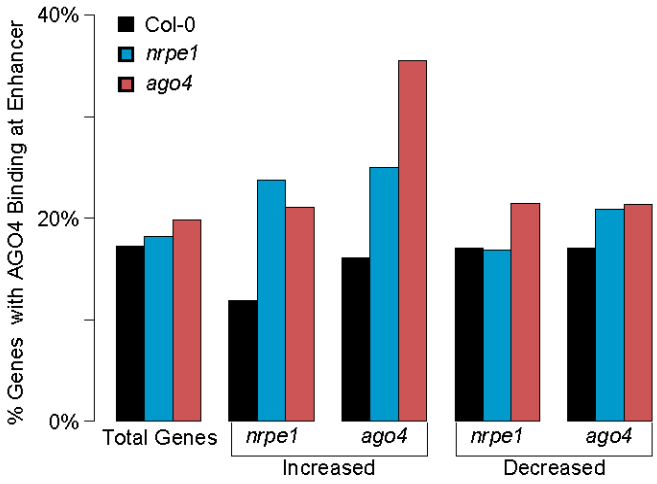


Figure 6.7 Repression of genes corresponds to RdDM repressed looping (supplementary to Figure 6.3)
Fraction of genes, which form chromosome loops with AGO4-bound direct RdDM targets. Total genes and genes increased or decreased in *ago4* or *nrpe1* are shown. Only genes with detectable looping were included. AGO4 binding is defined by AGO4 ChIP-seq signal (Col-0 > 10 reads and Col-0 / *ago4* > 2) at the distant regulatory region.

CHAPTER 7
Conclusion

Introduction

Long non-coding RNA (lncRNA) is a class of molecules where many significant functions are known, but actual mechanisms involving lncRNA are currently poorly understood. In recent years, as more functions of lncRNA are found, this topic has garnered a lot of attention (Figure 7.1). However, despite the increased attention much of the current research focuses on the downstream effects of lncRNA and leaves molecular mechanisms relatively unexplored. As research connected to lncRNA increases, understanding mechanisms involving these molecules becomes ever more pertinent. One important role of lncRNA is to direct chromatin modifications to control genomic elements such as potentially harmful transposons[1,2]. The aim of my research has been to understand long non-coding RNA specifically pertaining to its function in chromatin modification and transcriptional control.

One reason that progress in understanding mechanisms of lncRNA may be slow is the difficulty of genetic studies on this topic. In many organisms, the protein responsible for creating lncRNA is also necessary to create messenger RNA (mRNA); thus knockout of this protein results in many deleterious effects. In many cases, mutations in this protein are lethal to the organism[3]. In contrast, in *Arabidopsis thaliana* a specialized RNA polymerase is responsible for creating lncRNA, allowing generation of viable knockouts[3]. These knockouts make this system especially advantageous for studying molecular mechanisms involving lncRNA.

Studies of lncRNA show a function in directing chromatin modification[2,4]. By acting as platforms (i.e. scaffolds) for RNA binding proteins, lncRNA helps direct the activity of chromatin modifying enzymes[2,4]. At the beginning of the work presented in this dissertation, only a general framework of this process was known. It was shown that AGO4 binds siRNA and lncRNA produced by Pol V and that these components are

necessary for directing the *de novo* methyltransferase DRM2, giving this pathway the name RNA-directed DNA Methylation (RdDM)[1]. SPT5L was also shown to be important in RdDM, but the exact role of this protein was unknown[5]. Another protein, IDN2, was known to be involved in RdDM, but again the exact role was unknown[6]. Additionally, although RdDM was shown to be able to silence transposons, only a handful of targets was discovered[7]. The goal of my research was to understand the various roles of proteins involved in RdDM and to explore other functions of this pathway.

Findings

RdDM, also known as Transcripitonal Gene Silencing (TGS), functions in transposon silencing[1,4]; however, most transposons are located near centromeres where chromatin modifications are maintained independent of RdDM[1,8]. To investigate this contradiction, we examined AGO4 binding sites genome-wide (Chapter 2)[8]. AGO4 binds to chromosome arms rather than centromeric regions suggesting that RdDM functions outside of centromeric regions. Interestingly, AGO4 targets chromatin directly upstream of protein coding genes at what are thought to be regulatory regions (i.e. promoters). Sites bound by AGO4 are enriched for *de novo* DNA methylation. Placement of modifications like DNA methylation may alter the accessibility of chromatin for transcription factors and could potentially affect gene expression. Indeed, after testing several genes that displayed AGO4 binding signal in the promoter, I found gene expression changes in the *ago4* mutant. It is a novel finding that RdDM not only targets transposons in gene dense regions, but that it is involved in gene expression control. Many of these genes are part of stress response signaling pathways which suggests a function of RdDM in this process. In fact, components of RdDM are shown to be important for responding to stressful conditions[9,10]. Overall, I propose that control of these genes occurs by blockage of transcription factor binding, likely through chromatin modification.

This work also discovered a novel function of RdDM in altering chromatin structure. Screening for proteins that interact with components of RdDM identified an enzyme known to work in nucleosome positioning[11]. By looking at genome-wide nucleosome occupancy, I found nucleosomes whose occupancy is reduced in the *pol V*

mutant (*nrpe1*) (Chapter 3)[11]. We were then able to show that Pol V transcripts are bound by IDN2 which is important for placing or stabilizing nucleosomes at RdDM targets. These findings clarified IDN2's role and represents a novel function of RdDM; this pathway not only directs DNA methylation and histone modification, but also nucleosome positioning. Increasing nucleosome density could be a way in which the cell further enhances silencing so that chromatin accessibility to transcription factors is even more limited upon nucleosome placement.

Somewhat like the discovery of IDN2, another protein, SPT5L, was shown to be involved in RdDM, but its place in the pathway was unknown[5,6]. By examining chromatin binding of SPT5L in various mutants I found that SPT5L binding to chromatin is dependent on Pol V and that SPT5L and AGO4 binding to chromatin is independent of each other (Chapter 4)[12]. Previously SPT5L was proposed to be an intermediate between AGO4 and DRM2[5], but my findings indicate coordinate action of SPT5L and AGO4. I also showed that SPT5L differs from AGO4 in that SPT5L binds chromatin independent of siRNA. Additionally, I found that the *spt5l* mutant was necessary for DNA methylation only at some loci, which could indicate that the primary role of SPT5L lies elsewhere.

To further investigate the role of SPT5L in chromatin modification, I examined SPT5L binding sites genome-wide (Chapter 5). Like AGO4, SPT5L targets chromatin directly upstream of protein coding genes. I confirmed that SPT5L and AGO4 bind chromatin independent of each other, but also found exceptions to this rule at several sites. These loci correlate with feedback to Pol V transcripts, meaning that when Pol V transcription is altered AGO4 and SPT5L chromatin binding is affected. Interestingly, while AGO4 is necessary for Pol V dependent DNA methylation, defects in the *spt5l* mutant are much less. This minor role in DNA methylation is not due to limited SPT5L binding since SPT5L is found at most methylation sites dependent on Pol V. In other words, even though SPT5L is bound, methylation does not depend on this protein. This could suggest that SPT5L plays a different role than AGO4 in RdDM. When comparing the effects of various mutants on DNA methylation, *spt5l* is very similar to *idn2*. Correlation in the methylation pattern may indicate that these two proteins work together, or similarly in this pathway. Due to the finding that IDN2 is involved in

nucleosome positioning, I examined genome-wide nucleosome occupancy in *spt5l* and *ago4*. I found that Pol V dependent nucleosomes are dependent on both AGO4 and SPT5L. This clarifies the role of SPT5L in RdDM; SPT5L is more important for nucleosome positioning than for directing *de novo* DNA methylation.

It is evident that the RdDM pathway alters chromatin in a variety of ways: DNA methylation, histone modification, and nucleosome positioning[4]. Targeting of these modifications to regulatory regions occurs in this pathway which can control gene expression. Since regulatory regions often take part in long range enhancer-promoter interactions[13], I investigated whether RdDM is involved in controlling chromatin looping. Using genome wide maps of chromatin interactions (HiC)[14], I uncovered features of chromatin looping which are conserved between Arabidopsis and other eukaryotes[15,16] (Chapter 6). Simply put, genes with active expression are more likely to participate in long-distance interactions than inactive genes. Additionally, sites that have active histone modifications are more likely to loop than sites with repressive modifications. This is particularly intriguing as RdDM helps direct repressive chromatin modifications. Indeed, RdDM sites in wild-type conditions mostly do not engage in chromatin looping; however, knockout of *pol v* or *ago4* caused looping at these sites to increase. This indicates that RdDM modifies chromatin at regulatory regions so that enhancer-promoter interactions do not occur. When examining genes whose expression is controlled by RdDM I saw inhibition of looping in wild-type and increases in looping in the *pol v* and *ago4* mutants. Furthermore, these inhibited sites correspond to AGO4 binding sites indicating that AGO4 binds repressed chromatin interaction sites. Taken together these data indicate that RdDM can control gene expression long-distance by inhibiting enhancer-promoter interactions.

Implications

The finding that non-coding RNA is abundant in the nucleus in many eukaryotes[19–21] suggests that these molecules have important functions. One of these functions is to prevent transposons from causing mutation in the genome[22]. In addition, this work found that long non-coding RNAs in Transcriptional Gene Silencing (TGS) have the ability to control gene expression[8]. It is interesting that many of those genes

are linked to stress response pathways[8]. Having an overarching mechanism for responding to environmental stress as well as controlling harmful mutation may allow organisms to try to survive harsh conditions, but could also create advantageous mutations which help the species as a whole adapt and survive. Indeed transposons are thought to drive evolution by reshuffling genomic sequences[22,23]. Thus TGS may reign in mutation when it is not needed, but could release transposon control upon environmentally stressful conditions and thus increase mutagenic potential.

In addition to transposon control, investigating transcribed non-coding regions of the genome can increase our understanding of the coding regions of the genome. RNA Pol V and Pol II recently diverged in Arabidopsis and, though functionally separate, share several subunits[3]. Similarities between these RNA polymerases may allow insights from *pol v* knockouts to be applied to Pol II. Aside from transcription, SPT5L, involved in TGS / RdDM, is similar to the Pol II transcription factor, SPT5[5]. Like that of Pol V, details of SPT5L's role in RdDM may provide insights into SPT5.

Although investigation of gene expression is important, understanding mechanisms of RdDM is important in its own right. This has particular implications in early human development when *de novo* chromatin modifications are established[1,24]. Additionally, chromatin modifications are important features to consider for the generation and study of stem cells and investigating components of RdDM may improve studies in these fields[25].

In addition to these wider implications, the finding that SPT5L plays less of a role in DNA methylation, but more in nucleosome positioning further resolves the model of RdDM. The role of SPT5L in RdDM has been in question for years since its discovery; its position and function in the pathway have been mysterious[5]. By answering these questions, this work has filled in missing pieces and led to a greater understanding of the RdDM mechanism.

Perhaps even more intriguing is the role of RdDM in influencing chromatin looping. These findings indicate that chromatin modifications control gene expression long-range by altering enhancer-promoter interactions. This could suggest a link between lncRNA and insulator proteins. This notion is especially enticing in light of the finding in many organisms that even active enhancers often have lncRNA[26]. The role of

lncRNA in chromatin looping is a novel field of study and may provide dramatic insights into three dimensional genome organization and gene expression control.

Limitations

This research is meant to be applicable to a wide range of fields studying gene expression and/or chromatin modification. The Transcriptional Gene Silencing (TGS) pathway has many elements conserved across a broad range of species[1,27]. Certainly, transposon silencing is important for nearly all eukaryotic organisms[27,28]. However, the low transposon level in *Arabidopsis thaliana* is used to argue that this plant has been more successful than most at limiting transposition[28], which could mean that differences in TGS exist between Arabidopsis and other organisms. This is fully acknowledged and is seen by the separation of lncRNA production from mRNA production by Pol V and Pol II respectively in Arabidopsis[3]. Although differences may occur, Arabidopsis is a leading model organism in this field and knowledge from plants has been essential for insights into transposon and chromatin control in humans[1,4]. Despite the differences, general principles discovered in plants have greatly advanced our understanding of non-coding RNA in humans[1,27].

Much of the research presented here utilizes genome-wide studies to explore transcriptional silencing mechanisms. While this provides overall pictures of what is happening in the genome, locus variability can, and likely does occur. These individual loci may be functionally significant and merit further study; however, by generating these publicly available genome-wide maps, future studies of individual loci are greatly facilitated. These data should be used to illustrate genomic trends, but also to examine individual sites of interest.

Locus variability should especially be considered due to the finding that SPT5L binding and the effects to chromatin can vary significantly. While this variability may be explained by feedback between RdDM components, the cause of this feedback at some loci and not at others merits further investigation. These locus specific effects of SPT5L could explain why some RdDM targets have variations in their downstream effects on histone modification and gene expression[12,29].

In proposed models of TGS / RdDM, gene expression is controlled by inhibition of transcription factors or Pol II binding to chromatin[8]. It is likely that this occurs because of the chromatin modifications placed at RdDM targets, but direct inhibition by components of RdDM is possible. Likewise, control of chromatin looping may be through chromatin modification or through direct interference by RdDM components. Alternately, unknown proteins could be targeted to these loci dependent on RdDM and control expression. In essence, this work resolves much of the mechanism and function of RdDM in gene expression control, but we are only beginning to understand the connection between these two processes.

Future Directions

This work represents the first to implicate RdDM in chromatin looping and is one of the first to study the three dimensional genome organization of plants. Much more about long distance gene expression control is known in other organisms[15,16], thus studying the role of TGS / RdDM in these organisms may provide further insights into long-distance gene expression control. In other systems, lncRNA has been found at active enhancers which engage in chromatin looping important for gene expression[26,30] which suggests that lncRNA may have a larger role in genome organization than just TGS. Examining the role of these lncRNAs in chromatin looping and their relationship to insulator proteins may yield novel insights into gene expression control.

Identifying enhancer RNAs in Arabidopsis is also interesting and would allow a more direct comparison to lncRNAs involved in transcriptional silencing. Comparing these two functions could indicate what causes lncRNA to be either activating or silencing[2,26]. Discovery of ncRNAs involved in gene expression control could help identify Arabidopsis regulatory regions, which have typically been difficult to define. Once markers of regulatory regions are defined, manipulation of gene expression will be easier and locus specific studies can be performed in the correct context.

Although the data indicate involvement of RdDM in controlling chromatin looping at enhancers, very little is known of what specifically drives genome organization. In many eukaryotes, regions of the genome cluster to form topologically associated domains (TADs)[15,31]; however, TADs are not found in Arabidopsis or yeast. On the other

hand, these organisms do engage in long distance enhancer-promoter interactions (Chapter 6)[32,33]. TADs and other chromatin loops are thought to be mediated by insulator proteins[34], yet insulator proteins in Arabidopsis and yeast have not been identified. Identifying these proteins and resolving their relationship to RdDM or lncRNA would greatly advance this field.

Some insulator proteins in other systems have been found to avoid or rearrange nucleosomes[35]. It is suggested that nucleosome positioning is an important feature to consider in regards to chromatin looping[36]. Since two functions of RdDM identified in this work are nucleosome positioning and inhibition of chromatin looping, a connection between the two may exist. Future work should investigate nucleosome occupancy in regards to chromatin looping.

Concluding Remarks

From this work several genome-wide data sets have been made publicly available, including gene expression data in several mutants, nucleosome occupancy maps, chromatin binding locations of SPT5L and AGO4, as well as some of the first long distance chromatin interaction maps in Arabidopsis (see Appendix G, Table 7.1). While these data have been used to gain insights in TGS / RdDM, the scientific community can benefit from them in the future.

The aim of this work has been to understand the mechanism and functions of RdDM in transcriptional control. The main findings support a model where coordinate binding of SPT5L and AGO4 direct nucleosome positioning at transposons and regulatory regions. This pathway then inhibits regulatory regions from chromatin looping in order to control gene expression. These novel functions represent ways in which lncRNA can act in the genome and have changed the way we think of Transcriptional Gene Silencing.

References

1. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11,** 204–220 (2010).
2. Wierzbicki, A. T. The role of long non-coding RNA in transcriptional gene silencing. *Curr. Opin. Plant Biol.* **15,** 517–522 (2012).

3. Haag, J. R. & Pikaard, C. S. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nat. Rev. Mol. Cell Biol.* **12,** 483–492 (2011).
4. Matzke, M. A. & Mosher, R. A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15,** 394–408 (2014).
5. He, X.-J. *et al.* An effector of RNA-directed DNA methylation in arabidopsis is an ARGONAUTE 4- and RNA-binding protein. *Cell* **137,** 498–508 (2009).
6. Ausin, I., Mockler, T. C., Chory, J. & Jacobsen, S. E. IDN1 and IDN2 are required for de novo DNA methylation in Arabidopsis thaliana. *Nat. Struct. Mol. Biol.* **16,** 1325–1327 (2009).
7. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135,** 635–648 (2008).
8. Zheng, Q. *et al.* RNA polymerase V targets transcriptional silencing components to promoters of protein-coding genes. *Plant J.* (2012). doi:10.1111/tpj.12034
9. Popova, O. V., Dinh, H. Q., Aufsatz, W. & Jonak, C. The RdDM pathway is required for basal heat tolerance in Arabidopsis. *Mol Plant* **6,** 396–410 (2013).
10. Wei, W. *et al.* A role for small RNAs in DNA double-strand break repair. *Cell* **149,** 101–112 (2012).
11. Zhu, Y., Rowley, M. J., Böhmdorfer, G. & Wierzbicki, A. T. A SWI/SNF Chromatin-Remodeling Complex Acts in Noncoding RNA-Mediated Transcriptional Silencing. *Mol. Cell* **49,** 1–12 (2013).
12. Rowley, M. J., Avrutsky, M. I., Sifuentes, C. J., Pereira, L. & Wierzbicki, A. T. Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. *PLoS Genet.* **7,** e1002120 (2011).
13. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14,** 288–295 (2013).
14. Belton, J.-M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58,** 268–276 (2012).
15. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148,** 458–472 (2012).
16. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489,** 109–113 (2012).
17. Singer, S. D., Liu, Z. & Cox, K. D. Minimizing the unpredictability of transgene expression in plants: the role of genetic insulators. *Plant Cell Rep.* **31,** 13–25 (2012).
18. Meyer, P. Transgenes and their contributions to epigenetic research. *The International Journal of Developmental Biology* **57,** 509–515 (2013).
19. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell* **44,** 667–678 (2011).
20. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
21. Geisler, S. & Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.* **14,** 699–712 (2013).
22. Fedoroff, N. V. Transposable Elements, Epigenetics, and Genome Evolution. *Science* **338,** 758–767 (2012).

