Title
Mechanisms of DNA methylation control and epigenome engineering

Author
Liu, Wanlu

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Mechanisms of DNA methylation control and epigenome engineering

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy in

Molecular Biology

by

Wanlu Liu

2018

ABSTRACT OF THE DISSERTATION

Mechanisms of DNA methylation control and epigenome engineering

by

Wanlu Liu

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2018

Professor Steven Erik Jacobsen, Chair

Cytosine DNA methylation is an evolutionarily conserved epigenetic mark that plays critical roles in diverse biological processes, including gene and transposon silencing and imprinting. In mammals, DNA methylation mostly occurs in the symmetric dinucleotide CG sites. In the model plant *Arabidopsis thaliana*, DNA methylation frequently occurs at cytosine bases in all sequence contexts (CG, CHG and CHH, where H represents A, C or T).

In *Arabidopsis*, *de novo* DNA methylation is established by a process known as RNA-directed DNA methylation (RdDM). RdDM in plants not only requires the upstream production of 24-nucleotide (nt) small interfering RNAs (siRNAs) and the downstream recruitment of *de novo* DNA methyltransferase DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2), the production of RNA polymerase V (Pol V)-dependent intergenic non-coding (IGN) transcripts plays a crucial role likely in serving as scaffolds for siRNAs binding. Pol V is

required for DNA methylation and gene silencing and has been shown to be transcriptionally active *in vitro*. However, the characteristics of Pol V transcripts is poorly understood probably due to its low abundance. In this dissertation, to better understand the features of Pol V transcripts *in vivo*, I will first describe the application of a technique modified from global nuclear run-on (GRO) assay to characterize nascent Pol V transcripts at genome-wide level. With this technique, we captured Pol V nascent transcripts and we uncovered a novel mechanism of ARGONAUTE4/6/9 (AGO4/6/9) dependent, small-RNA-guided co-transcriptional slicing of nascent Pol V transcripts.

With the fast development of genome editing recent years, epigenetic modification including targeted DNA methylation and demethylation also becomes attractive for its capability of stably regulate gene expression. In order to develop site-specific and efficient tools for DNA methylation targeting, we tethered artificial zinc finger protein recognizing specific DNA sequence to various RdDM proteins (ZF-RdDM) in *Arabidopsis*. With this tool, we studied the hierarchy of action within RdDM pathway by testing their ability to target methylation in different mutant backgrounds. Also, at thousands of ZF-RdDM off target sites, we characterized the ectopic siRNAs production, Pol V recruitment and DNA methylation establishment and found that simultaneously recruiting both arms of the RdDM pathway, siRNA biogenesis and Pol V recruitment, dramatically enhanced targeted methylation. We then also developed a tool to target DNA demethylation in plants by fusing the catalytic domain of the human demethylase TEN-ELEVEN TRANSLOCATION1 (TET1cd) and an artificial zinc finger protein or CRISPR/dCas9 system.

Finally, I will discuss DNA methylation landscape in human embryonic stem cells (hESCs). hESCs are morphologically and transcriptionally similar to stem cells derived from the

mouse post-implantation epiblast. Thus, hESCs are typically considered to exhibit 'primed' pluripotency. Various culture conditions have been developed to promote maintenance and self-renewal of hypomethylated 'naive' hESCs. We have discovered that reverting primed hESCs to naive hESCs results in a Stage Specific Embryonic Antigen 4 (SSEA4)-negative population with a transcriptional program resembling the human pre-implantation epiblast. However, we also discovered that the methylation landscape of naive hESCs *in vitro* is distinct from human epiblast *in vivo* with a lost 'memory' of methylation state at primary imprints and human oocyte.

The dissertation of Wanlu Liu is approved.

Sriram Kosuri

Jeffrey Aaron Long

Matteo Pellegrini

Wei Wang

Steven Erik Jacobsen, Committee Chair

University of California, Los Angeles

2018

I dedicated this dissertation to my lovely husband, Congxi, for standing by myside throughout my graduate school and life. This work is also dedicated to my parents, Yuhua Wu and Xijun Liu, who have always loved me unconditionally.

TABLE OF CONTENTS

FIGURES AND TABLES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

ACKNOWLEDGMENTS

Firstly, I would like to express my special thanks to my graduate advisor and wedding officiant, Professor Steve Jacobsen. A mentor is someone who allows you to see the hope inside yourself. Steve has been the one inspiring me the beauty of life and science throughout my graduate career. With his guidance, I am excited and motivated about science throughout my PhD. His advice on both research and my career have been priceless. Of the many valuable conversations Steve and I had, one of them particularly imprinted me till today. One day, I asked Steve what I should do after graduation, instead of telling me what to do, he asked me back what kind of life I would like to pursue. Inspired by his question, the word 'adventurous' appeared in my head. I hope to take this word with me as I advance to the next stage and I wish I can keep adventuring in both science and the real world. I would also like to thank other members of my committee, Professor Jeff Long, Professor Matteo Pellegrini, Professor Sriram Kosuri and Professor Wei Wang for their insightful comments and encouragement.

I would also like to thank my marvelous fellow lab mates in the Jacobsen lab. Without their discussions, help, collaborations, guidance and support, I can never finish this fantastic journey. Specially I would like to thank Javier Gallego-Bartolome for helping train me at bench, William Pastor for introducing me to the human embryonic stem cell world, Suhua Feng, Jake Harris, Peggy Kuo, Linda Yen, Magdalena Potok, Ashot Papikian, Zhenhui Zhong, Jason Gardiner, Basudev Ghoshal, Marco Morselli, Pop Wongpalee, Qikun Liu, Martin Groth, Sylvain Bischof, Michael Christopher, Ming Wang, Yan Xue, Jixian Zhai, Isreal Ausin and Haifeng Wang for simulating scientific discussion. I would also like to thank Mahnaz Akhavan and Ming Chan for all their support.

Last but not least, I am also more than grateful to my fabulous collaborators, Sascha Duttke and Jonathan Hetzel from UCSD, Di Chen and Professor Amander Clark from UCLA. I had tremendous luck to work closely with Amander. Her brilliance, inspiration, passion and courage for science all set an excellent role model for me as a female scientist.

Chapter 1:

Chapter One is a version of the research article published in *Nature Plants*, 4, 181-188 entitled "RNA-directed DNA methylation involves co-transcriptional small RNA-guided slicing of Pol V transcripts in Arabidopsis". This paper is authored by Wanlu Liu, Sascha H. Duttke, Jonathan Hetzel, Martin Groth, Suhua Feng, Javier Gallego-Bartolome, Zhenhui Zhong, Hsuan Yu Kuo, Zonghua Wang, Jixian Zhai, Joanne Chory and Steven E. Jacobsen. We thank members of the Jacobsen lab for insightful discussion and M. Akhavan for technical assistance. We also thank Life Science Editors for editing assistance. High throughput sequencing was performed at UCLA BSCRC BioSequencing Core Facility. W.L. is supported by Philip J. Whitcome Fellowship from the UCLA Molecular Biology Institute and a scholarship from the Chinese Scholarship Council. Z.Z. is supported by a scholarship from the Chinese Scholarship Council. Group of J.Z. is supported by the Thousand Talents Program for Young Scholars and by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2016ZT06S172). This work was supported by NIH grant GM60398 to S.E.J. and NIH grant R01GM094428 and R01GM52413 to J.C. S.E.J. and J.C. are Investigators of the Howard Hughes Medical Institute.

Chapter 2:

Chapter Two is the version of a manuscript under revision entitled "Co-targeting RNA Polymerases IV and V promotes efficient de novo DNA methylation in Arabidopsis". This paper is authored by Javier Gallego-Bartolome, Wanlu Liu, Hsuan Yu Kuo, Suhua Feng, C. Jake Harris, Soo-Young Park, Joanne Chory and Steven E. Jacobsen. We thank Linda Yen for plasmids, Dr. Qikun Liu for seeds, Dr. Jixian Zhai for bioinformatics advice, Ceejay Lee, Raksha Dutt, Octavio Zaragoza-Rodriguez, Nathan Cai, Francois Yap, and Breanna Buhay for technical support, and members of the Jacobsen lab for insightful discussion. J.G.B. was supported partially by Human Frontier Science Program Fellowship (LT000425/2012-L). W.L. is supported by Philip J. Whitcome Fellowship from the UCLA Molecular Biology Institute and a scholarship from the Chinese Scholarship Council. This work was supported by a Bill and Melinda Gates Foundation grant (OPP1125410) to S.E.J. S.E.J. and J.C. are Howard Hughes Medical Institute investigators.

Chapter 3:

Chapter Three is the version of a manuscript published in *Proceedings of the National Academy of Sciences of the United States of America* doi:10.1073/pnas.1716300115 entitled "Targeted DNA demethylation of the Arabidopsis genome using the human TET1 catalytic domain". This paper is authored by Javier Gallego-Bartolome, Jason Gardiner, Wanlu Liu, Ashot Papikian, Basudev Ghoshal, Hsuan Yu Kuo, Jenny Miao-Chi Zhao, David J. Segal and Steven E. Jacobsen. We thank Dr. Zachary Nimchuk for the pMOA vector, and Truman Do for technical support. High-throughput sequencing was performed in the University of California, Los Angeles Broad Stem Cell Research Center BioSequencing Core Facility. This research was supported by the Bill & Melinda Gates Foundation (OPP1125410). W.L. is supported by the Philip J. Whitcome

Fellowship from the University of California, Los Angeles Molecular Biology Institute and a scholarship from the Chinese Scholarship Council. D.J.S. is supported by NIH CA204563. S.E.J. is an Investigator of the Howard Hughes Medical Institute.

Chapter 4:

Chapter Four is the version of a manuscript published in *Cell Stem Cell 18:323-329* entitled "Naive Human Pluripotent Cells Feature a Methylation Landscape Devoid of Blastocyst or Germline Memory". This paper is authored by William A. Pastor, Di Chen, Wanlu Liu, Rachel Kim, Anna Sahakyan, Anastasia Lukianchikov, Kathrin Plath, Steven E. Jacobsen, and Amander T. Clark. We thank the UCLA Broad Stem Cell Research Center (BSCRC) Flow Cytometry core for flow and FACS assistance, the UCLA BSCRC High Throughput Sequencing Core, Sriharsa Pradhan from N.E.B. for donating anti-DNMT1 antibody, and Steven Peckman from the UCLA BSCRC for consenting couples for embryo donation for hESC derivation. We thank Colin Shew, Beatrice Sun, and Tiasha Shafiq for help with experiments and the Fall 2015 UCLA Biomedical Research Minor 5HB undergraduate class for useful discussions. W.A.P. was supported by the Jane Coffin Childs Memorial Fund for Medical Research and a UCLA BSCRC Postdoctoral Training Fellowship. D.C. is supported by a UCLA BSCRC Postdoctoral Training Fellowship. W.L. is supported by the Philip J. Whitcome Fellowship from the UCLA Molecular Biology Institute and a scholarship from the Chinese Scholarship Council. This work was supported by the NIH R01 HD079546 (A.T.C.), CIRM RB4-06133 (K.P.), and P01 GM099134 (K.P.). Funds for human embryo banking and derivation of new hESC lines were provided by the UCLA Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research. No federal grant funding was used for work with human embryo's or derivation of new hESC lines. No payment

was provided to embryo donors for their generous gift of surplus embryos to stem cell research. S.E.J. is an investigator of the Howard Hughes Medical Institute.

VITA

Education:

2013-present Graduate Student (Advisor: Steven E. Jacobsen)
University of California, Los Angeles, CA, U.S.A.

2013 Bachelor of Medicine, Zhejiang University, China

Published manscripts:

**Denoted equal contribution.*

1. Pastor, W.A., et al. (2018) *Nature cell biology 20.5 (2018): 553-564.*
2. **Liu, W.***, Duttke, S. H.*, et al. (2018) *Nature Plants*, 4.3 (2018): 181.
3. Gallego-Bartolome, J.* , Gardiner, J.*, **Liu, W.*,** Papikian, A.*, et al. *Proc. Nat. Acad. Sci. U. S. A.,* 115.9 (2018): E2125-E2134.
4. Zhang, Y.*, Harris, C. J.*, Liu Q.*, et al. *Proc. Nat. Acad. Sci. U. S. A.,* (2018): 201716300.
5. Chen, D. et al. *Biology of Reproduction* 97.6 (2017): 850-861.
6. Clark, A. T., et al. *Stem Cell Report* 9.1 (2017): 329-341.
7. Ausin, I., et al. *Proc. Nat. Acad. Sci. U. S. A.,* 113.50 (2016): E8106-E8113.
8. Li, S., et al. *Proc. Nat. Acad. Sci. U. S. A.,* 113.35 (2016): E5108-E5116.
9. Pastor, W. A.*, Chen, D.*, **Liu, W.*, et al.** *Cell stem cell* 18.3 (2016): 323-329.
10. Harris, C. J.*, Husmann, D.*, et al. *PLoS Genet*. 12.5 (2016): e1005998.

11. Zhai, J.*, Bischof, S.* et al. *Cell* 163.2 (2015): 445-455.

12. Pastor, W.A., et al. *Nature communications* 5 (2014): 5795.

13. Jiang, Y. et al. *Acta biomaterialia* 9, no. 9 (2013): 8089-8098.

14. Zhang, S. et al. *Stem cells and development* 22, no. 1 (2012): 90-101.

Previous research experience:

2011	Summer research scholar, Stowers Institute for Medical Research, MO, U.S.A.
2012	CSST program, University of California, Los Angeles, CA, U.S.A.
2013	Research intern, Stowers Institute for Medical Research, MO, U.S.A.

Selected Honors:

2013	China Scholarship Council Scholarship
2015	Philip J. Whitcome Fellowship
2017	Philip J. Whitcome Fellowship

INTRODUCTION

Epigenetics

The genetic information in eukaryotic cells is stored in DNA and packaged into nucleosomes which consist of a segment of DNA sequence wrapped around eight core histone proteins including H2A, H2B, H3 and H4 (Kornberg, 1974). Serving as the basic structure of chromatin, modifications on DNA and histones provide additional layer of genetic information involved in gene expression and DNA replication regulation. Such epigenetic modification is a reversible process and can be deposited and removed from the chromatin. While DNA methylation mostly refers to the addition of a methyl group to cytosine or adenine, histones can be post-translational modified in various ways such as methylation, acetylation, phosphorylation, ubiquitination and sumoylation (Shiio and Eisenman, 2003; Strahl and Allis, 2000). When located in gene promoter, DNA methylation commonly acts to repress gene expression. Histone modification can be associated with both actively transcription or repression depends on the location and type of modification occurs on the histone tail (Strahl and Allis, 2000). Those chromatin modifications are epigenetic marks that can be mitotically inherited and regulate various biological processes such as gene expression and DNA replication without altering the underlying base sequence, and thus referred as 'epigenetics'(Deichmann, 2016).

In this dissertation, DNA methylation exclusively refers to cytosine methylation. The genome wide landscape of DNA methylation and histone modifications are defined as epigenome in this dissertation. Epialleles refers to loci with same underline DNA sequence but different epigenome landscapes.

DNA methylation in plants and mammals

Methylation of cytosine at position 5 (5mC) with the presence of methyl-group donor SAM and DNA methyltransferase is known as DNA methylation. Molecular and genetics studies in mammals and plants have shown that DNA methylation is associated with various biological processes such as silencing of genes and transposable elements (TEs) (Law and Jacobsen, 2010). In addition, DNA methylation also play crucial roles in developmental processes such as genomic imprinting and X chromosome imprinting(Li and Zhang, 2014). Targeted mutation of the DNA methyltransferase in mouse embryonic stem cells lead to embryonic lethality suggesting its essential role for development of mammals(Li et al., 1992). Consistent with these important roles, aberrant DNA methylation has been shown to linked with human diseases including various cancers (Cooper and Youssoufian, 1988; Rideout et al., 1990).

DNA methylation in mammals predominantly occurs in the CG context with ~70-80% of CG dinucleotides methylated throughout the genome (Ehrlich et al., 1982; Li and Zhang, 2014). In plants, DNA methylation is mainly established in three different DNA sequence contexts including CG, CHG and CHH (where H represents any base except G) by DNA methyltransferase (Law and Jacobsen, 2010). CG and CHG sites are symmetrical cytosine sites while CHH sites are asymmetrical sites. Unlike the mammalian genome which is heavily methylated in CG context, the genome-wide DNA methylation levels in *Arabidopsis thalian* is on average lower, where the DNA methylation is approximately 24% for CG, 6.7% for CHG and 1.7% for CHH (Cokus et al., 2008). In mammalian genome, certain regions contain high occurrence of CG dinucleotide sequences which are defined as CpG islands, are dense clusters of methylation-free CG dinucleotide sites and are often found near gene promoters (Cedar and Bergman, 2009; Larsen et al., 1992; Suzuki and Bird, 2008). In contrast, CpG islands is not

observed in plants. DNA methylation is predominantly observed over the heterochromatic regions in *Arabidopsis thaliana* (Cokus et al., 2008; Lister et al., 2008). Since transposable elements (TEs) are highly enriched in heterochromatic regions, another distinct difference in DNA methylation pattern between mammals and plants is that TEs are highly methylated in all three contexts (CG, CHG, CHH). Mutants cause loss of DNA methylation in *Arabidopsis thaliana* leads to transcriptional reactivation of certain TEs across the genome indicating the essential role of DNA methylation in silencing TEs (Hirochika et al., 2000; Lister et al., 2008; Stroud et al., 2013).

Notably, DNA methylation is widely conserved in many organisms including some lower eukaryotes such as *Neurospora* and invertebrates (Aramayo and Selker, 2013; Elgin and Reuter, 2013). However, some well-studied model organisms such as *Drosophila melanogaster* and *Caenorhadbitis elegans* lack DNA methylation. One potential explanation is that some histone modifications may have replaced its roles in those organisms (Nanty et al., 2011).

De novo DNA methylation in plants and mammals

In mammals, DNA methylation patterns are established during embryonic development by the DNA methyltransferase 3 (DNMT3), a family of *de novo* methyltransferases including DNMT3A and DNMT3B. Both of them are responsible for the methylation pattern establishment during early embryo development (Okano et al., 1999). Followed by the of genome wide demethylation after fertilization, *de novo* DNA methylation occurs around implantation when the inner cell mass cells start to differentiate to ectoderm (Smallwood and Kelsey, 2012). Even though DNMT3A and DNMT3B are both required for *de novo* DNA methylation, their function during development is distinct in some way. For example, in germ cell, DNMT3A instead of

DNMT3B plays essential role in *de novo* methylation of most imprinted loci (Kaneda et al., 2004; Kato et al., 2007; Sasaki and Matsui, 2008; Smallwood and Kelsey, 2012). Another member in the DNMT3 family, DNMT3L have sequence similarity to the PHD and catalytic domains of DNMT3A and DNMT3B (Bourc'his et al., 2001). Although DNMT3L lacks the critical methyltransferase motifs and thus is catalytically inactive (Chen et al., 2005), it is also required for *de novo* methylation establishment at most imprinted loci (Bourc'his et al., 2001; Hata et al., 2002; Kaneda et al., 2004; Smallwood and Kelsey, 2012).

In plants, all context of DNA methylation is established through a small RNA guided process namely RNA-directed DNA methylation (RdDM) first described by Wassenegger et al. in 1994 (Wassenegger et al., 1994). Deriving from longer double-stranded RNAs (dsRNAs), the small interfering RNAs (siRNAs) are able to provide sequence-specific guides for DNA methylation while long non-coding RNAs help assembly of other factors involved in catalysis of DNA methylation (Wierzbicki et al., 2008). The production of these RNAs depends on two plant specific non-canonical DNA-dependent RNA polymerases evolved from RNA polymerase (Pol) II, known as Pol IV and Pol V (Matzke et al., 2015). The current model for RdDM involves several sequential steps including the upstream synthesis of 24-nucleotide (nt) siRNAs by Pol IV and downstream synthesis of non-coding transcripts by Pol V. To produce 24-nt siRNAs, Pol IV is recruited to heterochromatic regions via SAWADEE HOMEODOMAIN HOMOLOG 1 (SHH1) (Law et al., 2014) and transcribes 30-40-nucleotide (nt) single-stranded RNAs (Blevins et al., 2015; Li et al., 2015; Zhai et al., 2015). Those precursor Pol IV transcripts are then processed by RNA-dependent RNA polymerases 2 (RDR2) into double-stranded RNAs (Haag et al., 2012; Xie et al., 2004a). These dsRNAs are then primarily cleaved into 24-nt siRNAs by

DICER-LIKE3 (DCL3) (Blevins et al., 2015; Qi et al., 2005; Xie et al., 2004b; Zhai et al., 2015) and loaded into ARGONAUTE 4 (AGO4) (Li et al., 2006; Qi et al., 2006; Zilberman et al., 2003). The production of a second set of non-coding RNAs by Pol V is coupled with DDR complex. The DDR complex is consist of the CLSY1-related chromatin remodeler DEFECTIVE IN RNA-DIRECTED DNA METHYLATION 1(DRD1) (Kanno et al., 2004) , a structural maintenance of chromosomes solo hinge protein DEFECTIVE IN MERISTEM SILENCING 3(DMS3) (Ausin et al., 2009; Kanno et al., 2008) and RNA-DIRECTED DNA METHYLATION 1 (RDM1) (Gao et al., 2010) which is a plant-specific protein that may involves in multiple part of the RdDM pathway. AGO4-associated 24-nt siRNAs is thought to base-pair with the nascent Pol V transcript and recruit DOMAIN REARRANGED METHYLTRANSFERASE 2 (DRM2), a homolog of the mammalian DNMT3 methyltransferase, to catalyze *de novo* methylation. During the base-pairing of 24-nt siRNAs and nascent Pol V transcripts, results from Chapter 1 indicate nascent Pol V transcripts is co-transcriptional sliced by AGO4/6/9 (Liu et al., 2018). In addition, this co-transcriptional slicing of Pol V transcripts also requires KOW DOMAIN-CONTAINING TRANSCRIOTPION FACTOR 1 (KTF1; also known as SPT5L) (Huang et al., 2009; Liu et al., 2018). The SU(VAR)3-9 histone methyltransferase (SUVH) family homologues SUVH2 and SUVH9 may act in the downstream steps of RdDM. Through their SRA (SET and RING-associated) domain, SUVH2 and SUVH9 are able to bind DNA methylation thus recruit Pol V to pre-existing methylation (Jing et al., 2016; Johnson et al., 2014; 2008; Kuhlmann and Mette, 2012; Liu et al., 2014).

DNA methylation maintenance

CG methylation is maintained by DNMT1during mitosis in mammals. UHRF1 (ubiquitin-like plant homeodomain and RING finger domains 1), a SRA domain containing protein binds to hemimethylated CG dinucleotide(Bostick et al., 2007). Interacting with UHRF1, DNMT1 localize to replication fork and act on hemimethylated DNA generated during DNA replication to restore them to fully methylated state(Kim et al., 2009). Another chromatin-remodeling factor LSH1 (lymphoid-specific helicase 1, also known as HELLS) is also required for the maintenance of CG methylation even though the mechanism is still unknown (Dennis et al., 2001; Huang et al., 2004).

In plants, CG methylation is maintained in a similar manner as in mammals requiring METHYLTRANSFERASE 1 (MET1, the DNMT1 homolog(Vongs et al., 1993)), VARIANT IN METHYLATION family of SRA domain proteins (VIM, the UHRF1 homolog (Woo et al., 2008; 2007)) and DECREASED IN DNA METHYLATION 1(DDM1, homolog of LSH (Vongs et al., 1993)). Maintenance of CHG methylation in *Arabidopsis thaliana* is thought to be a model of self-reinforcing loop between histone and DNA methylation (Ebbs and Bender, 2006; Inagaki et al., 2010; Johnson et al., 2007). CHG methylation is largely maintained by DNA methyltransferase CHROMOMETHYLASE 3 (CMT3) (Stroud et al., 2014). CHG methylation is then recognized SUVH4 (also known as KYP) histone methyltransferase which catalyze histone 3 lysine 9 dimethylation (H3K9me2) which is required for the maintenance of CHG methylation (Enke et al., 2011; Law and Jacobsen, 2010). On the other hand, asymmetric CHH methylation must be continually re-established by the action of DRM2 and CHROMOMETHYLASE 2 (CMT2) (Stroud et al., 2014; Zemach et al., 2013).

DNA demethylation in plants and mammals

DNA methylation is a dynamic modification which can be removed in both passive and active manners. Passive DNA demethylation happens via the absence of DNA methylation maintenance during replication while active DNA demethylation removes 5mC via an enzymatic process (Kohli and Zhang, 2013). In mammals, active 5mC is removed by Ten-eleven translocation (TET) family enzymes including TET1, TET2 and TET3 (Ito et al., 2011). TET proteins oxidize 5mC to 5-hydroxymethylcytosine (5hmC) which is further oxidized into 5-fomylcytosine and 5-carboxylcytosine (Ito et al., 2011). In plants, active DNA demethylation involves a DNA glycosylase REPERSSOR OF SILENCING 1 (ROS1) through a base excision repair pathway (Zhang and Zhu, 2012).

Epigenome engineering

With the fast development of genome editing recent years, epigenetic modification also become attractive for its capability of stable gene expression silencing by epigenetic mechanism. Epigenome editing is defined as the targeted alteration of chromatin marks at specific genomic loci to affect gene expression without changing the DNA sequence (Kungulovski and Jeltsch, 2016). Epigenome editing tools usually include two parts. The function of first part comprises designed DNA recognition domains (artificial zinc finger proteins (ZFs), transcription activator-like effectors (TALEs) or deactivated CRISPR/Cas9 complex) to bind specific DNA sequence (Kungulovski and Jeltsch, 2016). The second part usually is EpiEffectors including chromatin-modifying enzyme such as DNA methyltransferase, histone modification enzyme or transcription activator. Epigenome editing is a promising approach for stable gene regulation, with many

potential applications in both basic research as well as therapeutic implications.

Severing as a model for epigenetic targeting, several tandem arrays zinc finger proteins can bind to specific DNA sequence through the recognition of ~3 nucleotides of DNA by each zinc finger (Bhakta and Segal, 2010; Pavletich and Pabo, 1991; Segal et al., 2003; Wolfe et al., 2000). The first attempt at DNA methylation targeting was conducted by fusing the prokaryotic DNA methyltransferase M. SssI to the ZF Zif268 to silence specific genes(Xu and Bestor, 1997). Using bacterial DNA methyltransferase, following up studies confirmed that DNA methylation could be effectively targeted in yeast (Carvin et al., 2003a; 2003b). Study in human cells also demonstrated that DNA methylation targeting was able to repress the expression of targeted genes(Li et al., 2007). The *Arabidopsis FWA* gene was initially identified from late-flowering epiallelic mutants *fwa* (Kinoshita et al., 2004). The late-flowering phenotype is associated with the overexpression of *FWA* gene caused by the heritable hypomethylation of two tandem repeats around *FWA* gene transcription starting sites(Kinoshita et al., 2007; Soppe et al., 2000). Recent work demonstrated that by fusing SUVH9 to an artificial zinc finger protein designed to recognize the 18-nt sequence inside the two small repeats (CGGAAAGATGTATGGGCT) over the *FWA* promoter is able to restore CG, CHG and CHH methylation in *fwa* mutant (Johnson et al., 2014). Accordingly, the abnormal overexpression of *FWA* is repressed and the late flowering phenotype was reverted (Johnson et al., 2014).

Identified in the plant pathogen *Xanthomonas*, TALEs has been shown to bind DNA sequence with high activity and specificity (Mussolino et al., 2014). The DNA-binding domain (DBD) of TALEs contains a module each comprised a highly conserved 34 amino acid, among which the

identity of the amino acids in position 12 and 13 (RVD or repeat variable di-residues) determines specific binding to a certain DNA base pair (Boch and Bonas, 2010; Moscou and Bogdanove, 2009). Recent work demonstrate that fusing TALE to the catalytic domain of DNMT3A is capable to induce DNA methylation at target sites (Bernstein et al., 2015). However, due to its sensitivity to 5mC in their DBD, TALEs may be limited in developing tools for DNA methylation or demethylation targeting (Valton et al., 2012).

CRISPR (clustered regularly interspaced short palindromic repeats) system derived from *Streptococcus pyogenes* has recently been developed for genome editing in a broad range of organisms. As an immune mechanism against phages in bacteria and archaea, the CRISPR/Cas9 system can specifically recognize DNA sequence determined by a short small guide RNA (sgRNA/gRNA) and cut DNA sequence through the endogenous endonuclease activity of Cas9(Garneau et al., 2010). By introducing point mutations in the endonuclease domains of Cas9, CRISPR/deactivated Cas9 (dCas9) can be used as RNA-guided DNA-binding protein (Qi et al., 2013). With the simplicity of the CRISPR/Cas9 system, scientists have also fused nuclease-deactivated Cas9 (dCas9) to various effector proteins to modulate gene expression and enable epigenome editing in mammalian system including both methylation and demethylation targeting (Hilton et al., 2015; Liu et al., 2016; Thakore et al., 2015).

REFERENCES

Aramayo, R., and Selker, E.U. (2013). Neurospora crassa, a model system for epigenetics research. Cold Spring Harb Perspect Biol *5*, a017921–a017921.

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

Bernstein, D.L., Le Lay, J.E., Ruano, E.G., and Kaestner, K.H. (2015). TALE-mediated epigenetic suppression of CDKN2A increases replication in human fibroblasts. J. Clin. Invest. *125*, 1998–2006.

Bhakta, M.S., and Segal, D.J. (2010). The generation of zinc finger proteins by modular assembly. Methods Mol. Biol. *649*, 3–30.

Blevins, T., Podicheti, R., Mishra, V., Marasco, M., Tang, H., and Pikaard, C.S. (2015). Identification of Pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding de novo DNA methylation in Arabidopsis. Elife *4*, e09591.

Boch, J., and Bonas, U. (2010). Xanthomonas AvrBs3 family-type III effectors: discovery and function. Annu Rev Phytopathol *48*, 419–436.

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.

Bourc'his, D., Xu, G.L., Lin, C.S., Bollman, B., and Bestor, T.H. (2001). Dnmt3L and the establishment of maternal genomic imprints. Science *294*, 2536–2539.

Carvin, C.D., Dhasarathy, A., Friesenhahn, L.B., Jessen, W.J., and Kladde, M.P. (2003a). Targeted cytosine methylation for in vivo detection of protein-DNA interactions. Pnas *100*, 7743–7748.

Carvin, C.D., Parr, R.D., and Kladde, M.P. (2003b). Site-selective in vivo targeting of cytosine-5 DNA methylation by zinc-finger proteins. Nucl. Acids Res. *31*, 6493–6501.

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

Chen, Z.-X., Mann, J.R., Hsieh, C.-L., Riggs, A.D., and Chédin, F. (2005). Physical and functional interactions between the human DNMT3L protein and members of the de novo methyltransferase family. J. Cell. Biochem. *95*, 902–917.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215–219.

Cooper, D.N., and Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. Hum. Genet. *78*, 151–155.

Deichmann, U. (2016). Epigenetics: The origins and evolution of a fashionable topic. Dev. Biol. *416*, 249–254.

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.

Ebbs, M.L., and Bender, J. (2006). Locus-specific control of DNA methylation by the Arabidopsis SUVH5 histone methyltransferase. The Plant Cell *18*, 1166–1176.

Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., McCune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucl. Acids Res. *10*, 2709–2721.

Elgin, S.C.R., and Reuter, G. (2013). Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. Cold Spring Harb Perspect Biol *5*, a017780–a017780.

Enke, R.A., Dong, Z., and Bender, J. (2011). Small RNAs prevent transcription-coupled loss of histone H3 lysine 9 methylation in Arabidopsis thaliana. PLoS Genet. *7*, e1002350.

Gao, Z., Liu, H.-L., Daxinger, L., Pontes, O., He, X., Qian, W., Lin, H., Xie, M., Lorković, Z.J., Zhang, S., et al. (2010). An RNA polymerase II- and AGO4-associated protein acts in RNA-directed DNA methylation. Nature *465*, 106–109.

Garneau, J.E., Dupuis, M.-È., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67–71.

Haag, J.R., Ream, T.S., Marasco, M., Nicora, C.D., Norbeck, A.D., Pasa-Tolic, L., and Pikaard, C.S. (2012). In vitro transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. Molecular Cell *48*, 811–818.

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

Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nature Biotechnology 2016 34:10 *33*, 510–517.

Hirochika, H., Okamoto, H., and Kakutani, T. (2000). Silencing of retrotransposons in arabidopsis and reactivation by the ddm1 mutation. The Plant Cell *12*, 357–369.

Huang, J., Fan, T., Yan, Q., Zhu, H., Fox, S., Issaq, H.J., Best, L., Gangi, L., Munroe, D., and Muegge, K. (2004). Lsh, an epigenetic guardian of repetitive elements. Nucl. Acids Res. *32*, 5019–5028.

Huang, L., Jones, A.M.E., Searle, I., Patel, K., Vogler, H., Hubner, N.C., and Baulcombe, D.C. (2009). An atypical RNA polymerase involved in RNA silencing shares small subunits with RNA polymerase II. Nat. Struct. Mol. Biol. *16*, 91–93.

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.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science *333*, 1300–1303.

Jing, Y., Sun, H., Yuan, W., Wang, Y., Li, Q., Liu, Y., Li, Y., and Qian, W. (2016). SUVH2 and SUVH9 Couple Two Essential Steps for Transcriptional Gene Silencing in Arabidopsis. Mol Plant *9*, 1156–1167.

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. Current Biology *17*, 379–384.

Johnson, L.M., Du, J., Hale, C.J., Bischof, S., Feng, S., Chodavarapu, R.K., Zhong, X., Marson, G., Pellegrini, M., Segal, D.J., et al. (2014). SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. Nature *507*, 124–128.

Johnson, L.M., Law, J.A., Khattar, A., Henderson, I.R., and Jacobsen, S.E. (2008). SRA-domain proteins required for DRM2-mediated de novo DNA methylation. PLoS Genet. *4*, e1000280.

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900–903.

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

Kanno, T., Mette, M.F., Kreil, D.P., Aufsatz, W., Matzke, M., and Matzke, A.J.M. (2004). Involvement of putative SNF2 chromatin remodeling protein DRD1 in RNA-directed DNA methylation. Current Biology *14*, 801–805.

Kato, Y., Kaneda, M., Hata, K., Kumaki, K., Hisano, M., Kohara, Y., Okano, M., Li, E., Nozaki, M., and Sasaki, H. (2007). Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. Hum. Mol. Genet. *16*, 2272–2280.

Kim, J.K., Samaranayake, M., and Pradhan, S. (2009). Epigenetic mechanisms in mammals. Cell. Mol. Life Sci. *66*, 596–612.

Kinoshita, T., Miura, A., Choi, Y., Kinoshita, Y., Cao, X., Jacobsen, S.E., Fischer, R.L., and Kakutani, T. (2004). One-way control of FWA imprinting in Arabidopsis endosperm by DNA methylation. Science *303*, 521–523.

Kinoshita, Y., Saze, H., Kinoshita, T., Miura, A., Soppe, W.J.J., Koornneef, M., and Kakutani, T. (2007). Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats. The Plant Journal *49*, 38–45.

Kohli, R.M., and Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. Nature *502*, 472–479.

Kornberg, R.D. (1974). Chromatin structure: a repeating unit of histones and DNA. Science *184*, 868–871.

Kuhlmann, M., and Mette, M.F. (2012). Developmentally non-redundant SET domain proteins SUVH2 and SUVH9 are required for transcriptional gene silencing in Arabidopsis thaliana. Plant Mol. Biol. *79*, 623–633.

Kungulovski, G., and Jeltsch, A. (2016). Epigenome Editing: State of the Art, Concepts, and Perspectives. Trends Genet. *32*, 101–113.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. Genomics *13*, 1095–1107.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics *11*, 204–220.

Law, J.A., Du, J., Hale, C.J., Feng, S., Krajewski, K., Palanca, A.M.S., Brian, S., Patel, D.J., and Jacobsen, S.E. (2014). 126_SHH1Nature. Nature *498*, 385–389.

Li, C.F., Pontes, O., El-Shami, M., Henderson, I.R., Bernatavichute, Y.V., Chan, S.W.-L., Lagrange, T., Pikaard, C.S., and Jacobsen, S.E. (2006). An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in Arabidopsis thaliana. Cell *126*, 93–106.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915–926.

Li, E., and Zhang, Y. (2014). DNA Methylation in Mammals. Cold Spring Harb Perspect Biol *6*, a019133–a019133.

Li, F., Papworth, M., Minczuk, M., Rohde, C., Zhang, Y., Ragozin, S., and Jeltsch, A. (2007). Chimeric DNA methyltransferases target DNA methylation to specific DNA sequences and repress expression of target genes. Nucl. Acids Res. *35*, 100–112.

Li, S., Vandivier, L.E., Tu, B., Gao, L., Won, S.Y., Li, S., Zheng, B., Gregory, B.D., and Chen, X. (2015). Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in Arabidopsis reveals features and regulation of siRNA biogenesis. Genome Res. *25*, 235–245.

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.

Liu, W., Duttke, S.H., Hetzel, J., Groth, M., Feng, S., Gallego-Bartolome, J., Zhong, Z., Kuo, H.Y., Wang, Z., Zhai, J., et al. (2018). RNA-directed DNA methylation involves co-transcriptional small-RNA-guided slicing of polymerase V transcripts in Arabidopsis. Nat Plants *4*, 181–188.

Liu, X.S., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R.A., and Jaenisch, R. (2016). Editing DNA Methylation in the Mammalian Genome. Cell *167*, 233–247.e17.

Liu, Z.-W., Shao, C.-R., Zhang, C.-J., Zhou, J.-X., Zhang, S.-W., Li, L., Chen, S., Huang, H.-W., Cai, T., and He, X.-J. (2014). The SET domain proteins SUVH2 and SUVH9 are required for Pol V occupancy at RNA-directed DNA methylation loci. PLoS Genet. *10*, e1003948.

Matzke, M.A., Kanno, T., and Matzke, A.J.M. (2015). RNA-Directed DNA Methylation: The Evolution of a Complex Epigenetic Pathway in Flowering Plants. Annu Rev Plant Biol *66*, 243–267.

Moscou, M.J., and Bogdanove, A.J. (2009). A simple cipher governs DNA recognition by TAL effectors. Science *326*, 1501–1501.

Mussolino, C., Alzubi, J., Fine, E.J., Morbitzer, R., Cradick, T.J., Lahaye, T., Bao, G., and Cathomen, T. (2014). TALENs facilitate targeted genome editing in human cells with high specificity and low cytotoxicity. Nucl. Acids Res. *42*, 6762–6773.

Nanty, L., Carbajosa, G., Heap, G.A., Ratnieks, F., van Heel, D.A., Down, T.A., and Rakyan, V.K. (2011). Comparative methylomics reveals gene-body H3K36me3 in Drosophila predicts DNA methylation and CpG landscapes in other invertebrates. Genome Res. *21*, 1841–1850.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257.

Pavletich, N.P., and Pabo, C.O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252*, 809–817.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173–1183.

Qi, Y., Denli, A.M., and Hannon, G.J. (2005). Biochemical specialization within Arabidopsis RNA silencing pathways. Molecular Cell *19*, 421–428.

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

Rideout, W.M., Coetzee, G.A., Olumi, A.F., and Jones, P.A. (1990). 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. Science *249*, 1288–1290.

Sasaki, H., and Matsui, Y. (2008). Epigenetic events in mammalian germ-cell development: reprogramming and beyond. Nature Reviews Genetics *9*, 129–140.

Segal, D.J., Beerli, R.R., Blancafort, P., Dreier, B., Effertz, K., Huber, A., Koksch, B., Lund, C.V., Magnenat, L., Valente, D., et al. (2003). Evaluation of a modular strategy for the construction of novel polydactyl zinc finger DNA-binding proteins. Biochemistry *42*, 2137–2148.

Shiio, Y., and Eisenman, R.N. (2003). Histone sumoylation is associated with transcriptional repression. Pnas *100*, 13225–13230.

Smallwood, S.A., and Kelsey, G. (2012). De novo DNA methylation: a germ cell perspective. Trends Genet. *28*, 33–42.

Soppe, W.J.J., Jacobsen, S.E., Alonso-Blanco, C., Jackson, J.P., Kakutani, T., Koornneef, M., and Peeters, A.J.M. (2000). The Late Flowering Phenotype of fwa Mutants Is Caused by Gain-of-Function Epigenetic Alleles of a Homeodomain Gene. Molecular Cell *6*, 791–802.

Strahl, B.D., and Allis, C.D. (2000). The language of covalent histone modifications. Nature *403*, 41–45.

Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D.J., and Jacobsen, S.E. (2014). Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat. Struct. Mol. Biol. *21*, 64–72.

Stroud, H., Greenberg, M.V.C., Feng, S., Bernatavichute, Y.V., and Jacobsen, S.E. (2013). 122_ComplexRegulation. Cell *152*, 352–364.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nature Reviews Genetics *9*, 465–476.

Thakore, P.I., D'Ippolito, A.M., Song, L., Safi, A., Shivakumar, N.K., Kabadi, A.M., Reddy, T.E., Crawford, G.E., and Gersbach, C.A. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nat. Methods *12*, 1143–1149.

Valton, J., Dupuy, A., Daboussi, F., Thomas, S., Maréchal, A., Macmaster, R., Melliand, K., Juillerat, A., and Duchateau, P. (2012). Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. J. Biol. Chem. *287*, 38427–38432.

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

Wassenegger, M., Heimes, S., Riedel, L., and Sänger, H.L. (1994). RNA-directed de novo methylation of genomic sequences in plants. Cell *76*, 567–576.

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.

Wolfe, S.A., Nekludova, L., and Pabo, C.O. (2000). DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct *29*, 183–212.

Woo, H.R., Dittmer, T.A., and Richards, E.J. (2008). Three SRA-domain methylcytosine-binding proteins cooperate to maintain global CpG methylation and epigenetic silencing in Arabidopsis. PLoS Genet. *4*, e1000156.

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.

Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D., Jacobsen, S.E., and Carrington, J.C. (2004a). Genetic and functional diversification of small RNA pathways in plants. PLoS Biol. *2*, E104.

Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D., Jacobsen, S.E., and Carrington, J.C. (2004b). Genetic and functional diversification of small RNA pathways in plants. PLoS Biol. *2*, E104.

Xu, G.-L., and Bestor, T.H. (1997). Cytosine methylation targetted to pre-determined sequences. Nature Genetics *17*, 376–378.

Zemach, A., Kim, M.Y., Hsieh, P.-H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L., and Zilberman, D. (2013). The Arabidopsis Nucleosome Remodeler DDM1 Allows DNA Methyltransferases to Access H1-Containing Heterochromatin. Cell *153*, 193–205.

Zhai, J., Bischof, S., Wang, H., Feng, S., Lee, T.-F., Teng, C., Chen, X., Park, S.Y., Liu, L., Gallego-Bartolome, J., et al. (2015). A One Precursor One siRNA Model for Pol IV-Dependent siRNA Biogenesis. Cell *163*, 445–455.

Zhang, H., and Zhu, J.K. (2012). Active DNA Demethylation in Plants and Animals. Cold Spring Harb Symp Quant Biol *77*, 161–173.

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

CHAPTER 1

RNA-directed DNA methylation involves co-transcriptional small-RNA-guided slicing of polymerase V transcripts in *Arabidopsis*

Contributions

W.L., J.H., S.H.C.D., and S.F. performed GRO-seq experiments. M.G. performed ChIP-seq experiments. W.L. and Z.Z. performed small RNA-seq experiments. W.L. performed all bioinformatics analyses. W.L. and S.E.J. wrote the manuscript.

ABSTRACT

Small RNAs regulate chromatin modifications such as DNA methylation and gene silencing across eukaryotic genomes. In plants, RNA-directed DNA methylation (RdDM) requires 24-nucleotide small RNAs (siRNAs) that bind ARGONAUTE4 (AGO4) and target genomic regions for silencing. RdDM also requires non-coding RNAs transcribed by RNA polymerase V (Pol V) that probably serve as scaffolds for binding of AGO4-siRNA complexes. Here we used a modified global nuclear run-on (GRO) protocol followed by deep sequencing to capture Pol V nascent transcripts genome-wide. We uncovered unique characteristics of Pol V RNAs, including a uracil (U) common at position 10. This uracil was complementary to the 5' adenine found in many AGO4-bound 24-nucleotide siRNAs and was eliminated in a siRNA-deficient mutant as well as in the *ago4/6/9* triple mutant, suggesting that the +10U signature is due to siRNA-mediated co-transcriptional slicing of Pol V transcripts. Expression of wild-type AGO4 in *ago4/6/9* mutants was able to restore slicing of Pol V transcripts, but a catalytically inactive AGO4 mutant did not correct the slicing defect. We also found that Pol V transcript slicing required SUPPRESSOR OF TY INSERTION 5-LIKE (SPT5L), an elongation factor whose function is not well understood. These results highlight the importance of Pol V transcript slicing in RNA-mediated transcriptional gene silencing, which is a conserved process in many eukaryotes.

INTRODUCTION

DNA methylation is an evolutionarily conserved epigenetic mark associated with gene silencing that plays a key role in diverse biological processes. In plants, DNA methylation is mediated by small RNAs that target specific genomic DNA sequences in a process known as RNA-directed DNA methylation (RdDM). RdDM involves RNA polymerase (Pol) IV and Pol V, both of which evolved from Pol II, and plays crucial roles in transposon silencing and maintenance of genome integrity (Law and Jacobsen, 2010). The current model for RdDM involves several sequential steps. First, Pol IV initiates the biogenesis of siRNAs by producing 30- to 40-nt ssRNA (Blevins et al., 2015; Li et al., 2015; Zhai et al., 2015). These ssRNAs are then made double stranded by RNA-dependent RNA polymerase 2 (RDR2) (Haag et al., 2012; Xie et al., 2004), processed into 24-nt siRNA by DCL3 (Qi et al., 2005), and loaded into the effector protein AGO4 (Li et al., 2006; Qi et al., 2006; Zilberman et al., 2003). A second set of non-coding transcripts, generated by Pol V, has been proposed to serve as a targeting scaffold for the binding of AGO4-associated siRNAs through sequence complementarity (Wierzbicki et al., 2008). Ultimately, AGO4 targeting recruits the DRM2 DNA methyltransferase to mediate *de novo* methylation of cytosines in all sequence contexts (CG, CHG, and CHH, where H represents A, C, or T) (Zhong et al., 2014). Pol V is required for DNA methylation and silencing, and has been shown to be transcriptionally active *in vitro*. A recent study of RNAs co-immunoprecipitation (RIP) with Pol V showed Pol V-associated RNAs at thousands of locations in the genome (Böhmdorfer et al., 2016). However, shearing was used in the library preparation protocol, which meant that many features of the individual Pol V transcripts were lost (Böhmdorfer et al., 2016). Thus, several characteristics of Pol V transcripts and how they mediate RdDM remain poorly characterized (Wierzbicki et al., 2008; 2009).

RESULTS

Identification of nascent Pol V transcripts genome-wide.

To enable a detailed analysis of Pol V transcripts at single nucleotide resolution, we used a modified global nuclear run-on assay (Core et al., 2008; Hetzel et al., 2016) followed by deep sequencing (GRO-seq) in *Arabidopsis* (Figure 1-1A). This technique captures nascent RNA from engaged RNA polymerases in a strand specific manner. Uniquely mapping paired end reads were obtained from two independent experiments (Figure 1-2A) prepared from wild-type Columbia (Col-0) plants (Table 1-1). GRO-seq captures transcriptionally engaged RNA polymerases (Core et al., 2008; Hetzel et al., 2016), and although we selected against full length capped Pol II transcripts (Figure 1-1A), we still observed a background level of signal over Pol II transcribed protein-coding genes. Thus, in order to specifically identify Pol V-dependent nascent transcripts, we also performed GRO-seq in a Pol V mutant (*nrpe1*) as well as in a Pol IV/Pol V double mutant (*nrpd1/e1*). We coupled this with a genome-wide map of the chromatin association profile of Pol V, using ChIP-seq with an endogenous antibody against NRPE1, the largest catalytic subunit of Pol V. Combining Pol V ChIP-seq and GRO-seq in Col-0, *nrpe1*, and *nrpd1/e1*, we identified GRO-seq reads that mapped to Pol V regions, including those at previously defined individual Pol V intergenic non-coding (IGN) transcripts (Wierzbicki et al., 2008)(Figure 1-1B). As expected, we found that GRO-seq signals generated from Pol V occupied regions were largely eliminated in the *nrpe1* mutant, while signals over mRNA regions in the *nrpe1* mutant remained unchanged (Figure 1-2B,C), confirming that we had indeed identified Pol V-dependent nascent transcripts. In addition to the tight spatial co-localization of Pol V ChIP-seq and GRO-seq signals, we also observed a positive correlation between the two in signal intensity (Figure 1-2D). However, Pol V-dependent GRO-seq signals were much more

narrowly defined compared to signals from Pol V ChIP-seq, thereby providing a higher resolution view of Pol V transcription (Figure 1-1C). Unlike Pol II transcripts, which are primarily transcribed from one strand (Figure 1-1B, Figure 1-3A), Pol V-dependent transcripts were present roughly equally on both strands (Figure 1-1B, Figure 1-3B). RdDM has been shown to be enriched at short transposons as well as at the edges of long transposons (Zemach et al., 2013). Consistent with Pol V occupancy at long transposon edges (Zhong et al., 2012), we found that Pol V-dependent GRO-seq transcripts were also preferentially localized over those regions (Figure 1-3C, Figure 1-2E).

To investigate the relationship between Pol IV activity and Pol V transcript production, we performed Pol V ChIP-seq and GRO-seq in the *nrpd1* mutant, which specifically eliminates Pol IV activity. Although many Pol V transcripts were eliminated in the *nrpd1* mutant (Figure 1-4A), most remained (Figure 1-4B). Based on whether or not Pol V ChIP-seq signal remained in *nrpd1*, we classified Pol V regions into Pol IV/V-codependent regions (1,903 sites) or Pol IV-independent Pol V regions (2,365 sites). As expected, both the GRO-seq signal and the Pol V ChIP-seq signal were largely eliminated in *nrpd1* at Pol IV/V-codependent sites, while the signals at Pol IV-independent sites largely remained (Figure 1-4C,D).

The reason that some Pol V transcripts are dependent on Pol IV activity is likely because the RdDM pathway is a self-reinforcing loop (Law and Jacobsen, 2010). For example, although Pol V is required for DNA methylation and silencing, Pol V recruitment to chromatin requires preexisting DNA methylation via the methyl DNA binding proteins SUVH2 and SUVH9 (Johnson et al., 2014). We therefore hypothesized that the reason that Pol IV is required for Pol

V activity at only some genomic sites is because it plays a larger role in DNA methylation maintenance at this subset of sites. To test this, we analyzed cytosine methylation levels as well as 24-nt siRNAs abundance at both the Pol IV/V-codependent and Pol IV-independent sites. If Pol IV actively maintains DNA methylation at specific genomic sites to enable Pol V recruitment and transcription, then loss of Pol IV should have a more dramatic effect on the methylation levels at these sites. Indeed, Pol IV/V-codependent sites showed significantly higher 24-nt siRNAs levels as well as substantial reductions of all types of cytosine methylation in *nrpd1*, while Pol IV-independent sites showed fewer 24-nt siRNAs and less reduction in DNA methylation (Figure 1-4E,F). This is likely because the other DNA methylation maintenance pathways involving MET1, CMT3, and CMT2 are active at these loci, and compensate for the loss of methylation in the Pol IV mutant. In summary, these results show that even though Pol IV and Pol V work closely together in the RdDM pathway, Pol V can transcribe independently of Pol IV at many sites in the genome. Previous studies of Pol IV transcripts have shown them to be exceedingly rare in wild type because of their efficient processing into siRNAs by DICER enzymes (Blevins et al., 2015; Li et al., 2015; Zhai et al., 2015). However, it remains possible that trace levels of Pol IV transcripts could be present in our GRO-seq libraries. Thus, in order to uniquely focus on the characteristics of Pol V transcripts without any complication of the presence of small amounts of Pol IV transcripts, we focused our remaining analysis on Pol IV-independent Pol V regions.

Pol V transcripts show evidence of small RNA dependent slicing.

Because our GRO-seq method did not include the fragmentation step typical of traditional GRO-seq (Core et al., 2008), it was possible to estimate the length of Pol V nascent transcripts and

assess their 5' nucleotide composition. We observed a range of read lengths from 30- to 90-nt long with a peak at around 50-nt, and detected very few reads longer than about 120-nt (Figure 1-5A). Nascent Pol V transcripts observed in *nrpd1* GRO-seq showed a similar size distribution (Figure 1-6A). GRO-seq involves an *in vitro* nuclear run-on step in which the reaction is limited by time and nucleotide concentration, meaning that the run-on is unlikely to proceed to the natural 3' end of the transcript. Thus, the average Pol V transcript length measured here is likely an underestimate of the true length of Pol V transcripts *in vivo*. Using Pol V RIP-seq, Bohmdorfer et al. recently estimated the median Pol V transcript length to be around 200 nucleotides. However, since a fragmentation step was included in their RIP protocol, this was also an estimation (Böhmdorfer et al., 2016). Nevertheless, Pol V transcripts are clearly at least 50-nt long on average, which is significantly longer than Pol IV transcripts, which have been estimated to be around 30- to 40-nt long (Blevins et al., 2015; Zhai et al., 2015).

Eukaryotic and bacterial RNA polymerases preferentially initiate transcription at purines (A or G), commonly with a pyrimidine (C or T) present at the -1 position with respect to the transcription start site(Blevins et al., 2015; Li et al., 2015; Smale and Kadonaga, 2003; Sollner-Webb and Reeder, 1979; Zecherle et al., 1996; Zhai et al., 2015). However, instead of this expected enrichment at Pol V transcript 5' ends, we observed a strong U preference (on average 53.41%) at nucleotide +10 across six Col-0 biological replicates (Figure 1-5B, Figure 1-6B). This characteristic was unlikely to be an artifact of the GRO-seq procedure since no such preference was observed in transcripts that mapped to mRNA regions (Figure 1-6C,D). In order to test whether the +10U signature was specific to nascent RNAs with certain lengths, we examined the nucleotide preferences within different size ranges. We found a +10U signature in

all size ranges tested from 30-nt RNAs to RNAs longer than 70-nt, with the strongest signature in 40- to 50-nt long reads (Figure 1-6E-I).

In *Arabidopsis*, AGO4 shows slicer activity *in vitro* and interacts directly with Pol V (El-Shami et al., 2007; Qi et al., 2006). In addition, AGO4-associated 24-nt siRNAs are highly enriched for 5' adenines (Havecker et al., 2010; Mi et al., 2008). Therefore, we hypothesized that the 5' end of Pol V transcripts is often defined by an AGO4 slicing event, and that the U at position 10 in Pol V transcripts corresponds to a 5' A in AGO4 24-nt siRNAs (Figure 1-5C). We plotted the sequence composition of previously published AGO4-associated 24-nt siRNAs (Wang et al., 2011) that mapped to our identified Pol V transcript sites and observed a strong 5' enrichment for A (80.53%) (Figure 1-5D). If Pol V transcripts are sliced at 10-nt from the AGO4-siRNAs 5' end, we should detect sense-antisense siRNA-Pol V transcript pairs separated by 10-nt and a corresponding 10-nt of complementary sequence (Figure 1-5C). We plotted the distance between each AGO4-siRNAs 5' end and the 5' end of its Pol V transcript neighbors on the opposite strand. Consistent with our hypothesis, we found a strong peak of AGO4-associated 24-nt siRNAs 5' ends at 10 nucleotides downstream from the Pol V 5' end (Figure 1-5E). Overall, 78.07% of AGO4-associated 24-nt siRNAs had a Pol V-dependent transcripts partner detected in GRO-seq whose 5' end could be mapped 10 nucleotides away on the complementary strand.

