FUNCTION ANALYSIS OF THE ARABIDOPSIS EPIGENOME THROUGH INTEGRATING GENOME-WIDE PROFILES AND CELL-TYPE-SPECIFIC APPROACHES

by

CHONGYUAN LUO

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Plant Biology

written under the direction of

Eric Lam

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

[May, 2012]

ABSTRACT OF THE DISSERTATION

FUNCTION ANALYSIS OF THE ARABIDOPSIS EPIGENOME THROUGH INTEGRATING GENOME-WIDE PROFILES AND CELL-TYPE-SPECIFIC APPROACHES

By CHONGYUAN LUO

Dissertation Director:
Eric Lam

The eukaryotic epigenome is a dynamic ensemble of chromatin modifications and chromatin structure variations. Its complexity demands experimental methods for the global profiling of epigenomic features as well as informatics tools that facilitate data analyses. I developed Chromatin Immunoprecipitation coupled with high-throughput-sequencing (ChIP-seq) to interrogate histone modifications in the model plant *Arabidopsis thaliana*. Using the ChIP-seq and RNA-seq methods, I produced profiles of nine histone modifications and the transcriptome from aerial tissues of mature plants. With ANchored CORrelative Pattern (ANCORP) method, our analysis has delineated 42 chromatin states with distinct chromatin modification patterns. Selected states were tested by re-ChIP assay to validate the co-localization of histone modifications. I identified the enrichment of Gene Ontology, microRNAs and transposable elements in certain chromatin states, which suggests the regulation of these loci with chromatin-related mechanisms. To derive hypotheses regarding the interactions between the epigenome and transcriptome, State-Specific-Effects-Analysis (SSEA) was developed to incorporate

quantitative information related to chromatin states to improve the sensitivity of correlative analyses. Combining ANCORP and SSEA, I identified correlations between the quantity of Natural Antisense Transcripts (NATs) and the enrichment of H3K36me2 and 5mC marks. Genetic analyses identified Polymerase Associated Factors as potential regulators for NAT abundance. I further observed evidence for both nuclear- and cytoplasmically-localized NATs. Although some nuclear-localized NATs are known to regulate chromatin-related functions, our results suggest that NATs may not commonly regulate the cognate locus in cis.

Differentiation of distinct cell types is fundamental to the sophisticated body-plan and lifestyle of multicellular organisms. Therefore studying the epigenome in specific cell types is critical for elucidating the function of epigenetic regulation in plant development. I experimented with two different cell/nucleus isolation techniques – Fluorescence Activated Cell Sorting (FACS) and Isolation of Nuclei TAgged in specific Cell Types (INTACT) for performing epigenomic profiling in specific plant cell types. I showed that Arabidopsis root cells labeled with GFP can be effectively isolated with FACS. I initiated the effort to establish a high-throughput and Gateway-compatible INTACT platform for investigating interactions between key regulators and chromatin modifications in root cell layers. Our strategy and progress in this front will be described.

Acknowledgement

I have had 5-years of exciting time pursuing my Ph.D degree under the direction of Dr. Eric Lam. Dr. Lam has guided me entering the enterprise of biological research by encouraging me to be competitive and taking up the challenge of cutting-edge research. Through his efforts for establishing the epigenome and other projects in the lab, Dr. Lam has taught me an important lesson that successful research projects need to be driven with strong passion, dedication and discipline. Ever since I entered the lab, I was impressed by Dr. Lam's enthusiasm for developing new technologies to address fundamental biological questions as well as the need of the world. This philosophy has substantially influenced my view of biological research and will continue to make impact on my future careers. I am also very much grateful for Dr. Lam's advice that technical developments are meaningful only when they are driven by reasonable scientific aims. Dr. Lam emphasized that having a clear long-term biological question would be most important for a productive scientific career. Although I completely agree with these advice, I may need years of additional experience and perhaps frustrations to come close to the fine balance between the scientific and technical aspects of biological researches.

The experience of working with Dr. Lam was extremely enjoyable. I had much independence and freedom to choose or develop my favorite strategy to solve problems. Dr. Lam is always ready to help and provide critical suggestions in case that I encountered difficulties. With his exceptional hard working, Dr. Lam has helped me through a number of writing projects including manuscripts, meeting abstracts or fellowship applications. Overall, Dr. Lam has shown me, through his daily working style, the characters needed to become a researcher that is capable of tackling the most

challenging biological question in the world. This lesson is equally or perhaps more important than the knowledge and skills that I have learned from Dr. Lam and his lab.

It would not possible for me to complete this dissertation without the help of many others. I deeply appreciate the support given by the members of my Ph.D thesis committee – Dr. Hugo Dooner, Dr. Gojko Jelenkovic and Dr. Marc Gartenberg. I am truly grateful for their critical reviews of my progress, encouragements and career advices. I always feel fortunate that I had the chance working together with Dr. Todd Michael and Dr. Randall Kerstetter for about two years. Dr. Michael encouraged me to apply my informatics skills to address biological questions, which opened up a whole new avenue for my research. Dr. Kerstetter has provided critical help to me in the development of ChIP-seq approach with his exceptional knowledge and experience in high-throughput-sequencing. I enjoyed all the interactions with Dr. Michael Lawton in the past six years. I am particularly grateful that Dr. Lawton invited me to be one of the lecturers for the course 'Concepts in Biotechnology and Genomics'. The experience was very helpful for me to develop my teaching skills. I like to give my special thank to Dr. Faith Belanger. During my rotation in Dr. Belanger's lab, she helped me to begin learning molecular biology and phylogenetic analysis with her extraordinary kindness and patience. I further thank Drs. David Sidote, Remy Bruggmann and Anirvan Sengupta for their generous help in experimental and informatics aspects.

I appreciate the support that I have had from lab members of Drs. Eric Lam, Michael Lawton and Rong Di. Much of my training in plant molecular biology was from Drs. Naohide Watanabe and Faye Rosin, both were members of the Lam lab. I like to thank Drs. Jean Luc Cacas, Anica Amini and Thomas Widiez for their helpful

suggestions and fruitful discussions. In the end, I am deeply thankful for the support and understanding given by my family.

Table of Contents

Lists of tables

List of illustrations

1 Introductions

1.1 Epigenetics

Epigenetics has been a research area attracting much interests in the past two decades. Multiple definitions of epigenetics have been proposed, which are mainly different in the requirement of heritability for the regulatory mechanism (Bird 2007). In addition, certain definitions suggest epigenetic regulations are mainly mediated by chromatin components while others do not. As a term of classical developmental biology, epigenetics describes the study for the generation of numerous cell and tissue types from the single-cell zygote without a change in the DNA sequence (Waddington 1957). This definition was incredibly broad and encompasses all the developmental regulatory mechanisms that do not involve the change of genetic materials. Since the definition was proposed more than 50 years ago, the authors would not be possible to anticipate the significance of chromatin structure in the regulation of development. Some other definitions of epigenetics, such as the one proposed in (Russo et al, 1996), placed a stringent requirement that the regulation needs to be either mitotically or meiotically heritable (Russo et al., 1996). Although recent discussions of epigenetics are commonly related to chromatin structures, chromatin is in fact not necessary for the generation of heritable patterns. It is well known that positive feedback loops can create heritable cellular states. As an extreme example, the switch for lamda phage from lysogenic to lytic growth involves no chromatin structures and was mainly driven by positive feedbacks (Lewin, 2006). In addition, many histone modifications may not be considered as epigenetics marks under the definition of epigenetics that strictly require their heritability. Significant nucleosome replacement can be detected at actively transcribed

genes as well as epigenetic regulatory elements such as the binding sites of polycomb repressive complexes (Deal et al., 2010). There is also limited evidence supporting the direct recycling of modified histones to the nearby regions of newly synthesized daughter DNA strands. Nevertheless, DNA methylation and heterochromatic silencing that involves H3K9me2 as well as polycomb repressive mechanism are likely 'true' epigenetic systems. The evidence and potential mechanisms for these regulatory systems to form long-term transcription or silencing memory will be discussed in section 1.8.

Regardless of the debate on whether histone modifications can serve as the carrier for long-term epigenetic information, it is nevertheless clear that histone modifications can regulate numerous nuclear processes including but not limited to transcription, DNA replication or damage responses. In order to avoid the overly semantic arguments of the definition, I adopted a more 'contemporary' and relaxed definition as proposed by Adrian Bird of epigenetics in this dissertation – "the structure adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states." In the following sections, I will discuss several major epigenetic systems to review their biogenesis, molecular and biological functions as well as their potential to carry long-term regulatory memories.

1.2 Cytosine methylation (5mC)

1.2.1 Types of cytosine methylation and the corresponding maintenance mechanisms

Cytosine methylation has been found in a wide range of eukaryotes include fungi, plants, invertebrates and vertebrates. However this chemical modification of DNA appears to be lost in some lineage of fungi and invertebrates (Zemach et al., 2010; Feng

et al., 2010). For example, model organisms such as the budding yeast *S. cerevisiae* and the fruit fly *D. melanogaster* are devoid of cytosine methylation (Zemach et al., 2010; Feng et al., 2010). Cytosine methylation can be found in three sequence contexts – CG dinucleotide, CHG or CHH (H stands for any of A, T, C or G). CG methylation often appears as a pair of symmetric 5mC on both strands of DNA and is maintained by the DMNT1 methyltransferase (Chan et al., 2007; Law and Jacobsen 2010). DMNT1 is most active on hemi-methylated sites (cytosine is methylated on one DNA strand but not the other) and faithfully restores them to fully methylated sites. Consistent with the enzymatic characteristics of DNMT1, CG sites often show all-or-none patterns of methylation – sites are either not methylated in any cells or consistently methylated in most cells (Lister et al., 2008; Lister et al., 2009). In both animals and plants, the maintenance of CG methylation also requires a factor containing SET- or RING-associated (SRA) domain (Law and Jacobsen 2010). The factor was identified as UHRF1 (ubiquitin-like plant homeodomain and RING finger domain 1) in mammals and VIM1 (variation in methylation 1) in plants (Bostick et al., 2007; Sharif et al., 2007; Woo et al., 2007). UHRF1 binds hemi-methylated CG sites as well as fully methylated CG sites with lower affinity and was essential for the association of DNMT1 to chromatin (Bostick et al., 2007; Sharif et al., 2007). The SRA domain of VIM1 binds both CG and CNG sites that are fully methylated (Woo et al., 2007). The affinity between VIM1 and methylated CNG may have co-evolved with the plant CMT-KYP machinery. Since the affinities of VIM1 with hemi-methylated CG or CHG sites have not been reported, it is not clear whether VIM1 functions through a mechanism that is comparable to UHRF1 (Woo et al., 2007; Law and Jacobsen 2010). In addition to UHRF1, a SWI/SNF2 chromatin

remodeling factor named LSH1 (lymphoid-specific helicase 1) in mammals or DDM1 (decreased DNA methylation 1) in plants was required for the maintenance of DNA methylation (Vongs et al., 1993; Dennis et al., 2001).

CHG methylation is largely specific for plants and is established by the positive feedback between the plant-specific methyltransferase CMT3 (Chromomethylase 3) and the H3K9-specific methyltransferase KYP (Kryptonite, Chan et al., 2007; Law and Jacobsen 2010). The positive feedback between the two factors is established through the binding of H3K9me2 by CMT3 via its chromodomain and the recognition of CHG methylation by KYP with a SRA domain (Lindroth et al., 2004; Johnson et al., 2007; Law and Jacobsen 2010). Plant CHG methylation generally appears as symmetric methylation pairs. A considerable amount of CHG methylation was also found in human embryonic stem cells (hESC, Lister et al., 2009). However the human CHG methylation was highly asymmetric – 98% of CHG methylation was methylated on only one DNA strand (Lister et al., 2009). This observation was consistent with the absence of CMT in vertebrates and the maintenance mechanism for the asymmetric CHG methylation remains elusive.

In plants, CHH methylation is mediated by DNMT3 type de-novo methyltransferases DRM1 and 2 (domain rearranged methyltransferase 1/2). CHH methylation in plants was established primarily through persistent RNA-directed-DNA-methylation (RdDM) guided by 24nt siRNA (Chan et al., 2007; Law and Jacobsen 2007). CHH methylation is absent from differentiated animal cell types but was found to be abundant in hESC (Lister et al., 2009). CHH methylation showed an 8-10 base periodicity that is similar to the length of a single turn of DNA helix. The observation

supported that animal CHH methylation was catalyzed by DNMT3A/DNMT3L complex because the two active sites in the complex were separated with a distance of 8-10 nucleotides (Lister et al., 2009).

1.2.2 *de novo* DNA methylation in plants and mammals

The system of *de novo* DNA methylation shows substantial differences between plants and mammals. One major reason may be that DNA methylation undergoes two rounds of global resetting during mammal development. In contrary, there is no apparent reprogramming of plant DNA methylome throughout the life cycle except for in the endosperm, which does not directly contribute to the genetic material of offsprings (Sasaki et al., 2008; Feng et al., 2010b). Therefore, the scale of de novo methylation is much more pronounced in animal cells and its importance is apparent. The mechanistic details and significance of plant de novo methylation was not known until the discovery of RNA-direct-DNA-Methylation (RdRM) and the subsequent characterization of RdRM pathway (Matzke et al., 2009; Law and Jacobsen, 2010). The activity of RdDM is largely restricted to the heterochromatin and dispersed repetitive elements in euchromatic arms. The RdDM pathway does not appear to mediate the CG methylation found in about 1/3 of *Arabidopsis* genes (Zhang et al., 2006c; Lister et al., 2008).

The *de novo* methylation in plants is directed by 24nt siRNAs to their homologous regions and can target cytosines in all sequence contexts (CG, CHG and CHH). The precursors of these siRNA are transcribed by the plant specific RNA polymerase IV. The precursors can then be converted to double-strand RNA by RDR2 (RNA-dependent RNA Polymerase 2, Matzke et al., 2009; Law and Jacobsen, 2010). The double-strand RNA can be sliced by Dicer 3 to generate 24nt siRNAs and bound by

AGO4 (ARGONAUTE 4, Matzke et al., 2009; Law and Jacobsen, 2010). The AGO4-siRNA complex may be recruited to RdDM targets through a combination of mechanisms that include the activity of another plant specific RNA polymerase - Pol V. Pol V has been shown to produce long-non-coding RNAs at the target region of RdDM and such activity is essential for the *de novo* methylation (Wierzbicki et al., 2008). AGO4-siRNA complexes may be recruited to Pol V transcribing regions through the pairing between siRNA and Pol V-dependent nascent RNAs (Wierzbicki et al., 2009). It is also possible that AGO4-siRNA complexes can be recruited through protein-protein interactions. The largest subunit of Pol V- NRPE1 possesses a WG/GW-rich domain that can directly interact with AGO4 (El-shami et al., 2007). A homolog of transcription elongation factor SPT5 – KTF1 or SPT5L may provide another docking site for the binding of AGO4 to chromatin. KTF1/SPT5L is able to bind AGO4 and nascent RNA produced by Pol V simultaneously. Therefore it may serve as an adaptor protein that bridges AGO4 with chromatin (He et al., 2009). However, a recent work argued that KTF1/SPT5L and AGO4 are independently recruited to chromatin, although both of them are needed for the *de novo* methylation (Rowley et al., 2011). The late stages of RdDM require factors including DRD1 (defective in RNA-directed DNA methylation) and DMS3 (defective in meristem silencing 3), which contains a structural-maintenance-of-chromosomes hinge domain (Kanno, et al., 2004; Kanno et al., 2008; Ausin et al., 2009). These two proteins are indispensible for the generation of Pol V transcripts even though they do not regulate the accumulation of siRNA.

de novo methylation in mammals is responsible for restoring global DNA methylation following the massive demethylation events in primordial germ cells (PGCs)

and in early embryo. *de novo* methylation is carried out by DNMT3A/B and is assisted by a non-catalytic paralogue DNMT3L (Law and Jacobsen, 2010). DNMT3A/B may be guided to target sequences by a plethora of signals including attraction or repelling mechanisms – 1) DNMT3 can be recruited by sequence-specific transcription repressors such as the oncogenic PML-PAR to specific genomic regions and function as a transcription co-repressor (Klose and Bird, 2006). Similarly, the Myc mediated repression of a target gene $p21^{cip1}$ depends on the catalytic activity of DNMT3a (Klose and Bird, 2006). These results are consistent with the view that certain cis-elements can be instrumental in determining the DNA methylation pattern. This type of mechanism is further supported by a recent work showing that an additional copy of Nanog promoter inserted in the mouse genome can recapitulate the methylation pattern of the endogenous copy, suggesting regulatory sequences can autonomously determine the methylation sites (Lienert et al., 2011). 2) DNMT3 may interact with either histone modifications or histone modifying enzymes for the target selection. DNMT3L specifically interacts with the unmethylated H3K4 and the interaction was strongly disrupted by H3K4 methylations (Ooi et al., 2007). This biophysical behavior of DNMT3L explains the hypo-methylation observed at promoter regions as well as some CpG islands (Ooi et al., 2007). An interaction between DNMTs with H3K27me3 was also speculated in cancer cells (Schlesinger et al., 2007). Sites that are modified with H3K27me3 in normal cells are prone to be targeted for *de novo* methylation (Schlesinger et al., 2007). 3) *de novo* methylation targeting transposable elements during gametogenesis was guided at least partially by the 25-30 bps long piRNAs.

1.2.3 Genome-scale patterns of DNA methylation

The genome-wide patterns of 5mC have been studied at the single-base resolution in a diverse set of organisms. In plants, CG methylation was found in both actively expressed genes and transposons while CHG or CHH methylation were exclusively localized in transposons or repetitive sequences (Zhang et al., 2006c; Lister et al., 2008). Genic CG methylation is enriched towards the 3'-end of gene bodies and is depleted around transcription start sites (TSS) and transcription termination sites (TTS, Zhang et al., 2006c; Zilberman et al., 2007; Lister et al., 2008). CG methylation preferentially associates with long and actively expressed genes (Zhang et al., 2006c; Zilberman et al., 2007; Luo and Lam, 2010). Similar 5mC patterns were also observed in rice Oryza. sative, suggesting the distribution and perhaps the regulation as well as functions of 5mC are conserved between dicots and monocots (Zemach et al., 2010; Feng et al., 2010). The profiling of 5mC in vertebrates revealed patterns that are relevant to plants together with specific distinctions. CG sites in vertebrates were in general heavily methylated. For example, 74% of CG sites in mouse were methylated compared to 24% in *Arabidopsis* (Lister et al., 2009). CG methylation was depleted around TSS, which corresponds with the presence of CpG islands in the promoter of many active genes. DNA methylation in vertebrates may be regulated in a cell-type specific pattern. All DNA methylation (99.98%) are essentially in CG context in IMR90 human fetal lung fibroblasts cells (Lister et al., 2009). However, 17.3% and 7.2% of 5mC were found to be in CHG and CHH contexts respectively in hESC (Lister et al., 2009). The level of CHG and CHH methylations were elevated in genic regions and positively correlated with gene

expression levels (Lister et al., 2009). So far little is known about the functional significance of these non-CG methylation in hESC.

Among organisms that have been profiled for 5mC, several invertebrate showed distinct patterns of 5mC compared to plants or vertebrates (Zemach et al., 2010; Feng et al., 2010). 5mC was enriched in actively expressed genes in the several invertebrate examined and were depleted around TSS, which was similar as plants or animals (Zemach et al., 2010; Feng et al., 2010). However, repetitive sequences were generally hypomethylated in invertebrates even for transposon-rich organisms such as silk moth (Zemach et al., 2010; Feng et al., 2010). Therefore, 5mC may not be indispensable for suppressing transposon activities in invertebrates. Different mechanisms may have evolved in these lineages for controlling transposon activities.

1.2.4 Regulatory functions of DNA methylation

The most well documented function of 5mC was the suppression of transcription initiations when present adjacent to TSS. Although 5mC has been reported to impede the DNA binding activity of certain transcription factors independent of chromatin structures, the generality of this mechanism was not known (Watt and Molloy, 1988). The transcription repression activity of 5mC may be primarily mediated by various proteins that are capable of binding 5mC. In mammals, 5mC binding proteins include MeCP2 (methyl CpG binding protein 2), MBD1-4 (methyl-CpG-binding domain 1-4) and a non-classical 5mC binding protein Kaiso (Klose and Bird, 2006). The disruption of individual MBD proteins caused specific phenotypes in a diverse range of tissue or cell types. Rett syndrome frequently associated with MeCP2 mutations predominantly in the MBD domain (Kriaucionis and Bird, 2003). Knock-out mice having deletion in MeCP2 gene

also showed phenotypes closely resembling the human Rett syndrome (Klose and Bird, 2006). The disruption of MBD2 caused a drastic defect in the differentiation of T-helper cells by inducing the ectopic expression of IL-4 (Hutchins et al., 2002). Although certain functional redundancies may exist between MBD proteins, the observations collectively suggest that 5mC may be involved in diverse cellular processes through the differential interpretations of 5mC with distinct MBD proteins.

The transcription repression that associated with 5mC can be mediated either by the 5mC mark or by DNA methyltransferases. In the earlier scenario, co-repressor complexes can be recruited through the recognition of 5mC by MBD proteins. For example, 5mC binding protein Kaiso was capable of interacting with the N-CoR (nuclear receptor corepressor) complex that associates with HDAC3 (histone deacetylase complex 3, Yoon et al., 2003). MeCP2 has been suggested to interact with histone deacetylase Sin3A complex to suppress the transcription of BDNF (brain-derived neurotrophic factor) in neurons (Martinowich et al., 2003). DNA methyltransferases have been shown to directly cause transcription repressions that are independent from their catalytic activities (Klose and Bird, 2006). DNMT1, DNMT3A and DNMT3B have all been shown to associate with histone methyltransferase, deacetylase or chromatin remodeling factors (Klose and Bird, 2006).

1.3 Cytosine hydroxylmethylation

5-Hydroxymethylcytosine (hmC) is a type of DNA modification first discovered in brains, purkinje neurons and ESCs (Tahiliani et al., 2009; Kriaucionis and Heintz,

2009). hmC was converted from 5mC by TET1 perhaps along with its homologs TET2 and 3 (Tahiliani et al., 2009). The abundance of hmC was estimated to be ~4% compared to 55-60% for 5mC of all cytosines in the CG context (Tahiliani et al., 2009). Antibodies specifically against this modification have enabled the genome-wide mapping of hmC through hMeDIP-seq (hydroxymethylated DNA immunoprecipitation followed by high-throughput sequencing, Williams et al., 2011; Ficz et al., 2011). Alternative approaches involving series of chemical modifications of hmC were also developed including GLIB (glucosylation, periodate oxidation, biotinylation) and anti-CMS. The anti-CMS method requires the conversion of hmC to 5-methylenesulphonate (CMS) by sodium bisulphite treatment for the detection by anti-CMS antibody (Pastor et al., 2011). Since hmC cannot be distinguished from 5mC by bisulfite sequencing, genome-wide profiling of hmC at single-base resolution has not been reported (Huang et al., 2010). Recently, a novel approach based on PacBio SMRT (single molecule, real-time) sequencing platform was developed allowing the hmC mapping at single-base resolution (Song et al., 2011). The method took advantage of the fact that a hmC derivative HS-N_3-5-gmC causes extensive stalls of DNA polymerase during the sequencing that can be capture by the SMRT technique (Song et al., 2011). TET1 ChIP-seq and hMeDIP-seq experiments have shown that the TET1 binding and hmC are both enriched at TSSs of high and intermediate CG density promoters (HCP and ICP, Williams et al., 2011; Pastor et al., 2011; Ficz et al., 2011). The pattern of hmC thus contrasts that of 5mC, which is commonly depleted at HCP and ICPs. The depletion of TET1 led to increases of 5mC in the TET1-bound regions, suggesting TET1 was responsible for the hypo-methylated states of its target regions (Williams et al., 2011; Pastor et al., 2011; Ficz et al., 2011). Tracking the level of

hmC and 5mC during differentiations found that genes down-regulated during the process showed significant decreases of hmC and concomitant increases of 5mC at the promoter (Ficz et al., 2011). The results suggested that the equilibrium between 5mC and hmC may contribute to the dynamic regulation of pluripotent and lineage specific genes (Ficz et al., 2011). In mouse ESCs, the binding of TET1 and hmC were found to be enriched in H3K4me3/H3K27me3 bivalent chromatin states (Wu et al., 2011; Pastor et al., 2011). TET1 was shown to highly co-localize with PRC2 complex and modulate the chromatin binding of PRC2 (Wu et al., 2011). Although no direct interactions were detected between TET1 and PRC2, the hypo-methylated states maintained by TET1 may be critical for the binding PRC2 to chromatin (Wu et al., 2011).

1.4 Polycomb Repressive Complexes and H3K27me3

1.4.1 Components of polycomb repressive complexes

Polycomb group (PcG) genes were originally discovered to be essential for the repression of Drosophila homeotic gene expressions (Schwartz and Pirrotta, 2007; Simon and Kingston, 2009). PcG proteins are critical for the development and cell differentiations of Drosophila, mammals and plants. After decades of active researches, it is now known that PcG proteins contain at least three protein complexes – PRC1, PRC2 (polycomb repressive complex 1 and 2), PhoRC (Pho repressive complex).

PRC2 is responsible for the di- and tri- methylation of H3K27 and contains four core components –EZH1 and EZH2 in human or Enhancer of Zeste (E(Z)) in Drosophila, Suppression of Zeste 12 (SUZ12), EED in human or ESC and ESCL in flies, RBAP48

and RBAP46 in human or Nucleosome remodeling factor 55 (p55) in flies (Schwartz and Pirrotta, 2007; Simon and Kingston, 2009; Margueron and Reinberg, 2011). The activities of PcG complexes can commonly be modified by differential usage of subunit paralogs. The two human E(Z) homolog EZH1 and EZH2 show distinct behaviors in multiple aspects (Margueron et al., 2008; Margueron and Reinberg, 2011). EZH1 primarily expresses in differentiated cells whereas EZH2 strictly presents in actively dividing cells (Margueron et al., 2008). The PRC2 complex containing EZH1 has low H3K27 methyltransferase activity however is capable of repressing transcriptions and inducing chromatin compactions (Margueron et al., 2008). In contrary, the EZH2 containing PRC2 possess active histone methyltransferase activity (Margueron et al., 2008).

Homologs of all four PRC2 core components can be found in the *Arabidopsis* genome. *Arabidopsis* has three E(Z) homologs – Curly Leaf (CLF), Swinger (SWN) and MEDEA (Zheng and Chen, 2011). MEDEA is specifically expressed in endosperm and is involved in the seed development (Kinoshita et al., 1999). CLF and SWN are expressed broadly in vegetative and reproductive tissues. Three proteins - Vernalization2 (VRN2), Fertilization Independent Seed2 (FIS2) and Embryonic Flower2 (EMF2) are *Arabidopsis* homologs of SUZ12 (Zheng and Chen, 2011). The gene family encoding p55 has expanded to five members in *Arabidopsis*. A single gene Fertilization Independent Endosperm (FIE) encodes the homolog of EED subunit (Ohad et al., 1999).

The fly PRC1 contains four core proteins – Polycomb (PC), Posterior sex combs (PSC), polyhomeotic (PH) and RING (Simon and Kingston, 20090. The PRC1 gene families were substantially expanded in mammals. Five PC homologs – CBX2, 4, 6, 7

and 8, two PH homologs – PH1 and 2, three PSC homologs – BMI1, MEL18 and NSPC1, and two RING homologs – RING1A and B can be found in mammal genomes (Schwartz and Pirrotta, 2007; Simon and Kingston, 2009). PRC1 is capable of binding H3K37me3 through the chromodomain in PC. The RING1B and BMI1 subunits of PRC1 can ubiquitylate histone H2A, which was essential for the transcription repression mediated by PRC1 (Cao et al., 2005; Wang et al., 2004). Plant genomes contain no apparent homolog of PC. Instead, a distant homolog of HP1 – LHP1 has been shown to bind and co-localize with H3K27me3 in vivo with its chromodomain and was considered as a functional counterpart of PC in animals (Turck et al., 2007; Zhang et al., 2007). RING homologs AtRING1A and B were recently identified in *Arabidopsis*. AtRING1A and B were found to be essential for the repression of KNOTTED-like homeobox (KNOX) transcription factors (Xu and Shen, 2008). The interactions between AtRING1A/B and LHP1 as well as E(Z) homolog CLF support that AtRING1A/B function similar as RING in flies (Xu and Shen, 2008). Plants PSC homologs were recently identified in *Arabidopsis* as AtBMI1A and B (Bratzel et al., 2010). AtBMI1A/B were shown to interact with EMF1 and LHP1 and were able to uniquitylate histone H2A in vitro and in vivo (Bratzel et al., 2010).

The PhoRC consists of PHO or PHOL together with SFMBT (Scm-like with four MBT domain-containing protein 1) in flies (Klymenko et al., 2006). PHO and PHOL are homologs of the mammalian Yin-Yang 1 (YY1) transcription factor. PHO protein is the only DNA-binding protein among all PcG proteins and was critical for the recruitment of PRC1 and 2 to their targets. SFMBT was able to bind mono- or di- methylated H3K9 and

H4K20 (Klymenko et al., 2006). So far there was no PhoRC protein homologs identified in plants.

1.4.2 The recruitment of polycomb repressive complexes to regulatory targets

Multiple mechanisms have been proposed for the recruitment of PRCs to their targets. These mechanisms include the interaction between trans- factors such as PHO with cis- elements- polycomb response elements (PREs) or the recently discovered long-non-coding RNAs (lncRNAs). It is unlikely a common mechanism would explain all the PRC recruitment events and the responsible factor can be dependent on genomic contexts and cellular states.

The recruitment of PRCs through polycomb response elements (PREs) was the first mechanism identified. The analyses of fruit fly Hox and Engrailed regulatory sequences have located PREs as regions containing several hundreds base pairs (Schwartz and Pirrotta, 2007). However, the activities of PRE could not be assigned to certain consensus sequences (Simon and Kingston, 2009). PHO and its homolog PHOL were zinc finger DNA-binding proteins and appeared to be critical for the binding of PRC to chromatin (Simon and Kingston, 2009). Knocking-down PHO by RNAi caused the dislodging of PRC2 component E(z) and PRC1 subunit PC from their binding sites (Wang et al., 2004b). Genome-wide profiling of PcG proteins confirmed that PHO significantly co-localizes with PRC1 and PRC2, which is consistent with the view that PHO mediates the binding of other PcG proteins (Oktabe et al., 2008, Kwong et al., 2008; Schuettengruber et al., 2009). The role of the mammalian PHO homolog YY1 in PcG protein recruitment is less apparent because YY1 and components of PRC2 showed limited overlaps of their genomic binding profiles in ES cells (Squazzo et al., 2006). The

same work suggested that central ES cell regulators including OCT4, SOX2 and NANOG may contribute to the recruitment of PcG proteins (Squazzo et al., 2006). A interaction between polycomb binding and CpG islands was also speculated from comparing the binding pattern of PRC1/2 with CpG island distributions (Ku et al., 2008).

lncRNAs has recently emerged as a promising mediator for the recruitment of PcG proteins. The HOTAIR generated by the HOXC locus is the first lncRNA that was shown to physically interact with PRC (Rinn et al., 2007). The suppression of HOTAIR caused the de-repression of HOXD expression along with the decrease of H3K27me3 enrichment (Rinn et al., 2007). A following profiling of PRC2 associated RNA found that 20% of lncRNA were bound by PRC2 (Khalil et al., 2009). lncRNAs were also found to associate with other chromatin modifying complexes, suggesting the interaction between RNA and protein may be common for the specific targeting of chromatin modifying complexes (Khalil et al., 2009). A recent characterization of the HOTTIP lncRNA generated from the HOXA locus showed that HOTTIP is essential for the chromatin-looping-induced gene activation through its binding of WDR5 (Wang et al., 2011). The mechanism that lncRNAs bind to protein regulators, presumably through secondary structures, remains to be defined. It is plausible that lncRNAs can serve as molecular scaffolds to spatially coordinate nuclear processes, which is comparable to the role of rRNAs in the formation of ribosomes.

1.4.3 Repression of transcription by polycomb repressive complexes

Multiple hypotheses regarding the mechanism of polycomb mediated transcription repressions have been tested. Purified or reconstituted PRC1 complexes have been shown to induce chromatin condensation *in vitro* (Francis et al., 2004).

However little evidence suggests such mechanism is responsible for transcription repressions in vivo. Unlike the H3K9me2 histone mark that associates with the heterochromatin, H3K27me3 mark does not co-localize with strong DAPI stain in the nucleus. Therefore it is unlikely that PRC complexes induce chromatin condensations at the scale comparable to the constitutive heterochromatin. Further, with the *in vivo* photobleaching experiment and metabolic labeling of histones, it was shown that PC, PH and histone proteins are all highly dynamic in the regions that modified by H3K27me3 (Ficz et al., 2005; Deal et al., 2010). Therefore in term of the kinetics, chromatin domains regulated by PRC are not in the state of static condensation. In addition, using a engineered gene cassette containing a hsp26 promoter driving LacZ gene under the regulation of Ubx PRE element, it was shown that the polycomb mediated silencing do not inhibit the binding of heat shock transcription factors, TBP or RNA polymerase II (Dellino et al., 2004). Therefore, reducing accessibilities of regulatory proteins to the targeted chromatin domain is unlikely the mechanism for PRC mediated transcription repressions.

Mounting evidences suggest that PRC may regulate transcriptions predominantly at the transition stage between the initiation and elongation stages of transcriptions, probably through the E3 ubiquitin ligase activity of PRC1 components RING and BMI (Simon and Kingston, 2009). The investigation of the bivalent chromatin state that are simultaneously modified with H3K4me3 and H3K27me3 in ES cells showed that the bivalent loci were engaged with RNA polymerase II (Pol II) but were nevertheless poorly expressed (Stock et al., 2007). The Pol II C-terminal domains (CTD) at these loci were modified with Ser-5 phosphorylation that is characteristic for early transcription

elongations (Stock et al., 2007). However, Ser-2 phosphorylation of Pol II CTD was not detected at bivalent loci, suggesting the RNA polymerases were paused before entering the elongation phase (Stock et al., 2007). The conditional knockout of the RING1B protein caused the depletion of histone H2A monouniquitination (H2Aub) together with the de-repression of polycomb target genes (Stock et al., 2007). The role of H2Aub in transcription repressions was further supported by the analysis of a co-repressor complex nuclear receptor corepressor 1 (NCOR1, Simon and Kingston, 2009). H2Aub was shown to prevent the recruitment of a general transcription elongation factor FACT and impede the release of Pol II into the phase of productive elongations (Zhou et al., 2008). Therefore the results together suggest that PRC complexes may suppress transcription at least partially through the deposition of H2Aub and the resulting polymerase pausing.

1.5 Histone H3 lysine 4 methylations

The histone H3 lysine 4 can be modified with one, two or three methyl- groups. The global patterns of H3K4 methylation are different between organisms. In yeast and plants, H3K4me1 was enriched at the mid- to 3'- region of gene bodies (Pokholok et al., 2005; Zhang et al., 2009). In human however, H3K4me1 peaked around TSS and was suggested as a signature for active enhancers along with H3K27Ac and H3K4me3 (Barski et al., 2007; Heintzman et al., 2009). In both plants and animals, H3K4me2 largely overlapped with H3K4me3 and was enriched around the TSS of actively expressed genes (Barski et al., 2007, Zhang et al., 2009). In contrary, H3K4me2 primarily decorated gene bodies in yeast (Pokholok et al., 2005). In all organisms that have been

examined, H3K4me3 faithfully tracked the TSS region of actively expressed genes (Ruthenberg et al., 2007).

All three types of H3K4 methylation were catalyzed by SET1 subunit of the COMPASS complex in yeast. SET1 is the only H3K4 methyltransferase in this organism. H3K4 methylation also depended on the Ser-5 phosphorylation of the C-terminal domain of the RNA-polymerase II as well as the Polymerase Associated Complexes (Li et al., 2007; Ruthenberg et al., 2007). Histone H2B monoubiquitination mediated by Rad6/Bre1 was also required for the H3K4 methylation. The gene family of H3K4 methyltransferase was substantially expanded in higher plants and mammals. At least four SET domain proteins were identified to have the H3K4 methyltransferase activity in the model plant *Arabidopsis* (Alvarez-Venegas et al., 2003; Saleh et al., 2008; Tamada et al., 2009; Guo et al., 2010; Berr et al., 2010). Mammalian genomes contain at least ten H3K4 methyltransferases (Ruthenberg et al., 2007). Compared to the situation in the yeast that the knockout of SET1 completely removed all detectable H3K4me, the disruption of any single H3K4 methyltransferse in plants or animals caused modest loss of H3K4me and moderate phenotypes. For example, the disruption of *Arabidopsis* ATX1 caused the mis-regulation of merely a few hundred genes, although a set of flower homeotic genes were down-regulated, which led to certain specific phenotypes in flowers of *atx1* mutant (Alvarez-Venegas et al., 2003; Saleh et al., 2008). Recently, SDG2 was identified to be responsible for ~65% of H3K4me3 in *Arabidopsis*, albeit no decrease of H3K4me1 and H3K4me2 were observed under the knockout of SDG2 (Guo et al., 2010; Berr et al., 2010). Transcriptome analysis showed that ~500 genes were significantly mis-regulated in *sdg2* background (Berr et al., 2010). The limited impact on transcriptome caused by

the loss of H3K4me3 may be explained by that H3K4me3 is not absolutely essential for transcription initiations or elongations.

In spite of the strict association between H3K4me3 and active expressions, the molecular function of H3K4me3 is less well defined. Histone methylations do not affect the charging of histone tails and therefore are unlikely to directly change the physical properties of chromatin. It is likely that the majority of the H3K4me3 functions are mediated by 'effector' proteins that can recognize this histone mark. The recognition of H3K4me3 by chromatin remodeling factor CHD1 or the component of histone acetyltransferase complex ING3-5 may explain the effect of transcription activations associated with H3K4me3 (Ruthenberg et al., 2007). The binding of H3K4me3 by the TAF3 subunit of TFIID suggest that H3K4me3 may guide the anchoring of general transcription machinery (Vermeulen et al., 2007). This type of mechanism may be utilized to form certain transcriptional memories that are mediated by H3K4me3. The molecular functions of certain effectors of H3K4me3 suggest that this histone mark may function in processes irrelevant to transcription activations. The interaction between H3K4me3 and CHD1 was shown to facilitate the chromatin association of U2 snRNP spliceosome components (Sim et al., 2007). A particular intriguing effector protein for H3K4me3 is ING2 that can recruit the Sin3/HDAC histone deacetylase complex (Shi et al., 2006). This interaction has been implicated to mediate the 'shut-down' of active transcriptions during cell death by guiding co-repressor complexes to H3K4me3, which is the signature of active transcriptions (Shi et al., 2006).

1.6 Heterochromatin marks – histone H3 lysine 9 dimethylation and histone H3 lysine 27 monomethylation

The formation of heterochromatin is essential for maintaining the silencing of transposable elements and ensuring the proper function of centromeres (Grewal and Jia, 2007). In fission yeast, the heterochromatin was maintained by chromodomain protein HP1 homolog Swi6 and histone H3K9 methyltransferase SU(VAR)3-9 homolog Clr4 (Grewal and Jia, 2007). Swi6 bound to the characteristic heterochromatic histone mark H3K9me2 and further recruited Clr4 to reinforce the heterochromatin stability (Grewal and Jia, 2007). The RNAi machinery may be fundamental for the targeting of protein factors to the heterochromatic regions (Volpe and Martienssen, 2011). Non-coding RNAs are produced from both strands of the centromere in S.pombe (Kloc et al., 2008; Chen et al., 2008). The transcription activity of centromeric repeats was cell cycle regulated and peaked at G1 to S phase, which correlated with the decrease of H3K9me2 and Swi6 binding in the centromere (Kloc et al., 2008; Chen et al., 2008). The processing of these bidirectional centromeric transcripts by the RNAi machinery was critical for the heterochromatic silencing. Presumably the double-strand RNA formed by the bidirectional centromeric transcripts can be processed into siRNA by Dicer and guide the RNA-induced initiation of transcriptional silencing complex (RITS) to heterochromatic regions (Verdal et al., 2004). The processing of long centromeric transcripts to siRNA may also involve RNA-directed RNA polymerase complex (RDRC) that can be recruited by RITS (Motamedi et al., 2004). Once RITS was established at heterochromatin regions, the complex can further recruit the Clr4 methyltransferase complex (ClrC), which include Clr4 and other factors (Zhang et al., 2008b). The stability of heterochromatin formation

may be further reinforced by the binding of H3K9me with the chromodomain of Clr4 (Zhang et al., 2008b).

A distinctive feature of the heterochromatin formation in plants was the involvement of CHG DNA methylation. CHG methylation was catalyzed by the plant specific methyltransferase CMT3, which can bind H3K9me with its chromodomain (Lindroth et al., 2004). The H3K9 methyltransferase KYP can in turn recognize CHG methylation with the SRA domain (Johnson et al., 2007). The putative positive-feedback between CHG and H3K9me marks may contribute to the stability of plant heterochromatin, which was antagonized by the jmjC-domain containing H3K9me demethylase IBM1 (Saze et al., 2008; Inagaki et al., 2010). The plant RNAi pathway that involves 24nt siRNA clearly contributed to certain aspects of the heterochromatin formation (Volpe and Martienssen, 2011). However it was not known how KYP and CMT3 are recruited by the RNAi machinery such as the AGO4-siRNA complex.