23. O'Donnell, K. A. & Burns, K. H. Mobilizing diversity: transposable element insertions in genetic variation and disease. *Mobile DNA* **1,** 21 (2010).
24. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14,** 204–220 (2013).
25. Dyachenko, O. V., Schevchuk, T. V., Kretzner, L., Buryanov, Y. I. & Smith, S. S. Human non-CG methylation: are human stem cells plant-like? *Epigenetics* **5,** 569–572 (2010).
26. Ørom, U. A. & Shiekhattar, R. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell* **154,** 1190–1193 (2013).
27. Girard, A. & Hannon, G. J. Conserved themes in small-RNA-mediated transposon control. *Trends Cell Biol.* **18,** 136–148 (2008).
28. Huang, C. R. L., Burns, K. H. & Boeke, J. D. Active Transposition in Genomes. *Annual Review of Genetics* **46,** 651–675 (2012).
29. Böhmdorfer, G. *et al.* RNA-directed DNA methylation requires stepwise binding of silencing factors to long non-coding RNA. *Plant J.* **79,** 181–191 (2014).
30. Lai, F. *et al.* Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* **494,** 497–501 (2013).
31. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).
32. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465,** 363–367 (2010).
33. Grob, S., Schmid, M. W., Luedtke, N. W., Wicker, T. & Grossniklaus, U. Characterization of chromosomal architecture in Arabidopsis by chromosome conformation capture. *Genome Biol.* **14,** R129 (2013).
34. Phillips-Cremins, J. E. & Corces, V. G. Chromatin insulators: linking genome organization to cellular function. *Mol. Cell* **50,** 461–474 (2013).
35. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4,** e1000138 (2008).
36. Li, M., Belozerov, V. E. & Cai, H. N. Modulation of chromatin boundary activities by nucleosome-remodeling activities in Drosophila melanogaster. *Mol. Cell. Biol.* **30,** 1067–1076 (2010).

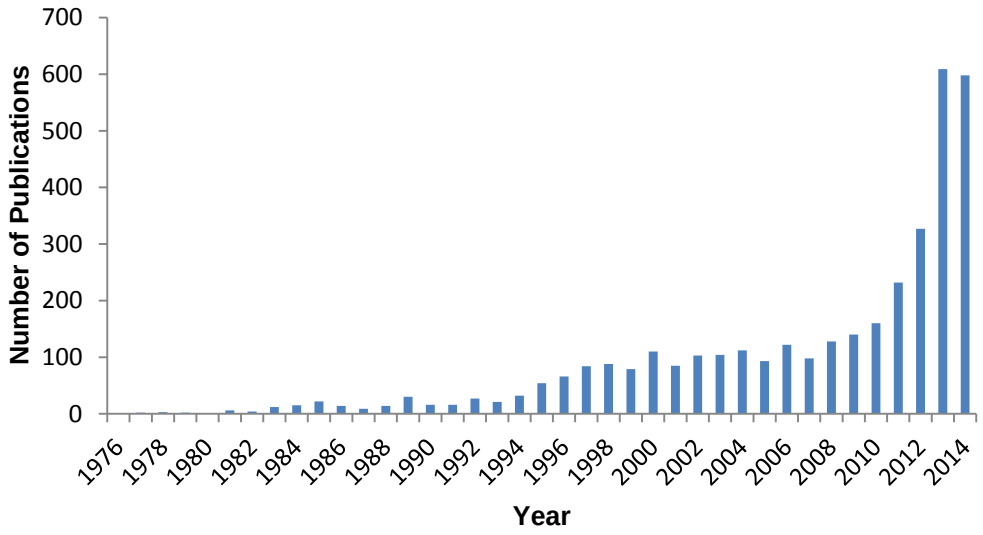


Figure 7.1 Publications about long non-coding RNA by year
As the abundance of long non-coding RNA in the genome is becoming increasingly evident, interest in these molecules is rising. Numbers of publications searchable for the term long non-coding RNA or some variation thereof are shown for each year.

APPENDIX A

Supplemental Information for Chapter 1

Table 1.1 Abbreviations

Abbreviation	Full name	Notes
AGO4	ARGONAUTE 4	Binds siRNA and lncRNA and helps direct chromatin modifications.
CMT3	CHROMOMETHYLASE 3	Maintenance Methyltransferase in CHG context.
DMS3	DEFECTIVE IN MERISTEM SILENCING 3	Part of the DDR complex important for Pol V binding to chromatin.
DMS4	DEFECTIVE IN MERISTEM SILENCING 4	May help Pol V bind to chromatin.
DRD1	DEFECTIVE IN RNA-DIRECTED DNA METHYLATION 1	Part of the DDR complex important for Pol V binding to chromatin.
DCL3	DICER-LIKE 3	Cleaves dsRNA into 24 nt siRNA.
DNMT1	DNA METHYLTRANSFERASE 1	Maintenance Methyltransferase in humans.
DNMT3B	DNA METHYLTRANSFERASE 3 BETA	*de novo* Methyltransferase in humans.
DRM2	DOMAINS REARRANGED METHYLTRANSFERASE 2	Places *de novo* DNA methylation.
dsRNA	double stranded RNA	Often cleaved into siRNA.
H3Ac	Histone 3 Acetylation	Generally a mark of active chromatin.
H3K9me2	Histone 3 Lysine 9 di-methylation	Generally a mark of inactive chromatin.
HOTAIR	HOX ANTISENSE INTERGENIC RNA	Well studied lncRNA involved in guiding chromatin modifications.
HEN1	HUA ENHANCER 1	Methylates small RNA.

IDN2	INVOLVED IN DE NOVO 2	Interacts with SWI3B to help position nucleosomes.
lncRNA	long non coding RNA	Used as structural components, diversions, or scaffolds for proteins.
MET1	METHYLTRANSFERASE 1	Maintenance Methyltransferase in CG context.
miRNA	micro RNA	21 nucleotide product important for Post Transcriptional Gene Silencing.
NRPD1	NUCLEAR RNA POLYMERASE D 1	Largest subunit of Pol IV.
NRPE1	NUCLEAR RNA POLYMERASE E 1	Largest subunit of Pol V.
Pol II	Polymerase II	Polymerase important for messenger RNA.
Pol IV	Polymerase IV	Creates lncRNA as precursors for siRNA.
Pol V	Polymerase V	Creates lncRNA as scaffolds for protein binding.
RDR2	RNA-DEPENDENT RNA POLYMERASE 2	Creates dsRNA from Pol IV transcripts.
RdDM	RNA-directed DNA Methylation	Another name for TGS.
RDM1	RNA-DIRECTED DNA METHYLATION 1	Part of the DDR complex important for Pol V binding to chromatin.
SHH1	SAWADEE HOMEODOMAIN HOMOLOG 1	May help recruit Pol IV to chromatin.
siRNA	small interfering RNA	24 nucleotide product which binds AGO4 and can direct silencing.
SUVH2	SU(VAR)3-9 HOMOLOG 2	May help Pol V bind to chromatin.
SUVH9	SU(VAR)3-9 HOMOLOG 9	May help Pol V bind to chromatin.
SPT5	SUPRESSOR OF TY INSERTION 5	Helps guide Pol II through nucleosomes.
SPT5L / KTF1	SUPRESSOR OF TY INSERTION 5 - LIKE / KOW DOMAIN CONTAINING TRANSCRIPTION FACTOR 1	Binds chromatin dependent on Pol V and helps guide chromatin modifications.
SWI/SNF	SWITCH/SUCROSE NONFERMENTABLE	Proteins important in nucleosome remodeling.
SWI3B	SWITCH3B	Part of a complex important for nucleosome positioning
TAD	Topologically Associated Domain	Distinct clusters of looped

		DNA with similar chromatin features.
TF	Transcription Factor	Proteins which bind DNA and influence transcription.
TGS	Transcriptional Gene Silencing	Important for transposon silencing and gene expression control.
TSS	Transcriptional Start Site	The environment near the TSS can impact gene expression.
UHRF1	UBIQUITIN-LIKE CONTAINING PHD AND RING FINGER DOMAINS 1	Recognizes hemimethylated DNA.
Xist	X-inactive specific transcript	lncRNA involved in X-inactivation.

APPENDIX B

Supplemental Information for Chapter 2

Supplemental Materials and Methods

Chromatin Immunoprecipitation

ChIP was performed as described [1] with slight modifications. 3 grams of 2-3 week old seedling tissue was harvested and crosslinked with 0.5% formaldehyde for 2 min followed by 8 min vacuum infiltration. Glycine was added to 80 mM and vacuum reapplied for 1 min then 4 min. Crosslinked tissue was rinsed with water and frozen in liquid nitrogen. Nuclei were extracted by grinding frozen tissue into powder using a mortar and pestle, suspended in 25 ml of Honda Buffer (20 mM HEPES-KOH pH 7.4, 0.44 M sucrose, 1.25% Ficoll, 2.5% Dextran T40, 10 mM $MgCl_2$, 0.5% Triton X-100, 5 mM DTT, 1 mM PMSF, 1% plant protease inhibitors (Sigma)), filtered through two layers of Miracloth and centrifuged at 2000 x g for 15 min. Nuclear pellets were washed three times with 1 ml of Honda buffer, resuspended in Nuclei Lysis Buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS, 1 mM PMSF, 1% Plant Protease Inhibitors from Sigma) and DNA was fragmented to the average size of 250 bp by 8 pulses of sonication each 10 seconds long with 1 minute pauses in between pulses using Fisher Scientific 100 Sonic Dismembrator at power setting 1. After centrifugation at 15,000 x g for 10 min, the supernatant was diluted 10-fold with 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl pH 8.0, 167 mM NaCl. 50 µl Dynabeads Protein A (Invitrogen) and the appropriate antibody were added and samples were incubated for 8 h at 4 ˚C on a rotating mixer. Bead-antibody complexes were washed 5 times, 5 min each, with binding/washing buffer (150 mM NaCl, 20 mM Tris-HCl pH 8.0, 2 mM EDTA, 1% Triton X-100, 0.1% SDS, 1 mM PMSF and twice for 5 min each with TE. Samples for ChIP-seq were eluted with RIP elution buffer (100 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS) for 20 min at 65 ˚C and were digested with 20 µg of proteinase K (Invitrogen) overnight at 60 ˚C. An equal volume of phenol/chloroform/isoamyl alcohol pH 6.7

25:24:1 was added to extract DNA, followed by addition of an equal volume of chloroform/isoamyl alcohol 24:1 and subsequent precipitation by addition of 2 volumes 100% EtOH, 0.1 volume 3 M Sodium Acetate and 4 µl Glycoblue (Ambion). Precipitated samples were washed once with 70% EtOH and resuspended in 30 µl TE. Other ChIP samples were eluted using 100 µl of 10% (w/v) Chelex (Bio Rad) resin, in water, added to the beads and crosslinking was reversed at 99 ˚C for 10 min. Samples were digested with 20 µg of proteinase K (Invitrogen) for 2 h at 60 ˚C followed by heat-inactivation at 95 ˚C for 10 min. ChIP samples were amplified in triplicate in Applied Biosystems 7500 real time PCR machine and obtained data were analyzed using comparative Ct relative to inputs. All ChIP-real time PCR experiments were replicated in two or three independent biological repeats, which yielded very similar results.

ChIP-seq library construction

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

High-throughput sequencing

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

Pre-processing and mapping of sequencing reads

In essence, all reads were pre-processed and mapped to the *Arabidopsis* genome using a pipeline as previously described [2] with slight modifications. The detailed procedures are described below:

Trimming of 3'-adaptors.

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

Reducing to NR-tags.

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

Mapping to *Arabidopsis* genome.

The trimmed and untrimmed reads were mapped to *Arabidopsis thaliana* genome (TAIR9 assembly) independently using the Bowtie program [3], with parameters tuned to allow ≤6% of seed mismatches (using 34 nt seeds), ≤8% of total mismatches and all valid alignments are reported. A subsequent parsing step was implemented to enforce these restraints, as well as to require insert lengths of ≥15 nt or ≥30 nt for the trimmed and untrimmed reads, respectively. It is of importance that the actual “insert length” for untrimmed reads was determined according to their alignments, by implementing a one-dimensional dynamic programming algorithm that could identify the most possible insert-fragment length based on output from Bowtie. Finally, we filtered only “best-stratum” alignments that contain ≤4% more mismatches compared to the best-hits for any given read.

Summary of mapping and clone-number information.

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

Calling AGO4-bound peaks

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

We then called different sets of peaks using the CSAR [4] R package, with non-default parameters set as: w = 250, considerStrand = "Sum", uniquelyMapped = FALSE, backg = 0, norm = 2e10. Taken together, these parameters extend all mapped reads up-to 250 nt (average size of the ChIP-seq fragments used to construct the sequencing libraries), merge them from both strands, normalize, and finally call AGO4-bound peaks. All calls required significant fold-enrichment between test and control with FDR < 0.05; the FDR was achieved by randomly permuting the mapped reads from test samples 10 times using the CSAR [4] package. As a result, 5 sets of peaks were called:

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

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

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

D = Col-0 ChIP vs. *ago4* ChIP

E = *nrpe1* ChIP vs. *ago4*-ChIP

The A, B, C peak sets are traditional ChIP against input calls and D, E sets are "direct-comparison" peaks using the *ago4* null mutant sample that were included in this study to eliminate effect of non-specific binding of DNA to the AGO4 antibody. We then defined Pol V-dependent and Pol V-independent peaks using the following "peak-arithmetic". Specifically, Pol V-dependent peaks are defined as F – G and Pol V-independent peaks are F∩G, where F = (A – B)∩D and G = (C – B)∩E. This peak-arithmetic was designed to identify high-quality peaks enriched for both ChIP vs. input and WT vs. mutant comparisons and minimize the effects of non-specific interactions. All the peak-arithmetic was performed using BEDTools [5], with the overlapping proportion being no less than half (-f = 0.5) of the peaks being compared.

Filtering Pol V-dependent and Pol V-independent peaks

Our *ago4* mutant plants were originally identified [6] in the Landsberg (Ler-1) ecotype of *Arabidopsis*, which was subsequently back-crossed to Col-0 plants for 3 successive times. As a consequence, the *ago4* plants could still contain some proportion of the genome that originates from Ler-1, and thus the calling procedure of Pol V-dependent and Pol V-independent peaks could include ecotype biases. Therefore, we further filtered out peaks that could originate specifically from Ler-1. In essence, any peak that either 1) cannot be mapped to Ler-1 draft genome (Ler-1

unmappable) or 2) can be better mapped to Ler-1 draft genome (Ler-1 better mapped) was recognized as potentially originating from Ler-1, and thus discarded from further analysis. More specifically, we first pulled out all raw reads as well as their qualities (in FASTQ format) in all called Pol V-dependent or Pol V-independent peaks, then re-mapped them to both the Col-0 genome and Ler-1 draft genome as described above. Only the *ago4* ChIP library was used as a proxy for this analysis. It is of note that we used the "standard" assembly of the 2011-08-25 release of the Ler-1 genome from the 1001 genomes project (http://1001genomes.org/), which was in draft status and still lacked a significant portion of the genome compared to the Col-0 genome sequence (TAIR9 assembly). We also re-mapped the reads to Col-0 genome while retaining read quality information, so that the mapping quality between the Col-0 and Ler-1 genomes for any given read could be better distinguished. All mapping criteria were kept identical as described above. By comparing the alignments for the Col-0 and Ler-1 genomes for each read, we defined "Ler-1 unmappable peaks" as those that contain <50% reads mapping to Ler-1 genome relative to Col-0, and "Ler-1 better mapped peaks" as those that contain more reads that can either exclusively or better map to Ler-1 compared to the Col-0 genome. A read is deemed as "better mapped" to a genome if the best hit to that genome contains fewer mismatches, and if it is a tie (same number of mismatches) they are further resolved by comparing the total quality scores over all mismatch sites.

Sampling of random peaks as negative controls

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

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

To distinguish the AGO4 peaks that are completely-dependent from those that are partially-dependent on Pol V activity, we calculated the total number of reads within all Pol V-dependent peaks for all ChIP samples, then plotted the log-odds of enrichment for Col-0 ChIP vs. *ago4* ChIP against *nrpe1* ChIP vs. *ago4* ChIP on a scatter-plot (Fig. 1C). The Pol V-completely dependent peaks were defined as those with less than 2 fold enrichment when comparing *nrpe1* ChIP vs. *ago4* ChIP (|abs(log-odds)| < 1), whereas

Pol V-partially dependent peaks were defined as all of the remaining peaks. This partitioning of peaks is based on the assumption that Pol V-completely dependent peaks should show no significant difference in clone-abundance between *nrpe1* and *ago4* samples. As expected, most Pol V-dependent peaks are completely dependent, and we didn't separate these peaks in further analyses for convenience, since partially dependent peaks are an insignificant fraction of the total AGO4 peaks.

Classification and annotation of AGO4 peaks

All AGO4-peaks were classified according to their genomic coordinates compared to known genetic elements on the *Arabidopsis* genome using the GFF annotation file downloaded from TAIR9 FTP repository for varies kinds of elements, including protein-coding genes (exons and introns), rRNAs, tRNAs, miRNAs, snoRNAs, snRNAs, ncRNAs, pseudogenes, and transposable elements (TEs). We also defined gene promoters as the upstream 1 kb regions of the transcription-start sites (TSS) of protein-coding genes. Additionally, we also searched the *Arabidopsis* genome (TAIR9 assembly) for more repetitive elements using the RepeatMasker program (RepeatMasker-open-3.2.8) (http://www.repeatmasker.org) with the repeat libraries from RepBase (release14.06) [7]. We used the RepeatMasker annotated repeats because the TEs annotated by TAIR don't have detailed class or family information. Other than the TEs, the RepeatMasker (RMSK) program could also identify repeat-rRNAs (RMSK-rRNAs) and tandem-repeats (RMSK-TRs).