To determine whether the slicing-associated U signature at position 10 was dependent on 24-nt siRNAs, which are transcribed by Pol IV, we examined the Pol V transcript sequence composition in the Pol IV mutant *nrpd1*. We found that in *nrpd1* the U preference at position 10 was completely abolished (Figure 1-5F,G). Instead, we observed the conventional +1 A/U and a

-1 U/A 5' signature (Figure 1-5F) similar to other RNA polymerases (Blevins et al., 2015; Hetzel et al., 2016; Li et al., 2015; Vo Ngoc et al., 2017; Zecherle et al., 1996; Zhai et al., 2015), and also similar to mRNA GRO-seq reads in wild type or the *nrpd1* mutant (Figure 1-6C,D). These results strongly support the hypothesis that the +10U signature is due to 24-nt siRNAs dependent slicing of Pol V transcripts.

AGO4, AGO6, and AGO9 are required for the slicing of Pol V transcripts.

Given that AGO4 is the main ARGONAUTE involved in RdDM, we tested whether AGO4 is also required for slicing of Pol V transcripts by performing GRO-seq in the *ago4-5* mutant in the Col-0 background (*ago4/Col-0*) and the *ago4-4* mutant in the Ws background (*ago4/Ws*). We observed that the +10U slicing signature of Pol V transcripts was reduced 13.26% in *ago4-5* relative to wild-type Col-0 and 12.37% in *ago4-4* relative to wild-type Ws (Figure 1-5B, Figure 1-7A-C,I). The remaining slicing signature in *ago4* mutants is likely due to redundancy of AGO4 with two other close family members, AGO6 and AGO9 (Eun et al., 2011; Mi et al., 2008). Therefore, we also performed GRO-seq in the *ago4-4/ago6-2/ago9-1* (*ago4/6/9*) triple mutant background (Wang and Axtell, 2016). The +10U signature in *ago4/6/9* mutants was completely abolished (Figure 1-7D,I) suggesting a complete lack of slicing.

Previous work showed that the Asp-Asp-His (DDH) catalytic motif of AGO4 is required for slicing of RNA transcripts *in vitro* (Qi et al., 2006). We therefore performed GRO-seq in plants containing either a wild-type AGO4 transgene (wtAGO4) expressed in *ago4/Ws* or the *ago4/6/9* mutant triple mutant, or a slicing defective AGO4 (D742A) mutant expressed in *ago4/Ws* or the *ago4/6/9* triple mutant (Wang and Axtell, 2016). We found that the wild-type AGO4 transgene

largely complemented the +10U slicing signature in the *ago* mutants, while the AGO4 D742A catalytic mutant failed to restore the +10U signature (Figure 1-7E-I). To rule out the possibility that the elimination of the +10U Pol V slicing signature in the *ago* mutants is caused by elimination of the +1A nucleotide preference of 24-nt siRNAs, we analyzed previously published small RNA-seq datasets corresponding to the same collection of *ago* mutant/transgene combinations (Wang and Axtell, 2016). We found that all mutants and mutant/transgene combinations retained a strong enrichment of A at position 1 of the 24-nt siRNAs (Figure 1-8A-H). These results further support the hypothesis that the +10U signature is due to Pol V transcript slicing, and that slicing is abolished in *ago4/6/9* triple mutants, although we cannot rule out minor levels of slicing that do not involve U-A pairing or by other AGO proteins.

SPT5L is required for the slicing of Pol V transcripts.

There are a number of proteins in the RdDM pathway whose precise function is unknown but that act at some point downstream of the biogenesis of siRNAs, including SUPPRESSOR OF TY INSERTION 5 – like/ KOW DOMAIN-CONTAINING TRANSCRIPTION FACTOR 1 (SPT5L) (Bies-Etheve et al., 2009; Greenberg et al., 2011; He et al., 2009; Huang et al., 2009; Rowley et al., 2011), DOMAINS REARRANGED METHYLTRANSFERASE3 (DRM3) (Zhong et al., 2015), INVOLVED IN DE NOVO2 (IDN2) (Ausin et al., 2009), IDN2-LIKE1 and 2 (IDL1 and 2) (Ausin et al., 2012; Zhang et al., 2012) SNF2-RING-HELICASE-LIKE1 and 2 (FRG1 and 2) (Groth et al., 2014), and SU(VAR)3-9 RELATED2 (SUVR2) (Han et al., 2014; Stroud et al., 2013). Mutations in these genes all show a partial reduction of DNA methylation associated with the RdDM pathway, rather than a complete loss of RdDM as seen in strong mutant such as *nrpd1* or *nrpe1* (Ausin et al., 2009; 2012; Bies-Etheve et al., 2009; Greenberg et

al., 2011; Groth et al., 2014; Han et al., 2014; He et al., 2009; Huang et al., 2009; Rowley et al., 2011; Stroud et al., 2013; Zhang et al., 2012; Zhong et al., 2015). To examine if any of these components are involved in the slicing of Pol V transcripts we performed GRO-seq in mutant backgrounds including *spt5l, drm3*, *idn2*, *idn2/idl1/idl2*, *frg1/frg2*, and *suvr2*. We observed that all mutants retained a strong +10U slicing signature (Figure 1-9A-E, Figure 1-10A) except for the *spt5l* mutant, which completely eliminated the slicing signature (Figure 1-9F, Figure. 1-10A). A trivial explanation for the lack of +10U slicing signature in *spt5l* would be that this mutant eliminated 24-nt siRNAs or eliminated the enrichment of A at the 5' nucleotide of 24-nt siRNAs. However, we found only a moderate (though significant) reduction of 24-nt siRNA abundance (Figure 1-10B) (Bies-Etheve et al., 2009; Greenberg et al., 2011; He et al., 2009; Huang et al., 2009) and a strong remaining +1A nucleotide preference (Figure 1-10C,D) in *spt5l*. These results reveal a novel role for SPT5L in the slicing of Pol V transcripts.

We also analyzed the effect of each of the mutants on the overall levels of Pol V GRO-seq signals (Figure 1-10E), and as a control examined their effects on the background levels of GRO-seq signals at the top 1,000 expressed Pol II genes (Figure 1-8I). While the *drm3*, *idn2*, *idn2/idl1/idl2*, *frg1/frg2*, and *suvr2* mutants showed only minor effects on overall Pol V transcript levels, *spt5l* showed a strong reduction.

This reduction was even greater than that seen in the Pol IV mutant *nrpd1*, a strong RdDM mutant which shows a much greater reduction in DNA methylation than in *spt5l* (Stroud et al., 2013). This result suggests that SPT5L plays a role in Pol V transcript stability and/or production. SPT5L is a homolog of the Pol II elongation factor SPT5 (Bies-Etheve et al., 2009). It has been shown to interact with the Pol V complex, but its precise role in the RdDM pathway has been

unclear (Bies-Etheve et al., 2009; Greenberg et al., 2011; He et al., 2009; Huang et al., 2009; Rowley et al., 2011). Our finding that both slicing and Pol V transcript levels are affected in *spt5l* suggests that SPT5L plays a dual role in the processing and utilization of Pol V transcripts.

DISCUSSION

In this work we show that Pol V transcripts are frequently sliced in a siRNA- and SPT5L-dependent manner. Because the slicing signature is present in Pol V transcripts that are in the process of transcribing, it is clear that this slicing is occurring co-transcriptionally. AGO4 mutations that affect the catalytic residues required for slicing show a partial loss of RdDM similar to *spt5l* mutants (Qi et al., 2006; Wang and Axtell, 2016), suggesting that the slicing step is required for efficient RNA-directed DNA methylation. However, it is also clear that slicing is not required for all RdDM, since *spt5l* mutants appear to abolish slicing, and yet show only a partial loss of CHH methylation at RdDM sites (Bies-Etheve et al., 2009; Greenberg et al., 2011; He et al., 2009; Rowley et al., 2011). AGO4 can also physically interact with DRM2, which provides an alternative mechanism by which AGO4/siRNA complexes can promote RdDM. This suggests a dual mechanism by which AGO4 can promote DRM2 activity, through both Pol V transcript slicing and through interaction with DRM2 (Model Figure 1-10F).

SPT5L contains a region rich in WG repeats (called the AGO hook) that is capable of binding to AGO4 (Bies-Etheve et al., 2009). AGO4 also interacts with a similar WG repeat region within the largest subunit of Pol V (El-Shami et al., 2007). It has been recently shown that deletion of the WG repeats of SPT5L, or deletion of the WG repeats of Pol V, still allow AGO4 recruitment and RdDM. However, simultaneous deletion of both WG repeat regions abolishes RdDM,

indicating that the WG-rich domains of SPT5L and Pol V are redundantly required for AGO4 recruitment (Lahmy et al., 2016). This genetic redundancy also indicates that SPT5L's role in AGO4 recruitment is unlikely to account for its requirement for Pol V transcript slicing. SPT5L is therefore a multifunctional protein mediating a number of steps in RdDM including AGO4 recruitment, and, as shown here, Pol V slicing and Pol V transcript abundance or stability (Model Figure 1-10F)

In *Drosophila*, similar slicing patterns were observed in the AGO3-rasiRNA 'ping-pong' pathway in which AGO3 directs cleavage of its cognate mRNA target across from nucleotides 10 and 11, measured from the 5' end of the small RNA guide strand, followed by the generation of secondary small RNAs from mRNA targets (Brennecke et al., 2007; Gunawardane et al., 2007). Thus, one hypothesis is that sliced Pol V RNAs are further trimmed to generate secondary small RNAs, as was previously proposed (Qi et al., 2006). However, we did not observe evidence suggesting secondary RNA production, suggesting that AGO4 slicing of Pol V transcripts does not result in the production of secondary small RNAs (data not shown). This is consistent with a recent study suggesting that AGO4 dependent siRNAs result from RdDM feedback rather than from secondary siRNA production (Wang and Axtell, 2016).

Our results also shed light on the long debate over the mechanism of action of AGO/siRNA complexes and whether the siRNAs target the nascent Pol V RNA or whether they bind directly to the DNA (Lahmy et al., 2016; Wierzbicki et al., 2008). Our results demonstrating siRNA-mediated slicing of Pol V nascent transcripts clearly supports an RNA targeting model whereby the siRNAs target the nascent Pol V RNA rather than binding directly to the DNA. This is also

supported by the conclusive data in fission yeast suggesting siRNA/RNA interactions (Noma et al., 2004; Shimada et al., 2016; Zofall et al., 2012). Once the AGO4-siRNAs have bound to nascent Pol V RNAs and slicing has occurred, one possibility is that the resulting sliced RNAs or siRNA/sliced RNA duplexes play a signaling role, perhaps through specific RNA binding proteins, in the targeting of the DRM2 methyltransferase to methylate chromatin (Model Figure 1-10F). This model is attractive because slicing represents the integration of the activities of the upstream Pol IV driven siRNA biogenesis pathway and the downstream Pol V driven non-coding RNA biogenesis pathway, which could provide additional accuracy and specificity for DNA methylation targeting. Another possibility is that slicing promotes the recycling of AGO/siRNA complexes, and/or Pol V transcripts to promote iterative cycles of targeting of DNA methylation through AGO4-DRM2 interactions (Zhong et al., 2014). Future studies aimed at understanding the biochemical details of the interaction of AGO4-bound siRNAs and Pol V targets are likely to shed additional light on the mechanisms of DNA methylation control.

MATERIALS AND METHODS

Plant Materials and Growth

The *A. thaliana* accession Columbia (Col-0) was used as the wild-type genetic background for this study unless specified. The mutant alleles of *nrpd1-4* (SALK_083051) (Herr et al., 2005), *nrpe1-12* (SALK_033852) (Pontier et al., 2005)*, spt5l-1* (SALK_001254)(Bies-Etheve et al., 2009), *drm3-1* (SALK_136439) (Zhong et al., 2015), *idn2-1* (SALK_012288) (Ausin et al., 2009), *suvr2-1*(SAIL_832_E07) (Groth et al., 2014), and *ago4-5 (described in* (Greenberg et al., 2011)*)* used in this study have been characterized previously and were in the Col-0 background. The double mutant for NRPD1 and NRPE1 was made by crossing *nrpd1-4* (SALK_083051) and

nrpe1-11 (SALK_029919) as described(Pontier et al., 2005). *frg1/2 (SALK_027637, SALK_057016)* double mutants were described as before (Groth et al., 2014). *idn2-1*, *idnl1-1* (SALK_075378), and *idnl2-1* (SALK_012288) triple mutant were described before(Ausin et al., 2012). Ws, *ago4/Ws*, *ago4/ago6/ago9*, *ago4/wtAGO4*, *ago4/D742A*, *ago4/6/9/wtAGO4*, and *ago4/6/9/D742A* were described as before(Wang and Axtell, 2016). All plants were grown on soil under long day conditions (16 hours light, 8 hours dark). Inflorescence tissues with both floral buds and open flowers were collected and used for the GRO-seq procedure. T-DNAs were confirmed by PCR-based genotyping.

Nuclei Isolation

Approximately 10 grams of inflorescence and meristem tissue was collected from plants and immediately placed in ice cold grinding buffer (300 mM sucrose, 20 mM Tris, pH 8.0, 5 mM MgCl2, 5 mM KCl, 0.2% Triton X-100, 5 mM β-mercaptoethanol, and 35% glycerol). Nuclei were isolated as described previously (Hetzel et al., 2016). Briefly, samples were ground with an OMNI International General Laboratory Homogenizer at 4°C until well homogenized, filtered through a 250 μm nylon mesh, a 100 μm nylon mesh, a miracloth, and finally a 40 μm cell strainer before being split into 50 ml conical tubes. Samples were spun for 10 minutes at 5,250*g*, the supernatant was discarded, and the pellets were pooled and resuspended in 25 ml of grinding buffer using a Dounce homogenizer. The wash step was repeated at least once more and nuclei were resuspended in 1 ml of freezing buffer (50 mM Tris, pH 8.0, 5 mM MgCl2, 20% glycerol, and 5 mM β-mercaptoethanol).

GRO-seq

Approximately $5x10^6$ nuclei in 200 µl of freezing buffer were run-on in 3x NRO-reaction buffer (Hetzel et al., 2016). For GRO-seq in Ws, *ago4/Ws*, *ago4/ago6/ago9*, *ago4/wtAGO4*, *ago4/D742A*, *ago4/6/9/wtAGO4*, and *ago4/6/9/D742A*, approximately $3x10^5$ to $5x10^5$ nuclei were used. To minimize run-on length, the limiting CTP concentration was reduced to a final concentration of 20 nM. Reactions were stopped after 5 minutes to minimize run on length (~5-15 nts) while still incorporating brUTP by addition of 750 µl TRIzol LS(Fisher Scientific) and RNA was purified according to the manufacturer's manual. Without fragmentation or Terminator treatment, nascent RNA was enriched twice for BrU by αBrdU (Santa Cruz Biotechnology sc-32323AC) and immunoprecipitated as described in Hetzel et al. 2016 (Hetzel et al., 2016). Subsequently, sequencing libraries were prepared from precipitated RNA using TruSeq Small RNA Library Prep kit following manufacturer instructions (Illumina). For most GRO-seq libraries, 14 cycles of PCR were used to amplify the libraries and products ranging from 100 to 500 bp were size selected by agarose gel, except for replicate 1 and 2 of *spt5l* (replicate 3 was prepared the same way as all other GRO-seq libraries), where products were size selected by double SPRI bead purification (ratio of Ampure beads to library: 0.5:1 to 1.1:1). The libraries were sequenced on either Illumina HiSeq 2000 or 2500 platform.

ChIP-seq

Chromatin immunoprecipitation was performed from 2 grams of formaldehyde crosslinked flower tissue as previously described (Zhong et al., 2012), except that half of the input was immunoprecipitated with 3 µg of affinity purified anti-NRPE1 antibody generated by Covance that recognizes the peptide N-CDKKNSETESDAAAWG- C (Ream et al., 2009), and the other half was immunoprecipitated with pre-immune serum as control. DNA libraries for Illumina

sequencing were generated using the Ovation Ultralow V2 system (NuGEN), and the libraries were sequenced on a HiSeq 2000 platform for single-end 50 bp, following the manufacturers' instructions.

Small RNA-seq

Total RNA was first extracted with Zymo Direct-zol RNA mini Prep kit (ZRC200687) followed by a size selection of RNA on a 15% Urea TBE Polyacrylamide gel (Invitrogen, EC6885BOX). Gels containing 15- to 30-nt were cut for small RNA library. After gel elution, Illumina TruSeq Small RNA kit (RS-200-0012) was used for making small RNA library. Agilent D1000 ScreenTape (5067-5582) was then used for checking the size and quality of final libraries.

Bioinformatic Analysis

GRO-seq analysis

Qseq files from the sequencer were demultiplexed and converted to fastq format with a customized script for downstream analysis. For GRO-seq data, paired-end reads were first trimmed for Illumina adaptors and primers using Cutadapt (v 1.9.1). After trimming, reads less than 10 bp long were removed with a customized Perl script. Paired-end reads were then separately aligned to the reference TAIR10 genome using Bowtie (v1.1.0) (Langmead et al., 2009) by allowing only unique hit (-m 1) and up to 3 mismatches (-v 3). Paired reads aligned to positions within 2,000 bp to each other were considered as correct read pairs, and reads aligned to Watson or Crick strands were separated by a customized Perl script.

ChIP-seq analysis

Qseq files from the sequencer were demultiplexed and converted to fastq format with a customized script for downstream analysis. Fastq reads were aligned to the Arabidopsis reference genome (TAIR10) with Bowtie (v1.0.0) (Langmead et al., 2009), allowing only uniquely mapping reads with fewer than two mismatches, and duplicated reads were combined into one read. NRPE1 ChIP-seq peak were called using MACS2 (v 2.1.1.) (Zhang et al., 2008) in Col-0 and *nrpd1*, respectively, with default parameters using ChIP-seq with pre-immune serum in each condition as control. ChIP-seq metaplots were plotted using NGSplot (v 2.41.4) (Shen et al., 2014).

Identification of Pol V-dependent transcripts from GRO-seq data

In order to remove signals from annotated gene regions, we only included GRO-seq reads aligned to defined Pol V occupied regions. Pol V ChIP-seq peak regions were split into 100 bp bins and the reads from GRO-seq in each bin were counted. To call Pol V-dependent transcripts, the R package DESeq2 (Anders and Huber, 2010) was used applied. Only bins with at least 4-fold enrichment in Col-0 compared to the *nrpe1* and *nrpd1/e1* mutant and FDR less than 0.05 were retained. Bins within 200 bp of each other were then merged into Pol V-dependent transcripts clusters. To characterize Pol IV dependency on those Pol V-dependent transcripts clusters, we checked NRPE1 binding in *nrpd1* mutant. If a Pol V-dependent transcripts cluster was not bound by NRPE1 in *nrpd1* mutant while also had a RPKM (Reads Per Kilobase Million) of GRO-seq in *nrpd1* greater than 2, then this site was classified as Pol IV/V codependent. On the other hand, if a Pol V-dependent transcripts cluster was also bound by NRPE1 in *nrpd1* mutant while had a RPKM of GRO-seq in *nrpd1* less than 1, then this site was classified as Pol IV-independent Pol V sites.

AGO4 RIP-seq and total small RNA analysis

Qseq files for small RNA-seq from the sequencer were demultiplexed and converted to fastq format with a customized script for downstream analysis. Raw AGO4 RIP-seq data were obtained from previously published datasets (GSM707686)(Wang et al., 2011) . Reads were then trimmed for Illumina adaptors using Cutadapt (v 1.9.1) and mapped to the TAIR10 reference genome using Bowtie(v1.1.0) (Langmead et al., 2009) allowing only one unique hit (-m 1) and zero mismatch.

Whole Genome Bisulfite Sequencing (WGBS) analysis

Processed WGBS data of Col-0 and *nrpd1* were obtained from previously published datasets (GSE39901, GSE38286) (Stroud et al., 2013). CG, CHG, and CHH methylation over different regions were extracted using a customized Perl script.

Data availability

High-throughput sequencing data that support the findings in this study can be accessed through the Gene Experssion Omnibus (GEO) database with accession number GSE108078 and GSE100010.

FIGURE LEGENDS

Table 1-1. Sequencing information summary.

Summary of sequenced ChIP-seq, GRO-seq and sRNA-seq data used in this chapter.

Figure 1-1. Capturing Pol V-dependent transcripts with GRO-seq.

(A) Procedure for constructing *Arabidopsis* GRO-seq library, which captures nascent Pol V transcripts. 7meG-capped transcripts generated by Pol II are excluded by selective ligation to the 5′ monophosphorylated (5′Pi) RNAs generated by Pol I, IV, and V. **(B)** Screenshot of CG, CHG, and CHH methylation in wild-type Col-0, Pol V ChIP-seq in Col-0, and GRO-seq in Col-0, *nrpe1*, and *nrpd1/e1* over the previously identified Pol V locus IGN5 (Wierzbicki et al., 2008). For CG, CHG, and CHH methylation, y-axis indicate the percentage of methylation. Plus (+) and Minus (-) indicate the strandness of GRO-seq signal. (**C)** Metaplot of Pol V ChIP-seq signal over input and ratio of GRO-seq signal in Col-0 to *nrpe1* graphed over the centers of Pol V occupied regions defined by Pol V ChIP-seq.

Figure 1-2. Modified GRO-seq is able to capture nascent Pol V-dependent transcripts.

(A) Scatterplot of signals from two independent GRO-seq experiments in Col-0. The Pearson's correlation coefficient is calculated and shown on the plot. **(B)** Metaplot showing GRO-seq signals over Pol V-occupied regions in Col-0 and *nrpe1*. **(C)** Metaplot showing GRO-seq signals over annotated genes in Col-0 and *nrpe1*. **(D)** Scatterplot of normalized signals from Pol V ChIP-seq versus GRO-seq in Col-0. The Pearson's correlation coefficient is calculated and shown on the plot. **(E)** Genome browser screenshot for CG, CHG, and CHH methylation in Col-0, Pol V ChIP-seq signals in Col-0, and GRO-seq signals in Col-0, *nrpe1*, and *nrpd1/e1* of a

representative long TE and a representative short TE. Plus (+) and Minus (-) indicate the strandness of GRO-seq signal.

Figure 1-3. Characteristics of Pol V-dependent transcripts.

(A) Distribution of ratios of plus strand GRO-seq signals over minus strand GRO-seq signals in Col-0 over the top 500 expressed mRNAs. **(B)** Distribution of ratios of plus strand GRO-seq signals over minus strand GRO-seq signals in Col-0 over the top 500 Pol V enriched regions defined by Pol V ChIP-seq. **(C)** Pol V ChIP-seq signals over inputs and the ratio of GRO-seq signal in Col-0 to *nrpe1* plotted over Pol V-associated transposons with different lengths.

Figure 1-4. Characterization of Pol IV/V-codependent sites and Pol IV-independent Pol V sites.

(A-B) Genome browser screenshot for Pol V ChIP-seq signals in Col-0 and GRO-seq signals in Col-0, *nrpe1, nrpd1,* and *nrpd1/e1* of a representative Pol IV/V-codependent site (**A**) and Pol IV-independent Pol V site (**B**) . Plus (+) and Minus (-) indicate the strandness of GRO-seq signal. **(C-D)** Heatmap of log2 ratio of GRO-seq in Col-0 vs. *nrpe1*, GRO-seq in *nrpd1* vs. *nrpd1*, Pol V ChIP signals in Col-0, and Pol V ChIP-seq signals in *nrpd1* plotted over Pol IV/V-codependent sites (**C**) and Pol IV-independent Pol V sites (**D**). **(E)** Boxplot of CG, CHG, and CHH methylation difference in *nrpd1* vs. Col-0. **p-value* < *0.05* (Welch Two Sample t-test). **(F)** Normalized 24-nt siRNAs abundance in Col-0 over Pol IV/V-codependent sites and Pol IV-independent Pol V sites. **p-value* < *0.05* (Welch Two Sample t-test).

Figure 1-5. Pol V transcripts is sliced in a small RNA dependent manner.

(A) Size distribution of nascent transcripts in Col-0 over Pol V-dependent regions. All replicates for Col-0 GRO-seq were merged for this plot. **(B)** The relative nucleotide bias of each position in the upstream and downstream 20-nt of nascent transcripts captured in Col-0. All replicates for Col-0 GRO-seq were merged for this plot. **(C)** A predicted model indicating the first 10-nt of AGO4/6/9 associated small RNAs show complementarities to the first 10-nt of sliced nascent transcripts over Pol V-dependent regions captured in GRO-seq library. **(D)** The relative nucleotide bias of each position for all AGO4-associated 24-nt siRNAs over regions that generated Pol V-dependent transcripts. **(E)** Frequency map of the separation of 5' of Pol V-dependent RNAs mapping to AGO4-associated 24-nt siRNAs on the opposite strand. **(F)** The relative nucleotide bias of each position in the upstream and downstream 20-nt of nascent transcripts captured in *nrpd1*. **(G)** The percentage of U presented over genomic average at position 10 from the 5' ends of nascent transcripts captured with GRO-seq in Col-0, *nrpd1*, *nrpe1*, and *nrpd1/e1*.

Figure 1-6. Pol V transcripts with different lengths are sliced.

(A) Size distribution of nascent transcripts in *nrpd1* over Pol V-dependent regions. Replicates were merged for this plot. **(B)** The percentage of U presented over genomic average at position 10 from the 5' ends of nascent transcripts captured with GRO-seq in six biological replicates for Col-0. **(C,D)** The relative nucleotide bias of each position in the upstream and downstream 20-nt of nascent RNAs generated from the top 1,000 expressed annotated gene regions in Col-0 (**c**) and *nrpd1* (**D**). Replicates were merged for plot (**C,D**). **(E-I),** The relative nucleotide bias of each

position in the upstream and downstream 20-nt of nascent transcripts of 30- to 40-nt long (**E**), 40- to 50-nt long (**F**), 50- to 60-nt long (**G**), 60- to 70-nt long (**H**) and 70-nt and longer (**I**) captured in Col-0. Replicates were merged for plot (**E-I**).

Figure 1-7. Slicing of Pol V transcripts requires AGO4/6/9.

(A-H) The relative nucleotide bias of each position in the upstream and downstream 20-nt of nascent transcripts captured in Ws (**A**), *ago4/Col-0* (**B**), *ago4/Ws* (**C**), *ago4/6/9* (**D**), *ago4/wtAGO4* (**E**), *ago4/D742A* (**F**), *ago4/6/9/wtAGO4* (**G**) and *ago4/6/9/D742A* (**H**). Replicates were merged for plot (**A-H**). **(I)** The percentage of U presented over genomic average at position 10 from the 5' end of nascent transcripts captured with GRO-seq in Col-0, *ago4/Col-0*, Ws, *ago4/Ws*, *ago4/6/9*, *ago4/wtAGO4*, *ago4/D742A*, *ago4/6/9/wtAGO4*, and *ago4/6/9/D742A*.

Figure 1-8. **24nt-siRNAs retain strong enrichment of A at position 1 for *ago4*, *ago4/6/9* mutant and *ago4* or *ago4/6/9* mutant expressing wtAGO4 or D742A.**

(A-H) The relative nucleotide bias of each position for 24-nt siRNAs over Pol V dependent regions in Col-0 (**A**), Ws (**B**), *ago4/Ws* (**C**), *ago4/wtAGO4* (**D**), *ago4/D742A* (**E**), *ago4/6/9* (**F**), *ago4/6/9/wtAGO4* (**G**) and *ago4/6/9/D742A* (**H**). **(I),** Boxplot of normalized GRO-seq signals from top 1,000 expressed annotated gene in Col-0, *nrpd1*, *nrpe1*, *nrpd1/e1*, *spt5l*, *drm3*, *frg1/2*, *idn2/idl1/idl2*, *idn2*, and *suvr2*. N.S., not significant.

Figure 1-9. Slicing signature of Pol V transcripts is eliminated in *spt5l* mutants.

(A-F) The relative nucleotide bias of each position in the upstream and downstream 20-nt of nascent transcripts captured in *idn2* (**A**), *idn2/idl1/idl2* (**B**), *drm3* (**C**), *suvr2* (**D**), *frg1/2* (**E**), *spt5l* (**F**). Replicates were merged for plot (**A-F**).

Figure 1-10. SPT5L is required for slicing of Pol V transcripts.

(A) The percentage of U presented over genomic average at position 10 from the 5' end of nascent transcripts captured with GRO-seq in Col-0, *spt5l*, *drm3*, *frg1/2*, *idn2/idl1/2*, *idn2*, and *suvr2*. **(B)** Normalized 24-nt siRNAs abundance in Col-0, *spt5l*, and *nrpd1*. **p-value < 0.05* (Welch Two Sample t-test). **(C,D)** The relative nucleotide bias of each position for all 24-nt siRNAs in Col-0 (**C**) and *spt5l* (**D**) generated over Pol V-dependent regions. **(E)** Nascent transcripts abundance over Pol V-dependent regions in Col-0, *nrpd1*, *nrpe1*, *nrpd1/e1*, *spt5l*, *drm3*, *frg1/2*, *idn2/idl1/2*, *idn2*, and *suvr2*. **p-value < 0.05* (Welch Two Sample t-test). **(F)** Proposed model for slicing of Pol V transcripts.

Table 1-1

Arabidopsis TAIR10 Genome

GROseq libraries (up to three mismatches, uniquely mapped)

Lib_ID	Type	genotype	replicate	Tissue	Sequencing	Total sequenced reads pair	reads pairs after filtering*	Uniquely-mapped left reads	Left reads discard due to multiple hit	Uniquely-mapped right reads	Right reads discard due to multiple hit	Valid reads pair
1	GRO-seq	Col-0	rep1	floral	PE100	97540139	87544808	53944585	26935593	56874153	20147460	24009496
2	GRO-seq	Col-0	rep2	floral	PE100	89632786	83490393	54050903	27168692	56575203	20324463	27066083
3	GRO-seq	*nrpd1*	rep1	floral	PE100	100540272	93204478	60085313	26602684	60364968	16469288	29316932
4	GRO-seq	*nrpd1*	rep2	floral	PE100	85930695	81848692	60096484	26833251	60158950	16089875	28593729
5	GRO-seq	*nrpe1*	rep1	floral	PE100	89433552	77281072	47384235	22726070	47418961	22709073	5181555
6	GRO-seq	*nrpe1*	rep2	floral	PE100	72267313	56790743	12809855	25590731	12726940	25456807	21690567
7	GRO-seq	*nrpd1e1*	rep1	floral	PE100	84937617	66905457	37946123	21363691	37972795	21080374	32084768
8	GRO-seq	*nrpd1e1*	rep2	floral	PE100	94426460	89362713	66364506	17844421	66036871	17510932	15753960
9	GRO-seq	*spt5l*	rep1	floral	PE50	166782259	166163620	63723111	87928371	63200149	88298346	20184273
10	GRO-seq	*spt5l*	rep2	floral	PE50	145737481	145098088	61431476	74135048	60586496	73984055	15866054
11	GRO-seq	*drm3*	rep1	floral	PE100	67773999	51643762	24587370	15166519	24388706	15874018	7174834
12	GRO-seq	*drm3*	rep2	floral	PE100	69686841	54739280	26547694	15406999	26314710	16107445	7523902
13	GRO-seq	*suvr2*	rep1	floral	PE100	71683810	58746819	24408756	21224859	24311436	23302461	5325226
14	GRO-seq	*suvr2*	rep2	floral	PE100	72903823	59937563	23261289	22141477	23191204	24362584	5720098
15	GRO-seq	*idn2/idl1/idl2*	rep1	floral	PE100	68092721	54099814	25484120	18749212	25315413	19834765	7846784
16	GRO-seq	*idn2/idl1/idl2*	rep2	floral	PE100	44944013	33593605	15496029	11462696	15387774	12126936	5733973
17	GRO-seq	*frg1/2*	rep1	floral	PE100	50838500	42335401	21291681	25235727	21084861	15057089	9670635
18	GRO-seq	*frg1/2*	rep2	floral	PE100	70386263	60079022	32959357	19500370	32412369	20485391	11730327
19	GRO-seq	*ago4-5*	rep1	floral	PE100	60391602	47616656	23595205	16515139	23322100	17314272	7662834
20	GRO-seq	*ago4-5*	rep2	floral	PE100	64665716	53127838	27806159	18335155	27386817	18997560	7514061
21	GRO-seq	*idn2*	rep1	floral	PE100	67281997	55081821	29854417	16561579	29571714	17430153	7113535
22	GRO-seq	*idn2*	rep2	floral	PE100	61468502	47480586	25448503	13983013	25159179	14465968	6509755
23	GRO-seq	Col-0	rep3	floral	PE100	55987939	42789222	18615120	15192889	18431502	15930138	5611951
24	GRO-seq	Col-0	rep4	floral	PE100	44933752	34443933	15652100	11603448	15425520	12185934	5381793

25	GRO-seq	*spt5l*	rep3	floral	PE100	78107137	57652390	33293184	18717257	32217500	18587349	14548666
26	GRO-seq	Col-0	rep5	floral	PE100	83510185	45544295	12222621	18535116	12669854	18993929	4353285
27	GRO-seq	Col-0	rep6	floral	PE100	89941085	49398907	13266808	20955943	13675822	20788342	4626321
28	GRO-seq	Ws	rep1	floral	PE100	97595463	36357707	6157205	17197312	6528891	17859216	5019289
29	GRO-seq	Ws	rep2	floral	PE100	103502495	48839967	13032832	18959938	13263113	19372466	5206264
30	GRO-seq	*ago4-4/Ws*	rep1	floral	PE100	83266696	32500019	4220907	16223701	4696115	15670838	1347380
31	GRO-seq	*ago4-4/Ws*	rep2	floral	PE100	103527868	35009989	5382366	18390623	5763119	17798076	1718150
32	GRO-seq	*ago4/wtAGO4*	rep1	floral	PE100	95001034	39892165	8358507	17466560	8696443	16669219	3032212
33	GRO-seq	*ago4/wtAGO4*	rep2	floral	PE100	99993176	36291682	8882618	16004659	9060927	15453897	3345580
34	GRO-seq	*ago4/D762A*	rep1	floral	PE100	91597220	44361530	7534976	24159480	8566286	22580421	2010798
35	GRO-seq	*ago4/D762A*	rep2	floral	PE100	107458615	42549962	6882067	23735633	7588831	22607507	1933406
36	GRO-seq	*ago4/6/9*	rep1	floral	PE100	114891255	36282941	4737653	19653094	5145776	20051435	1495014
37	GRO-seq	*ago4/6/9*	rep2	floral	PE100	119342710	43725214	9796918	22167845	10020444	22234199	2677546
38	GRO-seq	*ago4/6/9/wtAGO4*	rep1	floral	PE100	110504191	42534828	7856758	20324300	8126257	20972182	1641485
39	GRO-seq	*ago4/6/9/wtAGO4*	rep2	floral	PE100	77374006	34246861	8234112	13085174	8534557	13454233	2948378
40	GRO-seq	*ago4/6/9/D762A*	rep1	floral	PE100	105765179	45767953	11561168	23336842	11335595	23027208	4305021
41	GRO-seq	*ago4/6/9/D762A*	rep2	floral	PE100	87324013	39115960	9701624	17313123	9945097	17390848	3445693

*Adaptors and sequencing primers were trimmed in raw reads pair. After trimming, only reads pairs with length greater than 10bp

Arabidopsis TAIR10 Genome							
ChIPseq libraries (up to two mismatches, uniquely mapped)							
Lib_ID	Type	genotype	IP	Tissue	Sequencing	Raw reads	Uniquely-mapped
45	ChIP-seq	Col-0	pre-immune	floral	SE50	22274759	15752226 (70.72%)
46	ChIP-seq	Col-0	NRPE1	floral	SE50	20248789	15645938 (77.27%)
47	ChIP-seq	*nrpd1*	pre-immune	floral	SE50	17952606	12625263 (70.33%)
48	ChIP-seq	*nrpd1*	NRPE1	floral	SE50	11137634	8437601 (75.76%)

Arabidopsis TAIR10 Genome									
small RNAseq libraries (no mismatch allowed, uniquely mapped)									
Lib_ID	Type	genotype	Tissue	Sequen	Raw reads	Uniquely-mapped	21nt count	22nt count	24nt count

49	sRNA-seq	Col-0	floral	SE50	36913777	10586467 (28.68%)	678818	799141	5930723
50	sRNA-seq	Col-0	floral	SE50	34895400	7790318 (22.32%)	676330	681217	4141587
51	sRNA-seq	*spt5l*	floral	SE50	17632754	5103329 (28.94%)	336055	244292	3077292
52	sRNA-seq	*spt5l*	floral	SE50	28721993	7850079 (27.33%)	832372	715266	3776870

Figure 1-1

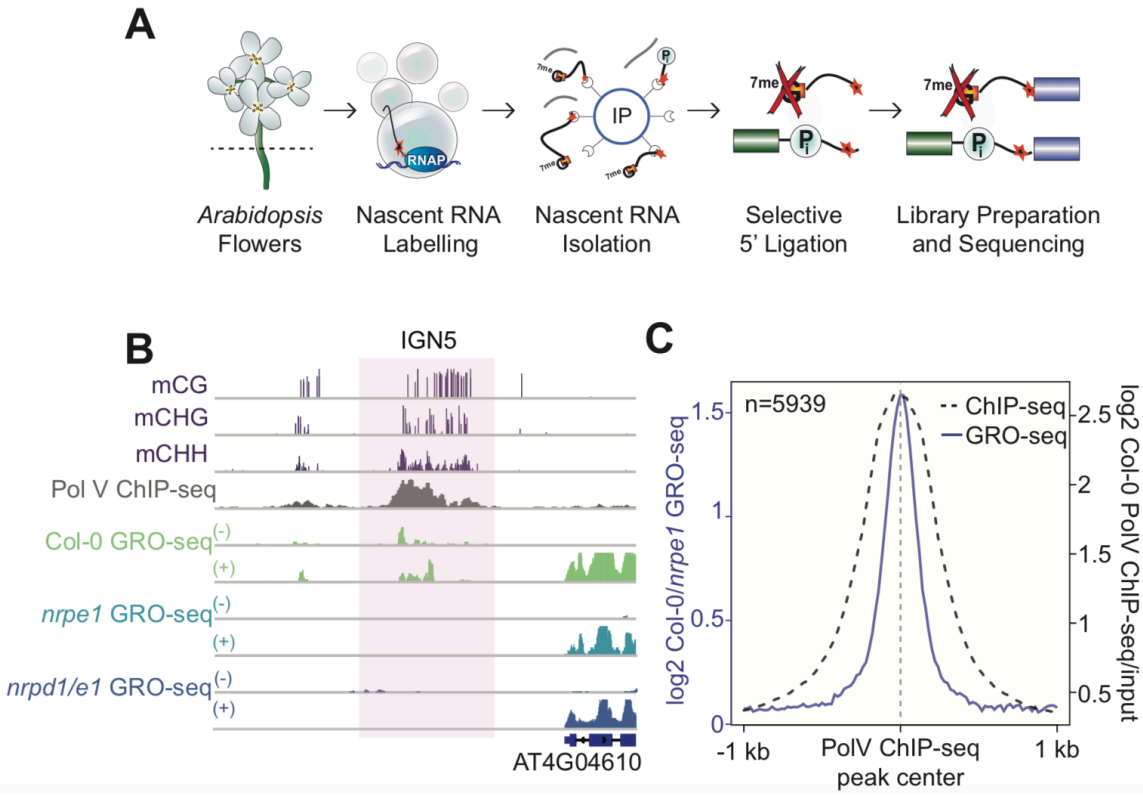
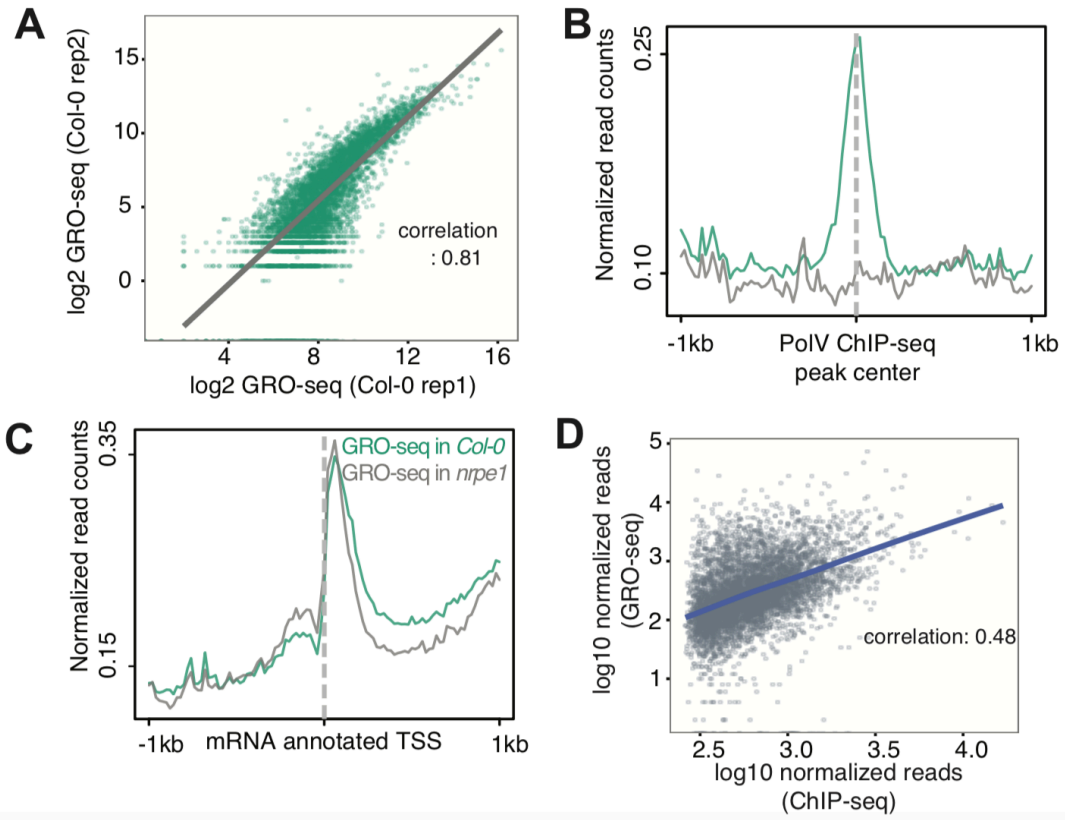

A
Arabidopsis Flowers
Nascent RNA Labelling
RNAP
Nascent RNA Isolation
IP
Selective 5' Ligation
7me
Library Preparation and Sequencing
B
IGN5
mCG
mCHG
mCHH
Pol V ChIP-seq
Col-0 GRO-seq
nrpe1 GRO-seq
nrpd1/e1 GRO-seq
AT4G04610
C
n=5939
ChIP-seq
GRO-seq
log2 Col-0/nrpe1 GRO-seq
log2 Col-0 PolV ChIP-seq/input
-1 kb
PolV ChIP-seq peak center
1 kb

Figure 1-2


A
log2 GRO-seq (Col-0 rep2)
log2 GRO-seq (Col-0 rep1)
correlation
: 0.81
B
Normalized read counts
0.25
0.10
-1kb
PolV ChIP-seq
peak center
1kb
C
Normalized read counts
0.35
0.15
GRO-seq in Col-0
GRO-seq in nrpe1
-1kb
mRNA annotated TSS
1kb
D
log10 normalized reads
(GRO-seq)
correlation: 0.48
log10 normalized reads
(ChIP-seq)

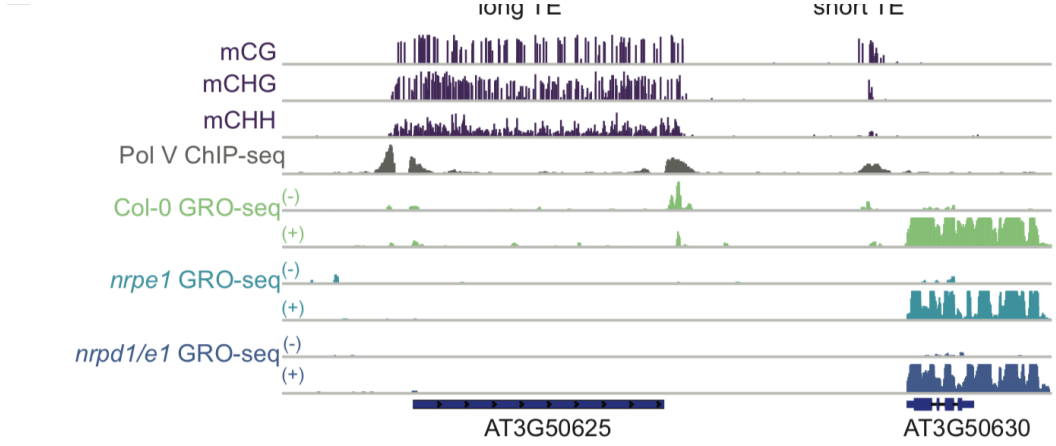

long TE
short TE
mCG
mCHG
mCHH
Pol V ChIP-seq
Col-0 GRO-seq
nrpe1 GRO-seq
nrpd1/e1 GRO-seq
(-)
(+)
AT3G50625
AT3G50630

Figure 1-3

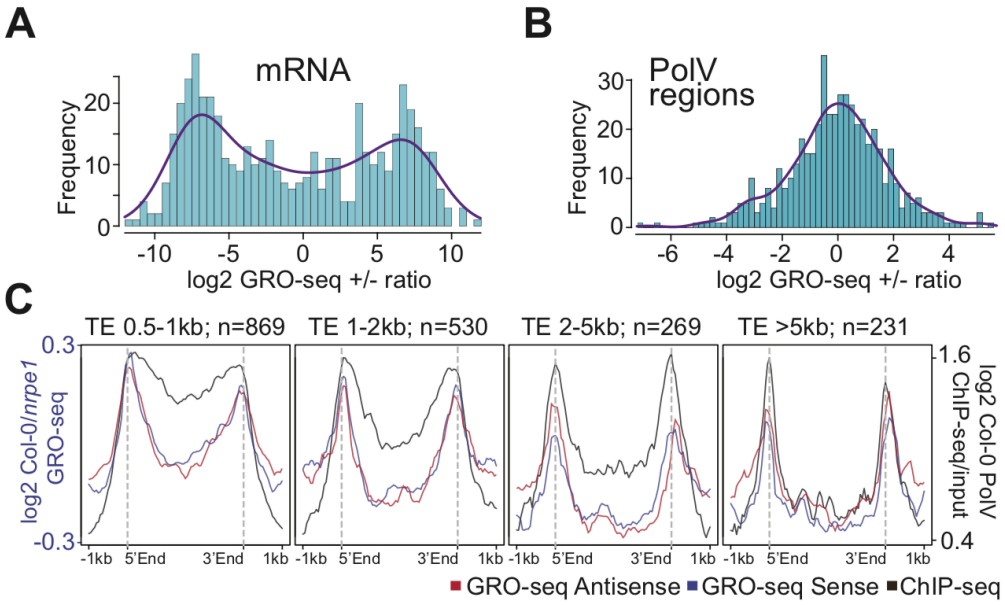

A
mRNA
Frequency
log2 GRO-seq +/- ratio
B
PolV regions
Frequency
log2 GRO-seq +/- ratio
C
TE 0.5-1kb; n=869
TE 1-2kb; n=530
TE 2-5kb; n=269
TE >5kb; n=231
log2 Col-0/nrpe1 GRO-seq
log2 Col-0 PolV ChIP-seq/input
-1kb
5'End
3'End
1kb
GRO-seq Antisense
GRO-seq Sense
ChIP-seq

Figure 1-4

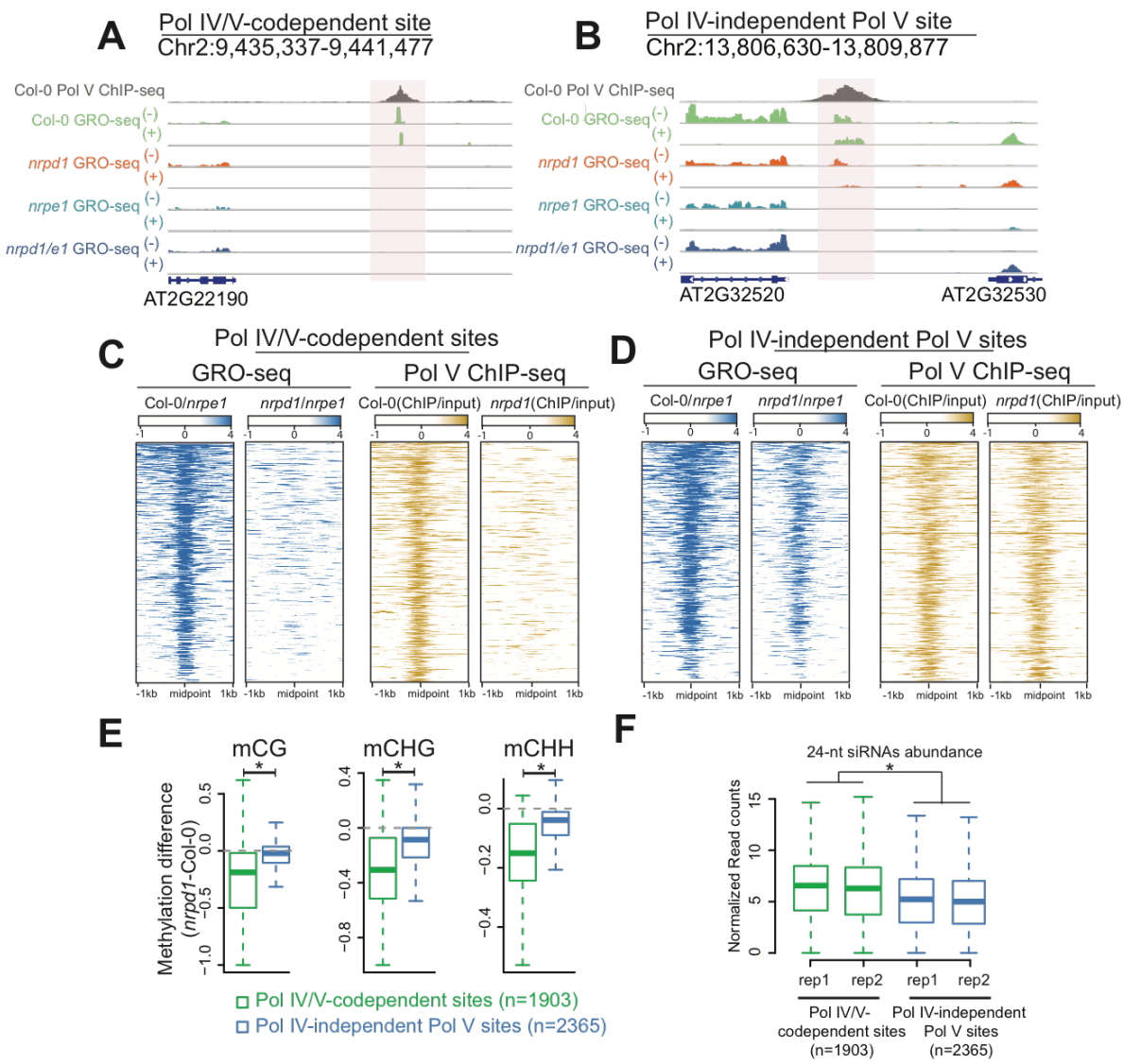

A
Pol IV/V-codependent site
Chr2:9,435,337-9,441,477
Col-0 Pol V ChIP-seq
Col-0 GRO-seq
nrpd1 GRO-seq
nrpe1 GRO-seq
nrpd1/e1 GRO-seq
AT2G22190
B
Pol IV-independent Pol V site
Chr2:13,806,630-13,809,877
AT2G32520
AT2G32530
C
Pol IV/V-codependent sites
GRO-seq
Pol V ChIP-seq
Col-0/nrpe1
nrpd1/nrpe1
Col-0(ChIP/input)
nrpd1(ChIP/input)
-1kb midpoint 1kb
D
Pol IV-independent Pol V sites
E
mCG
mCHG
mCHH
Methylation difference (nrpd1-Col-0)
Pol IV/V-codependent sites (n=1903)
Pol IV-independent Pol V sites (n=2365)
F
24-nt siRNAs abundance
Normalized Read counts
rep1
rep2
Pol IV/V-codependent sites (n=1903)
Pol IV-independent Pol V sites (n=2365)

Figure 1-5

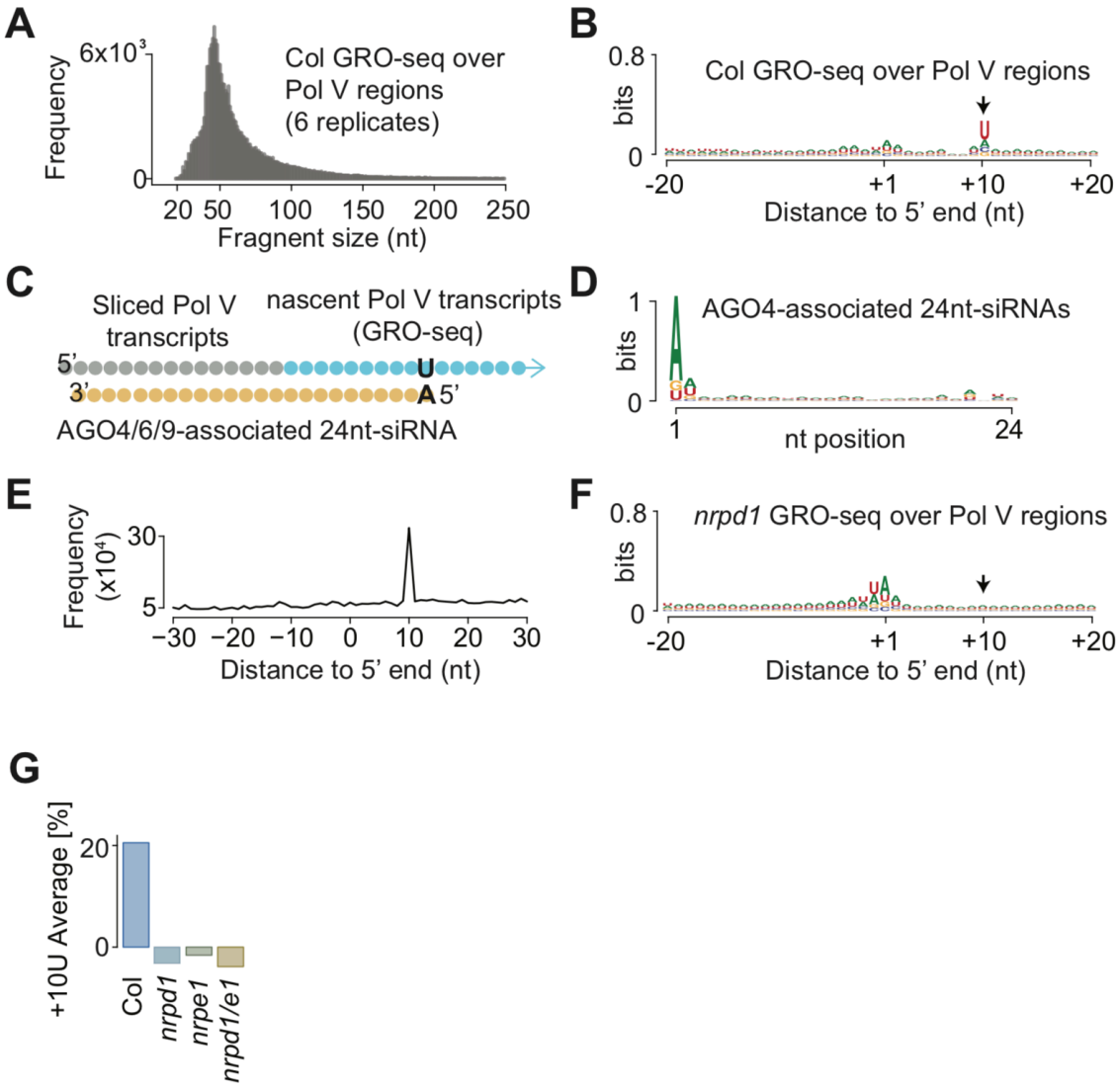

A
6x10³
Frequency
0
Col GRO-seq over
Pol V regions
(6 replicates)
20 50 100 150 200 250
Fragnent size (nt)
B
0.8
bits
0
Col GRO-seq over Pol V regions
-20 +1 +10 +20
Distance to 5' end (nt)
C
Sliced Pol V
transcripts
nascent Pol V transcripts
(GRO-seq)
5'
3'
U
A 5'
AGO4/6/9-associated 24nt-siRNA
D
1
bits
0
AGO4-associated 24nt-siRNAs
1
nt position
24
E
Frequency
(x10⁴)
30
5
−30 −20 −10 0 10 20 30
Distance to 5' end (nt)
F
0.8
bits
0
nrpd1 GRO-seq over Pol V regions
-20 +1 +10 +20
Distance to 5' end (nt)
G
+10U Average [%]
20
0
Col
nrpd1
nrpe1
nrpd1/e1