H3K27 monomethylation was identified as a specific heterochromatin mark in *Arabidopsis* (Jacob et al., 2009). The histone mark was deposited by the redundantly acting ATXR5 and ATXR6 (Jacob et al., 2009). Although the disruption of H3K27me1 caused the reactivation of certain heterochromatic loci, H3K27me1 appeared to be largely independent from H3K9me or DNA methylation (Jacob et al., 2009). Further characterizations of the *atxr5/atxr6* double mutants identified that H3K27me1 was crucial for the proper replication of heterochromatin DNA (Jacob et al., 2010). Re-replications of transposon and repetitive DNA were observed in the *atxr5/atxr6* mutant (Jacob et al., 2010). The function of H3K27me1 in the transcription regulations and its interaction with other silencing mechanisms remains to be defined.

1.7 Histone H3 lysine 36 methylation

Histone H3 lysine 36 methylation was mediated by SET2 histone methyltransferase and its homologs. The deposition of H3K36me required the Ser-2 phosphorylation of the C-terminal domain of the RNA-polymerase II, Polymerase Associated Complexes and Histone H2B monoubiquitination (Li et al., 2007). In yeast and animals, H3K36me2 and H3K36me3 were both enriched towards the 3'- gene bodies of actively expressed genes. In plant species including *Arabidopsis* and maize, H3K36me3 was found at TSS regions occupying overlapping domains as H3K4me3 (Wang et al., 2009; Charron et al., 2009). H3K36 methylation has been found to associate with two major molecular functions. 1) In yeast, H3K36me was recognized by the EAF3 subunit of the Rpd3S histone deacetylase complex. The recruitment of Rpd3S led to the hypo-acetylation of gene body regions and was essential for the suppression of intragenic cryptic promoters (Li et al., 2007b; Lickwar et al., 2009). 2) H3K36me3 was enriched in the exons of C.elegan and mammals, suggesting potential regulatory interactions between the histone mark and RNA splicing (Kolasinska-Zwierz et al., 2009). The speculation was later supported by works using human cells showing that H3K36me can be recognized by chromodomain protein MRG15 and subsequently regulated alternative splicings (Luco et al., 2010).

1.8 The heritability of chromatin modifications

As aforementioned, convincing mechanisms for the perpetuation of long-term regulatory memories were only identified for 5mC, heterochromatin silencing and polycomb repressive mechanisms. The maintenance DNA methyltransferase DNMT1 directly interacts with the DNA replication fork and faithfully copies the symmetric CG

methylation onto the newly synthesized DNA strand (Chuang et al., 1997). Since the DNA methylome does not undergo major reprogramming throughout plant life-cycles, changes of 5mC patterns are frequently carried over into the offspring and can persist for multiple generations. Plant epi-alleles caused by the variation of 5mC patterns may be one of the best examples for heritable epigenetic regulations (Jacobsen and Meyerowitz, 1997). Due to the global resetting of DNA methylome at the stage of primordial germ cells (PGCs) and in early embryo of mammals, it is unclear if any specific 5mC pattern established in the earlier generation can be transmitted to the offspring.

The RNAi machinery has been postulated to play critical roles to reinforce and stabilize heterochromatic silencing in S.pombe and multi-cellular organisms. As previously discussed, the S.pombe centromeric transcripts are actively transcribed during S-phase (Kloc et al., 2008; Chen et al., 2008). siRNAs derived from these transcripts may facilitate the formation of heterochromatin in the newly synthesized genomes (Kloc et al., 2008; Chen et al., 2008). Plants also adopted similar mechanisms during gametogenesis to maintain the silencing of transposable elements (TEs) in the next generation. Athila TEs are specifically reactivated in the pollen vegetative nucleus (VN) and produce abundant siRNAs (Slotkin et al., 2009). These siRNA originated from vegetative nucleus can move into the sperm presumably to reinforce the silencing of TEs (Slotkin et al., 2009). The RNAi-dependent propagation of heterochromatin suggests that epigenetic regulations can perpetuate with mechanisms other than the direct recognition and replication of chromatin marks.

A mechanism for the propagation of polycomb-mediated-silencing after cell divisions has been postulated through the analysis of PRC2 subunit EED (Margueron et

al., 2009). The carboxy-terminal domain of EED was shown to specifically bind H3K27me3 and stimulate the methyltransferase activity of PRC2 (Margueron et al., 2009). The authors suggested a model that histones associated with H3K27me3 are locally 'recycled' during DNA replications, although the repressive signal may be diluted by the newly incorporated histone proteins. The H3K27me3 provided by 'old' histones can stimulate the methyltransferase activity of PRC2 to catalyze the H3K27me3 mark on neighboring 'new' histones to maintain the integrity of polycomb repressive domains.

Histone modifications associated with active chromatins are generally considered to reflect the activity of RNA polymerases (Henikoff and Shilatifard, 2011). As discussed in sections 1.5 and 1.7, the methyltransferases of H3K4me (COMPASS) and H3K36me (SET2) are recruited by the elongating RNA polymerase II at least in S.cerevisiae. The recruitment mechanism for COMPASS, SET2 or histone acetytransferase complexes are likely to be more sophisticated in multi-cellular organisms. However no apparent 'self-propagation' mechanism has been identified for these 'active' modifications. 'Active' histone modifications may shape the epigenome through antagonizing the deposition of repressive chromatin marks. For example, H3K4me3 can inhibit the binding of DNMT3L and prevent the ectopic methylation of promoter regions (Ooi et al., 2007). Recent work also suggests that H3K36me2 can inhibit the histone methyltransferase activity of PRC2 (Yuan et al., 2011). It remains to be tested whether these proposed mechanisms reflect any *in vivo* interactions. Nevertheless, the results suggest possible mechanisms other than 'self-propagation' that allow 'active' chromatin marks contributing to the heritable epigenome.

2 Defining the genomic position-dependent interaction networks of chromatin regulators in *Arabidopsis thaliana*

2.1 Introductions

The distributions of chromatin marks and physical binding sites of chromatin regulators can now be quantitatively measured by techniques such as ChIP-chip and ChIP-seq. However, it remains to be addressed whether the recruitment of certain combinations of chromatin regulators as well as gene expressions can be affected by the genomic position *per se*. Over the past 20 years of plant transcriptional regulation studies, both positive and negative evidences have been described regarding the tentative 'position effects'. Observations supporting position effects include that transposable elements that are highly silenced tend to accumulate in the pericentromeric heterochromatin and regions adjacent to the NORs (Nuclealor Organizer Region, *Arabidopsis* Genome Initiative, 2000); The nucleolar dominance mediated rDNA expression is locus-dependent – ectopic rDNA copies introduced by T-DNA and incorporated outside of NOR can escape from the nucleolar dominance mediated silencing (Lewis et al., 2007). In addition, protein-coding genes localized in heterochromatic arms are in general less actively transcribed compared to genes in euchromatic arms (Schmid et al., 2005). These observations suggest that genomic positions contribute to the overall transcription activity and regulatory mechanisms. In contrary, several works using transgene-tagging approaches to assess the transcriptional potentials across the genome argued against the existence of any 'position effects' (Schubert et al., 2004; Nagaya et al., 2005).

In order to evaluated any potential position effects at a greater resolution compared to previous studies, the laboratory of Dr. Eric Lam has generated a collection of 277 Chromatin Charting (CC) transgenic lines with an identical transgene cassette (including a convenient luciferase reporter gene [Luc] driven by CaMV 35S promoter, a NPTII selection marker and ~64 copies of LacO [lac operon sequence]) incorporated in different regions of the *Arabidopsis* genome (Rosin et al., 2008). When presented with LacI-GFP recombinant protein either through stably integrated transgene expressions or transient expressions, the tandem LacO elements can be specifically bound by LacI-GFP enabling the *in vivo* visualization and tracking of the integration locus for the transgene (Kato and Lam, 2001; Lam et al., 2004). After determining the relative luciferase activities (RLA) of all 277 lines, Rosin et al., found about 20% of these lines showed less than 50% or higher than 150% of the RLA for the reference line CCP4.20 (Rosin et al., 2008). These lines showing deviating transgene expressions were designated as 'position effect lines' (Rosin et al., 2008). However, the mechanism that underlies these significant differential expressions of the Luc gene remained elusive. The highly repressed Luc expression in CCP4.211 line was shown to be mediated by the DNA methylation pathway involving DDM1 and MET1 (Rosin et al., 2008).

In order to test if different CC lines showing repressed Luc expressions were suppressed by a common mechanism or distinct mechanisms that vary between specific lines (e.g. DNA methylation or polycomb mediated silencing), we set out to knock down a panel of chromatin regulators individually in four different CC backgrounds and observe the changes of RLA induced by RNAi. The four CC lines used in this work were CCP4.211, CCT383, CCT 396 and CCT431. The CC construct in CCP4.211 integrated

within a highly heterochromatic environment that was surrounded by rDNA from 5'-end and Ty3-Gypsy retroelements from 3'-end (Figure 2.1D, Rosin et al., 2008). The integration sites of CC constructs for CCT383, CCT396 and CCT431 were in euchromatic regions that were populated by actively expressed genes (Figure 2.1D, Rosin et al., 2008). Interestingly, CCP4.211 and CCT383 showed similar RLAs, even though the transgenes in the two lines were integrated in drastically different contexts (Rosin et al., 2008). The Luc genes in CCT396 and CCT431 were actively expressed in leaf tissues – close to the level of the reference line CCP4.20 (Rosin et al., 2008). Importantly, the Luc genes expressed in a highly organ-specific manner in CCT396 and CCT431 (Rosin et al., 2008). Therefore, addressing the epigenetic regulatory mechanisms that were targeted to the CC transgene in CCT396 and CCT431 would help resolving the mechanistic basis of the organ/tissue specific silencing.

We have chosen 8 epigenetic regulators (epi-regulators) to be repressed by the RNAi approach in the four CC lines (Table 2.1). As discussed earlier in the general introduction, DRM2 is the major *de novo* DNA methyltransferase involved in RdDM (Matzke et al., 2009). CMT3 is the DNA methyltransferase responsible for CHG methylation and is able to interact with histone H3 methylation through its chromodomain (Lindroth et al., 2004; Law and Jacobsen, 2010). DRD1 is a SNF2 like protein that is essential for the association of RNA polymerase V with chromatin (Kanno et al., 2004; Wierzbicki et al., 2008). CLF is one of the *Arabidopsis* homolog of E(z) (Enhancer of zeste), which is a histone methyltransferase specific for H3K27. LHP1 has been shown to specifically bind H3K37me3 both *in vitro* and *in vivo* and mediate the transcription repression induced by H3K27me3 (Turck et al., 2007; Zhang et al., 2007).

MOM1 is required for the transcription repression of certain repetitive sequences in *Arabidopsis* without affecting DNA methylation (Vaillant et al., 2006). Recent works have revealed a genetic synergy between MOM1 and RNA polymerase V in silencing a set of genomic loci, suggesting MOM1 may be involved in the RdDM process (Yokthongwattana et al., 2010). Since the only difference between independent CC lines was the genomic position where the construct integrated, we reasoned that differential responses to regulator RNAi can, at least partially, be attributed to the effects of genomic regions and local sequence contexts.

Table 2.1. Chromatin regulators that were suppressed by RNAi in CC lines.

Chromatin regulator being suppressed	Associated pathway	Biochemical activity
MOM1 (Morpheus Molecule)	Unknown	Unknown
CMT3 (Chromomethyltransferse)	CHG DNA methylation, Heterochromatin Maintenance	Methyltransferase specific for CNG context
DRM2 (Domain Rearranged Methyltransferase 2)	RNA Dependent DNA Methylation	Methyltransferase for *de novo* DNA methylation
DRD1 (Defective in RNA-Directed DNA Methylation)	RNA Dependent DNA Methylation	SWI/SNF2 Chromatin Remodeling Factor
SUVH2 (Su(Var) Homolog 2)	RNA Dependent DNA Methylation	Putative histone methyltransferase
CLF (Curly Leaf)	Polycomb Silencing	H3K27 methyltransferase
LHP (Like Heterochromatin Protein 1)	Polycomb Silencing	Associate with H3K27me3 with Chromo Domain
HD1 (Histone Deactylase 1)	Unknown	Histone Deacetylase

2.2 Materials and methods

2.2.1 Generation of RNAi transgenic plants

RNAi constructs were generated by ChromDB (http://www.chromdb.org/) and distributed by *Arabidopsis* Biological Resource Center (ABRC, www.arabidopsis.org, Gendler et al., 2008). The accessions for RNAi constructs targeting each epigenetic regulator were: MOM1 (CD3-528), DRM2 (CD3-557), HD1 (CD3-515), DRD1 (CD3-646), LHP1 (CD3-647), CMT3 (CD3-649), CLF (CD3-538) and SUVH2 (CD3-539). The RNAi constructs were transformed into *Agrobacterium tumefaciens* GV3101/pMP90. *Arabidopsis* plants were transformed with *Agrobacterium* using the floral dipping method (Zhang et al., 2006). For the selection of transgenic plants, T1 seeds were geminated on 0.5X MS medium supplemented with 1% sucrose, 0.8% agar, 10 μg/ml glufosinate ammonium and 200 μg/ml carbenicillin. Resistant seedlings were identified ~10 days (including 2 days of imbibing) after the germination and were transferred to vertical growth allowing the extension of roots on the medium surface.

2.2.2 Bioluminescence imaging of living *Arabidopsis* plants

12-14 days old *Arabidopsis* plants growing vertically were sprayed homogenously with 1 mM D-luciferin sodium salt supplemented with 0.01% Triton X-100. The plants were incubated at room temperature for 10 min to allow the penetrance of luciferin. Before imaged for bioluminescence, plants were kept in closed imaging chamber for 5 min to quench the delayed autofluorescences. The imaging was performed with Lumazone Fluorescence Automated System (MAG Biosystems) with 7 min exposures. Images were processed by MAG Biosystems Software Ver. 7.5.5.0 and MAG Biosystems Lumazone Analyzer 2.0. The RLA of each plant was quantified by the average pixel

intensities within the areas circled by the plant tissue outline. Imaging backgrounds were determined by the average pixel intensity of regions not covered with plants and were subtracted from the sample measurements.

2.2.3 Molecular quantifications of transcript abundances

The abundances of NPTII transcripts were determined by northern blots. Total RNA was isolated from 12-day-old plants with Plant RNA Purification Reagents (Invitrogen). 10 μg RNA samples were resolved on 1% denaturing agarose gel and were blotted onto a Hybond N^+ nylon membrane (GE Healthcare). The probe was prepared using a PCR product amplified with primers EL1069 and EL1070 from the genomic DNA of CC lines (primer sequences listed in Table 2.2). The probe preparation was performed using RadPrime DNA Labelling System (Invitrogen) with [^{32}P]dCTP (MP Biomedical). The radioactive probe was hybridized with membrane at 68°C using MiracleHyb Hybridization Solution (Strategene).

For quantifying the expression of luciferase genes and endogenous loci, RNA samples were treated with RQ1 DNase (Promega) to remove the residual DNA. 400 ng total RNA was used for each reverse transcription reaction using Improm-II Reverse Transcription System (Promega). Primers sequences used for PCR amplications were listed in Table 2.2.

Table 2.2. List of Primers

Target	Primer	Sequence (5'->3')	Cycles	Application
LHP1	EL3163	GGCGTTCGATTGTACTTGAGA	35	RT-PCR
(AT5G17690)	EL3164	TGTAAGACCCAATGGCCTTC		
MOM1	EL3209	ACAAATCCAGGTCTGCGTTC	33	RT-PCR
(AT1G08060)	EL3210	ACCACTTCATTGTTGCTCTGC		
CMT3	EL3169	TTCCCAAAGCATATCCAAGG	33	RT-PCR
(AT1G69770)	EL3170	CACAACGCCATTTCAAAAGTT		
DRD1	EL3198	TAAGGCAGGACTTTCGAGGA	33	RT-PCR
(AT2G16390)	EL3199	TTGTTGGAGCATCTCGCATA		
DRM2	EL3165	CAAGACCAACTCGGCTTACC	35	RT-PCR
(AT5G14620)	EL3166	TGGCCAGAATGGGATAAAAG		
SUVH2	EL3254	GCTTGCCGTTCACTTCATCT	33	RT-PCR
(AT2G33290)	EL3255	CCCTCCACTGGATTTCTCAA		
CLF	EL3250	TGGGTCTACCAACAGAAGGTG	33	RT-PCR
(AT2G23380)	EL3251	CTGAAATTCGCCAACCATTC		
HD1	EL3159	CTCCACTCCCACTGGGTTTA	33	RT-PCR
		TGGAGTTGCACTTGGAGTTG		

(AT4G38130)	EL3160			
LUCIFERASE	EL3373	TTTCTTGCGTCGAGTTTTCC	32	RT-PCR
	EL3374	AACACCCCAACATCTTCGAC		
NPTII	EL2056	ACAACAGACAATCGGCTGC	31	RT-PCR
	EL2057	TGCTCTTCGTCCAGATCATCC		
NPTII	EL1069	GGGCACAACAGACAATCG	30	Northern
	EL1070	GTAGCCAACGCTATGTCC		Blot Probe
ACT2	EL3031	ATCCAAGCTGTTCTCTCCTTG	28	RT-PCR
(AT3G18780)	EL3033	AGAGCTTCTCCTTGATGTCTC		
AG	EL3396	CCGATCCAAGAAGAATGAGC	37	RT-PCR
(AT4G18960)	EL3397	CTAACTGGAGAGCGGTTTGG		
FLC	EL3394	GCAAGCTTGTGGGATCAAAT	37	RT-PCR
(AT5G10140)	EL3395	TTTGTCCAGCAGGTGACATC		
180bps-	EL3398	ACCATCAAAGCCTTGAGAAGCA	35	RT-PCR
Repeats	EL3399	CCGTATGAGTCTTTGTCTTTGTATCTTCT		
CyP40	EL3400	AGACGAGTACTGCCTTGCGT	29	RT-PCR
(At2G15790)	EL3401	TTCTTTCTTGATACCAGCGTCA		
AtSN1	EL3385	ACCAACGTGTTGTTGGCCCAGTGGTAAATC	37	RT-PCR
	EL3385	AAAATAAGTGGTGGTTGTACAAGC		
AtMu1	EL3392	CCGAGAACTGGTTGTGGTTT	37	RT-PCR

	EL3393	GCTCTTGCTTTGGTGATGGT		
Athila LTR	EL3388	TGTTTCATCCACGTTTCATCTC	37	RT-PCR
	EL3389	AGCAATAAGCGCAACTAATCC		

2.2.4 Clustering analysis and visualizations

The clustering analysis on epigenetic regulators based on the change of RLA after the suppression of the regulator with RNAi was performed with Cluster 3.0 (de Hoon et al., 2004). Clustering was performed with Euclidean distances and centroid linkage. The head map was generated with Java Treeview 1.1.3 (Saldanha, 2004).

2.3 Results

2.3.1 Characterization of transgene expressions in the four CC lines

The measurement of RLAs reflects the abundance of the catalytically active Luc enzyme. We further determined the transcript abundances for the Luc gene and the selection marker (NPTII) gene with RT-PCR and northern blot respectively (Figure 2.1B and C). RLAs quantified with the bioluminescence imaging showed an apparent correlation with the abundances of Luc transcripts (Figure 2.1A), which confirmed that Luc enzyme assays accurately reflect the transcription activity of Luc gene. A general correlation between the abundances of Luc and NPTII transcripts was found – CC lines showing higher Luc transcripts level generally associated with greater NPTII expression and *vice versa* (Figure 2.1B and C). However, relative to the NPTII expression in other CC lines, NPTII gene in CCT383 expressed at a higher level compared to the Luc gene in the same line (Figure 2.1B and C). The NPTII expression in CCT383 was substantially greater than in CCP4.211 and was comparable to CCP4.20. In contrary, the Luc gene in CCT383 expressed similarly as that in CCP4.211 and was much more repressed compared to in CCP4.20. The bioluminescence imaging has also confirmed our previous

observation that the root tissue of CCT431 and CCT396 showed substantially lower RLAs than in the shoot tissue (Rosin et al., 2008).

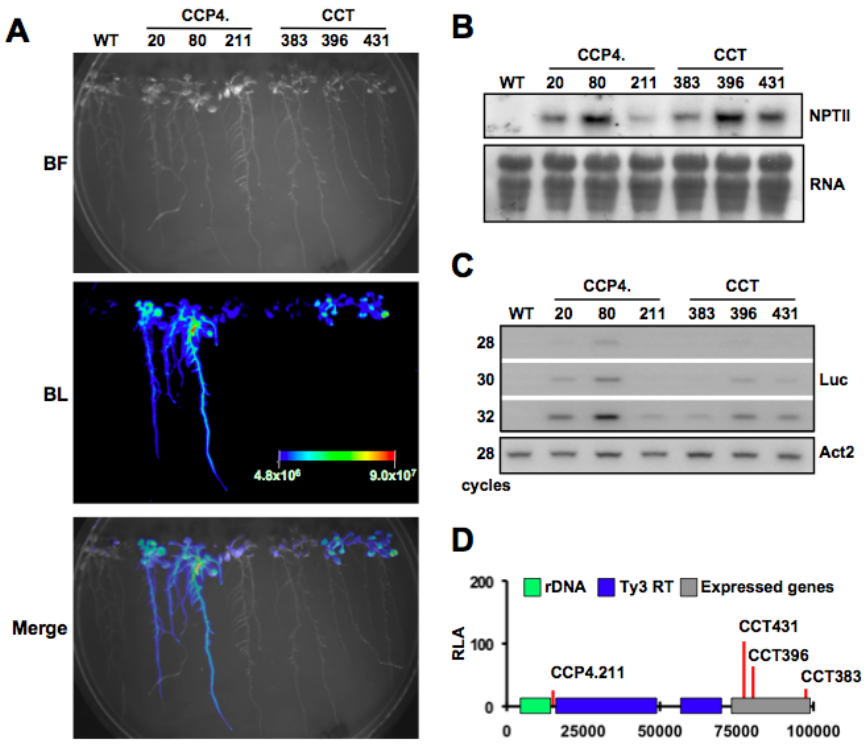

Figure 2.1. Luc and NPTII expression in CCP4.211, CCT431, CCT396 and CCT383.

(A) Bioluminescence imaging of CC lines. CCP4.20 and CCP4.80 contain actively expressed Luc gene and served as reference lines for 'normal' and 'high' Luc expressions. BF: bright field. BL: bioluminescence. Merge: The merge of bright field and bioluminescence images. (B) Northern blot quantification of NPTII transcripts in CC lines. RNA panel shows the methylene-blue stain of the membrane as the loading control. (C) RT-PCR quantification of Luc transcripts in CC lines. Results obtained from 28, 30 and 32 cycles of amplifications were shown. (D) The location of the transgene cassette for the four CC lines. The heights of red bars indicate the RLAs of corresponding CC lines.

2.3.2 Test the specificity and efficacy of epi-regulator RNAi through analyzing the mis-expression of endogenous loci

In order to empirically test the specificity and efficacy of our RNAi suppressions, we determined the expressions of loci that were known to be mis-regulated under the perturbation of epi-regulators included in this study. Two independent T1 lines for each RNAi construct in the background of CCP4.211 were used for this analysis (Figure 2.2). A number of known mis-regulations were successfully recapitulated with our RNAi lines. For example, AG (Agamous) and FLC (Flowering Locus C) were up-regulated in RNAi lines that suppressed LHP1 or CLF (Figure 2.2). LHP1 and CLF are the component of plant PRC1 and PRC2 complexes respectively, which repress AG and FLC through the histone mark H3K37me3 (Zheng and Chen, 2011). In addition, the suppression of MOM1 reactivated centromeric 180bp repeats and CyP40, which were both shown to be reactivated in the *mom1* background (Figure 2.2, Habu et al., 2006). The specificity of our RNAi approach was supported by the differential responses of certain loci to the suppression of epi-regulators. For example, the RNAi of LHP1 and CLF did not induce any transcription reactivation with heterochromatic loci such as Cyp40, AtMu1 or AtSN1 (Figure 2.2), which is consistent with the euchromatic localization of H3K27me3 mark. Similarly, the suppression of MOM1 or DRD1 had little effect on protein coding genes FLC and AG. Therefore we concluded that the RNAi approach for the suppression of epi-regulators performed as expected with regard to efficacy as well as specificity.

Our analysis also revealed regulatory relations that had not been previously reported. We found that AG was repressed by HD1 together with non-CHG methylation pathway including CMT3 and DRM2 (Figure 2.2). In addition, AtSN1 was activated by

the suppression of RdDM components DRD1 and DRM2 (Figure 2.2). This result is consistent with the finding that NRPD1 and RDR2, which are responsible for the generation of siRNA, were also required for the repression of AtSN1 (Bäurle et al., 2007).

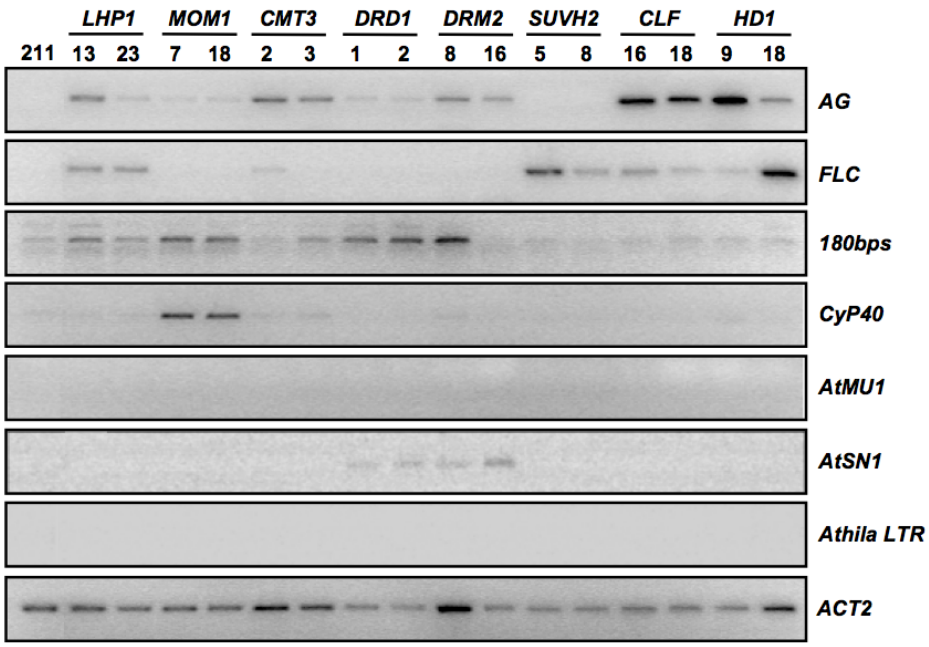

Figure 2.2. The expressions of endogenous loci under the suppression of epi-regulators with RNAi.

Two independent T1 plants for each RNAi construct in CCP4.211 background were included in the analysis. The line number for the T1 plants were indicated underneath the epi-regulator being suppressed. AG – Agamous; FLC – Flower Locus C; Cyp40 – AT2G15790; 180bps – centromeric satellite repeat. AtMu1 – *Arabidopsis* mutator like element 1; AtSN1 – SINE retroelement; ACT2 – Actin 2.

2.3.3 The set of epi-regulators that control Luc expressions in CC lines were tissue specific

Eight epi-regulators were individually suppressed by RNAi in the four CC lines (32 RNAi-CC combinations) to derive the set of regulators that modulate the expression of Luc gene in each line. The changes of RLA upon RNAi indicated the functional significance of the suppressed regulator for regulating the particular CC locus. A schematic description of the experimental procedure was shown in Figure 2.3A. RLAs were determined in 15-26 independent T1 plants together with their parental line through bioluminescence imaging, which captured RLA in both shoot and root tissues. In many RNAi-CC combinations, a subset of T1 plants showed significant changes of RLA compared to their parental lines, suggesting the CC locus is regulated by the factor being repressed. For these RNAi-CC combinations, the results shown in Figure 2.3B and 2.4 were obtained from six representative T1 plants showing significantly different RLAs from the parental line. In other combinations that no significant change in RLA was found in any T1 plants compared to the parental line, we concluded that the suppressed regulator was dispensable for regulating the CC locus. In this case, RLAs of six randomly chosen plants would be reported (Figure 2.3B and 2.4).

For each CC background, we observed that RLAs in shoots and roots responded differently to the RNAi of certain epi-regulators. For example, RNAi of MOM1 effectively reversed the suppression of RLA in the shoots but not in roots of CCP4.211 (Figure 2.3B and 2.4). No significant change of RLA was caused by suppressing any of the studied epi-regulators in shoots of CCT431 and CCT396, whereas the RNAi of multiple factors caused pronounced reactivation of RLA in the root tissue for both lines

(Figure 2.4). This result provided a mechanistic explanation for the root specific repression of RLAs in CCT431 and CCT396 – the expression pattern may be established through the root specific targeting of Luc genes in the two CC lines for epigenetic silencing.

CCT383 showed repressed RLA in both shoots and roots (Figure 2.1). Although all the eight epi-regulators targeted for RNAi were involved in the repression of Luc gene in the shoots of CCT 383 (Figure 2.3B), suppressing none of them led to the reactivation of RLA in the root (Figure 2.4). It is thus likely that the factors responsible for repressing RLA in roots of CCT383 were not included in the current study. Therefore distinct epigenetic mechanisms were involved for the repression of Luc gene in shoots and roots for CCT383.

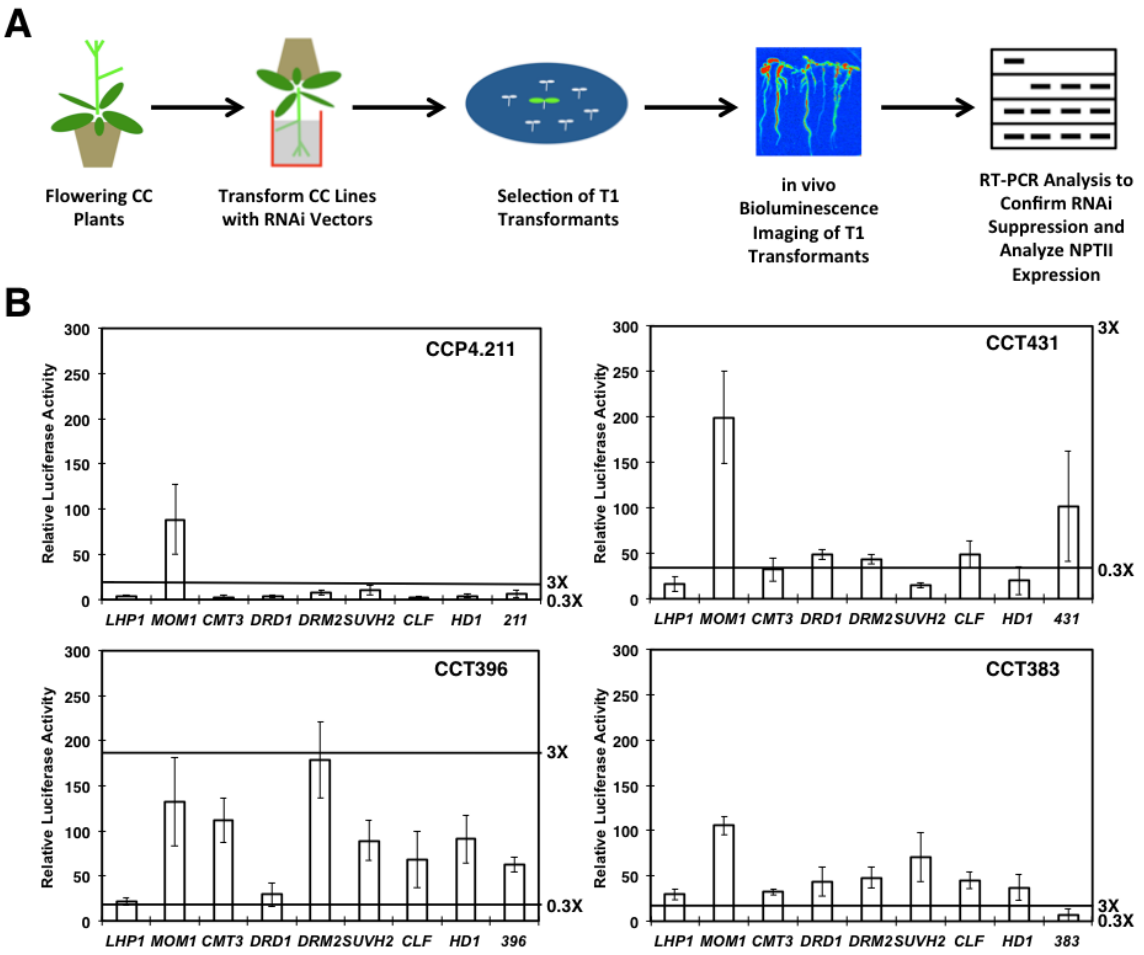

Figure 2.3. RLA changes induced by RNAi suppression of epi-regulators in shoots of CC lines.

(A) The strategy outline for determining the epi-factors that regulate the transgenic locus in CC lines. (B) Changes of RLA induced by RNAi of epi-regulators in the shoot tissue of CC lines. RLAs greater than 3× or less than 0.3x of the parental RLA were considered to be significantly up- or down- regulated respectively. The thresholds were indicated on the right side of each panel.

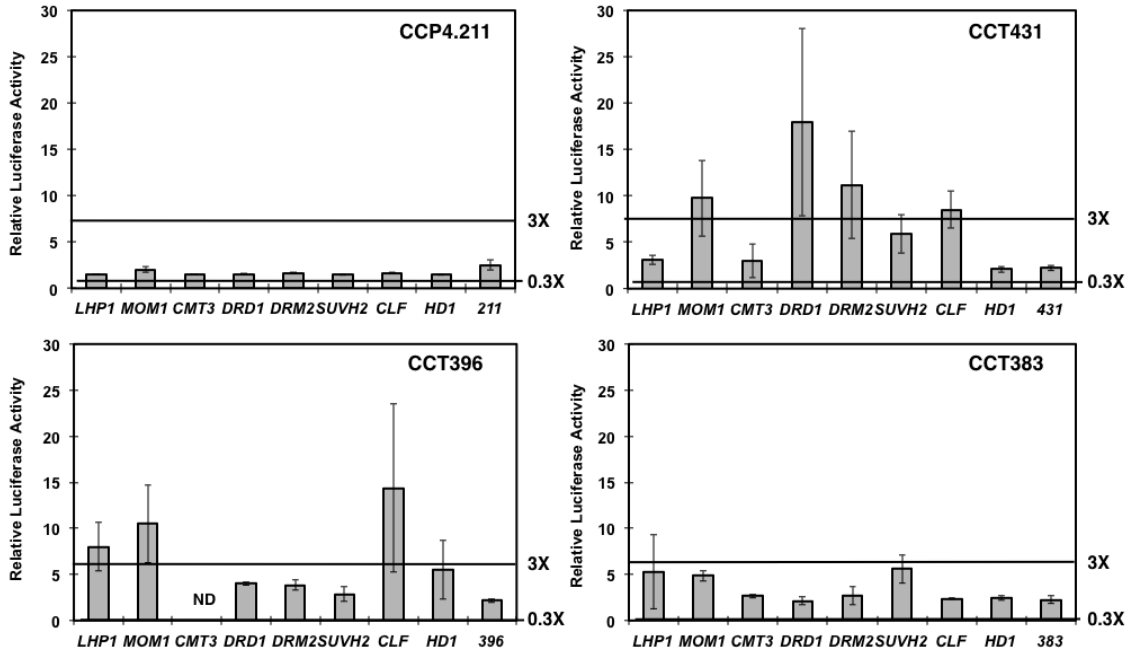

Figure 2.4. Changes of RLA induced by RNAi of epi-regulators in the root tissue of CC lines.

RLAs greater than 3× or less than 0.3x of the parental RLA were considered to be significantly up- or down- regulated respectively. The thresholds were indicated on the right side of each panel.

2.3.4 Evidences for genomic-location dependent epigenetic regulation of transgene expressions

The distinct responses of RLAs to RNAi in the shoots of CCP4.211 and CCT 383 provided a good example for a common transgenic construct being regulated by distinct epigenetic mechanisms. Among the eight tested regulators, only MOM1 was involved in the regulation of Luc gene in CCP4.211 (Figure 2.3B). The CC construct in CCP4.211 located in a heterochromatic region surrounded by rDNA and retroelements, which resembled the endogenous targets of MOM1 (Figure 2.1D). Consistent with the genomic context of the CC construct, Luc expression of CCP4.211 was not dependent on polycomb components such as CLF or LHP1. Although CCT383 showed a similarly repressed RLA as CCP4.211, the suppression mechanism appeared to be much distinct. The Luc expression of CCT383 was inhibited by a plethora of factors including all eight tested epi-regulators (Figure 2.3B). Since the CC construct in CCT383 did not locate adjacent to any repetitive sequences, there was no surprise that PcG proteins (i.e. LHP1 and CLF) together with other pathways contributed to the suppression of Luc expression.

If the epigenetic mechanisms that control CC locus were indeed determined by the integration location and/or the genomic context, we expected that both Luc and NPTII genes would be controlled by similar sets of epi-regulators. To test this conjecture, the transcript abundances of Luc and NPTII were determined in shoot tissues of RNAi lines derived from CCP4.211 and CCT383 (Figure 2.5). Interestingly, the suppression of MOM1 but no other epi-regulators activated NPTII expression in CCP4.211, which resembled the changes of RLA induced by RNAi in this background (Figure 2.5A). CCP4.211-MOM1 also showed corresponding enhancement of kanamycin resistance

compared to the parental line (data not shown). Therefore, Luc and NPTII expressions of CCP4.211 were controlled by a common epi-regulator, which suggested that the regulatory mechanisms of Luc and NPTII were determined by a common signal such as the genomic context.

NPTII expressions were not elevated in any CCT383-RNAi lines, which contrasted the responses of Luc expression to RNAi (Figure 2.5B). This may not be surprising because the NPTII gene of CCT383 expressed at a level similar as the reference line CCP4.20 whereas the Luc expression was considerable lower than in CCP4.20 and resembled the level of CCP4.211. The result indicates that the regulation of Luc and NPTII in CCT383 were independently established. While Luc gene was placed under the suppression by multiple mechanisms, there was little epigenetic silencing of NPTII expression mediated by the tested epi-regulators.

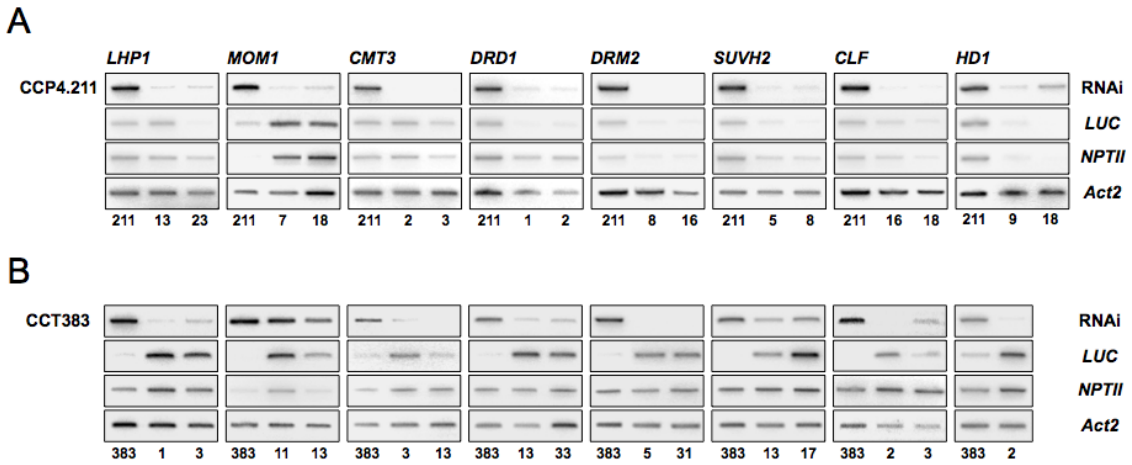

Figure 2.5. RT-PCR quantification of epi-regulator, Luc and NPTII transcripts in RNAi lines derived from CCP4.211 (A) and CCT383 (B).

2.3.5 Functional predictions of the interactions among epi-regulators

Our results have shown that many transgenic or genomic loci are regulated by multiple epi-regulators, sometimes clearly from different pathways. This observation supported our speculation that different pathways may function in a corporative manner to regulate any given gene expression events in the genome. The knockdowns of regulators functioning in the same pathway are likely to produce similar molecular phenotypes (e.g. change of gene expressions). For example, the RNAi of PcG proteins LHP1 and CLF both activated FLC and AG. We reasoned that the physical and genetic interactions of epi-regulators can be implicated by the comparison of molecular phenotypes associated with the RNAi lines. Therefore, we hierarchical clustered the eight epi-regulators based on the RNAi induced expression changes of both transgenic and endogenous loci (Figure 2.6). Importantly, DRD1 and DRM2 that were both known to be involved in the RdDM pathway were readily joint into a cluster because of identical molecular phenotypes (Figure 2.6). Although the power of the current work is limited by the small amount of regulators being suppressed and few loci being tested, our results nevertheless suggested the possibility to predict interactions of epi-regulators through bioinformatics integrations of the molecular phenotype data under the perturbation of regulators.