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

Characterizing smRNA profiles near AGO4 peaks

We downloaded published smRNA-IP and total smRNA datasets [8] for both AGO4 and AGO1 from *Arabidopsis* seedlings (accession: GSE28591) for our analysis. Raw reads were dumped from the NCBI SRA, processed, and mapped to the *Arabidopsis*

genome as described for our ChIP-seq libraries. Then smRNA-IP or total smRNA reads were searched within the AGO4 peaks as well as their flanking regions (2 kb of both upstream and downstream), and the base-wise coverage for every peak was determined for 24 nt and 21 nt reads, respectively. Finally, the coverage values were normalized by the total mapped reads of each specific sized library, and averaged across all AGO4 peaks.

Characterizing cytosine methylation in AGO4 peaks

To characterize the cytosine methylation (mC) in AGO4 peaks, we used the published single-nucleotide mC datasets [9] provided by Dr. Ryan Lister. The original mC site coordinates were based on the TAIR8 assembly, so we transformed them into TAIR9 coordinates using the Perl script provided by TAIR. The mC sites were searched within all AGO4 peaks as well as NC-peaks, and the mC density was calculated and compared between AGO4 peaks and NC-peaks for CG, CHG and CHH methylation types or altogether. We also used the recently published single-nucleotide mC datasets [10] to directly compare the mC density between Col-0 and *nrpe1* mutant plants. It is of note that in this comparison a corresponding Col-0 wild type dataset was used.

Characterizing the class and family of transposable elements in AGO4 peaks

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

Displaying the chromosome-distribution of AGO4 peaks

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

Characterizing AGO4 binding profiles around TSS

The log fold-change profile of ChIP-seq reads between Col-0 and *ago4* samples was generated using the CEAS program [11] with relative positions to the TSS of all protein-coding genes.

Characterizing nucleosome profiles around AGO4 peaks

To characterize the nucleosome profiles, we used published MNase-seq datasets [12] from NCBI GEO (accessions GSE21673, GSM543296), and merged raw reads from all 6 replicate runs. The MNase-seq reads were mapped to the *Arabidopsis* genome using Bowtie using the same parameters as described for our ChIP-seq libraries, with the exception that we only kept uniquely mapping reads. The log fold-change profile of MNase-seq reads between Col-0 and *ago4* samples were generated using the CEAS program [11] with relative positions to the TSS of all protein-coding genes. Well-positioned nucleosomes were then called as previously described [13], and the nucleosome-density profiles were determined near all AGO4-peaks or for only promoter overlapping peaks, respectively. It is of note that all mapped MNase-seq reads were extended to 147 nt before calling the well-positioned nucleosomes, which is the known average nucleosome size for eukaryotic genomes.

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

To identify significantly enriched biological processes for AGO4-bound promoters, the corresponding gene IDs (TAIR AGI) of these promoters were extracted and subjected to the GOEAST online Batch-Genes analysis tool [14] with an FDR < 0.05, and other parameters set as default.

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

DMRs identified by Dowen *et al.* (2012) were overlapped with AGO4 peaks using PeakAnalyzer [15]. p-values were derived from 1000 random permutations.

Detection of Pol V-dependent transcripts

Total RNA was isolated from 100 mg 2.5 weeks old seedlings (Col0, *nrpe1*, *ago4*) using the Plant RNeasy Mini kit (Qiagen) including on-column DNase treatment. To remove any potential residual DNA, 1 unit of Turbo DNase (Ambion) was added to 1 µg of total RNA and heat-inactivated after incubation at 25°C for 15 minutes. For cDNA synthesis, 500 ng of the DNase-treated RNA were converted to cDNA using the Random Primer Mix (NEB) and Superscript III Reverse Transcriptase (Invitrogen) following manufacturers' instructions. To detect potential contaminations by genomic

DNA, we also prepared control samples lacking reverse transcriptase. Subsequent real time PCR reactions were performed using Platinum Taq (Invitrogen) on a Bio-Rad CFX Connect Real-Time System.

Supplemental References

1. Rowley, M. J., Avrutsky, M. I., Sifuentes, C. J., Pereira, L. & Wierzbicki, A. T. Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. *PLoS Genet.* **7,** e1002120 (2011).
2. Zheng, Q. *et al.* Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS Genet.* **6,** e1001141 (2010).
3. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
4. Muiño, J. M., Kaufmann, K., van Ham, R. C., Angenent, G. C. & Krajewski, P. ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. *Plant Methods* **7,** 11 (2011).
5. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
6. Zilberman, D., Cao, X. & Jacobsen, S. E. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299,** 716–719 (2003).
7. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–467 (2005).
8. Wang, H. *et al.* Deep sequencing of small RNAs specifically associated with Arabidopsis AGO1 and AGO4 uncovers new AGO functions. *Plant J.* **67,** 292–304 (2011).
9. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133,** 523–536 (2008).
10. Wierzbicki, A. T. *et al.* Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. *Genes Dev.* **26,** 1825–1836 (2012).
11. Shin, H., Liu, T., Manrai, A. K. & Liu, X. S. CEAS: cis-regulatory element annotation system. *Bioinformatics* **25,** 2605–2606 (2009).
12. Chodavarapu, R. K. *et al.* Relationship between nucleosome positioning and DNA methylation. *Nature* **466,** 388–392 (2010).
13. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458,** 362–366 (2009).
14. Zheng, Q. & Wang, X.-J. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.* **36,** W358–363 (2008).
15. Salmon-Divon, M., Dvinge, H., Tammoja, K. & Bertone, P. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics* **11,** 415 (2010).

Table 2.1 Oligonucleotides used in chapter 2

Locus	Application	Primer 1	Primer 2
ACTIN	ChIP-qPCR, qRT-PCR	GAGAGATTCAGATGCCCAGAAGTC	TGGATTCCAGCAGCTTCCA
AT1G12700	ChIP-qPCR	CTTCCCCACCAAATGAATGTT	GAAACCAATACTACTGATGGAGATGC
	PCR	CTACAAACATCCATGGAGACG	AAATATGACATGTGACACGAGGA
AT1G12730	ChIP-qPCR,qRT-PCR	GGATTTAACGACATTTTTCCCTTCA	GGCTTAGGGCCCGTAACTAATAAAT
	PCR	GACTTAACGACGCGTATAATGTGG	TGTGTGAATTAGCGAAACTTGTCA
AT1G19400	PCR	CCTCGGATCTTTGGAGCATT	TTTCTTGGAGCTTTCACATCTGTT
AT1G27770	ChIP-qPCR	GAAATTTTTACAAAAGTAGACGAAAAA	TTTGACAAACAATGAGAGCAAATC
AT1G35160	PCR	TTGCTCAATCAACAGAAATTACAAAA	TGAAAAAGGTGGAGAAAGAAAGAGA
AT1G49490	PCR	TGAGGCTAAAATGATGATAAAATCCA	AACCATGTCTCTGCATATTCAATC
AT1G51805	PCR	CTTGGAGCAATCATTTTATGGTTTT	TAAACCAATGATTGAATGACAACTGTG
AT1G54120	ChIP-qPCR	GGTTCGATTCCTGACCACCA	TAAGGGTCCCCTGGAACTGG
	PCR	AAACCCAGGCATCGTGGTCT	TTTATAATTTAGTGTCCTGACCACACA
AT1G55535	PCR	TCCAAGATTGAGGCCAAATTA	AAAAGGAGTGGCCAAGTTGGAA
AT1G56090	ChIP-qPCR	CGCCCGTGTATTTGTTTTCG	TGCCTAAACGTGTCACATTCC
	qRT-PCR	TCCAGTGAAGAAGAAGACGATGAA	GAGGTGGCGACTCATCACTG
AT1G66580	ChIP-qPCR, qRT-PCR	ATGTGGATGGATTTGTTAACCCTCT	GAAGAACGAAGAACAATGTGTTGAC
AT1G67120	ChIP-qPCR	TTTTAAAAAGTTCGGTTTCCTTC	TCTCTGAAAGATGAAAGAGAGAGAAGA
AT1G79120	ChIP-qPCR	TAGATAGTTTTCTTCTCGTCGAACTCA	GAAGCAACAATGAGTCCCCTGT
AT1TE36560	ChIP-qPCR	TCGATGACAGTCGATAACTCATTTT	AAATCTCTATCCATTGCACATGCTC
	PCR	CATTTTAGAGCATGTGCAATGGA	CCAATAACCAAACGGTTAACCAAA
AT1TE73075	ChIP-qPCR	GAACCTCTCTCTATCTCCTTCATTTTTT	TCAATGAGATACTCTCCCACTAGAA
AT2G01735	PCR	CAAATCTGAAGTCGAACCCAAAA	GTCGGATTCGGGTAAAATTCG
AT2G21840	ChIP-qPCR	CGAGCTTCACTTTTGGGAGTTC	CGATATCCAAACCCATAATTGACC
	qRT-PCR	TTCGAGGGTGACTCGAGCTT	TCCGTAGAACGACACACCACA
AT2G30740	ChIP-qPCR	AATCGTCTTCCCCGCGGTTT	TGCTTGATGATGAAGACGGAGA
AT2G36490	ChIP-qPCR, qRT-PCR	CGTTTGTTTATGTAGGGCGAAAG	TAAAACTTTTCCCGCCAACCA
	qRT-PCR	GGAAACATGTCCAGCGCTTT	TGGAAGAGAAGCAGTTTCAGCA
AT2G46130	ChIP-qPCR	AGGAGAGAAGGAGAGTTAAATTTCTCG	GTGATTACACCTGTCCAATCATCC
AT3G06410	PCR	ACGAACCAGAGGGCTCATTG	TGGATCTTGTGCTATGCTCCAA
AT3G27690	ChIP-qPCR	TTGGTCTTTGTTTCAAAGTACACATGA	GCGCTAACATTTGGGGTACG
	PCR	TTGATGATGCTTACCCATATAAATGTT	TGGATGAGGTGTTATTCCAAAAATG
AT3G28100	ChIP-qPCR	TGGTCCGTTTAACGTCCAAT	GACCGTTCATTAAACCCCAAAA
	PCR	ACTAATAAGGTGTCAAGTGGTCCGTTT	TGGCTTTACGAGTTTCATGACTCC
AT3G30380	PCR	GAGGGAAAGATGATCCGTCAA	TGCACAAAATGACTATGAATTGTAAA
AT3G46700	ChIP-qPCR	CCGGTAATTCCGCTAAAAACAA	TTGTCTCCTTTGAAAGATTATGGAA
AT3G48131	ChIP-qPCR	TCAGTTGCAAGAAGACGACGA	TGGGCTATAAAGAGGCCCAAT
	qRT-PCR	TGCGATCATGTGTTTTCTCTTTTC	GCAAATCGATCTTCACAACGAA
	PCR	GGCAACAAGAAGTAGAGCAAATCG	AAAAAGGAATGTGGAGAGATGAAA
AT4G08310	PCR	GGGTCGGGTTCGGTTAAAA	CAAACCCGAACCCAAAACTAAA
AT4G10570	ChIP-qPCR	CGTTTTCTTAATATTTGTATTTTTCC	TGCTTCTGTTCCTTTTTGTTTGA
AT4G11330	ChIP-qPCR	TCTGGAAATTCAAACTCAAAGACC	GTGGATCCCGCCTCTAGAAAA
AT4G31770	PCR	ATGCCAACGTTGACTCACGA	TCATGAGTTTGGGAATGGTTTT
AT4TE23930	ChIP-qPCR	GGCTGGTGCAACGTGATATG	CTGGCTAGGTCGACCGGGTA
AT4TE36990	PCR	GAACATGAGATGTAATCAAGGGCATA	TTGAATATTTGTGGCATAACATGGA
AT5G01225	PCR	TTCTCATTCAAATTTTCTTGTTTGACA	GGAGCATAGCACAAGGTCCAA
AT5G07250	PCR	CCGGTTTTTGTGGTACGTGTT	GGGTATCAAGTCCAGAAGTTTAGACA
AT5G18640	ChIP-qPCR	TCGAGTTTTGATTATTGTAAGGGTTT	GAGGGTCCCAATTTGTTTGTC
AT5G19257	ChIP-qPCR	TCCATATAAAGAGAAACCGAGTAGGG	AACCACGGTTTTGTAGGGTTTTT

	PCR	TTCAAAGATGGAGTTTCACGTGTC	CGACTTCCGTAACACCCATT
AT5G27860	PCR	TGGAAAAGAATTGAGAGAATGATCTTG	TTCAACCTTTGTTGTTTATTTGTCCA
AT5G28620	PCR	TTGACCAATTATATTTCACCACA	TGCCATATGTTCTTTTCTTCTGA
AT5G52070	ChIP-qPCR	CATCTGATTCTTAACACCACCTACTCA	ATGTCCTGAGCTGCCACGTT
	qRT-PCR	TTGAAGCTGCTGTGTTGGACA	CCAATCAATTCGAACGATAAGCTC
	PCR	GACTTAACGACGCGTATAATGTGG	TGTGTGAATTAGCGAAACTTGTCA
AT5G58510	PCR	AGAGATCCGCTTCGGGAAAG	AGAAACCATTGATAGAGATGGTCTTAG
AT5TE60680	PCR	CAACGAATCAGCCAACTCAGAA	GAGAAGCCTTCAAACCCTAAA

APPENDIX C

Supplemental Information for Chapter 3

Supplemental Materials and Methods

Yeast two hybrid

The full length cDNA of IDN2 was generated by PCR and cloned into pAS2 vector (Clontech). Yeast Y190 cells containing pAS2-IDN2 plasmid were transformed with *Arabidopsis* yeast two hybrid cDNA library, (ABRC stock #CD4-22 [1]), and screened on dropout medium lacking leucine, tryptophan, and histidine but containing 50mM 3-aminotriazol. To test the interaction between two proteins in yeast, the full length cDNAs were cloned into pENTR/D-Topo vector (Invitrogen) to produce entry clones according to manufacturer's instructions. All the entry constructs were subsequently transferred to destination vector pGADT7-GW or pGBKT7-GW [2], and the pGADT7/pGBKT7 empty vectors served as negative controls. All the pGBKT7-based constructs were transformed into yeast strain Y187, and all the pGADT7-based constructs were transformed into yeast strain Y190. Yeast mating of Y187 and Y190 was performed according to Clontech Yeast Protocols Handbook (1999).

Generation of transgenic plants

The full length cDNA or genomic DNA including promoter regions of *SWI3B* and *IDN2* were cloned into pENTR/D-Topo vector (Invitrogen) to produce entry clones according to manufacturer's instructions. The resulting entry plasmids were incubated with the destination vectors: pMDC107 [3], pEarleyGate103 [4], pEarleyGate302 [4], or pZY35S302 with the Gateway LR Clonase™ II Enzyme Mix (Invitrogen) to obtain *IDN2p:IDN2-GFP* (pMDC107-*IDN2*), *IDN2p:IDN2-M8-GFP* (pMDC107-*IDN2-M8*), *SWI3Bp:SWI3B-FLAG* (pEarleyGate-*SWI3B*), *35S:SWI3B-GFP* (pEarleyGate103-*SWI3B*), *35S:IDN2-FLAG* (pZY35S302-*IDN2*), *35S:IDN2-GFP* (pEarleyGate103-*IDN2*), *35S:SWI3B-FLAG* (pZY35S302-*SWI3B*), and *35S:IDN2-M8-FLAG* (pZY35S302-*IDN2-M8*). To generate a binary vector pZY35S302 for 35S-driven expression of C-terminally

FLAG-tagged proteins, the Gateway cassette, FLAG nucleotide sequence and OCS 3' were amplified from pEarleyGate302 [4] using Pfu DNA Polymerase (Agilent Technologies). Primers used for this and other PCR amplifications are shown in Table S1. After *Kpn*I and *Hind*III double digestion, the Gateway cassette was inserted into *Kpn*I and *Hind*III digested pCHF1 vector [5]. All constructed plasmids were introduced into the GV3101 strain of *Agrobacterium tumefaciens* and transformed into *Arabidopsis* plants by the floral dip method [6] or infiltrated into tobacco leaves [7].

Protein co-immunoprecipitation

Infiltrated tobacco leaves or 3-week-old *Arabidopsis* rosette leaves were ground into fine powder in liquid nitrogen, extracted using lysis buffer (50mM Tris-HCl pH8.0, 150mM NaCl, 10mM EDTA, 10% glycerol, 1mM PMSF, 1% Plant Protease Inhibitor (Sigma), 0.5% Triton-X100 and centrifuged at 8000 rpm at 4˚C for 10 min. Resulting protein extracts were incubated with anti-GFP antibody (MBL 598, 1:1000 dilution) and 50µl of 50% slurry of Protein A agarose beads (Invitrogen). Beads were washed 3 times with the lysis buffer, and the bound proteins were eluted with 2x SDS buffer. Gel blots were analyzed using monoclonal anti-GFP antibody (Covance, MMS-118P), or monoclonal anti-FLAG antibody (Stratagene, 200472).

Chromatin-immunoprecipitation (ChIP)

ChIP was performed as described [8] with the following modifications: proceeding washes with Honda buffer, nuclei were washed once in 1 ml MNase reaction buffer (10 mM Tris-Cl pH 8, 15 mM NaCl, 60 mM KCl, 1 mM $CaCl_2$; centrifugation at 1900 g, 5 minutes at 4°C) and resuspended in 1 ml MNase reaction buffer. 250 µl aliquots of nuclei were incubated with 600 Kunitz units of Micrococcal Nuclease (NEB) at 37°C for 10 min, then sonicated with two 10 second long pulses (1 minute intervals) with a Fisher Scientific Sonic Dismembrator Model 100 (power setting 1). Immunoprecipitation was performed using 50 µl Dynabeads protein A (Invitrogen) and 2.5 µl histone H3 antibody (ab1791, Abcam) or affinity purified anti-SWI3B antibody at 4°C over-night. After reversion of crosslinking, samples were incubated with 20 µg proteinase K (Invitrogen) at 65°C for 2 hours. Rabbit polyclonal anti-SWI3B antibody was raised against a C-terminal portion of the SWI3B protein (aa 248-469) expressed in bacteria and affinity purified. H3 ChIP-seq samples were treated similarly, but without MNase treatment and

were sonicated eight times with 10 second long pulses. Library generation and Illumina sequencing were performed by the University of Michigan Sequencing Core.