Figure 1-6

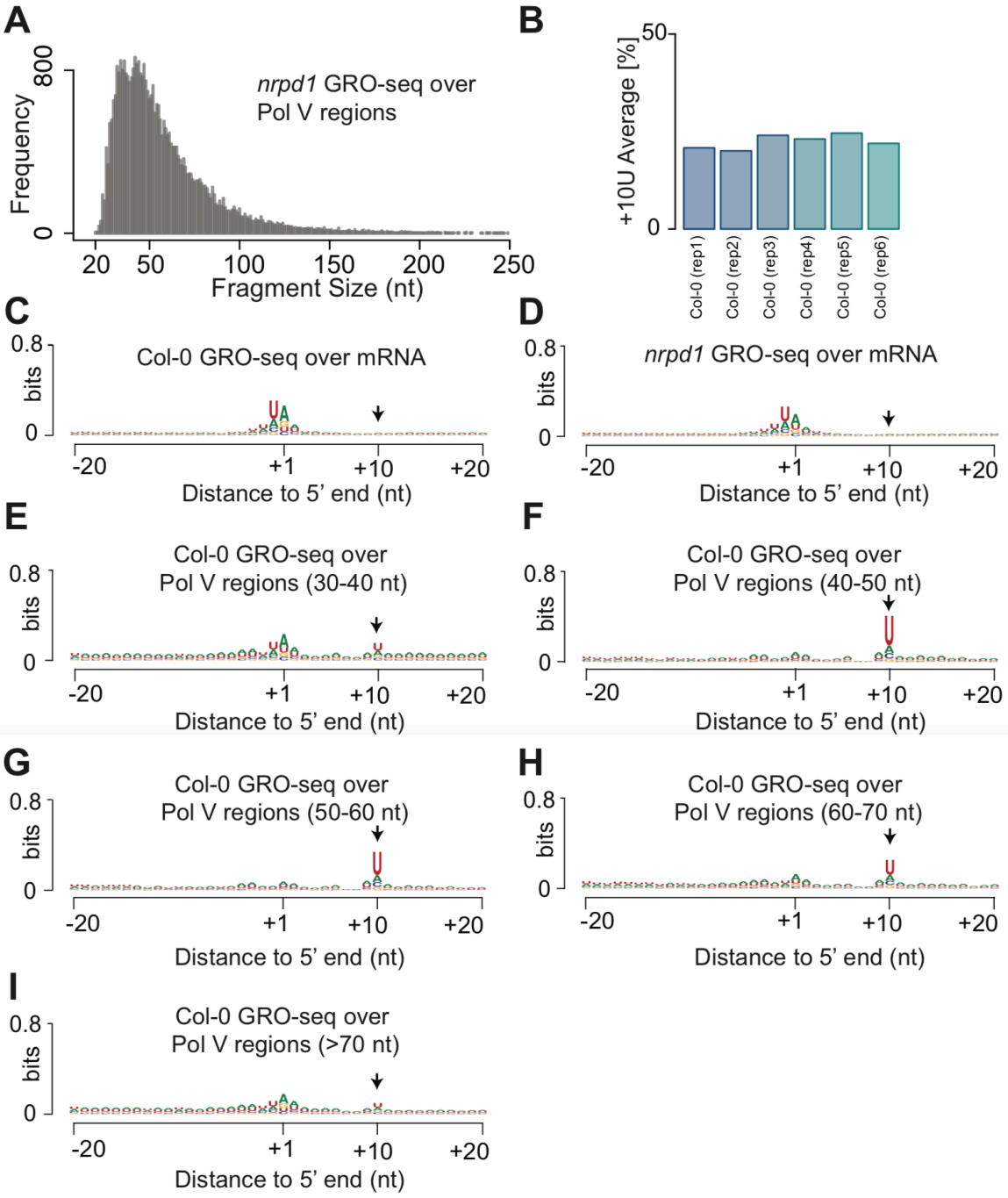

A
Frequency
800
0
nrpd1 GRO-seq over
Pol V regions
20
50
100
150
200
250
Fragment Size (nt)
B
+10U Average [%]
50
0
Col-0 (rep1)
Col-0 (rep2)
Col-0 (rep3)
Col-0 (rep4)
Col-0 (rep5)
Col-0 (rep6)
C
Col-0 GRO-seq over mRNA
bits
0.8
0
-20
+1
+10
+20
Distance to 5' end (nt)
D
nrpd1 GRO-seq over mRNA
E
Col-0 GRO-seq over
Pol V regions (30-40 nt)
F
Col-0 GRO-seq over
Pol V regions (40-50 nt)
G
Col-0 GRO-seq over
Pol V regions (50-60 nt)
H
Col-0 GRO-seq over
Pol V regions (60-70 nt)
I
Col-0 GRO-seq over
Pol V regions (>70 nt)

Figure 1-7

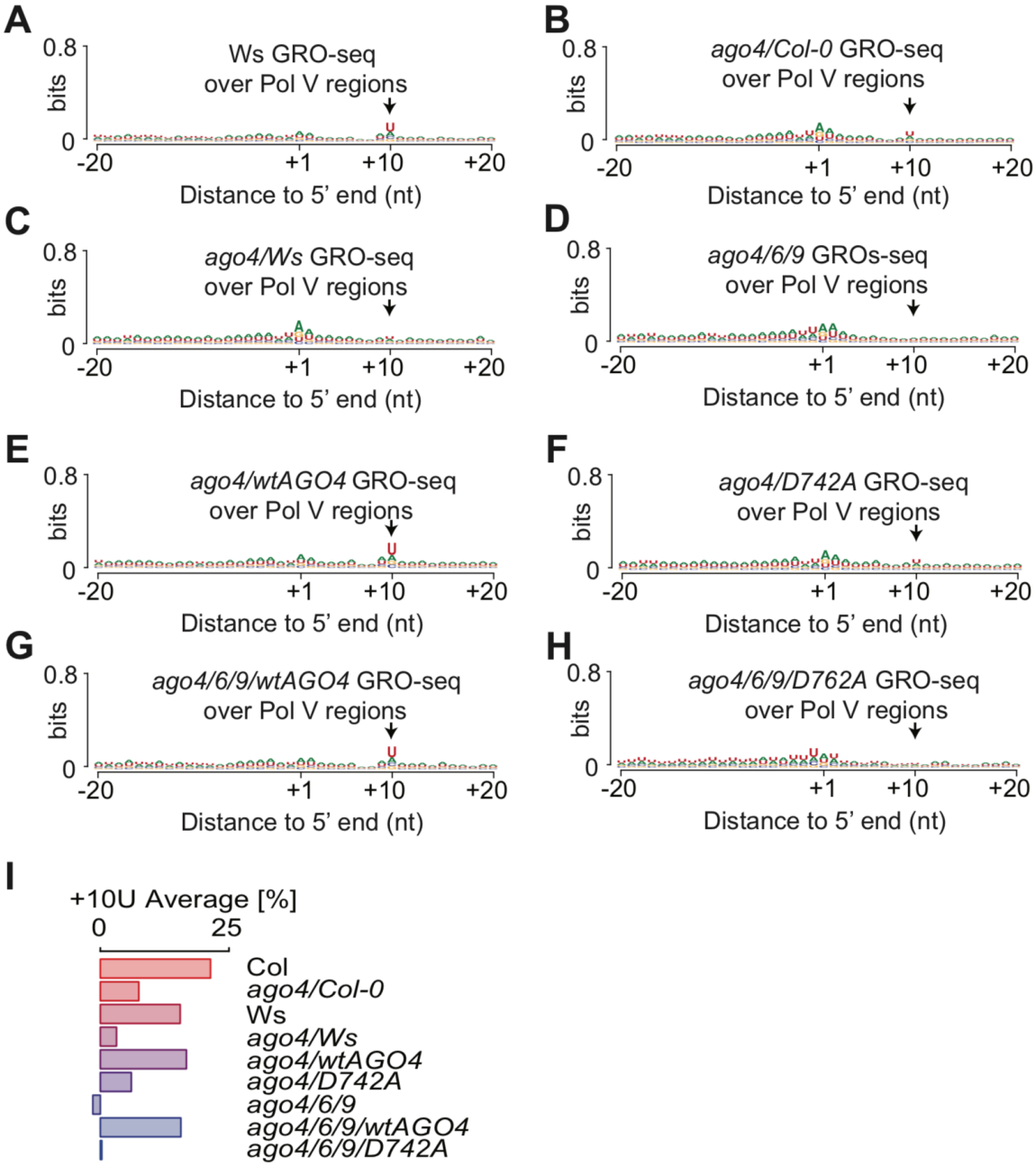

A
Ws GRO-seq
over Pol V regions
B
ago4/Col-0 GRO-seq
over Pol V regions
C
ago4/Ws GRO-seq
over Pol V regions
D
ago4/6/9 GROs-seq
over Pol V regions
E
ago4/wtAGO4 GRO-seq
over Pol V regions
F
ago4/D742A GRO-seq
over Pol V regions
G
ago4/6/9/wtAGO4 GRO-seq
over Pol V regions
H
ago4/6/9/D762A GRO-seq
over Pol V regions
bits
0.8
0
-20
+1
+10
+20
Distance to 5' end (nt)
I
+10U Average [%]
0
25
Col
ago4/Col-0
Ws
ago4/Ws
ago4/wtAGO4
ago4/D742A
ago4/6/9
ago4/6/9/wtAGO4
ago4/6/9/D742A

Figure 1-8

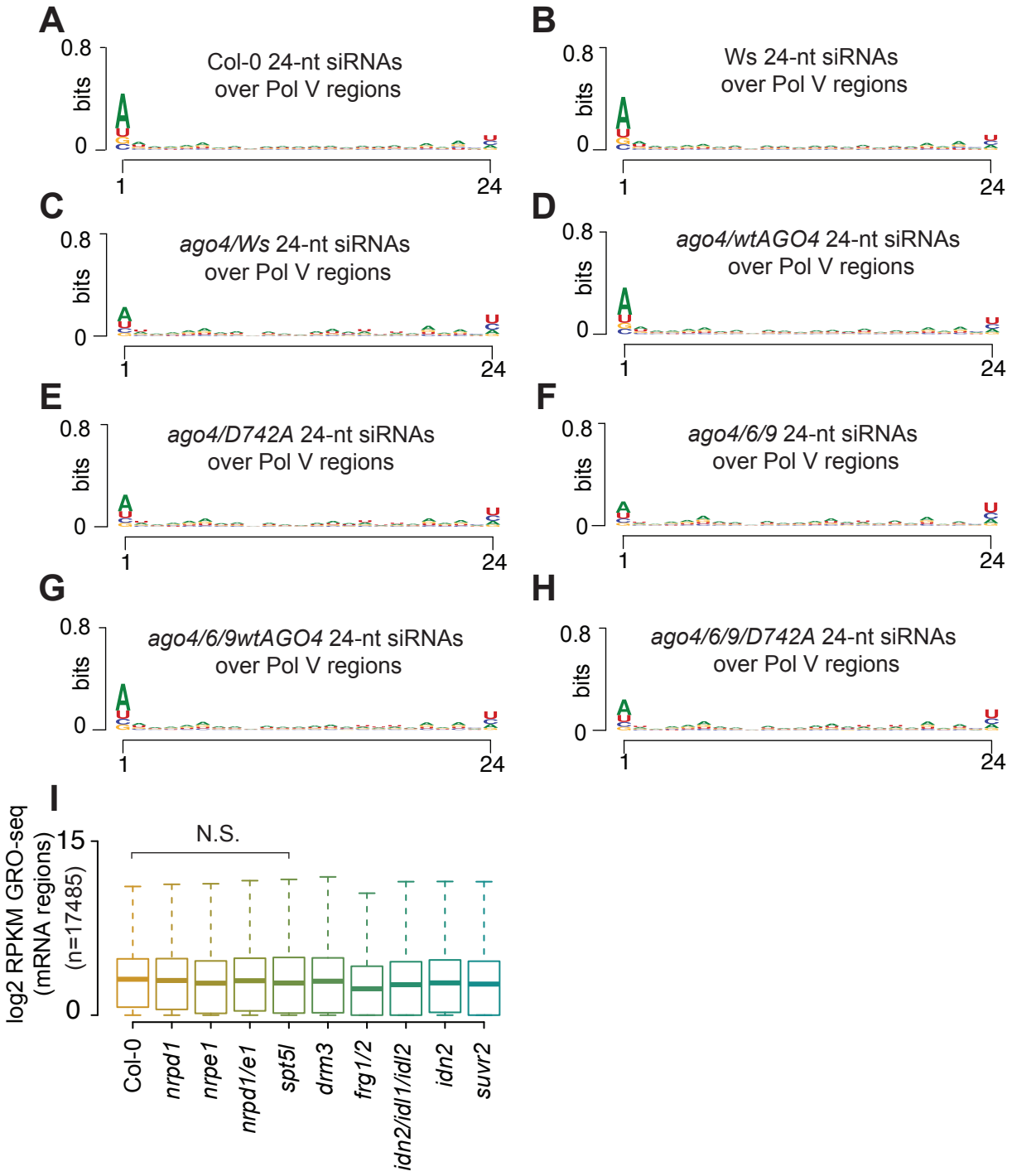

A
Col-0 24-nt siRNAs over Pol V regions
B
Ws 24-nt siRNAs over Pol V regions
C
ago4/Ws 24-nt siRNAs over Pol V regions
D
ago4/wtAGO4 24-nt siRNAs over Pol V regions
E
ago4/D742A 24-nt siRNAs over Pol V regions
F
ago4/6/9 24-nt siRNAs over Pol V regions
G
ago4/6/9wtAGO4 24-nt siRNAs over Pol V regions
H
ago4/6/9/D742A 24-nt siRNAs over Pol V regions
bits
0.8
0
1
24
I
log2 RPKM GRO-seq (mRNA regions) (n=17485)
N.S.
15
0
Col-0
nrpd1
nrpe1
nrpd1/e1
spt5l
drm3
frg1/2
idn2/idl1/idl2
idn2
suvr2

Figure 1-9

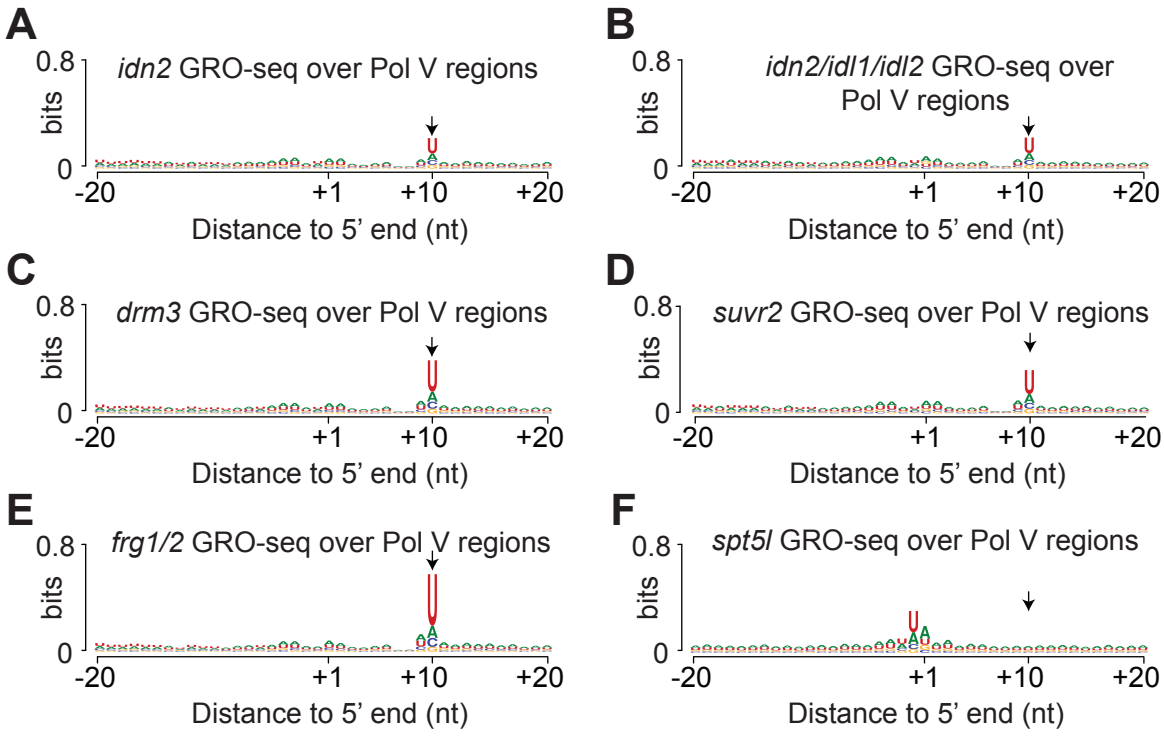
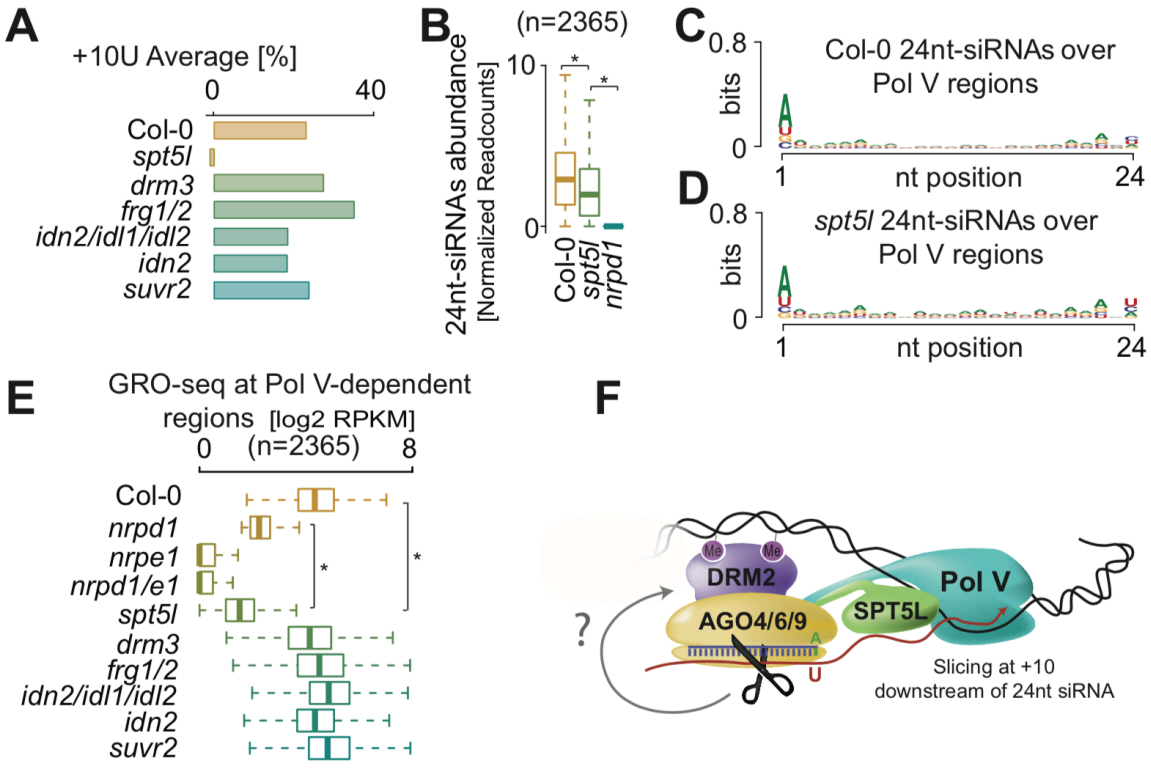

A
idn2 GRO-seq over Pol V regions
B
idn2/idl1/idl2 GRO-seq over Pol V regions
C
drm3 GRO-seq over Pol V regions
D
suvr2 GRO-seq over Pol V regions
E
frg1/2 GRO-seq over Pol V regions
F
spt5l GRO-seq over Pol V regions
bits
0.8
0
-20
+1
+10
+20
Distance to 5' end (nt)

Figure 1-10


A
+10U Average [%]
0
40
Col-0
spt5l
drm3
frg1/2
idn2/idl1/idl2
idn2
suvr2
B
(n=2365)
24nt-siRNAs abundance [Normalized Readcounts]
10
0
Col-0
spt5l
nrpd1
C
Col-0 24nt-siRNAs over Pol V regions
0.8
bits
0
1
nt position
24
D
spt5l 24nt-siRNAs over Pol V regions
0.8
bits
0
1
nt position
24
E
GRO-seq at Pol V-dependent regions [log2 RPKM]
0
(n=2365)
8
Col-0
nrpd1
nrpe1
nrpd1/e1
spt5l
drm3
frg1/2
idn2/idl1/idl2
idn2
suvr2
F
DRM2
Me
Me
Pol V
AGO4/6/9
SPT5L
?
A
U
Slicing at +10
downstream of 24nt siRNA

REFERENCES

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

Ausin, I., Greenberg, M.V.C., Simanshu, D.K., Hale, C.J., Vashisht, A.A., Simon, S.A., Lee, T.-F., Feng, S., Española, S.D., Meyers, B.C., et al. (2012). INVOLVED IN DE NOVO 2-containing complex involved in RNA-directed DNA methylation in Arabidopsis. Proc. Natl. Acad. Sci. U.S.a. *109*, 8374–8381.

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

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

Blevins, T., Podicheti, R., Mishra, V., Marasco, M., Tang, H., and Pikaard, C.S. (2015). Identification of Pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding de novo DNA methylation in Arabidopsis. Elife *4*, e09591.

Böhmdorfer, G., Sethuraman, S., Rowley, M.J., Krzyszton, M., Rothi, M.H., Bouzit, L., and Wierzbicki, A.T. (2016). Long non-coding RNA produced by RNA polymerase V determines boundaries of heterochromatin. Elife *5*, 1325.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell *128*, 1089–1103.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

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.

Eun, C., Lorković, Z.J., Naumann, U., Long, Q., Havecker, E.R., Simon, S.A., Meyers, B.C., Matzke, A.J.M., and Matzke, M. (2011). AGO6 functions in RNA-mediated transcriptional gene silencing in shoot and root meristems in Arabidopsis thaliana. PLoS ONE *6*, e25730.

Greenberg, M.V.C., Ausin, I., Chan, S.W.-L., Cokus, S.J., Cuperus, J.T., Feng, S., Law, J.A., Chu, C., Pellegrini, M., Carrington, J.C., et al. (2011). Identification of genes required for de novo DNA methylation in Arabidopsis. Epigenetics *6*, 344–354.

Groth, M., Stroud, H., Feng, S., Greenberg, M.V.C., Vashisht, A.A., Wohlschlegel, J.A., Jacobsen, S.E., and Ausin, I. (2014). SNF2 chromatin remodeler-family proteins FRG1 and -2

are required for RNA-directed DNA methylation. Proc. Natl. Acad. Sci. U.S.a. *111*, 17666–17671.

Gunawardane, L.S., Saito, K., Nishida, K.M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., and Siomi, M.C. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. Science *315*, 1587–1590.

Haag, J.R., Ream, T.S., Marasco, M., Nicora, C.D., Norbeck, A.D., Pasa-Tolic, L., and Pikaard, C.S. (2012). In vitro transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. Molecular Cell *48*, 811–818.

Han, Y.-F., Dou, K., Ma, Z.-Y., Zhang, S.-W., Huang, H.-W., Li, L., Cai, T., Chen, S., Zhu, J.-K., and He, X.-J. (2014). SUVR2 is involved in transcriptional gene silencing by associating with SNF2-related chromatin-remodeling proteins in Arabidopsis. Cell Res. *24*, 1445–1465.

Havecker, E.R., Wallbridge, L.M., Hardcastle, T.J., Bush, M.S., Kelly, K.A., Dunn, R.M., Schwach, F., Doonan, J.H., and Baulcombe, D.C. (2010). The Arabidopsis RNA-directed DNA methylation argonautes functionally diverge based on their expression and interaction with target loci. The Plant Cell *22*, 321–334.

He, X.-J., Hsu, Y.-F., Wierzbicki, A.T., Pontes, O., Pikaard, C.S., Liu, H.-L., 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.

Herr, A.J., Jensen, M.B., Dalmay, T., and Baulcombe, D.C. (2005). RNA polymerase IV directs silencing of endogenous DNA. Science *308*, 118–120.

Hetzel, J., Duttke, S.H., Benner, C., and Chory, J. (2016). Nascent RNA sequencing reveals distinct features in plant transcription. Proc. Natl. Acad. Sci. U.S.a. *113*, 12316–12321.

Huang, L., Jones, A.M.E., Searle, I., Patel, K., Vogler, H., Hubner, N.C., and Baulcombe, D.C. (2009). An atypical RNA polymerase involved in RNA silencing shares small subunits with RNA polymerase II. Nat. Struct. Mol. Biol. *16*, 91–93.

Johnson, L.M., Du, J., Hale, C.J., Bischof, S., Feng, S., Chodavarapu, R.K., Zhong, X., Marson, G., Pellegrini, M., Segal, D.J., et al. (2014). SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. Nature *507*, 124–128.

Lahmy, S., Pontier, D., Bies-Etheve, N., Laudié, M., Feng, S., Jobet, E., Hale, C.J., Cooke, R., Hakimi, M.-A., Angelov, D., et al. (2016). Evidence for ARGONAUTE4-DNA interactions in RNA-directed DNA methylation in plants. Genes Dev. *30*, 2565–2570.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics *11*, 204–220.

Li, C.F., Pontes, O., El-Shami, M., Henderson, I.R., Bernatavichute, Y.V., Chan, S.W.-L., Lagrange, T., Pikaard, C.S., and Jacobsen, S.E. (2006). An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in Arabidopsis thaliana. Cell *126*, 93–106.

Li, S., Vandivier, L.E., Tu, B., Gao, L., Won, S.Y., Li, S., Zheng, B., Gregory, B.D., and Chen, X. (2015). Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in Arabidopsis reveals features and regulation of siRNA biogenesis. Genome Res. *25*, 235–245.

Mi, S., Cai, T., hu, Y., Chen, Y., Hodges, E., Fangrui, N., Liang, W., Shan, L., Huanyu, Z., Chengzu, L., et al. (2008). Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. Cell *133*, 116–127.

Noma, K.-I., Sugiyama, T., Cam, H., Verdel, A., Zofall, M., Jia, S., Moazed, D., and Grewal, S.I.S. (2004). RITS acts in cis to promote RNA interference-mediated transcriptional and post-transcriptional silencing. Nature Genetics *36*, 1174–1180.

Pontier, D., Yahubyan, G., Vega, D., Bulski, A., Saez-Vasquez, J., Hakimi, M.-A., Lerbs-Mache, S., Colot, V., and Lagrange, T. (2005). Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. Genes Dev. *19*, 2030–2040.

Qi, Y., Denli, A.M., and Hannon, G.J. (2005). Biochemical specialization within Arabidopsis RNA silencing pathways. Molecular Cell *19*, 421–428.

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

Ream, T.S., Haag, J.R., Wierzbicki, A.T., Nicora, C.D., Norbeck, A.D., Zhu, J.-K., Hagen, G., Guilfoyle, T.J., Pasa-Tolic, L., and Pikaard, C.S. (2009). Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. Molecular Cell *33*, 192–203.

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.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics *15*, 284.

Shimada, Y., Mohn, F., and Bühler, M. (2016). The RNA-induced transcriptional silencing complex targets chromatin exclusively via interacting with nascent transcripts. Genes Dev. *30*, 2571–2580.

Smale, S.T., and Kadonaga, J.T. (2003). The RNA polymerase II core promoter. Annu. Rev. Biochem. *72*, 449–479.

Sollner-Webb, B., and Reeder, R.H. (1979). The nucleotide sequence of the initiation and termination sites for ribosomal RNA transcription in X. laevis. Cell *18*, 485–499.

Stroud, H., Greenberg, M.V.C., Feng, S., Bernatavichute, Y.V., and Jacobsen, S.E. (2013). Comprehensive Analysis of Silencing Mutants Reveals Complex Regulation of the *Arabidopsis* Methylome. Cell *152*, 352–364.

Vo Ngoc, L., Cassidy, C.J., Huang, C.Y., Duttke, S.H.C., and Kadonaga, J.T. (2017). The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. Genes Dev. *31*, 6–11.

Wang, F., and Axtell, M.J. (2016). AGO4 is specifically required for heterochromatic siRNA accumulation at Pol V-dependent loci in Arabidopsis thaliana. The Plant Journal.

Wang, H., Zhang, X., Liu, J., Kiba, T., Woo, J., Ojo, T., Hafner, M., Tuschl, T., Chua, N.-H., and Wang, X.-J. (2011). Deep sequencing of small RNAs specifically associated with Arabidopsis AGO1 and AGO4 uncovers new AGO functions. The Plant Journal *67*, 292–304.

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. Nature Genetics *41*, 630–634.

Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D., Jacobsen, S.E., and Carrington, J.C. (2004). Genetic and functional diversification of small RNA pathways in plants. PLoS Biol. *2*, E104.

Zecherle, G.N., Whelen, S., and Hall, B.D. (1996). Purines are required at the 5' ends of newly initiated RNAs for optimal RNA polymerase III gene expression. Mol. Cell. Biol. *16*, 5801–5810.

Zemach, A., Kim, M.Y., Hsieh, P.-H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L., and Zilberman, D. (2013). The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell *153*, 193–205.

Zhai, J., Bischof, S., Wang, H., Feng, S., Lee, T.-F., Teng, C., Chen, X., Park, S.Y., Liu, L., Gallego-Bartolome, J., et al. (2015). A One Precursor One siRNA Model for Pol IV-Dependent siRNA Biogenesis. Cell *163*, 445–455.

Zhang, C.-J., Ning, Y.-Q., Zhang, S.-W., Chen, Q., Shao, C.-R., Guo, Y.-W., Zhou, J.-X., Li, L., Chen, S., and He, X.-J. (2012). IDN2 and its paralogs form a complex required for RNA-directed DNA methylation. PLoS Genet. *8*, e1002693.

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

Zhong, X., Du, J., Hale, C.J., Gallego-Bartolome, J., Feng, S., Vashisht, A.A., Chory, J., Wohlschlegel, J.A., Patel, D.J., and Jacobsen, S.E. (2014). Molecular mechanism of action of plant DRM de novo DNA methyltransferases. Cell *157*, 1050–1060.

Zhong, X., Hale, C.J., Law, J.A., Johnson, L.M., Feng, S., Tu, A., and Jacobsen, S.E. (2012). DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. Nat. Struct. Mol. Biol. *19*, 870–875.

Zhong, X., Hale, C.J., Nguyen, M., Ausin, I., Groth, M., Hetzel, J., Vashisht, A.A., Henderson, I.R., Wohlschlegel, J.A., and Jacobsen, S.E. (2015). Domains rearranged methyltransferase3 controls DNA methylation and regulates RNA polymerase V transcript abundance in Arabidopsis. Proc. Natl. Acad. Sci. U.S.a. *112*, 911–916.

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

Zofall, M., Yamanaka, S., Reyes-Turcu, F.E., Zhang, K., Rubin, C., and Grewal, S.I.S. (2012). RNA elimination machinery targeting meiotic mRNAs promotes facultative heterochromatin formation. Science *335*, 96–100.

CHAPTER 2

Co-targeting RNA Polymerases IV and V promotes efficient de novo DNA methylation in Arabidopsis.

Contributions

Conceptualization, J.G.B., W.L., and S.E.J.; Methodology, J.G.B., W.L. and S.F.; Investigation, J.G.B., W.L., H.Y.K , S.F., C.J.H., S.Y.P.; Resources, J.C.; Writing, J.G.B., W.L., and S.E.J..

ABSTRACT

The RNA-directed DNA methylation (RdDM) pathway in plants controls gene expression and genome integrity via cytosine methylation. The ability to site-specifically manipulate methylation via RdDM would shed light on the mechanisms and applications of epigenetics to control gene expression. Here, we identified diverse RdDM proteins that are sufficient to target *de novo* methylation and silencing in Arabidopsis when tethered to an artificial zinc finger (ZF-RdDM). We studied their order of action within the RdDM pathway by testing their ability to target methylation in different mutant backgrounds. Also, we evaluated ectopic siRNA biogenesis, RNA Polymerase V (Pol V) recruitment, DNA methylation, and gene expression at thousands of ZF-RdDM binding sites. We found that simultaneously recruiting both arms of the RdDM pathway, siRNA biogenesis and Pol V recruitment, dramatically enhanced targeted methylation. This work defines how RdDM components establish heterochromatin and enables site-specific, epigenetic gene regulation via targeted DNA methylation in plants.

INTRODUCTION

Cytosine DNA methylation controls diverse processes in many eukaryotes, including gene expression, genome organization and integrity, chromatin architecture, and cell specification. Epigenome manipulation bears exciting possibilities for basic research and applied crop engineering, such as generating DNA methylation-based epialleles of important trait genes. Improving our knowledge of the pathways that trigger methylation in plants is critical, not only for understanding how methylation is established and controlled, which in turn regulates gene expression and cell function, but also for improving the current DNA methylation-targeting toolset.

In plants, cytosines can be methylated within three different contexts: CG, CHG, or CHH (where H is A, T, C). DNA methylation is enriched in heterochromatic regions, where it plays an important role in silencing transposable elements (TEs) and genes(Law and Jacobsen, 2010). Methylation establishment requires the plant-specific RNA-directed DNA methylation (RdDM) pathway, which acts via the *de novo* DNA methyltransferase DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) (Cao and Jacobsen, 2002; Matzke et al., 2015). The RdDM pathway can be divided into two major arms, each dependent on a plant-specific RNA polymerase. RNA polymerase IV (Pol IV) is required to generate siRNAs from target loci, and RNA polymerase V (Pol V) is required to generate non-coding scaffold RNAs that recruit the DNA methylation machinery (Matzke et al., 2015)(Figure 2-1A).

Pol IV generates transcripts (p4-RNAs) that are thought to be converted into double stranded RNAs (dsRNA) by the RNA-dependent RNA polymerase 2 (RDR2) and subsequently processed into 24-nt siRNAs by DICER-LIKE3 (DCL3) (Herr et al., 2005; Li et al., 2015; Onodera et al., 2005; Xie et al., 2004) (Zhai et al., 2015). In the absence of DCL3, other DICER-

LIKE proteins, DCL1, DCL2, and DCL4 can process p4-RNAs into 21-nt or 22-nt siRNAs that trigger *de novo* methylation by RdDM (Bond and Baulcombe, 2015; Henderson et al., 2006). Mutations in *NRPD1*, the catalytic subunit of Pol IV, lead to a virtually complete loss of 24-nt siRNAs genome wide (Mosher et al., 2008; Zhang et al., 2007) and a strong loss of DNA methylation at RdDM sites (Stroud et al., 2013). Pol IV accessory proteins include the poorly understood CLASSY SWI2/SNF2 chromatin remodeler (Smith et al., 2007), and SAWADEE HOMEODOMAIN HOMOLOG 1 (SHH1), which binds to the repressive histone mark H3K9 methylation associated with DNA methylation and is required for Pol IV recruitment at a subset of RdDM sites (Law et al., 2014). siRNAs for RdDM can be alternatively generated from Pol II transcripts in "non-canonical RdDM" (Cuerda-Gil and Slotkin, 2016). Pol II transcripts are processed into dsRNAs by RNA-DEPENDENT POLYMERASE 6 (RDR6) and subsequently cleaved into siRNAs by DCL2/DCL4 or DCL3 (Allen et al., 2005; Marí-Ordóñez et al., 2013) (Nuthikattu et al., 2013). siRNAs are loaded into ARGONAUTE 4 (AGO4) or its homologs AGO6 and AGO9 (Matzke et al., 2015).

Pol V, together with a number of accessory proteins, generates longer non-coding RNAs at target loci (Böhmdorfer et al., 2016; Liu et al., 2018){Matzke:2015ez}, (Figure 2-1A). The DNA methylation reader proteins SU(VAR)3-9 homologues SUVH2 and SUVH9 recruit Pol V to pre-existing DNA methylation (Johnson et al., 2014; Liu et al., 2014), and the DDR complex, consisting of RNA-DIRECTED DNA METHYLATION 1 (RDM1), DEFECTIVE IN MERISTEM SILENCING 3 (DMS3), and DEFECTIVE IN RNA-DIRECTED DNA METHYLATION 1 (DRD1), is required globally for Pol V occupancy on chromatin (Law and Jacobsen, 2010; Zhong et al., 2012). siRNA-loaded AGO4 interacts with Pol V through its C-terminal domain (El-Shami et al., 2007) (Li et al., 2006) and it is thought that homologous

pairing between siRNAs and Pol V RNAs leads to AGO4-mediated recruitment of DRM2, although many aspects of these molecular details remain unknown (Zhong et al., 2014).

Other factors implicated in RdDM include the Microrchidia (MORC) ATPases, MORC1 and MORC6, that act as heterodimers to mediate gene silencing (Moissiard et al., 2014) (Moissiard et al., 2012). Unlike RdDM mutants, *morc* mutants show reactivation of many methylated regions without a corresponding loss of DNA methylation, and thus appear to act primarily downstream of DNA methylation at most loci. However, a small number of RdDM loci are transcriptionally derepressed in *morc* mutants, and, at those loci only, *morc* mutants show a loss of DNA methylation (Harris et al., 2016). In addition, different studies have described physical interactions of MORC1 and MORC6 proteins with the RdDM proteins SUVH2, SUVH9, IDN2, and DMS3 (Jing et al., 2016; Liu et al., 2014; 2016; Lorković et al., 2012), although the functional relevance of these interactions and the specific role of MORCs in RdDM remain unclear.

The imprinted gene *FLOWERING WAGENINGEN (FWA)* (Soppe et al., 2000) is repressed by promoter DNA methylation in wild-type plants, except in the central cell and endosperm where maternal allele-specific demethylation triggers its expression (Kinoshita et al., 2004). *FWA* is aberrantly expressed and demethylated in DNA methylation-deficient mutants (Soppe et al., 2000), creating *fwa* epialleles that are heritably maintained in crosses between the DNA methylation-deficient mutants and wild-type Col-0 plant (Johnson et al., 2014). Ectopic expression from *fwa* epialleles disrupts the flowering time master regulator *FLOWERING LOCUS T* (*FT*), causing a strong late flowering phenotype (Ikeda et al., 2007). We previously showed that tethering the RdDM component SUVH9 to an artificial zinc finger that targets the *FWA* promoter (ZF108) could induce DNA methylation at an unmethylated *fwa* epiallele,

restoring heritable *FWA* repression and the early flowering phenotype (Johnson et al., 2014). Whether other components of the RdDM pathway are sufficient to trigger *de novo* DNA methylation and silencing is unknown.

Here, we found that ZF108 fusions with various components of the RdDM pathway can promote DNA methylation at *FWA*, as well as at thousands of additional loci targeted by ZF108. Importantly, co-targeting of Pol V and Pol IV synergistically enhanced target methylation, revealing that siRNA biogenesis and recruitment of the DNA methylation machinery to target loci are largely independent and both important for efficient methylation and silencing. Further, we utilized our collection of ZF108 fusions to dissect the primary role and hierarchy of action of RdDM pathway components, providing unprecedented mechanistic insight into heterochromatin formation. Thus, our findings provide an efficient approach to study and manipulate DNA methylation specifically at targeted loci in Arabidopsis.

RESULTS

Novel zinc finger fusion proteins that promote *FWA* methylation.

We utilized the targeting approach described previously (Johnson et al., 2014) to test RdDM components for their ability to promote *FWA* methylation when fused to ZF108. Ten different fusion proteins restored an early flowering phenotype and *FWA* DNA methylation in T1 transformed *fwa* plants, including components of the first (SHH1, NRPD1, RDR2, Figure 2-1) and second (SUVH9, RDM1, DMS3, Figure S2) arms of the RdDM pathway, MORC6 and MORC1 (Figure S3), and the catalytic domain of the tobacco DRM2 DNA methyltransferase (Figure S4). We confirmed the early flowering phenotype observed in T1 plants, and ruled out the possibility that early flowering was caused by plant stress or other causes by systematically scoring the flowering time of T2 plants descended from the four earliest T1 plants (Figures 1-4).

We also confirmed DNA methylation at the *FWA* promoter in representative T2 lines for all of the fusions that showed early flowering. Specifically, we amplified three regions of the *FWA* promoter from bisulfite-treated DNA using primers that incorporate Illumina adaptors, followed by multiplexed Illumina sequencing (BS-PCR-seq) (Figures 1-4).

To study the hierarchy of action of RdDM components in *de novo* methylation, we assessed the function of each zinc finger fusion in distinct RdDM mutant backgrounds. First, we established a collection of RdDM mutants in the unmethylated *fwa* background, and then transformed these plants with various fusion proteins (Figures 1-4, S1-S4). Combining gain-of-function ZF108 fusions with loss-of-function RdDM mutations offered a unique approach to interrogate the hierarchy of action of RdDM proteins in *de novo* heterochromatin formation.

Ectopic Methylation induced by RdDM "Arm 1": siRNA biogenesis.

Targeting by NRPD1

Ectopic expression of ZF108 fused to the Pol IV subunit NRPD1 caused early flowering and *FWA* promoter methylation in the *fwa* background (Figure 2-2A,B, Figure 2-1B). Loss of *shh1* did not block NRPD1-ZF108 targeted *FWA* methylation and silencing, consistent with SHH1 acting upstream of Pol IV recruitment (Law et al., 2014) (Figure 2-2A,B, Figure 2-1B). Similar results were obtained in the *clsy1* mutant background (Figure 2-2A,B, Figure 2-1B), indicating that this chromatin remodeling protein is dispensable when Pol IV is artificially targeted to chromatin. However, NRPD1-ZF108 failed to trigger early flowering in the *rdr2* mutant, consistent with previous observations that RDR2 is needed for p4-RNA production (Blevins et al., 2015; Li et al., 2015; Zhai et al., 2015) (Figure 2-2A, Figure 2-1B), and consistent with its proposed role downstream of Pol IV in the production of dsRNAs for siRNA biogenesis.

NRPD1-ZF108 could also induce *fwa* methylation in a *dcl3* single mutant, as well as in the *dcl2 dcl4* double mutant, indicating that different DCLs can process p4-RNAs into siRNAs that are competent for RdDM (Figure 2-2A,B, Figure 2-1B). This is consistent with the observation that DCL2 and DCL4 can produce 21-22nt siRNAs in non-canonical Pol II-RDR6 RdDM (Cuerda-Gil and Slotkin, 2016). Moreover, we observed NRPD1-dependent *FWA* methylation in *dcl2 dcl3 dcl4* triple mutant plants. However, contrary to all other mutant backgrounds, we only observed early flowering plants in the T2, but not the T1 generation (Figure 2-2A,B, Figure 2-1B), indicating that methylation is less efficient compared to single or double *dcl* mutants. A similar observation was reported using VIGS to target methylation to *FWA* (Bond and Baulcombe, 2015) and suggests that DCL1 is capable of producing siRNAs to mediate RdDM. This result is also consistent with the observation that *dcl2 dcl3 dcl4* mutant has only partial RdDM defects whereas *nrpd1* mutants completely eliminate RdDM methylation (Stroud et al., 2013).

To analyze siRNA biogenesis in different DCL mutant backgrounds, we performed small RNA sequencing (sRNA-seq) in T2 lines. We did not detect siRNAs derived from *FWA* in the *fwa* epiallele, indicating that NRPD1-ZF108 can initiate *FWA* methylation without preexisting siRNAs (Figure 2-2C, Figure 2-1C). NRPD1-ZF108 triggered the production of all forms of siRNAs at *FWA*, though mostly 24-nt siRNAs (Figure 2-2C, Figure 2-1C). siRNAs were observed a few hundred nucleotides from the ZF108 binding sites (Figure 2-1C), consistent with the observed DNA methylation (Figure 2-2B). The *dcl3* mutant displayed reduced 24-nt siRNAs upon expression of NRPD1-ZF108 relative to the wild-type and to the *dcl2 dcl4* double mutant, suggesting that most NRPD1-ZF108-dependent siRNAs at *FWA* are processed by DCL3 (Figure 2-2C, Figure 2-1C). Low levels of siRNA of different sizes were generated from *FWA* in *dcl2*

dcl3 dcl4, suggesting that DCL1, the only remaining dicer enzyme, can cleave NRPD1-ZF108-dependent p4-RNAs (Figure 2-2C, Figure 2-1C).

NRPD1-ZF108 triggered early flowering and methylation in an *ago4* mutant, although non-CG methylation levels were dramatically reduced (Figure 2-2A,B, Figure 2-1B). However, NRPD1-ZF108 did not trigger *FWA* silencing in an *ago4 ago6 ago9* triple mutant, indicating that AGO6 and/or AGO9, can substitute for the function of AGO4 (Figure 2-2A, Figure 2-1B). Importantly, NRPD1-ZF108 failed to trigger early flowering in the Pol V mutant *nrpe1,* or in *drm1 drm2* double mutants (*DRM1* is a lowly expressed homolog of *DRM2*), consistent with a requirement for these RdDM components downstream of siRNA biogenesis (Figure 2-2A, Figure 2-1B). Lastly, *cmt3* did not block *FWA* methylation induced by NRPD1-ZF108 (Figure 2-2A,B, Figure 2-1B), consistent with previous results showing that CMT3 is not required for *de novo* methylation (Chan et al., 2004).

Targeting by RDR2

RDR2-ZF108 induced DNA methylation and silencing of *FWA,* with a methylation pattern similar to that induced by NRPD1-ZF108 (Figure 2-2B,D, Figure 2-1D). However, RDR2-ZF108 failed to trigger early flowering in the *nrpd1* mutant (Figure 2-2D, Figure 2-1D), consistent with the strong association of RDR2 with the Pol IV complex and a role for RDR2 in converting p4-RNAs into dsRNA. RDR2-ZF108 behaved similarly to NRPD1-ZF108 in all other tested mutant backgrounds, except for its ability to induce *FWA* silencing in the *rdr2* mutant, as predicted (Figure 2-2B,D, Figure 2-1D).

Targeting by SHH1

SHH1-ZF108 could trigger methylation and silencing of *FWA*, though somewhat less efficiently than NRPD1-ZF108 or RDR2-ZF108 (Figure 2-2B,E, Figure 2-1E). As expected for a Pol IV

recruitment factor, SHH1-ZF108 could not induce *FWA* silencing in *nrpd1* or *rdr2* mutants, or in *nrpe1* and *drm1 drm2* mutants (Figure 2-2E, Figure 2-1E). Interestingly, SHH1 could induce *FWA* methylation in *clsy1*, suggesting that SHH1 can act independently of this SNF2 family chromatin remodeling factor (Figure 2-2B,E, Figure 2-1E). Contrary to NRPD1-ZF108 and RDR2-ZF108, SHH1-ZF108-targeted methylation was concentrated in a smaller region flanking the ZF108 target sequence (Figure 2-2B). However, methylation was more extensive when SHH1-ZF108 was targeted in an *shh1* mutant (Figure 2-2B), which correlated with an enhanced frequency of early flowering plants in the T1 generation in *shh1* (Figure 2-1E). This finding suggests that endogenous SHH1 competes with SHH1-ZF108 for Pol IV targeting.

Ectopic methylation induced by RdDM "Arm 2": Pol V transcription and methylation targeting.

Targeting by SUVH9

Pol V recruitment to chromatin is essential for RdDM and mutants defective in Pol V recruitment, such as the *suvh2 suvh9* double mutant or the *dms3, drd1, or rdm1* (DDR complex) single mutants, show a complete loss of RdDM (Johnson et al., 2014) (Zhong et al., 2012). The ZF108-SUVH9 fusion could target *FWA* silencing in wild-type plants, but not in any of the DDR complex single mutants, *nrpe1,* or *drm1 drm2*, positioning SUVH9 upstream of DDR/Pol V (Figure 2-3A, B, Figure 2-4A). Although SUVH9 can interact with MORC6 (Jing et al., 2016) (Liu et al., 2014), it was able to efficiently trigger methylation in a *morc6* mutant (Figure 2-3A, B, Figure 2-4A), indicating that SUVH9 can act independently of this factor.

Targeting by DMS3

DMS3 was the most potent RdDM-ZF108 fusion, triggering early flowering and DNA methylation at a consistently high frequency (Figure 2-3B,C, Figure 2-4B). Its activity was

blocked in the *nrpe1* mutant (Figure 2-3C, Figure 2-4B), consistent with the role of DMS3 as a component of the DDR complex needed for Pol V recruitment (Zhong et al., 2012). DMS3 was efficient in targeting methylation in the *suvh2 suvh9* double mutant and in the *morc6* mutant, positioning DMS3 downstream of these components (Figure 2-3B, C, Figure 2-4B).

DMS3 was unable to target methylation in plants containing a mutation in another DDR component, *DRD1* (Figure 2-3C, Figure 2-4B). However, it could target methylation (although less efficiently) in plants containing a mutation in the third DDR component, *RDM1* (Figure 2-3B,C, Figure 2-4B). One interpretation of this result is that RDM1 functions in the recruitment or stabilization of the DDR complex to chromatin, a function that can be replaced by artificially tethering DMS3 to chromatin.

Unexpectedly, DMS3 caused early flowering and methylation in the *nrpd1* mutant, suggesting that successful *de novo* methylation could be established in the absence of siRNAs (Figure 2-3B,C, Figure 2-4B). To test this hypothesis, we profiled small RNAs in lines expressing DMS3-ZF108 in wild-type or *nrpd1* mutant backgrounds (Figure 2-3D, Figure 2-4C). As mentioned above, *FWA* locus siRNAs were not present at the unmethylated *fwa* epiallele, indicating that pre-existing *FWA* siRNAs were not required for DMS3-ZF108 to target methylation. We observed high levels of 24-nt siRNAs, as well as some 21-nt and 22-nt siRNAs, over the ZF108 binding site in *fwa* plants expressing DMS3-ZF108, but not in DMS3-ZF108 *nrpd1* mutant plants (Figure 2-3D, Figure 2-4C). However, we did observe very low levels of short RNAs in the 21-24 nt size range in the *nrpd1* background. Since *nrpd1* is a null Pol IV mutant, these short RNAs may not be siRNAs and may instead be Pol V transcripts or degradation products. Consistent with this idea, these short RNAs lacked an enrichment for A at the 5' end, a known characteristic of RdDM associated siRNAs (Mi et al., 2008) (Figure 2-4D).

Furthermore, DMS3-ZF108 also targeted *FWA* methylation in an *rdr1 rdr6* double mutant and an *rdr1 rdr2 rdr6* triple mutant (backgrounds lacking the RdDM factor RDR2 and other related RDR genes), reinforcing the idea that DMS3 may induce methylation in the absence of siRNAs (Figure 2-3B,C, Figure 2-4D). While it seems unlikely, we cannot rule out however that trace levels of siRNAs from some unknown source are involved in the process.

Given that both SUVH9 and DMS3 are involved in targeting Pol V to chromatin (Johnson et al., 2014), it is somewhat surprising that ZF108-SUVH9 induced *FWA* silencing and methylation much less efficiently (only 2 early flowering plants within the four T2 populations measured) than DMS3-ZF108 in the *nprd1* mutant background (Figure 2-3A). However, ZF108-SUVH9 also induced *FWA* methylation less efficiently than DMS3-ZF108 in the wild-type (Figure 2-3A, Figure 2-4A), suggesting that Pol IV siRNA biogenesis is needed for efficient *FWA* methylation when Pol V targeting is limited.

As expected, DMS3-ZF108 failed to target methylation in *drm1 drm2* double mutant (Figure 2-3B) but, surprisingly, a number of independent transgenic lines exhibited a mild early flowering phenotype (Figure 2-3C, Figure 2-4B), suggesting that DMS3-ZF108 can suppress *FWA* without inducing DNA methylation. To confirm this hypothesis, we performed RNA sequencing of three independent early flowering T2 lines and controls. As expected, DMS3-ZF108 caused a complete loss of *FWA* mRNA in the *fwa* (but otherwise wild-type) background (Figure 2-4E). Consistent with the early flowering phenotype, we observed a partial repression of *FWA* by DMS3-ZF108 in the *drm1 drm2* background (Figure 2-4E), confirming that DMS3-ZF108 can repress expression in a DNA methylation-independent manner. One possible mechanism to explain this result is that DMS3-ZF108 might be so efficient at recruiting Pol V that this could interfere with Pol II transcription. We therefore performed ChIP-seq of Pol V in

DMS3-ZF108 plants. Since Pol V is normally recruited by DNA methylation, we also profiled DMS3-ZF108 *drm1 drm2* plants to ensure that Pol V recruitment was not a secondary consequence of DNA methylation targeting. Indeed, we observed robust recruitment of Pol V in both backgrounds, suggesting that DMS3-ZF108 can target Pol V recruitment even in the absence of DNA methylation (Figure 2-3E). As a comparison, we also profiled Pol V in ZF108-SUVH9 plants. ZF108-SUVH9 also recruited Pol V to *FWA* in both wild-type and *drm1 drm2* mutants, but did so less efficiently and in a narrower region than in DMS3-ZF108 plants, even though parallel ChIP-seq analysis of the DMS3-ZF108 and ZF108-SUVH9 fusion proteins showed similar signals over *FWA* (Figure 2-3E). These results show that DMS3 is a more powerful recruiter of Pol V than SUVH9, which might explain why DMS3-ZF108 can cause repression of *FWA* expression in a DNA methylation-independent manner.