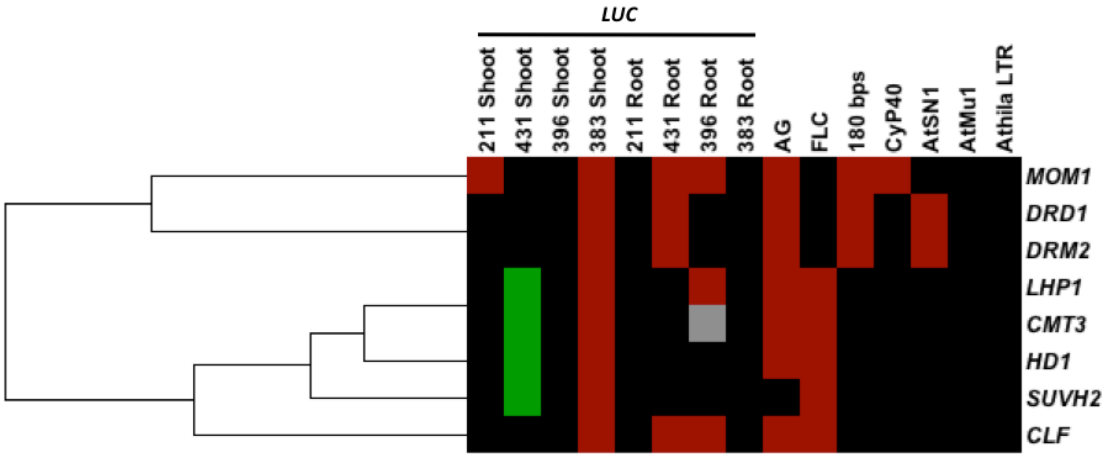

Figure 2.6. Functional predictions of the interactions among epi-regulators.

Heat maps representing qualitative changes of RLA caused by RNAi-mediated silencing of epi-regulators (listed to the right of the heat maps) in four CC lines and two tissue types (listed on the top of heat maps). More than three-fold higher or less than three-fold lower than the RLA of parental CC lines were considered significant up-regulations (red in heat maps) or down-regulations (green in heat maps). Changes in the expressions of endogenous loci were determined by RT–PCR. Black blocks indicate insignificant changes of RLA and expression levels of endogenous loci, while missing data are shown as gray blocks. The gene tree at the left of the heat map represents the similarity between effects (changes of RLA and endogenous loci expression) of RNAi-mediated silencing of different epigenetic regulators. Longer internode distance indicates less similarity while shorter internode distance indicates higher similarity.

2.4 Discussions

Hormone signaling and transcription factor networks are long known to drive organ, tissue or cell type specific gene expressions. In recent years, PcG mechanism was found to be critical for the specific expression pattern of certain developmental regulators, suggesting epigenetic factors may contribute to gene regulations in the context of plant developments. In this work, we expanded the previous discovery by showing that multiple epi-regulators cooperated to direct the specific expression of a transgene in shoots or roots. Both CCT431 and CCT396 showed severe root specific silencing of Luc expression. The RNAi of MOM1, DRD1, DRM2 and CLF led to the reactivation of RLA in the root of CCT431. The root specific repression of RLA in CCT396 was mediated by LHP1, CLF and MOM1. Since RLAs in shoots of CCT431 and CCT396 were not significantly affected in these RNAi lines, the results suggest that the organ specific Luc expression in the two lines can at least be partially explained by the differential regulation of Luc by epi-regulators. Tissue specific epigenetic regulations of Luc expression were also observed in CCP4.211 and CCT383 that showed no apparent tissue specific Luc expressions. MOM1 was involved in the repression of Luc in the shoots but not roots of CCP4.211. Seven of eight tested epi-regulators participated in the suppression of Luc in the shoots of CCT383 whereas none of them showed apparent regulatory function in the roots. Further analyses will be needed to define if endogenous loci can be regulated in a similar fashion by epigenetic mechanisms. In addition, it is unknown how the epigenetic silencing is established specifically in certain tissues. Presumably some epi-regulators can interact with the tissue-type determining transcription network, although little evidence supporting this speculation has been revealed.

The unique resource used in this study – a common transgenic construct integrated at different genomic locations, has allowed the testing of the interdependency between genomic locations and epigenetic regulations. If genomic locations indeed have a major impact on the regulation of gene expressions, we expected that the two marker genes Luc and NPTII that were separated by a ~2kb LacO region would behave similarly regarding to either expression levels or epigenetic regulatory mechanisms. We have observed a general correlation between the expression level of Luc and NPTII in the four CC lines investigated, although CCT383 was an apparent exception showing a low Luc expression but a NPTII expression comparable to the reference line CCP4.20. Through the analysis of CCP4.211, we further found that the Luc and NPTII genes of CCP4.211 were both specifically repressed by MOM1 whereas the two marker genes were regulated distinctly in CCT383. Collectively, we envisage a model that certain location-dependent mechanisms do exist to coordinate the epigenetic regulations of multiple genes located within a genomic region, which led to the correlated expression of Luc and NPTII in CCP4.211, CCT431 and CCT396. The uncoupled expression of Luc and NPTII in CCT383 may be explained by the stochastic cellular defense mechanisms triggered during the transposition of Ds elements, which may have caused the silencing of Luc but not NPTII gene.

The functional interactome approach described in this work has similarities with methods such as connectivity map or MSigDB (Lamb et al., 2006; Subramanian et al., 2005), which evaluate the molecular mechanism of regulators or drugs through profiling molecular phenotypes such as gene expressions. Compared to molecular interactomes that are constructed through yeast two-hybrid or protein chip methods, function

interactomes are derived from indirect information such as gene expression patterns. Nevertheless, the approach offers the advantages of making measurements *in vivo* and the compatibility with large-scale experiments. In support for the feasibility of the approach, we do observed the clustering of some factors that were known to work in the same pathway, such as RdDM regulators DRM2 and DRD1. The amount of test points scored in this work was very small and would not provide sufficient resolution and statistical strength in predicting true protein-protein interactions. However, with the feasibility of profiling global gene expressions under the knockout or RNAi of epi-regulators, we expect the large-scale functional interactome can be a complementary approach for protein interactomes and would provide significant insights into the organization of epigenetic regulatory networks.

3 The development of ANCORP (ANchored CORrelative Patterns) as a platform for epigenomic data integrations and visualizations.

3.1 Introduction

The epigenome of multicellular eukaryotic cells contain more than 50 distinct types of histone modifications, several types of DNA modifications together with other structural variations such as nucleosome positioning and high-order conformations. Therefore comprehensive descriptions of epigenomes require highly sophisticated datasets, which contrasts the linear form of genomic sequence information. Epigenomic and transcriptomic profiling are being carried out in various cell types and conditions, which would further increase the dimension of data to be analyzed. All these complexities have presented substantial challenges to the integration and visualization of epigenomic data. Since numerous studies have supported the extensive interactions between chromatin modifications (Li et al., 2007), it is plausible to speculate that the ensemble of chromatin modifications present in the area instead of individual modifications determines the status of a genomic region. However, classical analytic approaches commonly fall short in assembling multiple chromatin modifications into an integral chromatin state and thus impede patterns extractions and hypothesis formations. Combining the heatmap method and the chromatin state determination through Hidden Markov Model, we have developed ANchored CORrelative Patterns (ANCORP) as a platform for the data integration and visualization. ANCORP method consists of two steps - 1) the generation of an anchored scaffold through ordering annotated genes by a certain criteria such as the integrated chromatin states and 2) the plotting of other features

as correlative patterns using the same order of genes as the anchored scaffold. Several characteristics of heatmap made it highly suitable for the presentation of epigenomic data -1) the usage of color for representing quantitative measurements allowed the display of three dimensions of information in a two-dimension color plot. 2) Through the usage of visualization softwares such as Java Treeview (Saldanha, 2004), heatmap can be easily zoomed in to a single gene or zoomed out to provide a bird eye view of the whole genome. ANCORP has enabled intuitive visualizations and comparisons across multiple epigenomic profiles and facilitated the generation of testable hypotheses. Several hypotheses derived from ANCORP have been supported by independent studies, which suggest that ANCORP can be an effective approach in exploring epigenomic data.

3.2 Material and methods

3.2.1 ChIP-chip and gene expression microarray data

Genome-wide ChIP-chip data for H3K4me3, H3K27me3 and H3K36me2 were downloaded from NCBI GEO GSE7907 (Oh et al., 2008). ChIP-chip data for histone modifications were quantile normalized with histone H3 ChIP-chip using TileMap (Ji and Wong, 2005). TileMap was used for the identification of significantly modified chromatin regions with Hidden Markov Model (HMM) method. Regions greater than 100bp with continuous posterior probability > 0.5 were determined as significantly modified domains. Adjacent domains locate less than 300 bps apart were joined. The global measurement of nucleosome density was determined by comparing histone H3 ChIP-chip with 'Input' sample, which was obtained by the hybridization of *Arabidopsis* genomic DNA to the tiling array. The profile of DNA methylation was downloaded from NCBI GEO GSE5974 (Zilberman et al., 2007). The determination of significantly

methylated domain was performed as previously described (Zilberman et al., 2007). The gene expression data of *Arabidopsis* shoot tissue was downloaded from NCBI GEO GSE9648 (Matsui et al., 2008). The transcriptome profile of wild type Col-0 plants described in the work was used for the current analysis. The transcriptome data was normalized by MAT (model-based analysis of tiling arrays, Johnson et al., 2006). The profile of H2A.Z in *Arabidopsis* root tissue was downloaded from NCBI GEO GSE12212 (Zilberman et al. 2008).

3.2.2 Heatmap visualizations

The chromatin modifications or gene expression patterns between 1kb upstream to 5kb downstream of a gene was represented by 120 bins in matrix for heatmap generations. The value for each bin represents the average value of a 50 bp window. Hierarchical clustering was performed on the data matrix with Cluster 3.0 using single-lineage method (de Hoon et al., 2004). Custom Perl scripts were used to generate matrix for correlative patterns in the same row order as the anchored scaffold. Heatmaps were generated from the data matrix with Java Treeview 1.1.3 (Saldanha, 2004). The two-channel RGB merging of heatmaps was perfomed with ImageJ 1.38 (Abramoff et al., 2004).

3.2.3 Determination of chromatin states at the level of transcription units

Chromatin states of annotated genes were defined by overlapping genic regions with modified chromatin domains identified by TileMap as previously described. The chromatin state for a given gene was represented as four binary digits each indicating the enrichment state of H3K4me3, H3K36me3, H3K27me3 and 5mC respectively. The data

matrix containing the chromatin states of 28,244 genes were hierarchical clustered by Cluster 3.0 with the single-lineage method (de Hoon et al., 2004).

3.3 Results

3.3.1 ANCORP revealed correlations between chromatin modifications

The basic procedure of ANCORP is to first order all or a subset of the genes according to a criterion of interest. For example, 28,244 *Arabidopsis* genes were hierarchically clustered according to the pattern of H3K4me3 between 1kb upstream and 5kb downstream of TSS. The clustering result was displayed in Figure 3.1A and was regarded as the anchored scaffold. Figure 3.1A contains 28,244 rows with each row corresponding to a single gene aligned at TSS. The intensity of H3K4me3 enrichment was represented by colors ranging from blue to yellow. Figure 3.1B-E were generated by plotting the patterns of H3K36me2, H3K27me2, 5mC and transcript abundances with the same order of genes as in Figure 3.1A. We referred to Figure 3.1B-E as the correlative patterns of Figure 3.1A. The correlations between the four chromatin modifications and gene expression can be readily derived from the figures. H3K4me3 and H3K36me2 were enriched in a similar set of genes, which were coincidentally also actively expressed (Figure 3.1A, B and E). It was apparent from the figures that H3K4me3 localized at the 5'- of gene bodies while H3K36me2 peaked towards the 3'- ends (Figure 3.1A and B). The signals for H3K4me3 and H3K27me3 were largely exclusive (Figure 3.1A and C). As expected, genes enriched with H3K27me3 were poorly expressed (Figure 3.1C and E). There was no apparent pattern emerged for 5mC being plotted as a correlative pattern for H3K4me3, suggesting the two chromatin modifications are largely independent (Figure 3.1A and D).

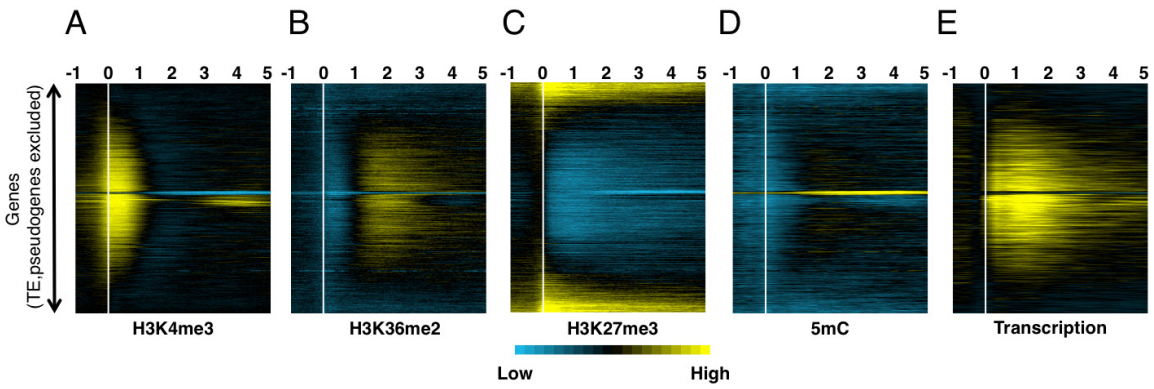

Figure 3.1. Correlative analyses of chromatin modifications and gene expressions using the pattern of H3K4me3 as the anchored scaffold.

Arabidopsis genes were hierarchical clustered according to their H3K4me3 patterns between 1 kb upstream and 5 kb downstream of TSS (A). The patterns of H3K36me2 (B), H3K27me3 (C) , 5mC (D) and sense transcripts (E) were plotted using the same order of genes as in (A). The positions of TSS were indicated by straight white lines.

3.3.2 Multi-channel visualizations identified gene-size dependent patterns of chromatin modifications.

With ANCORP method, we analyzed the correlations between gene lengths and chromatin modifications. The patterns of H3K4me3, H3K36me3, H3K27me3 and 5mC were plotted with genes ordered by their lengths (Figure 3.2A-D). We found H3K4me3 was depleted from the shortest group of genes but the enrichment around TSS was otherwise independent from gene lengths (Figure 3.2A). The depletion of H3K4me3 coincided with the enrichment of H3K27me3 in the shortest group of genes (Figure 3.2A and C). H3K27me3 was known to be established by PRC2 complex and was involved in the repression of developmental regulators. It is therefore intriguing to test if the length of genes can be perceived by certain mechanisms and mediate the recruitment of PRC2. Both H3K36me2 and 5mC appeared to be more abundant in longer genes (Figure 3.2B and D). The two marks differed in their patterns for genes longer than 3 kb. The regions in 3'-end of gene bodies that covered by H3K36me2 were proportional to the lengths of genes, which again suggested the signaling between gene lengths and chromatin modifying processes (Figure 3.2B, Oh et al., 2008). 5mC was depleted within the 1kb region downstream of TSS that was commonly modified by H3K4me3 (Figure 3.2D). This observation is consistent with the discovery that H3K4me3 can inhibit *de novo* DNA methylation through preventing the binding of DNMT3L in mammals (Ooi et al., 2007). The spatial correlation between H3K36me2 and 5mC suggests that the two chromatin marks may perform synergistic or complementary functions in transcription units.

The patterns for pairs of chromatin states were merged to explicitly analyze the spatial relations of chromatin modifications (Figure 3.2E-H). H3K4me3 and H3K36me2

are characteristic markers of early and later stages of transcription elongations respectively. In S.cerevisiae, H3K4me3 depends on Ser-5 phosphorylation of RNA polymerase II C-terminal domain (CTD) while Ser-2 phosphorylation of CTD is indispensable for the H3K36me (Li et al., 2007). Consistently, H3K4me3 and H3K36me2 covered essentially non-overlapping regions in gene bodies (Figure 3.2E). Interestingly, for genes longer than 3kb, substantial portions of gene bodies were modified by neither H3K4me3 nor H3K36me2 (Figure 3.2E). This region was instead decorated by 5mC (Figure 3.2G and H).

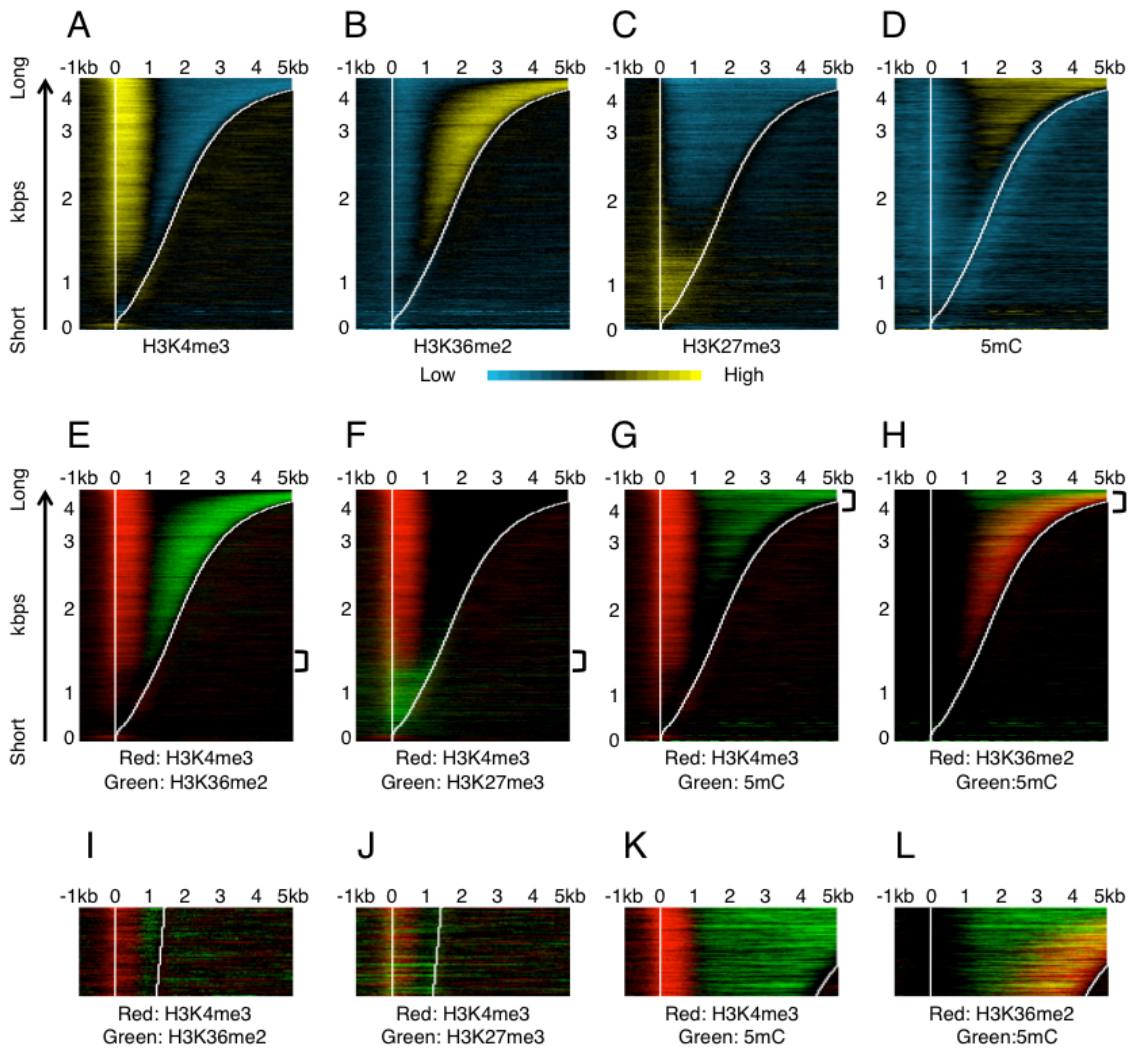

Figure 3.2. Chromatin modification patterns were correlated with gene lengths.

(A-D) Patterns of H3K4me3, H3K36me2, H3K27me3 and 5mC were plotted between 1kb upstream to 5 kb downstream of gene bodies. *Arabidopsis* genes were ordered by their length with the longest and shortest genes placed at the top and bottom of the charts respectively. The straight and curved white lines indicated the TSS and TTS of genes respectively. (E-H) Merged images for pairs of chromatin modifications indicated at the bottom of each image. (I-L) Magnified image for the region indicated with blanket in (E-H).

3.3.3 The determination of chromatin states with four chromatin modifications

To achieve higher order of integrations for chromatin modifications profiles, we defined chromatin states for each annotated genes to describe the ensemble of the four chromatin modifications that decorated the genic region. The 16 chromatins states were visualized by clustering all genes according to their four digits binary chromatin states (Figure 3.3A). Transposons and pseudogenes were excluded from the current analysis due to the generally poor annotation information associated with these types of loci. The clustering results provided substantial insights into the global organization of *Arabidopsis* chromatin states. The genes that were modified by H3K36me2 were essentially a subset of genes that were enriched of H3K4me3 (Figure 3.3A). 87.6% (10,284/11,737) of H3K36me2 modified genes also associated with H3K4me3 (Figure 3.3A). The observation is consistent with the notion that H3K36me2 is a characteristic mark for active chromatin and H3K4me and H3K36me are both guided by the phosphorylation states of Pol II CTD and Polymerase Associated Factors (PAF, Li et al., 2007). The vast majority (90.6%) of actively expressed genes that were associated with 5mC were also marked by H3K36me2 (Figure 3.3A). The apparent overlapping between genes modified by H3K36me2 and 5mC was consistent with that both H3K36me2 and 5mC were enriched in longer genes, although the biological significance of this overlapping was not known. H3K27me3 clearly occupied distinct gene spaces compared to the other three chromatin modifications (Figure 3.3A). Only 22.8% (1,572/6,904) of genes modified by H3K27me3 also associated with other chromatin modifications. The state that was simultaneously enriched of H3K4me3 and H3K27me3 (K4/K27me3 bivalent) is particular interesting. K4/K27me3 bivalent state is prevalent in mammalian ES cells and

marked many important developmental regulator genes (Bernstein et al, 2006; Mikkelsen et al, 2007). The bivalent state primarily associates with repressed expressions and paused RNA polymerases and may resolve into K4me3 or K27me3 monovalent states during differentiations (Bernstein et al, 2006; Mikkelsen et al, 2007). We have determined 630 genes that were associated with the K4/K27me3 bivalent state. However since the leaf tissue used for the ChIP-chip experiments contained many different cell types, it remained to be tested if the K4/K27me3 bivalent state indeed represent chromatin modified by the two marks instead of superimposing patterns observed in distinct cell types.

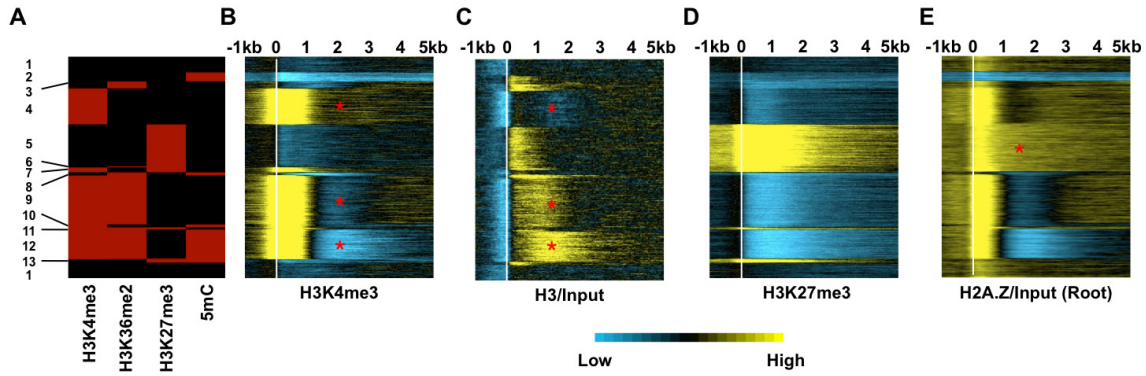

Figure 3.3. Predicting interactions among chromatin modifications with chromatin states as the anchored scaffold.

(A) Chromatin states determined by four chromatin modifications – H3K4me3, H3K36me2, H3K27me3 and 5mC. The clustered image contained 28,244 rows each correspond to the chromatin states of an annotated gene. Significant enrichment of a chromatin modification was represented by red while black indicated insignificant associations. 13 chromatin states that were visible from the image were each given a state number indicated on the left. (B-E) The patterns of H3K4me3, histone H3, H3K27me3 and H2A.Z between -1 kb upstream of 5 kb downstream of TSS were visualized as correlative patterns of (A). Red asterisk in the figures indicates the regions being discussed in the text.

3.3.4 The prediction of the interactions between chromatin modifications with ANCORP

In order to explore if the pattern of individual chromatin modifications was modulated by the integrated chromatin states, each chromatin modification was visualized quantitatively as the correlative patterns of the chromatin states defined by the four modifications (Figure 3.3). 1) Distinct patterns of H3K4me3 were observed between states 4, 9 and 12. Although the three states associated with comparable level of H3K4me3 around TSS, state 4 showed significantly greater H3K4me3 enrichments in 3'-gene body regions compared to state 9 and 12 (Figure 3.3B, red asterisks). Given that state 9 was enriched of H3K36me2 and state 12 was further enriched of 5mC, the correlations suggest that H3K36me2 and 5mC may synergistically repress H3K4me3 in gene bodies. 2) Chromatin states showed substantially differential nucleosome density in gene bodies. States 9 and to a further extent state 12 associated with much greater nucleosome density compared to state 4 as measured by histone H3 ChIP-chip (Figure 3.3C, red asterisks). Since nucleosome density, to a substantial extent, is determined by DNA sequence (Kaplan et al., 2009), we hypothesized that the nucleotide composition of genes in state 9 and 12 may enhance nucleosome depositions in gene body regions and subsequently promote the level of H3K36me2 and 5mC in the corresponding regions. 3) A conspicuous enrichment of H2A.Z was observed in state 5 that was highly modified by H3K27me3 and likely to be regulated by PcG proteins (Figure 3.3E). This correlation led us to speculate a regulatory interaction between H2A.Z and polycomb repressive complexes.

The three hypotheses we proposed were highly testable by genetic or biochemical approaches. For example, H3K4me3 enrichments at specific loci can be compared between wild-type and mutants that are deficient of H3K36me2 or 5mC with ChIP-qPCR assay to test the hypothesis 1. However, due to the limitation of time and resource, we performed literature searching to evaluate if the speculations were reasonable and further gauge the efficacy of ANCORP in deriving testable hypotheses. A human transcription factor MRG15 containing a chromodomain was shown to be capable of binding methylated histone H3K36 (Zhang et al., 2006a). Importantly, RBP2 – a jmjC domain protein that has histone H3K4 demethylase activity was identified as a component of MRG15 containing complex (Hayakawa et al., 2007). The two works together delineated a regulatory pathway that H3K36me can recruit RBP2 through MRG15 and subsequently remove H3K4me3 from 3'-gene bodies. This pathway was later validated using a human cell culture model and was shown to modulate transcripts splicing (Luco et al., 2010). Although the interaction between H3K36me2 and H3K4me3 remain to be tested in plants, the results in mammals supported that highly testable and valuable hypotheses can be generated by ANCORP method. Similarly, Chodavarapu and coauthors independently described the positive correlation between 5mC and nucleosome density (Chodavarapu et al., 2010). Their work further showed that DNA methylation exhibited a clear 10-base periodicity in both plants and human, which is consistently with the speculation that nucleosome positioning directs DNA methylation presumably through the structure of DNMT3A/DNMT3L heterodimer (Chodavarapu et al., 2010). Lastly, H2A.Z was found to be significantly co-localized with PcG protein Suz12 at important developmental regulator genes in mammalian ES cells (Creyghton et al., 2008). In addition, the binding

of PcG protein to their target and the incorporation of H2A.Z were interdependent to each other (Creyghton et al., 2008). Therefore, the result supported our speculation derived from ANCORP analysis that certain interactions may exist between H2A.Z and polycomb regulations.

3.4 Discussions

We consider ANCORP can improve or supplement existing analytical tools for epigenomic data in three major aspects - 1) By defining chromatin states at the transcription unit level and visualizing the quantitative measurement of chromatin modifications as correlative patterns, we were able to made observations that were specific for certain chromatin states. As we have shown, the state solely modified with H3K4me3 differed from the H3K4me3+H3K36me2 state in multiple aspects including greater H3K4me3 enrichments and lower nucleosome densities in 3'-gene bodies. These potentially functional relevant correlations would not be uncovered if given only the global distribution of H3K4me3 and H3K36me2 marks without the parsing of chromatin states. 2) The determination of chromatin states identified infrequent states that may reflect novel regulatory mechanisms. We have identified the states H3K36me2 and H3K36me2+5mC each containing less than 300 genes (Luo and Lam, 2010). Genes associated with these states were generally repressed and represented the examples that H3K4me3 was uncoupled from H3K36me2. Our recent works further showed that antisense transcripts were strongly enriched in these two states (see section 6 and 7). It would be intriguing to investigate the mechanism through which these states were established and the potential biological impacts. 3) The expandability of ANCORP –

multiple chromatin modifications being analyzed in a row enabled the cross-reference of epigenomic features that have little or no *a priori* biochemical or genetic information.

At this stage of the development, ANCORP also faced critical challenges that need to be addressed. Along with the increase of chromatin modification profiles being incorporated, the amount of defined chromatin states and the related visual comparisons would soon become unaccountable. In addition, the biological significance of the massive defined stages may be compromised due the accumulating statistical noises. To overcome these potential difficulties, certain fuzzy statistical methods may be used to reduce the dimension of chromatin states. Informatics tools for automatic pattern comparisons may also be helpful to substitute human eyes for the tedious task of correlation discovery and hypothesis constructions. Some of the extensions and improvements of the ANCORP method will be discussed in sections 6 and 7.

4 Development of ChIP-seq for the genome-wide profiling of histone modification patterns

4.1 Introductions

Chromatin immunoprecipitation (ChIP) followed by the microarray hybridization (ChIP-chip) or high throughput sequencing (ChIP-seq) are methods to interrogate the genome-wide localizations of particular types of chromatin related epitopes. These epitopes are typically specific modifications on histone proteins that are known to be epigenetically relevant, or sequence-specific DNA-binding proteins such as transcription factors. Moreover, ChIP-derived strategies have also been devised to "pull-down" cytosine-methylated DNA by using a 5mC-specific antibody and have been used successfully to display the genome-wide methylation landscape (Zhang et al., 2006c; Zilberman et al., 2007). The major steps of ChIP include crosslinking DNA with bound proteins and the subsequent enrichment of DNA associated with a particular protein by a specific antibody. The utilization of genomic tilling array or high throughput sequencing enables the global survey of DNA samples produced by ChIP. ChIP-seq has several advantages over the ChIP-chip method especially for large genomes. 1) ChIP-seq generally requires substantially less ChIPed DNA than ChIP-chip. A full flowcell of sequencing with SOLiD^{TM} 2.0 generating 400 millions tags needs less than 5 ng of library DNA. For ChIP-chip, ChIPed sample need to be amplified to generate more than 2 μg of DNA for each array (Park, 2009). Therefore with equal amounts of ChIPed samples, ChIP-seq requires less cycles for library amplifications that usually lead to superior quantifications over ChIP-chip. 2) ChIP-seq posses greater dynamic ranges as

compared to ChIP-chip. The dynamic range of any microarray platform is limited by background hybridization noises at the lower end and signal saturations at the higher end. In contrary, ChIP-seq can detect input DNA species at any abundance with high confidence that is limited only by the quantity of good quality sequence reads. 3) ChIP-seq has the special advantage to sequence multiplexed samples in one run through the 'barcode' or 'index' design, which is impossible for any microarray platform. Together with the ever-decreasing cost with concomitant increase in throughput for sequencing, ChIP-seq can be a more economical and rapid technique than ChIP-chip for obtaining the same amount of data with equivalent or better qualities.

We have developed the procedure to profile histone modifications in the *Arabidopsis* genome by using the SOLiD™ 2.0 or 3.0 sequencing platform. We aimed to develop an economical procedure without relying on commercial kits so that it would be affordable even for generating large-scale multiplexing libraries. As the quality and cost of performing ChIP-seq are highly related to the read length and sequencing depth, we have evaluated the impact of read length and sequencing depth on the quality of ChIP-seq experiment and provided our suggestions for the two parameters.

4.2 Materials and methods

4.2.1 Antibodies for ChIP-seq

H3K4me2 (Millipore, cat. # 07-030, lot. # DAM1503382), H3K4me3 (Millipore, cat. # 07-473, lot. # 27343), H3K9Ac (Millipore, cat. # 07-352, lot. # 31388), H3K9me2 (Abcam, cat. # ab1220, lot. # 625300), H3K18Ac (Abcam, cat. # ab1191), H3K27me1 (Millipore, cat. # 07-448, lot. # DAM1598790), H3K27me3 (Millipore, cat. # 07-449, lot.

DAM1514011), H3K36me2 (Abcam, cat. # ab9049-100, lot. # 614453), H3K36me3 (Abcam, cat. # ab9050-100, lot. # 573603).

4.2.2 ChIP (chromatin immunoprecipitation) assay

Prepare 25ml of nuclear isolation buffer (10mM Hepes pH=7.6, 1M sucrose, 5mM KCl and 5mM $MgCl_2$) at room temperature by adding 700 µl of 37% v/v formaldehyde, 750 µl 20% Triton X-100, 25 µl β-mercaptoethanol and 50 µl 0.2M PMSF. 0.5-1.0 g of *Arabidopsis* leaf tissue was homogenized by grinding in liquid nitrogen. The frozen powder was immediately transferred into the prepared nuclear isolation buffer and mix by gentle inversions. After 10 min of incubation at room temperature, add 1.7 ml 2M glycine and incubate for another 5 min to inactivate formaldehyde. Remove the debris by pass the homogenate through one layer of miracloth (Calbiochem, cat. # 475855-1R). Pellet the nuclei by centrifuge at 3000 × g for 10 min at 4°C. Resuspend the pellet in 300 µl nuclear isolation buffer and load onto 500 µl of nuclear separation buffer (10mM Hepes pH=7.6, 1M sucrose, 5mM KCl, 5mM $MgCl_2$, 5mM EDTA pH=8.0 and 15% Percoll). Separate nuclei from residual chloroplast by centrifuge at 3000 × g for 5 min at 4°C. Thoroughly resuspend the nuclei pellet in 600 µl of nuclear lysis buffer (50mM Tris-Cl pH=7.5, 1% SDS and 10mM EDTA pH=8.0). The fragmentation of chromatin was performed by sonicating the nuclear lysate 7 times with a Branson S150 sonicator with power setting 2. Each round of sonication lasts for 10 seconds and the sample was chilled on ice for several minutes between sonications. After the sonication, remove the debris by centrifuge at 13,000 × g for 3 min at 4°C and discard the pellet. Dilute the chromatin sample to 6 ml with ChIP dilution buffer (15mM Tris-Cl pH=7.5, 1% Triton X-100, 150mM NaCl and 1mM EDTA). For each ChIP reaction, 1ml of chromatin sample was

incubated with 2 μg of primary antibody and 30 μl of protein A agarose beads with rotation at 4°C for 2 hrs. The washing of protein A agarose beads was performed three times, each time for 10 min with ChIP dilution buffer at 4°C. Agarose beads were collected by centrifuge at 100 × g for 1 min at 4°C between washes. The beads were finally washed once briefly with 1 × TE (10mM Tris-Cl pH=7.5, 1mM EDTA pH=8.0). Bound chromatin was eluted from beads with 500 μl of ChIP elution buffer (0.1% $NaHCO_3$, 1% SDS) for 30 min at 65°C. Tubes were inverted every few minutes to keep the beads suspended. Transfer the elute into a fresh tube and add 20 μl of 5M NaCl. Perform the reverse-crosslinking by incubate the tube at 65°C overnight (more than 12 hrs). After the reverse-crosslinking, digest the protein with 20 μg of proteinase K for 2 hrs at 50°C. The ChIPed DNA was purified by phenol/chloroform extraction followed by ethanol precipitation.

4.2.3 Preparation of ChIP-seq libraries for SOLiD™ sequencing

Barcoded adaptors for generating sequencing libraries were ordered as single – strand custom oligos at desalted grade and were dissolved with nuclease-free water to a final concentration of 100 μM. Mix oligos corresponding to each strand of the adaptor with equal amount in a PCR tube. Perform the annealing with the presence of 1X PCR buffer using the following program on a thermo-cycler - 95°C 5 min - 72°C 5 min - 60°C 5 min - 50°C 3 min - 40°C 3 min - 30°C 3 min - 20°C 3 min - 10°C 3 min - 4°C ∞. Gel purification was essential for adaptors as the oligos were synthesized at the desalted grade and contains large amount of incompletely synthesized oligos. Purification of the annealed adaptors was performed by running 10 μg of the each barcoded adaptors on a 3% agarose gel. The predominant 50-nucleotide band was recovered with the QIAquick

Gel Extraction kit and elute in 30 µl. Adjust the concentration of P1 and P2 adaptor to 250 ng/µl and 500 ng/µl respectively and store at -20°C until ready for the ligation.

The fragmentation of chromatin with a probe sonicator, such as Branson S150 generate chromatin fragment with size range from 300-500 bps. The optimal insertion size of SOLiD™ 2.0 library is around 200 bps. Therefore the ChIPed DNA need to be further fragmented by the Covaris S2 before the ligation. Dilute each ChIPed DNA sample to 300 µl with water and transfer the entire sample into a T6 round bottom glass tube (Covaris Inc). Fill the tube with water to prevent the formation of air bubble during the Covaris treatment and snap on the cap. Mount the tube onto a Covaris S2 instrument and run the following program

Treatment 1 for 5 seconds		7 cycles
Duty cycle	0.5%	
Intensity	8	
Cycles/Burst	50	
Treatment 2 for 60 seconds		
Duty cycle	20%	
Intensity	8	
Cycle/Burst	200	

Transfer the sample into a clean microcentrifuge tube with an extended gel loading tip. The DNA was precipitated with 1 µl of glycogen, 1/10 volume of NaOAc and 2 volume of 100% ethanol. Wash with 80% ethanol and air dry the pellet before resuspend with 34 µl of H_2O.

The sonicated DNA samples were converted to phosphorylated blunt end molecules with End-It™ DNA End-Repair Kit (Epicentre) following the product instruction. Purify the End-repaired DNA with MinElute Reaction Cleanup Kit. Elute with 20 µl of EB buffer supplied in the kit. Using 10 µl of sample DNA, perform ligation with 1 µl of P1 adaptor (250 ng/µl) and 1µl Barcoded P2 adaptor (500 ng/µl) using Fast-Link™ DNA ligation kit (Epicentre). Extract the ligation reaction with Phenol/Chloroform/IAA (25:24:1) to remove the DNA ligase. Precipitate the ligation product with 1 µl of glycogen, 1/10 volume of NaOAc and 2 volume of 100% ethanol. After washing once with 80% ethanol, resuspend the pellet in 10 µl of H_2O. Load the entire ligation product onto a 6% native polyacrylamide TBE gel for the size selection of ligation products. Excise the gel piece containing 150-200 bp fragments on a UV-light box and vertically cut the gel piece into three slices. Use one gel slice for determining cycle numbers for library amplification and use the rest two slices for the library generation. It is important that the gel piece is sliced vertically rather than horizontally to ensure that DNA contained in the three pieces have identical size distribution. Gel slices can be stored at -20°C until ready to use.

To test the optimal condition for library amplification, transfer one gel slice into a PCR tube and crush the gel piece with a 200 µl pipetting tip. Set up a 100 µl PCR reaction using the gel slice as the template. Starting from the extension step of cycle 14,

pause the thermal cycler 2 to 3 seconds before the extension step finish and transfer 10 µl of the reaction into a clean tube. Repeat this procedure every two cycles until the PCR reaction is completed. Run all the PCR products after different cycles of amplification on a 6% native polyacrylamide gel and visualize through ethidium bromide staining. Choose the least cycle number with a robust library product appearing around 200 bps for the library amplification. Perform library amplifications with the cycle number selected for each library. Ethanol precipitate the whole PCR reaction and resuspend the pellet in 20 µl water. Load the entire product onto one lane of a 3% agarose gel and purify the band around 200 bps with QIAquick gel extraction kit (Qiagen). Elute the libraries in 15 µl to maximize the DNA concentration for accurate quantifications. Use 1 µl of each of the purified libraries to determine their concentration with a Nanodrop spectrophotometer.

4.2.4 Bioinformatic analysis of SOLiD^{TM} sequencing results

The ChIP-seq libraries were sequenced for 35 bps of color-codes. Barcode sequences were recognized allowing no mismatch. Sequencing tags were mapped to *Arabidopsis* TAIR8 reference genome using Corona Lite 4.2.1 (Life Technologies). Up to 3 mismatches were allowed for each 35 bps sequencing tag. Tags that match multiple genomic locations were randomly assigned to one of the candidate position. To correct for potential over-amplifications of the library, redundant tags (multiple tags mapped to a identical position) were removed except for one. For generating the genome-wide coverage patterns of ChIP signals, each tag was extended to 100 bps from its 3'- end to reflect the actual insertion size for the ChIP-seq libraries.