RNA analysis

For RT-PCR total RNA from inflorescences was extracted using RNeasy Plant Mini kit (Qiagen), and treated with DNase I (Invitrogen). Real time RT-PCR was performed using One-Step qRT-PCR kit (Invitrogen) according to the manufacturer's instructions. Pol V-dependent transcripts were assayed in RNA digested with 1 unit of Turbo DNase (Ambion) and reverse transcribed with Superscript III reverse transcriptase (Invitrogen) using random primers (Invitrogen) followed by real time PCR. For RNA-seq total RNA was extracted from 2.5 weeks old seedlings using RNeasy Plant Mini kit (Qiagen). rRNA was depleted from 8µg total RNA using RiboMinus Plant Kit for RNA-seq (Invitrogen). Library generation and Illumina sequencing was performed by the University of Michigan Sequencing Core.

DNA methylation analysis

DNA methylation tests using methylation sensitive restriction endonucleases were performed as described [9] and analyzed by PCR or real-time PCR.

MNase-seq

2g of 2.5-weeks old seedlings were ground to a fine powder in liquid nitrogen and resuspended in 15 ml Honda buffer (0.44 M Sucrose, 1.25 % Ficoll, 2.5 % Dextran T40, 20 mM HEPES-KOH pH7.4, 10 mM $MgCl_2$, 0.5% Triton X-100, 5 mM DTT, 1 mM PMSF, 1 % plant protease inhibitors (Sigma)). After filtering through two layers of Miracloth, the filter was washed in 10 ml Honda buffer. This washing buffer was then filtered through two fresh layers of Miracloth and the combined filtrates were centrifuged (2500 g, 15 minutes at 4°C). The pellet was washed four times in 1 ml Honda buffer (centrifugation at 2500 g, 15 minutes at 4°C) and 1 ml MNase reaction buffer (10 mM Tris-Cl pH 8, 15 mM NaCl, 60 mM KCl, 1 mM $CaCl_2$; centrifugation at 3000 g, 5 minutes at 4°C) and finally resuspended in 660 µl MNase reaction buffer. 100 µl aliquots of nuclei were incubated with Micrococcal Nuclease (NEB) at 20°C for 10 minutes. To terminate the reaction, 10 µl STOP buffer (100 mM EDTA, 100 mM EGTA), 10 µl 10 % SDS and 40 µg proteinase K (Invitrogen) were added followed by an incubation at 60°C for one hour. DNA was extracted with phenol-chloroform-isoamyl alcohol and

chloroform-isoamyl alcohol and precipitated with ethanol. The pellet was washed in 70% ethanol, resuspended in 30 µl TE and incubated with 1 U RNase cocktail (Ambion) at 37°C for one hour and then at 4°C over-night. DNA corresponding to the mononucleosomal fraction was purified (QIAEX II gel extraction kit, Qiagen) after separation on a 2% agarose gel and 20 ng of DNA was used for library generation. Library generation and Illumina sequencing was performed by the University of Michigan Sequencing Core.

Supplemental References

1. Kim, J., Harter, K. & Theologis, A. Protein-protein interactions among the Aux/IAA proteins. *Proc. Natl. Acad. Sci. U.S.A.* **94,** 11786–11791 (1997).
2. Lawit, S. J., O'Grady, K., Gurley, W. B. & Czarnecka-Verner, E. Yeast two-hybrid map of Arabidopsis TFIID. *Plant Mol. Biol.* **64,** 73–87 (2007).
3. Curtis, M. D. & Grossniklaus, U. A gateway cloning vector set for high-throughput functional analysis of genes in planta. *Plant Physiol.* **133,** 462–469 (2003).
4. Earley, K. W. *et al.* Gateway-compatible vectors for plant functional genomics and proteomics. *Plant J.* **45,** 616–629 (2006).
5. Fankhauser, C. *et al.* PKS1, a substrate phosphorylated by phytochrome that modulates light signaling in Arabidopsis. *Science* **284,** 1539–1541 (1999).
6. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J.* **16,** 735–743 (1998).
7. Voinnet, O., Rivas, S., Mestre, P. & Baulcombe, D. An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus. *Plant J.* **33,** 949–956 (2003).
8. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135,** 635–648 (2008).
9. Rowley, M. J., Avrutsky, M. I., Sifuentes, C. J., Pereira, L. & Wierzbicki, A. T. Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. *PLoS Genet.* **7,** e1002120 (2011).

Table 3.1 Oligonucleotides used in chapter 3

Target	Name	Sequence (5'-3')	Application
Primers for plasmid constructs			
IDN2	proIDN2-F	CACCTGTTGTAGTCGCTGTGATAC	amplify genomic IDN2
	IDN2-R	AGCCATTCCACGCTTGCGTTTCGC	
IDN2	IDN2-cacc-F	CACCATGGGAAGCACTGTGATTTTA	amplify IDN2 cDNA
	IDN2-R	AGCCATTCCACGCTTGCGTTTCGC	
IDN2	IDN2-pAS2-EcoRI-F	TGGAGGCCGAATTCATGGGAAGCACTGTGATT	generate pAS2-IDN2 construct
	IDN2-pAS2-PstI-R	TAGCTTGGCTGCAGCTAAGCCATTCCACGCTTG	
SWI3B	SWI3b-cacc-F	CACCATGGCCATGAAAGCTCCCGA	amplify SWI3B cDNA
	SWI3b-R	ACACTCTATTCTATCTTCAGTTTTCC	
SWI3A	SWI3a-cacc-F	CACCATGGAAGCCACTGATCCAAG	amplify SWI3A cDNA
	SWI3a-R	TTTCACGTACGTATGATCCCAACG	
SWI3D	SWI3d-cacc-F	CACCATGGAGGAAAAACGACGCGA	amplify SWI3D cDNA
	SWI3d-R	CGAAGAAACATTGTCTGAACCTG	
SWI3C	SWI3C-cacc-F2	CACCATGCCAGCTTCTGAAGATAGAAGAGG	amplify SWI3C cDNA
	SWI3C-R2	TAAGCCTAAGCCGGACCCTGAGCCTGAAC	
SWI3B	SWI3B-Pro-CACC-F	CACCTTAAGGCATGCGTTGAAGCAAAAGTT	amplify genomic SWI3B
	SWI3b-R	ACACTCTATTCTATCTTCAGTTTTCC	
Gateway	pEG300F	CGTCACGTCTTGCGCACTGATTTG	generate pZY35S302 vector
	pEG300R	GAACCCTGTGGTTGGCATGCAC	
TRUNCATIONS AND SITE-DIRECTED MUTAGENESIS			
IDN2	IDN2-274F	CACCATGAAATACCTTCAACAAGATCTTGCTG	generate IDN2-ΔE deletion
	IDN2-StopR	CTAAGCCATTCCACGCTTGCGTTTCGC	
IDN2	IDN2-721F	CACCATGGGAGAAAACTTGAGGAAGACGGG	generate IDN2-ΔF deletion
	IDN2-StopR	CTAAGCCATTCCACGCTTGCGTTTCGC	
IDN2	IDN2-964F	CACCATGAGTCACATTCAAAAGATAGTTG	generate IDN2-ΔG deletion
	IDN2-StopR	CTAAGCCATTCCACGCTTGCGTTTCGC	
IDN2	IDN2-cacc-F	CACCATGGGAAGCACTGTGATTTTA	generate IDN2-ΔH deletion
	IDN2-1521StopR	CTAATTTGTGTTCCATTCTTTCATAATGT	
IDN2	IDN2-cacc-F	CACCATGGGAAGCACTGTGATTTTA	generate IDN2-ΔI deletion
	IDN2-759StopR	CTATATAGTTTTCAGATCACCCGTCTTCC	
IDN2	IDN2-cacc-F	CACCATGGGAAGCACTGTGATTTTA	generate IDN2-ΔJ deletion
	IDN2-369StopR	CTAATGATCACAATCTTGAATAGGGTTTC	
IDN2	IDN2-del760-897(250-299)F	CTTGAGGAAGACGGGTGATCTGAAAACTATAATGG AAGAGAAGGAGAAGAATCAGCAAAAGC	generate IDN2-ΔA deletion
	IDN2-del760-897(250-299)R	GCTTTTGCTGATTCTTCTCCTTCTCTTCCATTATAGT TTTCAGATCACCCGTCTTCCTCAAG	
IDN2	IDN2-del964-1059(322-353)F	CGTGAGCTGAATGCTATACAAGAAAGAACAGCAAA GCGCGAAGTGCACAATGGAACCGAG	generate IDN2-ΔB deletion
	IDN2-del964-1059(322-353)R	CTCGGTTCCATTGTGCACTTCGCGCTTTGCTGTTCT TTCTTGTATAGCATTCAGCTCACG	
IDN2	IDN2-del1153-1320(385-440)F	GCATCTAAGAATAGCTCTCTTGAACTAGCTAAGCAC ATGGCATCAGATGGCGATGCTGAAG	generate IDN2-ΔC

	IDN2-del1153-1320(385-440)R	CTTCAGCATCGCCATCTGATGCCATGTGCTTAGCTAGTTCAAGAGAGCTATTCTTAGATGC	deletion
IDN2	IDN2-del1405-1512(469-504)F	CTTCAAAGATTTAGGTGAGAAGGAAGCACAAAACACAAATATCGGTGTTAAGAGAATGGGAG	generate IDN2-ΔD deletion
	IDN2-del1405-1512(469-504)R	CTCCCATTCTCTTAACACCGATATTTGTGTTTTGTGCTTCCTTCTCACCTAAATCTTTGAAG	
IDN2	IDN2-I325R/V329G/H332R-F	AGAACAATGAGTCACAGACAAAAGATAGGTGATGATCGTGAGAAATTGAAGAGG	generate IDN2 triple mutant
	IDN2-I325R/V329G/H332R-R	CCTCTTCAATTTCTCACGATCATCACCTATCTTTTGTCTGTGACTCATTGTTCT	
IDN2	IDN2-MM5on3-F	GATGATCGTGAGAAATTGGGGAGGCTGAGGGAGTCAGAGGGGAAGAAACGCGAAATCAAAGGTAATGAGTTGGCAAAGC	generate IDN2 octuple mutant (M8)
	IDN2-MM5on3-R	GCTTTGCCAACTCATTACCTTTGATTTCGCGTTTCTTCCCCTCTGACTCCCTCAGCCTCCCCAATTTCTCACGATCATC	
RNA DETECTION			
SWI3B	Swi3b-qRT-F2	CGGCGAAGTTGCGTTAGTTAAACA	real time RT-PCR
	Swi3b-qRT-R2	CCTCCAGACGTAGTTTCGGAAAGA	
ACTIN2	Actin2-A118	GAGAGATTCAGATGCCCAGAAGTC	real time RT-PCR RNA IP-qPCR
	Actin2-A119	TGGATTCCAGCAGCTTCCA	
soloLTR (At5TE35950)	soloLTR-F4	TCATGTTAAAACCGATTGCACCATTT	real time RT-PCR
	soloLTR-R4	CAAAAATTAGGATCTTGTTTGCCAGCTA	
soloLTR (At5G27845)	IG-up-F8	CGGAATGGGGAAATTTCAAGGACGC	real time RT-PCR
	IG-up-R8	CAGTGACGCTGTCACCCTCGAA	
At1TE51360	LTRCO1-F3	GCCGAATGGCTCATTAAGTACCTG	real time RT-PCR
	LTRCO1-R3	AAGTGGGTATTCGTGCGAAAAAGA	
At3TE51910	LTRCO3-F2	ATAACCTTCCCACGCTGCATTAGA	real time RT-PCR
	LTRCO3-R2	TGTGAGCCTGAAGGAGATGTTGAC	
At2TE78930	78930F1	TTGATTAATGATCGCGAAAAAGTA	real time RT-PCR
	78930FR1	TAATGAGTGTTGATCGGAAAGAGA	
At1TE58825	58825F1	ACTTACGCATCTCATTGTGTTGTT	real time RT-PCR
	58825R1	ATCCTCTCTTCCTTGTCATGATTC	
At4TE27915	27915F1	ATTCAATCGCTCCGGTAAAATCCT	real time RT-PCR
	27915R1	AGATCGTGGTCTCGTCTGTTTTCC	
At3TE47400	IG12F1	CGAAGCTTCCCACAAAATATCGTC	real time RT-PCR
	IG12R1	GAGGGAAGGAGAAGGAGCAGAATC	
TUBULIN8	JR147	GCTTACTAATCAAAGATGCGAGA	real time RT-PCR
	JR148	CTTGGTATCTTCCCGTCGAA	
UBQ10	GB473_UBQ10s_fw	CCATCACCCTTGAAGTGGAA	real time RT-PCR
	GB474_UBQ10s_rv	GATCTTGGCCTTGACGTTGT	
ROC3	GB469_ROC3s_fw	AAGGTTGGATCTGACTCTGGAA	real time RT-PCR
	GB470_ROC3s_rv	TCTGACCACAATCAGCAATGA	
25S rRNA	JR41	TGTTCACCCACCAATAGGGAA	RNA IP-qPCR
	JR42	TCAGTAGGGTAAAACTAACCTGTCTC	
IGN5	GB268_IGN5-A	ACATGAAGAAAGCCCAAACC	real time RT-PCR
	GB269_IGN5-A	GCCGAATAACAGCAAGTCCT	
IGN20	GB280_IGN20	AAGAACCGGACCAATACGG	real time RT-PCR
	GB281_IGN20	CCACCGCCTCTATTGAAATG	

IGN22	GB282_IGN22	TGGTCCATAGGTTCGGAATTT	real time RT-PCR
	GB283_IGN22	GGCATGGTTTGATATCAGGAG	
IGN25	GB288_IGN25	AAACCCACCTCTTTAGGTCCA	real time RT-PCR
	GB289_IGN25	GGCTTGGAGAGTCCAACAAT	
IGN26	GB290_IGN26	CGTTGTTCCGCCTAATTCTG	real time RT-PCR
	GB291_IGN26	GCCAGGAAACCCTAACTTCC	
IGN27	JA13	GGATTTAACGACATTTTTCCCTTCA	real time RT-PCR
	JA14	GGCTTAGGGCCCGTAACTAATAAAT	
IGN28	JA17	ATGTGGATGGATTTGTTAACCCTCT	real time RT-PCR
	JA18	GAAGAACGAAGAACAATGTGTTGAC	
IGN29	JA227	CGTTTGTTTATGTAGGGCGAAAG	real time RT-PCR RNA IP-qPCR
	JA228	TAAAACTTTTCCCGCCAACCA	
IGN30	GB402_PV-3	GTGTGATGATGTATCATTTATATGGAG	real time RT-PCR RNA IP-qPCR
	GB403_PV-3	ATATATGAAAATTGGCCTACACTCTC	
IGN31	GB416	CAATCTGGCACACACGAAAC	real time RT-PCR RNA IP-qPCR
	GB417	CAGGTTGGATCTGTTGACGA	
IGN32	GB424	CCGAAACCACAGCATGTAAT	real time RT-PCR RNA IP-qPCR
	GB425	TGAAATTTTCGCATCACAACA	
IGN33	GB418	TCTCTTAGGTTCCACCGGATT	real time RT-PCR RNA IP-qPCR
	GB419	CGGGTTTCATTCGTCTTCAT	

DNA METHYLATION ASSAYS

soloLTR (At5TE35950)	soloLTR-C-F(A211)(AluI)	ATAAAACTCGAAACAAGAGTTTTCTTATTGCTTTC	Chop qPCR, AluI
	soloLTR-C-R(A212)(AluI)	TAATGGTATTATTTTGATCAGTGTTATAAACCGGA	
siR02	siR02chop-qPCR-F(AluI)	ATAGTGCAGTTCCGAAACAGTAAACCAT	Chop qPCR, AluI
	siR02chop-qPCR-R(AluI)	TCAAAGTGAAAGTGGTTCTTGGGTTTAT	
At2TE78930	78930M1-F(AvaII)	ATCAATACAAGGTCCATCAACAAA	Chop qPCR, AvaII
	78930M1-R(AvaII)	GGGATTGAGGGTTTGAGTTTAGGG	
JA35/JA36	JA35	GGCGACCTTCTCGAGTTTCC	Chop qPCR
	JA36	CAAGAACCCCACCCCATACA	
IGN6	IGN6-A30(AluI)	GGGACATCTATTGGGTTTAGGCTGGATG	Chop qPCR, AluI,
	IGN6-A31(AluI)	TTTGTAATTCTCAGTTCGGGTATCTGCTTG	
IGN22	IGN22-A413(AvaII)	CAAAAATATTCACCCGCTACAAACAAAAA	Chop-qPCR, AvaII
	IGN22-A414(AvaII)	TCTTCCATTTGTGGGGCATGGT	
At3TE51910	51910M-F1(NlaIII)	TATTACATTGTCCCCCGCTATCA	Chop qPCR, NlaIII
	51910M-R1(NlaIII)	GGTGGAAGCATAAAGGATTAGGG	
AtSN1	AtSN1-A32(HaeIII)	ACCAACGTGCTGTTGGCCCAGTGGTAAATC	Chop PCR, HaeIII,
	AtSN1-A33(HaeIII)	AAAATAAGTGGTGGTTGTACAAGC	
IGN5	IGN5-A28(HaeIII)	TCCCGAGAAGAGTAGAACAAATGCTAAAA	Chop PCR, HaeIII,
	IGN5-A29(HaeIII)	CTGAGGTATTCCATAGCCCCTGATCC	
MEA-ISR	MEA-ISR-F(Sau3AI)	AAAAAGCTCTTTAAAATCCGAAAGTAAC	Chop PCR, Sau3AI
	MEA-ISR-R(Sau3AI)	ACATTGTGAAATCTAACCGGATTTTGGA	
PAR5	methpar5L	GGCGACCTTCTCGAGTTTCC	Chop PCR,

	methpar5R	CAAGAACCCCACCCCATACA	non cutting control for Sau3AI
ACTIN2	Actin2g-qPCR-F;	TTTATTTGCTGGATCTCGATCTTGTTTT	Chop qPCR, non cutting control for AluI, AvaII, NlaIII
	Actin2g-qPCR-R;	AAACCAAAAGATTTAGTGGAGGTTCACA	
ACTIN2	Actin2g-qPCR-F2;	AGTGTCGTACGTTGAACAGAAAGC	Chop qPCR, non cutting control for Sau3AI
	Actin2g-qPCR-R2;	GAGCTGCAAACACACAAAAAGAGT	
ACTIN2	ACTIN2-A65	CGAGCAGGAGATGGAAACCTCAAA	Chop PCR, non cutting control for HaeIII,
	ACTIN2-A66	AAGAATGGAACCACCGATCCAGACA	
NUCLEOSOME VALIDATION			
PVS1	JR339	GAAAATTAGAGAGTGAAACGAGAGCA	ChIP-qPCR
	JR340	TTTATTGGCCTGCCCTATTTG	
PVS2	JR377	CCTTCAAGGGGTGTGAAAAGA	ChIP-qPCR
	JR378	TCTCCTTCTTCGCTGCCAAA	
PVS3	JR379	CCCACAAAAATGGTTTTCCATC	ChIP-qPCR
	JR380	CAAGCCCAACATCTCGGAAA	
PVS4	JR381	CCCATTGGTCCATTTGGTGT	ChIP-qPCR
	JR382	GGGCCTGTAGTGGCCTTGTA	
PVS5	JR555	AGTTGGATGGAGTCCACGAC	ChIP-qPCR
	JR556	CGCTCTCTGCAATTTTGCTT	
PVS6	JR575	AAGGAGAAGAGACGAGTTGATGA	ChIP-qPCR
	JR576	TGCCTCTTGCGAAAACAACA	
ACTIN2	Actin2-A118	GAGAGATTCAGATGCCCAGAAGTC	ChIP-qPCR
	Actin2-A119	TGGATTCCAGCAGCTTCCA	
HSP70	A512	CTCTTCCTCACACAATCATAAACA	ChIP-qPCR
	A513	CAGAATTGTTCGCCGGAAAG	