DMS3-ZF108 targeted methylation was severely reduced in *ago4* and absent in the *ago4 ago6 ago9* triple mutant (Figure 2-3B,C, Figure 2-4B), indicating that an ARGONAUTE of the AGO4/6/9 clade is crucial for DMS3-dependent targeted methylation. This result, coupled with the fact that DMS3 appears to target methylation in a siRNA-independent manner, suggests that unloaded AGO protein may be sufficient to physically "bridge" Pol V and DRM2. This would be consistent with the known physical interactions between AGO4 and Pol V (El-Shami et al., 2007) (Li et al., 2006), and between AGO4 and DRM2(Zhong et al., 2014).

Targeting by RDM1

RDM1-ZF108 caused early flowering although with much lower efficiency than DMS3 (Figure 2-3F, Figure 2-4F). Consistent with the DMS3 results, RDM1 induced *FWA* methylation in *nrpd1*, *suvh2 suvh9*, and *morc6* mutants (Figure 2-3B,F, Figure 2-4F), further supporting the notion that Pol V recruitment through the DDR complex can be sufficient to initiate RdDM.

Interestingly, RDM1-ZF108 was more efficient when transformed into an *rdm1* mutant (Figure 2-3F, Figure 2-4F), suggesting that endogenous RDM1 might compete with RDM1-ZF108's ability to recruit/interact with the other DDR components. RDM1 was not able to cause early flowering in *drd1, dms3, nrpe1*, *and drm1 drm2* mutants (Figure 2-3F, Figure 2-4F), indicating that RDM1 is unable to recruit Pol V in the absence of the other DDR complex components.

Targeting by MORC6

MORC6 has been linked to RdDM, although its role is not well understood (Harris et al., 2016) (Jing et al., 2016; Liu et al., 2016) (Lorković et al., 2012). We found that MORC6-ZF108 triggered early flowering and induced *FWA* DNA methylation (Figure 2-5A,B, Figure 2-6A). In addition, MORC6-ZF108 targeted *FWA* methylation in a *nrpd1* mutant, but could not trigger silencing in mutants of the DDR complex, *nrpe1* or *drm1 drm2* (Figure 2-5A,B, Figure 2-6A). These results suggest that MORC6-ZF108 acts upstream of DDR to recruit Pol V activity. To provide more evidence for this hypothesis, we performed Pol V ChIP-seq as well as MORC6-ZF108 ChIP-seq in a wild-type background or a *drm1 drm2* mutant background. MORC6-ZF108 was indeed able to recruit Pol V to *FWA* in both backgrounds confirming that, like DMS3-ZF08 and ZF108-SUVH9, MORC6-ZF108 can recruit Pol V in a DNA methylation-independent manner (Figure 2-3E). MORC6-ZF108 was less efficient at recruiting Pol V than DMS3-ZF108, likely explaining why it did not cause early flowering in a *drm1 drm2* mutant background, unlike DMS3-ZF108 (Figure 2-5A, Figure 2-6A). On the other hand, MORC6-ZF108 was moderately more efficient at recruiting Pol V than ZF108-SUVH9 (Figure 2-3E), explaining why it was more efficient than ZF108-SUVH9 at targeting methylation in the *nrpd1* mutant.

Considering that both MORC6-ZF108 and ZF108-SUVH9 act upstream of DDR/Pol V activity, we tested the ability of MORC6-ZF108 to target methylation in *suvh2 suvh9* and found

that it did so efficiently (Figure 2-5A,B, Figure 2-6A). This result, together with the observation that ZF108-SUVH9 can target methylation in the *morc6* mutant (Figure 2-5A,B, Figure 2-6A), positions these two proteins in parallel pathways that utilize the DDR complex to recruit Pol V and establish DNA methylation.

MORC6 and *MORC1* have similar mutant phenotypes, and MORC6 and MORC1 form stable heteromers (Moissiard et al., 2014). As predicted, MORC1-ZF108 was also able to induce methylation and silencing of *FWA* (Figure 2-5B,C, Figure 2-6B).

Ectopic methylation by the DRM2 catalytic domain.

The last step in RdDM is the recruitment of the *de novo* methyltransferase DRM2 (Matzke et al., 2015). We found that full-length ZF108-DRM2 failed to trigger *FWA* silencing and methylation (data not shown). Therefore, we tested a fragment containing the methyltransferase domain of tobacco DRM2 (NtMTase) that had been previously crystallized (Zhong et al., 2014), reasoning that the N-terminus of DRM2 may contain negative regulatory domains as observed in mammalian DNA methyltransferases (Jeltsch and Jurkowska, 2016). Indeed ZF108-NtMTase induced *FWA* silencing and methylation (Figure 2-7A,B, Figure 2-8). Consistent with previous results, *cmt3* mutant did not affect ZF108-NtMTase activity. Unexpectedly however, ZF108-NtMTase activity greatly reduced in *nrpd1* suggesting that siRNA biogenesis is needed for its full activity (Figure 2-7A,B, Figure 2-8). Additionally, ZF108-NtMTase activity was completely blocked by either the *nrpe1* or *drm1 drm2* mutations (Figure 2-7A, Figure 2-8), suggesting that Pol V activity and endogenous DRM2 activity are critical for methylation establishment or amplification. Although the mechanism for this is unknown, it is possible that unknown chromatin marks targeted by Pol V might be needed for establishment. In addition, the Nt-MTase

fragment lacks the UBA domains present in the DRM2 N-terminus that are predicted to bind ubiquitin, and that have been shown to be important for its full activity (Henderson et al., 2010).

DMS3-ZF108 recruits Pol V to additional genomic sites.

Zinc fingers are rarely highly specific in their binding and we therefore sought to take advantage of "off-target" binding by ZF108 to study DNA methylation targeting at additional sites in the genome. We focused our initial analysis on the most efficient RdDM factor, DMS3-ZF108. The DNA methylation landscape of the *fwa* epimutant is chimeric since it was generated by crossing wild-type Col-0 with *met1* mutant plants. Because this would complicate any analysis of targeted DNA methylation changes, we re-transformed DMS3-ZF108 (with a FLAG tag), as well as a HA-tagged ZF108 control construct, into wild-type Col-0 plants and performed ChIP-seq to identify the locations of ZF108 binding.

DMS3-ZF108 and ZF108 had similar binding patterns and were found not only at *FWA* but also at thousands of additional loci (Figure 2-9A-C), and showed a preference for promoter regions (Figure 2-10A). When we ranked the ChIP-seq peaks based on their signal across the genome, the *FWA* peak ranked first or second in both DMS3-ZF108 and ZF108 control lines (Figure 2-9D); however there were many additional peaks with strong signals. The DMS3-ZF108 ChIP-seq peak intensities also strongly correlated with the presence of the ZF108 binding sequence (Figure 2-10B). A *de novo* cis motif analysis identified a core motif sequence corresponding to the inner zinc finger repeats of ZF108 as the most overrepresented (Figure 2-9E), suggesting that the external two zinc finger repeats do not play a major role in ZF108 binding to chromatin. Despite the fact that the ZF108 inner core motif is highly abundant in the genome, only 27.5% of loci containing this motif were occupied by ZF108 fusions (Figure 2-10C). When we analyzed the genome-wide distribution of the ZF108 motif with respect to the

presence or absence of ZF108 binding, we observed that ZF108 fusions tend to bind the motif when it is present in promoters and they tend to be excluded from motifs present in exons (Figure 2-10D). These differences might be due to differences in chromatin accessibility. Indeed, among the loci that contain a ZF108 binding motif, those bound by DMS3-ZF108 showed a more open chromatin structure as measured by ATAC-seq (Lu et al., 2016) (Figure 2-10E), which indicates that chromatin accessibility could be a major determinant for the ability of ZF108 to bind to its targets.

DMS3 appears to target DNA methylation by recruiting Pol V (Figure 2-3C, Figure 2-5C). To test the efficiency of DMS3-ZF108 in recruiting Pol V at different loci, we performed ChIP-seq of NRPE1, the largest subunit of Pol V. Strikingly, over 90% of DMS3-ZF108 binding sites gained a Pol V peak (Figure 2-9F,G). Moreover, consistent with the ZF108 binding profile, DMS3-dependent Pol V recruitment was more efficient over open chromatin regions like promoters and tended to be excluded from exons (Figure 2-10F,G).

The promiscuous nature of DMS3-ZF108 binding and its high efficiency of Pol V recruitment provide a unique opportunity to study the ability of Pol V to target methylation to thousands of loci. We first examined siRNA production over DMS3-ZF108 binding regions that recruited Pol V (n=9,941, Figure 2-9G). Compared to the high efficiency of recruiting Pol V (92.3%, Figure 2-9G), only 9.8% (n=972) of the Pol V–containing DMS3-ZF108 off-targets showed *de novo* accumulation of 24-nt siRNAs (Figure 2-11A). In addition, the loci in which siRNA production was stimulated by DMS3-ZF108 corresponded to those with the highest levels of Pol V recruitment (Figure 2-11A, Figure 2-12A,B). This suggests that high levels of Pol V recruitment are needed to engage the RdDM pathway and stimulate siRNA production.

To study targeted methylation at these sites, we analyzed whole genome DNA methylation of T2 and T3 plants expressing DMS3-ZF108, ZF108, as well as Col-0 controls. Of the DMS3-ZF108 loci containing Pol V and producing siRNAs, 46% were hypermethylated (Figure 2-11B). In addition, these hypermethylated sites showed a much higher accumulation of siRNAs (Figure 2-11C), and higher levels of Pol V signal (Figure 2-12C), compared to non-hypermethylated sites. Consistent with the genomic distribution of ZF108 and its correlation with open chromatin (Figure 2-10A,E), this set of siRNA-producing, hypermethylated loci was highly enriched over promoters and intergenic regions (Figure 2-12D), and the number of methylated loci increased with proximity to the transcriptional start site (TSS) (Figure 2-12E). We also examined the methylation levels in the two clusters of sites (siRNA producing and either hypermethylated or not) defined in Figure 6B across different generations (Figure 2-11D, Figure 2-12F,G). The siRNA-producing hypermethylated loci showed high levels of methylation in different sequence contexts in plants expressing DMS3-ZF108 but not in ZF108 plants (Figure 2-11D), and showed a slight increase between the T2 and T3 generations (Figure 2-11D, Figure 2-12F). The rest of the siRNA-producing loci that were not called as hypermethylated (n=521, Figure 2-11B) were indeed slightly hypermethylated but did not pass the cut off used to call differentially methylated regions (DMRs) (Figure 2-12G), indicating that most of the siRNA-producing loci are associated with methylation to some extent in DMS3-ZF108 plants.

DNA methylation generally represses gene expression and its effect is usually greater when it is close to the TSS (Zhong et al., 2012). In order to study effects on gene expression, we performed RNA-seq in DMS3-ZF108 and ZF108 plants. 63 genes were up-regulated and 35 were down-regulated in DMS3-ZF108 plants compared to ZF108 plants (Figure 2-11E). Of the 35 down-regulated genes, eight showed overlap with the hypermethylated regions bound by

DMS3-ZF108. Consistent with the observation that DNA methylation has a stronger impact on gene expression when it is close to the TSS (Zhong et al., 2012), seven of these eight downregulated genes had hypermethylation within a few hundred base pairs of the TSS (Figure 2-11F).

Co-targeting of DMS3 and NRPD1 enhances ectopic RdDM.

Only a small proportion of Pol V-containing DMS3-ZF108-bound sites displayed siRNAs and DNA methylation (Figure 2-11A,B). Thus, Pol V recruitment is not sufficient to recruit the entire RdDM pathway at most loci. We hypothesized that simultaneous targeting of Pol IV (via NRPD1) and Pol V (via DMS3) might stimulate full RdDM activity and therefore increase the number of additional ZF108 targets that become methylated. We first analyzed genome-wide DNA methylation and siRNA production in T1 plants expressing NRPD1-ZF108. Roughly 45% (4,831) of the 10,776 ZF108-bound loci produced 24-nt siRNAs. However, only 4.3% (n=204) of these siRNA-producing sites became methylated, representing an even lower efficiency for ectopic methylation than DMS3-ZF108 (Figure 2-11B, Figure 2-13A). Moreover, we observed only minor changes in gene expression in NRPD1-ZF108 plants compared to ZF108 control lines, none of which overlapped with genes with hypermethylated regions. This suggests that recruitment of the siRNA biogenesis machinery alone is not sufficient to target methylation at most loci, which is consistent with the inability of NRPD1-ZF108 to target *FWA* silencing in an *nrpe1* mutant (Figure 2-2A, Figure 2-1B).

To study the possible synergistic effect of recruiting Pol IV and Pol V simultaneously, we supertransformed NRPD1-ZF108 into DMS3-ZF108 or ZF108 control lines. While ZF108 control lines expressing NRPD1-ZF108 did not show phenotypic changes compared to wild-type plants, lines expressing both DMS3-ZF108 and NRPD-ZF108 showed a plethora of

developmental defects, such as abnormal leaf, inflorescence, and floral patterning (Figure 2-14A), and these plants were completely infertile. Analysis of siRNA-seq data showed an increase in the number of genomic sites producing siRNAs (Figure 2-13B), as well as an increase in the levels of 21-nt, 22-nt and 24-nt siRNAs compared to DMS3-ZF108 or NRPD-ZF108 fusions alone (Figure 2-14B). Strikingly, 2,186 siRNA-producing sites were hypermethylated in the supertransformants, including almost all hypermethylated regions detected in either DMS3 or NRPD1 ZF108 lines (Figure 7C). Thus, simultaneous targeting of Pol IV and Pol V dramatically enhances the efficiency of ectopic, site-specific methylation. To gain additional insight into the specific contribution of ZF108 binding strength, Pol V recruitment level, and siRNA abundance, we correlated their levels at each site with the levels of hypermethylation. This analysis showed that all three factors strongly correlated with the level of targeted methylation (Figure 2-14C).

We performed RNA-seq to identify genes whose expression might be affected by hypermethylation in plants expressing both DMS3-ZF108 and NRPD1-ZF108. We found 628 down-regulated and 1,073 up-regulated genes in these plants compared to ZF108 controls (Figure 2-13D). 102 down-regulated genes overlapped with hypermethylated regions, most of which are located within 1kb regions proximal to the TSS (Figure 2-13E). These results show that simultaneous targeting of Pol IV and Pol V dramatically enhanced the number of misregulated genes associated with targeted DNA methylation (summarized in Figure 2-13F), and further underscore that proximity to TSSs is a factor to consider when targeting DNA methylation to repress gene expression.

DISCUSSION

Our synthetic biology approach where we combined gain-of-function ZF-RdDM fusions together with loss-of-function mutations allowed us to determine the hierarchy of action of a number of

RdDM components, and also to identify RdDM factors that are most effective in site-specific targeting of ectopic methylation and silencing. The results confirmed the proposed roles of SHH1 acting upstream of Pol IV, the DDR complex acting upstream of Pol V, and SUHV9 acting upstream of the DDR complex and Pol V. The results also highlighted an essential role of ARGONAUTE proteins in methylation targeting, even in the absence of siRNAs. An unexpected finding was that the ZF fusion with MORC6 was effective in recruiting Pol V, leading to DNA methylation and silencing, since MORC6 appears to act primarily downstream of DNA methylation (Moissiard et al., 2012). *morc6* is not a traditional RdDM mutant since it does not show a loss of CHH methylation at the vast majority of RdDM sites, and thus plays little role in RdDM maintenance methylation (Harris et al., 2016). However, MORC6 and its dimerization partner MORC1 have been found to physically interact with some RdDM components (Jing et al., 2016; Liu et al., 2016; Lorković et al., 2012). Our interpretation of these results is that these MORCs may normally use their interaction with RdDM machinery as a mechanism for their own recruitment to facilitate their primary role in silencing that takes place downstream of DNA methylation. Artificial ZF tethering of MORCs to chromatin likely reverses the normal situation and allows MORCs to recruit the RdDM machinery *de novo*.

Of the ten factors that could successfully target *de novo* methylation to the *FWA* promoter, the DDR component DMS3 was the most effective. DMS3 and the DDR complex function in the recruitment of Pol V, which produces the non-coding RNAs required for the eventual targeting of DRM2 to methylate chromatin. Unexpectedly, ZF tethering of DMS3 to *FWA* could efficiently induce DNA methylation and silencing even in the *nrpd1* mutant that eliminates the activity of Pol IV and siRNA biogenesis. A plausible explanation for this result is that Pol V recruitment may be sufficient to recruit an ARGONAUTE protein of the AGO4/6/9

clade through physical interactions with its CTD (El-Shami et al., 2007; Li et al., 2006). AGO4 in turn, even in the absence of siRNAs, may recruit DNA methyltransferase activity through the known physical interaction between AGO4 and DRM2 (Zhong et al., 2014). On the other hand, we found that while tethering the Pol IV subunit NRPD1 to *FWA* was effective in promoting methylation and silencing, it could not do so in the Pol V mutant *nrpe1*, highlighting an essential role for Pol V activity in *de novo* methylation targeting.

Although siRNA biogenesis via Pol IV transcription and the production of non-coding RNAs by Pol V are distinct processes, there is also evidence for cross talk between them. For instance, the Pol V mutant *nrpe1* shows a reduction of siRNA biogenesis at some of loci (Mosher et al., 2008). Loss of DNA methylation in *nrpe1* likely reduces the associated H3K9 methylation, which in turn reduces SHH1 binding and NRPD1 recruitment (Law et al., 2014). Furthermore, the Pol IV mutant *nrpd1* shows a loss of about half of the Pol V RNAs in the genome (Liu et al., 2018). The likely explanation here is that the lack of siRNAs causes a loss of DNA methylation, reduced binding of the methyl binding proteins SUVH2 and SUVH9, and reduced Pol V recruitment (Johnson et al., 2014). Thus, the two major arms of the RdDM pathway are separate but also can have an influence on each other at many loci.

A genome wide analysis showed that the ZF used in this study, ZF108, could bind to thousands of loci in addition to *FWA*. We found that even though DMS3-ZF108 was highly efficient at recruiting Pol V to thousands of loci (summarized in Figure 2-13F), only a small fraction of these regions produced siRNAs and an even smaller proportion became methylated. This indicates that, in contrast to the *FWA* locus, Pol V recruitment is not sufficient to target methylation at most loci. On the other hand, NRPD1-ZF108 was efficient at recruiting siRNA production to thousands of loci, but the number of these sites gaining methylation was even

smaller than in DMS3-ZF108 plants (Figure 2-13F), indicating that recruitment of siRNA production alone is also not sufficient to target methylation at most loci. However, the simultaneous targeting of Pol IV and Pol V activities by combining NRPD1-ZF108 and DMS3-ZF108 fusions synergistically enhanced the efficiency and resulted in the methylation of thousands of loci. These results indicate that despite the known cross talk between them, the two major arms of the RdDM pathway driven by Pol IV and Pol V are recruited independently at most loci. This suggests that future strategies for targeting efficient RdDM should involve a combination of recruiting siRNA biogenesis and Pol V activity.

Previous attempts to target methylation and gene silencing to plant promoters have involved either the expression of hairpin RNAs that are processed by DICER proteins into siRNAs, or Virus Induced Gene Silencing (VIGS) in which plant promoter regions are incorporated into plant viruses that are processed by RNA-dependent RNA polymerases and DICERs into siRNAs (Bond and Baulcombe, 2015; Dalakouras et al., 2009; Jones et al., 1999; Mette et al., 2000; Sijen et al., 2001). The fact that these approaches have met with limited success is consistent with our finding that simultaneous targeting of Pol V and siRNA biogenesis via Pol IV was much more efficient than Pol IV targeting alone. In fact, given the central role of Pol V in DNA methylation establishment, it is somewhat surprising that NRPD1-ZF108 recruitment, expressed hairpins, or VIGS are effective at all. A possible explanation is that very high levels of siRNAs are able to utilize Pol II transcripts, rather than Pol V transcripts, to initially recruit ARGONAUTE/siRNA complexes and DRM2 to initiate methylation. Another possibility is the proposed pairing of AGO/siRNA complexes with DNA (Lahmy et al., 2016) could directly recruit DRM2. This initial seed of methylation would recruit the methyl-binding proteins SUVH2 and SUVH9, which recruit Pol V to initiate the full methylation cycle. In this

regard, it is enlightening that the Pol V mutant *nrpe1* blocked VIGS activity at *FWA* (Bond and Baulcombe, 2015), suggesting that even if Pol II transcripts can serve an initiating function, Pol V transcripts are still needed to establish significant methylation.

In summary, this work provides a theoretical framework for the design of efficient DNA methylation targeting in plants. The factors identified in this work could be used with other programmable DNA-binding targeting platforms, such as CRISPR systems, to improve locus specificity and ease of multiplexed targeting.

MATERIALS AND METHODS

Plant Materials and Growth

All plants in this study were grown in soil under long-day conditions (16h light/8h dark). Transgenic plants were obtained by agrobacterium-mediated floral dipping. T1 transgenic plants were selected on 1/2 MS medium + Glufosinate 50 μg/ml (Goldbio) or 1/2 MS medium + Hygromycin B 25 μg/ml (Invitrogen) in growth chambers under long day conditions and subsequently transferred to soil. Successive transgenic generations were directly sown on soil. Flowering time was scored by counting the total number of rosette and caulinar leaves. The following mutant were introgressed in the *fwa-4* epiallele previously described in Johnson et al, 2014: *shh1-1* (SALK_074540C), *clsy1-7* (SALK_018319), *nrpd1-4* (SALK_083051), *dcl3-1* (SALK_005512), *rdr1-1* (SAIL_672_F11); *rdr6-15* (SAIL_1277_H08), *rdr1-1* (SAIL_672_F11); *rdr2-1* (SAIL_617_H07); *rdr6-15* (SAIL_1277_H08), *morc6-3* (GABI_599B06), *suvh2* (Salk _079574); *suvh9* (Salk_048033), *rdm1-4* (EMS, (Gao et al., 2010) , *dms3-4* (SALK_125019C), *drd1-6* (EMS, (Kanno et al., 2005)), *ago4-5* (EMS, (Greenberg et

al., 2011)), *ago4-4* (FLAG_216G02); *ago6-2* (SALK_031553); *ago9-1* (SALK_127358), *drm1-2* (SALK_031705); *drm2-2* (SALK_150863), *cmt3-11* (SALK_148381). Double mutants between *fwa-d and rdr2-2* (SALK_059661)*, dcl2-1* (SALK_064627)*; dcl4-2* (GABI_160G05), *dcl2-1* (SALK_064627); *dcl3-1* (SALK_005512); *dcl4-2* (GABI_160G05), *nrpe1-1* (EMS, (Kanno et al., 2005)) were previously described in Bond et al, 2015.

Plasmid Construction

NRPD1-3xFlag-ZF: The ZF108 described in Johnson et al, 2014 was first cloned into the unique XhoI site of a modified pCR2 containing a 3xFlag and a Biotin Ligase receptor Peptide (BLRP) and separated by a unique XhoI site. The fragment containing 3xFlag_ZF108_BLRP was digested with AscI and ligated into AscI-digested pENTR-NRPD1, which contains the genomic sequence of NRPD1 (Law et al., 2011), to create pENTR-NRPD1-3xFlag-ZF108. The resulting plasmid was recombined into JP726 (Johnson et al., 2008) using LR clonase (Invitrogen) to create pEG302-NRPD1-3xFlag-ZF108.

DR2-3xFlag-ZF: The same modified pCR2 plasmid described above containing 3xFlag_ZF108_BLRP was used to clone 3xFlag_ZF108_BLRP into AscI-digested pENTR-RDR2, that contains the genomic sequence of RDR2 (Law et al., 2011), to create pENTR-RDR2-3xFlag-ZF108. The resulting plasmid was recombined into JP726 using LR clonase (Invitrogen) to create pEG302-RDR2-3xFlag-ZF108.

SHH1-3xMyc-ZF108: The ZF108 was ligated into the unique XhoI site of the plasmid pENTR-SHH1-3xMyc-BLRP, which contains the genomic sequence of SHH1 and a C-terminal 3xMyc-BLRP tag (Law et al., 2011) to create pENTR-SHH1-3xMyc-ZF108. In this particular construction, a shorter ZF108 sequence with only five tandem copies of the ZF108 repeats was

cloned instead of the six tandem copies present in ZF108. The resulting plasmid was recombined into JP726 using LR clonase (Invitrogen) to create pEG302-SHH1-3xMyc-ZF108.

ZF108-3xHA-SUVH9: The ZF108 was cloned into the unique XhoI site upstream of 3xHA in the pENTR-3xHA-SUVH9 plasmid described in (Johnson et al., 2008), that contains the genomic sequence of SUVH9 including a N-terminal 3xHA tag, to create pENTR-ZF108-3xHA-SUVH9. The resulting plasmid was recombined into JP726 using LR clonase (Invitrogen) to create pEG302-ZF108-3xHA-SUVH9.

DMS3-3xFlag-ZF108: The same modified pCR2 plasmid described above containing 3xFlag_ZF108_BLRP was digested with AscI to clone 3xFlag_ZF108_BLRP into AscI-digested pEG-DMS3, which contains a genomic sequence of DMS3 (Law et al., 2010) to create pEG-DMS3-3xFlag-ZF108.

RDM1-3xHA-ZF108: For this construct, we created the plasmid pENTR-RDM1 by cloning a genomic sequence of RDM1 including 350 base pairs of 5` promoter sequence into the pENTR/D plasmid (Invitrogen) using the primers FWD 5′-CACCATCATGGTATTGTAGACTAAAAC-3′, REV 5′-TTTCTCAGGAAAGATTGGGTCAATG-3′ The ZF108 was cloned into the unique XhoI site of a modified pCR2 plasmid containing 3xHA and BLRP separated by a unique XhoI site. The fragment containing 3xHA-ZF108-BLRP was digested with AscI and ligated into AscI-digested pENTR-RDM1 to create pENTR-RDM1-3xHA-ZF108. The resulting plasmid was recombined into JP726 using LR clonase (Invitrogen) to create pEG302- RDM1-3xHA-ZF108.

MORC6-3xHA-ZF108: The same modified pCR2 plasmid described above containing 3xHA-ZF108-BLRP was digested with AscI to clone 3xHA_ZF108_BLRP into a AscI-digested pENTR-MORC6 plasmid, which contains a genomic sequence of MORC6 (Moissiard et al.,

2012), to create pENTR-MORC6-3xHA-ZF108. The resulting plasmid was recombined into JP726 using LR clonase (Invitrogen) to create pEG302- MORC6-3xHA-ZF108.

MORC1_3xFlag_ZF108: The 3xFlag-ZF108-BLRP fragment in the modified pCR2 plasmid described above was digested with AscI and inserted in the single AscI site of the pENTR-MORC1 plasmid, which contains a genomic sequence of MORC1 (Moissiard et al., 2014), to create pENTR-MORC1-3xFlag-ZF108. The resulting plasmid was recombined into JP726 using LR clonase (Invitrogen) to create pEG302-MORC1-3xFlag-ZF108.

ZF108-3xFlag-9xMyc-AtDRM2: The ZF108 fragment was digested out from a ZF108-containing pUC57 plasmid with EcoRI and cloned into the unique EcoRI site of a modified plasmid pENTR-3xFlag-9xMyc-DRM2 originally described by (Henderson et al., 2010) that is situated upstream of the 3xFlag-9xMyc N-terminal tag that precedes the DRM2 genomic sequence to create pENTR-ZF108-3xFlag-9xMyc-DRM2. The resulting plasmid was recombined into JP726 using LR clonase (Invitrogen) to create pEG302-ZF108-3xFlag-9xMyc-DRM2.

ZF108_3xFlag_NtDRM2_Mtase: For this plasmid, a modified pMDC123 plasmid (Curtis and Grossniklaus, 2003) was created first, containing 1990bp of the promoter region of the Arabidopsis UBQ10 gene upstream of the BLRP-ZF108-3xFlag cassette present in one of the modified pCR2 plasmids described above. Both the UBQ10 promoter and BLRP-ZF108-3xFlag are upstream of the gateway cassette (Invitrogen) present in the original pMDC123 plasmid. The pENTR-NtMTase plasmid described in (Zhong et al., 2014) was used to deliver NtMTase into the modified pMDC123 by LR reaction (Invitrogen), to create pMDC123-ZF108-3xFlag-NtMTase.

ZF108-HA: For this plasmid, we removed the MORC6 coding region present in the pENTR-MORC6-3xHA-ZF108 plasmid described above. First we introduced StuI and ClaI sites upstream and downstream of MORC6 coding region by site-directed mutagenesis using the QuickChange II kit (Agilent). StuI-ClaI digested pENTR-MORC6-3xHA-ZF108 was treated with Klenow fragment (NEB) and re-ligated to create pENTR-pMORC6-3xHA-ZF108. The resulting plasmid was recombined into JP726 using LR clonase (Invitrogen) to create pEG302-pMORC6-3xHA-ZF108.

BS-PCR-seq

Leaf tissue from adult plants of representative T2 lines showing an early flowering time phenotype was collected. DNA was extracted following a CTAB-based method and converted using the EZ DNA methylation-lighting kit. To amplify the different regions, Pfu Turbo Cx (Agilent) were used together with primers containing the Illumina adaptors. The primers used for the different regions are:

REGION 1 FWD 5'-TCATATAAAAAAAAAATTAAATTTCATTTCACAATAACCATT-3',

REGION 1 REV 5'- GTATGGGYTTYGATAAAGAATATATGAGATTYT-3',

REGION 2 FWD 5'-CTCATATATACCTTATCCCATTCAACATTCATA-3',

REGEION 2 REV 5'- AAGATYTGATATTTGGYTGGAAAAAAYAATAATAAT-3',

REGION 3 FWD 5'- CRCTCTTTATCCCATTCAACATTCATAC-3',

REGION 3 REV 5'- TTTGGTTGAAAAAAATAATAAAAATTTGATTGTYAGTAT-3'

Different PCR products for the same sample were pooled and purified by Ampure beads. Libraries were made from purified PCR products by an Illumina NeoPrep automatic library

preparation machine or by Kapa DNA hyper kit with Illumina TruSeq DNA adapters or by NuGen Ultralow V2 kit. Libraries were sequenced on Illumina HiSeq 2000 or HiSeq 2500.

qseq files from the sequencer were demultiplexed and converted to fastq format with a customized script for downstream analysis. For BS-PCR-seq data, raw reads with designed BS-PCR primers were filtered followed by primer trimming with customized scripts. Trimmed reads were then aligned with BSMAP (v.2.74) (Xi and Li, 2009) to the reference TAIR10 genome by allowing up to 2 mismatches (-v 2), 1 best hit (-w 1) and aligning to both strands (-n 1). The methylation level at each cytosine was then extracted with BSMAP (methratio.py) scripts by allowing only unique mapped reads (-u). Reads with more than 3 consecutive methylated CHH sites were removed as described previously (Cokus et al., 2008). Methylation levels at each cytosine were calculated as #C/(#C+#T). To visualize the BS-PCR-seq data, only cytosines within designated regions around the FWA gene were kept and plotted with customized R scripts.

ChIP-seq

ChIPs were performed as described previously (Johnson et al., 2014) with minor modifications. 4gr of inflorescences were ground and crosslinked with 1% formaldehyde. Chromatin was sheared using a Bioruptor Plus (Diagenode) and immunoprecipitated with Flag M2 (Sigma) and HA 3F10 (Roche) commercial antibodies and NRPE1 antibody described in Johnson et al., 2014. Libraries were prepared using the NuGen Ovation Ultra Low System V2 1-16 kits following the manufacturer's instructions.

For ChIP-seq data, qseq files from the sequencer were demultiplexed and converted to fastq format with a customized script for downstream analysis. fastq reads were aligned to the Arabidopsis reference genome (TAIR10) with Bowtie (v1.0.0) (Langmead et al., 2009), allowing only uniquely mapping reads with fewer than two mismatches, and duplicated reads were combined into one read. NRPE1 ChIP-seq peak were called using MACS2 (v 2.1.1.) (Zhang et al., 2008) with NRPE1 pre-immune ChIP-seq as a control. For Flag and HA ChIP-seq, peak were called using MACS2 (v 2.1.1.) (Zhang et al., 2008) with Flag and HA ChIP-seq in Col-0 as controls. ChIP-seq metaplots were plotted using NGSplot (v 2.41.4) (Shen et al., 2014) (Shen et al., 2014). In order to identify predominant motifs in ZF108-associated ChIP-seq peaks, HOMER (Heinz et al., 2010) was applied to 200 bp around the ZF108 ChIP-seq peak midpoint.

Whole Genome Bisulfite Sequencing (WGBS)

Libraries for WGBS were made using 100ng of CTAB-extracted DNA from leaves of adult plants. Libraries were prepared using the Nugen Ultralow Methyl-seq kit and Qiagen Epitect Fast following the manufacturer's instructions.

For WGBS data, qseq files from the sequencer were demultiplexed and converted to fastq format with a customized script for downstream analysis. fastq reads were aligned using with BSMAP (v 2.74) (Xi and Li, 2009) to the reference TAIR10 genome by allowing up to 2 mismatches (-v 2), 1 best hit (-w 1) and aligning to both strands (-n 1). Methylation levels at each cytosine were then extracted with BSMAP (methratio.py) scripts by allowing only unique mapped reads (-u). Reads with more than 3 consecutive methylated CHH sites were removed as described previously (Cokus et al., 2008). Methylation levels at each cytosine were calculated as #C/(#C+#T). DMRs between DMS-ZF108 and ZF108, NRPD1-ZF108 and ZF108, DMS3-

ZF108 X NRPD1-ZF108 and ZF108 were calculated as before (Stroud et al., 2013). In general, transgenic lines with the same genotype were combined. DMRs were then defined with R package DMRcaller using the 100bp bins method as described before (Stroud et al., 2013). DMRs within 200bp of each other were merged for further analysis.

RNA-seq analysis

For *FWA* expression in DMS3-ZF T2 lines, RNA from 12 day-old seedlings grown on plates containing ½ MS+BASTA was extracted using Direct-zol kit (Zymo). For the rest of experiments, RNA from 4-5 week-old leaves was extracted using Direct-zol kit (Zymo). 75ng of total RNA was used to prepare libraries using the Neoprep stranded mRNA-seq kit (Illumina). Alternatively, 1μg of total RNA was used to prepare libraries using the TruSeq Stranded mRNA kit (Illumina).

For RNA-seq data, qseq files from the sequencer were demultiplexed and converted to fastq format with a customized script for downstream analysis. fastq reads were first aligned to TAIR10 gene annotation using Tophat (v 2.0.13) (Trapnell et al., 2009) by allowing up to two mismatches and only keeping reads that mapped to one location. When reads did not map to the annotated genes, the reads were mapped to the TAIR10 genome. Number of reads mapping to genes were calculated by HTseq (Anders et al., 2015) with default parameters. Expression levels were determined by RPKM (reads per kilobase of exons per million aligned reads). Differentially expressed genes were defined with R package DESeq (Anders and Huber, 2010) using a 2 fold change and FDR less than 0.05 as cut off.

small RNA-seq

RNA from flowers was extracted using the Zymo Direct-zol Kit. For siRNA libraries, 2ug total RNA was run in 15% UREA gels and small RNAs from 15 to 30bp were cut and precipitated. This RNA was used to prepare libraries using the Truseg small RNA kit (Illumina) following manufacturer's instructions.

For small RNA-seq data, qseq files from the sequencer were demultiplexed and converted to fastq format with a customized script for downstream analysis. fastq reads were trimmed for Illumina adaptors using Cutadapt (v 1.9.1) and mapped to the TAIR10 reference genome using Bowtie (v1.1.0) 51 (Langmead et al., 2009) allowing only one unique hit (-m 1) and zero mismatch.

ATAC-seq analysis

For ATAC-seq in Col-0, raw data from previously published data (GSM2260231) (Lu et al., 2016) were used in this paper. Data were processed as described previously. Then ATAC-seq metaplots were plotted using NGSplot (v 2.41.4) (Shen et al., 2014).

FIGURE LEGENDS

Figure 2-1. Methylation targeting with NRPD1, RDR2 and SHH1. Related to Figure 2-2.

(A) Model of RNA-directed DNA methylation. **(B)** Flowering time of Col-0 and *fwa* untransformed plants as well as T1 lines expressing the NRPD1-ZF108 transgene in different mutant backgrounds. **(C)** Screenshot with 21-nt, 22-nt and 24-nt siRNAs accumulation over the FWA promoter in two untransformed Col-0, *fwa* and *fwa dcl2 dcl3 dcl4* (*fwa dcl2/3/4*) plants as well as in two representative T2 lines expressing NRPD1-ZF108 in different mutant backgrounds introgressed into the *fwa* mutant. Methylation levels at different context (CG, CHG and CHH, where H is A, T, C) over the *FWA* promoter in wild-type Col-0, *fwa* and *fwa* transformed with NRPD1-ZF108 are shown. **(D)** Flowering time of Col-0 and *fwa* untransformed plants as well as T1 lines expressing the RDR2-ZF108 transgene in different mutant backgrounds. **(E)** Flowering time of Col-0 and *fwa* untransformed plants as well as T1 lines expressing the SHH1-ZF108 transgene in different mutant backgrounds. Although we observed a small number of SHH1-ZF108 T1 plants in a *drm1 drm2* mutant background that showed early flowering, these lines were all late flowering in the T2 suggesting that the T1 early flowering phenotype observed was caused by something other than *FWA* silencing such as stress.

Figure 2-2. NRPD1, RDR2 and SHH1 targeted methylation.

(A) Flowering time of 4 representative T2 lines expressing NRPD1-ZF108 in different mutant backgrounds introgressed in the *fwa* mutant. **(B)** DNA methylation over the *FWA* promoter measured by BS-PCR-seq in representative T2 lines expressing NRPD1-ZF108, RDR2-ZF108 and SHH1-ZF108 in different mutant backgrounds introgressed in the *fwa* mutant. CG, CHG and CHH methylation over three *FWA* promoter regions are depicted. **(C)** Normalized siRNA

accumulation over the 200bp covering the ZF108 binding sites in the *FWA* promoter in Col-0, *fwa* and *fwa x dcl2 dcl3 dcl4* (*fwa dcl2/3/4*) controls as well as in lines expressing NRPD1-ZF108 in different mutant backgrounds introgressed in the *fwa* mutant. Results show accumulation of 21nt, 22nt and 24nt siRNAs for two independent lines per background. **(D)** Flowering time of 4 representative T2 lines expressing RDR2-ZF108 in different mutant backgrounds introgressed into the *fwa* mutant. **(E)** Flowering time of 4 representative T2 lines expressing SHH1-ZF108 in different mutant backgrounds introgressed into the *fwa* mutant. **(A)**, **(B)**, **(D)**, **(F)** *wt* corresponds to single *fwa* mutant and the rest of named mutants are in *fwa* background. For all T2 flowering time experiments, flowering time of Col-0, *fwa, fwa nrpe1 and fwa drm1 drm2 (fwa drm1/2)* controls is shown.

Figure 2-3. SUVH9, DMS3 and RDM1 targeted methylation.

(A) Flowering time of 4 representative T2 lines expressing ZF108-SUVH9 in different mutant backgrounds introgressed into the *fwa* mutant. **(B)** DNA methylation over the *FWA* promoter measured by BS-PCR-seq in representative T2 lines expressing ZF108-SUVH9, DMS3-ZF108 and RDM1-ZF108 in different mutant backgrounds introgressed in the *fwa* mutant. CG, CHG and CHH methylation over three *FWA* promoter regions are depicted. For DNA methylation profiling of ZF108-SUVH9 in the *fwa nrpd1* background, the single early flowering plant from line 1 was used. **(C)** Flowering time of 4 representative T2 lines expressing DMS3-ZF108 in different mutant backgrounds introgressed into the *fwa* mutant. **(D)** Normalized siRNA accumulation over the 200bp covering the ZF108 binding sites in the *FWA* promoter in Col-0, *fwa* and *fwa nrpd1* controls as well as in lines expressing DMS3-ZF108 in different mutant backgrounds introgressed into the *fwa* mutant. Results show accumulation of 21nt, 22nt and 24nt

siRNAs for two independent lines per background. **(E)** ZF108 and Pol V ChIP-seq peaks over the *FWA* promoter. Upper: ZF108 Flag ChIP signal of DMS3-ZF108 in *fwa* and *fwa drm1 drm2 (fwa drm1/2)* backgrounds and non-transgenic line and ZF108 HA ChIP signal of ZF108-SUVH9 and MORC6-ZF108 in *fwa* and *fwa drm1 drm2 (fwa drm1/2)* backgrounds, and non-transgenic line. Lower: Pol V signal of DMS3-ZF108, ZF108-SUVH9 and MORC6-ZF108 in *fwa* and *fwa drm1 drm2 (fwa drm1/2)* backgrounds, as well as in *fwa* and *fwa drm1 drm2 (fwa drm1/2)* untransformed controls. **(F)** Flowering time of 4 representative T2 lines expressing RDM1-ZF108 in different mutant backgrounds introgressed into the *fwa* mutant. **(A)**, **(B)**, **(D)**, **(F)** *wt* corresponds to single *fwa* mutant and the rest of named mutants are in *fwa* background. For all T2 flowering time experiments, flowering time of Col-0, *fwa, fwa nrpe1 and fwa drm1 drm2 (fwa drm1/2)* controls is shown.

Figure 2-4. Methylation targeting with SUVH9, DMS3 and RDM1. Related to Figure 2-3.

(A) Flowering time of Col-0 and *fwa* untransformed plants as well as T1 lines expressing the ZF108-SUVH9 transgene in different mutant backgrounds. **(B)** Flowering time of Col-0 and *fwa* untransformed plants as well as T1 lines expressing the DMS3-ZF108 transgene in different mutant backgrounds. **(C)** Screenshot with 21-nt, 22-nt and 24-nt siRNAs accumulation over the *FWA* promoter in two untransformed Col-0, *fwa* and *fwa nrpd1* plants as well as in two representative T2 lines expressing DMS3-ZF108 in different mutant backgrounds introgressed into the *fwa* mutant. Methylation levels at different context (CG, CHG and CHH, where H is A, T, C) over the *FWA* promoter in wild-type Col-0, *fwa* and *fwa* transformed with DMS3-ZF108 are shown. **(D)** 5' nucleotide preference for 21- to 24-nt unique siRNAs produced over the 200bp covering the ZF108 binding sites in the *FWA* promoter in Col-0, DMS3ZF108 in *fwa* and

DMS3ZF108 in *fwa nrpd1* backgrounds. **(E)** FWA expression measured by RNA-seq in 3 independent pools of seedlings from untransformed Col-0 and *fwa* mutants and in 3 independent T2 lines expressing DMS3-ZF108 in the *fwa* and *fwa drm1 drm2 (fwa drm1/2)* backgrounds. **(F)** Flowering time of Col-0 and *fwa* untransformed plants as well as T1 lines expressing the RDM1-ZF108 transgene in different mutant backgrounds.

Figure 2-5. MORC-mediated targeted methylation.

(A) Flowering time of 4 representative T2 lines expressing MORC6-ZF108 in different mutant backgrounds introgressed into the *fwa* mutant. **(B)** DNA methylation over the *FWA* promoter measured by BS-PCR-seq in representative T2 lines expressing MORC6-ZF108 and MORC1-ZF108 in different mutant backgrounds introgressed in the *fwa* mutant. CG, CHG and CHH methylation over three *FWA* promoter regions are depicted. **(C)** Flowering time of 4 representative T2 lines expressing MORC1-ZF108 in different mutant backgrounds introgressed into the *fwa* mutant. **(A)**, **(B)**, **(C)**, *wt* corresponds to single *fwa* mutant and the rest of named mutants are in *fwa* background. For all T2 flowering time experiments, flowering time of Col-0, *fwa, fwa nrpe1 and fwa drm1 drm2 (fwa drm1/2)* controls is shown.

Figure 2-6. Methylation targeting with MORC. Related to Figure 2-5.

(A) Flowering time of Col-0 and *fwa* untransformed plants as well as T1 lines expressing the MORC6-ZF108 transgene in different mutant backgrounds. **(B)** Flowering time of Col-0 and *fwa* untransformed plants as well as T1 lines expressing the MORC1-ZF108 transgene in different mutant backgrounds.

Figure 2-7. DRM2 MTase targeted methylation.

(A) Flowering time of 4 representative T2 lines expressing ZF108-MTase in different mutant backgrounds introgressed into the *fwa* mutant. **(B)** DNA methylation over the *FWA* promoter of ZF108-MTase representative T2 lines measured by BS-PCR-seq. CG, CHG and CHH methylation over three regions from the *FWA* promoter are depicted. **(A)**, **(B)**, *wt* corresponds to single *fwa* mutant and the rest of named mutants are in *fwa* background. For all T2 flowering time experiments, flowering time of Col-0, *fwa, fwa nrpe1 and fwa drm1 drm2 (fwa drm1/2)* controls is shown.

Figure 2-8. Methylation targeting with DRM2 MTase. Related Figure 4.

Flowering time of Col-0 and *fwa* untransformed plants as well as T1 lines expressing the ZF108-MTase transgene in different mutant backgrounds.

Figure 2-9. DMS3-ZF108 efficiently recruits Pol V to thousands of loci.

(A) Screenshot of ZF108 ChIP-seq over *FWA* in DMS3-ZF108 and ZF108 control lines. ZF108 binding sites and sequence is depicted. **(B)** Screenshot of ZF108 ChIP-seq over two representative off target sites in DMS3-ZF108 and ZF108 control lines. Similar sequences to ZF108 designed binding sequence (17bp match) are depicted. '*' indicate nucleotide substitution. 'v' indicates nucleotide insertion. **(C)** Overlap between ChIP-seq peaks in DMS3. ZF108 and ZF108 lines using 2 fold change compared to control ChIP-seq in Col-0 as cut off. **(D)** ChIP-seq inflection curves for the two DMS3-ZF108 and ZF108 are shown. Peak intensity compared to control ChIP-seq in Col-0 is shown on the Y-axis and peak rank is shown on the X-axis. **(E)** Predominant motif identified by de novo motif analysis for DMS3-ZF108 and ZF108.

(F) Screenshot of ZF108 ChIP-seq and Pol V ChIP-seq in DMS3-ZF108 and Pol V ChIP-seq in Col-0 over a random genomic region. **(G)** Metaplot and heatmap of Flag and NRPE1 ChIP-seq signals in DMS3-ZF108 over off target sites with (upper panel) or without (lower panel) NRPE1 recruitment.

Figure 2-10. Characterization of DMS3-ZF108 off target sites. Related Figure 2-9.

(A) Pie chart of genomic annotation for peaks in common between FLAG-DMS3-ZF108 and HA-ZF108. **(B)** Bar plot showing the motif occurrence percentage over different deciles of DMS3-ZF108 peaks. **(C)** Pie chart showing the percentage of core motif-containing regions of ZF108 genome-wide bound by DMS3-ZF108. **(D)** Genomic annotation of core motif-containing regions with or without DMS3-ZF108 binding. **(E)** Metaplot of Col-0 ATAC-seq signals over core motif-containing regions with or without DMS3-ZF108 binding. **(F)** Metaplot of Col-0 ATAC-seq signals over ZF108 off target sites with or without NRPE1 recruitment in DMS3-ZF108. **(G)** Genomic annotation of ZF108 off target sites with (left) or without (right) NRPE1 recruitment in DMS3-ZF108.

Figure 2-11. DMS3-ZF108 targets methylation at hundreds of loci.

(A) Metaplot of NRPE1 ChIP-seq and 24nt siRNAs in DMS3-ZF108 over off target sites with (left) or without (right) 24nt siRNAs production. Shaped area around each curve represents standard errors. Y-axis represents Reads Per Kilobase Million (RPKM). **(B)** Pie chart showing the number of hyper methylated DMRs over 24nt siRNAs producing off target sites in DMS3-ZF108. **(C)** Boxplot of 24nt siRNAs levels (Reads Per Million, RPM) in DMS3-ZF108 over 24nt siRNA producing off target sites with (left) or without hyper methylated DMRs (right). **(D)** CG,

CHG, CHH methylation in T2 and T3 DMS3-ZF108 and ZF108 over 24nt siRNA producing off target sites with hyper methylated DMRs. *P < 0.05 (Welch Two Sample t-test). **(E)** Log2 RPKM scatterplot of differentially expressed genes in DMS3-ZF108 and ZF108. **(F)** Distance of targeted DNA methylation regions in DMS3-ZF108 to the nearest TSS of down regulated genes in DMS3-ZF108 compared with ZF108.

Figure 2-12. siRNA recruitment and DNA methylation targeting and to DMS3-ZF108 off target sites. Related to Figure 2-11.

(A) 24-nt siRNA deciles with NRPE1 recruitment over off target sites in DMS3-ZF108. **(B)** Boxplot of NRPE1 levels in DMS3-ZF108 over different 24-nt siRNAs deciles as shown in **(A)**. *P < 0.05 (Welch Two Sample t-test). **(C)** NRPE1 levels over 24-nt siRNA-producing off target sites with or without hyper methylated DMRs. *P < 0.05 (Welch Two Sample t-test). **(D)** Genomic annotation of off target sites with NRPE1 recruitment, 24-nt siRNA production, and targeted DNA methylation in DMS3-ZF108. **(E)** Frequency of off target sites with NRPE1 recruitment, 24-nt siRNA production, and targeted DNA methylation in DMS3-ZF108 within 2kb upstream and downstream of annotated genes. **(F)** Scatterplot of CG, CHG and CHH methylation in T2 and T3 DMS3-ZF108. Dashed lines provide visual assistance. **(G)** Boxplot of CG, CHG, CHH methylation in T2 and T3 DMS3-ZF108 and ZF108 over 24-nt siRNA-producing off target sites without hyper methylated DMRs. *P < 0.05 (Welch Two Sample t-test).

Figure 2-13. Co-targeting RNA Polymerases IV and V promotes efficient ectopic DNA methylation.

(A) Screenshot of CG, CHG, CHH methylation and normalized 24nt siRNA levels in DMS3-ZF108, NRPD1-ZF108, DMS3-ZF108 X NRPD1-ZF108 and ZF108 as well as Flag ChIP-seq signals in DMS3-ZF108 and HA ChIP-seq signals in ZF108 over representative targeted hyper methylated DMRs in DMS3-ZF108, NRPD1-ZF108, DMS3-ZF108 X NRPD1-ZF108 (left), hyper methylated DMRs in DMS3-ZF108, and DMS3-ZF108 X NRPD1-ZF108 (middle left), hyper methylated DMRs in NRPD1-ZF108 and DMS3-ZF108 X NRPD1-ZF108 (middle right), and hyper methylated DMRs in DMS3-ZF108 X NRPD1-ZF108 only (right). **(B)** Venn diagram of 24nt siRNAs producing off target sites in DMS3-ZF108, NRPD1-ZF108 and DMS3-ZF108 X NRPD1-ZF108. **(C)** Venn diagram of 24nt siRNAs producing off target sites with targeted DNA methylation in DMS3-ZF108, NRPD1-ZF108 and DMS3-ZF108 X NRPD1-ZF108. **(D)** Log2 RPKM scatterplot of differentially expressed genes in DMS3-ZF108 X NRPD1-ZF108 and ZF108. **(E)** Histogram (upper panel) and dot plot depicting the distance of hyper methylated ZF108 off target sites to the TSS of nearby down regulated genes in DMS3-ZF108 X NRPD1-ZF108 (lower panel). Some of the genes are associated multiple DMRs. **(F)** Multi-level pie chart of number of ZF108 off target sites, NRPE1 recruited sites in DMS3-ZF108, 24nt siRNA producing sites, hyper methylated DMR sites and repressed genes in DMS3-ZF108 (left), NRPD1-ZF108 (middle) and DMS3-ZF108 X NRPD1-ZF108 (right).

Figure 2-14. Methylation targeting with DMS3-ZF108 X NRPD1-ZF108. Related Figure 2-13.

(A) Abnormal phenotype of DMS3-ZF108 X NRPD1-ZF108. Upper left, middle left and middle right panels show 3 different DMS3-ZF108 X NRPD1-ZF108 mutant plants. Upper right panel shows a non-mutant plant. White scale in the lower corner of each image = 1cm. Lower left,

middle left and middle right panels show inflorescences of 3 different DMS3-ZF108 X NRPD1-ZF108 mutant plants. Lower right panel shows inflorescence of a non-mutant plant. **(B)** Boxplot of 24-nt, 22-nt and 21-nt siRNA levels in ZF108, DMS3-ZF108, NRPD1-ZF108 and DMS3-ZF108 X NRPD1-ZF108 over off target sites producing 24-nt siRNAs in DMS3-ZF108 (upper), NRPD1-ZF108 (middle), and DMS3-ZF108 X NRPD1-ZF108 (lower). **(C)** Boxplot of 10 deciles of methylation difference between DMS3-ZF108 X NRPD1-ZF108 and ZF108 (upper panel) and normalized 24-nt siRNA levels (Reads Per Million, RPM) in DMS3-ZF108 X NRPD1-ZF108 over the 10 methylation difference deciles (upper middle panel), normalized NRPE1 abundance in DMS3-ZF108 (lower middle panel), and normalized DMS3-ZF108 binding in DMS3-ZF108 (lower panel).