4.3 Results

4.3.1 ChIP-seq generated genome-wide profiling of nine histone modifications

The ChIP-seq experiments, including two and half flowcells of SOLiD™ sequencing, generated genome-wide profiles of nine histone modifications together with histone H3 and the input DNA (Table 4.1). Although each flowcell of sequencing would theoretically produce about 400 millions of tags, we only obtained on average 2-3 million non-redundant tags for each histone modification. The drastic reductions of useful tag quantities can be attributed to several factors – 1) Among the ~400 millions beads that were deposited onto the slides, roughly 60% of the beads had acceptable signal qualities. Poor signal qualities may be caused by closely located beads, which could lead to the mix-up of fluorescence signals. Polyclonal beads that associated with more than one type of DNA molecules can also contribute to un-interpretable sequencing signals. In addition, a fraction of beads that failed to amplify during emulsion PCR may have escaped from the beads enrichment procedure. These beads would not generate any signal during the sequencing. 2) In general about or less than 50% of tags with acceptable signals can be mapped onto the reference genome. Polyclonal beads can again be a substantial contributor for the issue, since the mixed sequencing result derived from multiple DNA copies would not match any location in the genome. We have also noticed that a considerable amount of un-mappable tags were caused by the contamination of libraries by PCR products generated from hetero- adaptor dimers. Although non-phosphorylated blunt-end adaptors are incapable ligating with each other, partial degradations of adaptors may cause the exposure of 5'-phospher and enabled dimer formations. 3) The nine ChIP-seq libraries had highly variable levels of redundancies. The libraries for H3K9Ac and

H3K27me3 had little redundancy - more than 80% of the sequencing tags from the two libraries corresponded to distinct DNA molecules. However some other libraries such as H3K36me2 or H3K18Ac were highly redundant (Table 4.1). On average each identical DNA copy were represented 5-7 times in the library of H3K36me2 or H3K18Ac, which drastically reduced the amount of useful tags. Insufficient starting DNA material and/or over-amplification of the library are potential causes of the redundant libraries.

Table 4.1. Summary of the ChIP-seq experiment

Histone modification	Numbers of mappable Tags (millions)	Non-redundant tags (millions)
H3K4me2	7.1	2.3
H3K4me3	7.7	2.2
H3K9Ac	1.9	1.6
H3K9me2	3.3	0.9
H3K18Ac	15.4	3.6
H3K27me1	8.6	1.1
H3K27me3	2.4	2.2
H3K36me2	10.1	1.8
H3K36me3	15.2	3.7
H3	19.9	5.8
Input	19.1	7.3

4.3.2 Evaluating the impact of sequencing depth on the profile of chromatin modifications.

The sequencing depth that will be needed to quantitatively profile a chromatin modification is dependent on two factors – 1) the size of the reference genome. A larger reference genome such as human or maize genome would require more tags to cover compared to smaller genomes such as that of *S. cerevisiae* or *Arabidopsis*. 2) The pattern of the chromatin modification being profiled. For example, profiling a spatially focused chromatin mark such as H3K4me3 that clusters around transcription start sites would not require many tags even for a large genome. In contrast, ChIP-seq of histone H3 for measuring genome-wide nucleosome density may need a much large number of tags to avoid sampling biases.

Since the sequencing depths of 2-3 million tags for each histone modification may not be considered 'ultra-deep' with current standard, we aimed to address if the depth would be sufficient to quantitatively describe the pattern of histone modifications. We have empirically evaluated the impact of variable sequencing depth on the resulting patterns of chromatin modifications. For a region of *Arabidopsis* chromosome 4, we compared the patterns of H3K36me3 that were generated with 500K, 1 million, 2 million and 3 million tags (Figure 4.1). The profiles have been scaled and presented as 'coverage per million tags' in order to compensate for the differences caused solely by total tag numbers. Except for a few minor differences as indicated by arrows in Figure 4.1 and the "noisier" outline of pattern in the 500K tag dataset, the patterns are essentially identical between the four panels. Similar results were obtained by analyzing other genomic regions or histone modifications (data not shown). We also compared the amount of

H3K36me3 peaks that can be significantly ($p<10^{-5}$) identified with different amount of ChIP-seq tags. The amount of identified peaks increased steadily along with the increase of sequencing depth until ~ 2 million tags and reached plateau between 2 and 3 million tags (Figure 4.2). Therefore, the result is consistent with the pattern comparison in Figure 4.1 and together showed that 2-3 million tags are sufficient to quantitatively describe histone modifications for the ~120 MB *Arabidopsis* thaliana genome.

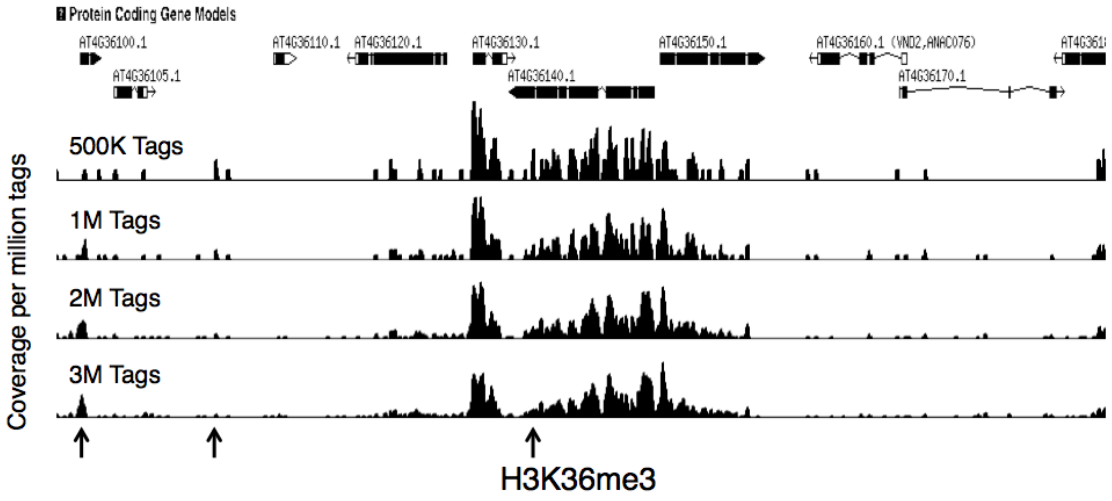

Figure 4.1. Evaluation of the effect of sequencing depth on the profiled pattern for the H3K36me3 mark.

Coverage of this genomic region by sequencing tags (H3K36me3 ChIP) is represented by 'coverage per million tags'. The profiles of H3K36me3 were generated by using 500K, 1 million, 2 million and 3 million of sequenced tags. The scales of the Y-axis are identical across the panels.

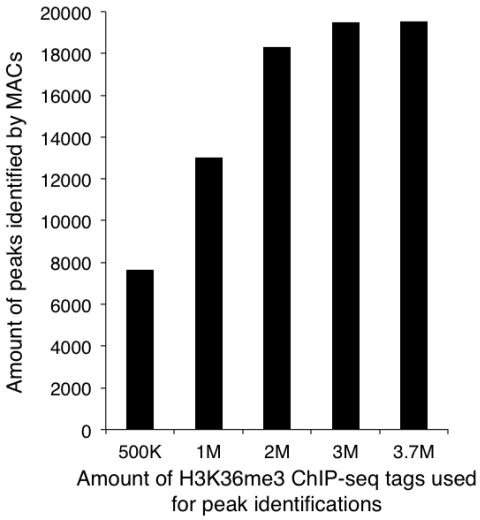

Figure 4.2. Amounts of H3K36me3 peaks identified by MACs using p value equals to 10^{-5} with 500K, 1M, 2M, 3M and 3.7M of randomly chosen ChIP-seq tags.

4.3.3 Evaluate the effect of read length on the alignment to *Arabidopsis* genome

The read length of sequencing tags is critical for the accuracy of aligning tags onto the reference genome. Shorter tags are more likely to match multiple positions in a genome and thus decrease the fidelity of the alignment. Longer tags, instead, would increase the cost of sequencing while they may not necessarily improve the confidence of the alignment. In addition, unlike RNA-seq where longer reads could provide useful information such as splice site variations, tag sequences in a ChIP-seq experiment would be identical with the reference genome and longer reads than necessary are usually of little value. The minimal read length that can generate satisfying alignment results is dependent on multiple factors including the genome size and the amount of repetitive sequences in the genome. Therefore it is essential to determine the relation between read length and alignment accuracy for each genome of interest in order to optimize the target tag sequence length in a ChIP-seq study.

To estimate the read length that is sufficient for aligning tags to the *Arabidopsis* genome, we plotted the read length versus the percentage of reads that were unique in the genome (Figure. 4.3). We found that more than 90% of all possible 26 nucleotide (nt) reads can be uniquely placed in this genome and the increase of this percentage was relatively minor for reads longer than 26 nts. Notably, for reads that are 50 nts long, there were still 5.9% of reads that were not unique in the genome. This analysis showed that sequence reads with length between 30 (91% are unique) and 35 (92.1% are unique) nts would be suitable for ChIP-seq with the *Arabidopsis* genome.

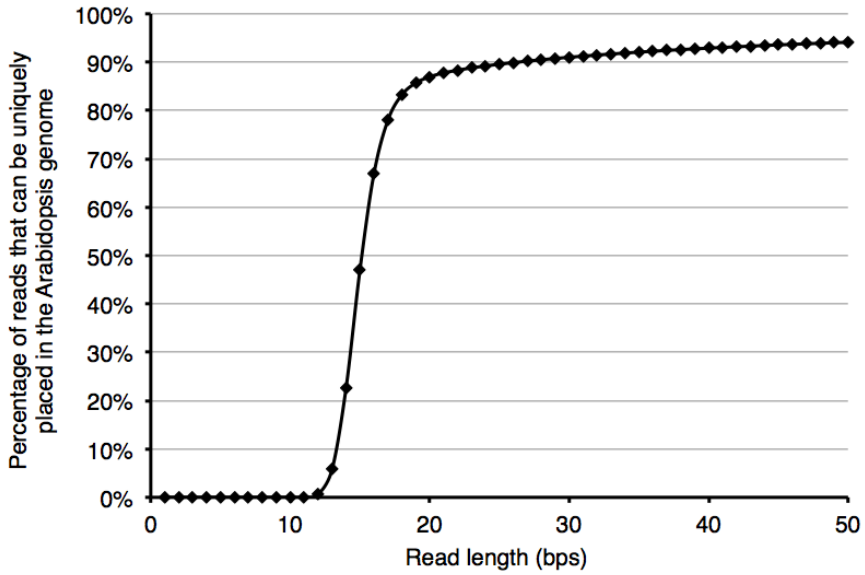

Figure 4.3. Evaluation of the effect of read length on the alignment with *Arabidopsis* genome.

The percentage of reads that can be matched to a unique position in the genome was calculated for reads with length from 1 to 50 bps.

4.3.4 Comparison between ChIP-seq profiles and published ChIP-chip profiles.

Several histone modifications have been profiled globally with ChIP-chip using the Affymetrix *Arabidopsis* Tiling 1.0 array. The profiles of H3K4me3 and H3K27me3 generated from published ChIP-seq results were compared with those generated by ChIP-chip (Figure 4.4, Oh et al., 2008). Both work used aerial tissues from 2-week-old *Arabidopsis* plants. We plotted the patterns of H3K4me3 and H3K27me3 marks as profiled by ChIP-seq or ChIP-chip in a 50 kbp region surrounding the *Agamous* locus. The profiles generated by the two techniques were highly similar and "sharper" ChIP-seq signals can be found at every H3K4me3 or H3K27me3 peaks that were detected with ChIP-chip, which suggests that ChIP-seq effectively captured the patterns that can be identified by ChIP-chip.

ChIP-seq may allow more accurate quantifications as shown by the differential enrichments of H3K4me3 between peaks A and B. An ~3 fold difference regarding the amplitude of peaks was detected between peaks A and B with ChIP-seq, whereas the two peaks were shown to be nearly equal when assayed with ChIP-chip even considering that the ChIP-chip results were presented in a logarithm form. A similar case was also shown by comparing peaks C and D. In addition, ChIP-seq results apparently showed greater signal to noise ratios as the quantification was not affected by background florescence signals as is the case for microarrays. For example, region E corresponded to the transcribed region of *Agamous* and was enriched with the H3K27me3 mark, which was also expected to be devoid of the H3K4me3 mark. Although few tags were found in region E for the H3K4me3 mark by ChIP-seq, ChIP-chip data for H3K4me3 has yielded a substantial amount of sporadic signals in this region. Overall, although we cannot rule

out that different antibodies and plant growth conditions caused the observed distinctions between ChIP-seq and ChIP-chip results, we can nevertheless conclude that ChIP-seq is capable to generate data with the quality equal or better than the ChIP-chip technique.

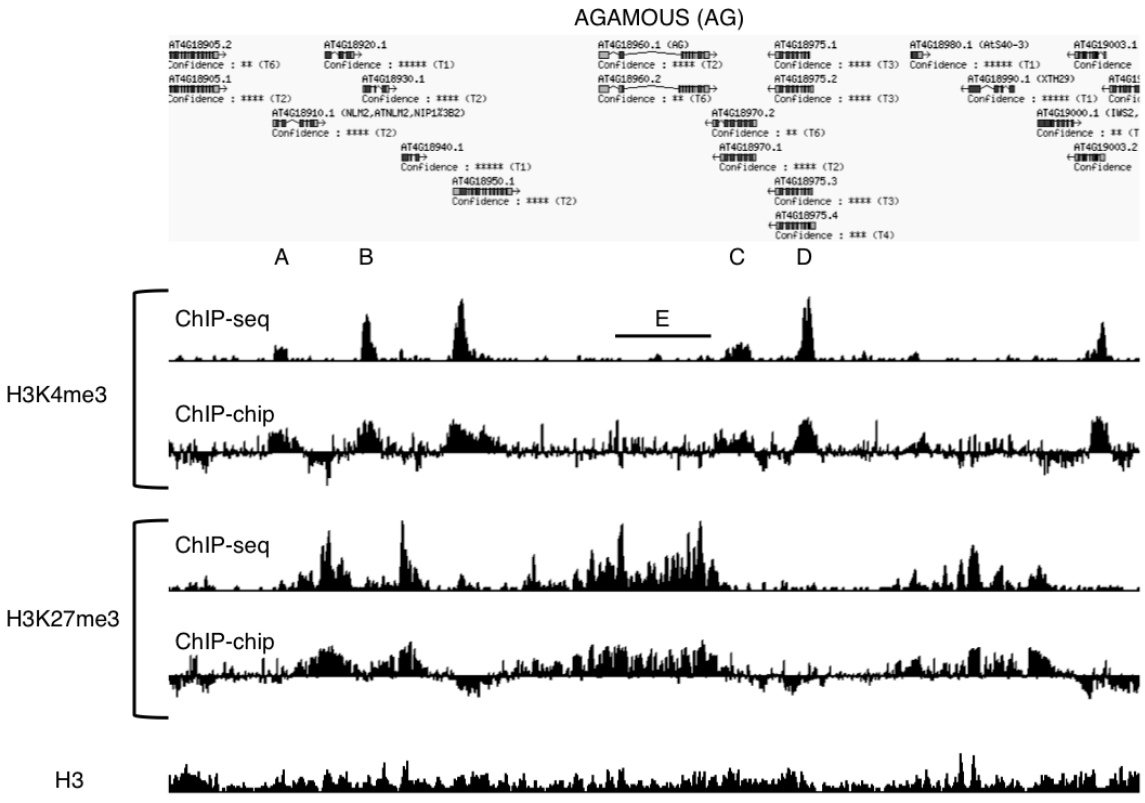

Figure 4.

Figure 4.4. Comparison of H3K4me3 and H3K27me3 profiles generated with ChIP-seq and ChIP-chip.

H3K4me3 and H3K27me3 patterns profiled with ChIP-seq and ChIP-chip were plotted for a 50 kbp region surrounding the *Arabidopsis Agamous* locus. Signals in ChIP-seq and ChIP-chip panels each indicate coverage by sequencing tags and log2 (H3K4me3 or H3K27me3 / H3) respectively. A H3 ChIP pattern produced by ChIP-seq was shown to control the nucleosome density in this region.

5 Discovery of Natural Antisense Transcripts (NATs) with strand-specific RNA-seq

5.1 Introductions

Genomes of eukaryotes are long known to produce natural antisense transcripts (Faghihi and Wahlestedt, 2009). However the global patterns of these enigmatic molecules were not known until the relatively recent development of microarray or RNA-seq methods. Antisense transcripts can be profiled by producing cRNA from the first-strand cDNA and hybridizing the cRNA with microarrays. However, since commonly used reverse transcriptase (RT) associate with a cryptic DNA-dependent-DNA-polymerase activity, second-strand cDNA and later the corresponding cRNA may occasionally be made and contribute to certain antisense signals (Perocchi et al., 2007). The impact of this cryptic activity of RT may vary depending on experimental designs (e.g. use random hexamer or oligo(dT) as the priming oligo, Matsui et al., 2008). In addition to the potential artifacts in the labeling procedure, microarray approaches suffer from signal noises, which may prevent the discovery of NATs with low abundance. Strand-specific RNA-seq experiment is a promising approach for the discovery of NATs and several different methods for library preparations have been developed (Levin et al., 2010). A comprehensive evaluation of library preparation strategies has identified that RNA-adaptor ligation and the dUTP second strand marking as the leading methods in terms of the specificity for identifying the origin strand for RNA-seq tags (Levin et al., 2010). The SOLiD™ Whole Transcriptome Kit protocol that used for our library preparation utilized a variation of the classical RNA-adaptor ligation method and is expected to offer comparable or superior performance.

Previous efforts in determining the antisense transcriptome of plants were performed with microarray platforms. Multiple works have reached a similar conclusion that roughly 30% of annotated loci were associated with NATs (Yamada et al., 2003; Stolc et al., 2005; Li et al., 2006; Matsui et al., 2008). In a recent analysis of *Arabidopsis* NATs, the full-length sequence of RD29A and CYP707A1 associated NATs were cloned and found to share the identical exon/intron junctions as the sense transcripts (Matsui et al., 2008). Further, using a transgenic line containing a T-DNA inserted in the 3'-UTR of CYP707A1 gene, the authors demonstrated that the NAT associated with this locus was not produced by a promoter located at the 3'- end of the gene (Matsui et al., 2008). These results support the speculation that certain NATs were generated by RNA-dependent RNA polymerases (RdRP) using mature mRNA as the templates. However, individually knocking out six RNA-dependent-RNA-Polymerases (RdRP) did not prevent the accumulation of the NAT associated with CYP707A1 (Matsui et al., 2008). It will be of substantial interest to determine the population of NATs that are produced by either DNA-dependent or RNA-dependent mechanisms.

In this section, we described our strand-specific RNA-seq experiment performed with SOLiD^{TM} sequencing platform. An informatics pipeline including a stringent and a relaxed definition was developed to identify NATs using RNA-seq data. Our efforts in determining the global pattern as well as the structure of individual NATs were also discussed.

5.2 Material and methods

5.2.1 RNA-seq experiment

Total RNA was isolated from leaves (7-12 mm) of 21-28 day old *Arabidopsis* plants with RNAqueous Midi kit (Ambion. Cat. # AM1911). Residual DNA in the RNA samples was removed by the treatment with amplification grade DNase (Invitrogen. Cat. # 18068-015). Two rounds of treatment with the Ribominus™ Plant Kit (Invitrogen cat. # A10838-08) were performed to deplete ribosomal RNA from the RNA preparation. The RNA-seq libraries were prepared with SOLiD™ Whole-Transcriptome Analysis Kit following the manual instruction. 50 bp of color-codes were sequenced for the RNA-seq library and the first 35 bp was used for mapping the tags to the reference genome. The mapping of RNA-seq tags was performed identically as ChIP-seq tags. Only tags mapped to a unique location in the genome were used for analysis to ensure the specificity of NAT identification. The raw RNA-seq data using PolyA RNA isolated from *Arabidopsis* seedlings was obtained from NCBI GEO accession GSE21323 and was processed similarly as our RNA-seq data.

5.2.2 Strand-specific RT-PCR assay for NAT detections

Total RNA isolated with Plant RNA Reagent (Invitrogen cat. # 12322-012) was subjected to two-round of RQ1 DNase digestions (Qiagen, cat. # 74904). RNA was cleaned up with RNeasy Plant Mini Kit (Qiagen, cat. # 74904) between the two rounds of DNase digestions. Roughly 2 μg of RNA was used for each reverse transcription (RT) reaction. RT reactions were primed with DNA oligos that can specifically anneal with the antisense transcripts. RT reactions were performed with Improm-II reverse transcriptase

(Promega, cat. # A3802) at 55°C. The quantity of NATs were determined with SYBR I qPCR assays using the cDNA as the templates.

5.2.3 Identification of selected NAT ends with 5'- and 3'- RACE

The RNA sample used for 5'- and 3'- RACE was identical as that used for strand-specific RT-PCR assays. The RACE was performed with 5'/3' RACE kit, 2nd generation (Roche, 03353621001) following the product manual. Briefly, for 5'-RACE, the synthesis of first-strand cDNA was primed with a NAT specific primer. The cDNA was then purified with High Pure PCR Product Purification Kit (Roche, 11732668001). The first-strand cDNA was then treated with terminal transferase and dATP for the polyA tailing. The RACE product was amplified with oligo(dT) and a NAT specific primer followed by another round of nested amplification. For 3'-RACE, oligo(dT) was used for priming the first-strand cDNA synthesis. The cDNA was subjected to the amplification with oligo(dT) primer and a NAT specific primer followed by another round of nested amplification. The RACE products were resolved on 1% agarose gel and the predominant bands were recovered from the gel for sequencing.

5.3 Results

5.3.1 RNA-seq performed with with SOLiD™ Whole Transcriptome Kit produced strand-specific results

The method used by SOLiD™ Whole Transcriptome Kit for generating RNA-seq library involved ligating distinct 5'- and 3'- adaptors with RNA molecules fragmented by RNase III (Figure 5.1A). The two adaptors were both RNA/DNA hybrids and had few bases of random protrusions on the DNA strand towards the insert. The usage of

RNA/DNA with protrusions on the DNA strand inhibited the generation of adaptor dimers and the ligation with blunt-ends double-strand DNA. However, since T4 RNA ligase is capable of catalyzing the ligation between RNA and single-strand DNA, RNA-seq library may be contaminated by the incomplete digested DNA containing sticky ends.

RNA-seq tags were processed identically as ChIP-seq tags as described in 4.2.3 except that only tags that mapped to a unique location in the genome were used for the analyses. The global antisense/sense ratio for exon regions was determined as 0.01, which was similar as the ratio of 0.008 reported by a RNA-seq experiment performed with Illumina RNA ligation procedure using *Arabidopsis* floral tissue (Lister et al., 2008). Through manual examinations, we could identify genes that associate with abundant antisense tags and genes that were actively expressed however associate with little antisense signals (Figure 5.1B). The differential abundances of antisense signals found with equally actively expressed genes suggested that the antisense tags were not non-specific artifacts but indeed reflects certain biological distinctions between loci. To evaluate the extent of potential DNA contamination, we examined numbers of genomic regions and found intergenic spaces were in general free of sequencing tags, suggesting the library contains little if any DNA contamination (data not shown). The results collectively supported that our RNA-seq was indeed strand-specific and contained no evidence of DNA contamination.

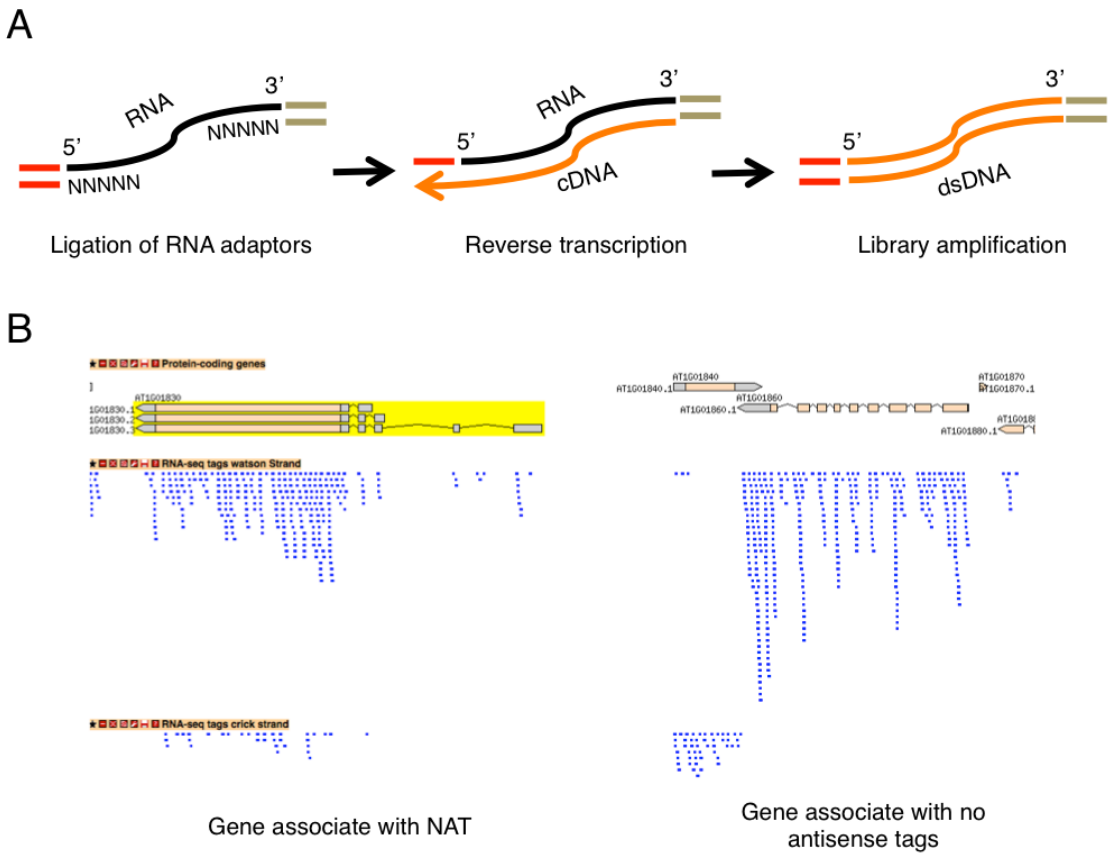

Figure 5.1. Strand-specific RNA-seq performed with SOLiD™ Whole Transcriptome Kit.

(A) Schematic description of procedure for generating RNA-seq libraries using SOLiD™ Whole Transcriptome Kit. (B) Examples of the RNA-seq data for genes either or not associated with antisense transcripts.

5.3.2 Determination of genomic regions associated with NATs

In order to determine genomic regions that associate with significant antisense signals, we first identified 'transcript-domains' by joining consecutive regions that associated with RNA-seq signals allowing a maximum signal gap of 20 bps. The 'transcript-domains' were then aligned with annotated *Arabidopsis* genes to identify NATs domains that were antisense to one or more annotated transcription units. Among the 4,939 NATs domains identified, 1,896 were caused by tail-to-tail overlapping 3'-UTR of protein-coding genes and the rest 3,043 domains likely corresponded to true NATs. The 3,043 NATs domains were mapped to 1,302 annotated genes. The amount of identified NATs was much less than that reported by other works using genome-tiling arrays, which commonly determine more than 5,000 NAT producing loci (Yamada et al., 2003; Matsui et al., 2008). The discrepancy is most likely caused by the insufficient sequencing depth of RNA-seq, which would lead to signal discontinuity for the scarce antisense transcripts. In order to accommodate for the discontinuity of NATs signals, we developed a relaxed definition of NATs - any locus containing more than two effective antisense tags within the region was considered as a NAT producing locus. The definition of effective tags was developed to avoid the identification of tail-to-tail protein coding genes as NATs. Tags that located at least 100 bps away from any annotated transcription unit in the opposite direction were considered 'effective'. 2,667 annotated loci were determined to associate with relaxed defined NATs.

Since our RNA-seq dataset containing 5.6 million tags cannot be considered 'ultra-deep' with current standards, we aimed to address if the sequencing depth would allow reliable discoveries of NATs. To this end, we analyzed an independent strand-specific

RNA-seq dataset produced by SOLiD^{TM} Whole Transcriptome Kit using PolyA RNA isolated from *Arabidopsis* seedlings (Deng et al., 2010). The dataset contained 3.3 millions of uniquely mapped and non-redundant tags. The sequencing of rRNA-depleted RNA and PolyA RNA yielded strikingly similar sets of NATs identified by either stringent or relaxed definition (Figure 5.2). The result also provided compelling evidence that many NATs may be polyadenylated. However, a trivial explanation of the result need to be tested – stretches of nucleotide A may be enriched in NAT transcripts, which can lead to the capture of NATs by oligo(dT) selection without actual 3'-end PolyA tails.

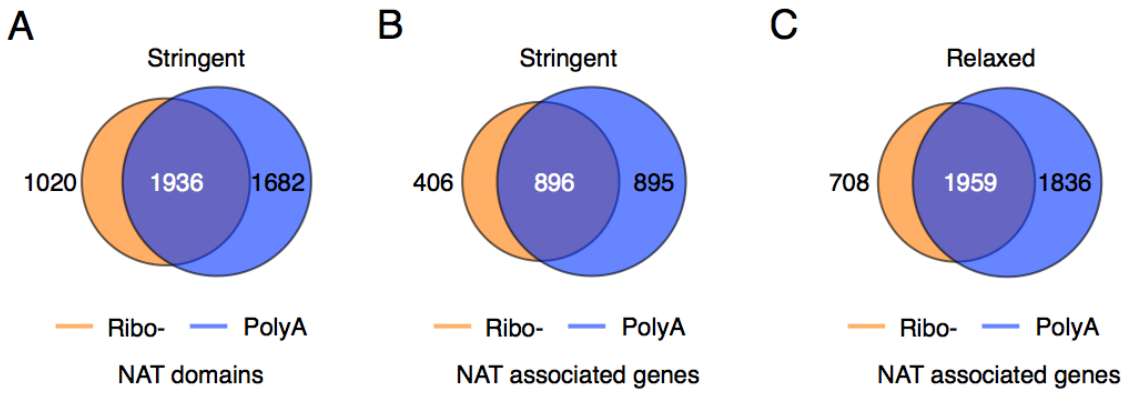

Figure 5.2. Comparison of NATs identified by the sequencing of rRNA-depleted and PolyA RNAs.

5.3.3 Global patterns of NATs

Previous analysis of *Arabidopsis* antisense transcripts found that certain NATs preserve the exon/intron junctions of the corresponding sense transcripts (Matsui et al., 2008). The result implicated that some NATs may be produced by RNA-dependent RNA polymerases using mature mRNA as the templates (Matsui et al., 2008). To test this speculation with our data, we determined the abundance of sense and antisense tags per length of transcription units, exons or intron (Figure 5.3). As expected, the frequency of sense tags associated with exons was ~6 time greater than that associated with introns (Figure 5.3A). To our surprise, exons also associated with much greater frequency of antisense tags compared to introns (Figure 5.3B), suggesting a large fraction of antisense tags may be produced using mRNA as the template. However, the interpretation of the result was complicated by two facts – 1) Due to the low sequence complexity of plant introns, the mapping of sequencing tags originated from intronic regions may be less efficient. 2) As we will discuss in section 8, antisense tags were enriched in genes that were relatively short and frequently contains no intron.

NATs overlapping with intronic regions were likely to be generated through DNA-dependent transcriptions. 1,815 NATs identified by the relaxed definition were found to contain at least one RNA-seq tag mapped within introns, which account for 36.5% (1,815/4,967) of all relaxed defined NATs. An example of NATs containing intronic tags was shown in Figure 5.4.

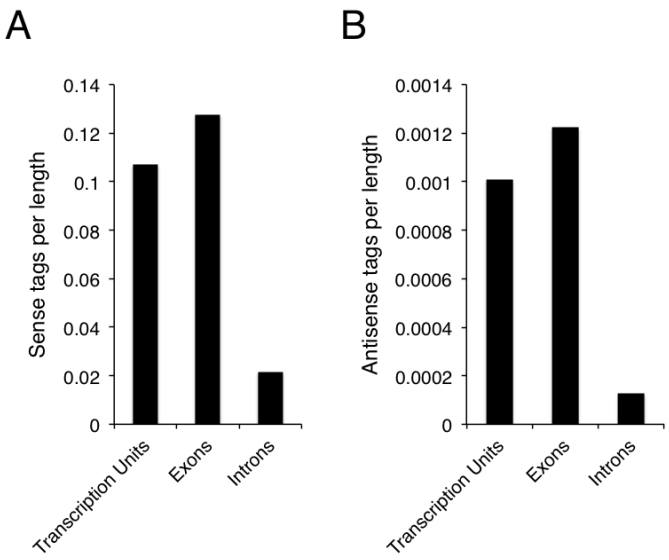

Figure 5.3. The abundance of sense and antisense tags per length of transcription units, exons or intron.

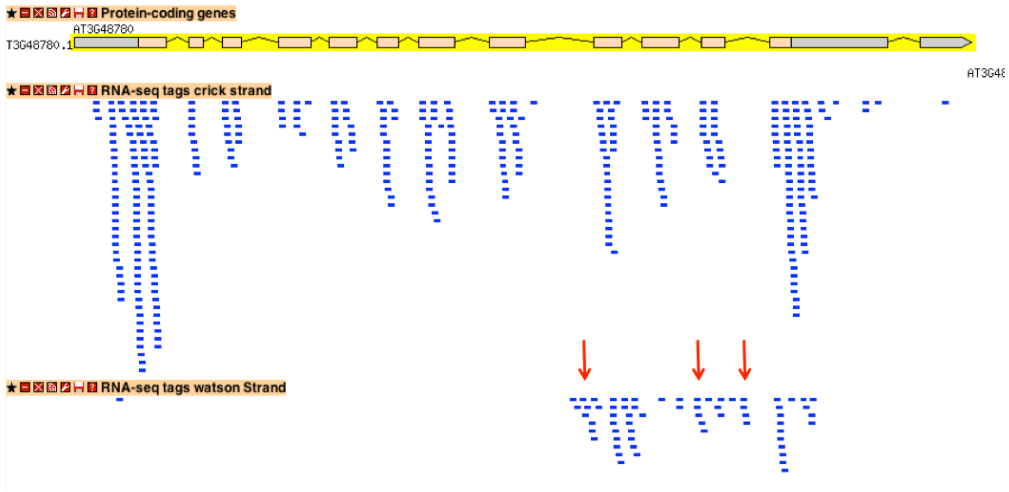

Figure 5.4. Examples of NATs showing no apparent similarity with the sense transcript with regard to exon/intron sturctures.

Clusters of antisense tags that mapped to the intron regions of the cognate loci were indicated by red arrows.

The global pattern of NATs within the region of transcription units were visualized by plotting the frequency of antisense tags aross the scaled transcription unit regions for loci associated with NATs identified with stringent or relaxed definition or no NATs (Figure 5.5A). Loci associated with NATs defined with the strignent definitition had the greatest frequency of antisense signals across the transcription units, which is consistent with that NATs defined with the stringent definition were generally supported by more antisense tags (Figure 5.5A). A steep decline of the antisense signal was observed around TSS region, which suggest the region is not commonly included in the NATs (Figure 5.5A). For NATs that were produced using the sense transcript as the template, it would be expected that regions upstream of TSS were depleted of antisense signals (Figure 5.5A). However for NATs that were generated by independent transcriptions, the results may implicate the antisense transcriptions rarely elongate beyond the TSS.

We next sought to define whether NATs associated with actively expressed loci or repressed loci. The percentages of genes associated with NATs were compared between annotated genes that were categorized into 5 bins according to their sense expressions. We found bins associated with more active expressions consistently contained higher fraction of NATs identified with either stringent or relaxed definition. Therefore, the results suggest that NATs is a feature for active chromatin and are found predominantly with active expressed genes.

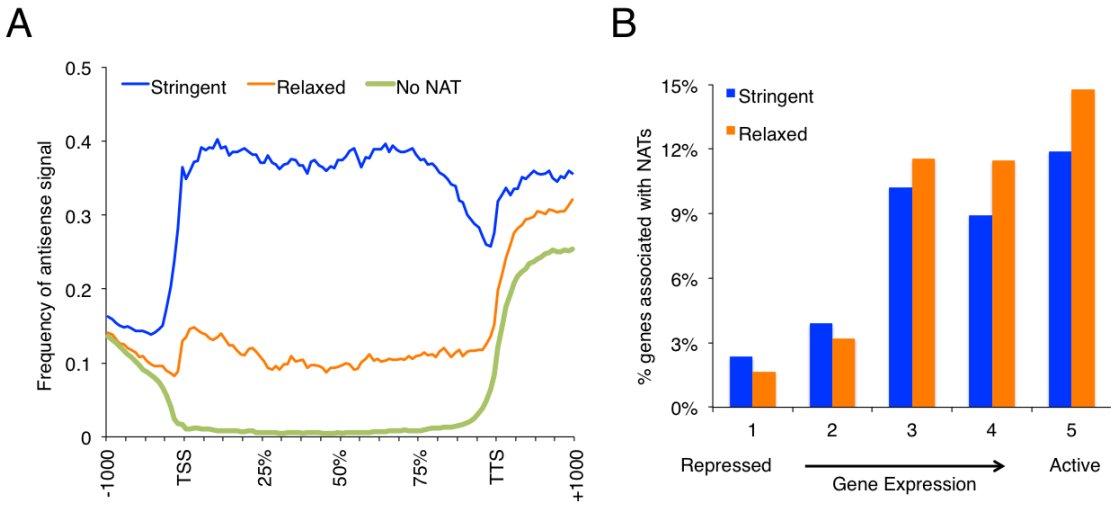

Figure 5.5. Global pattern of NATs.

(A) Average frequencies of antisense tags were plotted across scaled transcription units between 1 kb upstream of TSS and 1 kb downstream of TTS. Blue, brown and green lines each indicate the loci associated with NATs identified by stringent or relaxed definition or no NATs respectively. (B) The fraction of genes associate with NATs determined by the stringent or relaxed (stringent NATs excluded) definition for gene groups binned according to their expression levels.

5.3.4 Detection of NATs with the strand-specific RT-PCR assay

In order to validate the presence of NATs and quantify NATs in different genetic backgrounds, we aimed to develop a method to specifically detect the antisense transcripts. Due to the scarcity of antisense transcripts (~1% of total RNA-seq tags), northern blot is unlikely to provide the sufficient sensitivity for the detection. Therefore we chose to use strand-specific qRT-PCR assays for the quantification of NATs. During the set up of the assay, we found that cDNA corresponding to the exonic regions of actively expressed genes can be synthesized even without adding any priming oligo into the RT reactions. Presumably short double-stranded DNA fragments generated by DNase I treatments could melt during reverse transcriptions and serve as random primers for non-specific cDNA syntheses. Therefore we have chosen to quantify NATs that were mapped to intronic regions or intron/exon junctions, because the abundance of steady-state pre-RNA should be much less than mature mRNA and were thus unlikely to cause spurious reverse transcriptions.

The design of strand-specific qRT-PCR assays were shown in Figure 5.6A. Reverse transcriptions were primed with DNA oligos (arrows in Figure 5.6A) that specifically anneal with the antisense transcript. The NAT was then quantified with the primer pair A (Figure 5.6A). For each qRT-PCR assay, a mock RT reaction containing all reagents and RNA templates but the priming oligo (NP-RT) was performed to control for potential DNA contaminations or non-specific cDNA syntheses. To further rule out the possibility that the priming oligo may non-specifically bind to the sense pre-RNA, qPCR with the U primer pair was performed (Figure 5.6A). Since the two binding sites of U oligos either locate at upstream or across where the priming oligo binds, no products

should be generated if the RT reactions were performed specifically (Figure 5.6A). Testing the design with two loci showed that NATs could be specifically detected by the strand-specific RT-PCR assay. No PCR products were generated with primer set U with cDNA as the template (Figure 5.6B). Also no signal was detected with primer set A when NP-RT sample was used as the template (Figure 5.6B). To finally rule our the trivial possibility that the amplification efficiency of primer set U was drastically lower than primer pair A, the two pairs were used for amplifying different amount of gDNA and were shown to have similar efficiencies (Figure 5.6C).

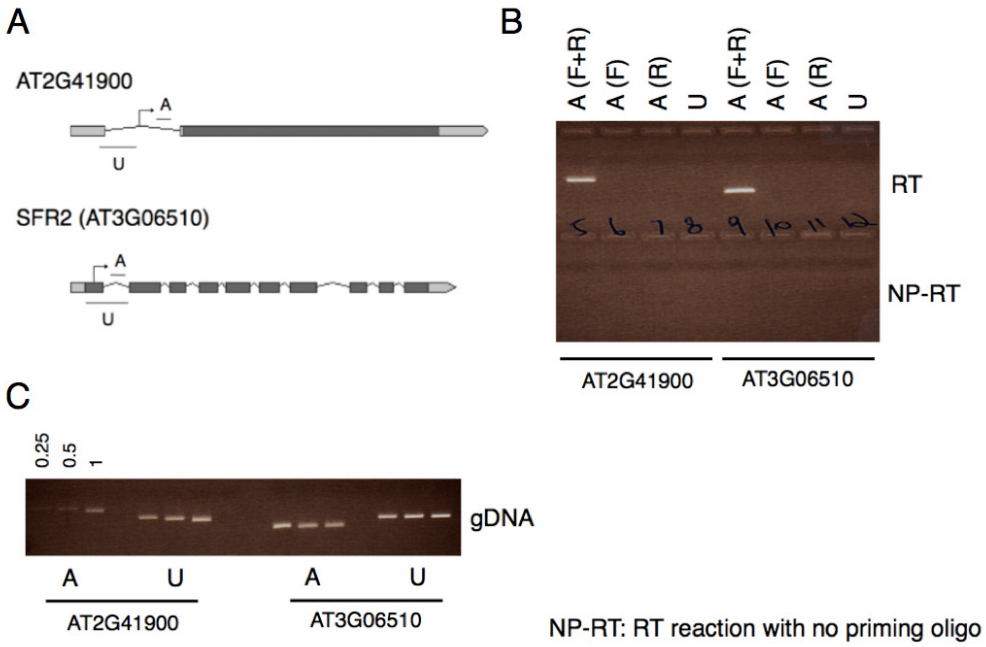

Figure 5.6. Strand-specific RT-PCR assay for the quantification of NATs.