APPENDIX D

Supplemental Information for Chapter 4

Supplemental Materials and Methods

Chromatin immunoprecipitation protocol

Three grams of above-ground tissue of 2-week old plants was crosslinked with 0.5% formaldehyde for 10 min by vacuum infiltration, followed by addition of glycine to 80 mM. Plants were rinsed with water, frozen in liquid nitrogen, ground into powder using a mortar and pestle, suspended in 25 ml of Honda Buffer (20 mM HEPES-KOH pH 7.4, 0.44 M sucrose, 1.25% ficoll, 2.5% Dextran T40, 10 mM $MgCl_2$, 0.5% Triton X-100, 5 mM DTT, 1 mM PMSF, 1% plant protease inhibitors (Sigma)), filtered through two layers of Miracloth and centrifuged at 2000 x g for 15 min. Nuclear pellets were washed three times with 1ml of Honda buffer, resuspended in Nuclei Lysis Buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS, 1 mM PMSF, 1% Plant Protease Inhibitors) and DNA was fragmented to the average size of 350-500 bp by 8 to 10 pulses of sonication each 10 seconds long with 1 minute pauses in between pulses using Fisher Scientific 100 Sonic Dismembrator at power setting 1. After centrifugation at 16000 x g for 10 min., the supernatant was diluted 10-fold with 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl pH 8.0, 167 mM NaCl. 25 µl of Protein A Agarose/Salmon Sperm DNA (Millipore) or Dynabeads Protein A (Invitrogen) and the appropriate antibody was added. Samples were then incubated for 8h or overnight at 4ºC on a rotating mixer. Bead-antibody complexes were washed 5 times, 5 min each, with binding/washing buffer (150 mM NaCl, 20 mM Tris-HCl pH 8.0, 2 mM EDTA, 1% Triton X-100, 0.1% SDS, 1 mM PMSF) and twice for 5 min each with 10 mM Tris-HCl pH 8.0, 1 mM EDTA. 100µl of 10% (w/v) Chelex (Bio Rad) resin, in water, was then added to the beads and crosslinking was reversed at 99 ºC for 10 min. Samples were digested with 20 µg of proteinase K (Invitrogen) for 1-2 h at 43-60 ºC followed by heat-inactivation at 95 ºC for 10 min. Alternatively, elution was performed twice with 50 µl

RIP elution buffer (100 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS) for 20 min at 65 ºC. Samples were digested with 20 µg of proteinase K (Invitrogen) for 6h - overnight at 60 ºC. An equal volume of phenol/chloroform/isoamyl alcohol pH 6.7 25:24:1 was added to extract DNA, followed by addition of an equal volume of chloroform/isoamyl alcohol 24:1 and subsequent precipitation by addition of 2 volumes 100% EtOH, 0.1 volume 3 M Sodium Acetate and 4ul Glycoblue (Ambion). Precipitated samples were washed once with 70% EtOH and resuspended in 100ul water or TE.

Table 4.1 Oligonucleotides used in chapter 4

Locus	TAIR Annotation	Primer Sequence	Method	Reference in Chapter 4
Actin	AT3G18780	GAGAGATTCAGATGCCCAGAAGTC	real-time PCR	Wierzbicki et al. 2008
		TGGATTCCAGCAGCTTCCA		
		CGAGCAGGAGATGGAAACCTCAAA	PCR	Wierzbicki et al. 2008
		AAGAATGGAACCACCGATCCAGACA		
TUB8	AT5G23860	GCTTACTAATCAAAGATGCGAGA	real-time PCR	Numa et al. 2010
		CTTGGTATCTTCCCGTCGAA		
AtSN1	AT3TE63860	CCAGAAATTCATCTTCTTTGGAAAAG	real-time PCR	Wierzbicki et al. 2008
		GCCCAGTGGTAAATCTCTCAGATAGA		
		ACCAACGTGCTGTTGGCCCAGTGGTAAATC	PCR	Wierzbicki et al. 2008
		AAAATAAGTGGTGGTTGTACAAGC		
solo LTR	AT5TE35950	GGATAGAGATGAATGATGGATAATGACA	real-time PCR	Wierzbicki et al. 2008
		TTATTTTGATCAGTGTTATAAACCGGATA		
		ATAAAACTCGAAACAAGAGTTTTCTTATTGCTTTC	PCR	Wierzbicki et al. 2008
		TAATGGTATTATTTTGATCAGTGTTATAAACCGGA		
IGN5	Between AT4TE10770 & AT4TE10775	AAGCCCAAACCATACACTAATAATCTAAT	real-time PCR	Wierzbicki et al. 2008
		CCGAATAACAGCAAGTCCTTTTAATA		
		TCCCGAGAAGAGTAGAACAAATGCTAAAA	PCR	Wierzbicki et al. 2008
		CTGAGGTATTCCATAGCCCCTGATCC		
IGN20	Between AT4G00400 & AT4G00413	GCGGTGGCTCGAGTCAAAA	real-time PCR	-
		CCTTCCTTTGTGTCGAATTAGTCCTA		
		TGTTAGCCAAAACCGACAAGAACC	PCR	-
		TTTGTCCTCGATTTTGTTCCTTCCT		
IGN22	AT4G01530	CGGGTCCTTGGACTCCTGAT	real-time PCR	-
		TCGTGACCGGAATAATTAAATGG		
		CAAAAATATTCACCCGCTACAAACAAAAA	PCR	-
		TCTTCCATTTGTGGGGCATGGT		
IGN23	AT4TE12070	GCCATTAGTTTTAGATGGACTGCAA	real-time PCR	-
		GGGCGAACCTGGAGAAAGTT		
		ACTGAAAATTGTAAACAAAGAAACGGCACTACA	PCR	-
		GATCGGTCCATAAACTTGTTGGGTTT		
IGN25	AT4TE22865	TCAAACCAAACCCCGAACTT	real-time PCR	-
		ATGCCAGAGCCTGGTGCTA		
		CTTCTTATCGTGTTACATTGAGAACTCTTTCC	PCR	-
		ATTCGTGTGGGCTTGGCCTCTT		
IGN26	Between AT4G11485 & AT4G11490	TTCCTGGCCGTTGATTGGT	real-time PCR	-
		CGTGACATTAGAAGCTCTACGAGAA		
		CTCTTTCAGTGCGACAGCCTCAT	PCR	-
		CGGCCAGGAAACCCTAACTTCC		
SPT5L mRNA	AT5G04290	TTCGTCTGCTGGTGGTTGTGCT	real-time PCR	-
		CCCGGTTTGTCATTGGTTTCTTTCT		

APPENDIX E

Supplemental Information for Chapter 5

Table 5.1 Oligonucleotides used in chapter 5

Locus ID	Primer 1	Primer 2
ACTIN	GGATAGAGATGAATGATGGATAATGACA	TTATTTTGATCAGTGTTATAAACCGGATA
IGN22	CGGGTCCTTGGACTCCTGAT	TCGTGACCGGAATAATTAAATGG
IGN25	TCAAACCAAACCCCGAACTT	ATGCCAGAGCCTGGTGCTA
IGN28	ATGTGGATGGATTTGTTAACCCTCT	GAAGAACGAAGAACAATGTGTTGAC
IGN29	CGTTTGTTTATGTAGGGCGAAAG	TAAAACTTTTCCCGCCAACCA
IGN34	ATGAATAACAAATTTGGAGTCGTC	CCCTTTCATCGACACTGACA
IGN35	GACGGACCAAACGATTTCAT	TTCCTCTTTGAGCTTGACCA
IGN36	CAGTTTTGGGTGCGGTTTAT	GACAAAAATTGCTTTAGACCATGA
IGN5A	CGTTTCTAGAAGAACTGATTGG	TTTGTTTAATAAAATTCTATCAGCTG
JA19	GTGGACTTTCCTTTTAGGCTGTTTT	CTTGATGATGAAGACGGAGACAGAT
JA213	CATCTGATTCTTAACACCACCTACTCA	ATGTCCTGAGCTGCCACGTT
JR1031	CGGTGGATAAAGTATACGCCACAT	TCACTGCCTTGTTTTGTCTGTGTC
JR1033	AGATGGTTGTGACAAAAAGGAAAA	TGCAATGGACATGCTCTAAGTGTT
JR1035	GGGTCTCAAAACCGTCATATTTTG	ACCTTTTGGTTATCGTTTGTGACG
JR1039	ACCATCTAATCAGGCCCGACTCTT	GTGCACTAGCACACTCAGCACCT
JR1041	ACCGGTTTTTGTGGTACGTGTTAC	TAGCTCCGTTACGATACAGTGTGC
JR341	CTTACTAGGCCTATCAAATTAAAGCA	CTATCAAGCGGGGCCATCAG
JR587	TGATTTGAAAACAAAAACAAGTAACGA	TCCCCAGTAAAGTCCTCACACC
JR687	GATATCCGGTTTTTCGGATCG	ATCCGCGGGTATCCGTATCT
JR693	CGGTCATGAGCATAACCAGATG	GAGTTTCTGGCCCAAAATATGC
soloLTR	GGATAGAGATGAATGATGGATAATGACA	TTATTTTGATCAGTGTTATAAACCGGATA
UBQ	CCATCACCCTTGAAGTGGAA	GATCTTGGCCTTGACGTTGT

APPENDIX F

Supplemental Information for Chapter 6

Table 6.1 Oligonucleotides used in chapter 6

Chromatin Loop	Primer 1	Primer 2
AT2TE22565 - AT3G03855	CGTTATCATCATCACCATTACTACCG	AGGAGAGAGAGAGGGAGAGAGATAGG
AT5TE93845 - AT4G10920	TAACCACAAATCCGCGTTTACTGT	TTTATGGTGAAAAATTAAGAGCCAAA
AT3TE70710 - AT4G17000	CACGTTATCAGCACGCTCTAAAAG	TCTAGATGATGGGCTTAGATGATAAGT

APPENDIX G

Supplemental Information for Chapter 7

Table 7.1 Publicly available datasets generated by this work

GEO ID	Chapter Data was Analyzed	Description	Method
GSE35381	Chapter 2 and 5	AGO4 binding sites	AGO4 ChIP-seq
GSE41143, GSE38401	Chapter 3	Nucleosome occupancy in Col-0 and *nrpe1*	MNase-seq, H3 ChIP-seq
GSE38400	Chapter 3	mRNA levels in Col-0, *nrpe1*, *ago4*, *idn2*, *swi3b*	RNA-seq
To be uploaded	Chapter 5	Nucleosome occupancy in Col-0, *spt5l*, *ago4*, and *idn2*	MNase/H3 ChIP-seq
To be uploaded	Chapter 5	SPT5L binding sites	SPT5L ChIP-seq
To be uploaded	Chapter 6	Long distance chromatin-interaction sites in Col-0, *nrpe1*, *ago4*, and *spt5l*	HI-C

APPENDIX H

Analysis of Long Non-coding RNAs Produced by a Specialized RNA Polymerase in *Arabidopsis thaliana*

The contents of this appendix are methods by which to study long non-coding RNAs in Arabidopsis and were published in the journal Methods in 2013. Gudrun Böhmdorfer contributed to sections on RIP and RT-PCR.

Abstract

Long non-coding RNAs (lncRNAs) play important roles in several processes including control of gene expression. In *Arabidopsis thaliana*, a class of lncRNAs is produced by a specialized RNA Polymerase V (Pol V), which is involved in controlling genome activity by transcriptional gene silencing. lncRNAs produced by Pol V have been proposed to serve as scaffolds for binding of several silencing factors which further mediate the establishment of repressive chromatin modifications. We present methods for discovery and characterization of lncRNAs produced by Pol V. Chromatin Immunoprecipitation coupled with deep sequencing (ChIP-seq) allows discovery of genomic regions bound by proteins in a manner dependent on either Pol V or transcripts produced by Pol V. RNA Immunoprecipitation (RIP) allows testing lncRNA-protein interactions at identified loci. Finally, real-time RT-PCR allows detection of low abundance Pol V transcripts from total RNA. These methods may be more broadly applied to discovery and characterization of RNAs produced by distinct RNA Polymerases.

Introduction

Long non-coding RNA (lncRNA), among other functions, is involved in directing chromatin modifications in order to control genes and transposons [1]. These modifications include *de novo* DNA methylation, histone modifications, and nucleosome

positioning in a pathway known as RNA-mediated transcriptional gene silencing or RNA-dependent DNA methylation (RdDM) [2,3,4]. Mutations in components of the RdDM pathway can cause increased transposon activity, faulty DNA repair, or misregulation of genes [4,5,6].

In *Arabidopsis thaliana*, a class of lncRNAs involved in RdDM is produced by RNA Polymerase V, a specialized DNA dependent RNA polymerase [7]. Pol V transcripts are thought to create scaffolds on which several factors bind in order to control chromatin states [3,8]. One of these factors is ARGONAUTE4 (AGO4), which binds to the Pol V C-terminal domain as well as to RNA and DNA [9,10]. AGO4 is guided to chromatin by not only Pol V and its transcripts, but also by 24 nucleotide small interfering RNA (siRNA) [4,8]. These siRNAs are generated from cleavage of double stranded transcripts produced by the coordinated activity of RNA Polymerase IV and RNA-dependent RNA Polymerase 2 [4,11]. AGO4 associates with siRNA and is guided to chromatin based on the sequence specificity of the siRNA and the localization of Pol V [3,5].
Another protein that binds to chromatin dependent on Pol V-produced lncRNA is SPT5-like (SPT5L) [12,13,14]. SPT5L works with AGO4 in a locus specific manner to direct the activity of the *de novo* DNA methyltransferase, DRM2 [2,13].

Although DNA methylation is a major repressive chromatin modification established in RdDM, it is not the only one. Our recent work implicated Pol V-produced lncRNA in nucleosome positioning by indirectly recruiting a SWI/SNF chromatin remodeling complex via the IDN2 protein [15]. Being involved in the establishment of DNA methylation and repressive histone modifications as well as changes in nucleosome positioning, Pol V-produced lncRNA has broad effects on chromatin status and has been proposed to control genome activity [3].

Discovery of loci under control of Pol V-produced lncRNAs relies on finding genomic regions bound by Pol V [16,17]. An additional approach is identifying binding sites of proteins recruited by Pol V transcripts [5]. Further investigation of the functional significance of those lncRNAs requires the ability to directly detect their presence and study their interactions with proteins [7,9,15,16,17]. We present methods, which allow the study of Pol V produced lncRNA as well as transcripts produced by other RNA

polymerases. Although methods we show have only been tested with plant tissue, they should be applicable to the wide array of eukaryotic organisms.

Chromatin Immunoprecipitation (ChIP)

One of the main mechanisms used by lncRNA to control genome activity is by guiding proteins to specific genomic loci [3]. Detecting where these proteins are bound can give insights into the function of lncRNA [9,13]. This approach may also be used in conjunction with high-throughput sequencing in order to discover new loci that are impacted by lncRNA [5,16,17]. Moreover, when performed in different mutant backgrounds, protein-DNA interaction assays may be used to further study molecular mechanisms involving these proteins.

Binding of specific proteins to DNA may be detected using Chromatin Immunoprecipitation (ChIP) [18]. In this method proteins and associated nucleic acids are precipitated using an antibody specific towards the protein of interest. DNA is further quantified and recovery of DNA above the background level is evidence of protein-DNA interaction.

An important step in ChIP is crosslinking with formaldehyde, which covalently fixes protein-DNA interactions but also makes it difficult to differentiate between direct and indirect interactions. The specificity of particular antibodies may cause variability between experiments studying different proteins; it is advisable to optimize the amount of antibody used for each protein studied. Additionally, other key points, such as formaldehyde concentration or DNA fragmentation intensity, may be optimized depending on the protein studied. This method outlines basic procedures for crosslinking, chromatin isolation, immunoprecipitation, and DNA isolation (Figure 8.1), and has been successfully used for several proteins that bind chromatin in a way dependent on lncRNA [5,7,9,13,15,16].