Figure 2-1

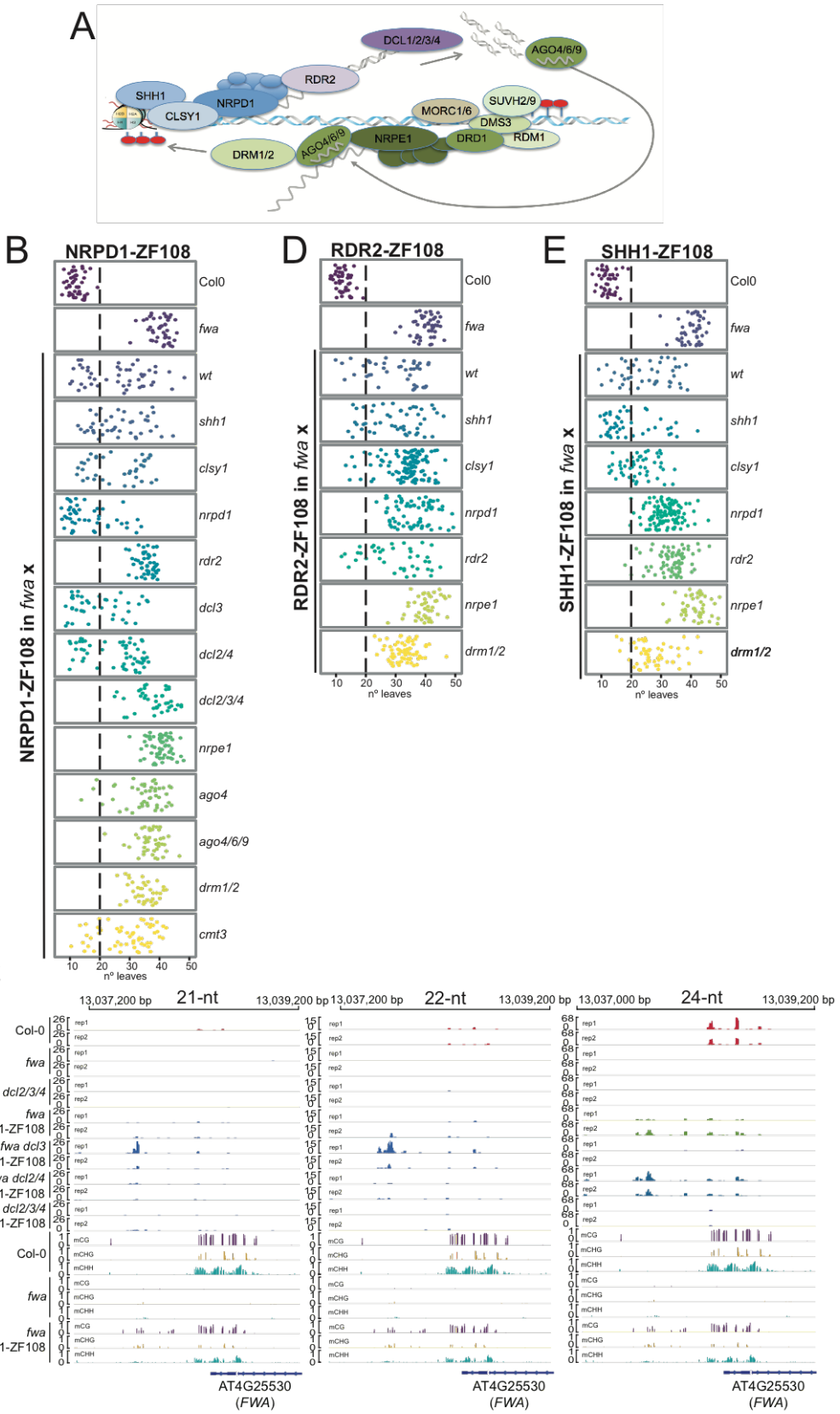

A
DCL1/2/3/4
AGO4/6/9
RDR2
SHH1
NRPD1
CLSY1
MORC1/6
SUVH2/9
DMS3
NRPE1
DRD1
RDM1
DRM1/2
AGO4/6/9
B
NRPD1-ZF108
NRPD1-ZF108 in fwa x
Col0
fwa
wt
shh1
clsy1
nrpd1
rdr2
dcl3
dcl2/4
dcl2/3/4
nrpe1
ago4
ago4/6/9
drm1/2
cmt3
10 20 30 40 50
nº leaves
D
RDR2-ZF108
RDR2-ZF108 in fwa x
Col0
fwa
wt
shh1
clsy1
nrpd1
rdr2
nrpe1
drm1/2
10 20 30 40 50
nº leaves
E
SHH1-ZF108
SHH1-ZF108 in fwa x
Col0
fwa
wt
shh1
clsy1
nrpd1
rdr2
nrpe1
drm1/2
10 20 30 40 50
nº leaves
C
13,037,200 bp
21-nt
13,039,200 bp
13,037,200 bp
22-nt
13,039,200 bp
13,037,000 bp
24-nt
13,039,200 bp
siRNAs
Col-0
fwa
fwa dcl2/3/4
fwa
NRPD1-ZF108
fwa dcl3
NRPD1-ZF108
fwa dcl2/4
NRPD1-ZF108
fwa dcl2/3/4
NRPD1-ZF108
Methylation
Col-0
fwa
fwa
NRPD1-ZF108
mCG
mCHG
mCHH
AT4G25530
(FWA)
AT4G25530
(FWA)
AT4G25530
(FWA)

Figure 2-2

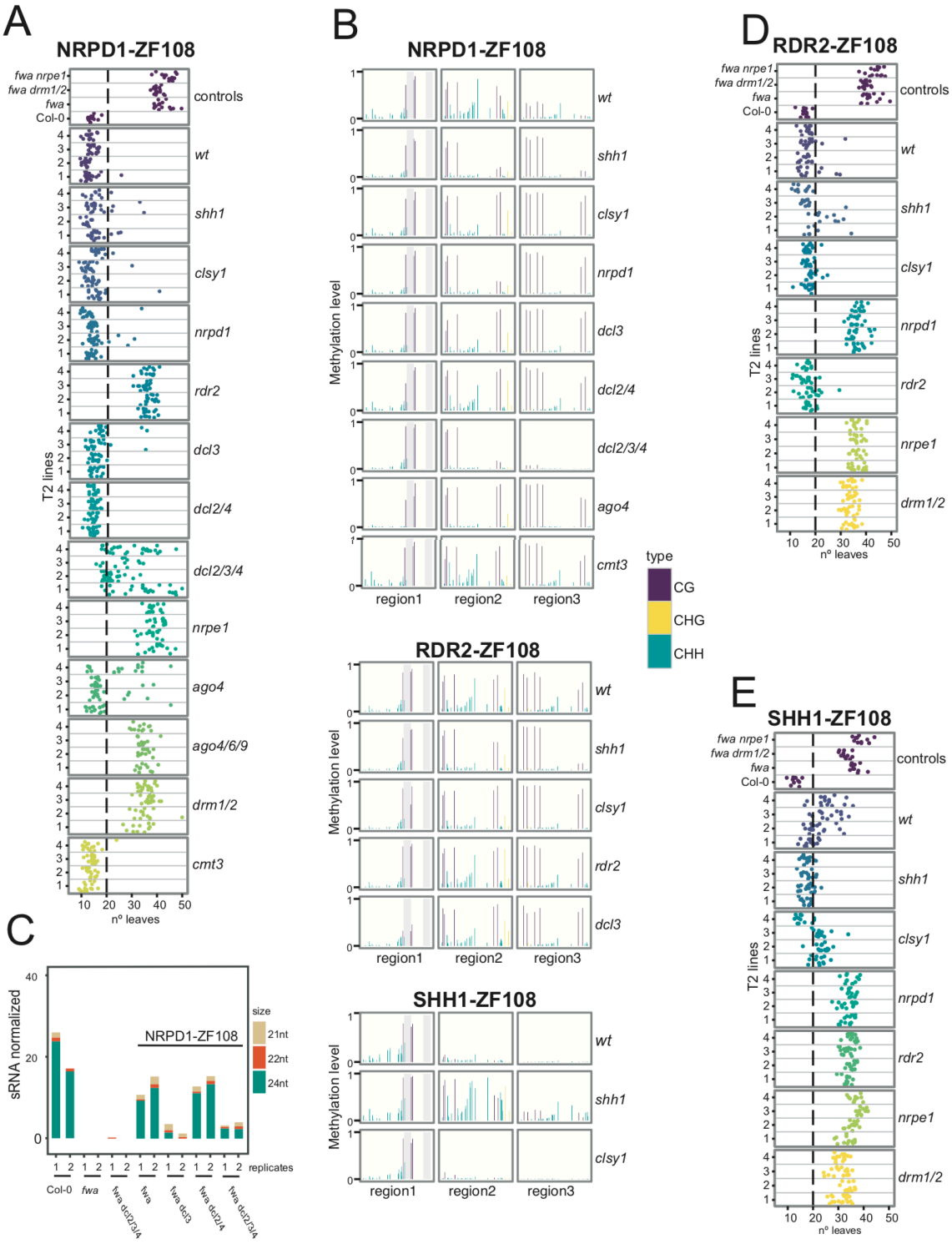

A
NRPD1-ZF108
fwa nrpe1
fwa drm1/2
fwa
Col-0
controls
wt
shh1
clsy1
nrpd1
rdr2
dcl3
dcl2/4
dcl2/3/4
nrpe1
ago4
ago4/6/9
drm1/2
cmt3
T2 lines
nº leaves
B
NRPD1-ZF108
RDR2-ZF108
SHH1-ZF108
Methylation level
region1
region2
region3
type
CG
CHG
CHH
C
sRNA normalized
NRPD1-ZF108
size
21nt
22nt
24nt
replicates
Col-0
fwa
fwa dcl2/3/4
fwa
fwa dcl3
fwa dcl2/4
fwa dcl2/3/4
D
RDR2-ZF108
E
SHH1-ZF108

Figure 2-3

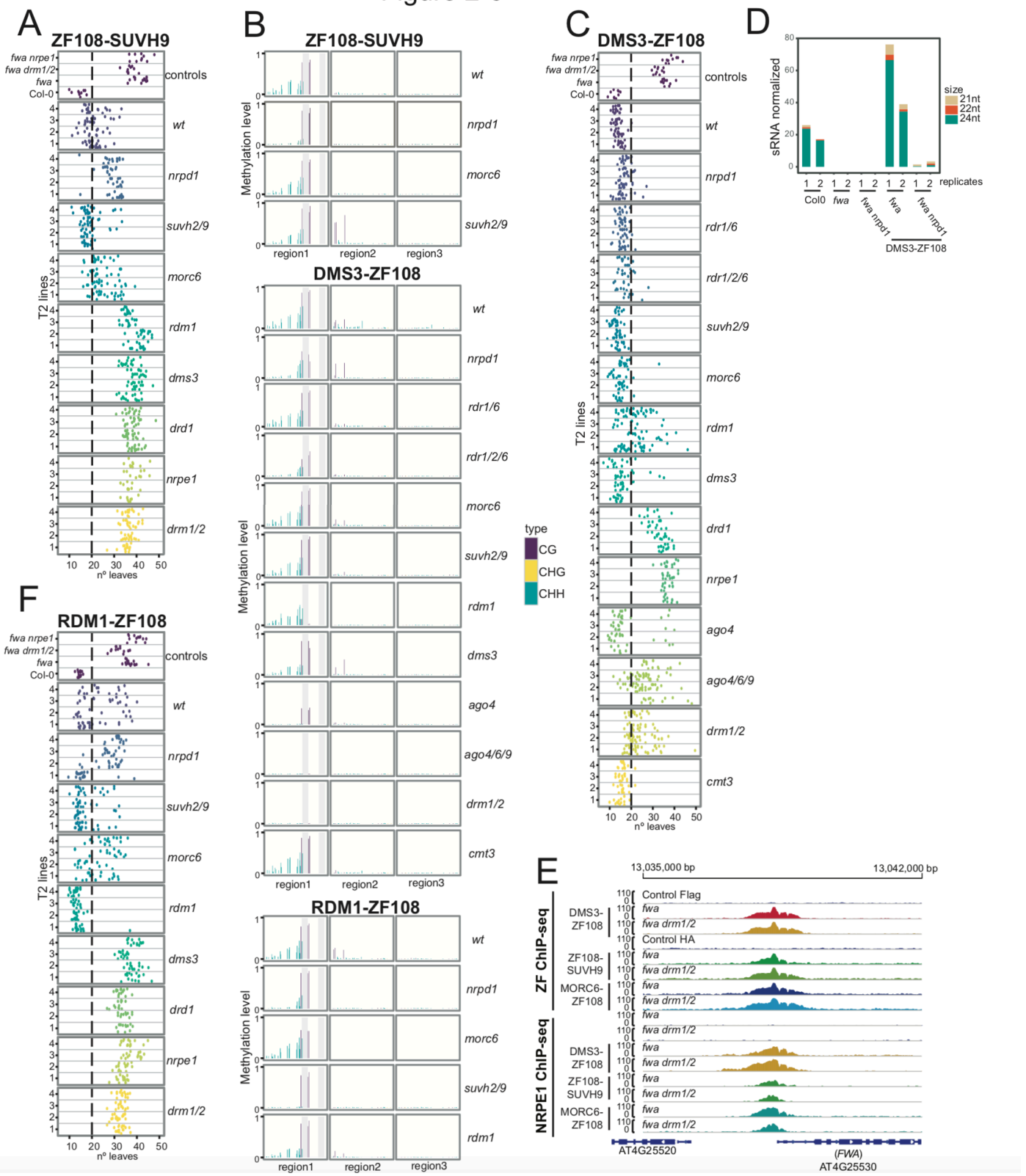

A
ZF108-SUVH9
fwa nrpe1
fwa drm1/2
fwa
Col-0
controls
wt
nrpd1
suvh2/9
morc6
rdm1
dms3
drd1
nrpe1
drm1/2
T2 lines
nº leaves
B
ZF108-SUVH9
Methylation level
region1
region2
region3
DMS3-ZF108
rdr1/6
rdr1/2/6
ago4
ago4/6/9
cmt3
RDM1-ZF108
type
CG
CHG
CHH
C
DMS3-ZF108
D
sRNA normalized
size
21nt
22nt
24nt
replicates
Col0
fwa
fwa nrpd1
fwa nrpd1
DMS3-ZF108
E
13,035,000 bp
13,042,000 bp
Control Flag
Control HA
ZF ChIP-seq
NRPE1 ChIP-seq
DMS3-ZF108
ZF108-SUVH9
MORC6-ZF108
AT4G25520
(FWA)
AT4G25530
F
RDM1-ZF108

Figure 2-4

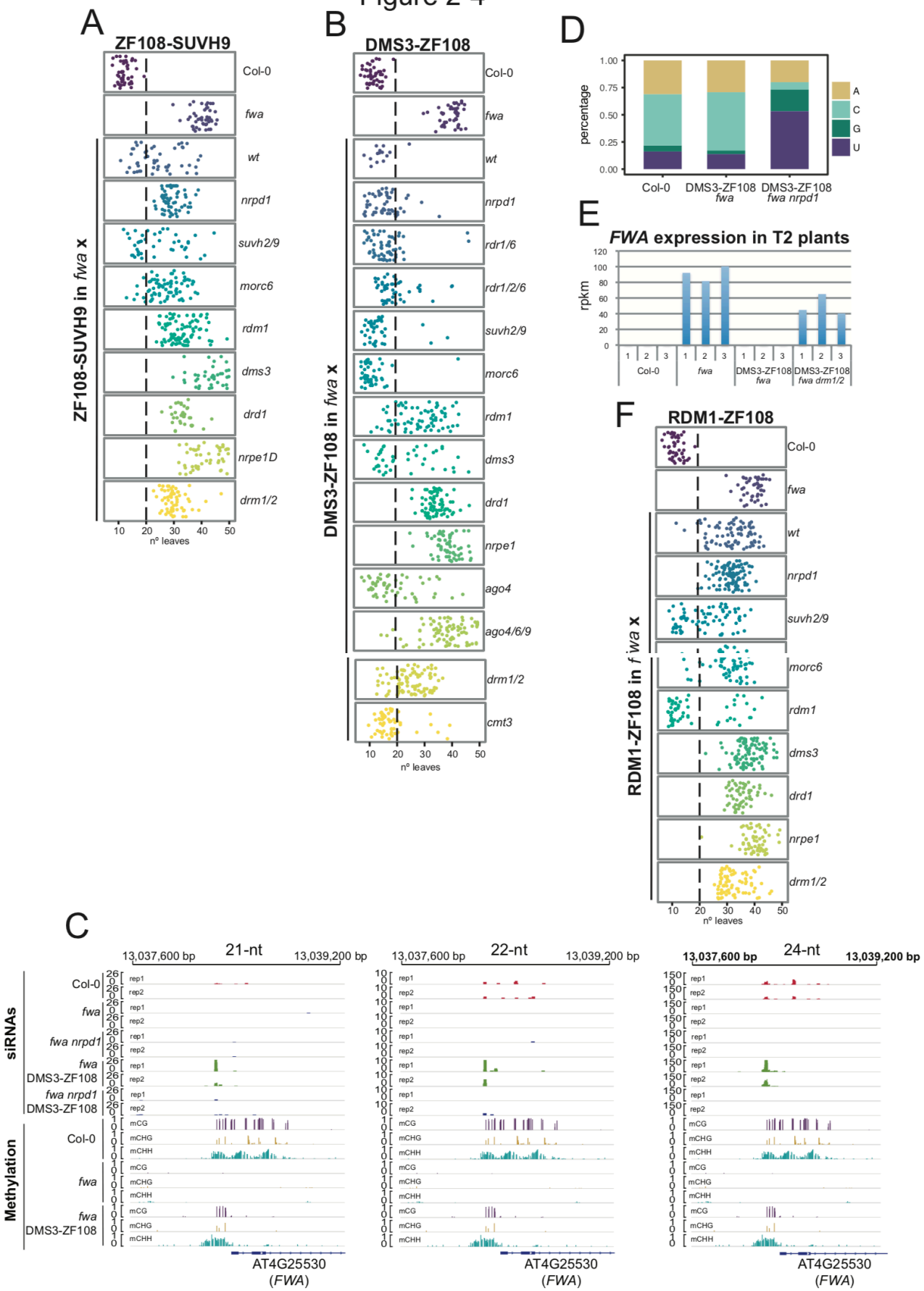

A
ZF108-SUVH9
ZF108-SUVH9 in fwa x
Col-0
fwa
wt
nrpd1
suvh2/9
morc6
rdm1
dms3
drd1
nrpe1D
drm1/2
10 20 30 40 50
n° leaves
B
DMS3-ZF108
DMS3-ZF108 in fwa x
Col-0
fwa
wt
nrpd1
rdr1/6
rdr1/2/6
suvh2/9
morc6
rdm1
dms3
drd1
nrpe1
ago4
ago4/6/9
drm1/2
cmt3
10 20 30 40 50
n° leaves
D
percentage
1.00
0.75
0.50
0.25
0.00
A
C
G
U
Col-0
DMS3-ZF108 fwa
DMS3-ZF108 fwa nrpd1
E
FWA expression in T2 plants
rpkm
120
100
80
60
40
20
0
1 2 3 1 2 3 1 2 3 1 2 3
Col-0
fwa
DMS3-ZF108 fwa
DMS3-ZF108 fwa drm1/2
F
RDM1-ZF108
RDM1-ZF108 in fwa x
Col-0
fwa
wt
nrpd1
suvh2/9
morc6
rdm1
dms3
drd1
nrpe1
drm1/2
10 20 30 40 50
n° leaves
C
21-nt
22-nt
24-nt
13,037,600 bp
13,039,200 bp
siRNAs
Methylation
Col-0
fwa
fwa nrpd1
DMS3-ZF108
fwa DMS3-ZF108
fwa nrpd1 DMS3-ZF108
rep1
rep2
mCG
mCHG
mCHH
AT4G25530
(FWA)

Figure 2-5

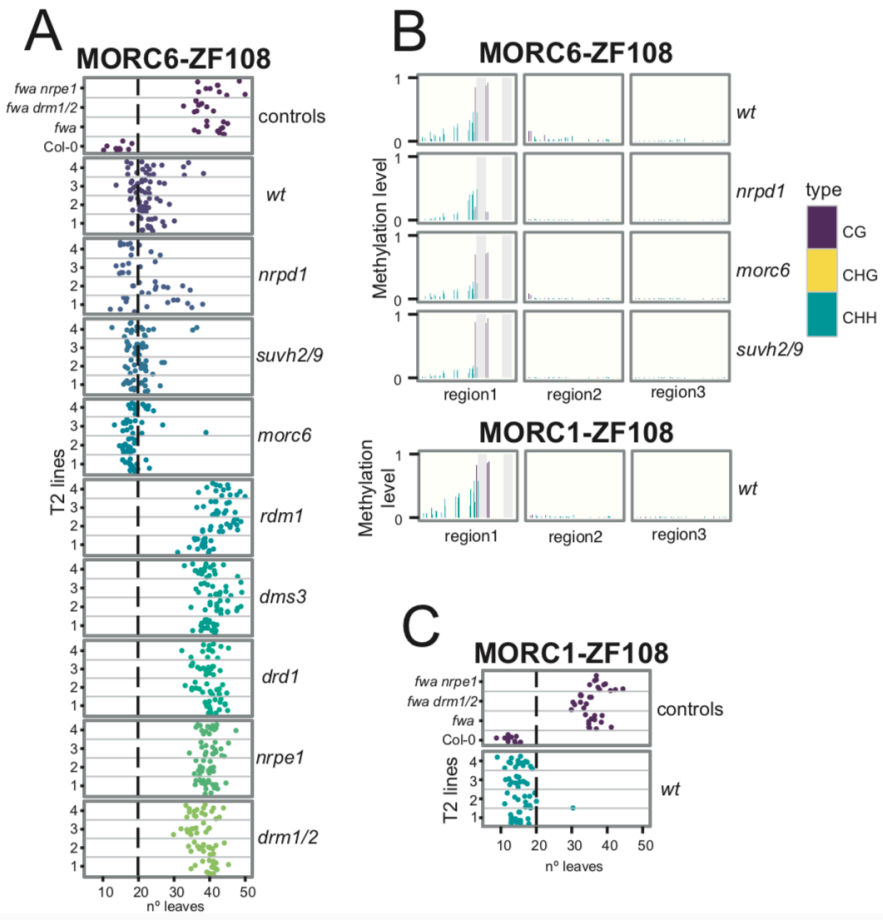

A
MORC6-ZF108
fwa nrpe1
fwa drm1/2
fwa
Col-0
controls
wt
nrpd1
suvh2/9
morc6
rdm1
dms3
drd1
nrpe1
drm1/2
T2 lines
n° leaves
B
MORC6-ZF108
Methylation level
region1
region2
region3
type
CG
CHG
CHH
MORC1-ZF108
C
MORC1-ZF108

Figure 2-6

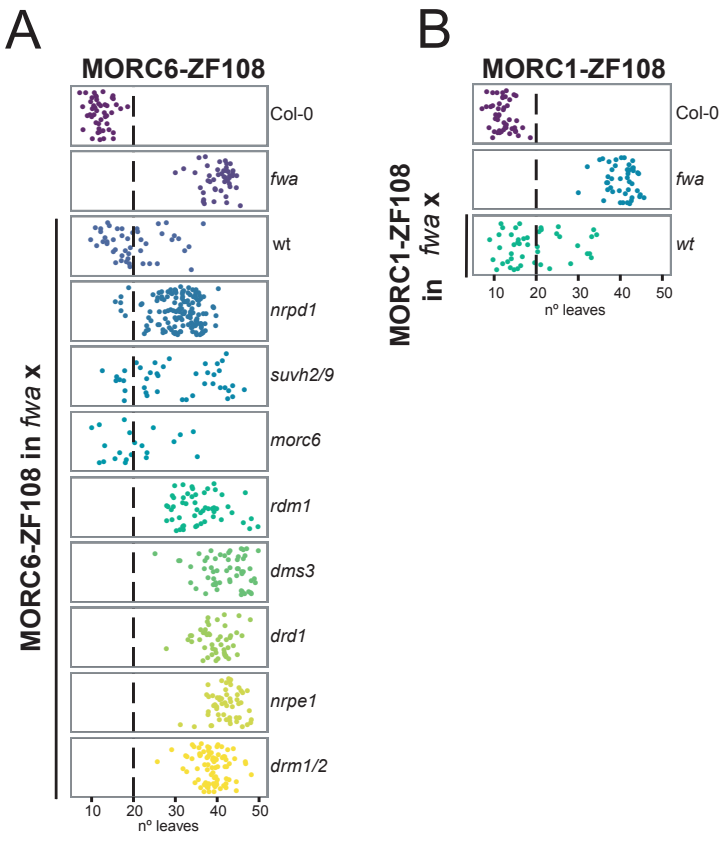

A
MORC6-ZF108
Col-0
fwa
wt
nrpd1
suvh2/9
morc6
rdm1
dms3
drd1
nrpe1
drm1/2
MORC6-ZF108 in fwa x
10
20
30
40
50
nº leaves
B
MORC1-ZF108
Col-0
fwa
wt
MORC1-ZF108 in fwa x
10
20
30
40
50
nº leaves

Figure 2-7

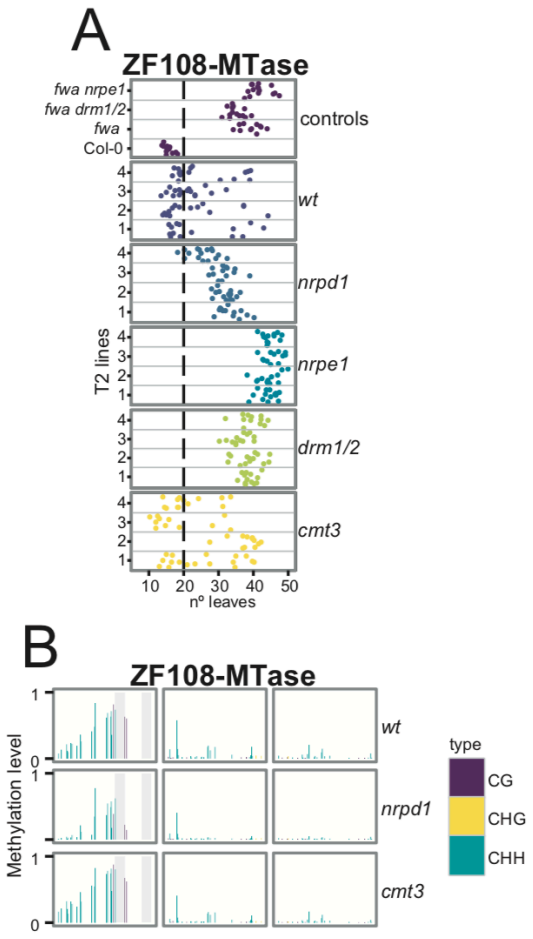

A
ZF108-MTase
fwa nrpe1
fwa drm1/2
fwa
Col-0
controls
wt
nrpd1
nrpe1
drm1/2
cmt3
T2 lines
nº leaves
B
ZF108-MTase
Methylation level
type
CG
CHG
CHH

Figure 2-8

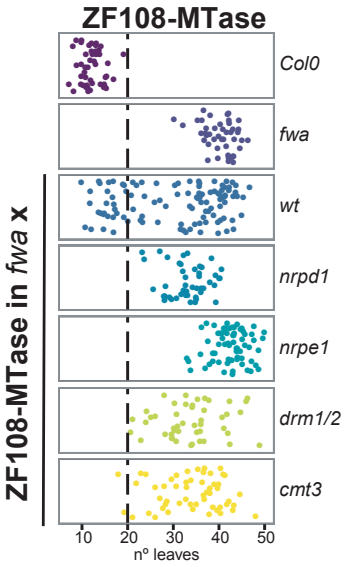

ZF108-MTase
ZF108-MTase in fwa x
Col0
fwa
wt
nrpd1
nrpe1
drm1/2
cmt3
10
20
30
40
50
n° leaves

Figure 2-9

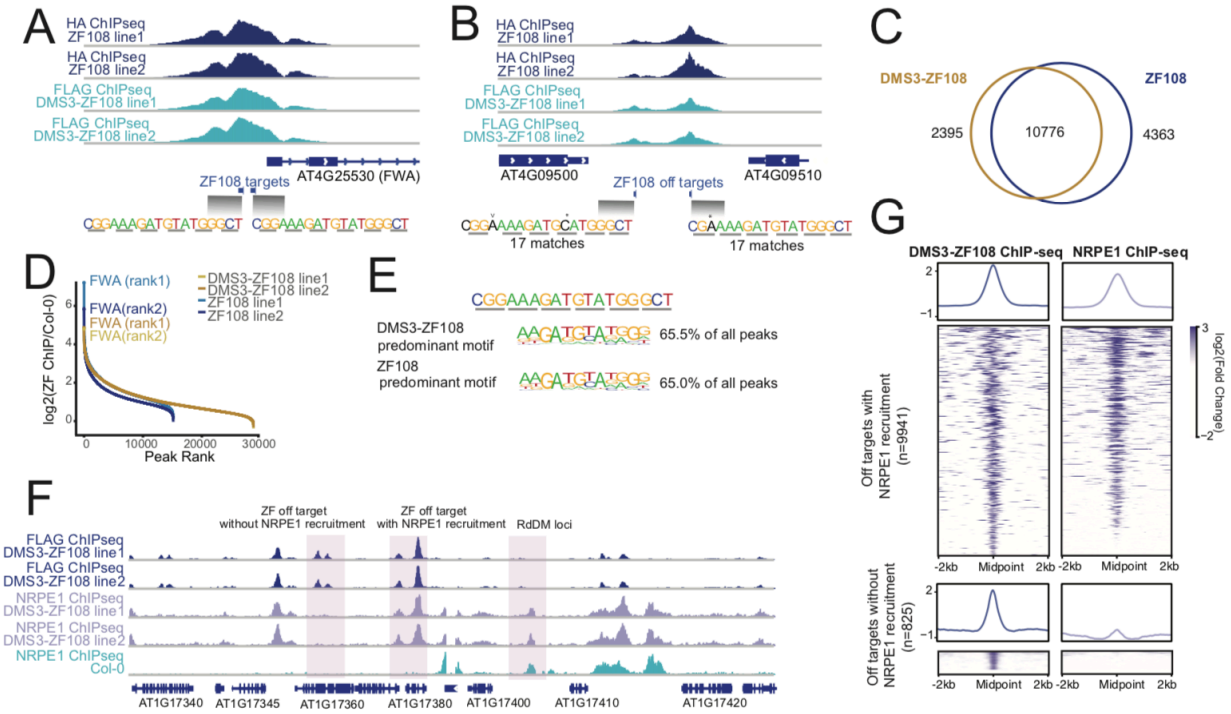

A
HA ChIPseq ZF108 line1
HA ChIPseq ZF108 line2
FLAG ChIPseq DMS3-ZF108 line1
FLAG ChIPseq DMS3-ZF108 line2
ZF108 targets
AT4G25530 (FWA)
B
AT4G09500
ZF108 off targets
AT4G09510
17 matches
17 matches
C
DMS3-ZF108
ZF108
2395
10776
4363
D
FWA (rank1)
FWA(rank2)
DMS3-ZF108 line1
DMS3-ZF108 line2
ZF108 line1
ZF108 line2
log2(ZF ChIP/Col-0)
Peak Rank
E
DMS3-ZF108 predominant motif
65.5% of all peaks
ZF108 predominant motif
65.0% of all peaks
F
ZF off target without NRPE1 recruitment
ZF off target with NRPE1 recruitment
RdDM loci
NRPE1 ChIPseq Col-0
AT1G17340
AT1G17345
AT1G17360
AT1G17380
AT1G17400
AT1G17410
AT1G17420
G
DMS3-ZF108 ChIP-seq
NRPE1 ChIP-seq
Off targets with NRPE1 recruitment (n=9941)
Off targets without NRPE1 recruitment (n=825)
log2(Fold Change)
-2kb
Midpoint
2kb

Figure 2-10

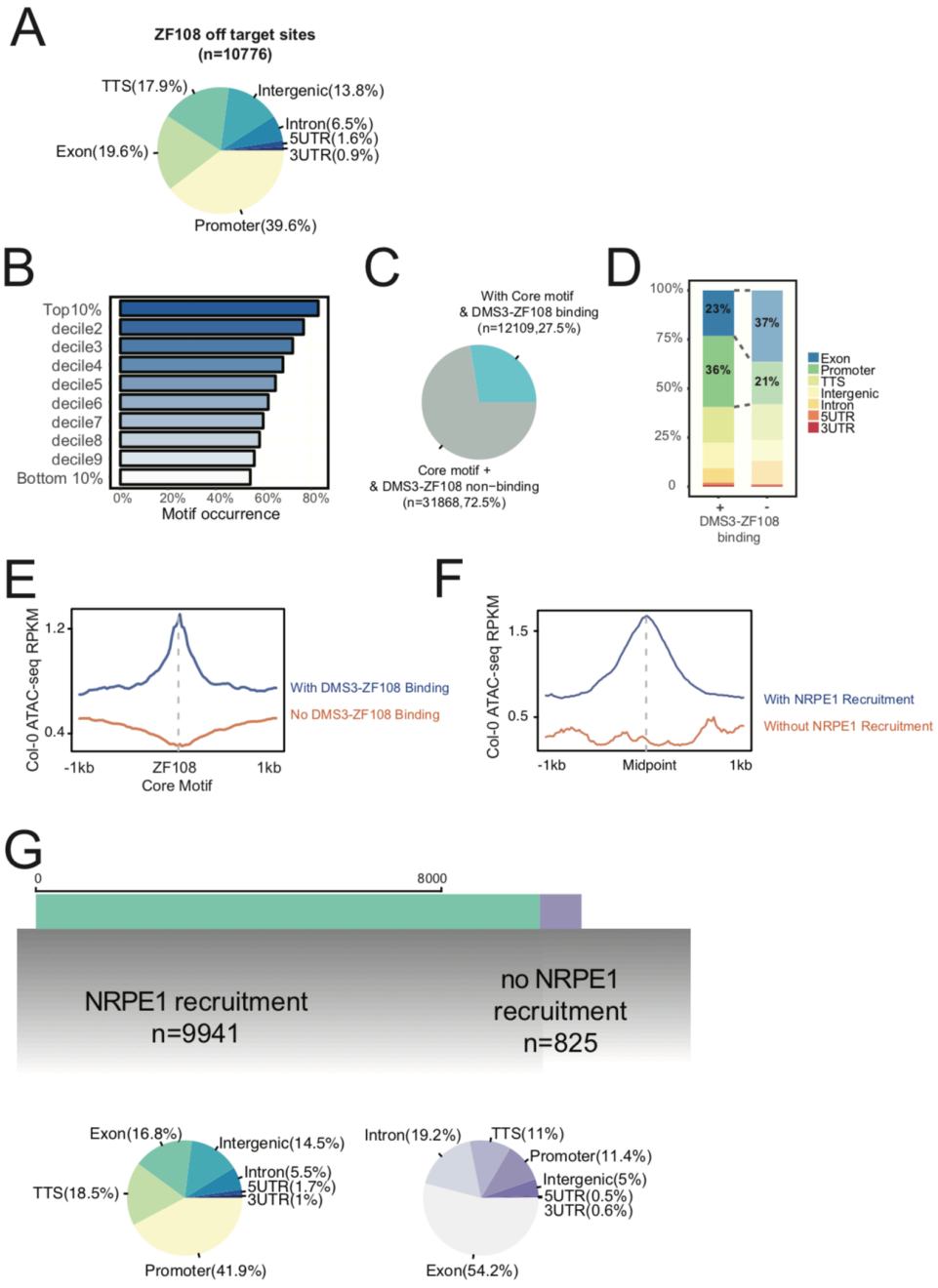

A
ZF108 off target sites
(n=10776)
TTS(17.9%)
Intergenic(13.8%)
Intron(6.5%)
5UTR(1.6%)
3UTR(0.9%)
Exon(19.6%)
Promoter(39.6%)
B
Top10%
decile2
decile3
decile4
decile5
decile6
decile7
decile8
decile9
Bottom 10%
0%
20%
40%
60%
80%
Motif occurrence
C
With Core motif
& DMS3-ZF108 binding
(n=12109,27.5%)
Core motif +
& DMS3-ZF108 non-binding
(n=31868,72.5%)
D
100%
75%
50%
25%
0
23%
37%
36%
21%
+
-
DMS3-ZF108
binding
Exon
Promoter
TTS
Intergenic
Intron
5UTR
3UTR
E
Col-0 ATAC-seq RPKM
1.2
0.4
-1kb
ZF108
Core Motif
1kb
With DMS3-ZF108 Binding
No DMS3-ZF108 Binding
F
Col-0 ATAC-seq RPKM
1.5
0.5
-1kb
Midpoint
1kb
With NRPE1 Recruitment
Without NRPE1 Recruitment
G
0
8000
NRPE1 recruitment
n=9941
no NRPE1
recruitment
n=825
Exon(16.8%)
Intergenic(14.5%)
Intron(5.5%)
5UTR(1.7%)
3UTR(1%)
TTS(18.5%)
Promoter(41.9%)
Intron(19.2%)
TTS(11%)
Promoter(11.4%)
Intergenic(5%)
5UTR(0.5%)
3UTR(0.6%)
Exon(54.2%)

Figure 2-11

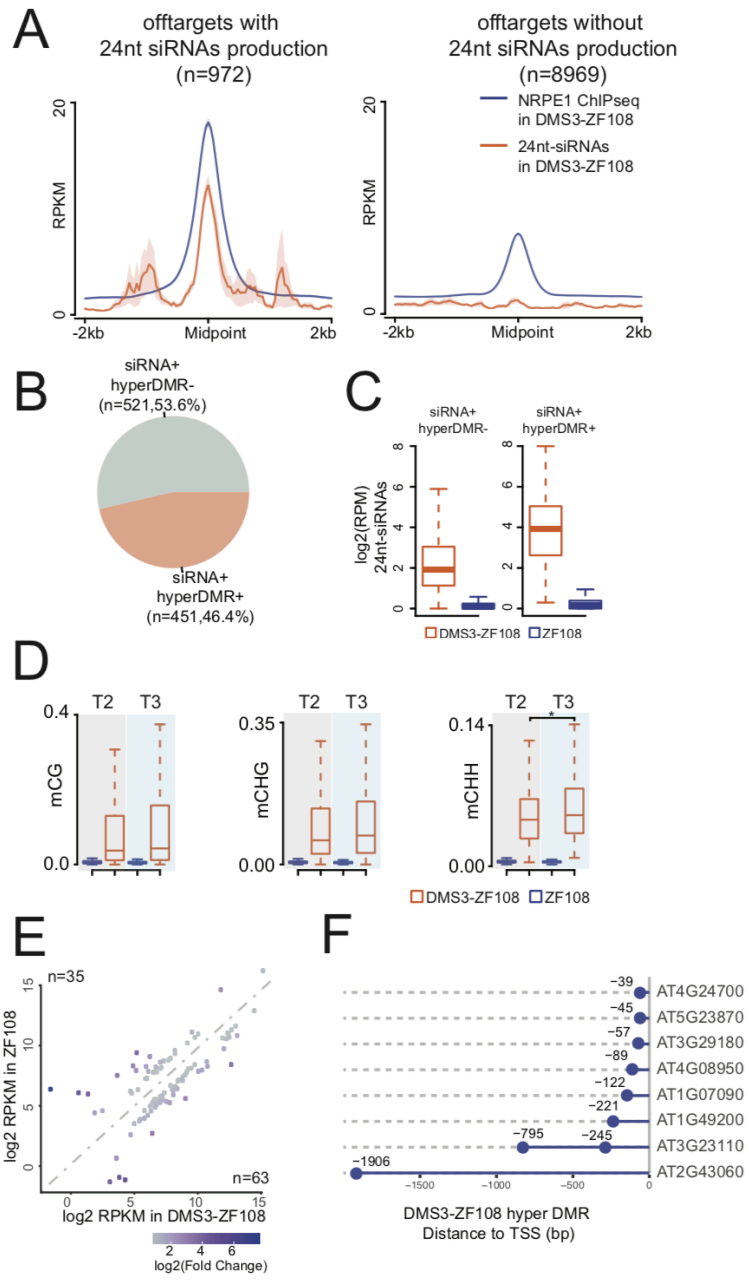

A
offtargets with
24nt siRNAs production
(n=972)
offtargets without
24nt siRNAs production
(n=8969)
NRPE1 ChIPseq
in DMS3-ZF108
24nt-siRNAs
in DMS3-ZF108
RPKM
-2kb
Midpoint
2kb
B
siRNA+
hyperDMR-
(n=521,53.6%)
siRNA+
hyperDMR+
(n=451,46.4%)
C
siRNA+
hyperDMR-
siRNA+
hyperDMR+
log2(RPM)
24nt-siRNAs
DMS3-ZF108
ZF108
D
T2
T3
mCG
mCHG
mCHH
DMS3-ZF108
ZF108
E
n=35
n=63
log2 RPKM in ZF108
log2 RPKM in DMS3-ZF108
log2(Fold Change)
F
AT4G24700
AT5G23870
AT3G29180
AT4G08950
AT1G07090
AT1G49200
AT3G23110
AT2G43060
DMS3-ZF108 hyper DMR
Distance to TSS (bp)

Figure 2-12

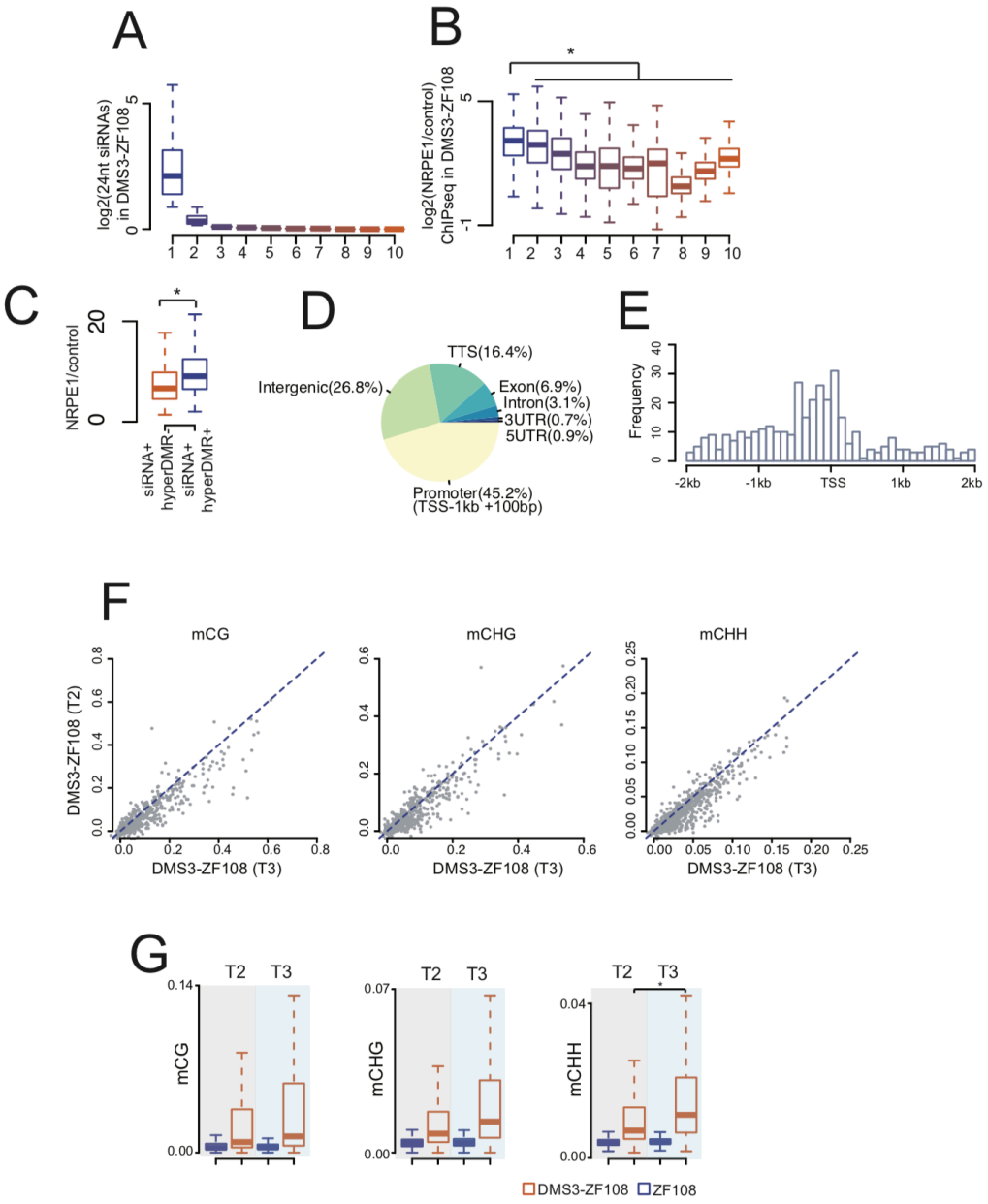

A
log2(24nt siRNAs) in DMS3-ZF108
B
log2(NRPE1/control) ChIPseq in DMS3-ZF108
C
NRPE1/control
siRNA+ hyperDMR-
siRNA+ hyperDMR+
D
TTS(16.4%)
Intergenic(26.8%)
Exon(6.9%)
Intron(3.1%)
3UTR(0.7%)
5UTR(0.9%)
Promoter(45.2%) (TSS-1kb +100bp)
E
Frequency
-2kb
-1kb
TSS
1kb
2kb
F
mCG
mCHG
mCHH
DMS3-ZF108 (T2)
DMS3-ZF108 (T3)
G
T2
T3
mCG
mCHG
mCHH
DMS3-ZF108
ZF108

Figure 2-13

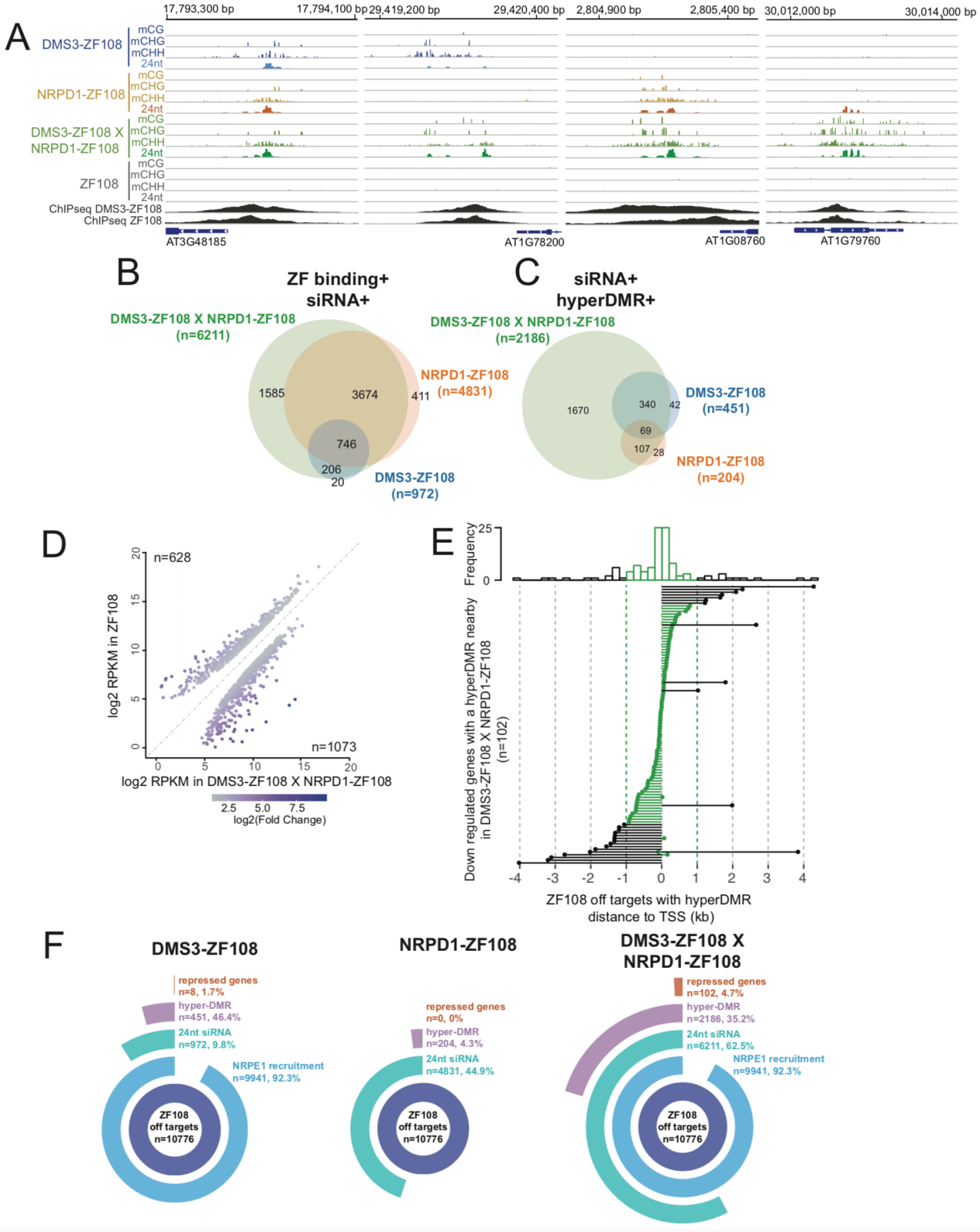
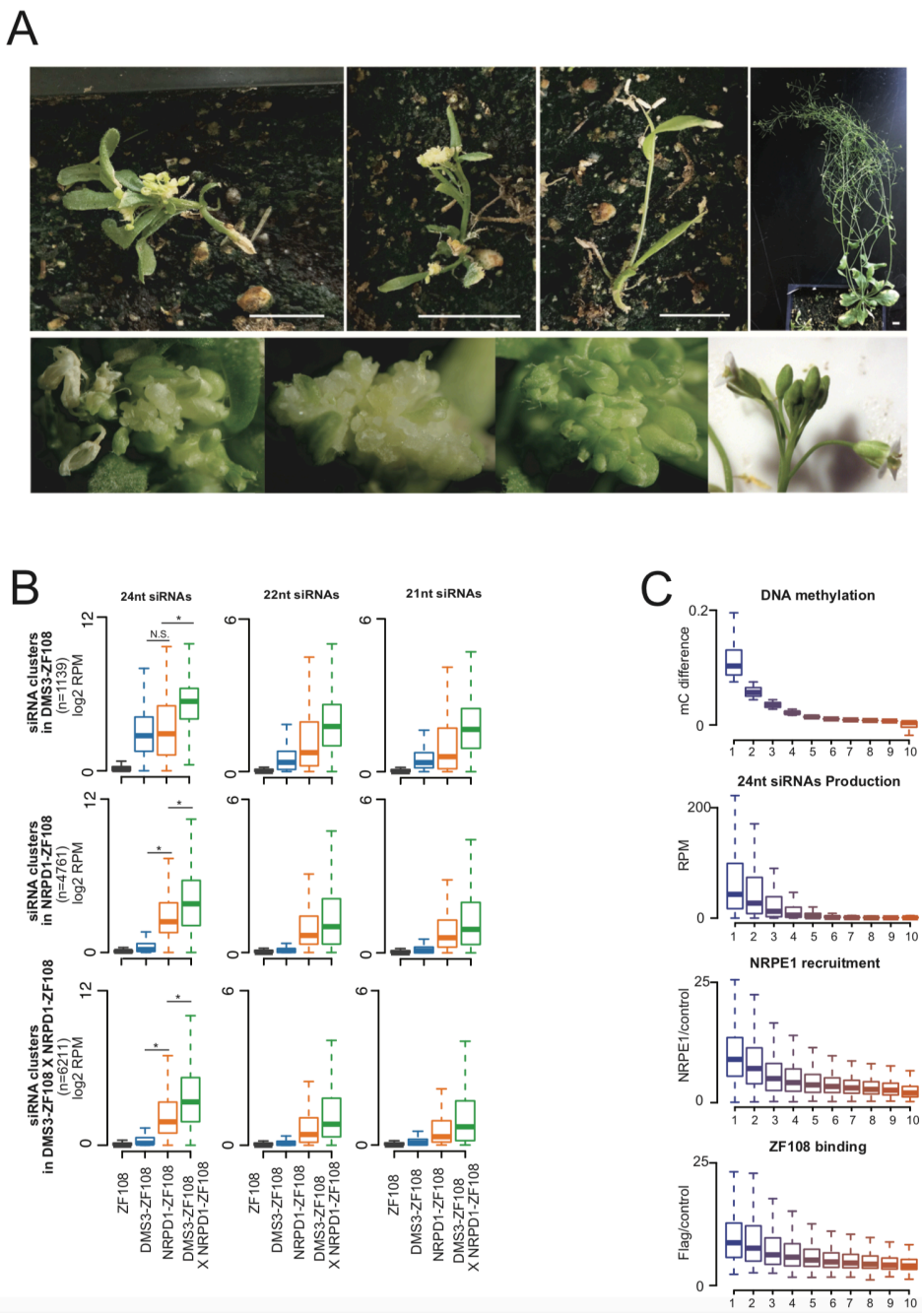

A
17,793,300 bp
17,794,100 bp
29,419,200 bp
29,420,400 bp
2,804,900 bp
2,805,400 bp
30,012,000 bp
30,014,000 bp
DMS3-ZF108
NRPD1-ZF108
DMS3-ZF108 X NRPD1-ZF108
ZF108
mCG
mCHG
mCHH
24nt
ChIPseq DMS3-ZF108
ChIPseq ZF108
AT3G48185
AT1G78200
AT1G08760
AT1G79760
B
ZF binding+ siRNA+
DMS3-ZF108 X NRPD1-ZF108 (n=6211)
NRPD1-ZF108 (n=4831)
DMS3-ZF108 (n=972)
1585
3674
411
746
206
20
C
siRNA+ hyperDMR+
DMS3-ZF108 X NRPD1-ZF108 (n=2186)
DMS3-ZF108 (n=451)
NRPD1-ZF108 (n=204)
1670
340
42
69
107
28
D
n=628
n=1073
log2 RPKM in ZF108
log2 RPKM in DMS3-ZF108 X NRPD1-ZF108
2.5 5.0 7.5
log2(Fold Change)
E
Frequency
Down regulated genes with a hyperDMR nearby in DMS3-ZF108 X NRPD1-ZF108 (n=102)
ZF108 off targets with hyperDMR distance to TSS (kb)
F
DMS3-ZF108
NRPD1-ZF108
DMS3-ZF108 X NRPD1-ZF108
repressed genes n=8, 1.7%
hyper-DMR n=451, 46.4%
24nt siRNA n=972, 9.8%
NRPE1 recruitment n=9941, 92.3%
repressed genes n=0, 0%
hyper-DMR n=204, 4.3%
24nt siRNA n=4831, 44.9%
repressed genes n=102, 4.7%
hyper-DMR n=2186, 35.2%
24nt siRNA n=6211, 62.5%
NRPE1 recruitment n=9941, 92.3%
ZF108 off targets n=10776

Figure 2-14


A
B
24nt siRNAs
22nt siRNAs
21nt siRNAs
siRNA clusters in DMS3-ZF108 (n=1139) log2 RPM
siRNA clusters in NRPD1-ZF108 (n=4761) log2 RPM
siRNA clusters in DMS3-ZF108 X NRPD1-ZF108 (n=6211) log2 RPM
N.S.
*
ZF108
DMS3-ZF108
NRPD1-ZF108
DMS3-ZF108 X NRPD1-ZF108
C
DNA methylation
mC difference
24nt siRNAs Production
RPM
NRPE1 recruitment
NRPE1/control
ZF108 binding
Flag/control

REFERENCES

Allen, E., Xie, Z., Gustafson, A.M., and Carrington, J.C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. Cell *121*, 207–221.

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

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Blevins, T., Podicheti, R., Mishra, V., Marasco, M., Tang, H., and Pikaard, C.S. (2015). Identification of Pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding de novo DNA methylation in Arabidopsis. Elife *4*, e09591.

Bond, D.M., and Baulcombe, D.C. (2015). Epigenetic transitions leading to heritable, RNA-mediated de novo silencing in Arabidopsis thaliana. Proc. Natl. Acad. Sci. U.S.a. *112*, 917–922.

Böhmdorfer, G., Sethuraman, S., Rowley, M.J., Krzyszton, M., Rothi, M.H., Bouzit, L., and Wierzbicki, A.T. (2016). Long non-coding RNA produced by RNA polymerase V determines boundaries of heterochromatin. Elife *5*, 1325.

Cao, X., and Jacobsen, S.E. (2002). Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing. Curr. Biol. *12*, 1138–1144.

Chan, S.W.-L., Zilberman, D., Xie, Z., Johansen, L.K., Carrington, J.C., and Jacobsen, S.E. (2004). RNA silencing genes control de novo DNA methylation. Science *303*, 1336–1336.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215–219.

Cuerda-Gil, D., and Slotkin, R.K. (2016). Non-canonical RNA-directed DNA methylation. Nat Plants *2*, 16163.