(A) Design of strand-specific RT-PCR for the detection of NATs. Arrows, A and U each indicates the position of oligo used for priming the RT reaction, qPCR amplicon for NAT detection and the control qPCR amplicon 'upstream' for testing the RT specificity, respectively. (B) Specific detection of the two NAT species with the strand-specific RT-PCR assays. RT: cDNA produced by the reverse transcription was used for template. NP-RT: reverse transcription was assembled without the priming oligo. (C) Comparison of PCR efficient between A and U pairs of primers.

5.3.5 Efforts of cloning NATs with 5'- and 3'- Rapid Amplfication of cDNA Ends (5'- and 3'- RACE).

Using a polyA-tailing based RACE system, we performed 5'- and 3'- RACE for seven putative NATs. Since strand-specific RT-PCR were established for each NAT, we sequenced the products of RT-PCR to confirm the specificity of the assays. The sequence information was further used for designing NAT specific primers for the RACE experiments.

Among the 5'-RACE products for the seven NATs, only the product corresponding to the NAT associated with AT3G06510 was specifically amplified as shown by the sequencing of PCR products (Figure 5.7A). The result suggested that the NAT associated with AT3G06510 was initiated in the second exon of the gene (Figure 5.8A). 3'-RACE of the same NAT did not yield any specific product. The 3'-end of the NAT was approached by scanning the genomic regions with serial strand-specific RT-PCR assays. The RT-PCR scanning experiment placed the 3'-end of the NAT at approximately ~300 bps or further upstream of the AT3G06510 TSS (Figure 5.8A).

Specific 3'-RACE products were obtained for NATs associated with AT1G04430, AT4G01250 and AT3G48360 (Figure 5.7B). The deduced 3'-end of NAT associated with AT1G04430 did not fall in a stretch of nucleotide A or T and was therefore likely corresponded to the true 3'-end of the RNA (Figure 5.8B). The putative 3'-end for NATs associated with AT4G01250 or AT3G48360 however, mapped to stretches of A. In addition, the two putative ends did not extend beyond the location of the oligo used for priming the strand-specific RT-PCR assays. Therefore these identified 3'-end were likely to be generated by the unspecific amplifications mediated by the binding of oligo(dT) to

A stretches. Interestingly, no splicing event was observed from all the partial sequences of NATs that we obtained.

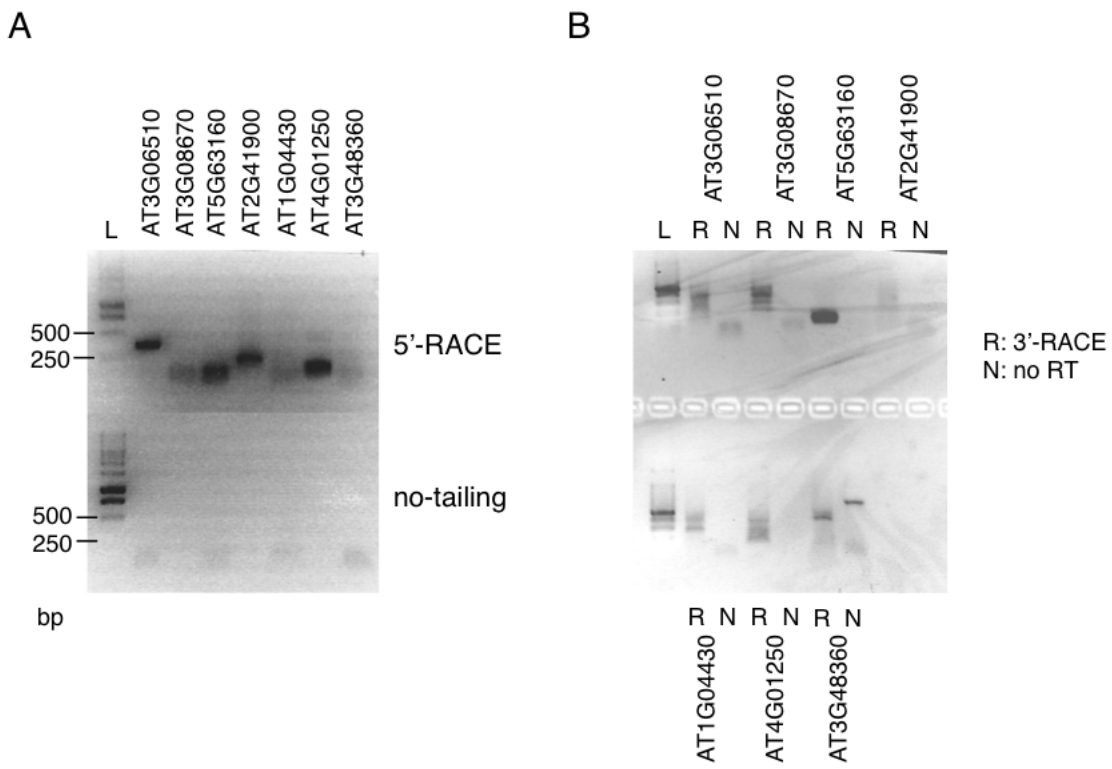

Figure 5.7. 5'- and 3'-RACE experiments for the identification of NAT ends.

(A) Products of 5'-RACE were resolved on a agarose gel. The specificity of the RACE was controlled by using cDNA that was not tailed with nucleotide adenine for the template of the nested PCR. (B) Products of 3'-RACE were resolved on a agarose gel. The specificity of the RACE was controlled by using the RNA preparation instead of cDNA as the template for amplification (no RT). L: marker for the size of DNA.

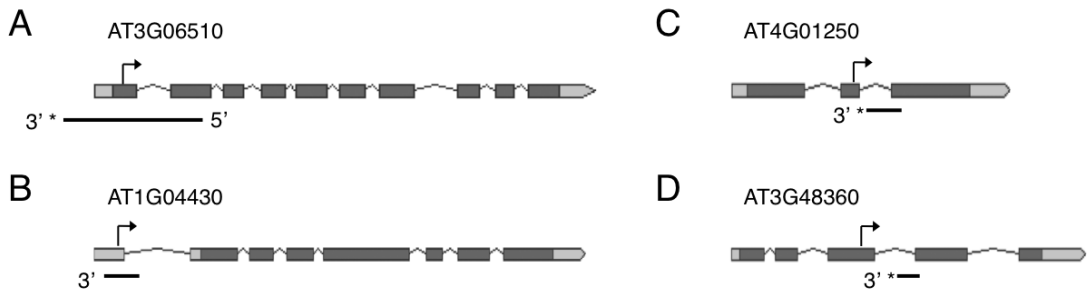

Figure 5.8. Deduced 5'- and 3'- ends of NATs identified by the RACE experiments.

Asterisks suggest the identified end was not likely to be the true ends of the RNA molecule. The 3'-end of AT3G06510 associated NAT was approximated by series of strand-specific RT-PCR. The identified 3'-ends of NATs associated with AT4G01250 and AT3G48360 were likely caused by the non-specific binding of oligo(dT) primer with stretches of A.

5.4 Discussions

The current work together with previous reports together suggest that plant cells produce multiple classes of NATs that may be different in their biogenesis, structures, subcellular localizations and biological functions (Matsui et al., 2008). The identification of NATs preserving the exon/intron junctions of the cognate sense transcripts suggested certain NATs are generated with mature mRNAs as the templates. Nevertheless, our successful detections of NATs produced from intronic regions supported that the NATs we identified were transcribed from DNA templates. The sequencing of more full-length NATs including any splicing junctions would be critical for further studying the biogenesis of these enigmatic molecules. This goal can be reached by either classical cloning and sanger sequencing or high-throughput-sequencing with extensive tag length and sequencing depth. In addition, the determination of phosphorylation states and capping characteristics of 5'- end as well as the nature of the 3'- tail of the RNA molecule would shed light on the biogenesis NATs. However classical RACE system can be easily interfered by low complexity sequences as we have shown. The usage of RACE system based on RNA adaptor ligations may be helpful to overcome this difficulty. Lastly, the abundance of NATs can be tested in mutants of DNA-dependent- and RNA-dependent- RNA polymerases or under the treatment of specific inhibitors for RNA polymerases to derive the polymerase that is responsible for the NATs productions.

6 Functional analysis of chromatin states defined with 10 chromatin modifications

6.1 Introductions

In section 3.3.3, we have described the *Arabidopsis* epigenome with chromatin states defined by four modifications – H3K4me3, H3K36me2, H3K27me3 and 5mC. However, histone proteins are known to carry numerous covalent modifications at different amino acid residues, although the functions for many of them remain elusive. Therefore it is not likely that the whole spectrum of epigenome information can be effectively described by merely four modifications. With the newly produced genome-wide maps of 9 histone modifications, we were able to refine the definition of chromatin states and pursue a more comprehensive description of *Arabidopsis* epigenome.

As we have discussed in section 3.4, the complexity of chromatin states defined by a large number of chromatin modifications can be enormous. If the state of a chromatin modification at a particular locus is described by a binary code (0 or 1 represent negative or positive enrichment respectively), theoretically up to 1,024 (2^{10}) states can be defined with 10 modifications. Since the determination of binary chromatin state codes involves arbitrary statistical cutoffs, noises may accumulate along with the fine division of chromatin states, which can lead to the identification of artificial states. Some previous works addressed this challenge by describing the whole epigenome by few states using fuzzy statistical approaches (Filion et al., 2010; Roudier et al., 2011). These highly abstract methods were effective in describing the predominant structural variations in the epigenome. For example, the four major chromatin states defined in Roudier et al., 2011 each correspond to actively expressed genes, polycomb targets, repetitive sequences and

regions associated with no chromatin modification. However because much quantitative information was disregarded after the statistical abstraction, this type of analysis may not be suitable for deriving functional hypotheses regarding chromatin modifications. Since the primary goal for our analysis is to construct testable hypotheses through correlative analyses, we chose to finely define the chromatin states utilizing the quantitative measurement of each modifications.

6.2 Material and methods

6.2.1 Determination of the chromatin state for each annotated transcription unit

Peaks for histone modifications were identified by MACS (Model-based Analysis of ChIP-seq) with p value equaled to 10^{-5} (Zhang et al., 2008). A gene was considered enriched of a particular histone modification if the genic region overlapped with a peak of the modification being interrogated. The chromatin state of a transcription unit was defined by the ensemble of the enrichment states for 10 chromatin modifications being analyzed in the work. The MeDIP-chip data of 5mC was obtained from NCBI GEO accession GSE5974. The bioinformatics processing and the identification of significantly methylated regions were performed as described (Zilberman et al., 2007; Luo and Lam, 2010). Briefly, values of MeDIP/Input were normalized by subtracting the bi-weight mean. Contiguous probes with $\log_2$(MeDIP/Input)>1.28 were defined as methylated regions. Hierarchical clusterings of the chromatin states was performed with Cluster 3.0 (De Hoon et al., 2004). The heatmap visualizations of clustering results and other data matrix were performed with Java TreeView 1.1.4r3 (Saldanha, 2004).

6.2.2 re-ChIP assay

For re-ChIP assay, 20 μg of the antibody used for the first round of ChIP was crosslinked with 40 μl protein-A agarose resin using Pierce crosslink immunoprecipitation kit (Pierce cat. # 26147). The protein-A agarose resin was then incubated with 3 ml of chromatin sample (in chromatin dilution buffer) at 4°C with rotation for 1 hour. The washing of protein-A resin was performed as the standard ChIP procedure described in section 4.2.2. Chromatin that bound by the antibody was eluted by 200 μl of 50mM Tris-Cl pH=7.5, 1% SDS at room temperature for 20 min. The eluted chromatin sample was diluted to 2 ml with ChIP dilution buffer and dispensed to 4 micro-centrifuge tubes. Each second round ChIP was performed with 2 μg of primary antibody and 30 μl protein-A agarose beads following the standard ChIP procedure.

6.3 Results

6.3.1 The determination of chromatin states at the transcription unit level with 10 chromatin modifications.

Using a simple integration of binary enrichment codes for 10 chromatin modifications, we determined 295 types of distinct chromatin states that associated with the 33,003 annotated genes in the *Arabidopsis* genome (Figure 6.1A). The number of distinct chromatin states was relatively small compared to the theoretical 1,024 (2^{10}) states that can be assembled from 10 chromatin modifications. As expected, the size of chromatin states (number of genes included in a particular state) clearly deviated from the random coincidence of 10 modifications (Table 6.1). Close to 2/3 of the identified states

(186 states) contained less than 10 genes and only 42 states were found to include more than 100 genes (Table 6.1). Due to the application of arbitrary statistical cutoffs, the identification of chromatin states contained substantial intrinsic noises. Therefore the significance of chromatin states associated with few genes was unapparent and we chose to focus on states containing more than 100 genes. The global correlations between chromatin modifications at the resolution of 300 bps were shown by the clustering of correlation coefficient matrix (Pearson, Figure 6.1B). Two tight clusters were found each corresponding with chromatin modifications that were specific for actively expressed or silenced loci, respectively. The former cluster included H3K4me3, H3K9Ac, H3K4me2 and H3K36me3 while the later cluster included H3K9me2, H3K27me1 and 5mC (Figure 6.1B). As shown earlier (Figure 3.1A and B, Figure 3.3A), H3K36me2 and H3K4me3 co-localized at a large number of targets. However the two histone marks occupied distinct regions within the genes bodies (Figure 3.1A and B) and were thus not identified as globally correlated in Figure 6.1B. This discrepancy highlighted the importance of performing correlative analysis in the level of transcription units. Transcription units are the natural context that chromatin modifications function to modulate the transcription activities. The identification of chromatin states at the level of single nucleosome, instead, fell short in the integration of chromatin structures with functional outputs.

Through analyzing the pair-wise relations between chromatin modifications, we identified an intriguing correlation between H3K18Ac and H3K27me3 (pearson r=0.44), although histone acetylations were not known to function in polycomb regulations.

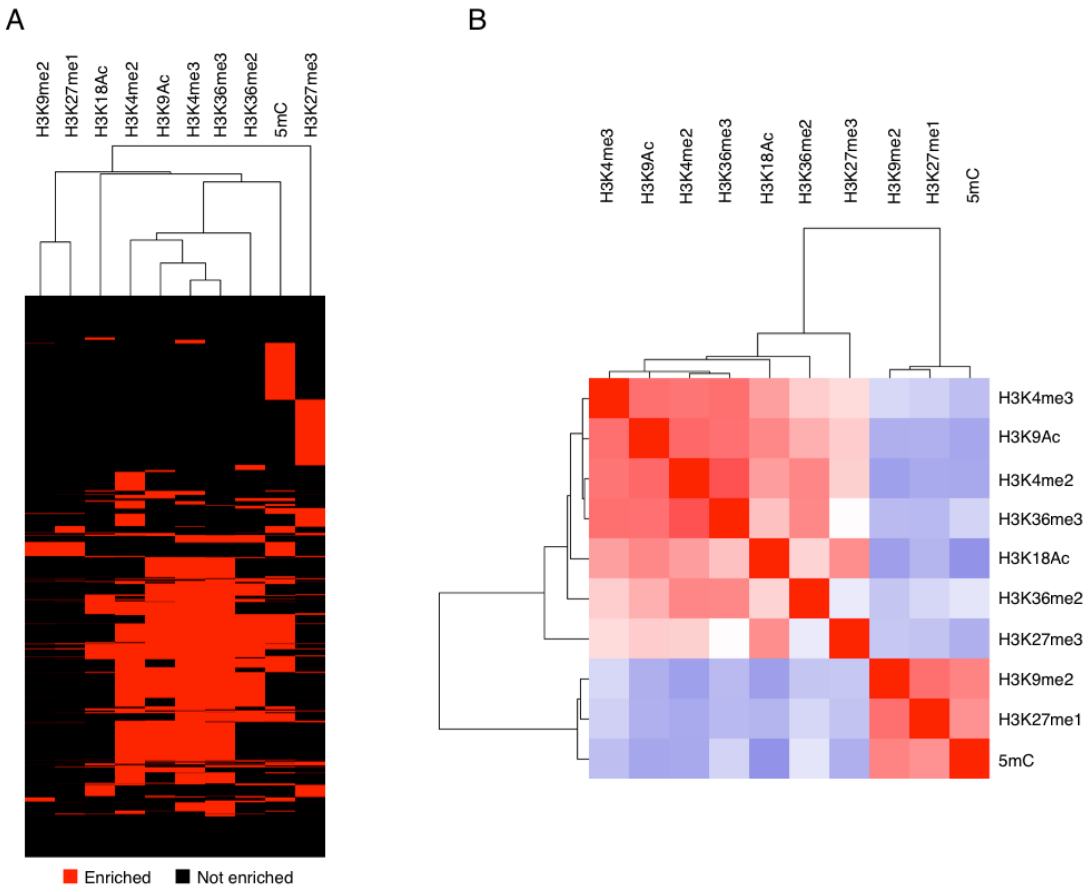

Figure 6.1. Hierarchical clustering of chromatin states and pairwise correlations between chromatin modifications.

(A) Bidirectional clustering of the chromatin states defined by 10 chromatin modifications. The extent of the correlations between the modifications was represented by the dendrogram above the clustering image. (B) Clustering of the pairwise correlation coefficient matrix (Pearson). The correlation coefficients were computed for the coverage pattern of chromatin modifications across the genome at a 300 bp resolution.

Table 6.1. List of the chromatin states associated with more than 100 annotated genes ordered by the amount of genes contained in the states.

H3K4me2	H3K4me3	H3K9Ac	H3K9me2	H3K18Ac	H3K27me1	H3K27me3	H3K36me2	H3K36me3	5mC	Number of Genes	Percentage of total genes
0	0	0	0	0	0	0	0	0	0	4882	14.79%
0	0	0	0	0	0	1	0	0	0	3851	11.67%
0	0	0	0	0	0	0	0	0	1	3307	10.02%
1	1	1	0	0	0	0	0	1	0	2208	6.69%
1	1	1	0	0	0	0	1	1	0	1467	4.44%
0	1	1	0	0	0	0	0	1	0	1103	3.34%
1	0	0	0	0	0	0	0	0	0	1092	3.31%
1	1	1	0	0	0	0	1	1	1	1064	3.22%
0	0	0	1	0	1	0	0	0	1	799	2.42%
1	0	0	0	0	0	1	0	0	0	717	2.17%
1	1	1	0	1	0	0	0	1	0	686	2.08%
0	1	1	0	0	0	0	1	1	0	642	1.95%
0	0	0	0	1	0	1	0	0	0	628	1.90%
1	1	0	0	0	0	0	0	1	0	595	1.80%
0	1	1	0	0	0	0	1	1	1	539	1.63%
1	1	0	0	0	0	0	1	1	0	519	1.57%
1	1	0	0	0	0	0	1	1	1	493	1.49%
0	1	0	0	0	0	0	0	1	0	477	1.45%
0	1	0	0	0	0	0	1	1	1	447	1.35%
0	1	0	0	0	0	0	1	1	0	394	1.19%
1	1	1	0	1	0	0	1	1	0	388	1.18%
0	0	0	0	0	1	0	0	0	1	377	1.14%
0	0	0	0	0	0	1	0	0	1	306	0.93%
1	1	1	0	0	0	0	0	0	0	298	0.90%
0	0	0	0	0	0	0	1	0	0	279	0.85%
1	1	1	0	0	0	0	0	1	1	261	0.79%
1	1	1	0	1	0	0	1	1	1	226	0.68%
0	1	0	0	0	0	0	0	0	0	203	0.62%
0	1	1	0	0	0	0	0	0	0	196	0.59%
0	0	0	1	0	0	0	0	0	1	196	0.59%

1	0	1	0	0	0	0	0	0	0	193	0.58%
0	1	1	0	1	0	0	0	1	0	191	0.58%
1	1	0	0	0	0	0	0	0	0	170	0.52%
1	0	0	0	0	0	0	0	1	0	165	0.50%
0	0	0	0	0	0	0	1	0	1	145	0.44%
1	0	0	0	0	0	0	1	0	0	142	0.43%
0	1	1	0	0	0	0	0	1	1	135	0.41%
0	0	1	0	0	0	0	0	0	0	125	0.38%
0	0	0	0	1	0	0	0	0	0	125	0.38%
1	1	0	0	0	0	0	0	1	1	122	0.37%

6.3.2 Verification of the chromatin states with re-ChIP assay

The model of chromatin states can only be valid if the co-markings of modifications indicated by the model indeed corresponded to physical co-localizations of the chromatin modifications. The co-localization of chromatin modifications was addressed by the development of a highly specific re-ChIP assay. Briefly, two rounds of ChIP were performed sequentially with different antibodies. The ChIPed chromatin from the first round of ChIP was eluted from the beads and served as the input for the second round of ChIP. The antibody for the first round of ChIP was crosslinked with the beads to prevent the co-elution of the antibody together with the chromatin. The specificity of the re-ChIP assay was proved by two evidences – 1) The second-round of ChIP yielded no detectable DNA if no antibody was added, which suggests little if any antibody used for the first-round ChIP was co-eluted with the chromatin (Figure 6.2). 2) re-ChIP using anti-H3K4me3 followed by anti-H3K27me3 antibodies yielded no significant signal for known H3K4me3 or H3K27me3 monovalent loci (Figure 6.2H and I). Therefore no cross-reactivity between the anti-H3K4me3 antibody and the H3K27me3 epitope or between the anti-H3K27me3 antibody and the H3K4me3 epitope were detected. We tested the co-localization of three 'active' histone marks – H3K4me3, H3K9Ac and H3K36me3 at the TSS region of three actively expressed genes (Figure 6.2A-C). H3K4me3+H3K9Ac or H3K4me3+H3K36me3 re-ChIP yielded comparable signals as the H3K4me3+H3K4me3 control for these regions, suggesting H3K9Ac or H3K36me3 were highly co-localized with H3K4me3 at all three tested loci. The results thus confirmed that the co-marking of multiple active marks is indeed a characteristic of actively expressed genes in *Arabidopsis*. As aforementioned, we observed an intriguing

correlation between a histone acetylation H3K18Ac and the repressive mark H3K27me3. With re-ChIP assay, we showed that the two marks indeed co-localized physically (Figure 6.2D-G).

The bivalent state of H3K4me3 and H3K27me3 (hereafter referred to K4/K27me3 bivalent state) was a particular interesting state to be tested by re-ChIP assay. The K4/K27me3 bivalent was known to associate with many development regulator genes in mammalian ES cell and resolve into either H3K4me3 or H3K27me3 monovalent states during differentiations (Bernstein et al., 2006; Mikkelsen et al., 2007). We identified 377 genes associated with the putative K4/K27me3 bivalent state with our dataset. The bivalency of 10 putative K4/K27me3 bivalent loci were tested by bi-directional re-ChIP assays, which inverted the order of antibodies used for the first and second round of ChIP (Figure 6.2H and I). The significance of the bivalency was determined by modeling the results obtained with control monovalent loci using a normal distribution ($p<10^{-4}$). Among the 10 tested loci, the bivalency of 3 loci was supported by the re-ChIP performed in both directions whereas the bivalency of the two other loci were supported by H3K4me3+H3K27me3 re-ChIP alone (Figure 6.2H and I). Our results therefore suggest that the putative K4/K27me3 group identified by our ChIP-seq experiment were in fact highly heterogeneous. It will be of great interests to combine the re-ChIP assay with high-throughput-sequencing to determine K4/K27me3 bivalent states at a global scale.

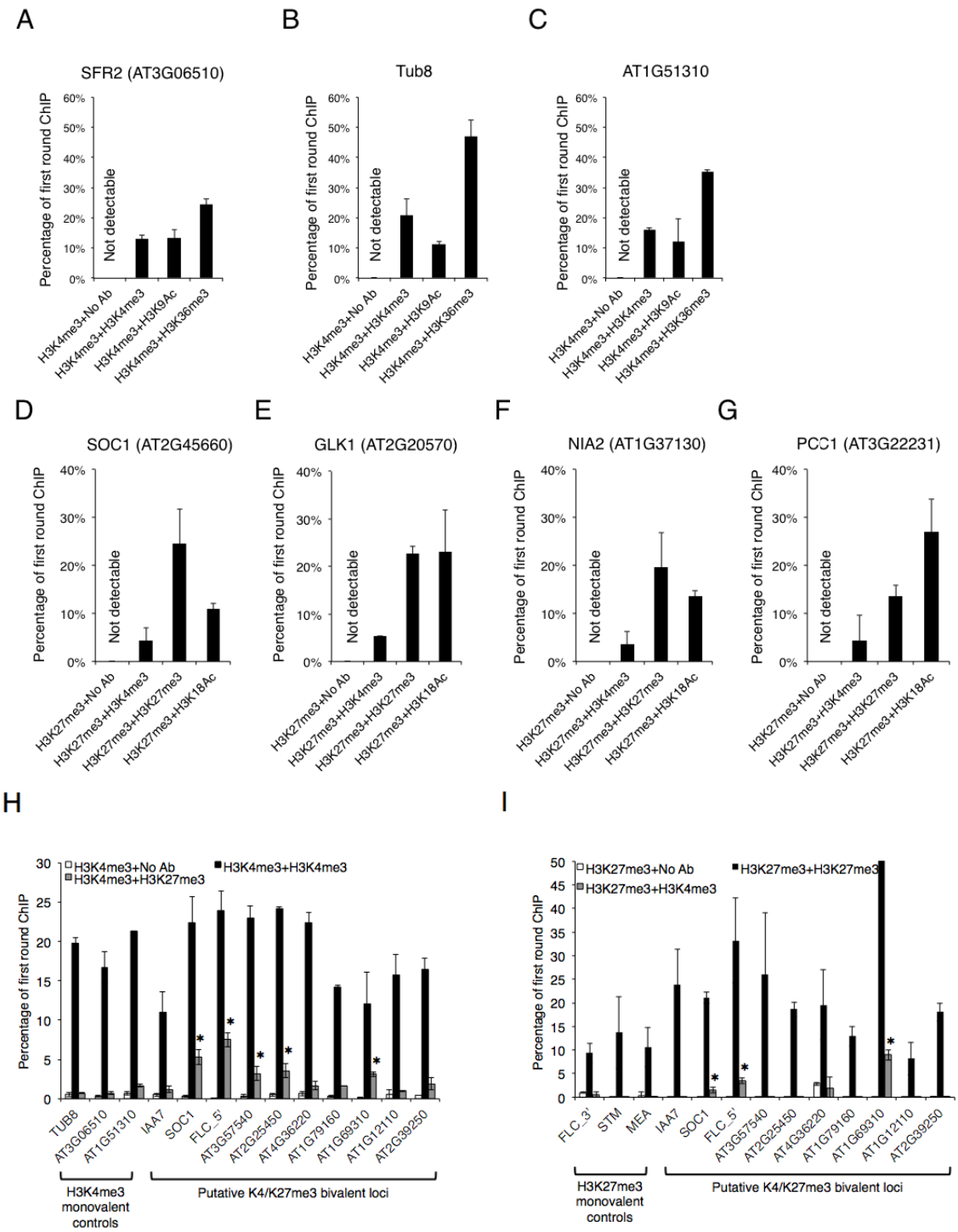

Figure 6.2. Validation of chromatin state models with re-ChIP assay.

(A-C) re-ChIP analysis of the TSS region for actively expressed genes. The first round of ChIP was performed with anti-H3K4me3 antibody followed with the second round of

ChIP using no antibody or anti-H3K9Ac, -H3K36me3 antibodies. (D-G) re-ChIP analysis of putative K4/K27me3 bivalent loci with anti-H3K4me3 antibody used in the first-round of ChIP. Anti-H3K4me3, H3K27me3 and H3K18Ac antibodies were used for the second round of ChIP. (H-I) Bidirectional re-ChIP analysis of putative K4/K27me3 bivalent loci.

6.3.3 Gene Ontology (GO) term enrichment analysis of chromatin states

Our current chromatin state model defined by the ensemble of 10 chromatin modifications enabled the functional analyses of chromatin states at an unprecedented resolution. The analysis was focused on the GO enrichments in related chromatin states that associated with partially overlapping set of modifications. We have identified examples that related chromatin states were enriched of a common GO term (Figure 6.3A). GO term 'structural constituent of ribosome' was enriched in five states modified by H3K4me3 and H3K36me3 but with variable association of several other chromatin modifications (Figure 6.3A). The comparison of gene expressions for the GO term 'structural constituent of ribosome' in the five states has shown that genes in the state 3 with this GO term expressed significantly lowly compared to those in state 1 or 2 (Figure 6.3B). A particular interesting finding for the GO term analyses was the differential enrichments of certain GO terms in related chromatin states. For example, states 1, 3, 4 and 5 in Figure 6.3C were all marked by H3K27me3 but had variable associations with H3K4me2, H3K18Ac or 5mC. Although commonly modified with the characteristic mark for polycomb regulation – H3K27me3, the four states showed rather distinct GO enrichment patterns (Figure 6.3C). Transcription factors were enriched in states 1 and 3 but not in 4 or 5; GO term 'zinc ion binding' was only enriched in state 1; 'hydrolase activity' was enriched in states 1 and 4 but not in state 3 (Figure 6.3C). The biological implication for the differential enrichment of GO terms in related states was elusive. However, certain combinations of chromatin modifications, such as H3K4me2+H3K27me3 or H3K18Ac+H3K27me3 may represent non-classical type of

bivalent chromatin states in addition to the well-described H3K4me3+H3K27me3 bivalent state.

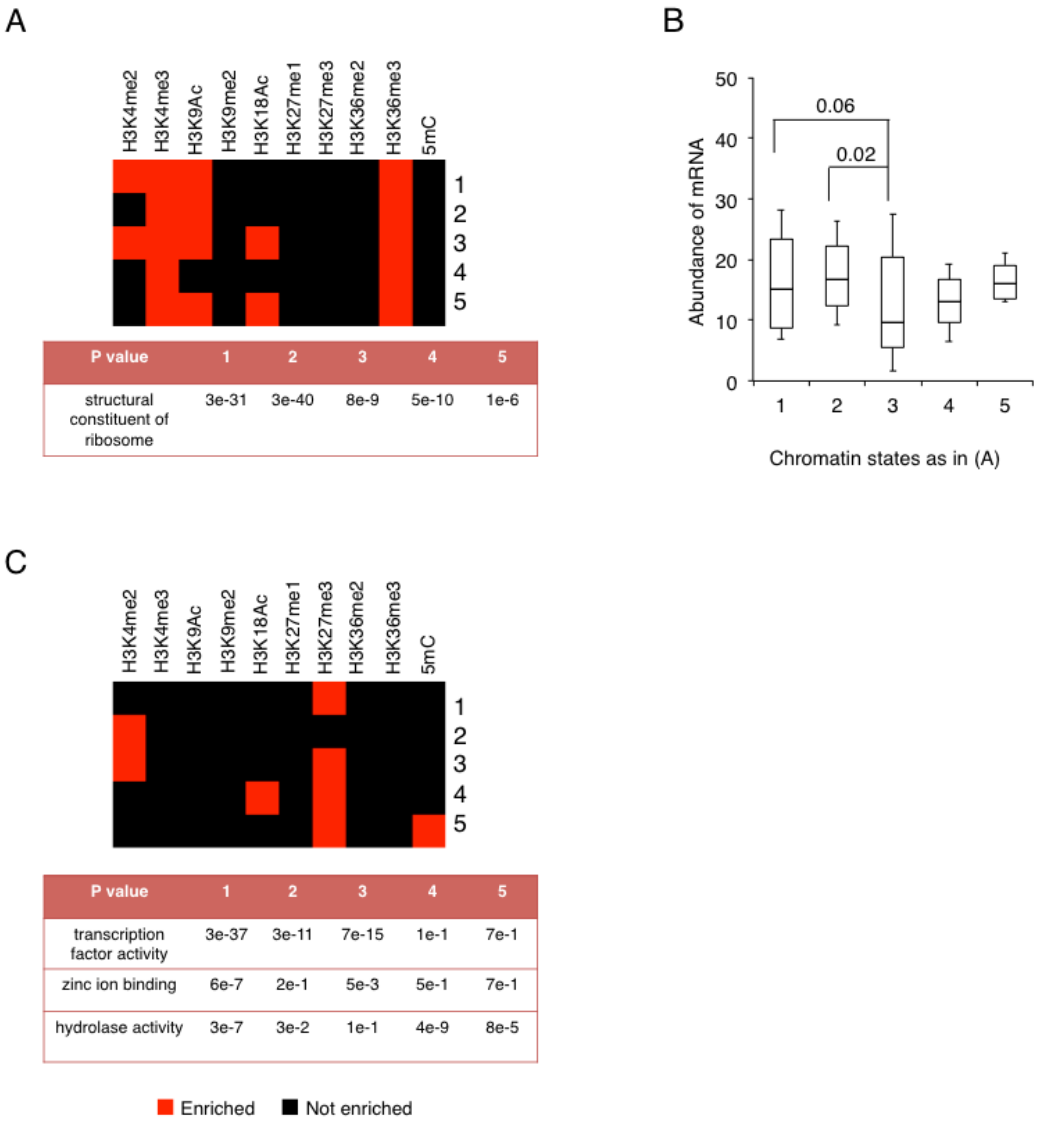

Figure 6.3. GO (gene ontology) enrichment analysis of the chromatin states.

(A) GO enrichment analysis of chromatin states enriched with H3K4me3 and H3K36me3. (B) Expression level of genes with GO term 'structural constituent of ribosome' in the five chromatin states shown in (A). (C) GO enrichment analysis of chromatin states associated with H3K27me3 but have variable enrichments of H3K4me2, H3K18Ac or 5mC.

6.3.4 Locus type enrichment analysis of chromatin states

Enrichment analysis for locus types, including various non-coding RNA species, transposable elements and pseudogenes were performed for the 42 major chromatin states. A conspicuous pattern observed from the analysis was that the vast majority of pre-tRNA (93%), miRNA (85%) and snoRNA (86%) associated with merely 3 states (Figure 6.4A). Interestingly, 53%, 26% and 6% of miRNA loci were modified by no modification, H3K27me3 or H3K4me3 alone, respectively (Figure 6.4A). The observation suggested that the interaction between the two antagonistic histone marks may be required for the regulation of miRNA expressions. To test this speculation, we aimed to compare the expression levels of miRNAs genes modified by H3K4me3, H3K27me3 or no modification. Since miRNA genes belonging to a same family generally produce identical or highly similar mature miRNAs, the expression of a particular miRNA gene was quantified by counting the amount of RNA-seq tags mapped to the pre-miRNA 'fold-back' region. Consistent with the speculation, miRNA genes modified by H3K4me3 were significantly more actively expressed than miRNA genes associated with H3K27me3 or no modification (Figure 6.4B). Therefore the results are consistent with the notion that H3K27me3 mediates the suppression of certain miRNA genes while H3K4me3 is associated with active expressions. Interestingly, we identified certain miRNA families contain members with distinct chromatin states (Figure 6.5), suggesting the differential regulation by histone modifications may contribute to the functional diversification of miRNA family members.

As expected, transposable elements (TE) were enriched in states modified with one or multiple modifications including H3K9me2, H3K27me1 or 5mC (Figure 6.4A).

Although all three chromatin marks were considered universal to heterochromatic regions, we have identified transposon families that were enriched in certain states but not others among the four states listed in the TE panel of Figure 6.4A. For example, LTR/Gypsy retro-element was strongly and specifically enriched the state 2. The LTR/Gypsy TE accounted for a strikingly 76% (597/785) of all the TEs associated with the state 2 (Table 6.2). This group of LTR/Gypse TE found in state 2 may predominantly localize in peri-centromeric regions, because TEs in state 2 were localized significantly closer to the centromere compared to TEs in other states (Figure 6.4C). Therefore the results suggest that the co-marking of H3K9me2 and H3K27me1 may be a characteristic of pericentromeric heterochromatin. The dispersed repetitive elements found throughout the chromosome arms may associate with lower level of H3K9me2, H3K27me1 or both. This speculation was supported by the finding that peri-centromeric repeats were modified with continuous and higher level of H3K9me2 compared to the repetitive sequences found in euchromatic arms (Bernatavichute et al., 2008).

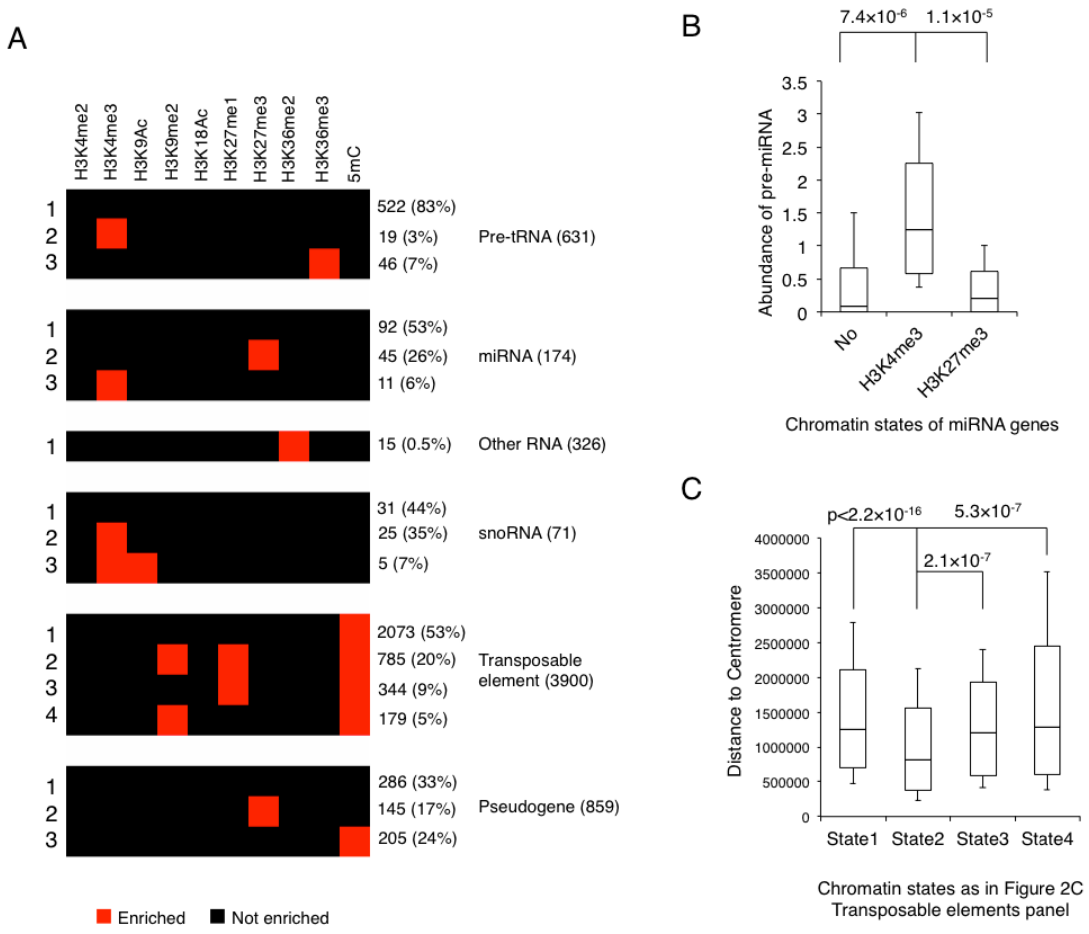

Figure 6.4. Locus type enrichment analysis of chromatin states.

(A) Significant enrichments ($p<10^{-5}$) of Pre-tRNA, miRNA, other RNA, snoRNA, transposable element and pseudogene in chromatin states. (B) Expression analysis of miRNA loci modified by H3K4me3, H3K27me3 or no modification. (C) Average distances between centromeres and transposable elements associated with states 1-4 in the transposable element panel of (A).

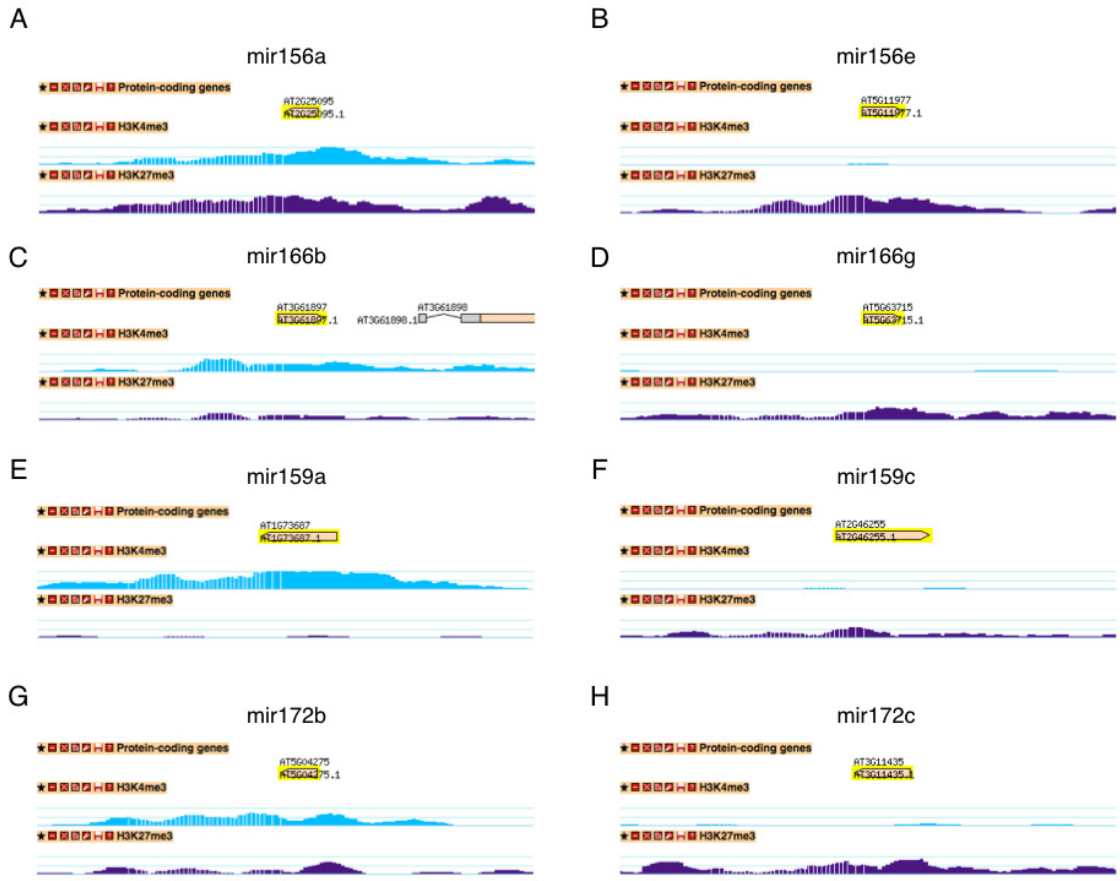

Figure 6.5. Chromatin states of members for miRNA family 156, 166, 159 and 172.