Crosslinking

Successful ChIP usually requires a relatively large amount of starting material, thus 3g of approximately 2.5 weeks old *Arabidopsis thaliana* seedlings are used for each sample. Crosslinking is performed with formaldehyde [19]. In order to limit spontaneous decrosslinking, samples are kept on ice as much as possible. Samples are

first washed once with ultrapure water to remove contaminants. Enough formaldehyde is added to completely cover the tissue and a vacuum is applied to infiltrate plants. To neutralize formaldehyde, glycine is added, mixed, and the plants are exposed to vacuum again. Samples are placed on ice and washed twice in ultrapure water. To make grinding easier, samples are lightly squeezed between paper towels in order to remove excess water and then frozen in liquid nitrogen. Samples may be stored at -80°C.

Chromatin Isolation

Frozen crosslinked plant material can now be directly used to isolate chromatin. Each step should be performed on ice. Samples are ground in liquid nitrogen to a fine powder and resuspended gently in ice cold Honda buffer (supplemented with DTT, PMSF, and Plant Protease Inhibitors). The resuspended powder is filtered through two layers of Miracloth, and the flow-through is collected in an empty tube. To increase yield, the used Miracloth is washed in Honda buffer and the resulting solution is filtered through a clean Miracloth and the two filtrates combined. Nuclei are collected by centrifugation and washed several times with cold Honda buffer to remove cellular debris. After washes, the nuclei are resuspended in Nuclei Lysis Buffer.

Nuclei are lysed and chromatin is fragmented by sonication on ice to an approximate average DNA length of between 250 and 500 bp. The fragmentation intensity should be experimentally optimized for a specific sonicator instrument prior to performing ChIP experiments (See Fig. 8.2A). Sonication optimization may be performed by skipping immunoprecipitation (steps 23-33) in our ChIP protocol and by checking the DNA fragment sizes by gel electrophoresis. Average fragment length of sonicated DNA is an important variable since it limits resolution of the entire ChIP assay.

After fragmentation, nuclear debris is pelleted by centrifugation and the supernatant containing fragmented chromatin is kept. The large amount of starting material means that aliquots can be taken and kept at -80°C after flash freezing in liquid nitrogen, providing several ChIP experiments from one chromatin isolation step. A smaller aliquot from each sample should also be made and stored for input controls.

Immunoprecipitation and Elution of DNA

Immunoprecipitation is done with magnetic beads conjugated to protein A (Dynabeads Protein A). In preparation for immunoprecipitation, the magnetic beads are washed and resuspended in B/W buffer supplemented with PMSF. All steps are performed on ice or at 4°C. Chromatin aliquots are thawed on ice and diluted with ChIP Dilution Buffer to reduce the concentration of SDS. The prepared magnetic beads are added, as is the antibody directed against the protein of interest. Immunoprecipitation is performed at 4°C rotating overnight. In addition to samples with antibody, samples without antibody can be included to determine the nonspecific background pulled-down directly by the beads.

The next day, the inputs are thawed on ice. Meanwhile, the beads are washed in B/W Buffer to remove non-precipitated chromatin. Another wash with TE is performed during which the beads are transferred to new tubes in order to reduce background by DNA bound nonspecifically to the tubes. After the washes, crosslinked protein-DNA complexes are eluted in Elution Buffer at 65°C. Inputs are also prepared by adding a small amount of chromatin to Elution Buffer. After transferring the eluate to a new tube, a second elution step is done to increase the yield of recovered DNA, and the eluates are combined. Inputs and ChIP samples are digested with Proteinase K at 60°C overnight. Extended heat treatment also reverses formaldehyde-induced crosslinks, which allows subsequent purification of DNA.

DNA Isolation

DNA is purified by phenol extraction. After centrifugation the aqueous phase (top layer) is transferred to a new 1.5ml tube. Subsequent extraction with chloroform-isoamyl alcohol helps to remove traces of phenol in the sample, which is essential if samples are used for library preparation. DNA is precipitated with 96% Ethanol in the presence of GlycoBlue (or a different carrier compatible with downstream applications) and washed with 70% Ethanol. The supernatant is carefully removed and, after air drying, the pellet is resuspended in water.

Alternately, if samples are to be used solely for real-time PCR analysis, 10% Chelex may be added directly to the magnetic beads after the final wash with TE [20]. Chelex should be added using a pipette tip cut at the 100µl mark while mixing the Chelex in between the addition to each sample. The tip is cut to ensure equal amounts

of beads are added to each sample. Samples can then be incubated at 99°C for ten minutes. Proteinase K is then added and samples are incubated at 65°C for 2 hours followed by 95°C for ten minutes. Samples should be centrifuged at maximum speed for 1 minute and the supernatant taken to avoid carryover of Chelex before PCR analysis.

Analysis and Interpretation

Purified immunoprecipitated DNA may be analyzed by real-time PCR. Signal levels are usually calculated relative to input for every specific locus using a relative delta-cT method [21,22]. Several controls are required to conclusively demonstrate protein binding to DNA.

1. Background level control shows signal that does not originate from specific antibody-epitope interactions. If a protein-specific antibody is used, a knock-out mutant line entirely lacking the protein of interest should be used as a background level control (Figure 8.2A – *nrpe1*). Alternatively, when an epitope tag-specific antibody is used to detect a tagged protein, a line which does not express the epitope-tagged protein also provides reliable background level information.

2. Loading control shows if technical issues arise during sample preparation. This has to be a locus where signal levels should be identical between samples. It can be beneficial to use more than one loading control locus. This control can either be loci where the protein does not bind (Figure 8.2B – *ACTIN2*) or loci where the protein is known to bind in all samples. Signal levels from the loading control may sometimes be used for further normalization of ChIP results.

3. No antibody control, in which ChIP is performed without an antibody or by using non-specific antibodies. It shows signal that does not originate from antibody interactions but is non-specifically carried over on the beads. Because most antibodies show non-specific interactions, this control often underestimates background levels and should not be used as a replacement to a background level control described above.

4. Positive control, which is a locus tested in all biological samples and based on pre-existing information is known to be bound by the protein of interest. This control is necessary if samples will be tested using high-throughput sequencing.

Protein binding to DNA is conclusively demonstrated if specific signal is detected above the background level signal and loading controls show no significant differences

between biological samples (Figure 8.2B). Additionally, specific signal should be significantly higher than no antibody controls and a positive control should also demonstrate signal above the background level. Every ChIP experiment should be performed in at least three independent biological repeats.

High-throughput sequencing

Purified immunoprecipitated DNA, which has been successfully tested on specific loci using real-time PCR, may be used to discover new loci bound by the protein of interest by high-throughput sequencing. Samples used for high-throughput sequencing should have high signal to background proportion on known binding sites, ideally higher than 8x. If this proportion is lower, data analysis may be difficult and therefore the ChIP protocol should be optimized to increase signal to background ratio. Sequencing libraries can be generated using the Illumina ChIP-seq Library Preparation protocol. Detailed description of library generation is beyond the scope of this paper and usually is performed by specialized facilities.

RNA-Immunoprecipitation (RIP) of Pol V Transcripts

Using a similar strategy as ChIP, protein-RNA interactions can be detected. When applied to studying interactions between proteins and Pol V-produced non-coding transcripts they provide a link between a protein's effect on chromatin and the ability to bind lncRNA. As for ChIP, binding is manifested as enrichment between conditions/genotypes when analyzed by real-time RT-PCR. Protein-RNA interactions are covalently fixed using reversible chemical crosslinking. RIP follows the same steps as ChIP: crosslinking, chromatin preparation and fragmentation, immunoprecipitation, and nucleic acid isolation (Figure 8.1). Differences include enzymatic elimination of DNA and precautions against loss of RNA.

This approach shares all the limitations of ChIP, including limited resolution and inability to conclusively distinguish direct from indirect interactions. It is however much more straightforward than CLIP [23] and may successfully be applied for specific questions, especially if proper controls are available. The protocol described below has been applied to study interactions of Pol V-produced lncRNAs with several proteins [9,12,15].

Procedure

Crosslinking is performed exactly as described for ChIP and crosslinked tissue can be used interchangeably between these methods. Chromatin isolation is also performed as described for ChIP with the addition of an RNase inhibitor to most buffers to prevent RNA degradation. Immunoprecipitation is also performed as in ChIP, with a few minor modifications. Inputs are prepared by adding undiluted chromatin to Elution Buffer, and the following steps are performed with both inputs and precipitated samples. Two elution steps are performed to retrieve RNA, the first is performed at room temperature for ten minutes, while the second is performed at 65°C for ten minutes. Proteinase K digestion is done at 55°C for one hour 15 minutes. RNA is isolated by extraction with acidic phenol:chloroform:isoamyl alcohol (pH 4.3) followed by centrifugation and ethanol precipitation as in ChIP. After final wash in 70% ethanol it is critical to remove as much of the supernatant as possible and air dry the pellet for as short as possible (preferably 2 min.).

Analysis and Interpretation

Using the reverse transcription method, precipitated RNA should immediately be converted to cDNA (after DNase I digestion) and then analyzed by real-time RT-PCR. Evidence of binding to RNA can be established based on controls as described for ChIP. If applied to studying a specific class of RNA, it is important to include an additional control, which allows determining background interaction of non-specific RNAs. In the case of Pol V-produced lncRNA this control may be a knock-out mutant in a Pol V subunit, which lacks Pol V transcripts and any detected signal may be attributed to RNAs produced by other RNA polymerases. A critical control is RT-PCR without reverse transcriptase (no RT) to detect DNA contaminations, which are the most common technical problem in RIP. An additional consideration is that in contrast to ChIP, signal levels in RIP may vary dramatically between tested loci due to differences in transcription levels. It is therefore possible that background signal levels detected at an unrelated locus may be much higher than specific signal. Input levels can provide additional information on RNA levels, however RNA obtained from input samples is often difficult to amplify. Normalization to inputs may be applied during analysis as shown in Figure 8.3A. Due to potential quality issues and variations in total RNA levels

between samples, inputs should be examined before deciding to normalize. For example, Pol V produced lncRNAs are eliminated in the Pol V mutant (*nrpe1*) making normalization to inputs in *nrpe1* impossible. For comparing independent biological replicates it is often necessary to normalize to wild type controls.

Real-Time RT-PCR of low abundance lncRNA

Following RIP, RNA is converted to cDNA and amplified by real-time PCR. In addition to analyzing RIP, this method may be used for detection of lncRNAs in total RNA samples. Long non-coding RNAs, including Pol V transcripts, are generally transcribed at low levels (some approximately 10,000 fold lower than *ACTIN2*), causing them to be difficult to detect. This method is sensitive enough to reliably and quantitatively detect Pol V transcripts and other low abundance RNAs. The protocol involves a DNase digestion to eliminate DNA, reverse transcription and real-time PCR. We recommend using random oligonucleotides as primers, however if strand-specificity is required, locus-specific primers should be used. The method we describe has been used to detect and quantify several Pol V-produced lncRNAs [5,15].

RNA Isolation and cDNA Synthesis

For detection of Pol V transcripts from total RNA, RNA isolation is performed with the RNeasy Plant Mini Kit from Qiagen. The optional on-column DNase I digestion step is followed to eliminate contaminating DNA. Alternately, RIP samples are used. Quality of total RNA should be tested using denaturing agarose electrophoresis (Fig. 8.3B) or using a Bioanalyzer. Quality of immunoprecipitated RNA is very difficult to assay directly at this step, since expected amounts of RNA recovered from RIP are below detection thresholds of currently available methods.

One microgram of total RNA is sufficient for each reverse-transcription reaction to reliably detect Pol V transcripts. To further eliminate any DNA contamination, RNA is treated with Turbo DNase in the presence of Ribolock RNase Inhibitor. RNA is then separated into two reactions per genotype: half of the RNA is transferred to a new tube to be used as a no-reverse-transcriptase control (no RT), while the other half of the sample is transferred to a separate tube for the RT reaction. To denature and anneal primers, the RNA is mixed with random primers and dNTPs and incubated at 65°C

followed by at least one minute on ice. For reverse-transcription, DTT, 5 x First Strand buffer, Ribolock, and Superscript III Reverse-Transcriptase are added. In case of the no-RT controls, water is added in place of Superscript III. Reverse transcription is performed at 50°C.

A critical technical issue is avoiding DNA contaminations, especially originating from previously handled PCR products. We recommend using filtered tips, disposable gloves and working in clean workspace to avoid possible contaminations.

Real-Time PCR of Long Non-Coding RNA

For real-time PCR, a small volume of the cDNA is added to a PCR reaction mix, which in addition to a Hot Start Taq Polymerase, corresponding buffer, magnesium, and dNTPs, also contains SYBR Green – a dye specific towards double-stranded DNA which allows real-time quantification of the PCR product. Depending on the real-time PCR instrument being used, it may be also necessary to add an internal reference dye. The reaction is run with one step at 95°C, followed by 40 cycles of 95°C, 55°C, and 72°C. It is recommended to finish each reaction with a melting curve to determine primer specificity and detect possible contaminations contributing to the signal.

Analysis and Interpretation

Obtained results should be analyzed using relative delta-cT method between samples [21,22], which is sufficient for comparing signal levels generated with a primer pair in various biological samples but does not allow quantitative comparison between different primer pairs. If different primer pairs have to be compared, data should be analyzed using a standard curve.

If detecting Pol V transcripts in total RNA, enrichment between wild-type and a mutant in a Pol V subunit (*nrpe1*) indicates presence of lncRNA (Figure 8.3C). This assay may be used to show lncRNA production or quantitatively compare lncRNA levels between genotypes. If RIP samples are used, protein-RNA interactions can be seen.

For proper data interpretation, the possibility of DNA contaminations should be excluded by performing controls without reverse transcriptase. We recommend performing every experiment in three biological repeats and testing every biological repeat with three PCR amplifications. In ChIP and RIP experiments it is common that overall signal levels are variable between independent experiments. If it is the case,

every repeat should be normalized to wild type prior to calculating averages and standard deviations for all available biological repeats.

ChIP-Sequence Analysis

ChIP-seq is performed to identify protein binding sites throughout the genome. In order to generate a list of binding locations, the raw sequencing reads must first be processed in a way that accurately maps to the *Arabidopsis* genome while controlling for quality. After reads have been mapped, enrichment scores can be generated and utilized to obtain genomic coordinates of binding sites (Figure 8.4). Several different algorithms exist for each step and decisions on the use of each algorithm can be difficult; each has advantages and individual decisions must be made specific to the type of protein studied and the question being asked. Presented here is a simple data analysis pipeline based on published work [5], presented in a way which should provide a starting point to biologists and allow further refinement for a specific dataset. We use the peak-calling algorithm, CSAR [24], because it was built specifically around the *Arabidopsis* genome and from our experience has a high rate of discovery with a low number of false positives. Several aligners can be used and our experiences with each have been excellent. For simplicity sake we use SOAP2 [25] because the output format can directly be read into CSAR.

Read Trimming and Alignment

Although sequencing generally produces quality reads, it may be useful to trim any low quality bases from either end. There are several ways to do this, but ConDeTri [26] provides a simple method of trimming N's from the 5' end and low quality bases from the 3' end of reads.

In order to align reads to the genome efficiently, alignment algorithms generally require an index of the reference genome sequence. Alignment packages usually include a separate algorithm for index creation. The *Arabidopsis* genome sequence can be downloaded from the TAIR website (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_chromosome_files/TAIR10_chr_all.fas) and indexed using 2bwt-builder in the SOAP2 package [25].

Index files will be given the suffix ".index". Alignments can now be performed using the SOAP2 aligner.

Peak Detection

Once mapping is complete, binding sites are called based on regional read enrichment in wild-type compared to mutant. Although a knockout mutant represents the true background signal obtained during the ChIP experiment, inputs may also be used to control for sequence or amplification bias during library generation. Peak calling through enrichment scores can be done using CSAR which was built and tested around the *Arabidopsis* genome. First the mapped reads are loaded in R in a format that CSAR is able to recognize. Reads are then artificially extended to the fragmentation length obtained from sonication, and the number of reads at each base is counted for both strands. To remove PCR amplification errors, only unique reads are kept. Scores are calculated by comparing two samples (i.e. wild-type versus mutant) and regions with enrichment are kept. Significant binding sites are determined by establishing an arbitrary cutoff score at a specified false discovery rate (FDR). Generally an FDR of 0.001 provides high quality peaks, but peak lists at various FDR values can be generated. The final list will have chromosomal coordinates, the enrichment scores, and the length of peaks included in the file.

Conclusions

The activity of lncRNA can be studied through a combination of techniques focused on indentifying loci controlled by RdDM. New targets of transcriptional gene silencing can be detected by ChIP-seq for proteins that bind chromatin in a way dependent on lncRNA. These loci can then be directly tested for lncRNA production and protein-RNA interactions. Additionally, by using a combination of these methods in various mutant backgrounds and with antibodies for different proteins, molecular mechanisms of RNA-mediated silencing pathways can be deciphered. As knowledge of lncRNA expands, methods such as these will become increasingly important to study its role in the regulation of genome activity.

Step-by-Step protocols

ChIP and RIP

Note: Use filter tips when working with RNA.

Note: Keep samples on ice as much as possible.

1. Harvest 3g of above ground tissues from 2.5 week old seedlings and place in 50ml tubes on ice with a hole punched in lids with scissors.
2. Rinse once with H_2O.
3. Add 0.5% formaldehyde so that tissue is immersed (up to 35ml).
4. Apply vacuum at 85 kPa (~25in. Hg) below standard atmospheric pressure for 2 minutes. Release and reapply for 8 minutes.

Note: Ensure a hole is placed in each lid or tubes may break from the pressure.

5. Add 1.25ml 2M glycine. Mix and apply vacuum at 85 kPa (~25in.Hg) for 1 minute. Release and reapply for 4 minutes.
6. Rinse twice in H_2O, squeeze dry between paper towels, wrap in aluminum foil, and freeze in liquid nitrogen.
7. Store at -80°C.

Note: Crosslinked tissue can be used for either ChIP or RIP.

8. Prepare 30ml Honda Buffer per sample. For every 30ml Honda Buffer, add 150µl 1M DTT, 300µl 100mM PMSF, and 300µl Plant Protease Inhibitor. For RIP, also add 240u Ribolock RNase Inhibitor.