Curtis, M.D., and Grossniklaus, U. (2003). A gateway cloning vector set for high-throughput functional analysis of genes in planta. Plant Physiol. *133*, 462–469.

Dalakouras, A., Moser, M., Zwiebel, M., Krczal, G., Hell, R., and Wassenegger, M. (2009). A hairpin RNA construct residing in an intron efficiently triggered RNA-directed DNA methylation in tobacco. The Plant Journal *60*, 840–851.

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.

Gao, Z., Liu, H.-L., Daxinger, L., Pontes, O., He, X., Qian, W., Lin, H., Xie, M., Lorković, Z.J., Zhang, S., et al. (2010). An RNA polymerase II- and AGO4-associated protein acts in RNA-directed DNA methylation. Nature *465*, 106–109.

Greenberg, M.V.C., Ausin, I., Chan, S.W.-L., Cokus, S.J., Cuperus, J.T., Feng, S., Law, J.A., Chu, C., Pellegrini, M., Carrington, J.C., et al. (2011). Identification of genes required for de novo DNA methylation in Arabidopsis. Epigenetics *6*, 344–354.

Harris, C.J., Husmann, D., Liu, W., Kasmi, F.E., Wang, H., Papikian, A., Pastor, W.A., Moissiard, G., Vashisht, A.A., Dangl, J.L., et al. (2016). Arabidopsis AtMORC4 and AtMORC7 Form Nuclear Bodies and Repress a Large Number of Protein-Coding Genes. PLoS Genet. *12*, e1005998.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell *38*, 576–589.

Henderson, I.R., Deleris, A., Wong, W., Zhong, X., Chin, H.G., Horwitz, G.A., Kelly, K.A., Pradhan, S., and Jacobsen, S.E. (2010). The de novo cytosine methyltransferase DRM2 requires intact UBA domains and a catalytically mutated paralog DRM3 during RNA-directed DNA methylation in Arabidopsis thaliana. PLoS Genet. *6*, e1001182.

Henderson, I.R., Zhang, X., Lu, C., Johnson, L., Meyers, B.C., Green, P.J., and Jacobsen, S.E. (2006). Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning. Nature Genetics *38*, 721–725.

Herr, A.J., Jensen, M.B., Dalmay, T., and Baulcombe, D.C. (2005). RNA polymerase IV directs silencing of endogenous DNA. Science *308*, 118–120.

Ikeda, Y., Kobayashi, Y., Yamaguchi, A., Abe, M., and Araki, T. (2007). Molecular basis of late-flowering phenotype caused by dominant epi-alleles of the FWA locus in Arabidopsis. Plant Cell Physiol. *48*, 205–220.

Jeltsch, A., and Jurkowska, R.Z. (2016). Allosteric control of mammalian DNA methyltransferases - a new regulatory paradigm. Nucl. Acids Res. *44*, 8556–8575.

Jing, Y., Sun, H., Yuan, W., Wang, Y., Li, Q., Liu, Y., Li, Y., and Qian, W. (2016). SUVH2 and SUVH9 Couple Two Essential Steps for Transcriptional Gene Silencing in Arabidopsis. Mol Plant *9*, 1156–1167.

Johnson, L.M., Du, J., Hale, C.J., Bischof, S., Feng, S., Chodavarapu, R.K., Zhong, X., Marson, G., Pellegrini, M., Segal, D.J., et al. (2014). SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. Nature *507*, 124–128.

Johnson, L.M., Law, J.A., Khattar, A., Henderson, I.R., and Jacobsen, S.E. (2008). SRA-domain proteins required for DRM2-mediated de novo DNA methylation. PLoS Genet. *4*, e1000280.

Jones, L., Hamilton, A.J., Voinnet, O., Thomas, C.L., Maule, A.J., and Baulcombe, D.C. (1999). RNA-DNA interactions and DNA methylation in post-transcriptional gene silencing. The Plant Cell *11*, 2291–2301.

Kanno, T., Huettel, B., Mette, M.F., Aufsatz, W., Jaligot, E., Daxinger, L., Kreil, D.P., Matzke, M., and Matzke, A.J.M. (2005). Atypical RNA polymerase subunits required for RNA-directed DNA methylation. Nature Genetics *37*, 761–765.

Kinoshita, T., Miura, A., Choi, Y., Kinoshita, Y., Cao, X., Jacobsen, S.E., Fischer, R.L., and Kakutani, T. (2004). One-way control of FWA imprinting in Arabidopsis endosperm by DNA methylation. Science *303*, 521–523.

Lahmy, S., Pontier, D., Bies-Etheve, N., Laudié, M., Feng, S., Jobet, E., Hale, C.J., Cooke, R., Hakimi, M.-A., Angelov, D., et al. (2016). Evidence for ARGONAUTE4-DNA interactions in RNA-directed DNA methylation in plants. Genes Dev. *30*, 2565–2570.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics *11*, 204–220.

Law, J.A., Ausin, I., Johnson, L.M., Vashisht, A.A., Zhu, J.-K., Wohlschlegel, J.A., and Jacobsen, S.E. (2010). A protein complex required for polymerase V transcripts and RNA-directed DNA methylation in Arabidopsis. Curr. Biol. *20*, 951–956.

Law, J.A., Du, J., Hale, C.J., Feng, S., Krajewski, K., Palanca, A.M.S., Brian, S., Patel, D.J., and Jacobsen, S.E. (2014). 126_SHH1Nature. Nature *498*, 385–389.

Law, J.A., Vashisht, A.A., Wohlschlegel, J.A., and Jacobsen, S.E. (2011). SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. PLoS Genet. *7*, e1002195.

Li, C.F., Pontes, O., El-Shami, M., Henderson, I.R., Bernatavichute, Y.V., Chan, S.W.-L., Lagrange, T., Pikaard, C.S., and Jacobsen, S.E. (2006). An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in Arabidopsis thaliana. Cell *126*, 93–106.

Li, S., Vandivier, L.E., Tu, B., Gao, L., Won, S.Y., Li, S., Zheng, B., Gregory, B.D., and Chen, X. (2015). Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in Arabidopsis reveals features and regulation of siRNA biogenesis. Genome Res. *25*, 235–245.

Liu, W., Duttke, S.H., Hetzel, J., Groth, M., Feng, S., Gallego-Bartolome, J., Zhong, Z., Kuo, H.Y., Wang, Z., Zhai, J., et al. (2018). RNA-directed DNA methylation involves co-transcriptional small-RNA-guided slicing of polymerase V transcripts in Arabidopsis. Nat Plants *4*, 181–188.

Liu, Z.-W., Shao, C.-R., Zhang, C.-J., Zhou, J.-X., Zhang, S.-W., Li, L., Chen, S., Huang, H.-W., Cai, T., and He, X.-J. (2014). The SET domain proteins SUVH2 and SUVH9 are required for Pol V occupancy at RNA-directed DNA methylation loci. PLoS Genet. *10*, e1003948.

Liu, Z.-W., Zhou, J.-X., Huang, H.-W., Li, Y.-Q., Shao, C.-R., Li, L., Cai, T., Chen, S., and He, X.-J. (2016). Two Components of the RNA-Directed DNA Methylation Pathway Associate with MORC6 and Silence Loci Targeted by MORC6 in Arabidopsis. PLoS Genet. *12*, e1006026.

Lorković, Z.J., Naumann, U., Matzke, A.J.M., and Matzke, M. (2012). Involvement of a GHKL ATPase in RNA-directed DNA methylation in Arabidopsis thaliana. Curr. Biol. *22*, 933–938.

Lu, Z., Hofmeister, B.T., Vollmers, C., DuBois, R.M., and Schmitz, R.J. (2016). Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. Nucl. Acids Res.

Marí-Ordóñez, A., Marchais, A., Etcheverry, M., Martin, A., Colot, V., and Voinnet, O. (2013). Reconstructing de novo silencing of an active plant retrotransposon. Nature Genetics *45*, 1029–1039.

Matzke, M.A., Kanno, T., and Matzke, A.J.M. (2015). RNA-Directed DNA Methylation: The Evolution of a Complex Epigenetic Pathway in Flowering Plants. Annu Rev Plant Biol *66*, 243–267.

Mette, M.F., Aufsatz, W., van der Winden, J., Matzke, M.A., and Matzke, A.J. (2000). Transcriptional silencing and promoter methylation triggered by double-stranded RNA. Embo J. *19*, 5194–5201.

Mi, S., Cai, T., hu, Y., Chen, Y., Hodges, E., Fangrui, N., Liang, W., Shan, L., Huanyu, Z., Chengzu, L., et al. (2008). Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. Cell *133*, 116–127.

Moissiard, G., Bischof, S., Husmann, D., Pastor, W.A., Hale, C.J., Yen, L., Stroud, H., Papikian, A., Vashisht, A.A., Wohlschlegel, J.A., et al. (2014). Transcriptional gene silencing by Arabidopsis microrchidia homologues involves the formation of heteromers. Pnas *111*, 7474–7479.

Moissiard, G., Cokus, S.J., Cary, J., Feng, S., Billi, A.C., Stroud, H., Husmann, D., Zhan, Y., Lajoie, B.R., McCord, R.P., et al. (2012). MORC family ATPases required for heterochromatin condensation and gene silencing. Science *336*, 1448–1451.

Mosher, R.A., Schwach, F., Studholme, D., and Baulcombe, D.C. (2008). PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. Proc. Natl. Acad. Sci. U.S.a. *105*, 3145–3150.

Nuthikattu, S., McCue, A.D., Panda, K., Fultz, D., DeFraia, C., Thomas, E.N., and Slotkin, R.K. (2013). The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. Plant Physiol. *162*, 116–131.

Onodera, Y., Haag, J.R., Ream, T., Costa Nunes, P., Pontes, O., and Pikaard, C.S. (2005). Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. Cell *120*, 613–622.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics *15*, 284.

Sijen, T., Vijn, I., Rebocho, A., van Blokland, R., Roelofs, D., Mol, J.N., and Kooter, J.M. (2001). Transcriptional and posttranscriptional gene silencing are mechanistically related. Curr. Biol. *11*, 436–440.

Smith, L.M., Pontes, O., Searle, I., Yelina, N., Yousafzai, F.K., Herr, A.J., Pikaard, C.S., and Baulcombe, D.C. (2007). An SNF2 protein associated with nuclear RNA silencing and the spread of a silencing signal between cells in Arabidopsis. The Plant Cell *19*, 1507–1521.

Soppe, W.J.J., Jacobsen, S.E., Alonso-Blanco, C., Jackson, J.P., Kakutani, T., Koornneef, M., and Peeters, A.J.M. (2000). The Late Flowering Phenotype of fwa Mutants Is Caused by Gain-of-Function Epigenetic Alleles of a Homeodomain Gene. Molecular Cell *6*, 791–802.

Stroud, H., Greenberg, M.V.C., Feng, S., Bernatavichute, Y.V., and Jacobsen, S.E. (2013). 122_ComplexRegulation. Cell *152*, 352–364.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics *10*, 232.

Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D., Jacobsen, S.E., and Carrington, J.C. (2004). Genetic and functional diversification of small RNA pathways in plants. PLoS Biol. *2*, E104.

Zhai, J., Bischof, S., Wang, H., Feng, S., Lee, T.-F., Teng, C., Chen, X., Park, S.Y., Liu, L., Gallego-Bartolome, J., et al. (2015). A One Precursor One siRNA Model for Pol IV-Dependent siRNA Biogenesis. Cell *163*, 445–455.

Zhang, X., Henderson, I.R., Lu, C., Green, P.J., and Jacobsen, S.E. (2007). Role of RNA polymerase IV in plant small RNA metabolism. Pnas *104*, 4536–4541.

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

Zhong, X., Du, J., Hale, C.J., Gallego-Bartolome, J., Feng, S., Vashisht, A.A., Chory, J., Wohlschlegel, J.A., Patel, D.J., and Jacobsen, S.E. (2014). Molecular mechanism of action of plant DRM de novo DNA methyltransferases. Cell *157*, 1050–1060.

Zhong, X., Hale, C.J., Law, J.A., Johnson, L.M., Feng, S., Tu, A., and Jacobsen, S.E. (2012). DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. Nat. Struct. Mol. Biol. *19*, 870–875.

CHAPTER 3

Targeted DNA demethylation of the *Arabidopsis* genome using the human TET1 catalytic domain.

Contributions

J.G.B., J.G., W.L., A.P. and S.E.J. designed research. J.G.B., J.G., W.L., A.P., B.G., H.Y.K. and J.M.C.Z. performed research. J.G.B., J.G., W.L., A.P., D.J.S. and S.E.J analyzed data. J.G.B., J.G., W.L., A.P. and S.E.J. wrote the paper.

SIGNIFICANCE

DNA methylation is an epigenetic modification involved in gene silencing. Studies of this modification usually rely on the use of mutants or chemicals that affect methylation maintenance. Those approaches cause global changes in methylation and make difficult the study of the impact of methylation on gene expression or chromatin changes at specific loci. In this study, we develop tools to target DNA demethylation in plants. We report efficient on-target demethylation and minimal effects on global methylation patterns, and show that in one case, targeted demethylation is heritable. These tools can be used to approach basic questions about DNA methylation biology, as well as to develop new biotechnology strategies to modify gene expression and create new plant trait epialleles.

ABSTRACT

DNA methylation is an important epigenetic modification involved in gene regulation and transposable element silencing. Changes in DNA methylation can be heritable and thus, can lead to the formation of stable epialleles. A well characterized example of a stable epiallele in plants is *fwa*, which consists of the loss of DNA cytosine methylation (5mC) in the promoter of the *FLOWERING WAGENINGEN* (*FWA*) gene, causing upregulation of *FWA* and a heritable late flowering phenotype. Here we demonstrate that a fusion between the catalytic domain of the human demethylase TEN-ELEVEN TRANSLOCATION1 (TET1cd) and an artificial zinc finger (ZF) designed to target the *FWA* promoter can cause highly efficient targeted demethylation, *FWA* upregulation, and a heritable late flowering phenotype. Additional ZF-TET1cd fusions designed to target methylated regions of the *CACTA1* transposon also caused targeted demethylation and changes in expression. Finally, we have developed a CRISPR/dCas9 based targeted demethylation system using the TET1cd and a modified SunTag system. Similar to the ZF-TET1 fusions, the SunTag-TET1cd system is able to target demethylation and activate gene expression when directed to the *FWA* or *CACTA1* loci. Our study provides tools for targeted removal of 5mC at specific loci in the genome with high specificity and minimal off-target effects. These tools provide the opportunity to develop new epialleles for traits of interest, and to reactivate expression of previously silenced genes, transgenes, or transposons.

INTRODUCTION

DNA methylation is involved in silencing genes and transposable elements (TE). In contrast to many organisms where methylation is largely erased and re-established in each generation (Bogdanović and Lister, 2017), changes in DNA methylation patterns in plants can be transmitted through sexual generations to establish stable epigenetic alleles (Agrawal et al., 1999; Cubas et al., 1999; Jacobsen and Meyerowitz, 1997; Kakutani, 1997; Soppe et al., 2000b). For example, complete loss of 5mC in the promoter of the *FWA* gene causes stable *fwa* epialleles that have been found in flowering time mutant screens (Koornneef et al., 1991) and in strong DNA methylation mutants (Kakutani, 1997; Kankel et al., 2003; Soppe et al., 2000a). This loss of 5mC at the *FWA* promoter activates *FWA* expression that is responsible for the late flowering phenotype observed in *fwa* epialles (Soppe et al., 2000a). DNA methylation in plants occurs in different cytosine contexts -CG, CHG and CHH- (where H is A, T, C) and is controlled by different DNA methyltransferases (DNMTs). MET1, a homolog of DNMT1, is responsible for the maintenance of symmetric methylation in the CG context (Law and Jacobsen, 2010). CMT3 and CMT2 are responsible for the maintenance of CHG and CHH methylation, respectively, at pericentromeric regions and long TEs (Bartee et al., 2001; Lindroth et al., 2001; Stroud et al., 2014; Zemach et al., 2013). Lastly, DRM2, a homolog of DNMT3, is involved in the maintenance of CHH at borders of long TEs in pericentromeric heterochromatin as well as small TEs in euchromatin (Zemach et al., 2013) (Cao and Jacobsen, 2002; Cao et al., 2003; Stroud et al., 2014; 2013) , and represents the last step of the *de novo* methylation pathway in plants called RNA-directed DNA methylation (RdDM) (Matzke et al., 2015). Plants also have an active DNA demethylation system driven by ROS1 and 3 other related glycosylase/lyase enzymes (Gong et al., 2002; Penterman et al., 2007; Zhu et al., 2007). These enzymes recognize DNA

methylcytosines and initiate DNA demethylation through a base excision repair process(Zhang and Zhu, 2012).

Thus far, studies aiming to understand the effect of DNA methylation on gene expression have relied on the use of mutants defective in genes involved in the DNA methylation machinery, or chemicals to inhibit methylation maintenance such as 5-azacytidine or zebularine (Baubec et al., 2009; Griffin et al., 2016; Kankel et al., 2003; Taylor and Jones, 1982). Both approaches, genetic and chemical, have the disadvantage of affecting DNA methylation at a genome-wide scale making it difficult to study the impact of DNA methylation on gene expression and chromatin architecture at specific loci. Therefore, it is important to create tools in plants that allow the manipulation of DNA methylation in a more locus-specific manner.

A previous study in *Arabidopsis* has shown that a fusion of the RdDM component SUVH9 to an artificial zinc finger (ZF108) designed to target the *FWA* promoter is able to target methylation to the *FWA* promoter, silencing *FWA* expression and rescuing the late flowering phenotype of the *fwa-4* epiallele (Johnson et al., 2014). Unfortunately, no equivalent tool has been developed in plants for targeted DNA demethylation.

In animals, controlled removal of 5mC by TEN-ELEVEN TRANSLOCATION1 (TET1) has been achieved through targeting the human TET1 catalytic domain (TET1cd) to specific regions of the genome by fusing it to DNA binding domains such as ZFs, TAL effectors or CRISPR-dCas9 (Amabile et al., 2016; Chen et al., 2014; Choudhury et al., 2016; Liu et al., 2016; Lo et al., 2017; Maeder et al., 2013; Morita et al., 2016; Okada et al., 2017; Xu et al., 2016).

TET1 causes demethylation of DNA through oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Wu and Zhang, 2017). This is followed by either the passive removal of methylation through the failure

of DNA methylation maintenance after DNA replication or the active removal of DNA methylation by glycosylase-mediated base excision repair (Kohli and Zhang, 2013). While plants do not contain TET enzymes, a previous study has shown that overexpression of the human TET3 catalytic domain in *Arabidopsis* can cause changes in DNA methylation levels at rDNA loci (Hollwey et al., 2016). However, both hypermethylation and hypomethylation were observed in this study making the results difficult to interpret, and only effects at rDNA loci were examined. This suggests that while TET enzymes may potentially be used in plants to manipulate DNA methylation, improved strategies are needed in order to use TET enzymes to manipulate 5mC in a locus-specific manner.

In this work, we describe the development of different tools to target locus-specific DNA demethylation in *Arabidopsis*. We have fused human TET1cd to artificial zinc fingers designed to target two different loci in the *Arabidopsis* genome. We have also adapted the CRISPR/dCas9 SunTag system to target DNA demethylation in plants (Morita et al., 2016; Tanenbaum et al., 2014). Using both targeting platforms —ZF or SunTag— we observe precise DNA demethylation and associated changes in gene expression over the targeted regions with only small effects on genome-wide methylation or gene expression. The development of tools for targeted demethylation in plants creates exciting avenues for the study of locus-specific effects of DNA methylation on gene expression and chromatin landscape. These tools should also allow for the generation of new epialleles, and the manipulation of TE expression levels to create insertional mutations and study genome evolution.

RESULTS

Expression of ZF108-TET1cd causes late flowering and *FWA* activation.

In animals, targeted removal of 5mC has been achieved by using the human TET1cd (Amabile et al., 2016; Chen et al., 2014; Choudhury et al., 2016; Liu et al., 2016; Maeder et al., 2013; Morita et al., 2016; Xu et al., 2016) (Lo et al., 2017; Okada et al., 2017). In order to test if TET1cd can be used in plants for targeted demethylation, we fused human TET1cd to ZF108 and expressed the fusion under the control of the constitutive *UBIQUITIN 10* (*UBQ10*) promoter from *Arabidopsis* (Figure 3-1A). ZF108 was previously shown to target DNA methylation to the promoter of the *FWA* gene when fused to the RdDM component SUVH9 (Johnson et al., 2014). The *FWA* promoter is normally methylated in wild type Col-0 plants, causing silencing of *FWA*. Demethylation of the promoter in *met1* mutants or *fwa-4* epialleles is heritable over generations, triggers the ectopic expression of *FWA* and causes a late flowering phenotype (Soppe et al., 2000b). Therefore, this methylation-dependent visual phenotype can be exploited as a readout for succesful targeted demethylation. We screened T1 plants expressing ZF108-TET1cd in the Col-0 background and found 25 out of 57 that displayed a late flowering phenotype suggesting *FWA* activation (Figure 3-1B,C).

We then studied the stability of the late flowering phenotype over generations by analyzing the flowering time of T3 lines that either retained the ZF108-TET1cd transgene (T3+) or had the transgene segregated away in the T2 generation (T3-). Both T3+ and T3- lines retained a late flowering phenotype, consistent with a loss of methylation at the *FWA* promoter. Importantly, control plants expressing a fusion of ZF108 to the fluorescent protein YPet (ZF108-YPet) (Nguyen and Daugherty, 2005) did not show any effect on flowering time, suggesting that the late flowering phenotype observed is not simply a consequence of ZF108 binding to the *FWA* promoter (Figure 3-1D).

To test if the late flowering phenotype observed was due to *FWA* upregulation, we performed RNA-seq of Col-0, *fwa-4* and four representative late flowering T1 plants expressing ZF108-TET1cd (Figure 3-1E), as well as 4 biological replicates of Col-0, and two representative T3 lines expressing ZF108-TET1cd or ZF108-YPet (Figure 3-1F). *FWA* expression was dramatically increased in ZF108-TET1cd as compared to Col-0 and ZF108-YPet and had a similar expression level as *fwa-4,* indicating that the late flowering phenotype observed was due to *FWA* overexpression (Figure 3-1E,F). A genome-wide gene expression analysis showed very few additional changes and revealed *FWA* as the most upregulated gene in the ZF108-TET1cd lines as compared to ZF108-YPet control lines (Figure 3-1G). These results suggest successful removal of methylation at the *FWA* promoter and, importantly, very few off-target effects due to ZF108-TET1cd expression.

Targeted demethylation at the *FWA* promoter is specific and heritable.

We then analyzed methylation levels at the *FWA* promoter by McrBC digestion in different ZF108-TET1cd late flowering T1 plants. All lines showed a large reduction in DNA methylation, similar to that observed in *fwa-4* plants (Figure 3-2). To confirm these results, we performed whole genome bisulfite sequencing (WGBS) of Col-0, four representative T1 ZF108-TET1cd plants, as well as two representative T3 ZF108-TET1cd lines including one T3+ and T3-. We observed complete demethylation over the *FWA* promoter in all four representative T1 lines, resembling the methylation pattern seen in *fwa-4* (Figure 3-3A, Figure 3-4A). Also, both T3+ and T3- lines showed complete demethylation of the *FWA* promoter (Figure 3-3A, Figure 3-4A), indicating that the targeted DNA demethylation is heritable, even in the absence of the transgene. Interestingly, loss of methylation spanned the entire methylated region of the *FWA*

promoter—approximately 500 base pairs—including cytosines a few hundred base pairs away from the ZF108 binding site. To assess the specificity of TET1cd mediated demethylation, we looked at methylation levels in a larger region flanking the *FWA* gene (Figure 3-4B), as well as analyzed genome-wide methylation levels (Figure 3-3B,C, Figure 3-5). We found that genome-wide CG, CHG, and CHH methylation levels were very similar to the wild type Col-0 control, indicating that targeted demethylation using ZF108-TET1cd was very specific. These results are consistent with the RNA-seq results presented in Figure 3-1G that showed very few changes in genome-wide expression patterns in plants expressing ZF108-TET1cd as compared to controls.

Targeted demethylation at the *CACTA1* promoter using ZFCACTA1-TET1cd fusions.

To test the ability of the ZF-TET1cd fusions to target demethylation at a heterochromatic locus, we fused TET1cd to two ZFs (ZF1CACTA1 and ZF2CACTA1) designed to target the promoter region of *CACTA1*, a TE that resides in an area of the genome with a very high level of DNA methylation and H3K9me2 (Kato et al., 2004; Miura et al., 2004). Five and nine independent T1 plants containing ZF1CACTA1-TET1cd and ZF2CACTA1-TET1cd, respectively, were screened for demethylation at the *CACTA1* promoter by McrBC. The ZF1CACTA1-TET1cd and ZF2CACTA1-TET1cd T1 lines showing the greatest demethylation were selected for further analysis by WGBS (Figure 3-6A,B). Compared to Col-0, ZF1CACTA1-TET1cd and ZF2CACTA1-TET1cd plants showed a loss of methylation in all three sequence contexts that extended up to 2 kilobases upstream of the ZF binding sites (Figure 3-6A,B). To assess the specificity of ZF1CACTA1-TET1cd and ZF2CACTA1-TET1cd targeted demethylation, we analyzed genome-wide methylation levels (Figure 3-7A) and methylation over all protein coding genes or TEs (Figure 3-7B). We found that methylation across the entire

genome was slighlty reduced as compared to the Col-0 control in both the ZF1CACTA1-TET1cd and ZF2CACTA1-TET1cd lines indicating a partial non-specific global demethylation.

Next, we performed RNA-seq to test if targeted demethylation had an impact on *CACTA1* expression. In both lines tested, a significant increase in *CACTA1* transcript levels was observed (Figure 3-6C), indicating that targeted demethylation at this region is sufficient to reactivate *CACTA1* expression.

To test heritability of targeted demethylation in these lines, we performed WGBS on T2 plants containing the transgene (+) or T2 plants that had segregated it away (-). Plants that had lost the ZF-TET1cd transgenes showed re-establishment of methylation to levels similar to Col-0 control (Figure 3-6D,E). This is in contrast to *FWA*, where methylation loss was stable in the absence of the transgene, and is likely a consequence of the incomplete removal of DNA methylation at the *CACTA1* region that is then able to attract the methylation machinery through self-reinforcing mechanisms (25). To study if this recovery of methylation in the absence of the transgene translates to the re-silencing of *CACTA1*, we analysed the expression of *CACTA1* in ZF1CACTA1-TET1cd (+) and (-) plants. Consistent with the methylation levels observed, *CACTA1* expression was detected in the presence of the transgene, while its expression was silenced to wild-type levels in the absence of ZF1CACTA1-TET1cd (Figure 3-6F).

Interestingly, T2 plants containing the transgenes showed an increase in global demethylation compared to T1 plants (Figure. 3-7A,C), indicating that the continuous presence of the transgene over generations can increase genome wide effects. Moreover, consistent with the recovery of methylation in the absence of the transgene observed within the *CACTA1* region, global methylation returned to wild-type levels when the transgene was lost (Figure. 3-7C,D).

Targeted demethylation at the *FWA* promoter using SunTag-TET1cd

While ZFs can efficiently target demethylation to specific loci in the genome, the design of new ZFs can be laborious and expensive. We therefore developed a plant-optimized CRISPR/dCas9-based SunTag-TET1cd system similar to one previously used to target demethylation in animals and shown to be more effective than direct fusions of TET1cd to dCas9 (Morita et al., 2016). In this system, dCas9 is fused to a C-terminal tail containing a variable number of tandem copies of peptide epitopes. In a separate module, a single chain variable fragment (scFv) antibody that recognizes the peptide epitopes is fused to a superfolder-GFP (sfGFP) followed by an effector protein (Tanenbaum et al., 2014) (Figure 3-8A). We adapted the SunTag-TET1cd system for use in *Arabidopsis* by expressing both the dCas9 and the scFv modules under the control of the constitutive *UBQ10* promoter. We created two versions of the epitope tail fused to dCas9, one containing a 22 amino acid linker separating each epitope similar to the one used in Morita et al. 2016, and one containing a 14 amino acid linker separating each epitope (Figure 3-8A). To preserve the components used in previous successful SunTag constructs (Morita et al., 2016), we cloned TET1cd downstream of the scFv-sfGFP module, and added two SV40 type NLSs to allow plant nuclear localization. We utilized a single gRNA (FWAgRNA-4) driven by the U6 promoter designed to target the ZF108 binding sequence in the *FWA* promoter. Two out of nine Col-0 transgenic plants containing SunTag-FWAgRNA-4-22aa-TET1cd (SunTagFWAg4-22aa) and two out of three Col-0 transgenic plants containing SunTag- FWAgRNA-4-14aa-TET1cd (SunTagFWAg4-14aa) showed a late flowering phenotype. Consistent with this phenotype, RNA-seq on two SunTagFWAg4-22aa and SunTagFWAg4-14aa T1 late flowering plants showed dramatic *FWA* overexpression similar to that of *fwa-4* (Figure 3-8B). Also, quantification of the flowering time of a representative T2 line expressing SunTagFWAg4-22aa

and one expressing SunTagFWAg4-14aa confirmed a late flowering phenotype similar to *fwa-4* plants (Figure 3-8C).

We then performed WGBS on two SunTagFWAg4-22aa and SunTagFWAg4-14aa T1 lines and T2 progeny that had the transgene (+) or had it segregated away (-) (Figure 3-8D, Figure 3-9A,B). In all cases, we observed efficient demethylation at the *FWA* promoter that was stable in the absence of the transgenes, suggesting that both SunTagFWAg4-22aa and SunTagFWAg4-14aa are able to target heritable demethylation at the *FWA* promoter. In order to study potential off-target effects, we examined methylation levels in a wider region surrounding *FWA* (Figure 3-9B), and also analyzed genome-wide methylation (Figure 3-8E, Figure 3-10A-C). Methylation levels over regions flanking *FWA* did not show significant changes compared to Col-0 (Figure 3-9B). Similarly, genome-wide DNA methylation levels were similar between the SunTagFWAg4-22aa and SunTagFWAg4-14aa plants and Col-0 control (Figure 3-7E, Figure 3-9A-C).

Targeted demethylation at the *CACTA1* promoter using SunTag-TET1cd

To test the ability of the SunTag22aa-TET1cd fusion to target demethylation at a heterochromatic locus, we designed a single gRNA (CACTA1g2) driven by the U6 promoter designed to target the same region that we targeted with the ZFCACTA1-TET1cd fusions (Figure 3-6). Six T1 plants containing SunTagCACTA1g2-22aa-TET1cd (SunTagCACTA1g2-22aa) were screened for demethylation at the *CACTA1* promoter by McrBC. The two plants showing the greatest demethylation were selected for further analysis by WGBS (Figure 3-11A). Consistent with the results obtained with ZFCACTA1, SunTagCACTA1g2-22aa plants showed a loss of methylation in all three sequence contexts that extended up to 2 kilobases upstream of the

gRNA binding site (Figure 3-11A,B), causing the upregulation of *CACTA1* expression (Figure 3-11C). Moreover, genome-wide methylation analysis indicated no observable differences between wild-type Col-0 and the SunTagCACTA1g2-22aa lines (Figure 3-11D, Figure 3-12). Overall, these results confirm that the SunTag approach is effective for targeting demethylation in plants without a major effect on global methylation levels. We also tested the impact of expressing our SunTag-TET1cd systems in wild-type Col-0 plants in the absence of a gRNA that directs the construct to a specific location. Flowering time of T1 plants expressing these constructs was unaffected (Figure 3-13A). Also, methylation levels at the *FWA* promoter or *CACTA1* region were similar to Col-0 (Figure 3-13B,C), and global methylation levels did not show any significant differences as compared to a Col-0 control (Figure 3-13D).

DISCUSSION

In this work, we present two independent methods for targeting DNA demethylation in *Arabidopsis*. We first fused the human TET1cd to an artificial zinc finger protein —ZF108— designed to target the *FWA* promoter. Col-0 plants expressing this construct showed highly specific demethylation and reactivation of *FWA* with virtually no genome-wide effects on DNA methylation or gene expression. In *Arabidopsis* plants grown under long day conditions (16h light/8d dark), flowering is established around 10-12 days after germination (Kardailsky et al., 1999). The fact that T1 plants expressing ZF108-TET1cd showed a late flowering phenotype indicates that demethylation of the *FWA* promoter occured during the early stages of development of the T1 plants. Surprisingly, the targeted demethylation at *FWA* comprised a large region, almost 500bp surrounding the ZF108 binding site. This could be due to direct access of ZF108-TET1cd to these cytosines. Another possibility is that loss of methylation in the distal

regions from the ZF108 binding site is a secondary effect of *FWA* reactivation or three-dimensional chromatin conformation that would place distal regions in proximity to the the targeted region where ZF108-TET1cd is bound.

We also generated new zinc fingers to target the promoter of the *CACTA1* transposable element whose expression is also controlled by DNA methylation (Kato et al., 2004). Importantly, this locus is located in pericentromeric heterochromatin which is associated with long stretches of chromatin that are highly enriched in DNA methylation and H3K9 methylation (Miura et al., 2004), which may represent a more challenging environment for targeted approaches. Two different ZFs targeting TET1cd to the *CACTA1* promoter triggered loss of methylation up to 2 kilobases away from the ZF binding sites. This data, together with the data obtained for *FWA,* indicates that ZF fusions to TET1cd can cause demethylation hundreds of base pairs away from the targeted sequence.

Contrary to the heritable loss of methylation in the *FWA* promoter, targeted demethylation at *CACTA1* disappeared when the ZFCACTA1 transgenes were segregated away, showing that unlike *FWA*, methylation was quickly re-established. The most likely explanation for this is that, contrary to the complete demethylation of the entire *FWA* methylated region, the incomplete demethylation of *CACTA1* leaves enough residual methylation to attract the RdDM machinery, probably via recruitment of Pol V by the methyl DNA binding proteins SUVH2 and SUVH9 (Johnson et al., 2014). In addition, the MET1 CG methyltransferase would likely perpetuate and potentially amplify any remaining methylated CG sites. In this scenario, heritable demethylation might be more efficiently achieved by targeting the TET1cd to multiple adjacent locations to achieve a more complete demethylation. Alternatively, *CACTA1* remethylation may occur because other methylated regions in the genome with sequences homologous to *CACTA1*

may be able to efficiently target remethylation *in trans* via siRNAs. Additional targeting experiments will be needed to determine frequency with which targeted demethylation can be heritable.

While targeted demethylation using ZF108-TET1cd was very specific and showed negligible changes in genome-wide methylation as compared to Col-0, lines expressing ZF1CACTA1-TET1cd, or ZF2CACTA1-TET1cd showed a varying amount of genome-wide hypomethylation. This variability highlights the importance to be selective with different ZFs, protein fusions, expression levels, and insertion events when using TET1cd to avoid genome-wide effects.

We also created a plant-optimized version of the SunTag TET1cd system and showed that it can be successfully implemented in plants for targeted DNA demethylation at the *FWA* and *CACTA1* loci. Similar to the results obtained using ZFs, we observed very high on-target demethylation and gene activation, with small effects on genome-wide methylation levels. Resembling the ZF-TET1cd fusions, the demethylation extended well beyond the targeted region reaching a region of approximately 2kb in the case of SunTagCACTA1g2 lines. Morita et al., 2016 reported that SunTag-TET1cd could also demethylate more than 200bp in mammalian cells. In this case it is reasonable to think that the TET1cd may be able to directly access long stretches of DNA considering the extension of the long epitope tail and the simultaneous recruitment of many molecules of TET1cd.

In summary, our results show highly efficient targeted demethylation in plants by using artificial zinc fingers or SunTag fused to TET1cd with limited off-target effects. As a result of their efficiency and specificity, they provide an ideal way to study the role of DNA methylation at specific loci and circumvent the need to use DNA methylation mutants or chemicals that

reduce methylation. Moreover, these tools may allow for the creation of new stable epialleles with traits of interest by activating genes normally silenced by DNA methylation. Other potential uses are for the reactivation of specific classes of transposons or the reactivation of previously silenced transgenes.

MATERIALS AND METHODS

Plant material and growth conditions

All the plants used in this study were in the Columbia-0 ecotype (Col-0) and were grown under long day conditions. The *fwa-4* epiallele was selected from a *met1* segregating population (Johnson et al., 2014). Transgenic plants were obtained by agrobacterium-mediated floral dipping(Clough and Bent, 1998). Plants were selected on 1/2 MS medium + Glufosinate 50 µg/ml (Goldbio), ½ MS medium + Hygromycin B 25 µg/ml (Invitrogen), or sprayed with Glufosinate (1:2000 Finale in water). Flowering time was scored by counting the total number of rossette and caulinar leaves.

Zinc finger design and cloning

Cloning of pUBQ10_ZF108_TET1cd.

For this purpose, a modified pMDC123 plasmid (Curtis and Grossniklaus, 2003) was created, containing 1990bp of the promoter region of the *Arabidopsis UBQ10* gene upstream of a cassette containing the ZF108, previously described in Johnson et al, 2014, and a 3xFlag tag. Both *UBQ10* promoter and ZF108_3xFlag are upstream of the gateway cassette (Invitrogen) present in the original pMDC123 plasmid. The catalytic domain of the TET1 protein (TET1cd) was amplified from the plasmid pJFA334E9, a gift from Keith Joung (Addgene plasmid # 4937) (Maeder et al., 2013), and cloned into the pENTR/D plasmid (Invitrogen) and then delivered into

the modified pMDC123 by an LR reaction (Invitrogen), creating an in-frame fusion of the TET1_cd cDNA with the upstream ZF108_3xFlag cassette (Figure 4-1A). Similarly, YPet was amplified from a YPet containing plasmid and cloned into the pENTR/D plasmid and then delivered to the modified pMDC123 by an LR reaction. Sequences of the modified pUBQ10_ZF108_3xFlag_TET1cd as well as pUBQ10_ZF108_3xFlag_YPet are provided in Dataset S1.

Cloning of pUBQ10_CACTA1ZF_TET1cd.

New ZFs were designed to bind 18bp sequences in the different targeted promoters. Amino acid sequences were obtained in silico using Codelt (http://zinc.genomecenter.ucdavis.edu:8080/Plone/codeit), selecting linker type "normal". The resulting amino acid sequence was plant codon optimized and synthesized by IDT. A modified pMDC123 plasmid (Curtis and Grossniklaus, 2003) was created, containing 1990bp of the promoter region of *Arabidopsis UBQ10* gene upstream of a cassette containing a HpaI restriction site and a 3xFlag tag. Both *UBQ10* promoter and 3xFlag are upstream of the gateway cassette (Invitrogen) present in the original pMDC123 plasmid. The TET1cd was delivered into the modified pMDC123 by an LR reaction (Invitrogen), creating an in-frame fusion of the TET1cd cDNA with the upstream 3xFlag cassette. The different CACTA1-ZF were plant codon optimized and synthesized by IDT and cloned in the HpaI restriction site in the modified pMDC123_3xFlag_TET1cd plasmid by In-Fusion (Takara). Sequences of the modified pMDC123_3xFlag_TET1cd as well as the different ZFs are provided in Dataset S1. In an effort to make these plasmids widely available for the academic community, the above plasmids are available through addgene using the corresponding addgene plasmid identification number: pUBQ10::ZF108_3xFlag_TET1CD(106432); pUBQ10::ZF1CACTA1_3xFlag_TET1cd

(106433);pUBQ10::ZF2CACTA1_3xFlag_TET1cd(106434);pUBQ10::ZF108_3xFlag_YPet (106441).

SunTag design and cloning

Nucleic acid sequences of SunTagFWAg4-22aa-TET1cd and SunTagFWAg4-14aa-TET1cd were either PCR amplified from Addgene plasmid # 60903 and # 60904, gifts from Ron Vale(Tanenbaum et al., 2014), or synthesized using GenScript services. The SunTag constructs were adapted from Tanenbaum et al., 2014 in order to create a dCas9-based demethylation system in plants. dCas9+epitope tail (GCN4x10), single chain variable fragment (scFv) antibody, and the sgRNA were cloned into a binary pMOA backbone vector (Barrell et al., 2002) using In-Fusion (Takara). Expression of dCas9+epitope tail and scFv was controlled by the *UBQ10* promoter, and the sgRNA was expressed using the U6 promoter.

The epitope tails fused to dCas9 consisted of 10 copies of the GCN4 peptide and either a 14 amino acid linker or a 22 amino acid linker separated each epitope. An extra SV40 type NLS was added to the dCas9+epitope sequence. Due to a lack of an effective NLS on the scFv-TET1cd fusion, two SV40 type nuclear localization signals were added for nuclear import of the antibody. These were preceded by 1xHA. Sequences of SunTagFWAg4-22aa-TET1cd and SunTagFWAg4-14aa-TET1cd are provided in Dataset S1. In an effort to make these plasmids widely available for the academic community, the above plasmids are available through addgene using the corresponding addgene plasmid identification number: SunTagFWAg4-22aa-TET1cd (106435); SunTagFWAg4-14aa-TET1cd (106436); SunTagCACTA1g2-22aa-TET1cd (106437); SunTagng22aa (106438); SunTagng14aa(106439);

Quantitative Real-time PCR（qRT-PCR)

RNA was extracted using Direct-zol RNA Miniprep kit (Zymo). For qRT-PCR involving ZF1CACTA1-TET1cd T2 plants 600ng of total RNA was used to prepare cDNA libraries using the SuperScript III First-Strand Synthesis SuperMix (Invitrogen). For qRT-PCR involving SunTagCACTA1g2-22aa plants 250ng of total RNA was used to prepare cDNA libraries using the SuperScript III. First-Strand Synthesis SuperMix (Invitrogen). qRT-PCR of the *CACTA1* transcripts was done using the oligos (*5′-* agtgtttcaatcaaggcgtttc *-3)* and (*5′-* cacccaatggaacaaagtgaac *-3)′*. Values were normalized to the expression of the house keeping gene *IPP2* using oligos (*5′-* gtatgagttgcttctccagcaaag *-3)* and (*5′-* gaggatggctgcaacaagtgt *-3).*

McrBC-qRT-PCR

CTAB-extracted DNA (1µg) was digested using the McrBC restriction enzyme for 4h at 37°C. As a non-digested control, 1 µg of DNA was incubated in digestion buffer without McrBC enzyme for 4h at 37°C. Quantitative Real-time PCR of the *FWA* promoter was done using the oligos (*5′-ttgggtttagtgtttacttg-3)* and (*5′-gaatgttgaatgggataaggta-3)′*. A control region methylated in Col-0 and unmethylated in *fwa-4* was amplified using the oligos (5′-*tgcaatttgtctgcttgctaatg*-3′) and (5′-*tcatttataatggacgatgcc*-3′). The ratio between the digested and non-digested samples was calculated.

RNAseq analysis

RNA was extracted using Direct-zol RNA Miniprep kit (Zymo). For RNAseq involving ZF108-TET1cd and ZF108-YPet plants, 75ng of total RNA was used to prepare libraries using the Neoprep stranded mRNA-seq kit (Illumina). For RNAseq involving, ZF1CACTA1-TET1cd, ZF2CACTA1-TET1cd, SunTagFWAg4-14aa and SunTagFWAg4-22aa plants, 1µg of total RNA was used to prepare libraries using the TruSeq Stranded mRNA-seq kit (Illumina). Reads were

first aligned to TAIR10 gene annotation using Tophat (Trapnell et al., 2009) by allowing up to two mismatches and only keeping reads that mapped to one location. When reads did not map to the annotated genes, the reads were mapped to the TAIR10 genome. Number of reads mapping to genes were calculated by HTseq (Anders et al., 2015) with default parameters. Expression levels were determined by RPKM (reads per kilobase of exons per million aligned reads) using customized R scripts.

Whole genome bisulfite sequencing analysis

DNA was extracted using a CTAB-based method and 100ng were used to make libraries using the Nugen Ultralow Methyl-seq kit (Ovation). Raw sequencing reads were aligned to the TAIR10 genome using BSMAP (Xi and Li, 2009) by allowing up to 2 mismatches and only retaining reads mapped to one location. Methylation ratio are calculated by #C/(#C+#T) for all CG, CHG and CHH sites. Reads with 3 consecutive methylated CHH sites were discarded since they are likely to be unconverted reads as described before(Cokus et al., 2008).

Metaplot of WGBS data

Metaplots of WGBS data were made using custom Perl and R scripts. Regions of interest were broken into 50 bins while flanking 1kb regions were each broken into 25bins. CG, CHG and CHH methylation levels in each bin were then determined. Metaplots were then generated with R.

Accession Numbers.

All sequencing data are available at Gene Expression Omnibus (GEO) with accession no. GSE109115.

FIGURE LEGENDS

Figure 3-1. ZF108-TET1cd expression causes heritable late flowering and *FWA* upregulation.

(A) Schematic representation of the ZF108-YPet (top) and the ZF108-TET1cd fusions (bottom). **(B)** Flowering time of Col-0, *fwa-4,* and ZF108-TET1cd T1 plants. **(C)** Col-0 plants and a representative ZF108-TET1cd T3 line grown side by side to illustrate the differences in flowering time. **(D)** Flowering time of Col-0, *fwa-4*, three independent lines containing ZF108-YPet and three independent lines containing ZF108-TET1cd. For each independent ZF108-TET1cd line, two different T3 populations were scored, one containing the ZF108-TET1cd transgene (+) and one that had the transgene segregated away in the T2 generation (-). Individual plants are depicted as colored dots. Leaf number corresponds to the total number of rosette and caulinar leaves after flowering. All plants above the dashed line are considered late flowering. **(E)** Bar graph showing *FWA* expression of one plant of Col-0, *fwa-4*, and four representative late flowering T1 plants expressing ZF108-TET1cd. **(F)** Bar graph showing *FWA* expression of four biological replicates of Col-0 plants and two representative T3 lines expressing ZF108-TET1cd and ZF108-YPet. **(G)** Scatterplot comparing gene expression of ZF108-TET1cd lines and ZF108-YPet lines. Values were calculated using four biological replicates of two independent lines for ZF108-TET1cd and ZF108-YPet. Gray dots indicate non-differentially expressed genes. Blue dots indicate differentially expressed genes. A 4-fold change and FDR less than 0.05 was used as a cutoff. *FWA* expression is highlighted in red.

Figure 3-2. McrBC-qRT-PCR indicates loss of methylation at *FWA* promoter in ZF108-TET1cd T1 plants.

McrBC-qRT-PCR analysis of methylation in 8 independent ZF108-TET1cd T1 plants, Col-0 and *fwa-4* controls. Oligos were designed to amplify the *FWA* promoter or a control region used to differentiate ZF108-TET1cd T1 plants in the Col-0 background from *fwa-4*.

Figure 3-3. Targeted demethylation at the *FWA* promoter is specific and heritable.

(A) Screenshot of CG, CHG, and CHH methylation levels over the *FWA* promoter in Col-0 and a representative ZF108-TET1cd T1 line (Top). Screenshot of CG, CHG, and CHH methylation levels over the *FWA* promoter in Col-0 and a representative ZF108-TET1cd T3 line for which WGBS was done in a plant containing the ZF108-TET1cd construct (ZF108-TET1cd-1 (+)) and in a plant that had segregated away the transgene already in the T2 generation (ZF108-TET1cd-1 (-)) (Bottom). Gray vertical lines indicate the ZF108 binding sites in the *FWA* promoter. 5′ proximal representation of the *FWA* transcribed region is depicted in blue with filled squares indicating the untranslated regions (UTRs) and diamond lines indicating introns. **(B)** Genome-wide distribution of CG methylation in two Col-0 plants and four representative T1 ZF108-TET1cd plants (Top) as well as one Col-0 plant and one T3 plant containing the ZF108-TET1cd-1 (+) and a T3 plant that had segregated away the transgene already in the T2 generation (ZF108-TET1cd-1 (-)) (Bottom). **(C)** Metaplot of CG, CHG, and CHH methylation levels over protein coding genes and TEs in Col-0, ZF108-TET1cd-1 (+) and ZF108-TET1cd-1 (-) T3 plants. Percent methylation is depicted on the Y-axis of all graphs.

Figure 3-4. ZF108-TET1cd specifically demethylates the *FWA* promoter.

(A) Screenshot for CG, CHG and CHH methylation levels over the *FWA* promoter in a second Col-0 control, three additional ZF108-TET1cd T1 plants and an additional ZF108-TET1cd-2 T3 line for which WGBS was done in a plant containing the ZF108-TET1cd construct (ZF108-TET1cd-2 (+)) and in a plant that had the transgene segregated away in the T2 generation (ZF108-TET1cd-2 (-)). The gray vertical lines indicate the ZF108 binding sites in the *FWA* promoter. 5′ proximal representation of the *FWA* transcribed region is depicted in blue with filled squares indicating the UTRs and diamond lines indicating introns. **(B)** A zoomed out screenshot for CG, CHG, and CHH methylation levels over *FWA* and the surrounding regions in Col-0 controls, ZF108-TET1cd T1 plants and ZF108-TET1cd T3 plants (with (+) or without (-) the transgene). The gray vertical line indicates the ZF108 binding sites in the *FWA* promoter. Percent methylation is depicted on the Y-axis.

Figure 3-5. ZF108-TET1cd specifically demethylates the *FWA* promoter.

(A) Genome-wide distribution of CHG and CHH methylation in two Col-0 and four representative ZF108-TET1cd T1 plants. **(B)** Genome-wide distribution of CHG and CHH methylation in Col-0, and a representative ZF108-TET1cd-1 T3 line for which WGBS was done in a plant containing the ZF108-TET1cd transgene (ZF108-TET1cd-1 (+)) and in a plant that had the transgene segregated away in the T2 generation (ZF108-TET1cd-1 (-)). **(C)** Genome-wide distribution of CG, CHG, and CHH methylation in Col-0, ZF108-TET1cd-2 (+) T3 plants and ZF108-TET1cd-2 (–) T3 plants. Percent methylation is depicted on the Y-axis of all graphs.

Figure 3-6. **Targeted demethylation of *CACTA1* using ZF-TET1cd fusions.**

(A) Screenshot showing CG, CHG, and CHH methylation over the *CACTA1* region in Col-0, and one T1 plant each for ZF1CACTA1-TET1cd and ZF2CACTA1-TET1cd. **(B)** Bar graphs depicting the methylation levels in the region comprising 200bp upstream and downstream of the ZF1CACTA1 or ZF2CACTA1 binding sites for Col-0 and one representative T1 plant of ZF1CACTA1-TET1cd and ZF2CACTA1-TET1cd plants. **(C)** Bar graph showing *CACTA1* expression in Col-0 and one representative T1 plant of ZF1CACTA1-TET1cd T1 plant and ZF2CACTA1-TET1cd T1. RPKM alues are indicated. **(D)** Screenshot showing CG, CHG, and CHH methylation over the *CACTA1* region in Col-0, and one T2 plant containing the transgene (+) and one that had segregated it away (-) for ZF1CACTA1-TET1cd and ZF2CACTA1-TET1cd lines. In **(A)** and **(D)** a red arrow indicates the ZF1 binding site and a purple arrow indicates the ZF2 binding site in the promoter region of *CACTA1*. A zoom in of the targeted region is shown (right). Percent methylation is shown for WGBS. **(E)** Bar graphs depicting the methylation levels in the region comprising 200bp upstream and downstream of the ZF1CACTA1 binding sites for Col-0 and one T2 plant containing the ZF1CACTA1-TET1cd transgene (+) or that had segregated it away (-) for ZF1CACTA1-TET1cd. **(F)** Bar graph showing relative expression by qRT-PCR of *CACTA1* over *IPP2* in Col-0, one T2 plant containing the transgene (+) or one that had it segregated away (-) for ZF1CACTA1-TET1cd.

Figure 3-7. ZF-TET1cd fusions targeting *CACTA1* show variable levels of non-specific loss of methylation.

(A) Genome-wide distribution of CG, CHG, and CHH methylation in Col-0 and one representative ZF1CACTA1-TET1cd and ZF2CACTA1-TET1cd T1 plant. **(B)** Metaplot showing CG, CHG, and CHH methylation levels over all protein coding genes and TEs in Col-0

and one representative ZF1CACTA1-TET1cd and ZF2CACTA1-TET1cd T1 plant. **(C)** Genome-wide distribution of CG, CHG, and CHH methylation in Col-0 and one T2 plant containing the transgene (+) and one that had segregated it away (-) for one representative ZF1CACTA1-TET1cd and ZF2CACTA1-TET1cd line. **(D)** Metaplot showing CG, CHG, and CHH methylation levels over all protein coding genes and TEs in Col-0 and one T2 plant containing the transgene (+) or one that had segregated it away (-) for one representative ZF1CACTA1-TET1cd and ZF2CACTA1-TET1cd line. Percent methylation is depicted on the Y-axis of all graphs.

Figure 3-8. Targeted demethylation at the *FWA* promoter using SunTag-TET1cd.

(A) Schematic representation of the SunTagFWAg4-22aa (left) and SunTagFWAg4-14aa (right) systems. **(B)** Bar graph of *FWA* expression in Col-0, *fwa-4*, and two T1 lines each for SunTagFWAg4-22aa and SunTagFWAg4-14aa. **(C)** Flowering time of Col-0, *fwa-4,* and one representative SunTagFWAg4-22aa and SunTagFWAg4-14aa T2 line. **(D)** Screenshot of CG, CHG and CHH methylation levels over the *FWA* promoter in Col-0, one representative SunTagFWAg4-22aa and SunTagFWAg4-14aa T1 line, and one representative T2 plant of the same lines containing the transgene (+) or that had segregated it away (-). A gray arrow indicates the FWAgRNA-4 binding site (FWAg4) in the *FWA* promoter. 5′ proximal representation of the *FWA* transcribed region is depicted in blue with filled squares indicating the UTRs and diamond lines indicating introns. **(E)** Genome-wide CG methylation levels in two Col-0 plants, two T1 lines each for SunTagFWAg4-22aa and SunTagFWAg4-14aa (upper), as well as one Col-0, one T2 plant containing the transgene (+) or one that had segregated it away (-) for representative

SunTagFWAg4-22aa and SunTagFWAg4-14aa lines. Percent methylation is depicted on the Y-axis.

Figure 3-9. SunTag-TET1cd lines specifically demethylate the *FWA* promoter.

(A) Screenshot of CG, CHG and CHH methylation levels over the *FWA* promoter in Col-0, a second representative SunTagFWAg4-22aa and SunTagFWAg4-14aa T1 line (results for the other representative line is shown in Fig. 4), and one T2 plant containing the transgene (+) and one that had segregated it away (-) for a second representative SunTagFWAg4-22aa line (results for the other representative line is shown in Fig. 4). **(B)** Zoomed out screenshot of CG, CHG, and CHH methylation levels over the *FWA* promoter and the surrounding regions in Col-0, two representative T1 plants each for SunTagFWAg4-22aa and SunTagFWAg4-14aa, as well as Col-0 and one T2 plant containing the transgene (+) or one that had segregated it away (-) for two representative SunTagFWAg4-22aa lines and one representative SunTagFWAg4-14aa line. A gray arrow indicates the FWAgRNA-4 binding site (FWAg4) in the *FWA* promoter. 5′ proximal representation of the *FWA* transcribed region is depicted in blue with filled squares indicating the UTRs and diamond lines indicating introns.