Table 6.2. Enrichments of transposon families in the four chromatin states shown in the transposable element panel in Figure 6.4A.

Brown shadings indicate significant enrichment with $p<10^{-5}$.

	State 1	State 2	State 3	State 4
LTR/Copia	9.E-11	1.E+00	1.E+00	4.E-03
DNA/En-Spm	1.E-02	1.E+00	2.E-04	4.E-01
LINE?	0.E+00	2.E-01	1.E-01	5.E-02
DNA	0.E+00	9.E-01	6.E-01	4.E-01
Unassigned	2.E-09	1.E+00	1.E+00	4.E-01
DNA/HAT	4.E-11	1.E+00	1.E+00	1.E+00
RathE1_cons	0.E+00	2.E-01	1.E-01	5.E-02
DNA/Harbinger	3.E-12	1.E+00	1.E+00	9.E-01
LTR/Gypsy	1.E+00	0.E+00	9.E-02	2.E-02
DNA/Pogo	0.E+00	5.E-01	3.E-01	2.E-01
DNA/MuDR	5.E-03	1.E+00	3.E-06	9.E-01
LINE/L1	0.E+00	1.E+00	1.E+00	6.E-01
RC/Helitron	0.E+00	1.E+00	1.E+00	1.E+00

6.3.5 The abundance of sense and antisense transcripts in chromatin states

To correlate the chromatin states with transcription outputs, the abundances of sense and antisense transcripts were plotted as the correlative patterns of the 295 chromatin states (Figure 6.6A and B). The chromatin states that showed active or repressed expressions were overall consistent with existing knowledge. Active expressions were found with the co-marking of H3K4me3, H3K9Ac and H3K36me3 (red arrows in Figure 6.6B). Three predominant states that showed repressed expressions were those enriched of 5mC, H3K27me3 (green and black arrows in Figure 6.6B) or no modification. In line with our previous analysis that NATs were more frequently associated with actively expressed genes, we observed abundant antisense signals in states associated with multiple active marks (red arrows in Figure 6.6B). We further identified two states showing conspicuous antisense signals (indicated by A and B in Figure 6.6B). The magnification of A and B regions showed that the two states were modified by H3K36me2 alone (arrows in Figure 6.6C) or together with 5mC (arrows in Figure 6.6D).

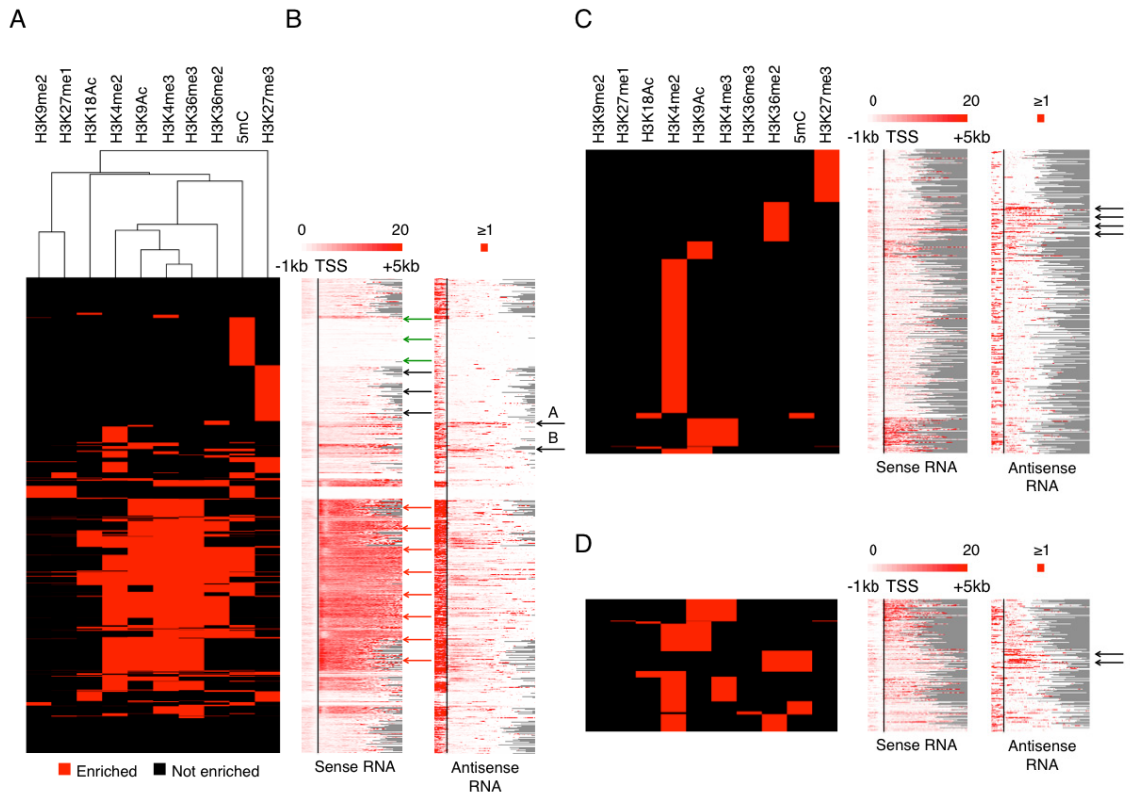

Figure 6.6. The analysis of sense and antisense transcripts as the correlative patterns of chromatin states.

(A) Hierarchical clustering of the chromatin states defined by 10 chromatin modifications as in Figure 6.1A. (B) The plotting of sense and antisense transcripts between 1 kb upstream to 5 kb downstream of TSS as correlative pattern of (A). (C and D) The magnification of the region indicated by the arrow A (C) and B (D), which correspond to the state modified by H3K36me2 alone or together with 5mC respectively. The arrows in (C) and (D) indicate the chromatin states being discussed in the text.

The correlation between chromatin states and transcript abundances was further analyzed by ordering the states according to the abundance of sense and antisense transcripts (Figure 6.7). The analysis with sense transcripts showed that the top 19 states showing active expressions in Figure 6.7A were all associated with H3K4me3, which confirmed that H3K4me3 is essential for active transcriptions. States showing active expressions were commonly co-marked by multiple 'active marks' including H3K4me3, H3K9Ac and H3K36me3 (Figure 6.7A). Among the top 10 states in Figure 6.7A, 7 of them were simultaneously modified by H3K4me3, H3K9Ac and H3K36me3. The observation therefore suggests that multiple active marks may function additively to support the transcriptions.

On examining the correlation between chromatin states and the abundances of antisense transcripts, we identified two types of states that associated with active antisense signals (Figure 6.7B). Consistent with the previous analysis showing NATs were more frequently discovered at actively expressed loci, states modified by one or more active marks were found at the top of the chart (e.g. rank 1, 3 and 5-10, Figure 6.7B). As shown by the visual correlative analysis (Figure 6.6), the states that either modified by H3K36me2 (4^{th}) alone with together with 5mC (2^{nd}) were among the states showing most abundant antisense transcripts (Figure 6.7B). The two states contained relatively few genes – states H3K36me2 and H3K36me2+5mC each contained 279 and 145 genes respectively and may thus be sensitive to annotations errors. We assessed the annotation quality of the genes within the two states that showed evidence of NATs with either cDNA or RNA-seq evidences. The results showed that 83% (53/64) and 91% (31/34) of genes in H3K36me2 and H3K36me2+5mC states respectively were supported

by at least one of the cDNA or RNA-seq evidences. Therefore, the observed enrichment of antisense transcripts in the two states were likely valid and were not caused by potential mis-annotations.

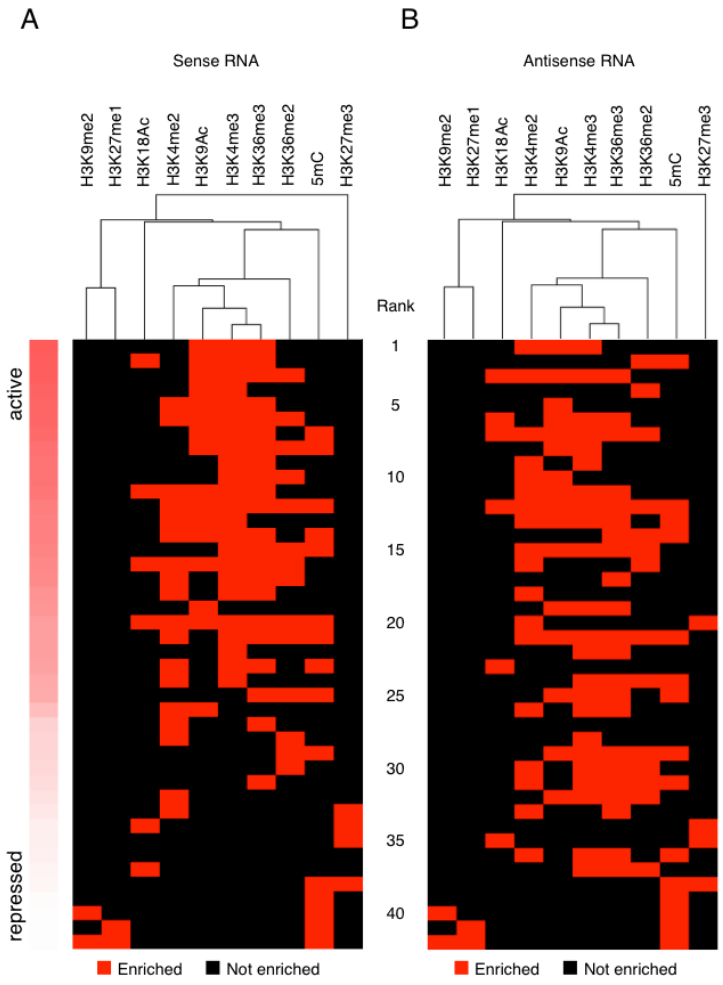

Figure 6.7. Chromatin states ordered by the abundance of sense (A) or antisense transcripts (B).

The quantity of sense transcripts were defined by the average amount of RNA-seq tags per mRNA length for a particular gene. The abundance of antisense transcripts corresponds to the amount of antisense tags per gene length.

6.4 Discussion

We defined chromatin states with the enrichment information for 10 chromatin modifications, which yielded only 42 major states containing more than 100 genes. The number of states was considerably little considering 1,024 theoretically possible states can be produced from the integration of 10 modifications. Compared to some previous works that described the epigenome with merely few states (Roudier et al., 2011), our chromatin state definition preserved more quantitative information for each individual modifications. GO term and locus type enrichment analyses were thus enabled at an unprecedented resolution. Taking advantage of the resolution for the analysis, we identified intriguing impacts of individual modifications on GO term and locus type enrichments. For example, transcription factors were significantly enriched in the state modified solely with H3K27me3 but not together with H3K18Ac. Similarly, LTR/Gypsy type of retroelements were strongly enriched only in the state co-marked by H3K9me2, H3K27me1 and 5mC. In many cases, it was not apparent why the presence or absence of a chromatin mark would interfere with the enrichment of certain function groups, especially for modifications that were poorly characterized such as H3K18Ac. Nevertheless, this type of analysis was only possible with the fine determination of chromatin states containing information for all individual marks.

The identification of two minor states – H3K36me2 and H3K36me2+5mC that were enriched of antisense transcripts further demonstrated the power of our fine chromatin states definition. These two states also associated with other intriguing features such as the unexpected enrichment of H3K36me2 around TSS as well as repressed expressions (Luo and Lam, 2010). Since the two states associated with merely few hundred genes,

they were not likely to be discovered with the highly abstractive summary of epigenome with few chromatin states.

With the analysis of locus types in chromatin states, we identified the striking enrichments of several non-coding RNA species in as few as three states. The enrichment of miRNA genes in states of no modification, H3K4me3 or H3K27me3 may be particularly functional relevant. miRNA and polycomb mechanisms were both known to be critical for the regulation of plant developments. Our results suggest that the expression of some miRNA may be modulated by the polycomb repressive mechanism. Intriguingly, we found evidences that members of a common miRNA family can associate with distinct chromatin states in the aerial tissue of *Arabidopsis* plants. Presumably the distinct chromatin states can contribute to the differential expression patterns of miRNA family members, which may allow the fine tuning of mature miRNA abundance in a given tissue or cell types.

Our bidirectional re-ChIP assay for the first time provided an estimation for the fraction of putative K4/K27me3 bivalent loci that are truly bivalent. The results showed that around or more than 50% of the putative bivalent loci are caused by the superimposition of opposing patterns presumably from different cell types. Therefore, re-ChIP-seq would be necessary to identify bivalent loci at a genome-wide scale. Another implication of the result is that the state of histone modifications can be drastically remodeled (from H3K4me3 monovalent to H3K27me3 monovalent or *vice versa*) between cell types. Except for few examples such as the vernalization response of FLC chromatin, dynamic remodelings of chromatin states are not well characterized in plants.

The adaptation of specific cell type isolation approaches will be needed to further address this question (Brady et al., 2007; Deal and Henikoff, 2010).

7 The development of State Specific Effects Analysis (SSEA) for the incorporation of chromatin state context information with correlative analysis

7.1 Introduction

To derive the function of a particular chromatin modification, commonly gene loci significantly modified by the modification were compared with the rest of the gene-space regarding certain outputs such as gene expressions. We refer to this type of relatively simple analysis as Global Effects Analysis (GEA). As shown in Figure 7.1A, the GEA of chromatin modification X compared between gene groups that were either or not enriched in X regardless of the chromatin contexts A, B or C that were defined by the modifications other than X. However, as chromatin in a given region can simultaneously associate with many modifications, this type of analysis did not account for the presence of other chromatin modifications other than the one being queried. To overcome this common caveat of correlative analyses, we introduced State Specific Effect Analysis (SSEA) for the incorporation of chromatin state contexts to improve the sensitivity of correlative analyses on epigenomic features. The principle of SSEA was to separately interrogate the function of X in distinct chromatin state contexts, such as A, B and C in Figure 7.1B, which were defined by the ensemble of chromatin modifications other an X in a given genomic region. In this section, we will compare the results obtained from GEA and SSEA regarding the function of H3K4me2 and H3K36me2 and emphasize the difference between the two strategies.

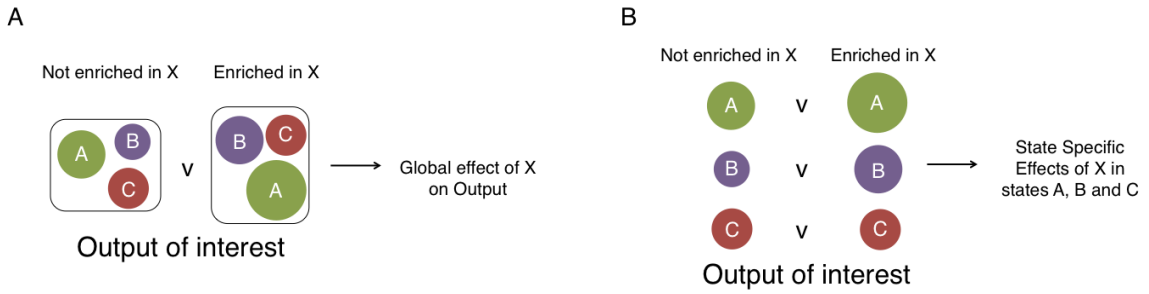

Figure 7.1. Schematic comparison of GEA (A) and SSEA (B).

X indicates the chromatin modification being interrogated. The distinct chromatin state contexts that are defined by chromatin modifications other than X were represented as A, B and C.

7.2 Results

7.2.1 GEA and SSEA of the correlations between H3K4me2 and H3K4me3 in TSS regions.

In order to analyze the impact of H3K4me2 enrichment on the pattern of H3K4me3 at TSS regions, we quantitatively compared H3K4me3 between genes that were significantly modified with H3K4me2 and the rest of the gene space. Such GEA identified a positive correlation between H3K4me2 and H3K4me3 – the average enrichment of H3K4me3 mark around TSS was 1.93 fold higher in genes enriched of H3K4me2 compared to the rest (Figure 7.2B). However a substantially different conclusion was reached with SSEA. SSEA was performed for H3K4me2 between 11 pairs of control and positive state pairs (Figure 7.2A). The states of all chromatin modifications except for H3K4me2 were identical between each control and positive state pairs (Figure 7.2A). The functional relevance of H3K4me2 enrichments was estimated by comparing various outputs of interests (e.g. gene expressions, patterns of other chromatin modifications) between control and the corresponding positive state pairs. In this example, the average abundance of H3K4me3 in TSS regions were presented as SSE=$\log_2$(positive/control) in Figure 7.2B. A SSE greater than zero indicated the enrichment of H3K4me2 in the positive state associated with more abundant H3K4me3 whereas a negative SSE suggested a reduced quantity of H3K4me3 was observed together with the enrichment of H3K4me2 in the positive state. The statistical significance of the correlations was estimated by randomly assigning genes from the whole gene space into the control and positive states for 10,000 times (scores in Figure 7.2B).

With SSEA, we identified 3 and 8 states showing positive and negative SSE respectively (Figure 7.2B). Notably, 7 out of 8 states pairs showing negative SSEs associated with significant level of H3K4me3 in both control and positive states (Figure 7.2A and B). On the other hand, H3K4me3 was not enriched in all 3 state pairs showing positive SSEs (Figure 7.2A and B). The observations suggest that the correlation between H3K4me2 and H3K4me3 is dependent of the enrichment state of H3K4me3. For actively expressed genes that are commonly modified by H3K4me3 at high level, the enrichment of H3K4me2 may negatively affect the abundance of H3K4me3 in TSS regions. This speculation is consistent with the fact that a single histone H3 lysine 4 cannot be di-methylated and tri-methylated simultaneously. Thus the increase of H3K4me2 may lead to a corresponding decrease in H3K4me3. The discrepancy between GEA and SSEA conclusions can be explained by the apparent enrichment of H3K4me3 modified genes in the group that associates with a significant level of H3K4me2. 75% (9321/12405) of genes enriched of H3K4me2 were modified by H3K4me3, whereas only 24% (4885/20598) were marked by H3K4me3 for genes that are not enriched of H3K4me2. Therefore without classifying genes according to their chromatin state contexts as SSEA did, the uneven distribution of H3K4me3 modified genes would necessarily lead to the conclusion that H3K4me2 and H3K4me3 were positively correlated at a global scale.

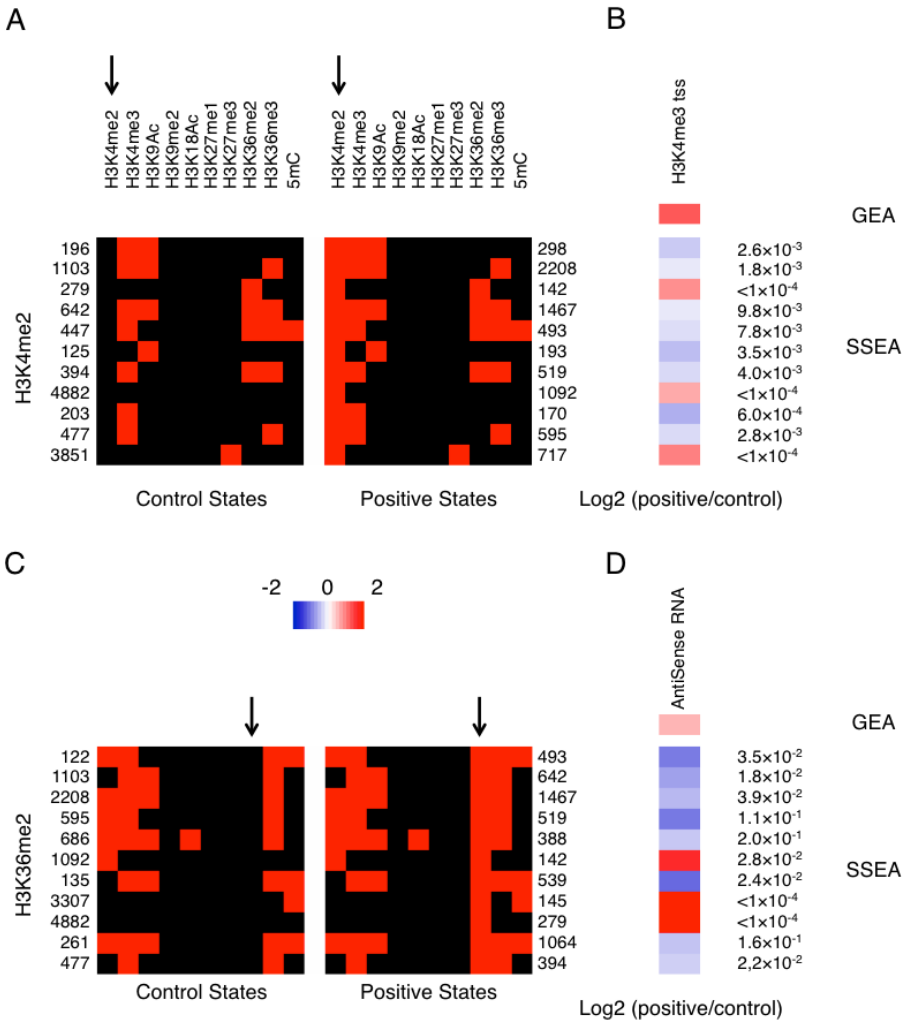

Figure 7.2. SSEA of the correlation between the state of H3K4me2 and the enrichment of H3K4me3 at TSS regions (A and B), the state of H3K36me2 and the abundance of antisense transcripts (C and D).

Arrows in (A) and (C) indicate the chromatin modifications being interrogated. The values in (B) and (D) represent the statistical significance of the observed difference between the control and positive states as determined by the permutation test. The numbers of genes associated with each state were indicated adjacent to the states in (A) and (C).

7.2.2 GEA and SSEA of the correlation between H3K36me2 and the abundance of antisense transcripts

The SSEA method was applied to analyze the functional relevance of H3K36me2. GEA identified a positive correlation between H3K36me2 and antisense transcripts with a $\log_2$(positive/control) equaled to 0.57. Among 11 control and positive state pairs, we identified 3 and 8 state pairs associated with positive and negative SSEs respectively (Figure 7.2C and D). Intriguingly, all 3 states pairs associated with positive SSEs were not modified by H3K4me3 (Figure 7.2C and D). The positive states for the 3 state pairs were H3K4me2+H3K36me2, H3K36me2 and H3K36me2+5mC, respectively (Figure 7.2C and D). The results were consistent with the previous analysis showing that chromatin states H3K36me2 and H3K36me3+5mC associated with abundance antisense transcripts (Figure 6.6). Correspondingly, all 8 state pairs associated with negative SSEs were modified by significant levels of H3K4me3 (Figure 7.2C and D). Collectively, the results suggest that the state of H3K4me3 may modulate the interaction between H3K36me2 and antisense transcripts. H3K36me2 may be supportive for the generation of antisense transcripts in the absence of H3K4me3 while be suppressive for antisense transcripts in genes modified by H3K4me3.

7.3 Discussions

A crucial motivation for defining chromatin states was the hypothesis that the functional output of a chromatin segment may be determined by the combinatorial patterns of all covalent modifications and other structural variations present in the region. The hypothesis suggests that the integrated impact of multiple chromatin modifications may be different from the simple addition of individual effects. For example, the co-

marking of multiple modifications can generate synergistic outputs. Alternatively, the effect of one modification can be modulated by the presence of other chromatin marks. These theoretical scenarios can describe certain known interactions between histone modifications. For example, the phosphorylation of histone H3 serine 10 can interfere with the binding of HP1 with heterochromatic mark H3K9me2 (Grewal and Jia, 2007). Therefore it is likely that the repressive function of H3K9me2 is inhibited by the presence of H3S10 phosphorylation. Importantly, without *a priori* knowledge, traditional correlative analysis such as GEA is unlikely to reveal the complex interactions between chromatin modifications. Analysis of H3S10ph or H3K9me2 individually may successfully assign the two marks to active and repressive chromatin respectively. However, the interaction between H3S10ph and H3K9me2 would not be discovered without the investigation of genomic regions that are modified together by H3S10ph and H3K9me2.

SSEA was designed to resolve the complex interplay between chromatin modifications through the incorporation of chromatin state contexts. We identified that the state of H3K4me3 may modulate the interaction between H3K36me2 and antisense transcripts, which resemble the regulation of H3K9me2 output by H3S10ph. As the name of SSEA indicated, the identification of chromatin states is the prerequisite for the application of SSEA. The minor H3K36me2 and H3K36me2+5mC states need to be first determined before the reported prediction can be made by SSEA. Although the predicted interactions between chromatin marks remain to be verified by genetic and biochemical approaches, our results nevertheless showed that SSEA is a effective approach for

harnessing chromatin state information and facilitate the functional study of chromatin modifications.

8 The identification of Polymerase Associated Factors (PAF) as potential regulators of Natural Antisense Transcripts (NATs).

8.1 Introduction

Except for few examples, the mechanism for NATs biogenesis is not well characterized in plants. The transcription of FLC associated NAT - COOLAIR was initiated by a cold responsive promoter located immediately downstream of the polyadenylation site of FLC (Swiezewski et al., 2009). The cloning of COOLAIR showed that the structure of the COOLAIR was not related to the exon/intron structure of the mature FLC transcript (Swiezewski et al., 2009). However, the cloning of NATs from other loci such as RD29A or CYP707A1 has identified NATs that contained no intron sequences and had identical exon-intron junctions as the mature sense transcript (Matsui et al., 2008). In support for the prevalence of this type of NATs, we found antisense tags were enriched in exons with a ratio comparable to the sense tags (Figure 5.3). As an initial step to further address the mechanism of NATs biogenesis, the full-length sequence of more NATs belonging to different classes need to be identified by a combination of RNA-seq and classical cloning and RACE approaches.

The molecular mechanisms that are involved in the regulation of NATs are just beginning to be revealed. The COOLAIR promoter was activated during cold and can confer vernalization-like epigenetic response to a heterologous transgene (Swiezewski et al., 2009). A genome-wide study of NATs using tiling array approach has discovered that many sense and antisense transcripts pairs were coordinately regulated under stress conditions (Matsui et al., 2008). Regarding the molecular pathways that may modulate

NATs productions, a recent work showed that antisense signals were enriched in genes targeted by miRNA cleavage (Luo et al., 2009). The author tested if the NATs were produced through a pathway similar as that generates trans-acting siRNA and found that the abundances of certain NATs can be modulated by factors involved in miRNA biogenesis (Luo et al., 2009). The nonsense-mediated mRNA decay pathway has also been found to affect NAT quantities at several hundred loci in the *Arabidopsis* genome (Kurihara et al., 2009). However since the *Arabidopsis* genome produces a large variety of NATs and likely include multiple classes, many more mechanisms may be involved in the regulation of NAT productions. Chromatin modifications were known to regulate the activities of alternative promoters, which may contribute to a subset of antisense transcriptions. H3K36me was known to recruit Rpd3S histone deacetylase complexes to gene bodies and suppress the activity of intragenic cryptic promoters (Li et al., 2007a; Li et al., 2007b). DNA methylation has also been shown to modulate the selection of alternative promoters in human cells (Maunakea et al., 2010). Given the fundamental role of chromatin and general transcription factors in regulating essentially every steps of transcription, it is tempting to investigate the correlations between chromatin modifications and NATs. With our SSEA analysis discussed in the previous section, we identified differential correlations between H3K36me2 and NATs that may be modulated by the state of H3K4me3 enrichment. In this section, we further carefully investigated the correlation between NATs and H3K36me2 as well as 5mC.

8.2 Material and methods

8.2.1 Determination of the nucleus/cytoplasm partition of NATs.

The nuclei isolation from *Arabidopsis* leaf tissue was performed as the ChIP procedure described in 4.2.2 without adding formaldehyde into the nuclear isolation buffer. RNA was isolated from the nuclear pellet and total tissue using Plant RNA Purification Reagents (Invitrogen). After two rounds of RQ1 DNase (Promega) treatment, roughly equal amount (~100 ng) of total and nuclear RNA were used for reverse transcriptions. The quantification was normalized with the abundance of nuclear encoded 18S rRNA as the internal control.

8.3 Results

8.3.1 NATs were depleted in the gene bodies of actively expressed genes that were significantly modified with H3K36me2 and/or 5mC.

To explore the correlation between NAT abundance and H3K36me2 and/or 5mC in the context of loci enriched of H3K4me3, a chromatin state cluster was generated containing the 14,205 genes modified with H3K4me3 but with differential enrichment of H3K36me2 and 5mC (Figure 8.1A). The patterns of sense and antisense transcripts were plotted as the correlative patterns of the chromatin state cluster (Figure 8.1B and C). The abundance of sense transcripts associated with states 1, 3 and 4 appeared to be similar by visual examinations (Figure 8.1B), whereas substantially less antisense signals were observed in states 3 and 4 compared to state 1 (Figure 8.1C). Further analysis showed that considerably more NATs determined by the stringent definition were discovered in state 1 than in state 3 or 4 (Figure 8.1G). However the fractions of genes associated with

NATs determined by the relaxed definition (excluding stringently defined NATs) were similar between the four states (Figure 8.1G). Since NATs defined by the stringent definition were commonly supported by more tags, the results suggest that NATs may accumulate to greater abundances for loci associated with the state 1. We compared the amount of antisense tags per gene length for loci associated with the four chromatin states (Figure 8.1I). As expected, states 2, 3 and 4 associated with significantly less antisense tags per gene length compared to the state 1 (Figure 8.1I). The expressions of genes associated with the four chromatin states were quantified as the amount of sense tags per mRNA length (Figure 8.1H). The test of statistical significances showed that loci associated with state 1 and 3 expressed at similar levels, whereas states 2 and 4 expressed at significant lower levels compared to the state 1 (Figure 8.1H). However, the difference between state 1 and 4 for the abundance of antisense tags (state 4/state 1 median ratio = 0.28) was much more substantial than that for sense transcripts (state 4/state 1 median ratio = 0.91). Therefore, the quantitative results support our visual observation that the enrichment of H3K36me2 and/or 5mC in the gene bodies of actively expressed genes inversely correlated with the abundance of NATs but not sense transcripts.

Since H3K36me2 was known to repress intragenic promoters through depleting histone acetylation, we examined whether the enrichment of H3K36me2 and 5mC may inverse correlate with ‘active’ chromatin marks in gene bodies. H3K4me3, H3K9Ac and H3K18Ac were plotted as the correlative patterns for the four chromatin states shown in the Figure 8.1A (Figure 8.1D-F). All the three histone modifications showed significantly less enrichments in states 3 and 4 for 3’-gene body regions compared to state 1 (Figure 8.1D-F and Figure 8.2). Therefore the data support our hypothesis that H3K36me2 and

5mC may synergistically deplete ‘active’ histone marks from gene bodies and consequently repress the production of antisense transcripts.

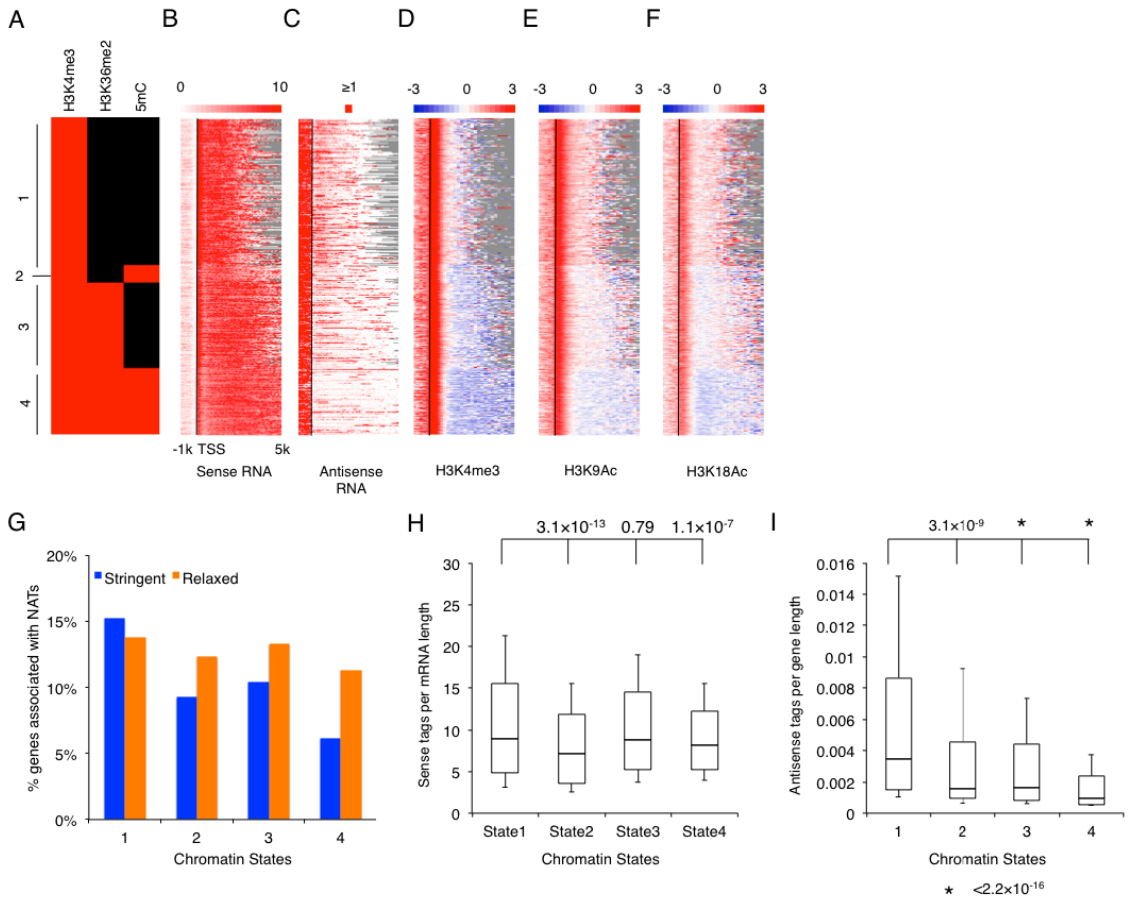

Figure 8.1. The abundance of sense, antisense transcripts and histone modifications in chromatin states modified by H3K4me3 but with differential associations of H3K36me2 and 5mC.

(A) The clustering of chromatin states defined by H3K4me3, H3K36me2 and 5mC for all H3K4me3 enriched genes. (B-F) Patterns of sense transcripts, antisense transcripts, H3K4me3, H3K9Ac and H3K18Ac were plotted as the correlative patterns of (A). (G) The fractions of loci associated with NATs defined with stringent or relaxed definitions in the four chromatin states shown in (A). (H) The quantity of sense transcripts per mRNA length for the four chromatin states shown in (A). (I) The quantity of antisense

tags per gene length in the four chromatin states shown in (A). The significance levels of the quantitative differences were determined with Mann-Whitney test.

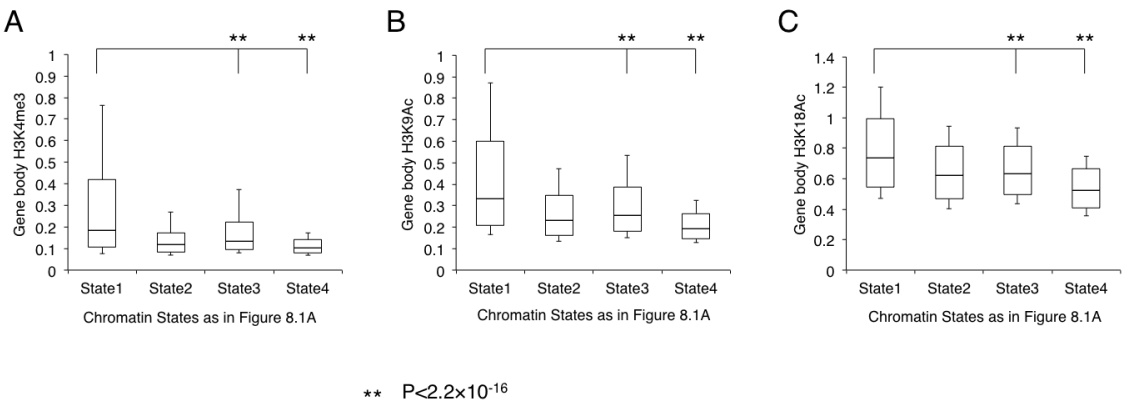

Figure 8.2. Enrichment of H3K4me3 (A), H3K9Ac (B) and H3K18Ac (C) in 3'-gene bodies (from 1,000 bps downstream of TSS to TTS) for genes associated with the four chromatin states shown in Figure 8.1A.

The significance levels of the quantitative differences were determined with Mann-Whitney test.

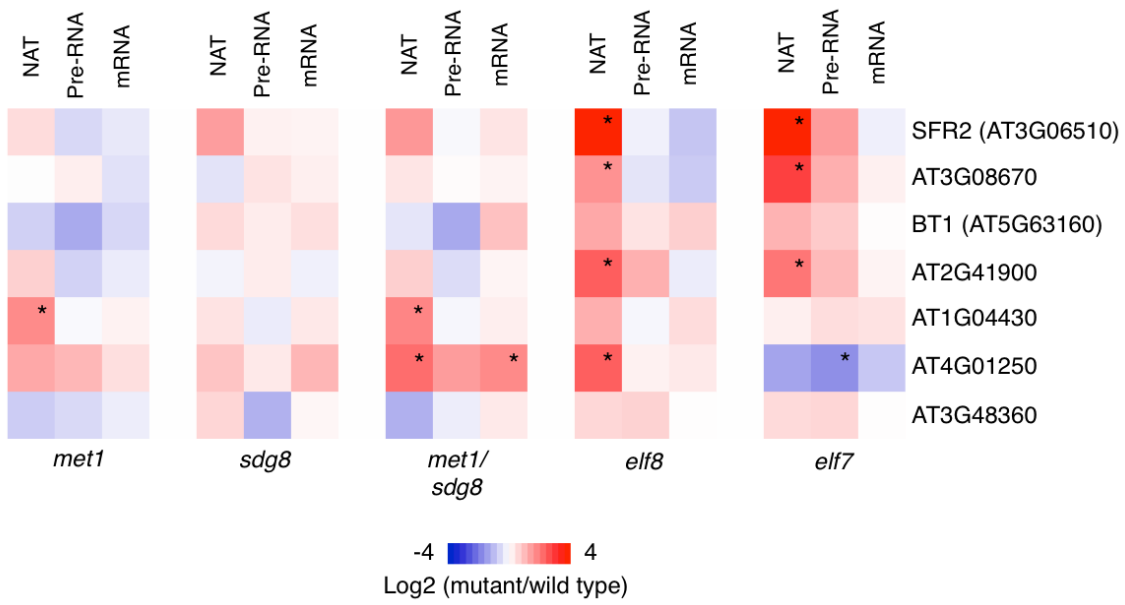

Figure 8.3. Quantifications of NAT, pre-RNA and mature mRNA in different genetic backgrounds with strand-specific qRT-PCR for 7 randomly chose loci that are enriched of H3K4me3.

Asterisks indicate greater than 3 fold up- or down- regulations compared to wild type plants.

8.3.2 Mutant analysis identified Polymerase Associated Factors (PAF) as potential regulators of NATs abundances.

To test whether H3K36me2 and 5mC were involved in the regulation of NATs productions, the abundances of NATs were determined in wild type Col, *met1-1*, *sdg8-2*, *met1-1/sdg8-2*, *elf8-1* and *elf7-3* with the strand-specific RT-PCR assays described in section 5.1.2. *met1-1* carried a loss-of-function mutation for *Arabidopsis* CpG methyltransferase - MET1; The gene encoding the major histone H3K36 methyltransferase - SDG8 was knocked-out by a T-DNA insertion in *sdg8-2*. *elf8-1* and *elf7-3* each disrupt the Paf and Ctr9 subunit of Polymerase Associated Factor (PAF) complexes. Subunits of the PAF complexes were included in the analysis because the mutant of Ski8 subunit for PAF complexes showed aberrant patterns of H3K36me2 as well as H3K4me3 at a global scale (Oh et al., 2008).

From the results obtained from NATs associated with 7 randomly chosen loci that were modified with H3K4me3, the abundance of NATs were upregulated for more than 3 folds at one locus in *met1-1*, two loci in *met1-1/sdg8-2*, four loci in *elf8-1* and three loci in *elf7-3* (Figure 8.3). The results showed that disrupting DNA methylation or H3K36me2 alone were not sufficient to cause significant mis-regulations of NATs. With McrBC-qPCR method that includes the specific digestion of methylated DNA with McrBC enzyme followed by qPCR quantification, we showed that the gene bodies of AT3G06510 and AT3G08670 were effectively demethylated in *met1-1* (Figure 8.4). Also *sdg8-2* led to drastic decreases of H3K36me2 in the two tested loci (Figure 8.5C and D). Therefore the lack of apparent changes regarding NATs abundance in *met1-1* and *sdg8-2* single mutants was not due to the inefficient removal of chromatin marks. We fount that a

significant percentage of NATs were over-accumulated for more than 3 folds in elf8-1 and elf7-3 (Figure 8.3). Importantly, three NATs showed coordinated changes of abundances in *elf8-1* and *elf7-3*, which supports that the effects were not subunit specific and were indeed caused by the disruption of PAF function (Figure 8.3). Therefore our results implicated that PAF is a potential regulator of NATs abundance and may affect a significant portion of genomic loci.