Note: For all RNA work solutions should be prepared with RNase-free water. We recommend using fresh miliQ water without any additional treatments.

9. Place 15ml and 10ml prepared Honda Buffer in 50ml tubes on ice. Keep leftover buffer on ice as well.
10. Grind frozen crosslinked tissue in liquid nitrogen into a fine powder and resuspend gently in 15ml prepared Honda Buffer. Keep on ice.

Note: Ensure that all large chunks are dissolved and that a fine powder is present in the buffer.

11. Cut two 6cm wide strips of Miracloth and form one into a funnel with two layers between outside and inside of the funnel.
12. Filter the resuspended tissue through funnel into empty 50ml tube.
13. Invert filter and place in prepared 10ml Honda Buffer. Resuspend as much tissue as possible by swirling Miracloth in buffer.

Note: Washing the Miracloth increases chromatin yield.

14. Using the second strip of Miracloth from step 11, form another funnel like the first and filter washed sample from step 13, combining the flow through with the previous one.
15. Centrifuge at 3000 g for 7.5 minutes at 4°C and discard supernatant.
16. Resuspend pellet with 1ml prepared Honda Buffer and transfer to a 1.5ml tube.
17. Centrifuge at 1900 g for 5 minutes at 4°C. Discard supernatant.
18. Wash twice with 1ml prepared Honda Buffer by centrifugation at 1900 g for 5 minutes at 4°C. If the pellet is still green perform one additional washing step.

Note: Use leftover Honda Buffer from step 9 for steps 16 and 18.

19. Resuspend nuclei gently in 550µl freshly prepared cold NLB (with 88U Ribolock for RIP).

20. Sonicate on ice 8 times for 10 seconds with 1 minute pauses at power setting 1. (Fisher Scientific Dismembrator - Adapt number of times based on the instrument used until chromatin is fragmented to approximately 250bp to 500bp on average).

Note: It is important to keep samples on ice during sonication to avoid overheating and possible protein degradation or loss of crosslinks.

21. Centrifuge samples in a microcentrifuge at maximum speed for 10 minutes at 4°C. Transfer supernatant to a new 1.5ml tube.
22. Take 100µl aliquots and flash freeze in liquid nitrogen. Also freeze one 10µl aliquot for inputs. Store at -80°C.

Note. While the preceding steps are specific for plant material, the remainder of the protocol may be applicable to chromatin samples obtained from other organisms.

23. Prepare 5ml B/W buffer by adding 50µl PMSF (and 200U Ribolock for RIP).
24. Using magnetic rack, wash 40µl per sample Dynabeads Protein A three times with 1ml B/W buffer from step 23. Resuspend in 110µl B/W from step 23 per sample.
25. Combine one 100µl aliquot of chromatin, 900µl ChIP Dilution Buffer (with Ribolock for RIP), 100µl washed magnetic beads, and 2.5µg - 5µg antibody.
26. Rotate at 4°C over-night.

Note: Amount of antibody and incubation time can be adjusted depending on the protein studied or the antibody used. In our experience, 2.5µg - 5µg antibody works well for most applications. Magnetic beads may be replaced with Protein A Agarose beads blocked with salmon sperm DNA, which give lower background levels for some antibodies; however, due to the presence of blocking DNA, samples obtained using those beads cannot be used for sequencing.

27. Prepare 5ml per sample B/W buffer by adding 50µl PMSF (and 200U Ribolock for RIP).

28. For ChIP wash the samples three times with 1ml B/W with rotation at 4°C for 5 minutes between washes. Follow with one wash in TE.
For RIP wash twice with 1ml B/W by inverting, magnetic separation, and removal of supernatant, followed by twice with 1ml B/W and rotating at 4°C for 5 minutes between washes.

Note: Use a magnetic separator buried in ice during washes to keep samples cold.

29. After the last wash transfer samples to new 1.5ml tubes, remove supernatant, and add 55µl Elution Buffer (with 22U Ribolock for RIP).
30. Prepare 1% input samples by adding 1µl chromatin extract to 110µl Elution Buffer (with 44U Ribolock for RIP).
31. Vortex all samples briefly and incubate them at 65°C for 20 minutes for ChIP (room temperature for 10 minutes for RIP).

Note: Vortexing vigorously helps elute the sample from the beads.

32. Place on magnetic separator, transfer supernatant to a new 1.5ml tube and add 55µl Elution Buffer to beads (with 22U Ribolock for RIP).

Note: Input samples do not need to be transferred.

33. Incubate at 65°C for 20 minutes (10 minutes for RIP), vortex, place on magnetic separator, and transfer supernatant to previous eluate.
34. Add 20µg RNA-grade Proteinase K and incubate at 60°C overnight for ChIP (55°C for 1 hour 15 minutes for RIP).
35. Add 110µl room temperature phenol:chloroform:isoamyl-alcohol 25:24:1 (pH 7.5 for ChIP, pH 4.3 for RIP).
36. Vortex well and centrifuge at maximum speed for 5 minutes.
37. Transfer 100µl aqueous phase to new 1.5ml tube.
38. For ChIP repeat steps 35-37 with 24:1 chloroform:isoamyl-alcohol.

39. Add 60µg GlycoBlue (or an equivalent amount of a different carrier compatible with downstream applications), 10µl 3M sodium acetate pH 5.3, and 250µl 96% ethanol.
40. Incubate at -80°C for 2 hours (overnight for RIP).
41. Centrifuge at maximum speed for at least 30 minutes at 4°C. Discard supernatant.
42. Add 300µl 70% ethanol and centrifuge at maximum speed for 10 minutes at 4°C. Discard supernatant. If traces of supernatant are left on tube walls, centrifuge again briefly and remove leftovers of the supernatant.
43. Air dry pellet for 2 minutes and resuspend in 30µl H_2O for ChIP (12µl RNase-free H_2O for RIP). ChIP samples may be stored for up to a year at -20°C, RIP samples may be stored at -80°C for a couple of days, however due to low stability of RNA should be processed as soon as possible.
44. ChIP DNA can be checked for enrichment of DNA levels between samples by real-time PCR or used for library generation and sequencing. RIP RNA can be reverse transcribed and used in quantitative PCR as in A.2.

Note: Alternately, if ChIP DNA will not be used for sequencing, steps 29-43 can be replaced with the following protocol using Chelex. Cut a pipette tip at 100µl and add 100µl 10% Chelex-100 to the beads. Prepare 1% inputs by adding 1µl of chromatin to 100µl 10% Chelex-100. Vortex well and incubate at 99°C for ten minutes. Cool samples to 65°C and add 20µg Proteinase K, incubating at 65°C for two hours. Deactivate Proteinase K at 95°C for ten minutes and spin samples at max speed for 1 minute. Supernatant can be diluted and analyzed by real-time PCR as in step 44.

Real-Time RT-PCR of low abundance lncRNA

Note: Use filtered tips, disposable gloves and clean workspace when working with RNA.

1. Isolate RNA using the Qiagen RNeasy Plant Mini Kit following instructions with the optional on-column DNase digestion. RIP samples from A.1 can also be used.

Note: Total RNA may be stored at -80°C up to 2 weeks.

2. To 1-5µg total RNA or the entire RIP sample from A.1 step 43, add 3units Turbo DNase, 24 units Ribolock, and H_2O up to 12µl.
3. Incubate at 25°C 30 minutes.
4. Add 3µl 25mM EDTA.
5. Incubate at 65°C 10 minutes.
6. Transfer 1/2 RNA to a new tube for no-RT control and transfer 1/2 RNA to a new tube for the RT reaction.
7. Add 0.4µl 500ng/µl random primers (Invitrogen) and 1µl 10mM dNTPs.
8. Incubate at 65°C 5 minutes. Keep on ice for at least 1 minute.
9. Add 1µl 0.1M DTT, 4µl 5x First Strand Buffer, 1µl 40U/µl Ribolock, and 1µl Superscript III Reverse Transcriptase (substitute water for Superscript III in no-RT controls). Add H_2O to a total volume of 20µl.
10. Incubate 25°C for 5 minutes, followed by 50°C for 1 hour, then 70°C for 15 minutes.

Note: cDNA may be stored at -20°C.

11. To 1µl cDNA add 2.5µl 10x Platinum Taq PCR buffer, 1.2µl 50mM $MgCl_2$, 0.5µl 10mM dNTPs, 0.25µl 25x SYBR Green, 0.2µl 25µM forward and reverse primer mix, 0.1µl Platinum Taq, and H_2O to a total volume of 25µl.

Note: 25X SYBR Green is prepared before hand by diluting 1:400 with H_2O and stored at 4°C in the dark.

12. Run PCR plate at 95°C for 2 minutes, followed by 40 cycles of 95°C for 15 seconds, 55°C for 20 seconds, 72°C for 45 seconds.

Simple ChIP-seq Data Analysis Pipeline

Note: Commands for each step are italicized. Ensure all software has been successfully downloaded and installed including Condetri, SOAP2, and CSAR.

Note: Perform steps 1 and 3 as well as 7 and 8 for each genotype/sample.
Note: Each command contains file names that will change depending on user preference. This example uses the name "wt.fastq" for the file obtained from sequencing and "TAIR10_chr_all.fas" for the genome reference file downloaded from the TAIR website.

1. Trim reads: *perl condetri_v2.2.pl -fastq1=wt.fastq -prefix=wt -rmN*

Note: wt.fastq contains the reads and quality scores from sequencing. The prefix parameter "-prefix=" is used to name the output file and should be changed with each sample. In step1 the output file will be named wt_trim.fastq. The parameter "-rmN" removes N base calls from the beginning of reads.

2. Build Genome Index (only needs to be done once): *./2bwt-builder TAIR10_chr_all.fas*
3. Align Reads to Genome: *./soap -a wt_trim.fastq -D TAIR10_chr_all.fas.index -o wt.soap*

Note: wt_trim.fastq is the file containing the trimmed reads from step 1. wt.soap is the name of the output file for this step.

4. Repeat Steps 1 and 3 for each sample changing the names of the appropriate files.
5. Start R: *R*
6. Load CSAR package: *> library(CSAR)*

Note: The symbols > or + denotes a new line and are displayed by R.

7. Load Mapped Reads for each genotype: *> wt <- loadMappedReads("wt.soap", format = "SOAP", header = FALSE)*

Note: In R results from each command are saved into temporary variables using the "<-" symbols. The results from step seven are saved into "wt" and these results are called upon in step eight.

8. Extend reads to fragmentation length, eliminate PCR amplification errors, and count reads: > *wt.nhits <- mappedReads2Nhits(wt, file = "wt.nhits" , chr = c("1", "2", "3", "4", "5"), chrL = c(30427671, 19698289, 23459830, 18585056, 26975502), w = 250L, considerStrand = "Sum", uniquelyMapped = TRUE)*

Note: Parameter w is the fragmentation length after sonication and artificially extends the reads to this length. considerStrand = "Sum" is used to count all the reads regardless of the strand they map to, but can be changed to "Minimum" if only reads with equal numbers on both strands are desired. uniquelyMapped is used to remove duplicate reads due to PCR amplification errors during library preparation. Each chromosome's length is specified by the chrL parameter.

9. Repeat steps 7 and 8 with each sample changing the names "wt", "wt.soap", and "wt.nhits" to "mutant", "mutant.soap", and "mutant.nhits" respectively.
10. Compare genotypes and calculate enrichment scores: > *score <- ChIPseqScore(mutant.nhits, wt.nhits)*
11. Create windows of enrichment between genotypes (these windows will later be filtered based on false discovery rate (FDR)): > *TotalPeaks <- sigWin(score)*
12. Run each of the following to perform permutations in order to calculate score cutoff based on FDR.

>dir.create("permutations")
>setwd("permutations")
> for (j in 1:20) {

+ permutatedWinScores(nn = j, mutant, wt, fileOutput = "perm", chr = c("1","2","3","4","5"), chrL = c(30427671, 19698289, 23459830, 18585056, 26975502), w = 250L, considerStrand = "Sum", uniquelyMapped = TRUE)
+ }
>permscores <- getPermutatedWinScores(file = "perm", nn=1:20)
>cutoff <- getThreshold(winscores=TotalPeaks$score, permutatedScores=permscores, FDR=.001)
>setwd("../")
13. Use cutoff value to create list of significant enrichment windows: *>FilteredPeaks <- TotalPeaks[TotalPeaks$score > cutoff$threshold,]*

Note: Step 12 uses a cutoff FDR of 0.001which is somewhat strict. Changing the FDR value in step 12 will change the sensitivity and specificity of peaks in the final list.

14. Write file with list of peaks: *>write.table(FilteredPeaks, sep="\t", file="wtvsmutant.txt", row.names=FALSE)*

Note: Exit R and examine the file using a word processor. The list of peaks is saved as "wtvsmutant.txt" and includes peaks where the FDR is less than 0.001. This list includes the coordinates of binding sites along with the position of the peak summit, the peak enrichment score, and the length of the binding region.

Equipment, Reagents and Buffers

Equipment

Vacuum chamber (ChIP and RIP), Isotemp Vacuum Oven Model 280A (Fisher Scientific), self-cleaning dry vacuum system (Welch)
Refrigerated centrifuge (ChIP and RIP), Thermo Scientific Sorvall Legend X1 Centrifuge, Eppendorf centrifuge 5424
Sonicator (ChIP and RIP), Fisher Scientific, Sonic Dismembrator Model 100
Rotator (ChIP and RIP), Thermo Scientific Labquake Tube Rotator

Magnetic separator (ChIP and RIP), Promega MagneSphere Technology Magnetic Separation Stand

Real-Time Thermal Cycler (ChIP, RIP, and RT-PCR), Biorad CFX Connect Real-time System

Reagents

DTT (ChIP, RIP, and RT-PCR), Fisher Scientific, BP172-5

Formaldehyde 0.5% (ChIP and RIP), Sigma-Aldrich, F1635-500ML

Glycine 2M (ChIP and RIP), Fisher Scientific, G46-1

PMSF 100mM (ChIP and RIP), Sigma-Aldrich, P7626-25G

Plant Protease Inhibitor [Sigma] (ChIP and RIP), Sigma-Aldrich, P9599-5ML

Miracloth (ChIP and RIP), VWR, 80058-394

Dynabeads Protein A Magnetic Beads [Invitrogen] (ChIP and RIP), Invitrogen, 100-01D

Protein/Tag specific antibody (ChIP and RIP)

RNA-grade Proteinase K 20mg/ml [Invitrogen] (ChIP and RIP), Invitrogen, 25530-049

25:24:1 Phenol-chloroform-isoamyl alcohol pH 7.5 (ChIP), Fisher Scientific, BP1752I-100

25:24:1 Phenol-chloroform-isoamyl alcohol pH 4.3 (RIP), Fisher Scientific, BP1754I-100

24:1 Chloroform-isoamyl alcohol (ChIP), Fisher Scientific (chloroform: C298500, isoamyl alcohol: BP1150-500; mix 24:1)

GlycoBlue 15mg/ml [Ambion] (ChIP and RIP), Fisher Scientific, NC9567599

3M NaOAc pH 5.2 (ChIP and RIP), Fisher Scientific (NaOAc: BP333-500, acetic acid: AC14893-0025)

96% EtOH (ChIP and RIP), Decon Labs Inc., 2701

Turbo DNase 2U/µl [Ambion] (RIP and RT-PCR), Fisher Scientific, NC9075048

Ribolock RNase Inhibitor 40U/µl [Fermentas] (RIP and RT-PCR), Fisher Scientific, FEREO0384

RNeasy Plant Mini Kit [Qiagen] (RT-PCR), Qiagen, 74904

25mM EDTA pH8 (RT-PCR), Fisher Scientific (EDTA: BP120500, NaOH: BP359-500)

Random Primer 500ng/µl [Invitrogen] (RT-PCR), life technologies, 48190-011

dNTPs 10mM [Promega] (RT-PCR), Fisher Scientific, PRU1515

Superscript III Reverse Transcriptase 200U/µl, 5 x FS buffer, 0.1M DTT [Invitrogen] (RT-PCR), life technologies, 18080-044

Platinum Taq 5U/µl, 10x Platinum Taq PCR Buffer, 50mM $MgCl_2$ [Invitrogen] (RT-PCR), life technologies, 10966-034

SYBR Green I Nucleic Acid Gel Stain - 10,000X concentrate in DMSO [Invitrogen] (RT-PCR), life technologies, S-7563

Chelex-100 (ChIP), Bio-Rad 142-1253

Nalgene Rapid Flow Sterile Disposable Filter Unit with PES Membrane [Thermo Scientific] (ChIP/RIP), Fisher Scientific, 09-741-02

Ultra-pure water (miliQ)

Buffers

Honda Buffer

0.44M Sucrose, Fisher Scientific, BP220-212

1.25% Ficoll 400, Sigma-Aldrich, F2637-100G

2.5% Dextran T40, Sigma-Aldrich, D1662-100G

20mM HEPES, KOH pH7.4, Fisher Scientific (HEPES: BP310-100, KOH: P250-1)

10mM $MgCl_2$, Fisher Scientific, M33-500

0.5% Triton X-100, Fisher Scientific, BP151-500

Prepare HEPES first in 350ml and set pH, then add other components.

Make 500ml and filter through 0.2 micron filter.

Store at 4°C.

Add the day of experiment:

5mM DTT, Fisher Scientific, BP1725

1mM PMSF, Sigma-Aldrich, P7626-25G

1% Plant Protease Inhibitors, Sigma-Aldrich, P9599-5ML

8U/ml Ribolock RNase Inhibitor (RIP only), Fermentas, FEREO0384

Binding/Washing Buffer (B/W)

150mM NaCl, Fisher Scientific, BP358-212

20mM Tris HCl pH8, Fisher Scientific (Tris: BP152-5, HCl: A144-500)

2mM EDTA pH8, Fisher Scientific (EDTA: BP120500, NaOH: BP359-500)
1% Triton X-100, Fisher Scientific, BP151-500
0.1% SDS, Fisher Scientific, BP166-500
Add water to 500ml and filter through 0.2 micron filter.
Store at 4°C.