Figure 3-10. FWA-targeted SunTag-TET1cd lines do not affect global DNA methylation levels.

(A) Genome-wide distribution of CHG and CHH methylation in two independent Col-0, and two representative T1 lines each for SunTagFWAg4-22aa and SunTagFWAg4-14aa. **(B)** Genome-wide distribution of CHG and CHH methylation in Col-0, one T2 plant containing the transgene (+) and one that had segregated it away (-) for the SunTagFWAg4-22aa and SunTagFWAg4-

14aa lines shown in Fig. 4D. **(C)** Genome-wide distribution of CG, CHG and CHH methylation in Col-0, one T2 plant containing the transgene (+) or one that had segregated it away (-) for a second representative line expressing SunTagFWAg4-22aa (SunTagFWAg4-22aa -1).

Figure 3-11. Targeted demethylation of *CACTA1* using SunTag-TET1cd.

(A) Screenshot of CG, CHG and CHH methylation levels over the *CACTA1* region in Col-0 and two representative SunTagCACTA1g2-22aa T1 lines. A gray arrow indicates the gRNA binding site in the promoter region of *CACTA1*. A zoom in of the targeted region is shown (right). **(B)** Bar graphs depicting the methylation levels in the region comprising 200bp upstream and downstream of the gRNA binding sites for Col-0 and two representative SunTagCACTA1g2-22aa T1 plants. **(C)** Bar graph showing relative expression by qRT-PCR of *CACTA1* over *IPP2* in Col-0 and two representative SunTagCACTA1g2-22aa T1 plants. **(D)** Genome-wide CG methylation levels in Col-0 and two representative SunTagCACTA1g2-22aa T1 plants. Percent methylation is depicted on the Y-axis.

Figure 3-12. CACTA1-targeted SunTag-TET1cd lines do not affect global DNA methylation levels.

Genome-wide distribution of CHG and CHH methylation in Col-0 and two representative SunTagCACTA1g2-22aa T1 lines.

Figure 3-13. SunTag-TET1cd lines with no gRNA do not affect global DNA methylation levels.

(A) Flowering time of Col-0, *fwa-4* controls and SunTagng-22aa and SunTagng-14aa T1 plants. **(B)** Screenshot of CG, CHG and CHH methylation levels over the *FWA* promoter in Col-0 and two representative T1 lines each for SunTagng-22aa and SunTagng-14aa. **(C)** Screenshot of CG, CHG and CHH methylation levels over the *CACTA1* region in Col-0 and two representative T1 lines each for SunTagng-22aa and SunTagng-14aa. **(D)** Genome-wide distribution of CG, CHG and CHH methylation in Col-0 and two representative T1 lines each for SunTagng-22aa and SunTagng-14aa.

Figure 3-1

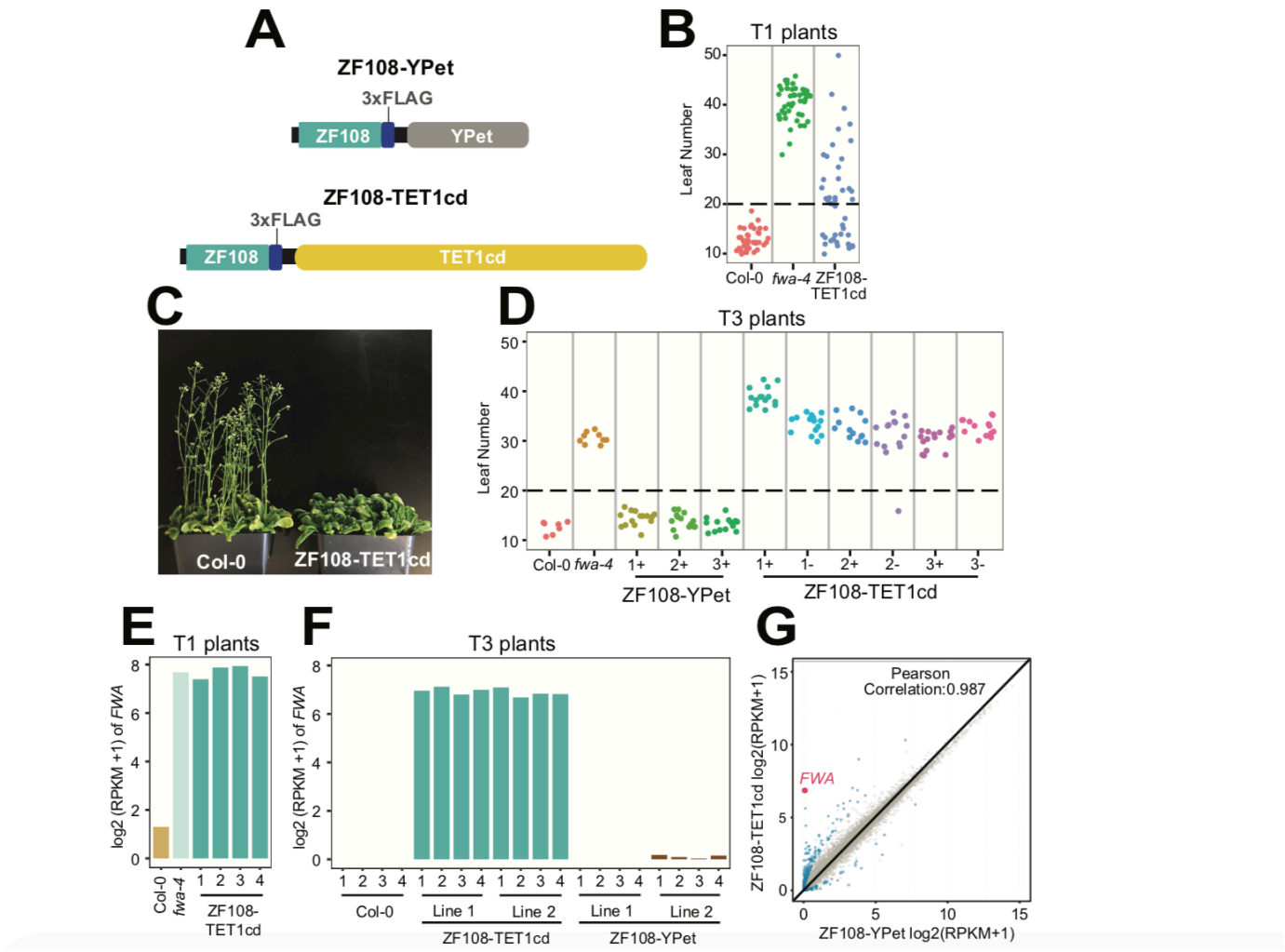

A
ZF108-YPet
3xFLAG
ZF108
YPet
ZF108-TET1cd
3xFLAG
ZF108
TET1cd
B
T1 plants
Leaf Number
Col-0
fwa-4
ZF108-TET1cd
C
Col-0
ZF108-TET1cd
D
T3 plants
Leaf Number
Col-0
fwa-4
ZF108-YPet
ZF108-TET1cd
E
T1 plants
log2 (RPKM +1) of FWA
Col-0
fwa-4
ZF108-TET1cd
F
T3 plants
log2 (RPKM +1) of FWA
Col-0
Line 1
Line 2
ZF108-TET1cd
ZF108-YPet
G
Pearson Correlation:0.987
FWA
ZF108-TET1cd log2(RPKM+1)
ZF108-YPet log2(RPKM+1)

Figure 3-2

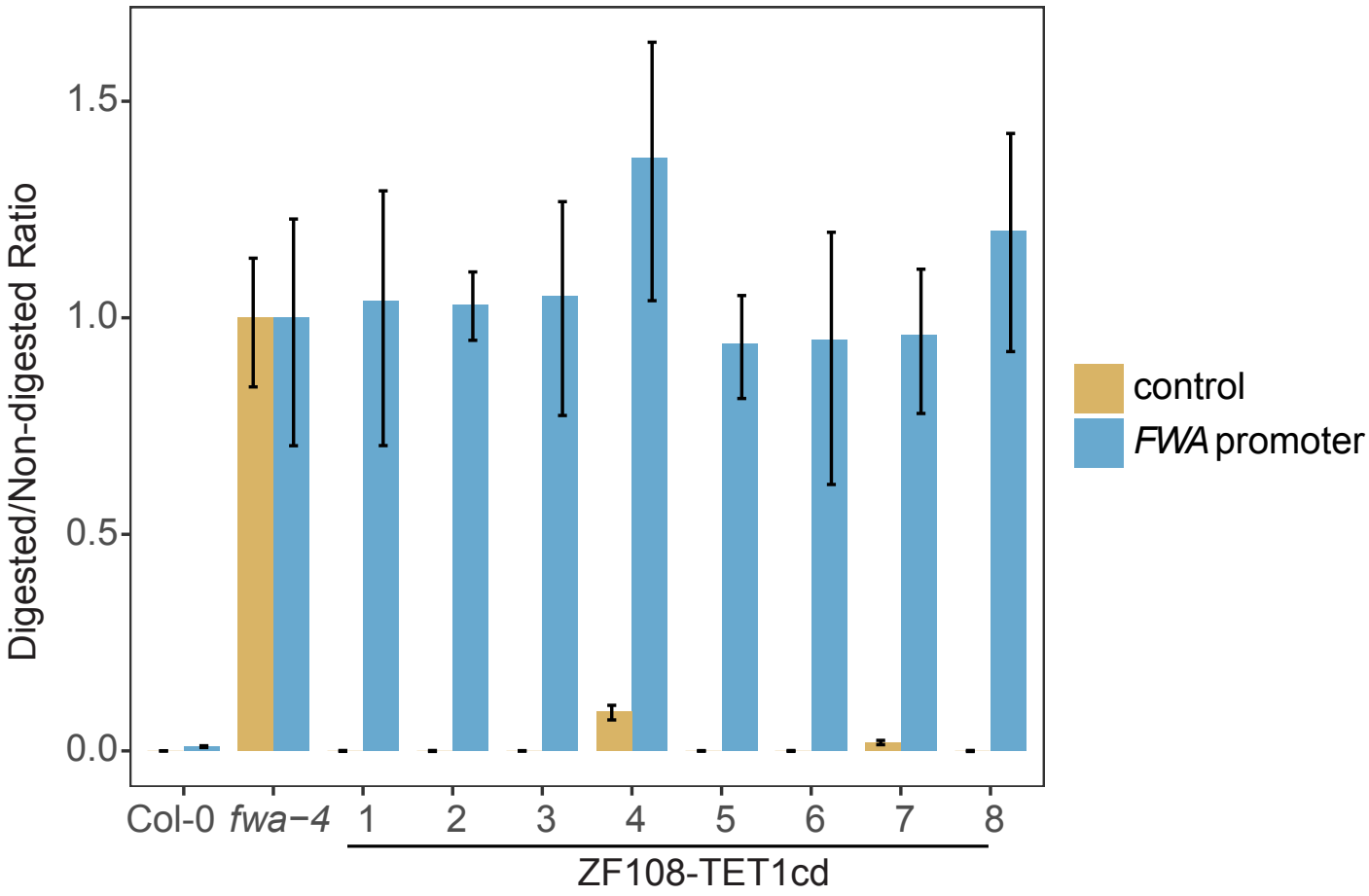

Digested/Non-digested Ratio
1.5
1.0
0.5
0.0
Col-0
fwa−4
1
2
3
4
5
6
7
8
ZF108-TET1cd
control
FWA promoter

Figure 3-3

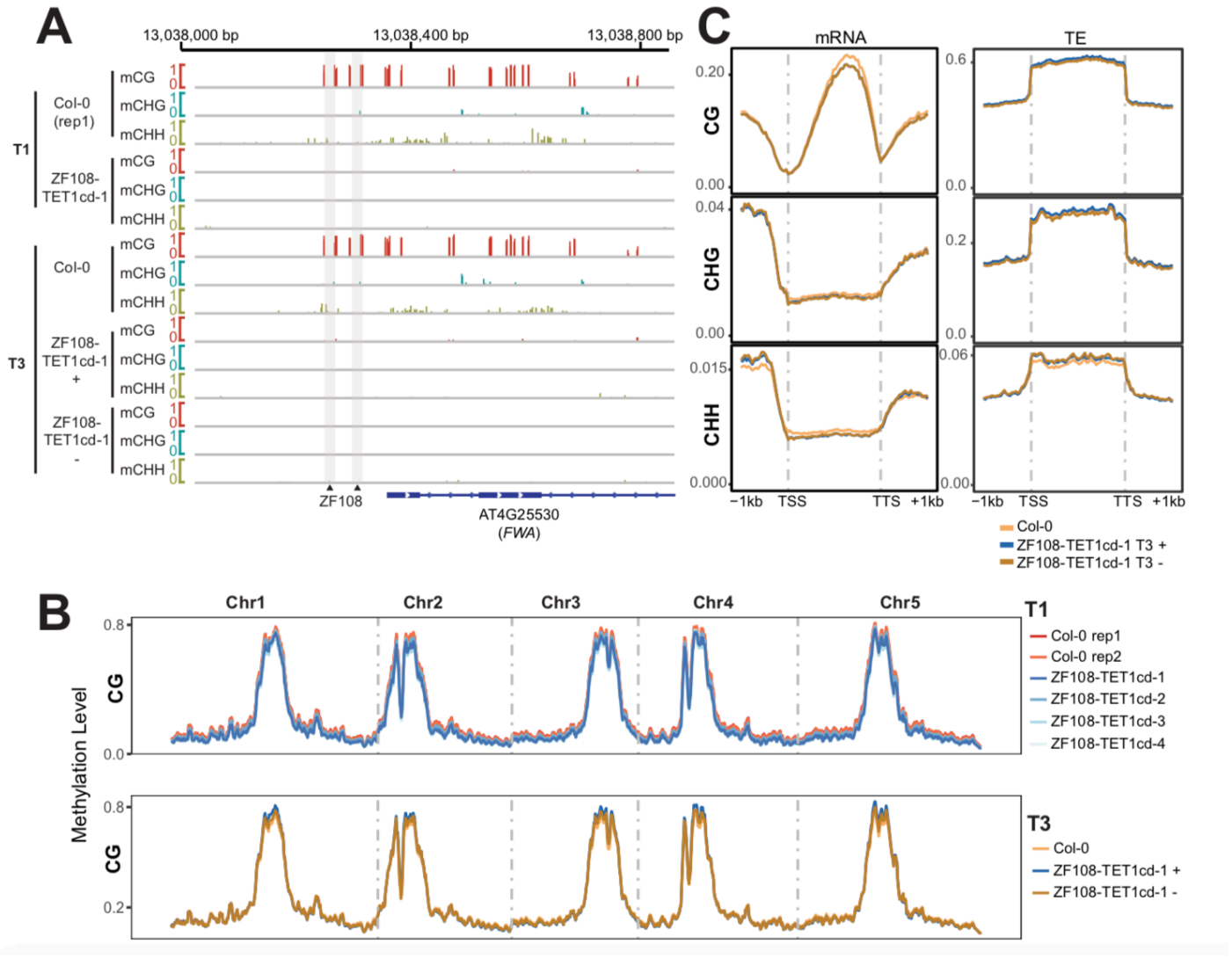

A
13,038,000 bp
13,038,400 bp
13,038,800 bp
T1
Col-0 (rep1)
ZF108-TET1cd-1
T3
Col-0
ZF108-TET1cd-1 +
ZF108-TET1cd-1 -
mCG
mCHG
mCHH
ZF108
AT4G25530
(FWA)
C
mRNA
TE
CG
CHG
CHH
−1kb
TSS
TTS
+1kb
Col-0
ZF108-TET1cd-1 T3 +
ZF108-TET1cd-1 T3 -
B
Chr1
Chr2
Chr3
Chr4
Chr5
Methylation Level
T1
Col-0 rep1
Col-0 rep2
ZF108-TET1cd-1
ZF108-TET1cd-2
ZF108-TET1cd-3
ZF108-TET1cd-4
T3
Col-0
ZF108-TET1cd-1 +
ZF108-TET1cd-1 -

Figure 3-4

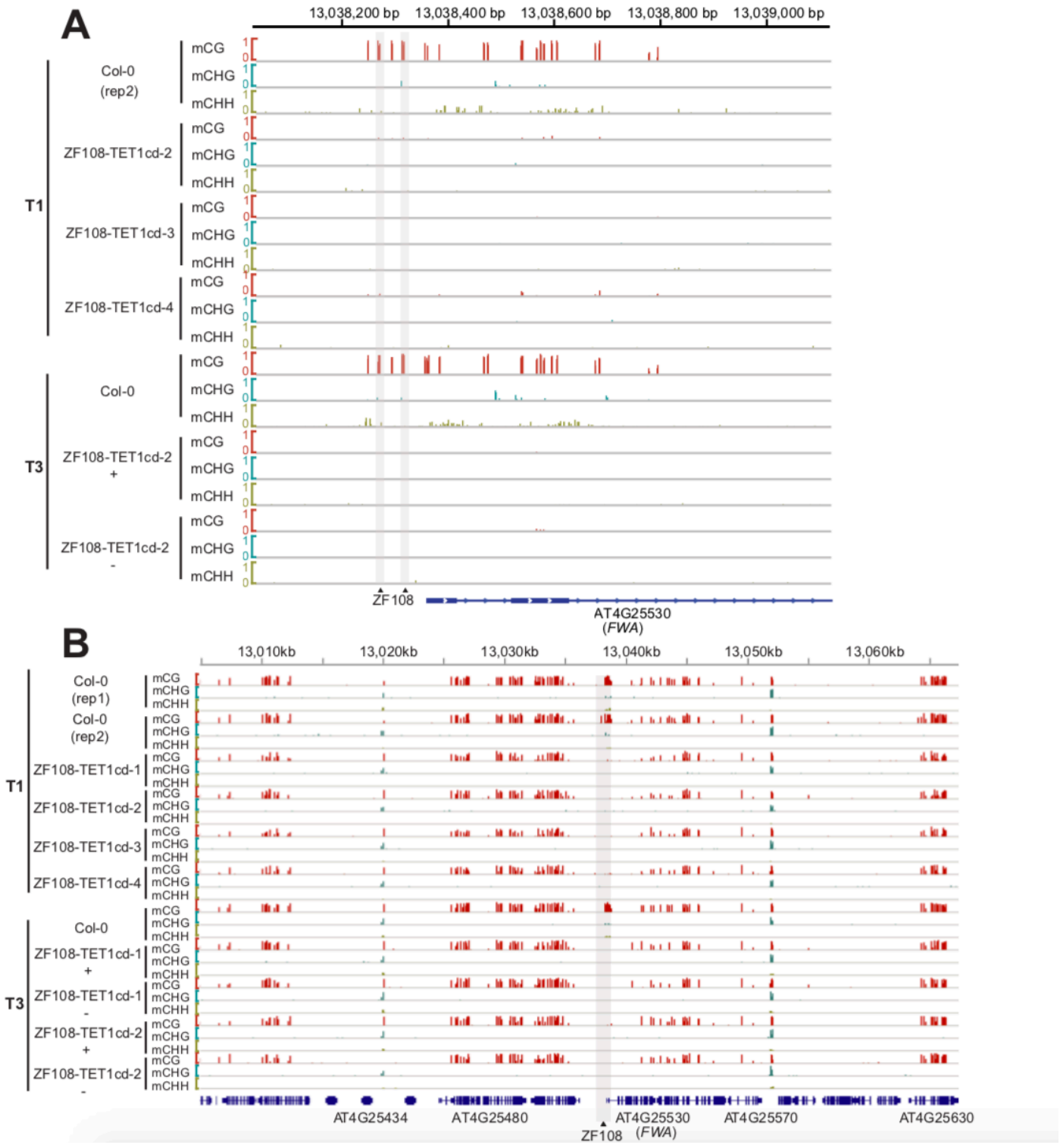

A
13,038,200 bp
13,038,400 bp
13,038,600 bp
13,038,800 bp
13,039,000 bp
T1
Col-0 (rep2)
ZF108-TET1cd-2
ZF108-TET1cd-3
ZF108-TET1cd-4
T3
Col-0
ZF108-TET1cd-2 +
ZF108-TET1cd-2 -
mCG
mCHG
mCHH
ZF108
AT4G25530 (FWA)
B
13,010kb
13,020kb
13,030kb
13,040kb
13,050kb
13,060kb
Col-0 (rep1)
Col-0 (rep2)
ZF108-TET1cd-1
ZF108-TET1cd-2
ZF108-TET1cd-3
ZF108-TET1cd-4
Col-0
ZF108-TET1cd-1 +
ZF108-TET1cd-1 -
ZF108-TET1cd-2 +
ZF108-TET1cd-2 -
AT4G25434
AT4G25480
AT4G25530 (FWA)
AT4G25570
AT4G25630
ZF108

Figure 3-5

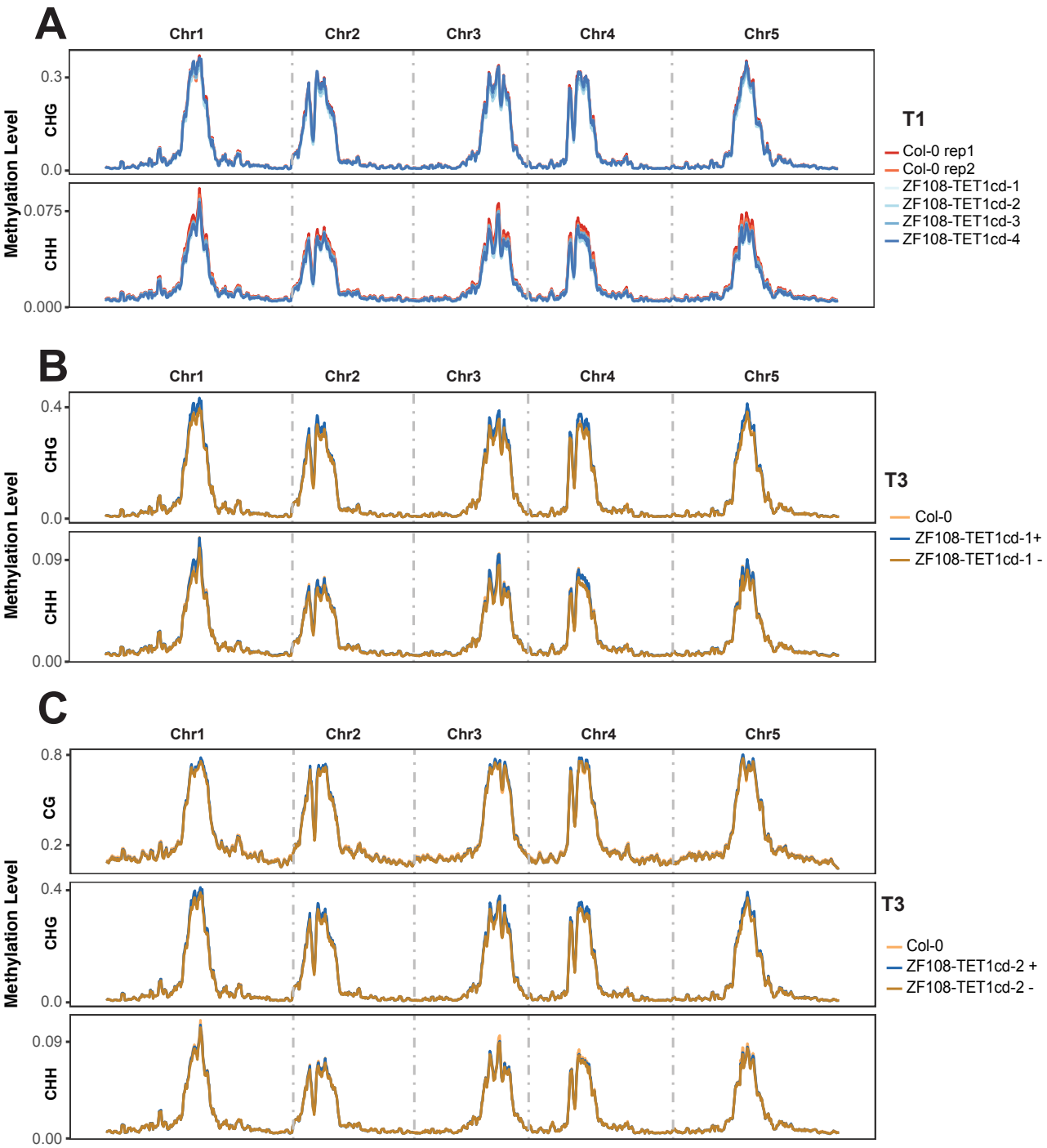

A
Chr1
Chr2
Chr3
Chr4
Chr5
Methylation Level
CHG
CHH
0.3
0.0
0.075
0.000
T1
Col-0 rep1
Col-0 rep2
ZF108-TET1cd-1
ZF108-TET1cd-2
ZF108-TET1cd-3
ZF108-TET1cd-4
B
0.4
0.09
0.00
T3
Col-0
ZF108-TET1cd-1+
ZF108-TET1cd-1 -
C
CG
0.8
0.2
ZF108-TET1cd-2 +
ZF108-TET1cd-2 -

Figure 3-6

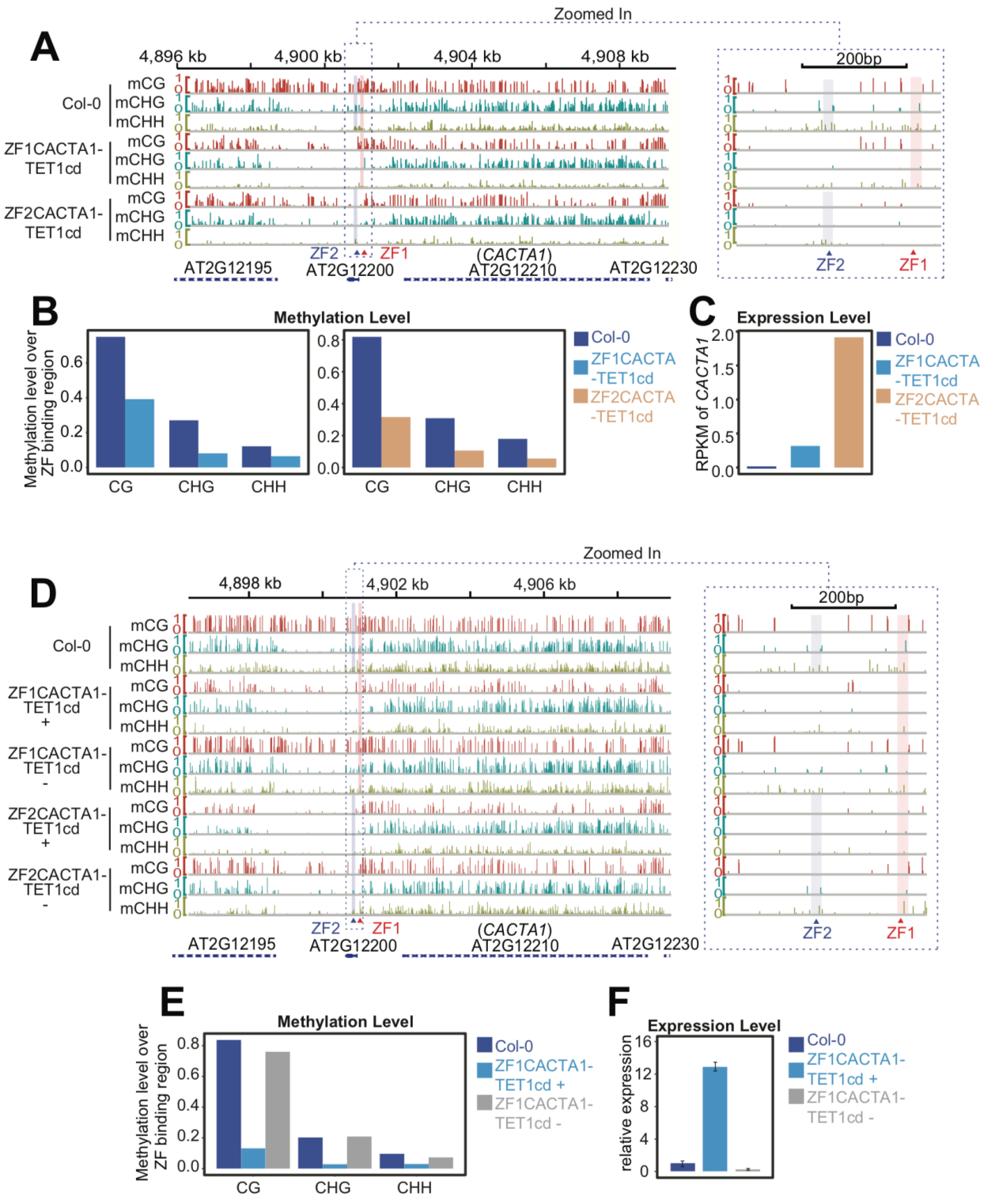

A
Zoomed In
4,896 kb
4,900 kb
4,904 kb
4,908 kb
200bp
Col-0
mCG
mCHG
mCHH
ZF1CACTA1-TET1cd
ZF2CACTA1-TET1cd
ZF2
ZF1
AT2G12195
AT2G12200
(CACTA1)
AT2G12210
AT2G12230
B
Methylation Level
Methylation level over ZF binding region
CG
CHG
CHH
Col-0
ZF1CACTA-TET1cd
ZF2CACTA-TET1cd
C
Expression Level
RPKM of CACTA1
D
Zoomed In
4,898 kb
4,902 kb
4,906 kb
200bp
ZF1CACTA1-TET1cd +
ZF1CACTA1-TET1cd -
ZF2CACTA1-TET1cd +
ZF2CACTA1-TET1cd -
E
Methylation Level
F
Expression Level
relative expression

Figure 3-7

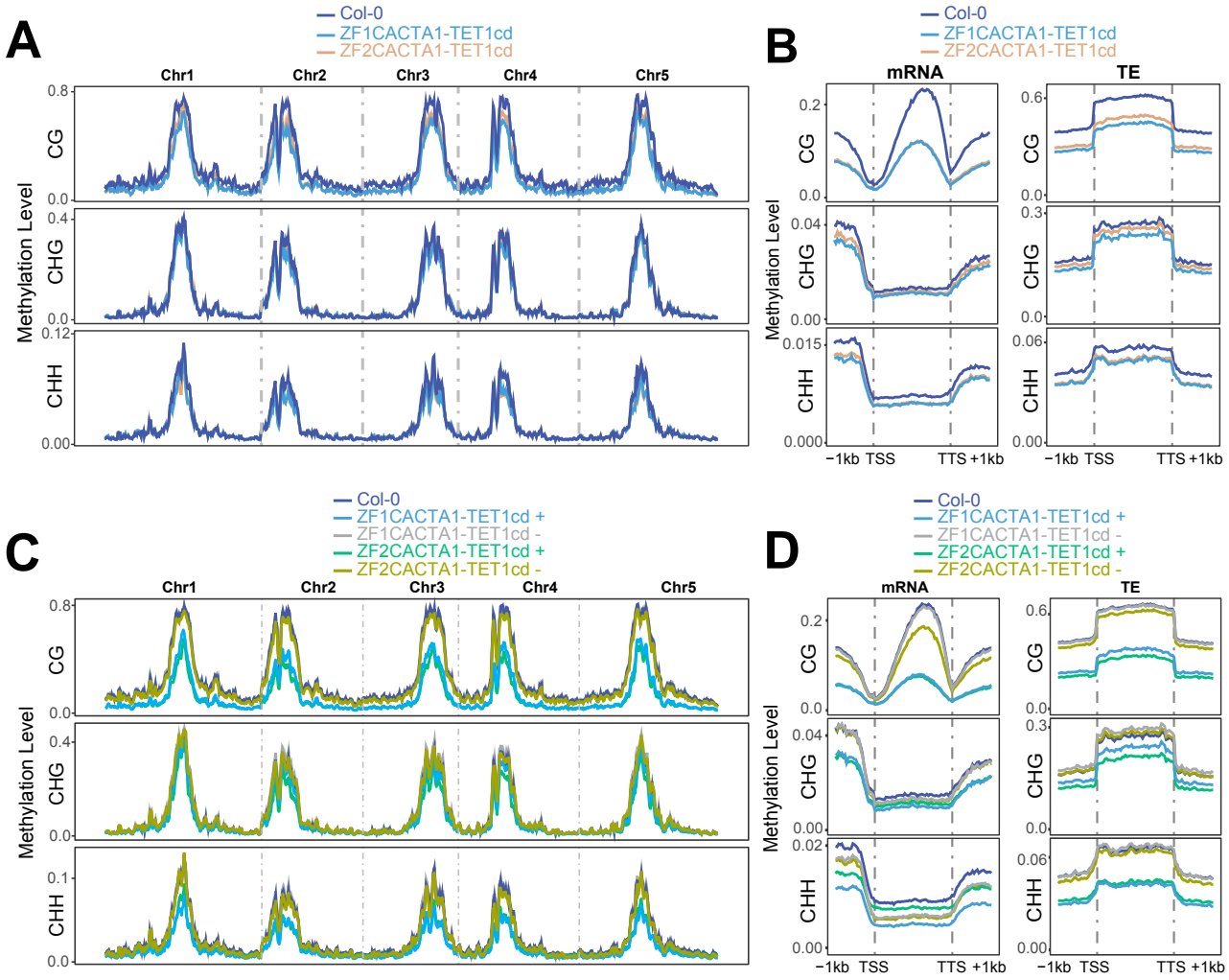

A
Col-0
ZF1CACTA1-TET1cd
ZF2CACTA1-TET1cd
Chr1
Chr2
Chr3
Chr4
Chr5
Methylation Level
CG
CHG
CHH
B
mRNA
TE
−1kb TSS
TTS +1kb
C
Col-0
ZF1CACTA1-TET1cd +
ZF1CACTA1-TET1cd -
ZF2CACTA1-TET1cd +
ZF2CACTA1-TET1cd -
D

Figure 3-8

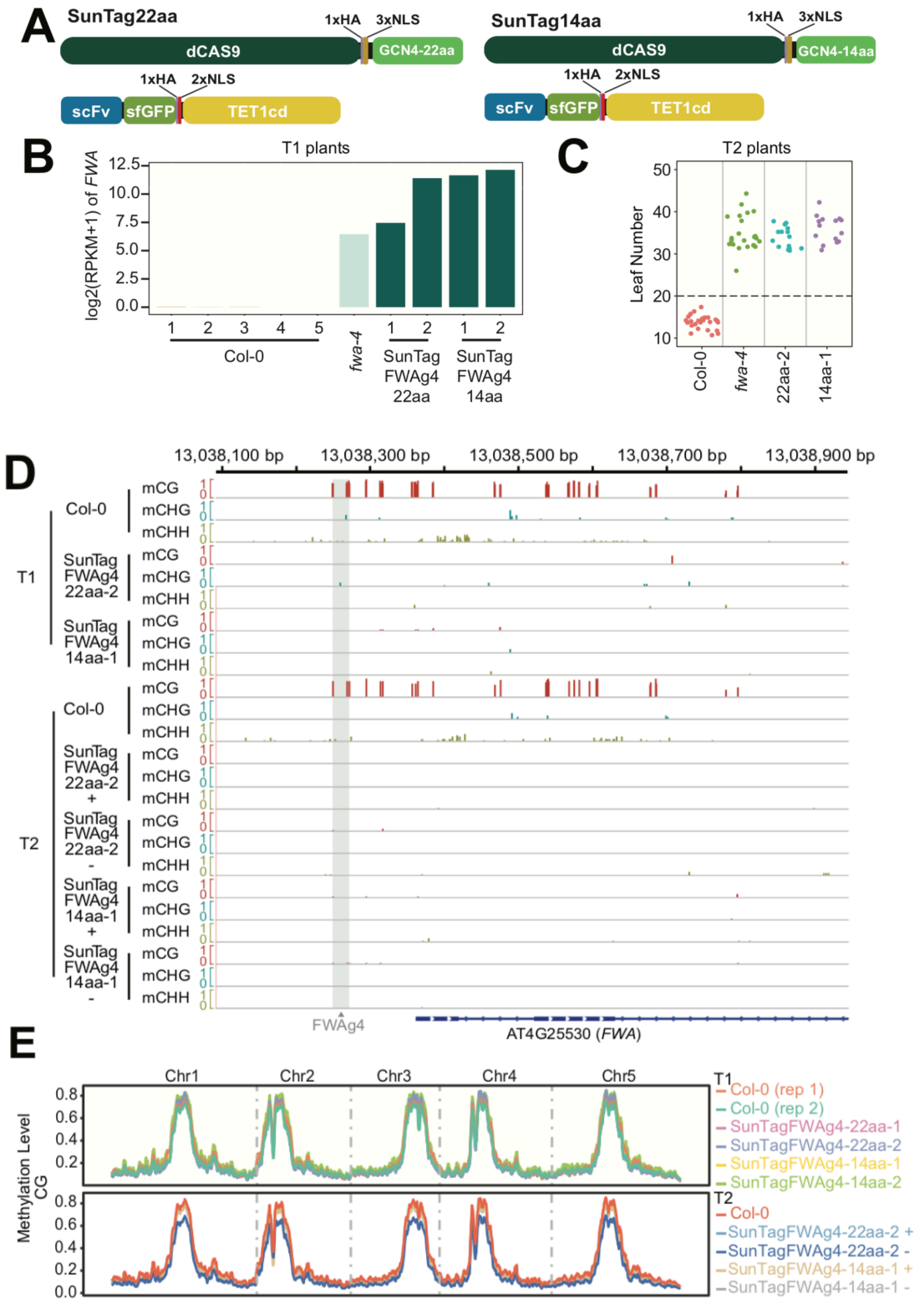
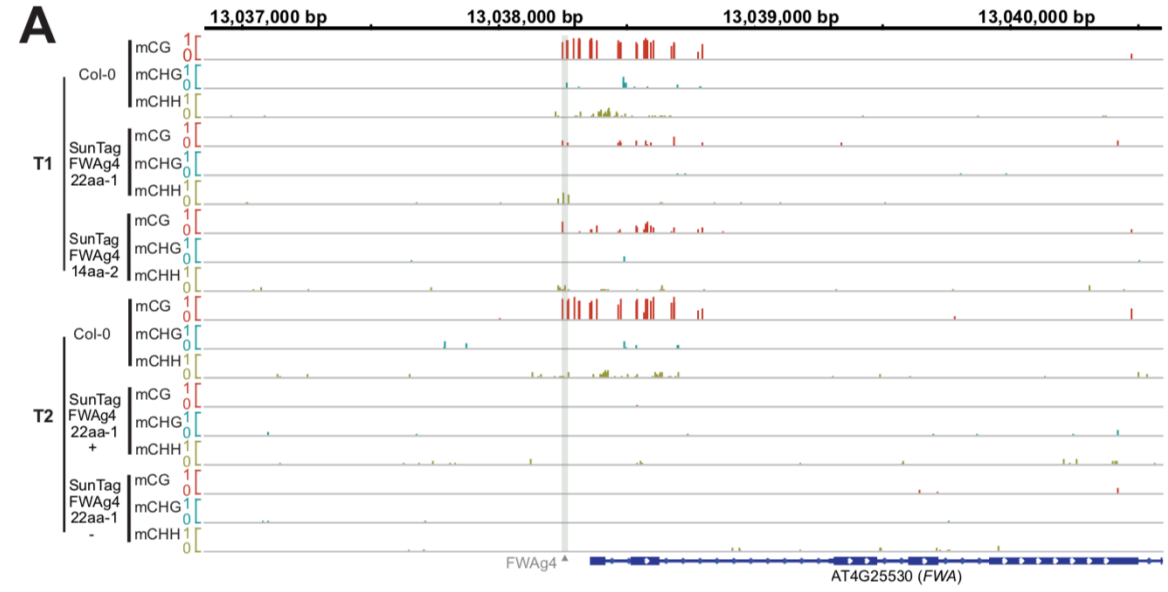

A
SunTag22aa
1xHA
3xNLS
dCAS9
GCN4-22aa
1xHA
2xNLS
scFv
sfGFP
TET1cd
SunTag14aa
1xHA
3xNLS
dCAS9
GCN4-14aa
1xHA
2xNLS
scFv
sfGFP
TET1cd
B
T1 plants
log2(RPKM+1) of FWA
12.5
10.0
7.5
5.0
2.5
0.0
1 2 3 4 5
Col-0
fwa-4
1 2
SunTag FWAg4 22aa
1 2
SunTag FWAg4 14aa
C
T2 plants
Leaf Number
50
40
30
20
10
Col-0
fwa-4
22aa-2
14aa-1
D
13,038,100 bp
13,038,300 bp
13,038,500 bp
13,038,700 bp
13,038,900 bp
T1
Col-0
mCG
mCHG
mCHH
SunTag FWAg4 22aa-2
mCG
mCHG
mCHH
SunTag FWAg4 14aa-1
mCG
mCHG
mCHH
T2
Col-0
mCG
mCHG
mCHH
SunTag FWAg4 22aa-2 +
mCG
mCHG
mCHH
SunTag FWAg4 22aa-2 -
mCG
mCHG
mCHH
SunTag FWAg4 14aa-1 +
mCG
mCHG
mCHH
SunTag FWAg4 14aa-1 -
mCG
mCHG
mCHH
FWAg4
AT4G25530 (FWA)
E
Chr1
Chr2
Chr3
Chr4
Chr5
Methylation Level CG
0.8
0.6
0.4
0.2
0.8
0.6
0.4
0.2
0.0
T1
Col-0 (rep 1)
Col-0 (rep 2)
SunTagFWAg4-22aa-1
SunTagFWAg4-22aa-2
SunTagFWAg4-14aa-1
SunTagFWAg4-14aa-2
T2
Col-0
SunTagFWAg4-22aa-2 +
SunTagFWAg4-22aa-2 -
SunTagFWAg4-14aa-1 +
SunTagFWAg4-14aa-1 -

Figure 3-9


A
13,037,000 bp
13,038,000 bp
13,039,000 bp
13,040,000 bp
T1
T2
Col-0
SunTag FWAg4 22aa-1
SunTag FWAg4 14aa-2
SunTag FWAg4 22aa-1 +
SunTag FWAg4 22aa-1 -
mCG
mCHG
mCHH
FWAg4
AT4G25530 (FWA)

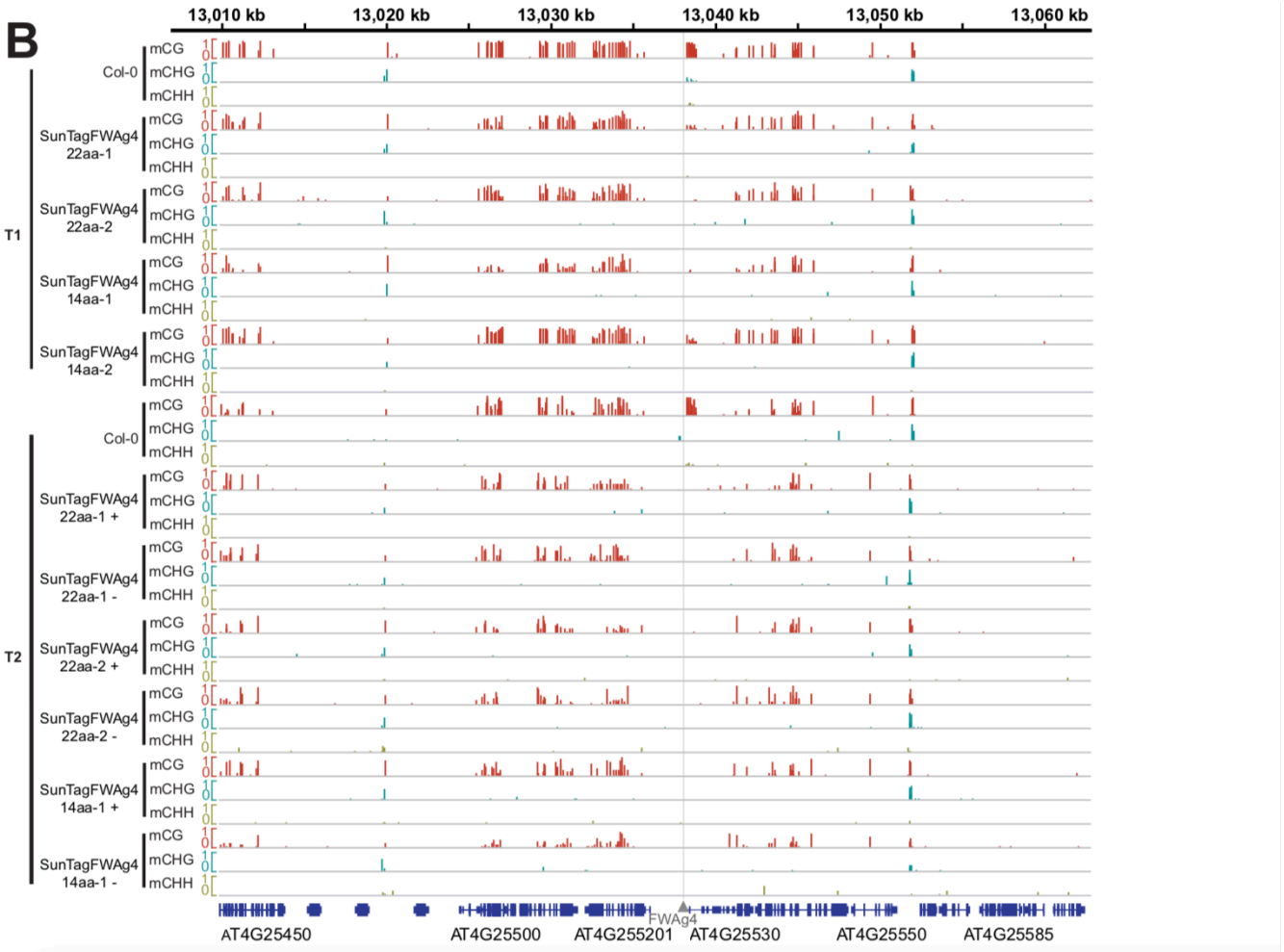

B
13,010 kb
13,020 kb
13,030 kb
13,040 kb
13,050 kb
13,060 kb
T1
T2
Col-0
SunTagFWAg4 22aa-1
SunTagFWAg4 22aa-2
SunTagFWAg4 14aa-1
SunTagFWAg4 14aa-2
SunTagFWAg4 22aa-1 +
SunTagFWAg4 22aa-1 -
SunTagFWAg4 22aa-2 +
SunTagFWAg4 22aa-2 -
SunTagFWAg4 14aa-1 +
SunTagFWAg4 14aa-1 -
mCG
mCHG
mCHH
FWAg4
AT4G25450
AT4G25500
AT4G255201
AT4G25530
AT4G25550
AT4G25585

Figure 3-10

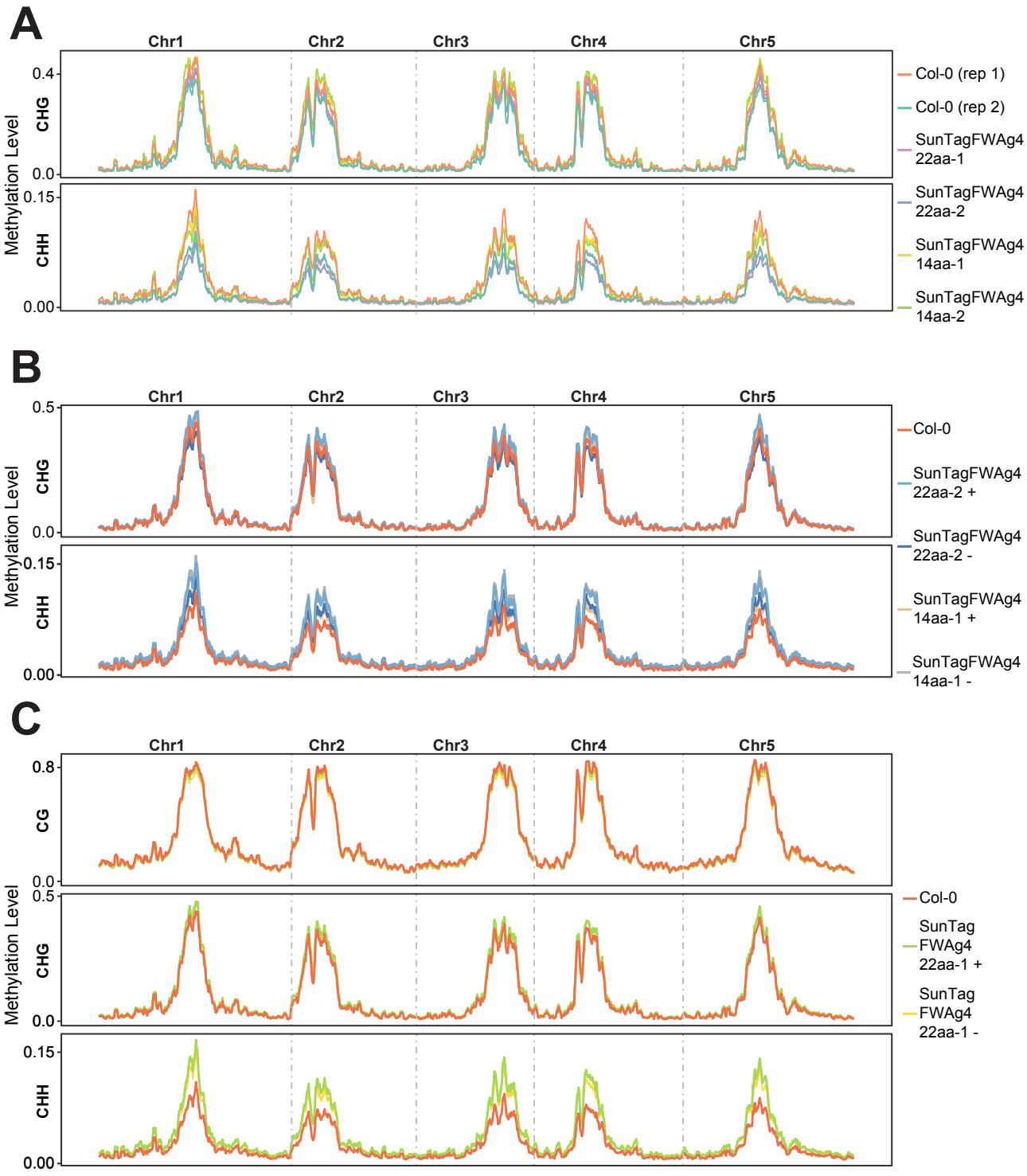

A
Chr1
Chr2
Chr3
Chr4
Chr5
Methylation Level
CHG
CHH
0.4
0.0
0.15
0.00
Col-0 (rep 1)
Col-0 (rep 2)
SunTagFWAg4 22aa-1
SunTagFWAg4 22aa-2
SunTagFWAg4 14aa-1
SunTagFWAg4 14aa-2
B
0.5
SunTagFWAg4 22aa-2 +
SunTagFWAg4 22aa-2 -
SunTagFWAg4 14aa-1 +
SunTagFWAg4 14aa-1 -
C
CG
0.8
SunTag FWAg4 22aa-1 +
SunTag FWAg4 22aa-1 -

Figure 3-11

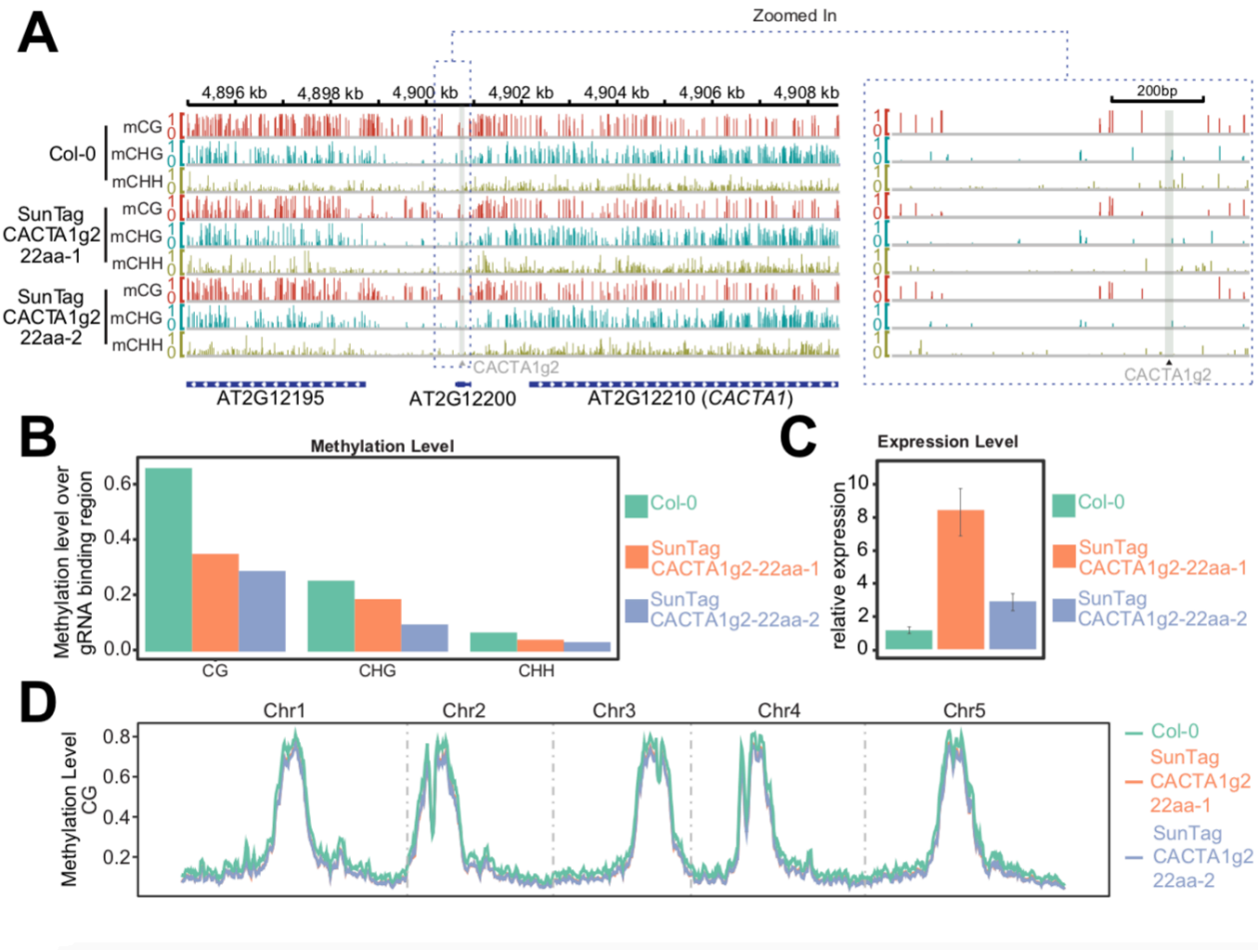

A
Zoomed In
4,896 kb
4,898 kb
4,900 kb
4,902 kb
4,904 kb
4,906 kb
4,908 kb
200bp
Col-0
mCG
mCHG
mCHH
SunTag CACTA1g2 22aa-1
SunTag CACTA1g2 22aa-2
CACTA1g2
AT2G12195
AT2G12200
AT2G12210 (CACTA1)
B
Methylation Level
Methylation level over gRNA binding region
CG
CHG
CHH
Col-0
SunTag CACTA1g2-22aa-1
SunTag CACTA1g2-22aa-2
C
Expression Level
relative expression
D
Chr1
Chr2
Chr3
Chr4
Chr5
Methylation Level CG

Figure 3-12

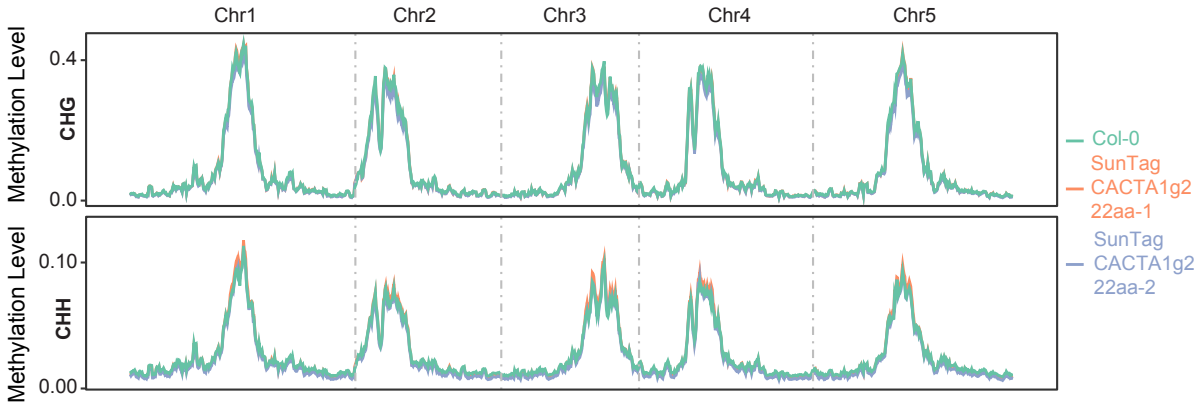
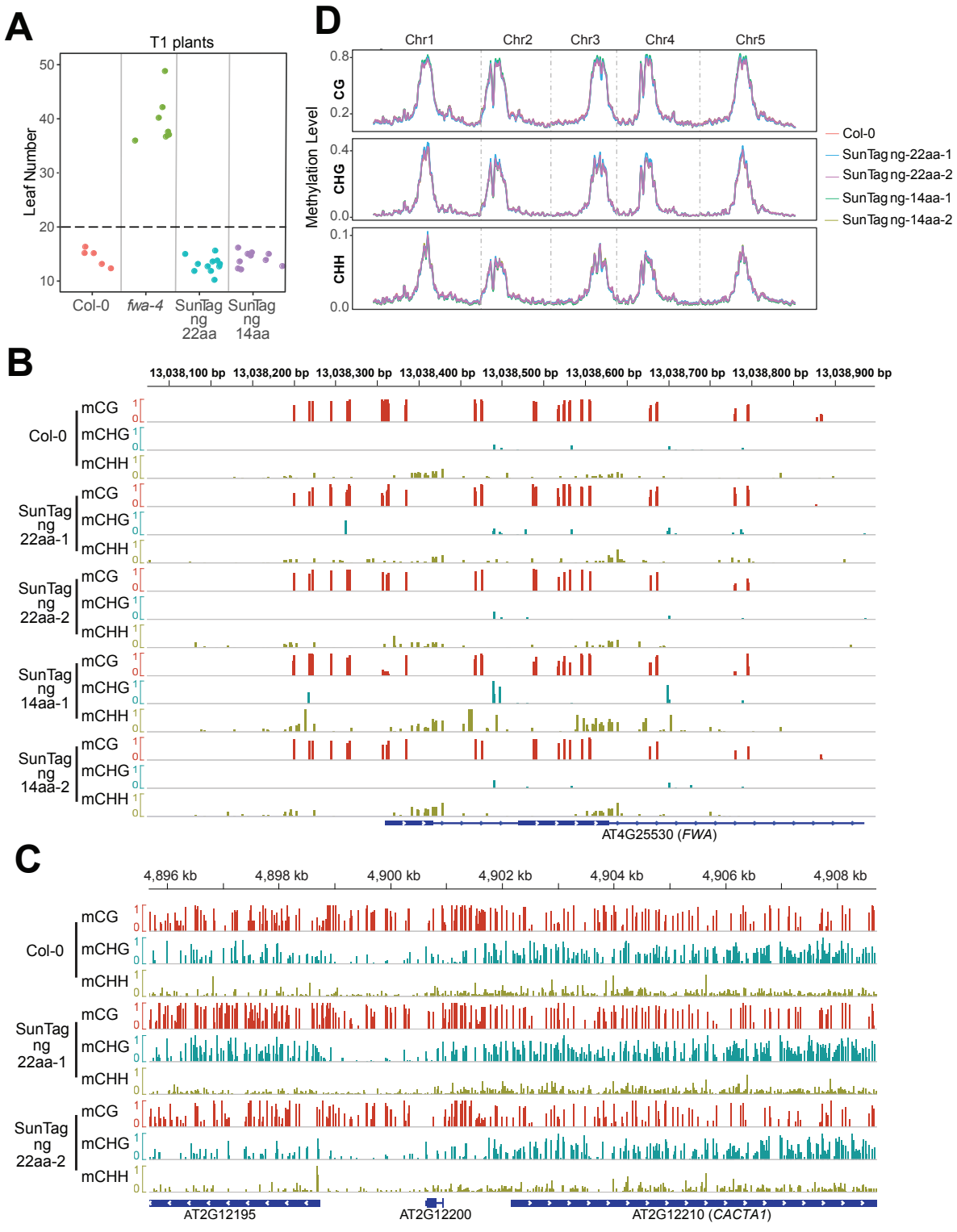

Chr1
Chr2
Chr3
Chr4
Chr5
Methylation Level
CHG
0.4
0.0
Methylation Level
CHH
0.10
0.00
Col-0
SunTag
CACTA1g2
22aa-1
SunTag
CACTA1g2
22aa-2

Figure 3-13


A
T1 plants
Leaf Number
50
40
30
20
10
Col-0
fwa-4
SunTag ng 22aa
SunTag ng 14aa
D
Chr1
Chr2
Chr3
Chr4
Chr5
Methylation Level
CG
CHG
CHH
0.8
0.2
0.4
0.0
0.1
Col-0
SunTag ng-22aa-1
SunTag ng-22aa-2
SunTag ng-14aa-1
SunTag ng-14aa-2
B
13,038,100 bp
13,038,200 bp
13,038,300 bp
13,038,400 bp
13,038,500 bp
13,038,600 bp
13,038,700 bp
13,038,800 bp
13,038,900 bp
Col-0
SunTag ng 22aa-1
SunTag ng 22aa-2
SunTag ng 14aa-1
SunTag ng 14aa-2
mCG
mCHG
mCHH
AT4G25530 (FWA)
C
4,896 kb
4,898 kb
4,900 kb
4,902 kb
4,904 kb
4,906 kb
4,908 kb
AT2G12195
AT2G12200
AT2G12210 (CACTA1)

REFERENCES

Agrawal, A.A., Laforsch, C., and Tollrian, R. (1999). Transgenerational induction of defences in animals and plants. Nature *401*, 60–63.