We further examined the mechanism that may mediate the over-accumulation of NATs. The level of H3K36me2, H3K9Ac, H3K18Ac and H4K12Ac were determined for the 3'-gene body region of two loci showing more than 3 folds over-accumulation of NATs (Figure 8.5). The results showed *sdg8-2* caused much more severe reductions of H3K36me2 compared to *elf8-1* without significantly affecting the level of NATs. Therefore the abundances of NATs may be modulated by PAF through a mechanism independent of the H3K36me2 mark. Further, we found no hyper-acetylation of either histone H3 or H4 associated with the NATs over-accumulation in *elf8-1* background, suggesting the depletion of active histone marks in the 3'-gene body may not be essential for repressed NATs productions.

Since COOLAIR was postulated to establish the stable repression at the FLC locus, we investigated the possibility that NATs may participate in the regulation of the cognate sense locus. The abundances of pre-RNA and mature mRNA were determined for the 7 tested loci in different genetic backgrounds. We observed no general correlation between the changes of transcript abundances from the two strands (Figure 8.3). Particularly, the NAT associated with AT3G06510 was upregulated for about 50 fold in *elf8-1*. However no significant changes of either sense pre-RNA or mature mRNA were

observed for this locus in *elf8-1*. Therefore, our results suggest that NATs may not be commonly involved in the regulation of cognate locus in cis.

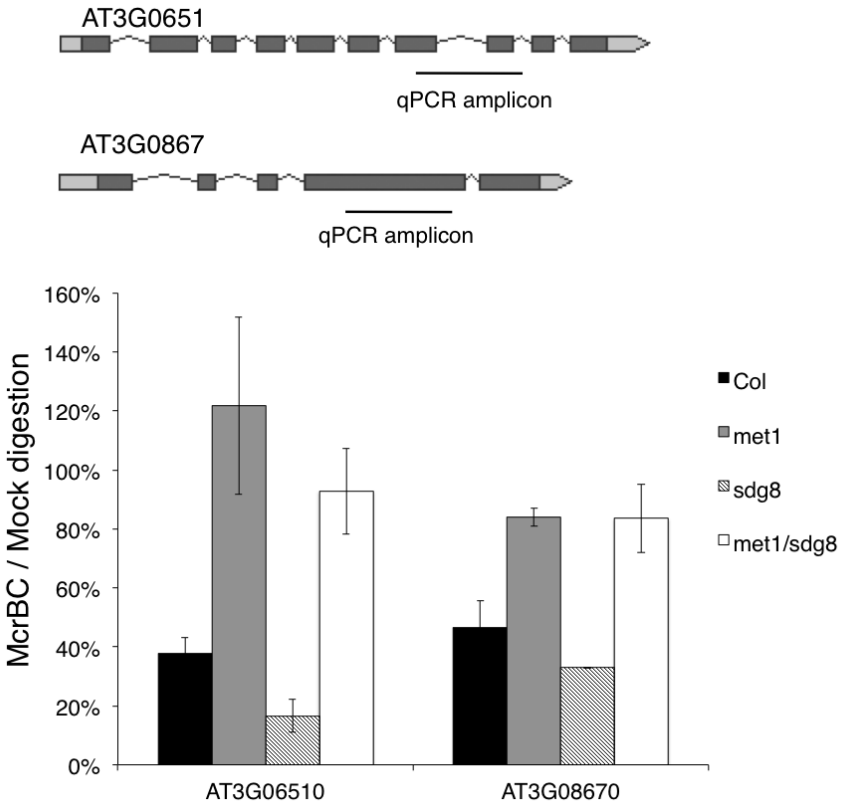

Figure 8.4. Quantitative determination of DNA methylation in the gene bodies of AT3G06510 and AT3G08670 for different genetic backgrounds.

The PCR amplicons used for the McrBC-qPCR assays were shown underneath the gene models.

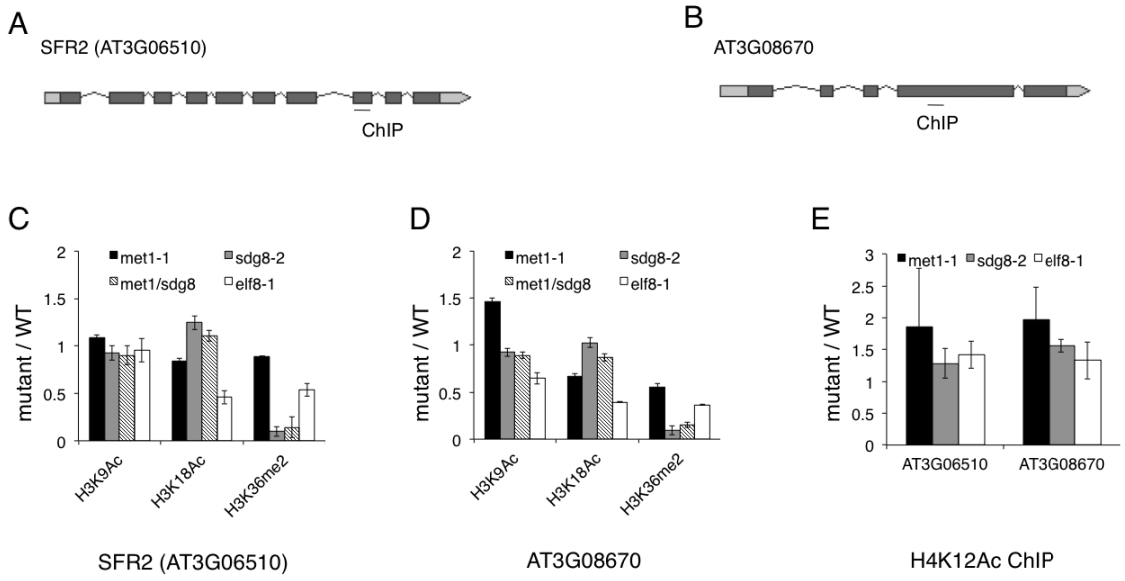

Figure 8.5. Determination of the enrichment for histone modifications in 3'-gene body regions of AT3G06510 and AT3G08670.

The qPCR amplicons used for ChIP-qPCR were shown underneath the gene models.

8.3.3 Subcellular localizations of NATs identified evidences for both nuclear- and cytoplasmically-localized NATs.

Certain lncRNAs including NATs have been speculated to modulate chromatin–related functions. It is thus expected that at least a portion of NATs should be partitioned into the nucleus to interact with chromatin and its binding proteins. We determined the abundance of NATs in RNA isolated from total tissue or nuclei preparation to estimate the subcellular localization of these RNA molecules. Among the 11 NATs that we have tested, 8 of them appeared to be localized in the nucleus with enrichment ratios comparable to their corresponding pre-RNAs, suggesting the 8 NATs were predominantly localized in the nucleus. The rest 3 NATs behave closely resemble the two mature mRNA controls. Therefore, we observed evidences for both nuclear- and cytoplasmically-localized NATs.

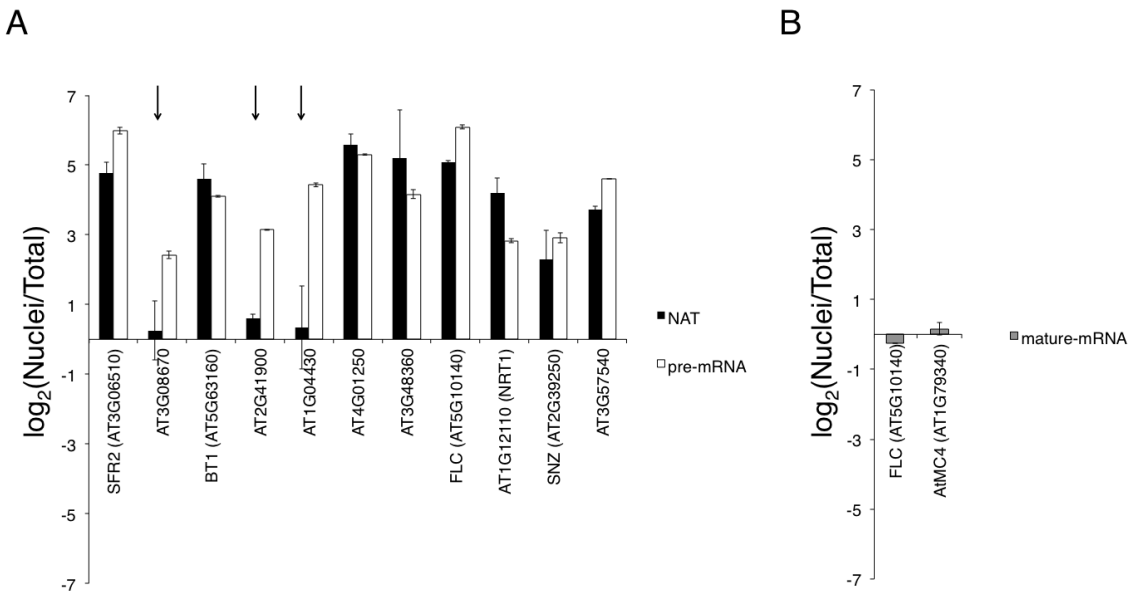

Figure 8.6. Determination of the subcellular localization of NATs.

The abundances of NATs were determined in similar amount of total and nuclear RNA with strand-specific qRT-PCR assays. The quantification was normalized by the abundance of the nuclear encoded 18S rRNA. Two qPCR probes specifically designed for the mature (spliced transcripts) of FLC and AtMC4 was used for controls.

8.4 Discussions

The mutant analysis described in this section was stimulated by the hypothesis that H3K36me2 and 5mC may suppress 'active' histone marks in 3'-gene bodies and subsequently inhibit the production of NATs. Although the enrichment of H3K36me and/or 5mC was inversely correlated with NATs abundance and multiple 'active' histone modifications, little of the proposed molecular mechanisms was supported by the genetic test. The perturbation of 5mC or H3K36me2 alone were clearly not sufficient for the mis-regulation of NATs. Also, the notable over-accumulation of NATs at AT3G06510 and AT3G08670 in *elf8-1* did not associate with any hyperacetylation of histone H3 or H4. Through comparing the extent of H3K36me2 depletion in *sdg8-2* and *elf8-1*, we further concluded that the over-accumulation of NATs was not dependent on H3K36me2 mark. At this point, the molecular mechanism for PAF-mediated NATs repression is completely unknown. Although we have demonstrated ANCORP and SSEA were effective methods for performing functional analysis of chromatin modifications, the results of the genetic analysis nevertheless suggested that clear correlations may not correspond with direct regulatory interactions. Due to the complexity of epigenetic regulation and the correlative nature of ANCORP and SSEA, any proposed interactions between epigenomic and transcriptomic components definitely need to be tested with genetic or biochemical approaches.

Consistent with the postulated function of some lncRNAs in regulating chromatin-related functions, a significant portion of NATs tested in the current work were localized in the nucleus. However for the establishment of strand-specific RT-PCR assays (section 5.3.4), we targeted only NATs showing intronic antisense tags for the analysis. This

strong selection of the type of NATs being analyzed may compromised the generality of the conclusion for NATs subcellular localizations. It will be critical to analyze the nuclear/cytoplasm partitioning of other types of NATs, especially for the type having identical exon/intron structure as the sense transcripts.

The biological functions of lncRNAs have been subjected to active debates. The characterization of lncRNAs in different organisms has reached similar models that lncRNAs can regulate the chromatin structure through the interaction with chromatin modifying complexes (Rinn et al., 2007; Khalil et al., 2009; Swiezewski et al., 2009; Huarte et al., 2010; Heo and Sung, 2011; Wang et al., 2011). lncRNAs have been shown to function either *in cis* or *in trans*. The COLDAIR lncRNAs was found to be crucial for the establishment of polycomb repression at the cognate FLC locus through the interaction with PRC2 complex (Heo and Sung, 2011). Instead, a systematic suppression of lncRNAs in ES cell showed that lncRNAs primarily function *in trans* (Guttman et al., 2010). From our analysis of NATs associated with 7 loci in different genetic backgrounds, we found no apparent correlation between the regulation of sense and antisense transcripts. Therefore our results implicate that NATs may not usually regulate the gene expression *in cis*. The functions of these NATs may be addressed through a combination of reverse genetics and biochemical approaches. It will be important to develop siRNA or artificial miRNA methods to suppress NATs and characterize the associated phenotypes. The identification of NATs interacting proteins or binding loci will be further needed for the interpretation of any observed phenotypes caused by NATs knock-down. The proteins that interact with lncRNA may be identified by ‘pull down’ assays using synthetic biotinylated RNA molecules (Huarte et al., 2010). For the identification of global

RNA/DNA interactions, the existing high throughput chromosome conformation capture (5C) method may be modified to ligate proximal DNA and RNA.

9 Development of cell-type specific epigenomics with isolation techniques for labeled cell or nucleus.

9.1 Introduction

The differentiation of distinct cell types is essential for the sophisticated body-plan and life styles of multicellular organisms. Mature human body contains over 200 distinct cell types while about 40 cell types can be defined in dicotyledon angiosperm such as *Arabidopsis* thaliana. As early as the first embryonic cleavage, the asymmetric cell division generates daughter cells with divergent cell fates. For example, the division of *Arabidopsis* zygote generates apical and basal daughter cells. The descendants of the apical cell contribute to the vast majority of plant bodies whereas only the quiescent center and the stem cells of central root caps were derived from the basal cell (Laux et al., 2004). The distinction of cell types provides the structure basis for the complexity of multi-cellular organisms. For example, each root layer launches specific transcriptomic responses under abiotic stresses (Dinneny et al., 2008); The suppression of DDM1 and the resulting de-suppression of transposable elements were detected in the vegetative nucleus but not sperm cells of pollens (Slotkin et al., 2009). Given the pervasiveness of cell type specific regulatory programs in any multi-cellular organisms, epigenomic studies using bulk tissue containing heterogeneous cell populations would not be sufficient to address the mechanisms and functions of epigenetic regulation in many biological processes.

Fluorescence Activated Cell Sorting (FACS) is the most established method for isolating specific cell types. Cells are separated to different fractions according to the

fluorescence intensity of one or multiple channels. The common approach for labeling plant cells with fluorescent tags is to generate stable transgenic lines expressing fluorescence protein gene driven by cell type specific promoters. For example, series of enhancer trap lines and transgenic lines carrying Green Fluorescence Protein (GFP) driven by specific promoters were generated, which enabled the labeling of all known cell types of *Arabidopsis* roots (Brady et al., 2007). This collection of the root-cell-type labeling lines has been used for the generation of root-cell-type-specific transcriptome profiles. Since plant cells are surrounded by the rigid cell walls, cells need to be released from plant tissues through the process of protoplasting that involves the digestion of cell walls. However, protoplasting generally suffers from low efficiency especially for inner layer cells that are less accessible for digesting enzymes. Although the efficiency of protoplasting is generally acceptable for performing transcriptome profiling, more cells may be needed for reproducible ChIP-seq experiments. ChIP usually recovers about or less than few percent of input DNA, presumably caused by the low efficiency of formaldehyde mediated crosslinking.

Recently, a novel approach for isolating nuclei from specific cell types was developed and named as Isolation of Nuclei TAgged in specific Cell Types (INTACT, Deal and Henikoff, 2010). INTACT method involves generating transgenic lines expressing a nuclear targeting fusion protein (NTF) containing a GFP protein fused with the WPP domain of *Arabidopsis* RAN GTPase activating protein 1 (RanGTP1) as well as a biotin ligase recognition pepetide (BLRP) allowing the in vivo biotinylaiton mediated by the E. Coli biotin ligase BirA (Deal and Henikoff, 2010). The NTF protein is localized to the external surface of nuclear envelop as guided by the WWP domain for the

exposure of biotin after cell lysis. The NTF is driven by a specific promoter to establish a cell type specific expression pattern while the BirA expression is driven by the constitutive Actin2 promoter (Act2p). To isolate nuclei carrying NTF label on the envelope, cells are first gently lysed to expose the naked nuclei. The labeled nuclei can then be enriched by the specific binding of biotin to streptavidin beads.

In order to establish the platform for cell type specific epigenome profiling, we experimented with specific cell or nuclei isolation using FACS and INTACT techniques, respectively. We showed that specific *Arabidopsis* root cell types can be effectively isolated by FACS. However except for epidermal and stele cells, FACS is not likely to produce sufficient quantity of cells for epigenomic profiling. We initiated the effort for setting up a high-throughout and Gateway-compatible INTACT system for addressing the interaction between developmental regulators and chromatin modifications. The progress for this project will be discussed.

9.2 Material and methods

9.2.1 Isolating specific cell populations from *Arabidopsis* root with FACS

Arabidopsis seedlings were grown vertically in large quantities to provide the material for root protoplasting. Roughly 20 µl of *Arabidopsis* seeds were germinated on each 60mm × 15mm square petridish. Gently place an autoclaved nylon mesh on the surface of solid growth medium (0.5 × MS salt, 1% sucrose solidified with 2.5g/L phytagel). Seeds were aligned into two horizontal lines on the nylon mesh to allow the convenient root harvesting. Followed by 2 days of imbibing at 4°C, the petridish were placed vertically in a regular plant growth chamber. The root tissue was harvested from

9-days-old (including two days of imbibing) plants by slicing the roots into fine pieces on the nylon mesh. Transfer the sliced root tissues into a 70 micron cell strainer (Falcon, 352350) and sink the cell strainer into a 60mm × 15mm petridish containing ~7ml of protoplasting solution A (600 mM Mannitol, 2mM $MgCl_2$, 0.1% BSA, 2mM $CaCl_2$, 2mM MES pH=5.5, 10mM KCl, 1.5% (w/v) cellulysin [Calbiochem, cat #. 219466] and 0.1% pectolyase [Sigma, cat #. 3026]). The tissue was incubated with solution A at room temperature for 1 hour with gentle swirling. After the incubation, transfer all the liquid together with any debris in the petridish into a 15 ml centrifuge tube and centrifuge at 500 × g for 5 min. After discarding the supernatant, the pellet was resuspended in 500 µl solution C (600 mM Mannitol, 2mM $MgCl_2$, 2mM $CaCl_2$, 10mM Hepes pH=7.6 and 10mM KCl). The cell suspension was passed through a 35 micron cell strainer (Falcon, cat #. 352235) to completely remove any debris. In order to crosslink the protoplasts, the volume of cell suspension was measured with pipette tips and formaldehyde was added to a final concentration of 1% (v/v). The cell suspension was then incubated at room temperature for 10 min. After the incubation, 2M glycine was added to the suspension to a final concentration of 125mM to stop the crosslinking reaction. The crosslinked cells are ready to be loaded onto the cell sorter.

The crosslinked cells were sorted according to the ratio between the emissions at 520 mm (GFP) and 585 mm (autofluorescence). The cells giving a 520mm/585mm ratio greater than a threshold would be collected. The GFP-positive cells were collected with 2X Nuclear Lysis Buffer (100mM Tris-HCl pH=7.5, 2% SDS and 20mM EDTA, pH=8.0) to allow immediate cell and nuclei lysis. The lysate can then be diluted by ChIP dilution buffer and used for ChIP assays as described in section 4.2.2.

9.2.2 Cloning for the generation of Gateway-compatible INTACT vectors.

The NTF gene in the original pBluescript-ADF8p::NTF was fused with the first six bases of the second exon for *Arabidopsis* ADF8 gene from 5'-end and thus has no independent starting codon (Deal and Henikoff, 2010) . The NTF-NOS (NOS – nopaline synthase transcription terminator) fragment was amplified from pBluescript-ADF8p::NTF vector with a BamHI-ATG-forward (5'-cgGGATCCatgGATCATTCAGCGAAAACCACA -3') and a reverse-XbaI primer (5'-gcTCTAGAcatctagtaacatagatgacaccgc-3'). The PCR product was digested with BamHI/XbaI and ligated into BamHI/XbaI digested pBlusescript SKII vector to generate the CY1 construct. The 'spacer' GUS coding sequence was amplified from pCCharting#9 with a SpeI-forward (5'- GGactagtATGGTCCGTCCTGTAGAAACC-3') and a reverse-XhoI-SstI primer (5'- cgGAGCTCctcgagTCATTGTTTGCCTCCCTGC-3'). The PCR product was cloned by EcoRV digested pBluescript SKII to generate the CY2 construct. The Gateway attR1-ccdB-Cm(R)-attR2 cassette (frame B) was amplified from pEarleyGate-100 plasmid with a HindIII-forward (5'-cgaagcttATCACAAGTTTGTACAAAAAAGCTGA-3') and a reverse-BglII primer (5'-cgagatctCACCACTTTGTACAAGAAAGCTGAA-3'). The PCR product was cloned into EcoRV digested pBluescript SKII plasmid to generate the CY3 construct. In order to place a TMV omega translational enhancer sequence in the 5'-end of the NTF gene, the EcoRI-BamHI-omega-BglII DNA was ordered as single strand oligos and annealed (5'-AATTAggatccTATTTTTACAACAATTACCAACAACAACAAACAACAAACAACATTACAATTACTATTTACAATTACAa-3' and 5'-gatctTGTAATTGTAAATAGTAATTGTAATGTTGTTTGTTGTTTGTTGTTGTTGGT

AATTGTTGTAAAAATAggatcct-3'). The annealed molecular contain 5'-EcorRI and 3'-BglII sticky ends. The double strand synthetic DNA was ligated with EcoRI/BamHI digested CY1 to generate the CY4 construct. The GUS coding sequence fragment was then excised from CY2 with SpeI/SstI digestion and subcloned into XbaI/SstI digested CY4 to generate the CY5 construct. To incorporate the Gateway cassette, attR1-ccdB-Cm(R)-attR2 cassette was excised from CY3 with HindIII and partial BglII digestion. The fragment was subcloned by the ligation with HindIII/BamHI digested CY5 to generate CY6. The gene expression construct was mobilized to the binary vector by subcloning the HindIII/XhoI fragment excised from CY6 into HindIII/SalI digested pCambia-1300 vector to generate CY7. To finally incorporate the Act2::BirA cassette, KpnI/SalI fragment excised from pCambia-3301-Act2::BirA was subcloned into KpnI/SalI digested pBluescript SKII (Deal and Henikoff, 2010) to generate CY8. The KpnI/SpeI fragment of CY8 containing the Act2::BirA cassette was ultimately subcloned into the KpnI/XbaI digested CY7 to generate the CY9 construct that is ready for recombining with pENTR/D/SD vectors containing promoter sequences.

Promoter sequences were subcloned into the pENTR/D/SD vector for further recombination. CaMV 35S promoter was amplified from pEarleyGate-100 vector with (5'- CACCTCCAATCCCACAAAAATCTGA-3' and 5'-CGTGTCCTCTCCAAATGAAA-3'). The 6XUAS-35S (-46..8) GAL4 responsive promoter was amplified from EL700 with (5'- CACCGCCGGTCGACTCTAGAGGAT-3' and 5'- CAGCGTGTCCTCTCCAAATG-3'). The AtMYB60 promoter driving specific guard cell expressions were amplified from *Arabidopsis* genomic DNA with (5'-CACCtggttgcactaagttcggttt-3' and 5'-tctctctcttcctctagatctctctga-3'). The CaMV 35S,

6XUAS-35S (-46..8) and AtMYB60p fragments were subcloned into PENTR/D/SD with TOPO reaction to generate CY10, CY11 and CY12 constructs respectively. CY10, CY11 and CY 12 plasmids were recombined with CY9 with Gateway LR Clonase II enzyme mix (Invitrogen cat #. 11791-100) to generate CY13, CY14 and CY15 constructs respectively.

9.3 Results

9.3.1 Isolation of GFP labeled root cell types with FACS for chromatin immunoprecipitations.

To determine the specificity of the cell isolation with FACS, we tested the gating parameter with wild-type and transgenic WEREWOLF::GFP-ER plants expressing endoplasmic reticulum (ER) localized GFP driven by WEREWOLF promoter that is specific for the non-hair root epidermal cells (Figure 9.1). The result showed that the gating condition was highly stringent because 0% of the protoplasts released from wild-type plants were identified to be GFP-positive (Figure 9.1A). With the identical gating parameter, 35.7% of the protoplasts released from WEREWOLF::GFP-ER plants was found to have significant GFP signals (Figure 9.1B).

The total preparation of protoplasts from WEREWOLF::GFP-ER plants and the population isolated by FACS was analyzed by epi-fluorescence microscope to empirically determine the percentage of GFP-positive cells in the two samples (Figure 9.2). Cells were identified through a combination of bright field imaging (Figure 9.2A and C) and DAPI staining (not shown). We found 83% (25/30) of cells imaged in the FACS-isolated fraction were GFP positive compared to only 28% (13/47) in the total

protoplast preparation. The latter number was consistent with the percentage of GFP-positive protoplasts (~35%) obtained from the WEREWOLF::GFP-ER line as measured by the flow cytometry (Figure 9.1). We thus conclude that our FACS isolation parameters were sufficiently stringent for purifying cell types to relatively high purity (i.e. >80%) using our existing *Arabidopsis* lines.

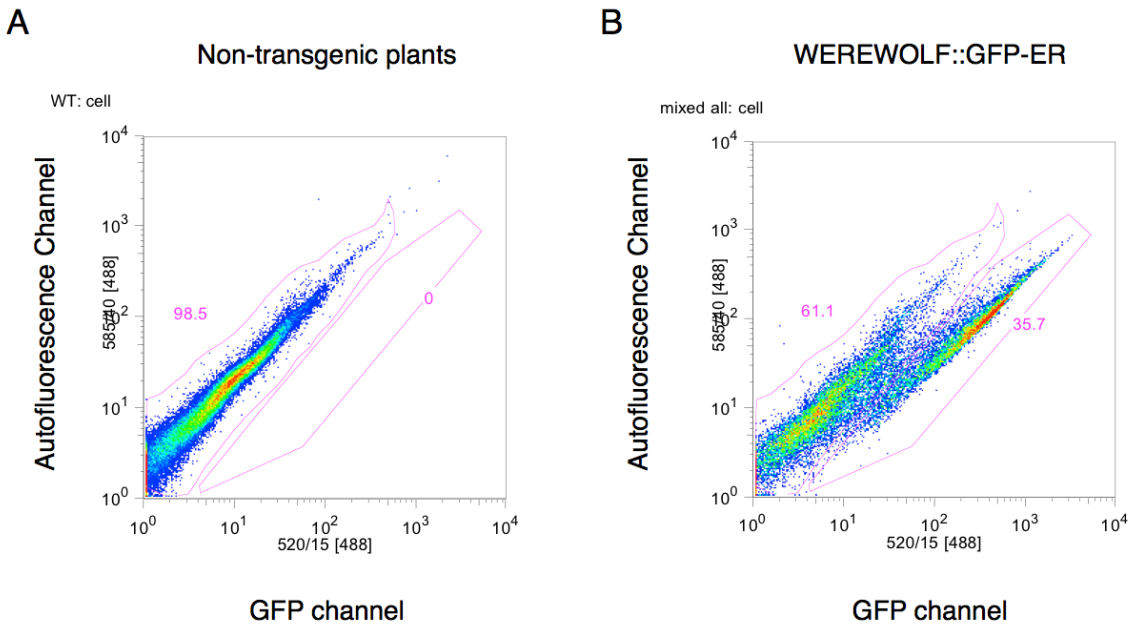

Figure 9.1. Flow cytometry of protoplasts released from wild-type and WEREWOLF::GFP-ER plants.

Protoplasts with values of 520mm/585mm emission fallen into the gate on the right were isolated as GFP-positive cells. Numbers adjacent to gating regions indicate the percentage of cells fallen into the corresponding gate. (The experiment was performed and analyzed by Jean Wang).

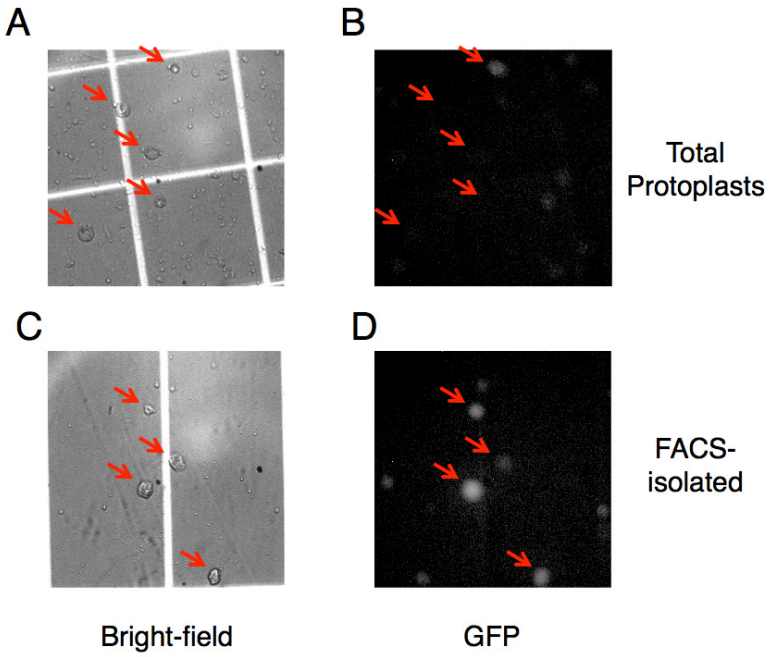

Figure 9.2. Measurement of the proportion of GFP-positive cells in total protoplasts (A, B) and FACS-isolated fraction (C, D) via bright-field and fluorescent microscopy.

Confidently identified cells were indicated with red arrowheads. The same viewing fields are shown in respective pairs of panels as indicated.

We estimated that 100,000 cells may be needed as the starting material for the construction of ChIP-seq libraries providing reproducible results. The yields of FACS-isolated protoplasts for major root cell layers (epidermis, cortex, endodermis and stele) were tested to estimate the amount of needed plant material and equipment running-times. 100,000 cells can be isolated from the epidermis and stele layer with 10 petridishes of *Arabidopsis* plants and 1 hour operation of the cell sorter. Obtaining sufficient cells for the cortex and endodermis layers were much more challenging. Protoplasts released from cortex or endodermis layers only accounted for ~0.3% of total protoplast prepared from *Arabidopsis* root. This was likely caused by the poor protoplasing efficiency for inner cell layers such as cortex or endodermis due to the reduced accessibility of digesting enzymes. It was evident that protoplasing was most efficient for epidermal cell, as solely the non-hair epidermal cells constitutes a disproportional ~35% of the total root protoplast preparation.

With the protoplasts isolated from the non-hair epidermal cells and stele cells, we determined the enrichment of H3K4me3 and H3K9Ac at the TSS region of Werewolf (WER) and Wooden Leg (WOL) genes (Figure 9.3). WER and WOL were specifically expressed in non-hair epidermal cells and stele cells, respectively. The TSS region of WER associated with a much greater H3K4me3 and to a less extend H3K9Ac enrichments in epidermal cells compared to stele cells, which is consistent with the specific expression of WER in epidermal cells (Figure 9.3A). Similarly, the TSS region of the stele specifically expressed WOL associate with stronger H3K9Ac enrichment in stele cells than in epidermal cells, although the enrichment of H3K4me3 was not so different at WOL between the two cell types. Therefore, we have demonstrated that

ChIP-qPCR assays targeting specific genomic loci can be performed with the protoplasts isolated by FACS for the non-hair epidermal and stele cells. However, the ChIP assays with FACS-isolated protoplasts yield minute amount of DNA and was not detectable by Agilent Bioanalyzer 2100 (data not shown). With the same ChIPed DNA sample, we also failed to construct ChIP-seq libraries using the standard protocol. Successful ChIP-chip experiments have been reported with as few as 1,000 cells (Dahl et al., 2009). The addition of a whole-genome-amplification step before the construction of sequencing library may help to overcome the difficulties caused by the low abundance of ChIPed DNA.

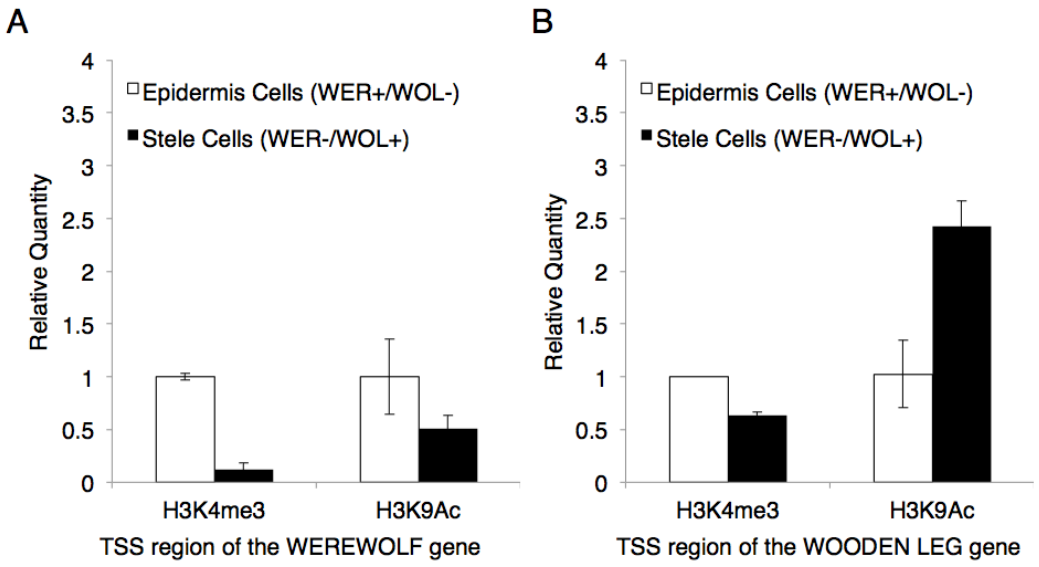

Figure 9.3. The enrichment of H3K4me3 and H3K9Ac marks at the TSS regions of WEREWOLF and WOODENLEG genes in epidermis and stele cells isolated with FACS.

100,000 cells labeled with WER::GFP-ER or WOL::GFP-ER were isolated with FACS. Approximately 15,000 cells were used for each ChIP reaction with antibody each against H3K4me3 or H3K9Ac. qPCR was performed to compare the enrichment of each modification at the Transcription Start Site (TSS) of the WER or WOL gene in the two cell types. The ChIP signals were normalized with that in the epidermal cells.

9.3.2 The establishment of high-throughput and Gateway compatible INTACT system for nuclei isolation from specific cell types.

The original INTACT system consisted of two separate vectors each containing the Act2p::BirA and NTF expression cassette respectively (Deal and Henikoff, 2010). The usage of two separate vectors may help to prevent the interference between the strong Act2 promoter and the promoter driving cell type specific expressions. However, the design placed substantial hurdle for assimilating the INTACT method into different genetic backgrounds because crossing or multiple transformations would be needed to generate plants containg both vectors. In addition, promoter sequences need to be subcloned into the construct by traditional restriction sites based approaches, which prevents the high-throughput generation of INTACT vectors targeting many different cell types.

To generate a convenient and high-throughput version of INTACT vector, we combined the Act2p::BirA and the NTF expression cassette into a single vector. To avoid any interference between the Act2 and cell type specific promoters, we placed the two promoters at the distal ends of the constructs (Figure 9.4). In addition, a ~1.8 kb GUS coding sequence was placed between the two cassettes as the 'spacer sequence' (Figure 9.4). To enable the convenient replacement of the cell type specific promoter that drives the NTF expression, a Gateway attR1-ccdB-Cm(R)-attR2 cassette (frame B) together with a Tobacco Mosaic Virus omega translation enhancer sequence was placed at the 5'-end of the NTF gene (Figure 9.4). The final construct was named CY9 and its structure was shown in Figure 9.4.

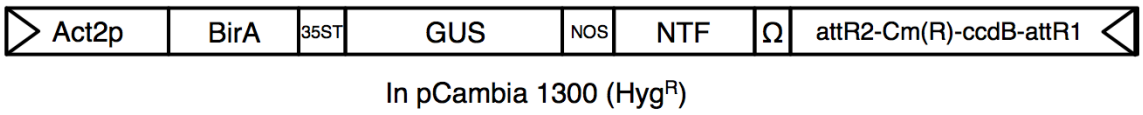

Figure 9.4. The CY9 construct for the Gateway-compatible INTACT implementation.

Act2p: *Arabidopsis* Actin 2 promoter. BirA: E.Coli biotin ligase. 35ST: CaMV 35S transcription terminator. GUS: beta-glucuronidase coding sequence. NOS: nopaline synthase transcription termination. NTF: nuclear targeting fusion protein. Ω: Tobacco Mosaic Virus translation enhancer sequence. attR2-Cm(R)-ccdB-attR1: Gateway attR2-Cm(R)-ccdB-attR1 cassette (frame B).

To test the performance of our Gateway-compatible INTACT system, CaMV 35S promoter, the guard cell specific AtMYB60 promoter and the GAL4 responsive UAS-35S(-46..8) promoter were cloned in the pENTR/D vector and recombined with CY9 to generate CY13, CY15 and CY14 constructs, respectively. The expression of NTF protein was tested with transient expression assays in tobacco leaves. The infiltration of Agrobacterium carrying CY13 construct produced constitutive fluorescence signals, suggesting the NTF protein was successfully expressed (Figure 9.5A). However, the NTF expressed from our vectors appeared be localized to both cytoplasm and nucleus, whereas the NTF expressed in *Arabidopsis* root epidermal cells were strictly localized to the nucleus envelops (Figure 9.5A and B, Deal and Henikoff, 2010). Interestingly, when NTF expression was driven by the weaker AtMYB60 promoter, a higher proportion of the fluorescence signal appeared to partition into the nucleus. The observation suggests that the cytoplasmic localization of NTF may be caused by the over-production of the protein. As nuclear envelope can only anchor certain amount of the NTF protein, the excess protein being produced may be trapped in the cytoplasm where they are produced. In addition, the AtMYB60 promoter did not appear to drive guard cell specific expression in tobacco leaf (Figure 9.5B). Instead, fluorescence signals were detected in guard cells as well as subsidiary cells (red arrows in Figure 9.5B). We detected no apparent fluorescence signal with the infiltration of Agrobacterium carrying CY14 UAS-35S(-46..8)::NTF construct, the UAS-35S(-46..8) promoter was not expected to drive significant expression without the presence of yeast GAL4 trans-activator. Therefore the result suggests that our construct design can support the specific expression pattern of NTF and no apparent activating effect from the constitutive Act2 promoter was observed.

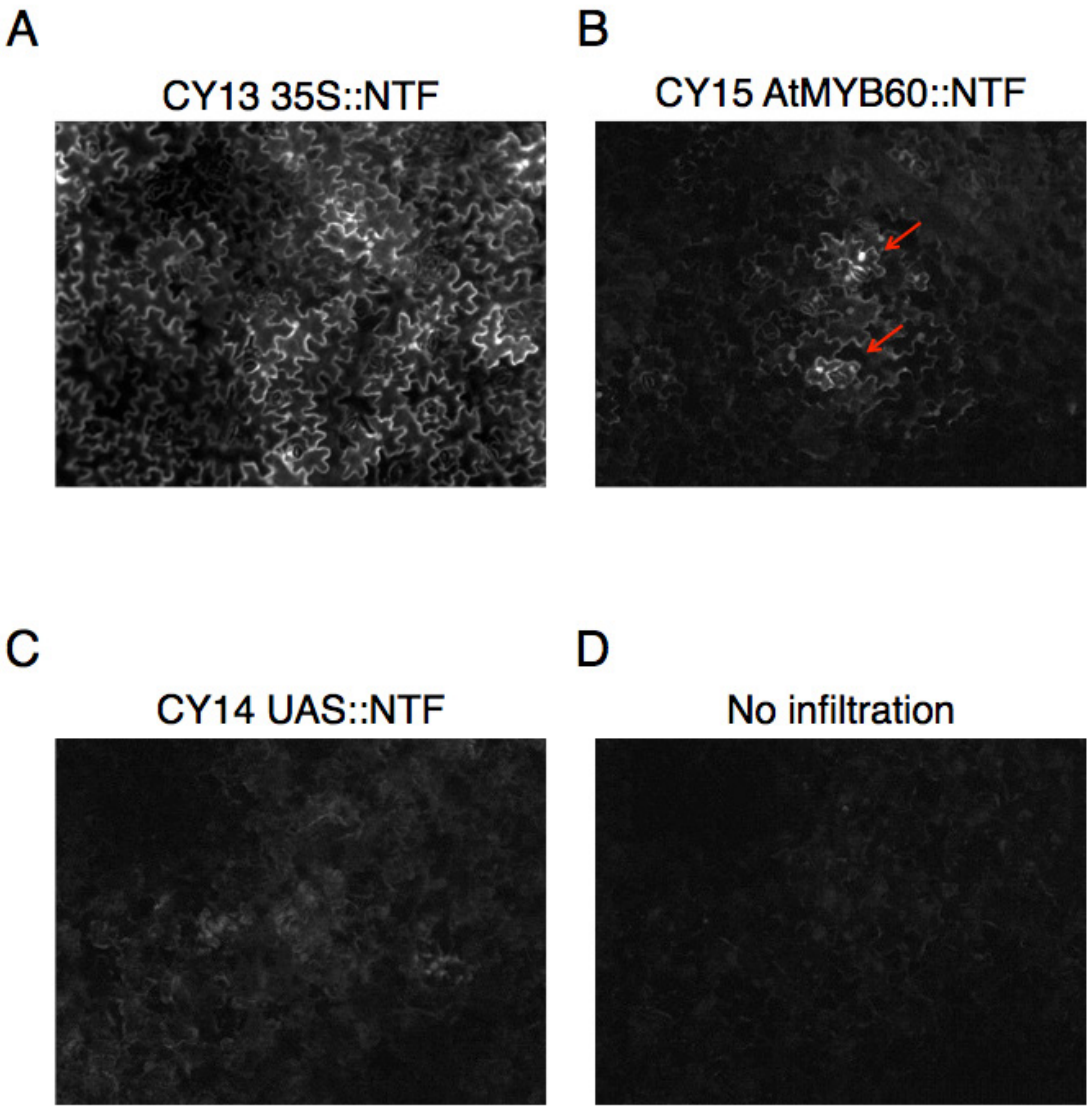

Figure 9.5. Transient expression of the assimilated INTACT vectors in tobacco leaf through the infiltration of Agrobacterium.