Add the day of experiment:
1mM PMSF, Sigma-Aldrich, P7626-25G
40U/ml Ribolock RNase Inhibitor (RIP only), Fermentas, FEREO0384

Nuclei Lysis Buffer (NLB)
50mM Tris-HCl pH8, Fisher Scientific (Tris: BP152-5, HCl: A144-500)
10mM EDTA, Fisher Scientific (EDTA: BP120500, NaOH: BP359-500)
1% SDS, Fisher Scientific, BP166-500

Add the day of the experiment:
1mM PMSF, Sigma-Aldrich, P7626-25G
1% Plant Protease Inhibitors, Sigma-Aldrich, P9599-5ML
160U/ml Ribolock RNase Inhibitor (RIP only), Fermentas, FEREO0384

ChIP-Dilution Buffer
Note: Tris-HCl and EDTA are added from filtered stock solutions.
1.1% Triton X-100, Fisher Scientific, BP151-500
1.2mM EDTA pH8, Fisher Scientific (EDTA: BP120500, NaOH: BP359-500)
16.7mM Tris-HCl pH8, Fisher Scientific (Tris: BP152-5, HCl: A144-500)
167mM NaCl, Fisher Scientific, BP358-212
350U/ml Ribolock RNase Inhibitor (RIP only), Fermentas, FERE00384
Store at 4°C for ease of use.

Elution Buffer
Note: Tris-HCl and EDTA are added from filtered stock solutions.

100mM Tris HCl pH8, Fisher Scientific (Tris: BP152-5, HCl: A144-500)

10mM EDTA, Fisher Scientific (EDTA: BP120500, NaOH: BP359-500)

1% SDS, Fisher Scientific, BP166-500

Add the day of the experiment:

400U/ml Ribolock RNase Inhibitor (RIP only) Fermentas, FERE00384

TE Buffer

10mM Tris-HCl pH8, Fisher Scientific (Tris: BP152-5, HCl: A144-500)

1mM EDTA, Fisher Scientific (EDTA: BP120500, NaOH: BP359-500)

List of Primers

IGN29: CGTTTGTTTATGTAGGGCGAAAG and TAAAACTTTTCCCGCCAACCA.

ACTIN2: GAGAGATTCAGATGCCCAGAAGTC and TGGATTCCAGCAGCTTCCA.

Acknowledgements

We thank Brian Gregory and Qi Zheng for sharing expertise in high throughput sequencing data analysis. This work has been supported by National Science Foundation grant MCB 1120271 to Andrzej Wierzbicki. and Austrian Science Fund (FWF) fellowship J3199-B09 to Gudrun Böhmdorfer.

References

1. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81,** 145–166 (2012).
2. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11,** 204–220 (2010).
3. Wierzbicki, A. T. The role of long non-coding RNA in transcriptional gene silencing. *Curr. Opin. Plant Biol.* **15,** 517–522 (2012).
4. Matzke, M., Kanno, T., Daxinger, L., Huettel, B. & Matzke, A. J. M. RNA-mediated chromatin-based silencing in plants. *Curr. Opin. Cell Biol.* **21,** 367–376 (2009).
5. Zheng, Q. *et al.* RNA polymerase V targets transcriptional silencing components to promoters of protein-coding genes. *Plant J.* (2012). doi:10.1111/tpj.12034
6. Wei, W. *et al.* A role for small RNAs in DNA double-strand break repair. *Cell* **149,** 101–112 (2012).
7. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135,** 635–648 (2008).

8. Haag, J. R. & Pikaard, C. S. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nat. Rev. Mol. Cell Biol.* **12,** 483–492 (2011).
9. Wierzbicki, A. T., Ream, T. S., Haag, J. R. & Pikaard, C. S. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat. Genet.* **41,** 630–634 (2009).
10. El-Shami, M. *et al.* Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components. *Genes & Development* **21,** 2539–2544 (2007).
11. Haag, J. R. *et al.* In vitro transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. *Mol. Cell* **48,** 811–818 (2012).
12. He, X.-J. *et al.* An effector of RNA-directed DNA methylation in arabidopsis is an ARGONAUTE 4- and RNA-binding protein. *Cell* **137,** 498–508 (2009).
13. Rowley, M. J., Avrutsky, M. I., Sifuentes, C. J., Pereira, L. & Wierzbicki, A. T. Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. *PLoS Genet.* **7,** e1002120 (2011).
14. Bies-Etheve, N. *et al.* RNA-directed DNA methylation requires an AGO4-interacting member of the SPT5 elongation factor family. *EMBO Rep.* **10,** 649–654 (2009).
15. Zhu, Y., Rowley, M. J., Böhmdorfer, G. & Wierzbicki, A. T. A SWI/SNF Chromatin-Remodeling Complex Acts in Noncoding RNA-Mediated Transcriptional Silencing. *Mol. Cell* **49,** 1–12 (2013).
16. Wierzbicki, A. T. *et al.* Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. *Genes Dev.* **26,** 1825–1836 (2012).
17. Zhong, X. *et al.* DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. *Nat. Struct. Mol. Biol.* **19,** 870–875 (2012).
18. Kuo, M. H. & Allis, C. D. In vivo cross-linking and immunoprecipitation for studying dynamic Protein:DNA associations in a chromatin environment. *Methods* **19,** 425–433 (1999).
19. Wong, S. S. *Chemistry of protein conjugation and cross-linking*. (CRC Press, 1991).
20. Nelson, J. D., Denisenko, O., Sova, P. & Bomsztyk, K. Fast chromatin immunoprecipitation assay. *Nucleic Acids Res.* **34,** e2 (2006).
21. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25,** 402–408 (2001).
22. Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc* **3,** 1101–1108 (2008).
23. König, J., Zarnack, K., Luscombe, N. M. & Ule, J. Protein-RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet.* **13,** 77–83 (2011).
24. Muiño, J. M., Kaufmann, K., van Ham, R. C., Angenent, G. C. & Krajewski, P. ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. *Plant Methods* **7,** 11 (2011).
25. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).

26. Smeds, L. & Künstner, A. ConDeTri--a content dependent read trimmer for Illumina data. *PLoS ONE* **6,** e26314 (2011).

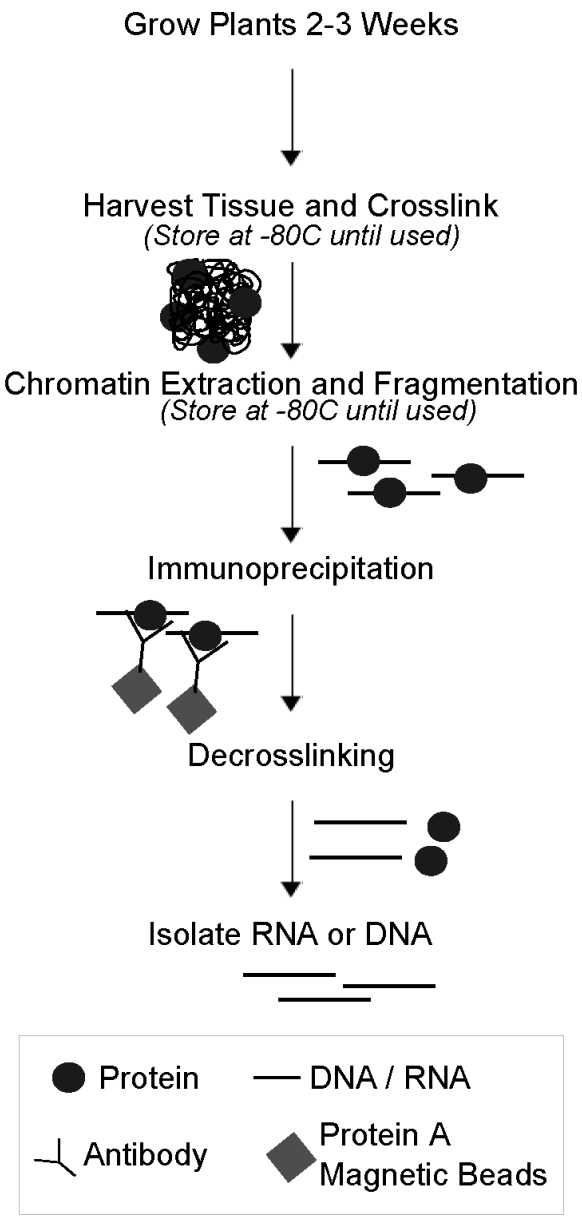


Figure 8.1 Overview of Chromatin-Immunoprecipitation (ChIP) and RNA-Immunoprecipitation (RIP)
DNA or RNA (black lines) is crosslinked to proteins (circles). Chromatin is extracted and fragmented (straight lines with circles). Fragments bound by the protein of interest are precipitated by an antibody (y-shaped) and magnetic beads (diamonds). RNA or DNA (straight lines) is purified from proteins (circles) and tested by real-time PCR.

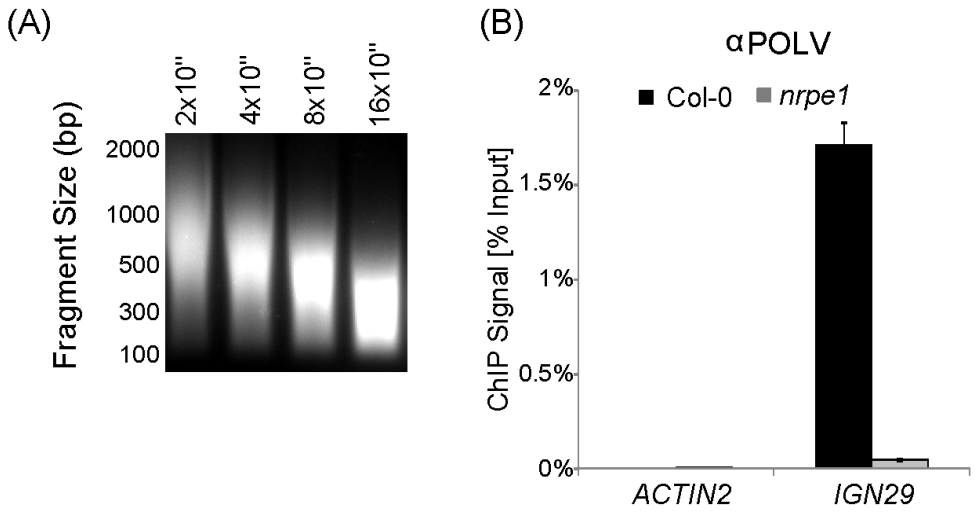


Figure 8.2 ChIP

(A) DNA fragmentation after sonication. Fragmentation can be checked by gel electrophoresis after decrosslinking. Optimal fragmentation can be determined by varying the sonication time as shown.

(B) ChIP-based detection of DNA bound by NRPE1, the largest subunit of Pol V. DNA binding is seen as enrichment in Col-0 wild-type compared to *nrpe1* mutant (which serves as a background level control) at the *IGN29* locus. *ACTIN2* serves as an unbound loading control to check specificity of immunoprecipitation. Bars represent average real time PCR signal normalized to inputs in Col-0 wild-type (black) and *nrpe1* (gray). Error bars represent standard deviations from three PCR amplifications.

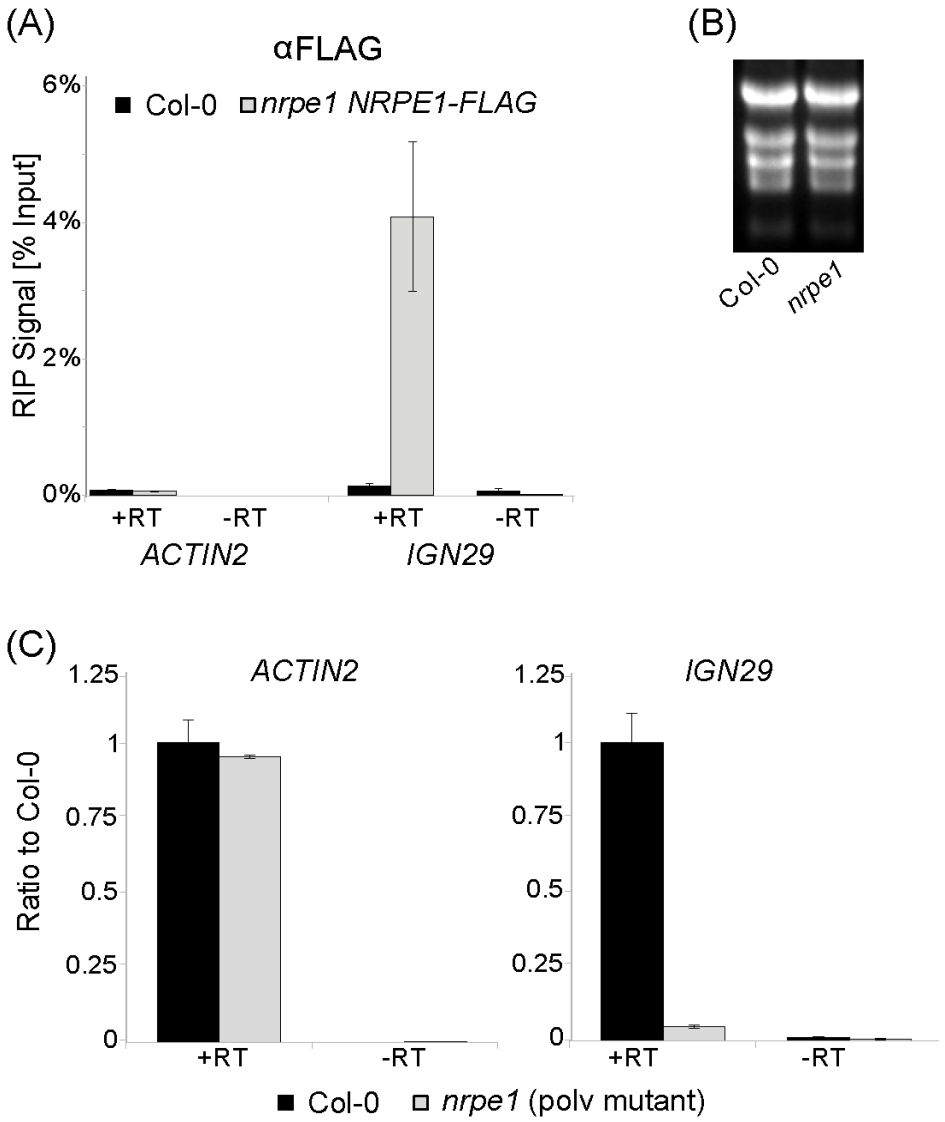


Figure 8.3 RIP and total RNA RT-PCR

(A) RIP-based detection of RNA bound by Pol V. RNA binding is seen by enrichment between FLAG tagged Pol V subunit (*nrpe1 NRPE1-FLAG*) compared to Col-0 wild-type (which serves as a background control) at *IGN29* [15]. *ACTIN2* serves as an unbound loading control to check efficiency of immunoprecipitation. Samples without reverse transcriptase (-RT) are used to check for signal resulting from DNA contamination. Bars represent average real time RT-PCR signal normalized to inputs. Error bars represent standard deviations from three PCR amplifications.

(B) Quality of RNA used in RT-PCR. RNA quality may be seen by gel electrophoresis as shown.

(C) Non-coding transcripts are seen by enrichment between Col-0 wild-type and the Pol V mutant (*nrpe1*) at *IGN29* [15]. *ACTIN2* serves as a loading control. Samples without reverse transcriptase (-RT) are used to check for signal resulting from DNA contamination. Bars represent average RNA signal as a ratio of mutant (*nrpe1* – gray) to Col-0 wild-type (black). Error bars represent standard deviations from three PCR amplifications.

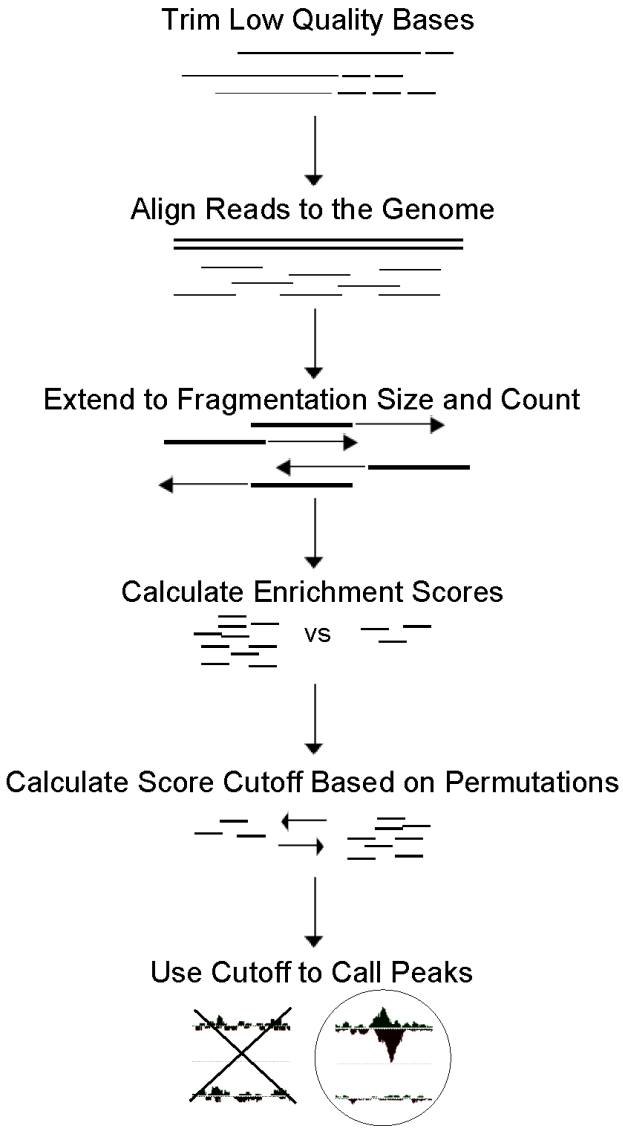


Figure 8.4 ChIP-sequencing analysis
Analysis of sequencing reads from ChIP involves quality control, mapping, and peak calling. Low quality reads and bases (dashed lines) are removed from analysis to improve alignment. Remaining reads are aligned to the genome. Reads are extended to the fragmentation size obtained in sonication to more precisely map binding sites (arrows). Enrichment scores and permutations are used to obtain statistical significance for enriched genomic regions.