10 Future directions

10.1 Next generation epigenomics – integrating chromatin modification maps with quantitative transcription networks.

Much efforts of this dissertation have been dedicated to studying the interactions between epigenomic and transcriptomic components. Experimental approaches including ChIP-seq and RNA-seq were applied to the model plant *Arabidopsis* to generate quantitative profiles. We further developed novel informatics methods to visualize and identify any correlations between epigenomic and transcriptomic features. However, the works addressed little of an important question – how do epigenetic mechanisms interact with the established transcription networks to direct various biological processes?

For many chromatin modifications that are pertinent to gene expression regulations, analyzing the mutant of the responsible chromatin-modifying enzyme provides the very first information for the function of the chromatin mark. For example, central flowering regulator FLC is down-regulated in the mutants of H3K4 methyltransferase ATX1 and ATXR7 as well as H3K36 methyltransferase SDG8, which suggests the histone marks H3K4me3 and H3K36me2 are essential for supporting the FLC expression (Zhao et al., 2005; Pien et al., 2008; Tamada et al., 2009). However, much remains unknown regarding the regulation of FLC by the two histone marks - 1) Chromatin modifying complexes rarely contain sequence-specific DNA binding activities. Therefore the recruitment of chromatin modifying complexes is commonly mediated by the general transcription machinery or the interaction with sequence-specific transcription factors. So far, no transcription factor or other chromatin bind protein has been identified to mediate

the recruitment of histone methyltransferases to the FLC locus. Although FRIGIDA (FRI) is known as a dominant activator of FLC that presumably functions at a co-transcriptional step (Geraldo et al., 2009), it is not known if FRI affects the recruitment of histone methyltransferases to FLC chromatin. 2) The stages of FLC transcription that are regulated by H3K4me3 and H3K36me2 are not determined. Histone modifications may affect any stage of transcriptions including the binding of sequence-specific factors, formation for RNA polymerase pre-initiation complex, releasing of RNA polymerase from promoter or elongation. H3K4me3 and H3K36me2 have also been implicated in the regulation of alternative splicing (Sims et al., 2007; Luco et al., 2010). To answer these questions, the transcription process at FLC locus or for the whole genome need to be tracked with finer resolution. For example, the promoter occupancy of RNA polymerase II or the splicing efficiency need to be determined in the mutants of histone methyltransferases to explicitly define the 'mode-of-action' for various histone modifications.

The challenge for placing chromatin modifications into the existing gene regulation network is not unique for FLC locus. Currently there is no gene specific or generic model to integrate epigenetic components with the well established transcription networks that mainly composed of sequence-specific transcription factors. Scarce information has been collected regarding the change of chromatin conformation caused by the binding of transcription factors. The impact of chromatin modification on the affinity of *trans-cis* interactions was poorly characterized as well. Consequently, it is difficult to determine if epigenetic mechanisms function upstream, downstream or in parallel with sequence-specific transcription factors. Depending on the specific locus, chromatin modifications

may either function as the effector of the transcription machinery or as the carrier of epigenetic information, even though the information may be transient. For example, H3K36me2 may be considered as the effector of RNA polymerase II when the mark was established co-transcriptionally to regulated alternative splicing (Luco et al., 2010). In other cases, the H3K27me3 deposited around the FLC locus during vernalization clearly store and transmit the memory of prolonged cold.

Through the past decade, we have obtained the capability of profiling chromatin modifications and other chromatin structural variations at whole-genome scales. Numbers of 'reference' epigenomic maps have been produced from major tissue/cell types of various model systems by ENCODE and modENCODE projects (ENCODE Project Consortium, 2007; Celniker et al., 2009). These genome-wide maps are priceless resources for further epigenetic researches. However since the maps were generated mostly with samples prepared from reference genotype under regular growth conditions, they provided limited information regarding the interaction between epigenetic regulations and transcription networks. In my perspective, the next generation epigenomics would be combining the well-established profiling techniques with the systematic perturbations of sequence-specific transcription factors as well as chromatin modifiers. Through experimentally determining the interdependency between transcription factors and chromatin modification and other structural variations, a more integrated view of epigenomes and transcription networks may be produced.

10.2 Hypothesis for the function of chromatin modifications

The packaging of genetic materials with chromatin structure is a critical distinction between eukaryotes and prokaryotes. The types of chromatin modifications generally increase along with the complexity of organisms. For example, S.cerevisiae contains no H3K27me3, which plays critical roles in cell differentiations in perhaps all animal lineages and plants. Further, the regulatory mechanisms of chromatin marks appear to be more sophisticated in multi-cellular organisms compared to in the yeast. Knocking out subunits of PAF complex causes the complete loss of H3K4me or H3K36me marks in S.cerevisiae. However the disruption of PAF in higher plant *Arabidopsis* cause only modest changes to the patterns of the two histone marks (Oh et al., 2008). It is therefore reasonable to speculate that the expansion of chromatin-related regulatory mechanisms contributes to the overall complexity of multi-cellular organisms. However as we discussed earlier in the section 1.1, chromatin-related mechanism is not the only choice for establishing complex regulatory programs. Prokaryotes have no chromatin structure but still utilize sophisticated transcription networks to cope with environmental challenges. So what is so unique about chromatin-related mechanisms that enabled the enormous complexity found in multi-cellular organisms? I like to envisage two unique advantages of chromatin-related mechanisms that may substantiate the regulatory potential of transcription networks.

1) The marking of transcription units with chromatin modifications provided the signatures for cells to index and categorize different types of loci. For example, in plants all transposable elements are marked by CHH and CHG methylation regardless of the particular genomic context. Such indexing mechanism may substantially improve the

regulatory efficiency by facilitate the target recognition of silencing machineries. It may be challenging to directly target co-repressors such as histone deacetylase complexes to numerous types of transposable elements with various degrees of degenerations. By marking transposons with CHH and CHG methylation through RNA-directed-DNA-methylation pathway and other undefined mechanisms, the task for silencing all different transposon families was simplified to targeting loci containing CHG methylation for transcription repressions. Further, H3K9 methyltransferase KYP can be recruited to CHG methylation through SRA domain and establish the conserved heterochromatic mark H3K9me2 to expand the docking potential.

2) Chromatin structure can mediate the crosstalk between sequence-specific factors through the transient storage and transmission of regulatory information. There is little evidence suggesting chromatin modifications other than 5mC, H3K9me or H3K27me3 can perpetuate information through mitotic or meiotic cycles (see section 1.8). Along with the continuous discovery of effectors for histone modifications, it become apparent histone modifications can transiently store and relay certain cellular information. The stability or ‘half-life’ of the signal mediated by chromatin modifications are difficult to estimate due to the constant presence of ‘writers’ and ‘erasers’ of chromatin marks. Nevertheless, the lack of long-term memory does not prevent chromatin structures from serving as an integration node for regulatory inputs. Such function of chromatin structures was elegantly shown by the classical study of gene regulation at yeast PHO5 promoter. One of the two binding sites for bHLH transcription factor PHO4 need to be first exposed through chromatin remodeling before PHO4 can bind and activate the PHO5 transcription. The example showed that by modulating the structure or perhaps

also the covalent modification of chromatin, one chromatin-bound factor can modulate the binding or activity of other transcription regulators. Indeed, complex enhancers are commonly found in multi-cellular organisms containing multiple regulatory elements. Presumably the enhancer can be bound by a number of regulatory proteins simultaneously to integrate the input from different cellular pathways. However, it is not clear if all the bound regulators would directly interact with the general transcription machinery. Alternatively, some transcription regulators can function through modifying chromatin structures to tune the binding or activity of either another regulator or the general transcription machinery. Although the described mechanism is theoretically plausible as supported by the analysis of certain model systems such as PHO5 regulation, the generality of such regulation remains to be tested. We speculate that the role of chromatin to integrate regulatory inputs from multiple transcription regulators can be common and may be essential for the formation of complex transcription networks in multi-cellular organisms. As discussed in 10.1, to test this speculation at a global scale, the dynamic binding of trans-regulators and changes in chromatin structure and covalent modifications need to be tracked with fine temporal and spatial resolutions.

11 Reference

Alvarez-Venegas, R., Pien, S., Sadder, M., Witmer, X., Grossniklaus, U. and Avramova, Z. (2003). ATX-1, an Arabidopsis homolog of trithorax, activates flower homeotic genes. Curr Biol. 13: 627-637.

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. (2000) Nature. 408:796-815.

Ausin, I., Mockler, T.C., Chory, J. and Jacobsen SE. (2009). IDN1 and IDN2 are required for de novo DNA methylation in Arabidopsis thaliana. Nat Struct Mol Biol. 16: 1325-1327.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell. 129: 823-837.

Bäurle, I., Smith, L., Baulcombe, D.C. and Dean, C. (2007). Widespread role for the flowering-time regulators FCA and FPA in RNA-mediated chromatin silencing. Science. 318: 109-112.

Bernatavichute, Y.V., Zhang, X., Cokus, S., Pellegrini, M. and Jacobsen SE. (2008). Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana. PLoS One. 3: e3156.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S.L. and Lander, E.S. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315-326.

Berr, A., McCallum, E.J., Ménard, R., Meyer, D., Fuchs, J., Dong, A. and Shen, W.H. (2010). Arabidopsis SET DOMAIN GROUP2 is required for H3K4 trimethylation and is crucial for both sporophyte and gametophyte development. Plant Cell. 22: 3232-3248.

Bird, A. (2007). Perceptions of epigenetics. Nature 447: 396-398.

Bostick, M., Kim, J.K., Estève, P.O, Clark, A., Pradhan, S. and Jacobsen, S.E. (2007). UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science 317: 1760-1764.

Brady, S.M, Orlando, D.A, Lee, J.Y, Wang, J.Y, Koch, J., Dinneny, J.R, Mace, D., Ohler, U. and Benfey, P.N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. Science. 318: 801-806.

Bratzel, F., López-Torrejón, G., Koch, M., Del Pozo, J.C. and Calonje, M. (2010). Keeping cell identity in Arabidopsis requires PRC1 RING-finger homologs that catalyze H2A monoubiquitination. Curr Biol. 20: 1853-1859.

Cao, R., Tsukada, Y. and Zhang, Y. (2005). Role of Bmi-1 and Ring1A in H2A ubiquitylation and Hox gene silencing. Cell. 20: 845-854.

Celniker, S.E., Dillon, L.A., Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston RH; modENCODE Consortium. (2009). Unlocking the secrets of the genome. Nature. 459: 927-930.

Chan, S.W., Henderson, I.R. and Jacobsen, S.E. (2005). Gardening the genome: DNA methylation in Arabidopsis thaliana. Nat Rev Genet. 6: 351-360.

Charron, J.B., He, H., Elling, A.A. and Deng, X.W. (2009). Dynamic landscapes of four histone modifications during deetiolation in Arabidopsis. Plant Cell. 21: 3732-3748.

Chen, E.S., Zhang, K., Nicolas, E., Cam, H.P., Zofall, M. and Grewal, S.I. (2008). Cell cycle control of centromeric repeat transcription and heterochromatin assembly. Nature. 451: 734-737.

Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J., Casero, D., Bernal, M., Huijser, P., Clark, A.T., Krämer, U., Merchant, S.S., Zhang, X., Jacobsen, S.E. and Pellegrini, M. (2010). Relationship between nucleosome positioning and DNA methylation. Nature. 466: 388-392.

Chuang, L.S., Ian, H.I., Koh, T.W., Ng, H.H., Xu, G. and Li, B.F. (1997). Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. Science. 277: 1996-2000.

Creyghton, M.P., Markoulaki, S., Levine, S.S., Hanna, J., Lodato, M.A., Sha, K., Young, R.A., Jaenisch, R. and Boyer, L.A. (2008). H2AZ is enriched at polycomb complex target genes in ES cells and is necessary for lineage commitment. Cell. 135: 649-661.

Dahl, J.A., Reiner, A.H. and Collas, P. (2009). Fast genomic muChIP-chip from 1,000 cells. Genome Biol. 10: R13.

de Hoon, M.J., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. Bioinformatics. 20, 1453-1454.

Deal, R.B., Henikoff, J.G. and Henikoff, S. (2010). Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. Science. 328: 1161-1164.

Deal, R.B and Henikoff, S. (2010). A simple method for gene expression and chromatin profiling of individual cell types within a tissue. Dev Cell. 18: 1030-1040.

Dellino, G.I., Schwartz, Y.B., Farkas, G., McCabe, D., Elgin, S.C. and Pirrotta, V. (2004). Polycomb silencing blocks transcription initiation. Mol Cell. 13: 887-893.

Deng, X., Gu, L., Liu, C., Lu, T., Lu, F., Lu, Z., Cui, P., Pei, Y., Wang, B., Hu, S. and Cao, X. (2010). Arginine methylation mediated by the Arabidopsis homolog of PRMT5 is essential for proper pre-mRNA splicing. Proc Natl Acad Sci U S A.107: 19114-19119.

Dennis, K., Fan, T., Geiman, T., Yan, Q. and Muegge, K. (2001). Lsh, a member of the SNF2 family, is required for genome-wide methylation. Genes Dev. 15: 2940-2944.

Dinneny, J.R., Long, T.A., Wang, J.Y., Jung, J.W., Mace, D., Pointer, S., Barron, C., Brady, S.M., Schiefelbein, J. and Benfey, P.N. (2008). Cell identity mediates the response of Arabidopsis roots to abiotic stress. Science. 320: 942-945.

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

ENCODE Project Consortium. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 447: 799-816.

Faghihi, M.A. and Wahlestedt, C. (2009). Regulatory roles of natural antisense transcripts. Nat Rev Mol Cell Biol. 10: 637-643.

Feng, S., Cokus, S.J., Zhang, X., Chen, P.Y., Bostick, M., Goll, M.G., Hetzel, J., Jain, J., Strauss, S.H., Halpern, M.E., Ukomadu, C., Sadler, K.C., Pradhan, S., Pellegrini, M. and Jacobsen, S.E. (2010a). Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A. 107: 8689-8694.

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

Ficz, G., Heintzmann, R. and Arndt-Jovin, D.J. (2005). Polycomb group protein complexes exchange rapidly in living Drosophila. Development. 132: 3963-3976.

Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S. and Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature. 472: 398-402.

Filion, G.J., van Bemmel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., de Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J. and van Steensel, B. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell 143: 212-224.

Francis, N.J., Kingston, R.E. and Woodcock, C.L. (2004). Chromatin compaction by a polycomb group protein complex. Science. 306: 1574-1577.

Gendler, K., Paulsen, T., and Napoli, C. (2008). ChromDB: the chromatin database. Nucleic Acids Res. 36, D298-302.

Geraldo, N., Bäurle, I., Kidou, S., Hu, X. and Dean C. (2009). FRIGIDA delays flowering in Arabidopsis via a cotranscriptional mechanism involving direct interaction with the nuclear cap-binding complex. Plant Physiol. 150: 1611-1618.

Grewal, S.I. and Jia, S. (2007). Heterochromatin revisited. Nat Rev Genet. 8: 35-46.

Guo, L., Yu, Y., Law, J.A. and Zhang, X. (2010) SET DOMAIN GROUP2 is the major histone H3 lysine [corrected] 4 trimethyltransferase in Arabidopsis. Proc Natl Acad Sci U S A. 107: 18557-18562.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J.L., Root, D.E. and Lander, E.S. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature. 477: 295-300.

Habu, Y., Mathieu, O., Tariq, M., Probst, A.V., Smathajitt, C., Zhu, T. and Paszkowski, J. (2006). Epigenetic regulation of transcription in intermediate heterochromatin. EMBO Rep. 7: 1279-1284.

Hayakawa, T., Ohtani, Y., Hayakawa, N., Shinmyozu, K., Saito, M., Ishikawa, F. and Nakayama, J. (2007). RBP2 is an MRG15 complex component and down-regulates intragenic histone H3 lysine 4 methylation. Genes Cells. 12: 811-826.

He, X.J., Hsu, Y.F., Zhu, S., Wierzbicki, A.T., Pontes, O., Pikaard, C.S., Liu, H.L., Wang, C.S., Jin, H. and Zhu, J.K. (2009). An effector of RNA-directed DNA methylation in arabidopsis is an ARGONAUTE 4- and RNA-binding protein. Cell 137: 498-508.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., Ching, K.A., Antosiewicz-Bourget, J.E., Liu, H., Zhang, X., Green, R.D., Lobanenkov, V.V., Stewart, R., Thomson, J.A., Crawford, G.E., Kellis, M. and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 459: 108-112.

Henikoff, S. and Shilatifard, A. (2011). Histone modification: cause or cog? Trends Genet. 27: 389-396.

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

Huang, Y., Pastor, W.A., Shen, Y., Tahiliani, M., Liu, D.R. and Rao, A. (2010). The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PLoS One. 5: e8888.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., Attardi, L.D., Regev, A., Lander, E.S., Jacks, T. and Rinn, J.L. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell. 142: 409-419.

Inagaki, S., Miura-Kamio, A., Nakamura, Y., Lu, F., Cui, X., Cao, X., Kimura, H., Saze, H. and Kakutani, T. (2010). Autocatalytic differentiation of epigenetic modifications within the Arabidopsis genome. EMBO J. 29: 3496-3506.

Jacob, Y., Feng, S., LeBlanc, C.A., Bernatavichute, Y.V., Stroud, H., Cokus, S., Johnson, L.M., Pellegrini, M., Jacobsen, S.E. and Michaels, S.D. (2009). ATXR5 and ATXR6 are H3K27 monomethyltransferases required for chromatin structure and gene silencing. Nat Struct Mol Biol. 16: 763-768.

Jacob, Y., Stroud, H., Leblanc, C., Feng, S., Zhuo, L., Caro, E., Hassel, C., Gutierrez, C., Michaels, S.D. and Jacobsen, S.E. (2010). Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. Nature. 466: 987-991.

Jacobsen, S.E. and Meyerowitz, E.M. (1997). Hypermethylated SUPERMAN epigenetic alleles in arabidopsis. Science. 277: 1100-1103.

Ji, H. and Wong, W.H. (2005). TileMap: create chromosomal map of tiling array hybridizations. Bioinformatics. 21: 3629-3636.

Johnson, W.E., Li, W., Meyer, C.A., Gottardo, R., Carroll, J.S., Brown, M. and Liu, X.S. (2006). Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci U S A. 103: 12457-12462.

Johnson, L.M., Bostick, M., Zhang, X., Kraft, E., Henderson, I., Callis, J. and Jacobsen, S.E. (2007). The SRA methyl-cytosine-binding domain links DNA and histone methylation. Curr Biol. 17: 379-384.

Kanno T, Mette MF, Kreil DP, Aufsatz W, Matzke M, Matzke AJ. (2004). Involvement of putative SNF2 chromatin remodeling protein DRD1 in RNA-directed DNA methylation. Curr Biol. 14: 801-805.

Kanno, T., Bucher, E., Daxinger, L., Huettel, B., Böhmdorfer, G., Gregor, W., Kreil, D.P., Matzke, M. and Matzke, A.J. (2008). A structural-maintenance-of-chromosomes hinge domain-containing protein is required for RNA-directed DNA methylation. Nat Genet. 40: 670-675.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J. and Segal, E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature. 458: 362-366.

Kato, N. and Lam, E. (2001). Detection of chromosomes tagged with green fluorescent protein in live Arabidopsis thaliana plants. Genome Biol. 2: RESEARCH0045.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., Regev, A., Lander, E.S. and Rinn, J.L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A. 106: 11667-11672.

Kinoshita, T., Yadegari, R., Harada, J.J., Goldberg, R.B. and Fischer, R.L. (1999). Imprinting of the MEDEA polycomb gene in the Arabidopsis endosperm. Plant Cell. 11: 1945-1952.

Kloc, A., Zaratiegui, M., Nora, E. and Martienssen, R. (2008). RNA interference guides histone modification during the S phase of chromosomal replication. Curr Biol. 18: 490-495.

Klose, R.J. and Bird, A.P. (2006). Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci. 31: 89-97.

Klymenko, T., Papp, B., Fischle, W., Köcher, T., Schelder, M., Fritsch, C., Wild, B., Wilm, M. and Müller, J. (2006). A Polycomb group protein complex with sequence-specific DNA-binding and selective methyl-lysine-binding activities. Genes Dev. 20: 1110-1122.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S. and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet. 41: 376-381.

Kriaucionis, S. and Bird, A. (2003). DNA methylation and Rett syndrome. Hum Mol Genet. 12: R221-227.

Kriaucionis, S. and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science. 324: 929-930.

Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X., Chi, A.S., Adli, M., Kasif, S., Ptaszek, L.M., Cowan, C.A., Lander, E.S., Koseki, H. and Bernstein, B.E. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet. 4: e1000242.

Kurihara, Y., Matsui, A., Hanada, K., Kawashima, M., Ishida, J., Morosawa, T., Tanaka, M., Kaminuma, E., Mochizuki, Y., Matsushima, A., Toyoda, T., Shinozaki, K. and Seki, M. (2009). Genome-wide suppression of aberrant mRNA-like noncoding RNAs by NMD in Arabidopsis. Proc Natl Acad Sci U S A 106: 2453-2458.

Kwong, C., Adryan, B., Bell, I., Meadows, L., Russell, S., Manak, J.R. and White, R. (2008). Stability and dynamics of polycomb target sites in Drosophila development. PLoS Genet. 4: e1000178.

Lam, E., Kato, N. and Watanabe, K. (2004). Visualizing chromosome structure/organization. Annu Rev Plant Biol. 55: 537-554.

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S.A., Haggarty, S.J., Clemons, P.A., Wei, R., Carr, S.A., Lander, E.S. and Golub, T.R. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 313: 1929-1935.

Laux, T., Würschum, T., and Breuninger, H. (2004). Genetic Regulation of Embryonic Pattern Formation. Plant Cell. 16: S190-S202.

Law, J.A. and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 11: 204-220.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A. and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. 7: 709-715.

Lewin, B. (2006). Gene IX. Jones and Bartlett Publishers.

Lewis, M.S., Pikaard, D.J., Nasrallah, M., Doelling, J.H. and Pikaard, C.S. (2007). Locus-specific ribosomal RNA gene silencing in nucleolar dominance. PLoS One 2: e815.

Li, B., Carey, M. and Workman, J.L. (2007a). The role of chromatin during transcription. Cell. 128: 707-719.

Li, B., Gogol, M., Carey, M., Pattenden, S.G., Seidel, C. and Workman, J.L. (2007b). Infrequently transcribed long genes depend on the Set2/Rpd3S pathway for accurate transcription. Genes Dev. 21: 1422-1430.

Li, L., Wang, X., Stolc, V., Li, X., Zhang, D., Su, N., Tongprasit, W., Li, S., Cheng, Z., Wang, J. and Deng, X.W. (2006). Genome-wide transcription analyses in rice using tiling microarrays. Nat Genet 38: 124-129.

Lickwar, C.R., Rao, B., Shabalin, A.A., Nobel, A.B., Strahl, B.D. and Lieb, J.D. (2009). The Set2/Rpd3S pathway suppresses cryptic transcription without regard to gene length or transcription frequency. PLoS One. 4: e4886.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F. and Schübeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. Nat Genet. 43: 1091-1097.

Lindroth, A.M., Shultis, D., Jasencakova, Z., Fuchs, J., Johnson, L., Schubert, D., Patnaik, D., Pradhan, S., Goodrich, J., Schubert, I., Jenuwein, T., Khorasanizadeh, S. and Jacobsen, S.E. (2004). Dual histone H3 methylation marks at lysines 9 and 27 required for interaction with CHROMOMETHYLASE3. EMBO J 23: 4286-4296.

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

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B. and Ecker, J.R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462: 315-322.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M. and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science. 327: 996-1000.

Luo, C. and Lam, E. (2010). ANCORP: a high-resolution approach that generates distinct chromatin state models from multiple genome-wide datasets. Plant J. 63: 339-351.

Luo, Q.J., Samanta, M.P., Köksal, F., Janda, J., Galbraith, D.W., Richardson, C.R., Ou-Yang, F. and Rock, C.D. (2009). Evidence for antisense transcription associated with microRNA target mRNAs in Arabidopsis. PLoS Genet. 5: e1000457.

Margueron, R., Li, G., Sarma, K., Blais, A., Zavadil, J., Woodcock, C.L., Dynlacht, B.D. and Reinberg, D. (2008). Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. Mol Cell. 32: 503-518.

Margueron, R., Justin, N., Ohno, K., Sharpe, M.L., Son, J., Drury, W.J. 3rd, Voigt, P., Martin, S.R., Taylor, W.R., De Marco, V., Pirrotta, V., Reinberg, D. and Gamblin, S.J. (2009). Role of the polycomb protein EED in the propagation of repressive histone marks. Nature. 461: 762-767.

Margueron, R. and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. Nature. 469: 343-349.

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

Matsui, A., Ishida, J., Morosawa, T., Mochizuki, Y., Kaminuma, E., Endo, T.A., Okamoto, M., Nambara, E., Nakajima, M., Kawashima, M., Satou, M., Kim, J.M., Kobayashi, N., Toyoda, T., Shinozaki, K. and Seki M. (2008). Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array. Plant Cell Physiol. 49: 1135-1149.

Matzke, M., Kanno, T., Daxinger, L., Huettel, B. and Matzke, A.J. (2009). RNA-mediated chromatin-based silencing in plants. Curr Opin Cell Biol 21: 367-376.

Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V.M., Rowitch, D.H., Xing, X., Fiore, C., Schillebeeckx, M., Jones, S.J., Haussler, D., Marra, M.A., Hirst, M., Wang, T. and Costello, J.F. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 466: 253-257.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E.S. and Bernstein, B.E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

Motamedi, M.R., Verdel, A., Colmenares, S.U., Gerber, S.A., Gygi, S.P. and Moazed, D. (2004). Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. Cell. 119: 789-802.

Nagaya, S., Kato, K., Ninomiya, Y., Horie, R., Sekine, M., Yoshida, K. and Shinmyo, A. Expression of randomly integrated single complete copy transgenes does not vary in Arabidopsis thaliana. (2005) Plant Cell Physiol. 46:438-444.

Nuber, U.A., Kriaucionis, S., Roloff, T.C., Guy, J., Selfridge, J., Steinhoff, C., Schulz, R., Lipkowitz, B., Ropers, H.H., Holmes, M.C. and Bird, A. (2005). Up-regulation of

glucocorticoid-regulated genes in a mouse model of Rett syndrome. Hum Mol Genet. 14: 2247-2256.

Oh S, Park S, van Nocker S. (2008). Genic and global functions for Paf1C in chromatin modification and gene expression in Arabidopsis. PLoS Genet. 4: e1000077.

Ohad, N., Yadegari, R., Margossian, L., Hannon, M., Michaeli, D., Harada, J.J., Goldberg, R.B. and Fischer, R.L. (1999). Mutations in FIE, a WD polycomb group gene, allow endosperm development without fertilization. Plant Cell. 11: 407-416.

Oktaba, K., Gutiérrez, L., Gagneur, J., Girardot, C., Sengupta, A.K., Furlong, E.E. and Müller, J. (2008). Dynamic regulation by polycomb group protein complexes controls pattern formation and the cell cycle in Drosophila. Dev Cell. 15: 877-889.

Ooi, S.K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D., Cheng, X and Bestor, T.H. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature. 448: 714-717.

Park PJ. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 10, 669-680.

Pastor, W.A., Pape, U.J., Huang, Y., Henderson, H.R., Lister, R., Ko, M., McLoughlin, E.M., Brudno, Y., Mahapatra, S., Kapranov, P., Tahiliani, M., Daley, G.Q., Liu, X.S., Ecker, J.R., Milos, P.M., Agarwal, S. and Rao, A. (2011). Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature. 473: 394-397.

Perocchi, F., Xu, Z., Clauder-Münster, S. and Steinmetz, L.M. (2007). Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. Nucleic Acids Res. 35: e128.

Pien, S., Fleury, D., Mylne, J.S., Crevillen, P., Inzé, D., Avramova, Z., Dean, C. and Grossniklaus, U. (2008). ARABIDOPSIS TRITHORAX1 dynamically regulates FLOWERING LOCUS C activation via histone 3 lysine 4 trimethylation. Plant Cell. 20: 580-588.

Pokholok, D.K., Harbison, C.T., Levine, S., Cole, M., Hannett, N.M., Lee, T.I., Bell, G.W., Walker, K., Rolfe, P.A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D.K. and Young, R.A. (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. Cell. 122: 517-527.

Rosin, F.M., Watanabe, N., Cacas, J.L., Kato, N., Arroyo, J.M., Fang, Y., May, B., Vaughn, M., Simorowski, J., Ramu, U., McCombie, R.W., Spector, D.L. Martienssen, R.A. and Lam, E. Genome-wide transposon tagging reveals location-dependent effects on transcription and chromatin organization in Arabidopsis. (2008) Plant J. 55:514-525.

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L., Després, B., Drevensek, S., Barneche, F., Dèrozier, S., Brunaud, V., Aubourg, S., Schnittger, A., Bowler, C.,

Martin-Magniette, M.L., Robin, S., Caboche, M. and Colot, V. (2011). Integrative epigenomic mapping defines four main chromatin states in *Arabidopsis*. EMBO J 30: 1928-1938.

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

Russo, V. E. A., Martienssen, R. A. & Riggs, A. D. (1996) Epigenetic Mechanisms of Gene Regulation. Cold Spring Harbor Laboratory Press, Woodbury.

Ruthenburg, A.J., Allis, C.D. and Wysocka, J. (2007). Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. Mol Cell. 25: 15-30.

Saldanha, A.J. (2004). Java Treeview--extensible visualization of microarray data. Bioinformatics. 20, 3246-3248.

Saleh, A., Alvarez-Venegas, R., Yilmaz, M., Le, O., Hou, G., Sadder, M., Al-Abdallat, A., Xia, Y., Lu, G., Ladunga, I. and Avramova, Z. (2008). The highly similar Arabidopsis homologs of trithorax ATX1 and ATX2 encode proteins with divergent biochemical functions. Plant Cell. 20: 568-579.

Sasaki, H. and Matsui, Y. (2008). Epigenetic events in mammalian germ-cell development: reprogramming and beyond. Nat Rev Genet 9: 129-140.

Saze, H., Shiraishi, A., Miura, A. and Kakutani, T. (2008). Control of genic DNA methylation by a jmjC domain-containing protein in Arabidopsis thaliana. Science. 319: 462-465.

Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff, B.E., Bergman, Y., Simon, I. and Cedar, H. (2007). Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet. 39: 232-236.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D. and Lohmann, J.U. A gene expression map of Arabidopsis thaliana development. (2005) Nat Genet. 37:501-506.

Schubert, D., Lechtenberg, B., Forsbach, A., Gils, M., Bahadur, S. and Schmidt, R. Silencing in Arabidopsis T-DNA transformants: the predominant role of a gene-specific RNA sensing mechanism versus position effects. (2004) Plant Cell. 16: 2561-2572.

Schuettengruber, B., Ganapathi, M., Leblanc, B., Portoso, M., Jaschek, R., Tolhuis, B., van Lohuizen, M., Tanay, A. and Cavalli, G. (2009). Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. PLoS Biol. 7: e13.

Schwartz, Y.B. and Pirrotta, V. (2007). Polycomb silencing mechanisms and the management of genomic programmes. Nat Rev Genet. 8: 9-22.

Sharif, J., Muto, M., Takebayashi, S., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., Tajima, S., Mitsuya, K., Okano, M. and Koseki, H. (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature 908-912.

Shi, X., Hong, T., Walter, K.L., Ewalt, M., Michishita, E., Hung, T., Carney, D., Peña, P., Lan, F., Kaadige, M.R., Lacoste, N., Cayrou, C., Davrazou, F., Saha, A., Cairns, B.R., Ayer, D.E., Kutateladze, T.G., Shi, Y., Côté, J., Chua, K.F. and Gozani, O. (2006). ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. Nature. 442: 96-99.

Simon, J.A. and Kingston, R.E. (2009). Mechanisms of polycomb gene silencing: knowns and unknowns. Nat Rev Mol Cell Biol. 10: 697-708.

Sims, R.J. 3rd, Millhouse, S., Chen, C.F., Lewis, B.A., Erdjument-Bromage, H., Tempst, P., Manley, J.L. and Reinberg, D. (2007). Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol Cell. 28: 665-676.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdzić, M., Becker, J.D., Feijó, J.A. and Martienssen, R.A. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell. 136: 461-472.

Song, C.X., Clark, T.A., Lu, X.Y., Kislyuk, A., Dai, Q., Turner, S.W., He, C. and Korlach, J. (2011). Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nat Methods. [Epub ahead of print]

Squazzo, S.L., O'Geen, H., Komashko, V.M., Krig, S.R., Jin, V.X., Jang, S.W., Margueron, R., Reinberg, D., Green, R. and Farnham, P.J. (2006). Suz12 binds to silenced regions of the genome in a cell-type-specific manner. Genome Res. 16: 890-900.

Stock, J.K., Giadrossi, S., Casanova, M., Brookes, E., Vidal, M., Koseki, H., Brockdorff, N., Fisher, A.G. and Pombo, A. (2007). Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nat Cell Biol. 9: 1428-1435.

Stolc, V., Samanta, M.P., Tongprasit, W., Sethi, H., Liang, S., Nelson, D.C., Hegeman, A., Nelson, C., Rancour, D., Bednarek, S., Ulrich, E.L., Zhao, Q., Wrobel, R.L., Newman, C.S., Fox, B.G., Phillips, G.N. Jr., Markley, J.L. and Sussman, M.R. (2005). Identification of transcribed sequences in Arabidopsis thaliana by using high- resolution genome tiling arrays. Proc Natl Acad Sci U S A 102: 4453-4458.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 102: 15545-15550.

Swiezewski, S., Liu, F., Magusin, A. and Dean, C. (2009). Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. Nature. 462: 799-802.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L. and Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 324: 930-935.

Tamada, Y., Yun, J.Y., Woo, S.C. and Amasino, R.M. (2009). ARABIDOPSIS TRITHORAX-RELATED7 is required for methylation of lysine 4 of histone H3 and for transcriptional activation of FLOWERING LOCUS C. Plant Cell. 21: 3257-3269.

Turck, F., Roudier, F., Farrona, S., Martin-Magniette, M.L., Guillaume, E., Buisine, N., Gagnot, S., Martienssen, R.A., Coupland, G and Colot, V. (2007). Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27. PLoS Genet. 3: e86.

Vaillant, I., Schubert, I., Tourmente, S. and Mathieu, O. (2006). MOM1 mediates DNA-methylation-independent silencing of repetitive sequences in Arabidopsis. EMBO Rep. 7: 1273-1278.

Verdel, A., Jia, S., Gerber, S., Sugiyama, T., Gygi, S., Grewal, S.I. and Moazed, D. (2004). RNAi-mediated targeting of heterochromatin by the RITS complex. Science. 303: 672-676.

Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W., van Schaik, F.M., Varier, R.A., Baltissen, M.P., Stunnenberg, H.G., Mann, M. and Timmers, H.T. (2007). Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell. 131: 58-69.

Volpe, T. and Martienssen, R.A. (2011). RNA interference and heterochromatin assembly. Cold Spring Harb Perspect Biol. 3: a003731.

Vongs, A., Kakutani, T., Martienssen, R.A. and Richards, E.J. (1993). Arabidopsis thaliana DNA methylation mutants. Science. 260: 1926-1928.

Wang, H., Wang, L., Erdjument-Bromage, H., Vidal, M., Tempst, P., Jones, R.S. and Zhang, Y. (2004a). Role of histone H2A ubiquitination in Polycomb silencing. Nature. 431: 873-878.

Watt, F. and Molloy, P.L. (1988). Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. Genes Dev. 2: 1136-1143.

Wang, L., Brown, J.L., Cao, R., Zhang, Y., Kassis, J.A. and Jones, R.S. (2004). Hierarchical recruitment of polycomb group silencing complexes. Mol Cell. 14: 637-646.

Wang, X., Elling, A.A., Li, X., Li, N., Peng, Z., He, G., Sun, H., Qi, Y., Liu, X.S. and Deng, X.W. (2009). Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. Plant Cell. 21: 1053-1069.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., Wysocka, J., Lei, M., Dekker, J., Helms, J.A. and Chang, H.Y. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature. 472: 120-124.

Wierzbicki, A.T., Haag, J.R. and Pikaard, C.S. (2008). Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. Cell 135: 635-648.

Wierzbicki, A.T., Ream, T.S., Haag, J.R. and Pikaard, C.S. (2009). RNA polymerase V transcription guides ARGONAUTE4 to chromatin. Nat Genet 41: 630-634.

Wu, H., D'Alessio, A.C., Ito, S., Xia, K., Wang, Z., Cui, K., Zhao, K., Sun, Y.E. and Zhang, Y. (2011). Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature. 473: 389-393.

Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A., Rappsilber, J. and Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature. 473: 343-348.

Woo, H.R., Pontes, O., Pikaard, C.S. and Richards, E.J. (2007). VIM1, a methylcytosine-binding protein required for centromeric heterochromatinization. Genes Dev 21: 267-277.

Xu, L. and Shen, W.H. (2008). Polycomb silencing of KNOX genes confines shoot stem cell niches in Arabidopsis. Curr Biol. 18: 1966-1971.

Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M., Pham, P., Cheuk, R., Karlin-Newmann, G., Liu, S.X., Lam, B., Sakano, H., Wu, T., Yu, G., Miranda, M., Quach, H.L., Tripp, M., Chang, C.H., Lee, J.M., Toriumi, M., Chan, M.M., Tang, C.C., Onodera, C.S., Deng, J.M., Akiyama, K., Ansari, Y., Arakawa, T., Banh, J., Banno, F., Bowser, L., Brooks, S., Carninci, P., Chao, Q., Choy, N., Enju, A., Goldsmith, A.D., Gurjal, M., Hansen, N.F., Hayashizaki, Y., Johnson-Hopson, C., Hsuan, V.W., Iida, K., Karnes, M., Khan, S., Koesema, E., Ishida, J., Jiang, P.X., Jones, T., Kawai, J., Kamiya, A., Meyers, C., Nakajima, M., Narusaka, M., Seki, M., Sakurai, T., Satou, M., Tamse, R., Vaysberg, M., Wallender, E.K., Wong, C., Yamamura, Y., Yuan, S., Shinozaki, K., Davis, R.W., Theologis, A. and Ecker JR. (2003). Empirical analysis of transcriptional activity in the Arabidopsis genome. Science 302: 842-846.

Yokthongwattana, C., Bucher, E., Caikovski, M., Vaillant, I., Nicolet, J., Mittelsten Scheid, O. and Paszkowski, J. (2010). MOM1 and Pol-IV/V interactions regulate the intensity and specificity of transcriptional gene silencing. EMBO J. 29: 340-351.

Yoon, H.G., Chan, D.W., Reynolds, A.B., Qin, J. and Wong, J. (2003). N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso. Mol Cell. 12: 723-734.

Yuan W, Xu M, Huang C, Liu N, Chen S, Zhu B. (2011). H3K36 methylation antagonizes PRC2-mediated H3K27 methylation. J Biol Chem. 286: 7983-7989.

Zemach, A., McDaniel, I.E., Silva, P. and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328: 916-919.

Zhang, P., Du, J., Sun, B., Dong, X., Xu, G., Zhou, J., Huang, Q., Liu, Q., Hao, Q. and Ding, J. (2006a). Structure of human MRG15 chromo domain and its binding to Lys36-methylated histone H3. Nucleic Acids Res. 34: 6621-6628.

Zhang, X., Henriques, R., Lin, S.S., Niu, Q.W. and Chua, N.H. (2006b). Agrobacterium-mediated transformation of Arabidopsis thaliana using the floral dip method. Nat Protoc. 1:641-646.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E. and Ecker, J.R. (2006c). Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell. 126: 1189-1201.

Zhang, X., Germann, S., Blus, B.J., Khorasanizadeh, S., Gaudin, V. and Jacobsen, S.E. (2007). The Arabidopsis LHP1 protein colocalizes with histone H3 Lys27 trimethylation. Nat Struct Mol Biol. 14: 869-871.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. and Liu, X.S. (2008a). Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9: R137.

Zhang, K., Mosch, K., Fischle, W. and Grewal, S.I. (2008b). Roles of the Clr4 methyltransferase complex in nucleation, spreading and maintenance of heterochromatin. Nat Struct Mol Biol. 15: 381-388.

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

Zhao, Z., Yu, Y., Meyer, D., Wu, C. and Shen, W.H. (2005). Prevention of early flowering by expression of FLOWERING LOCUS C requires methylation of histone H3 K36. Nat Cell Biol. 7: 1256-1260.

Zheng, B and Chen, X. (2011). Dynamics of histone H3 lysine 27 trimethylation in plant development. Curr Opin Plant Biol. 14: 123-129.

Zhou, W., Zhu, P., Wang, J., Pascual, G., Ohgi, K.A., Lozach, J., Glass, C.K. and Rosenfeld, M.G. (2008). Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation. Mol Cell. 29: 69-80.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T. and Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet. 39: 61-69.