Amabile, A., Migliara, A., Capasso, P., Biffi, M., Cittaro, D., Naldini, L., and Lombardo, A. (2016). Inheritable Silencing of Endogenous Genes by Hit-and-Run Targeted Epigenetic Editing. Cell *167*, 219–232.e14.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Barrell, P.J., Yongjin, S., Cooper, P.A., and Conner, A.J. (2002). Alternative selectable markers for potato transformation using minimal T-DNA vectors. Plant Cell, Tissue and Organ Culture *70*, 61–68.

Bartee, L., Malagnac, F., and Bender, J. (2001). Arabidopsis cmt3 chromomethylase mutations block non-CG methylation and silencing of an endogenous gene. Genes Dev. *15*, 1753–1758.

Baubec, T., Pecinka, A., Rozhon, W., and Scheid, O.M. (2009). Effective, homogeneous and transient interference with cytosine methylation in plant genomic DNA by zebularine. The Plant Journal *57*, 542–554.

Bogdanović, O., and Lister, R. (2017). DNA methylation and the preservation of cell identity. Curr. Opin. Genet. Dev. *46*, 9–14.

Cao, X., and Jacobsen, S.E. (2002). Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing. Curr. Biol. *12*, 1138–1144.

Cao, X., Aufsatz, W., Zilberman, D., Mette, M.F., Huang, M.S., Matzke, M., and Jacobsen, S.E. (2003). Role of the DRM and CMT3 Methyltransferases in RNA-Directed DNA Methylation. Current Biology *13*, 2212–2217.

Chen, H., Kazemier, H.G., de Groote, M.L., Ruiters, M.H.J., Xu, G.-L., and Rots, M.G. (2014). Induced DNA demethylation by targeting Ten-Eleven Translocation 2 to the human ICAM-1 promoter. Nucl. Acids Res. *42*, 1563–1574.

Choudhury, S.R., Cui, Y., Lubecka, K., Stefanska, B., and Irudayaraj, J. (2016). CRISPR-dCas9 mediated TET1 targeting for selective DNA demethylation at BRCA1 promoter. Oncotarget *7*, 46545–46556.

Clough, S.J., and Bent, A.F. (1998). Floral dip: a simplified method for Agrobacterium - mediated transformation of Arabidopsis thaliana. The Plant Journal *16*, 735–743.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215–219.

Cubas, P., Vincent, C., and Coen, E. (1999). An epigenetic mutation responsible for natural variation in floral symmetry. Nature *401*, 157–161.

Curtis, M.D., and Grossniklaus, U. (2003). A gateway cloning vector set for high-throughput functional analysis of genes in planta. Plant Physiol. *133*, 462–469.

Gong, Z., Morales-Ruiz, T., Ariza, R.R., Roldán-Arjona, T., David, L., and Zhu, J.-K. (2002). ROS1, a Repressor of Transcriptional Gene Silencing in Arabidopsis, Encodes a DNA Glycosylase/Lyase. Cell *111*, 803–814.

Griffin, P.T., Niederhuth, C.E., and Schmitz, R.J. (2016). A Comparative Analysis of 5-Azacytidine- and Zebularine-Induced DNA Demethylation. G3 (Bethesda) *6*, 2773–2780.

Hollwey, E., Watson, M., and Meyer, P. (2016). Expression of the C-Terminal Domain of Mammalian TET3 DNA Dioxygenase in Arabidopsis thaliana Induces Heritable Methylation Changes at rDNA Loci. Abb *07*, 243–250.

Jacobsen, S.E., and Meyerowitz, E.M. (1997). Hypermethylated SUPERMAN Epigenetic Alleles in Arabidopsis. Science *277*, 1100–1103.

Johnson, L.M., Du, J., Hale, C.J., Bischof, S., Feng, S., Chodavarapu, R.K., Zhong, X., Marson, G., Pellegrini, M., Segal, D.J., et al. (2014). SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. Nature *507*, 124–128.

Kakutani, T. (1997). Genetic characterization of late-flowering traits induced by DNA hypomethylation mutation in Arabidopsis thaliana. Plant J. *12*, 1447–1451.

Kankel, M.W., Ramsey, D.E., Stokes, T.L., Flowers, S.K., Haag, J.R., Jeddeloh, J.A., Riddle, N.C., Verbsky, M.L., and Richards, E.J. (2003). Arabidopsis MET1 Cytosine Methyltransferase Mutants. Genetics *163*, 1109–1122.

Kardailsky, I., Shukla, V.K., Ahn, J.H., Dagenais, N., Christensen, S.K., Nguyen, J.T., Chory, J., Harrison, M.J., and Weigel, D. (1999). Activation Tagging of the Floral Inducer FT. Science *286*, 1962–1965.

Kato, M., Takashima, K., and Kakutani, T. (2004). Epigenetic Control of CACTA Transposon Mobility in Arabidopsis thaliana. Genetics *168*, 961–969.

Kohli, R.M., and Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. Nature *502*, 472–479.

Koornneef, M., Hanhart, C.J., and van der Veen, J.H. (1991). A genetic and physiological analysis of late flowering mutants in Arabidopsis thaliana. Molec. Gen. Genet. *229*, 57–66.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics *11*, 204–220.

Lindroth, A.M., Cao, X., Jackson, J.P., Zilberman, D., McCallum, C.M., Henikoff, S., and Jacobsen, S.E. (2001). Requirement of CHROMOMETHYLASE3 for Maintenance of CpXpG Methylation. Science *292*, 2077–2080.

Liu, X.S., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R.A., and Jaenisch, R. (2016). Editing DNA Methylation in the Mammalian Genome. Cell *167*, 233–247.e17.

Lo, C.-L., Choudhury, S.R., Irudayaraj, J., and Zhou, F.C. (2017). Epigenetic Editing of Ascl1 Gene in Neural Stem Cells by Optogenetics. Scientific Reports 2017 7 *7*, 42047.

Maeder, M.L., Angstman, J.F., Richardson, M.E., Linder, S.J., Cascio, V.M., Tsai, S.Q., Ho, Q.H., Sander, J.D., Reyon, D., Bernstein, B.E., et al. (2013). Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. Nature Biotechnology 2016 34:10 *31*, 1137–1142.

Matzke, M.A., Kanno, T., and Matzke, A.J.M. (2015). RNA-Directed DNA Methylation: The Evolution of a Complex Epigenetic Pathway in Flowering Plants. Annu Rev Plant Biol *66*, 243–267.

Miura, A., Kato, M., Watanabe, K., Kawabe, A., Kotani, H., and Kakutani, T. (2004). Genomic localization of endogenous mobile CACTA family transposons in natural variants of Arabidopsis thaliana. Mol Genet Genomics *270*, 524–532.

Morita, S., Noguchi, H., Horii, T., Nakabayashi, K., Kimura, M., Okamura, K., Sakai, A., Nakashima, H., Hata, K., Nakashima, K., et al. (2016). Targeted DNA demethylation in vivo using dCas9–peptide repeat and scFv–TET1 catalytic domain fusions. Nature Biotechnology 2016 34:10 *34*, 1060–1065.

Nguyen, A.W., and Daugherty, P.S. (2005). Evolutionary optimization of fluorescent proteins for intracellular FRET. Nature Biotechnology 2016 34:10 *23*, 355–360.

Okada, M., Kanamori, M., Someya, K., Nakatsukasa, H., and Yoshimura, A. (2017). Stabilization of Foxp3 expression by CRISPR-dCas9-based epigenome editing in mouse primary T cells. Epigenetics & Chromatin 2017 10:1 *10*, 24.

Penterman, J., Zilberman, D., Huh, J.H., Ballinger, T., Henikoff, S., and Fischer, R.L. (2007). DNA demethylation in the Arabidopsis genome. Pnas *104*, 6752–6757.

Soppe, W.J., Jacobsen, S.E., Alonso-Blanco, C., Jackson, J.P., Kakutani, T., Koornneef, M., and Peeters, A.J. (2000a). The late flowering phenotype of fwa mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. Molecular Cell *6*, 791–802.

Soppe, W.J.J., Jacobsen, S.E., Alonso-Blanco, C., Jackson, J.P., Kakutani, T., Koornneef, M., and Peeters, A.J.M. (2000b). The Late Flowering Phenotype of fwa Mutants Is Caused by Gain-of-Function Epigenetic Alleles of a Homeodomain Gene. Molecular Cell *6*, 791–802.

Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D.J., and Jacobsen, S.E. (2014). Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat. Struct. Mol. Biol. *21*, 64–72.

Stroud, H., Greenberg, M.V.C., Feng, S., Bernatavichute, Y.V., and Jacobsen, S.E. (2013). Comprehensive Analysis of Silencing Mutants Reveals Complex Regulation of the *Arabidopsis* Methylome. Cell *152*, 352–364.

Tanenbaum, M.E., Gilbert, L.A., Qi, L.S., Weissman, J.S., and Vale, R.D. (2014). A Protein-Tagging System for Signal Amplification in Gene Expression and Fluorescence Imaging. Cell *159*, 635–646.

Taylor, S.M., and Jones, P.A. (1982). Changes in phenotypic expression in embryonic and adult cells treated with 5-azacytidine. Journal of Cellular Physiology *111*, 187–194.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Wu, X., and Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. Nature Reviews Genetics *18*, 517–534.

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics *10*, 232.

Xu, X., Tao, Y., Gao, X., Zhang, L., Li, X., Zou, W., Ruan, K., Wang, F., Xu, G.-L., and Hu, R. (2016). A CRISPR-based approach for targeted DNA demethylation. Cell Discovery 2016 2 *2*, 16009.

Zemach, A., Kim, M.Y., Hsieh, P.-H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L., and Zilberman, D. (2013). The Arabidopsis Nucleosome Remodeler DDM1 Allows DNA Methyltransferases to Access H1-Containing Heterochromatin. Cell *153*, 193–205.

Zhang, H., and Zhu, J.K. (2012). Active DNA Demethylation in Plants and Animals. Cold Spring Harb Symp Quant Biol *77*, 161–173.

Zhu, J., Kapoor, A., Sridhar, V.V., Agius, F., and Zhu, J.-K. (2007). The DNA Glycosylase/Lyase ROS1 Functions in Pruning DNA Methylation Patterns in Arabidopsis. Current Biology *17*, 54–59.

CHAPTER 4

Naive Human Pluripotent Cells Feature a Methylation Landscape Devoid of Blastocyst or Germline Memory

Contributions

The reversion and culture of hESCs in naive conditions was conducted by R.K., D.C., and A.S. hESC derivation from human blastocyst was undertaken by R.K. Experiments and data interpretation was conducted by D.C., W.A.P., A.L. and R.K. Computation analysis was conducted by W.L. and W.A.P. Conceiving and directing research was conducted by K.P., S.E.J. and A.T.C. Maintenance of University Compliance, including ESCRO, IRB and Biological Safegy, was overseen by A.T.C. The manuscript was written by W.A.P. and A.T.C.

ABSTRACT

Human embryonic stem cells (hESCs) typically exhibit 'primed' pluripotency, analogous to stem cells derived from the mouse post-implantation epiblast. This has led to a search for growth conditions that support self-renewal of hESCs akin to hypomethylated naive epiblast cells in human pre-implantation embryos. We have discovered that reverting primed hESCs to a hypomethylated naive state or deriving a new hESC line under naive conditions results in the establishment of Stage Specific Embryonic Antigen 4 (SSEA4)-negative hESC lines with a transcriptional program resembling the human pre-implantation epiblast. In contrast, we discovered that the methylome of naive hESCs *in vitro* is distinct from that of the human epiblast *in vivo* with loss of DNA methylation at primary imprints and a lost 'memory' of the methylation state of the human oocyte. This failure to recover the naive epiblast methylation landscape appears to be a consistent feature of self-renewing hypomethylated naive hESCs *in vitro*.

INTRODUCTION

Human embryonic stem cells (hESCs) are *in vitro* pluripotent cell types with the capacity for unlimited self-renewal and differentiation, making them critical models for understanding mechanisms required for human embryo development and differentiation. Although hESCs are derived from pre-implantation human blastocysts, they are morphologically and transcriptionally similar to murine epiblast stem cells (EpiSCs), which are derived from post-implantation mouse embryos. As such, hESCs and EpiSCs are said to exhibit a 'primed pluripotent state' while mouse ESCs derived from the pre-implantation blastocyst exhibit a 'naive pluripotent state' corresponding to an earlier stage of development (Nichols and Smith, 2009).

A number of culture conditions have recently been developed that promote maintenance and self-renewal of naive human pluripotent stem cells (Chan et al., 2013) (Gafni et al., 2013; Takashima et al., 2014; Theunissen et al., 2014; Ware et al., 2014). Each protocol generates cell types with slightly different molecular characteristics, which may reflect metastable states in the spectrum of naive to primed pluripotency. A recent meta-analysis of sequencing data indicates that two of these protocols generate cells with a close transcriptional resemblance to the human pre-implantation epiblast (Huang et al., 2014). In the first protocol, hESCs are transfected with KLF2 and NANOG and are cultured in media with titrated two inhibitors plus leukemia inhibitory factor and Gö6983 (t2iL+Gö) (Takashima et al., 2014). In the second protocol, primed cells can be reverted by being transferred to a media containing a cocktail of five inhibitors plus LIF, Activin, and/or Fibroblast Growth Factor 2 (5iLAF) (Theunissen et al., 2014). Using t2iL+Gö reversion of the H9 primed hESC line, it was shown that DNA methylation is globally reduced to the average level measured in human pre-implantation epiblasts (Takashima et al.,

2014), with additional locus-specific erosion in the 5' region of the LINE1 human specific (L1HS) retrotransposons (Gkountela et al., 2015). The DNA methylation profile of cells cultured in 5iLAF has never been evaluated.

RESULTS

Before studying the methylation pattern of 5iLAF cultured cells, we first wanted to confirm and characterize the naive phenotype. We performed n = 4 independent reversions of the hESC line UCLA1 (Diaz Perez et al., 2012) using 5iLAF (Theunissen et al., 2014). Upon the reversions we observed a mixture of small, round colonies similar to naive mESCs as well as flat, cobblestone-like colonies (Figures 4-1A,B). We evaluated one reversion using two classic human pluripotency surface markers called SSEA4 and TRA-1-81. Unlike primed UCLA1 hESCs, which are double positive for SSEA4 and TRA-1-81, the 5iLAF-reverted hESCs have a large fraction of double-negative cells (Figures 4-1A,B). Immunofluorescence staining showed that the SSEA4- and TRA-1-81-negative cells were still positive for OCT4 and NANOG (Figures 4-2A-F).

Next, we sorted the 5iLAF-cultured cells into SSEA4-positive and -negative populations using fluorescence-activated cell sorting (FACS) and re-plated the sorted cells onto MEFs in 5iLAF media (Figure 4-1C). We discovered that SSEA4-positive cells yielded mostly flat colonies, whereas SSEA4-negative cells yielded mostly round colonies. One passage after sorting, the SSEA4-negative population remained SSEA4 negative, indicating that this is a relatively stable state (Figure 4-2G). We then reverted two additional lines called UCLA4 and UCLA5 (Diaz Perez et al., 2012) and found that small, round colony morphology was always enriched in the

SSEA4-negative fraction whereas the SSEA4-positive cells yielded mostly flat, cobblestone colonies (Figures 4-2H–K).

In order to determine whether the heterogeneity in SSEA4 expression was also observed when deriving hESC lines completely under naive 5iLAF conditions, we derived n = 2 new hESC lines, which we have called UCLA19n and UCLA20n after thawing n = 7 day 5 vitrified human blastocysts. Colonies were uniformly round and flow cytometry revealed that UCLA19n was 85% SSEA4 negative (results not shown), whereas UCLA20n was almost completely SSEA4 negative (Figure 4-1D). In contrast, TRA-1-81 was expressed on a significant portion of SSEA4-negative cells in UCLA20n as well as reverted UCLA4 and UCLA5 hESC lines (Figure 4-1D, Figure 4-2I,K). Therefore, loss of TRA-1-81 is not a consistent marker of naive morphology, whereas absence of SSEA4 is a highly correlated feature of naive round colony morphology. In summary, reversion of primed hESCs in 5iLAF generates a heterogeneous mixture of colonies, with SSEA4-negative hESCs correlating with small round colony morphology similar to naive hESCs derived from the human pre-implantation blastocyst.

On the basis of morphology, we speculated that 5iLAF SSEA4-negative hESCs are the naive population and thus transcriptionally resemble the cells of the human pre-implantation epiblast. To address this, we performed RNA-seq of 5iLAF-cultured SSEA4-positive or SSEA4-negative fractions of UCLA1, and we compared them to SSEA4-positive primed UCLA1 hESCs at equivalent passages. We also performed RNA-seq of UCLA20n at passage 20 after derivation. We did not analyze UCLA19n as it was found to be 70% polyploid by passage 15. Consistent with the expression patterns of genes associated with naive pluripotency in mice, the 5iLAF

SSEA4-negative cells and UCLA20n had elevated levels of NANOG as well as a dramatic upregulation of KRUPPLE-LIKE FACTOR (KLF) family transcription factors and reduced expression of primed state master regulators such as ZINC FINGER OF THE CEREBELLUM (ZIC) family transcription factors and OTX2 (Buecker et al., 2014; Tang et al., 2011; 2016) (Yang et al., 2014) (Figure 4-1E). To further confirm the similarity of 5iLAF SSEA4-negative and UCLA20n hESCs to the human pre-implantation blastocyst, we used the previously published single-cell expression data from late pre-implantation epiblast and primed hESCs (Yan et al., 2013). We defined a set of "pre-implantation epiblast-specific" and "primed-specific" genes, which showed >4-fold difference in expression between these two cell types . Using these genes as a reference, we found that the 5iLAF SSEA4-negative hESCs and UCLA20n had global upregulation of naive epiblast-specific genes and downregulation of primed-specific genes (Figure 4-1F).

In contrast, the SSEA4-positive hESCs sorted from 5iLAF cultures had an intermediate expression pattern between primed and naive, suggesting that SSEA4-positive cells that stably self-renew in 5iLAF are partially reverted to the naive state (Figures 4-1E,F). Comparing to published datasets, we found that the SSEA4-negative population in UCLA1 and the new hESC line UCLA20n is analogous to the original 5iLAF hESC lines created by reverting WIBR2 (Theunissen et al., 2014) and to t2iL+Gö-cultured hESCs created by reverting H9 (Takashima et al., 2014). In contrast, other published naive methods showed a less pronounced shift toward the naive state and failure to repress primed markers (Chan et al., 2013) (Gafni et al., 2013) (Ware et al., 2014). Interestingly, lines generated by these methods are also reported to be SSEA4

positive. Given these results, we focused our methylation analysis on the 5iLAF and t2iL+Gö conditions.

To determine the methylation pattern of hESCs in 5iLAF, we performed whole-genome bisulfite sequencing (WGBS) on two to four independent sorts of SSEA4-negative or SSEA4-positive reverted UCLA1 cells, SSEA4-negative UCLA20n cells, and primed UCLA1 cells that had been in culture a similar length of time to the reverted lines. We discovered that, similar to the levels observed in t2iL+Gö (Gkountela et al., 2015) (Takashima et al., 2014), 5iLAF-cultured SSEA4-negative hESCs and UCLA20n had an average CG methylation level that resembled that of the human blastocyst (Figure 4-3A) (Okae et al., 2014).

In mammals the methylation pattern of the blastocyst is shaped by events during gametogenesis and early embryogenesis. The male pronucleus is selectively demethylated in early embryonic development, with only a few regions such as paternally methylated imprinted loci protected from DNA demethylation (Okae et al., 2014; Smith et al., 2012) (Smith et al., 2014). Thus in humans the methylation pattern of the blastocyst strongly resembles that of the oocyte (Figures 4-3B,C). In contrast, the methylation landscapes of SSEA4-negative 5iLAF-cultured hESCs, UCLA20n, and t2iL+Gö-cultured cells are only weakly correlated with the human blastocyst and human oocyte (Figures 4-3B,C). Naive cells, even if cultured by different methodologies or derived directly from the human blastocyst, converge toward a methylation pattern that is different from that of the pre-implantation human blastocyst (Figures 4-3B,C). A striking example of this trend is observed at 332 CpG islands identified previously as “transient maternal imprints:” sites that are highly methylated in oocytes and the maternal chromosomes of

blastocyst that lose methylation upon implantation (Smith et al., 2014). We discovered that reversion does not regenerate methylation at these sites, nor is methylation retained at these transient maternal imprints in the UCLA20n hESC line (Figure 4-4A).

An additional, striking deviation from oocyte and blastocyst methylation patterns in 5iLAF and t2iL+Go cultured cells occurred at stable imprints. These are regions where DNA methylation is established exclusively during germ-cell development. These methylated sites are protected from DNA demethylation during pre-implantation embryo development, remaining differentially methylated in somatic cells through the life of the organism and promoting a parent-of-origin-specific expression pattern in the neighboring genes. We examined DNA methylation at 29 maternally methylated stable primary imprints and 2 paternally methylated stable primary imprints (Okae et al., 2014) (Figure 4-3D). There is roughly 50% methylation in somatic tissue and slightly below 50% methylation in blastocysts, as expected. In the primed UCLA1 hESCs used in our study, the median methylation of these imprinted sites was close to 50%, though some imprints were hyper- or hypomethylated, similar to what has been observed previously for other hESC lines (Rugg-Gunn et al., 2007). Strikingly, the 5iLAF SSEA4-negative hESCs and UCLA20n had near complete loss of methylation from all 31 primary stable imprints evaluated in our study, with loss over many imprints also found in t2iL+Gö (Figures 4-3D,E). Taking advantage of single nucleotide polymorphisms (SNPs) present in the UCLA1 hESC line, we observed a shift upon reversion in 5iLAF from monoallelic to biallelic expression of several imprinted genes including H19 and SNRPN. (Figure 4-4B). In order to determine whether methylation could be restored at imprinted genes by reverting the naive hESCs back to a primed state, we cultured 5iLAF SSEA4-negative and primed UCLA1 cells in primed epiblast-like cell

(EpiLC) media (Hayashi et al., 2011) for 16 days. During this time, we discovered that 5iLAF SSEA4-negative cells showed a global shift toward expression of primed-specific genes and gained DNA methylation genome-wide (Figure 4-4C). However, increased methylation over imprinted regions was very modest, and biallelic expression was still observed (Figure 4-4D). Thus, when lost, imprinting is not re-established in cells cultured in primed conditions, a similar scenario to the rescue of global DNA methylation, but not imprint methylation, in Dnmt1 knockout ESCs by the re-expression of Dnmt1 (Holm et al., 2005). Furthermore, consistent with data observed in hESCs cultured in t2iL+Gö (Gkountela et al., 2015), young LINE elements also show dramatic promoter hypomethylation in 5iLAF (Figure 4-4E).

Given the problem with maintenance of imprint methylation in naive cells, we considered the possibility that the 5iLAF SSEA4-positive cells may represent a useful intermediate. However, we discovered that the SSEA4-positive cells showed intermediate levels of global and imprint methylation loss (Figures 4-3C,D), with biallelic expression of SNRPN and H19. We also analyzed the methylation loss imparted by naive human stem cell media (NHSM) (Gafni et al., 2013), which shows the smallest transcriptional shift toward naive pluripotency (Figures 4-1E,F). In order to directly compare our methylome data to findings of Gafni et al. (2013), we modified all whole-genome datasets to simulate the Reduced Representation Bisulfite Sequencing (RRBS) approach used by Gafni to measure DNA methylation. We discovered that imprint methylation was unperturbed in NHSM (Figure 4-4F), but very little global change in methylation was observed either (Figure 4-3F).

Consistent with an initial report of karyotypic instability in 5iLAF culture (Theunissen et al., 2014), we discovered that, 24 passages after reversion, the 5iLAF UCLA1 hESCs developed widespread karyotypic abnormalities, which was not observed in the first 13 passages following reversion (Figure 4-4G). Similarly, UCLA20n had evidence of trisomies at chromosomes 3, 7, 12, and 20 by passage 14 and as discussed above, UCLA19n was 70% polyploid at passage 15 (Figures 4-4G,H). Therefore, karyotypic instability may also be a frequent consequence of naive hESC culture.

To determine the cause of the cells' failure to maintain DNA methylation, we analyzed changes in RNA and protein levels of DNA methylation and demethylation machinery. We found that the RNA and protein levels of the de novo DNA methyltransferase DNMT3B dropped sharply in the 5iLAF SSEA4-negative cells, while DNMT3A was unchanged and DNMT3L increased dramatically relative to primed hESCs. UHRF1 RNA levels were slightly elevated in naive hESCs. However, at the protein level, we observed a 65% loss of UHRF1, and both DNMT1 RNA and protein levels were reduced by 50% in the naive state. Furthermore, expression of the 5mC oxidases TET1 and TET2 increased substantially in the naive state (Figures 4-3G,H).

DISCUSSION

In the current study we have shown that naive hESCs have a transcriptional program enriched in human pre-implantation-specific genes but with a global DNA methylation landscape that is distinct from the normal state of the human pre-implantation blastocyst. The negative effect of the loss of "transient imprints" and the failure to recapitulate the oocyte-like methylation pattern is unclear. However, the loss of stable primary imprints is potentially serious in human

pluripotent stem cell research. Correct imprinting is necessary for organism survival, and a number of rare human medical disorders have been linked to aberrant imprinting (Butler, 2009). Of note, murine embryonic germ cell lines are transcriptionally similar to murine ESCs but have widespread loss of imprints and contribute poorly to chimeras (Leitch et al., 2013; Oliveros-Etter et al., 2015) (Tada et al., 1998), demonstrating the importance of imprints in correct differentiation of pluripotent cells in vivo. We also observed, in accompaniment to the loss of imprints, extensive karyotypic abnormalities in cells after their prolonged culture in 5iLAF. Loss of DNA methylation has been linked to karyotypic instability (Haaf, 1995).

We note that methylation at the imprinted loci is clearly depressed relative to that of surrounding regions. This may reflect the observation that many imprinted loci are promoters or regulatory elements that are active in the blastocyst (Rugg-Gunn et al., 2007). Thus if methylation is partially eroded at the imprint, the relevant transcription factors bind and cause further demethylation (as is generally the case at these genetic elements). In other words, methylation may be a very weak barrier to locus activation in 5iLAF. Similar dynamics may be at work at L1HS elements.

Although we observed a reduction in DNMT3B protein in the naive cells, we propose that this has only modest effects on creating the 5iLAF methylome given that DNMT3A−/− DNMT3B−/− DKO primed hESCs maintain primary imprints and show only modest DNA demethylation even after extended culture (Liao et al., 2015). We therefore propose that a combination of impaired maintenance methylation and increased TET activity could explain the majority of the 5iLAF hypomethylation phenotype. In a cell type with impaired maintenance and

some continuous de novo methylation (imparted by DNMT3A and the remaining DNMT3B), DNA methylation levels will reach a steady state, but memory of previous methylation will be lost with DNA replication.

MATERIALS AND METHODS

Cell culture

Reversion and culture of cells was adapted from the published 5iLAF protocol (Theunissen et al., 2014). UCLA-derived human embryonic stem cell lines were routinely maintained in DMEM/F-12 (Life Technologies), 20% KSR (Life Technologies), 10ng/mL bFGF (Peprotech), 1% nonessential amino acids (Life Technologies), 2mM GlutaMAX (Life Technologies), penicillin-streptomycin (Life Technologies) and 0.1mM beta-mercaptoethanol (Sigma) and passaged with 1mg/mL collagenase type IV (Life Technologies). During maintenance, they were passaged once every seven days. To achieve reversion, two days post passage, medium was changed to DMEM/F-12, 15% FBS (Omega Scientific), 5% KSR (Life Technologies), 4ng/mL bFGF (Peprotech), 1% nonessential amino acids (Life Technologies), 1mM GlutaMAX (Life Technologies), penicillin-streptomycin (Life Technologies) and 0.1mM beta-mercaptoethanol (Life Technologies). On day 7 post-passage, cells were washed once with 1x dPBS (Life Technologies) and treated for 3 min. with 0.05% trypsin-EDTA (Life Technologies). Cells were dissociated into a single cell suspension, passed through a 40μm cell strainer and plated at a density of $2x10^5$ cells per 9.5 cm^2 in the 15% FBS containing medium with the addition of 10μM Y-27632 (Stemgent). Subsequent media changes were in the absence of Y-27632. Two days post plating, medium was changed to 5iLAF with daily changes thereafter. 5iLAF medium contained a 50:50 mixture of DMEM/F-12 (Life Technologies) and

Neurobasal (Life Technologies), with 1x N2 supplement (Life Technologies), 1x B27 supplement (Life Technologies), 8ng/mL bFGF (Peprotech), 1% nonessential amino acids (Life Technologies), 1mM GlutaMAX (Life Technologies), penicillin-streptomycin (Life Technologies), 0.1mM beta-mercaptoethanol (Life Technologies), 50μg/mL BSA (Sigma), 1μM PD0325901 (Stemgent), 1μM IM-12 (Enzo), 0.5μM SB590885 (R&D Systems), 1μM WH-4-023 (A Chemtek), 10μM Y-27632 (Stemgent), 20ng/mL Activin A (Peprotech), 20ng/mL rhLIF (Millipore) and 0.5% KSR (Life Technologies). At about 11 to 12 days post plating, cells were passaged using a 3 min. treatment with StemPro Accutase (Life Technologies) and replated after passing through a 40μm cell strainer in 5iLAF medium. Round naive colonies could be seen at this subsequent passage. Cultures were maintained in 5iLAF and passaged every 5-6 days using Accutase. All cultures were grown on a MEF layer seeded at a density of $1.5x10^6 – 2.5x10^6$ cells per 6-well plate. Cells were cultured in ambient oxygen and 5% CO_2.

Analysis of surface markers

In flow cytometry or FACS experiments, cells were detached by Accutase and centrifuged and washed with 1xFACS buffer (1xPBS 1% BSA). Antibodies used for staining include: PE-conjugated TRA-1-85 (R&D systems, FAB3195P), APC-conjugated anti-SSEA4 (R&D systems, FAB1435), and Alexa488-conjugated anti-TRA-1-81 (Cell Technologies, 60065). DAPI was added immediate prior to flow cytometry or FACS.

To generate stable SSEA4 negative or positive lines from 5iLAF cultures, live (DAPI negative) human ($TRA\text{-}1\text{-}85^+$) cells were sorted into $SSEA4^-$ and $SSEA4^+$ subpopulations. The sorted cells were then centrifuged at 1200RPM, re-suspended in 5iLAF media, and plated at a density of 300k/well of a 6-well plate.

To generate material for sequencing or Western blots, TRA-1-85^{+} SSEA4^{-} cells were sorted from the 5iLAF SSEA4^{-} culture, and TRA-1-85^{+} SSEA4^{+} cells were sorted from 5iLAF SSEA4^{+} or primed cultures. The cells were then centrifuged at 1200RPM five minutes and frozen, and DNA, RNA or protein extracted as described below.

Human hESC derivation in 5iLAF

Derivation of UCLA19n and UCLA20n were performed with vitrified day five human blastocysts under hypoxic conditions (5% O_2, 3%CO_2). A total of seven human blastocysts were used for these experiments. Blastocysts were received vitrified from the in vitro fertilization clinic following informed consent and thawed using Vit Kit-Thaw (Irvine Scientific) according to manufacturer protocol. The embryos were cultured in drops of Continuous Single Culture media (Irvine Scientific) supplemented with 20% Quinn's Advantage SPS Serum Protein Substitute (Sage Media) under mineral oil (Irvine Scientific) overnight at 37°C, 6% CO_2 and 5% O_2. The zona pellucida was removed using Tyrode's solution acidified (Irvine Scientific) before plating onto inactivated MEFs in 5iLAF media at passage (P) 0. Derivation success rate involved 5/7 blastocysts attaching to the MEFs at P0, and 2/7 giving rise to naive (n) hESC lines capable of self-renewal for at least 15 passages. Accutase was used to harvest the P0 blastocyst outgrowths at day 6 (UCLA19n) and day 9 (UCLA20n). UCLA20n was supplemented from P0-P3 with a 50:50 mix of MEF conditioned media (20% knockout serum replacer and 4ng/ml FGF2) and 5iLAF to promote colony outgrowth. Starting at P4 UCLA20n was maintained exclusively in 5iLAF on inactivated MEFs under normoxic conditions according to methods described above for reverted hESC lines. UCLA19n was cultured from P0-P14 in 5% O_2, 3%CO_2 in 5iLAF on MEFs. Pluripotent stem cell identity for UCLA19n was confirmed by round

dome-shaped colony morphology at all passages and positive immunofluorescence staining for TRA-1-81, OCT4 and NANOG at passage 5. Flow cytometry was performed at passage 7 revealing 60% Tra-1-81 positive and 85% SSEA4 negative cells. UCLA19n cultures were sent for karyotyping by Cell Line Genetics Inc. (Madison, WI) at passage 15 resulting in the discovery that UCLA19n was 70% polyploidy. UCLA20n was characterized by a round, dome-shaped colony morphology at all passages, together with SSEA4 negative staining. Array comparative genomic hybridization (aCGH) was performed at passage 14 by Cell Line Genetics Inc revealing gains at chromosome 3, 7 and 12. 5iLAF ESC derivations were also attempted under normoxic conditions using n=5 day five vitrified human blastocysts. Under these conditions, 4/5 blastocysts attached at P0, however none resulted in ESC lines.

Human embryo studies were approved by the full UCLA Institutional Review Board (IRB#11-002027) and the UCLA Embryonic Stem Cell Research Oversight (ESCRO) Committee (2007-005).

Culture in Epiblast like-cell (EpiLC conditions)

MEF-depleted hESCs were plated at 200k/well in Human Plasma Fibronectin (Invitrogen)-coated 12-well-plate for 16 days in EpiLC media. Media were changed everyday and EpiLCs were split every 4 days. EpiLC media is a 50:50 ratio of DMEM-F-12 (Life Technologies) and Neurobasal media (Life Technologies) with 1x N2 supplement (Life Technologies), 1x B27 supplement (Life Technologies), 1% KSR (Life Technologies), 10ng/mL bFGF (Peprotech), 20ng/mL Activin A (Peprotech), 10μm Y-27632 (Stemgent).

Western quantitation

Sorted cells were centrifuged 800g 5 minutes, then resuspended in 1xLaemmli buffer at 5k cells/μL and denatured for five minutes at 99°C. Samples were run a 10% Bis-Tris gel (ThermoFisher), transferred at 60-70V for 3 hours, and blocked with 1xOdyssey Blocking Buffer overnight (Licor). Primary and secondary antibody incubation was conducted in 1xOdyssey Blocking buffer 0.15% Tween. The following antibodies were used to stain and quantify DNA methyltransferase levels:

Antigen	Catalog Number	Manufacturer	Concentration
UHRF1	373750	Santa Cruz	1:500
DNMT1	20701	Santa Cruz	1:500
DNMT1	Gift from S. Pradhan	New England Biolabs	1:5000
DNMT3A	20703	Santa Cruz	1:1000
DNMT3A	13888	Abcam	1:1000
DNMT3B	2851	Abcam	1:1000
DNMT3L	39908	Active Motif	1:500

Because fluorescently labeled anti-mouse and anti-rabbit antibodies can be used simultaneously (provided they are conjugated to different fluorophores), multiple proteins were stained simultaneously. Both DNMT3A antibodies were used simultaneously to allow identification of the correct band. For other proteins, costaining with anti-H3 antibody (either Abcam 1791 or Abcam 10799) at 1:5000 was performed to confirm similar loading or to provide relative concentration. After antibody staining, the cells were washed four times with 1xPBS 0.1% tween. Fluorescently conjugated anti-mouse and anti-rabbit secondary antibody (Licor) were used at 1:20,000 concentration in 1xOdyssey Blocking buffer 0.15% Tween and incubated for 45 minutes. The blots were again washed 4x5minutes with 1xPBS 0.1% tween and then rinsed quickly with 1xPBS to remove detergent. The blots were then dried and imaged on an Odyssey Infrared Imager (Licor). Band quantitation was performed using the instrument software.

Immunofluorescence

Colonies of primed or naive cells were dissociated with Collagenase IV then fixed in 4% paraformaldehyde, embedded in paraffin, and then sectioned and mounted on slides. Slides were deparaffinized by successive treatment with xylene and 100%, 95%, 70% and 50% ethanol, and antigen retrieval was performed by incubation with 10mM Tris pH 9.0, 1mM EDTA, 0.05% Tween 95°C for 40 minutes. The slides were cooled and washed with 1xPBS and 1xTBS 0.05% tween. The samples were permabilized with 0.5% Triton X-100 in 1xPBS, then washed with 1xTBS 0.05% Tween and blocked with 10% donkey serum in 1xTBS-tween. Primary antibody incubation was conducted overnight in 10% donkey serum, using these antibodies:

Antigen	Catalog Number	Manufacturer	Concentration
Oct4	8628	Santa Cruz	1:100
Nanog	AF1997	R&D Systems	1:20
SSEA4	MC-813-70	DSHB	1:100
Tra-1-81	14-8883-82	eBiosciences	1:100

Samples were again washed with 1xTBS-tween and incubated with fluorescent secondary at 1:250 for 45 minutes, then washed and mounted using with ProLong Gold Antifade Mountant with DAPI (ThermoFisher). Slides were imaged on an LSM 780 Confocal Instrument (Zeiss).

DNA preparation.

DNA for bisulfite sequencing was extracted using the Quick gDNA Mini-Prep Kit (Zymo) and quantified using the Qubit dsDNA High Sensitivity Kit (Life Technologies).

RNA preparation.

RNA for RNA-seq was extracted using the RNeasy Micro Kit (Qiagen) and quantified using a Nanodrop ND-1000 (Nanodrop).

Library preparation

RNA sequencing libraries were prepared using the Nugen RNA-seq System V2 with 5-100ng starting material. Bisulfite sequencing libraries were prepared using the Ovation Ultralow Methyl-Seq Library System (Nugen). Unmethylated Lambda phage DNA (NEB) was spiked in at 0.25% input DNA quantity to determine conversion efficiency, which was 99.3%-99.5% for all libraries.

Sequencing

Libraries were sequenced on Illumina HiSeq instruments (Illumina).

RNAseq analysis

Analysis of individual gene expression

Reads were first aligned to hg19 gene annotation using Tophat (Trapnell et al., 2009) by allowing up to two mismatches and only keeping reads that mapped to one location. When reads did not mapped to the annotated genes, the reads were mapped to hg19 genome. Number of reads mapping to genes were calculated by HTseq (Anders et al., 2015) with default parameters. Expression levels were determined by RPKM (reads per kilobase of exons per million aligned reads). For RNAseq of published datasets (Chan et al., 2013), raw reads were processed exactly the same as described above.

Analysis of published array data

Different datasets were processed slightly differently. For Gafni et al, Ware et al and Theunissen et al, processed gene expression levels were downloaded from Gene Expression Omnibus (GEO) or European Bioinformatics Institute (EBI). Microarray probe ID were converted to gene symbol using Bioconductor packages in R. Different probes corresponding to same gene were randomly chosen for future processing. For the gene expression level of Takashima et al, raw expression datasets were downloaded from EBI database. Raw data were processed using Bioconductor packages in R. Affymetrix arrays were normalized using the RMA method, and genes with multiple probes were represented by the arithmetic mean value.

Heatmap on pluripotency genes

RPKM were obtained for each pluripotency genes. Heatmap was plotted over log2 fold changes of 5iLAF SSEA4 negative and 5iLAF SSEA4 positive comparing to Primed cells in R.

Analysis of "pre-implantation blastocyst epiblast" and "primed hESC" up-regulated genes

Pre-implantation blastocyst epiblast and primed hESC expression level (RPKM) as well as differential expressed genes list were obtained from published data (Yan et al., 2013). Genes with greater than 4 fold change as well as a FDR less than 0.05 in epiblast compared to primed hESC are defined as "pre-implantation blastocyst epiblast" up-regulated genes and vice versa.

Whole Genome Bisulfite Sequencing Analysis

Reads were split into 50 bp reads before mapping. Reads were aligned to the hg19 genome using BSMAP (Xi and Li, 2009) by allowing up to 2 mismatches and only retaining reads mapped to one location. Methylation ratio are calculated by #C/(#C+#T) at CG sites.

Metaplot of WGBS data

Metaplot of WGBS data were made using custom Perl and R scripts. Briefly, regions of interest were broken into 50 bins while flanking 1kb regions were each broken into 25bins. CG methylation level in each bin was then determined. Metaplots were then generated with R.

Analysis on imprints

Coordinates for stable primary imprints were obtained from published data (Okae et al., 2014). Transient maternal imprints were defined as CpG islands having higher methylation in blastocyst than sperm (>20% absolute difference), no substantial evidence of *de novo* methylation in blastocyst (<20% absolute increase between cleavage and blastocyst) and low methylation in later development (<20% methylation in brain), using methylation data from (Smith et al., 2014). Percent methylation over imprints was called using data from CG methylation levels were then calculated on those imprints by custom Perl scripts.

Repeat analysis on L1HS, L1PA2, L1PA3

Repeat annotation file of hg19 was downloaded from UCSC genome browser (http://genome.ucsc.edu/). For the metaplot of L1HS and L1PA2, only repeats longer than 6kb were retained for plotting.

Comparison to RRBS data

To compare WGBS sets to published RRBS data (Gafni et al., 2013), we used a custom Python script to filter mapped WGBS data and eliminate data from any CG that was not covered at least once in the RRBS sets. Any imprints that did not have at least one hundred methylation calls for CGs (e.g., if there is tenfold coverage of one CG, that is ten calls), were excluded from further analysis, so only sixteen of the thirty-one possible imprints were analyzed.

Alterations to images

Two brightfield microscopy images (Figure 4-1A,C lower image) were brightened using Adobe Photoshop in order to improve the visibility of the printed figures. Brightness was increased uniformly across the image.

FIGURE LEGENDS

Figure 4-1. 5iLAF SSEA4 negative subpopulation recapitulates naive expression pattern. **(A)** Upper: brightfield image of primed UCLA1 hESCs. Lower: Flow cytometry plot of primed UCLA1 hESCs stained for SSEA4 and TRA-1-81. **(B)** Upper: UCLA1 hESCs reverted in 5iLAF. A mixture of round and flat colonies are observed. Lower: Flow cytometry plot of 5iLAF cultured UCLA1 hESCs stained for SSEA4 and TRA-1-81. Scale bar indicates 200μm. **(C)** 5iLAF cells were sorted into $SSEA4^+$ and $SSEA4^-$ populations. Upon re-plating, the $SSEA4^+$ cells formed flat colonies and the $SSEA4^-$ cells formed round colonies (n=2 biological replicates). Scale bar= 100μm. **(D)** hESC line called UCLA20n, derived from a 5-day human blastocyst in 5iLAF. Left: Brightfield image. Scale bar indicates 200μm. Right: Flow cytometry plot of TRA-1-85+ (human) UCLA20n hESCs stained for SSEA4 and TRA-1-81. **(E)** Expression of genes identified by others as associating with the naive and primed states in mice. Expression level is determined by RNA-seq. For 5iLAF SSEA4 negative (neg) and primed hESCs, n=4. For 5iLAF SSEA4 positive (pos), n=2. Other data comes from published RNA-seq or microarray datasets. Methodology, cell type, and citation are indicated. (**F)** A set of "pre-implantation epiblast" and "primed" specific genes were defined based on published data. Expression of these genes is shown for various methodologies, relative to primed controls from the same dataset. UCLA20n was normalized to a primed UCLA1 library generated and sequenced at the same time.

Figure 4-2. Properties of 5iLAF SSEA4 negative and SSEA4 positive cells generated by reversion of primed hESCs. Related to Figure 4-1.

(A-C) Immunofluorescence for SSEA4 and OCT4. Note that all colonies are OCT4 positive. **A)** A colony of SSEA4 positive primed UCLA1 hESCs. **(B)** A colony of 5iLAF SSEA4 positive UCLA1 hESCs. **(C)** A colony of 5iLAF SSEA4 negative UCLA1 hESCs. **(D-F)** Immunofluorescence for TRA-1-81 and NANOG. Note that all populations are NANOG positive. **(D)** A colony of TRA-1-81 positive UCLA1 hESCs. **(E)** A colony of 5iLAF TRA-1-81 positive UCLA1 hESCs. **(F)** A colony of 5iLAF TRA-1-81 negative UCLA1 hESCs. **(G)** Flow cytometry of control primed and re-plated 5iLAF SSEA4 negative UCLA1 hESCs grown for one passage. Fluorescence of unstained cells is indicated in red and percentage showing positive staining is indicated. After re-plating, the vast majority of sorted SSEA4 negative cells remain SSEA4 negative. **(H-K)** Similar to UCLA1, SSEA4 positive 5iLAF cells from UCLA4 **(H)** and UCLA5 **(J)** yield flat colonies upon re-plating while the SSEA4 positive 5iLAF cells yield round colonies. **(I,K)** Unlike UCLA1, most cells in the SSEA4 negative subpopulation have high TRA-1-81 expression (compare to Figure 4-1B).

Figure 4-3. Naive hESCs fail to recapitulate naïve-specific methylation pattern.

(A) Average genome wide-CG methylation level in primed and 5iLAF UCLA1 hESCs, shown in comparison with published datasets. For 5iLAF SSEA4 negative (neg) and primed hESCs, n=3. For 5iLAF SSEA4 positive (pos), n=2. **(B)** DNA methylation is shown for a region of chromosome 10. Each bar indicates a single CG, and the height of the bar indicates the percentage of CG methylation. Where multiple CGs are too close to be visually rendered separately, an average value is shown. **(C)** Correlation plots relative to human oocyte using 100kb genome bins. **(D)** DNA methylation over stable primary imprints. The average

methylation level of each imprint in a given sample is represented as one point in the box and whisker point. **(E)** DNA methylation over the paternally imprinted *H19* locus. Each bar indicates a single CG, and the height of the bar indicates the fraction of CG methylation. Where multiple CGs are too close to be visually rendered separately, an average value is shown. **(F)** Total DNA methylation for three competing approaches for culturing naive cells. Because the Gafni 2013 data was generated by RRBS, only CGs that had coverage in the Gafni 2013 dataset are included in this analysis to make the data comparable. **(G)** Expression (RPKM) of DNA methyltransferases, DNMT cofactors, and Tet-family oxidases as measured by RNA-seq (n=4). **(H)** RNA and protein levels of DNA methyltransferases in 5iLAF SSEA4 neg UCLA1 hESCs relative to primed. RNA level is determined from RNA-seq data (n=4), protein level from quantitative westerns (UHRF1, n=6 Western blots; DNMT1, DNMT3A, DNMT3B n=2; DNMT3L n=1).

Figure 4-4. Methylation pattern in Naive hESCs. Related to Figure 4-3.

(A) DNA methylation over transiently imprinted CG islands. The average methylation level of each imprint in a given sample is represented as one point in the box and whisker point. **(B)** Reads mapped over an annotated SNP in the maternally imprinted SNRPN locus. Reads over each base are plotted, and the SNP sequence is indicated by color. Only one allele is expressed in the parent primed UCLA1 line, but both alleles are expressed in 5iLAF cells. **(C)** Global DNA methylation in naiveand primed cells before and after sixteen days of culture in EpiLC-like conditions to restore the primed state. **(D)** DNA methylation over imprints in naiveand primed cells before and after sixteen days of culture in EpiLC-like conditions to restore the primed state. Each imprint is represented as a single point in the box plot. Note the modest increase in methylation at imprints as the naivecells are converted to primed conditions, whereas the global

increase in methylation is much greater. **(E)** Hypomethylation of young LINE elements including L1 human specific (L1HS) and its descendent L1PA2 in 5iLAF SSEA4 negative UCLA1 hESCs, as shown by metaplot. Note the dramatic loss of methylation in the vicinity of the element promoter. **(F)** DNA methylation over imprints for three alternate approaches for culturing naivecells. Because the Gafni 2013 data was generated by RRBS, only CGs that had coverage in the Gafni 2013 dataset are included in this analysis to make the data comparable. Only sixteen stable imprints had sufficient coverage for robust analysis. **(G)** Karyotyping results from reverted UCLA1 lines and new lines derived from blastocyst in 5iLAF. **(H)** Comparative Genomic Hybridization (CGH) data is shown over two chromosomes for the UCLA20n line cultured in 5iLAF. Most chromosomes showed normal karyotype (e.g chromosome 6, left), but several showed regions of elevated DNA content consistent with aneuploidy (e.g. chromosome 12, right).

Figure 4-1

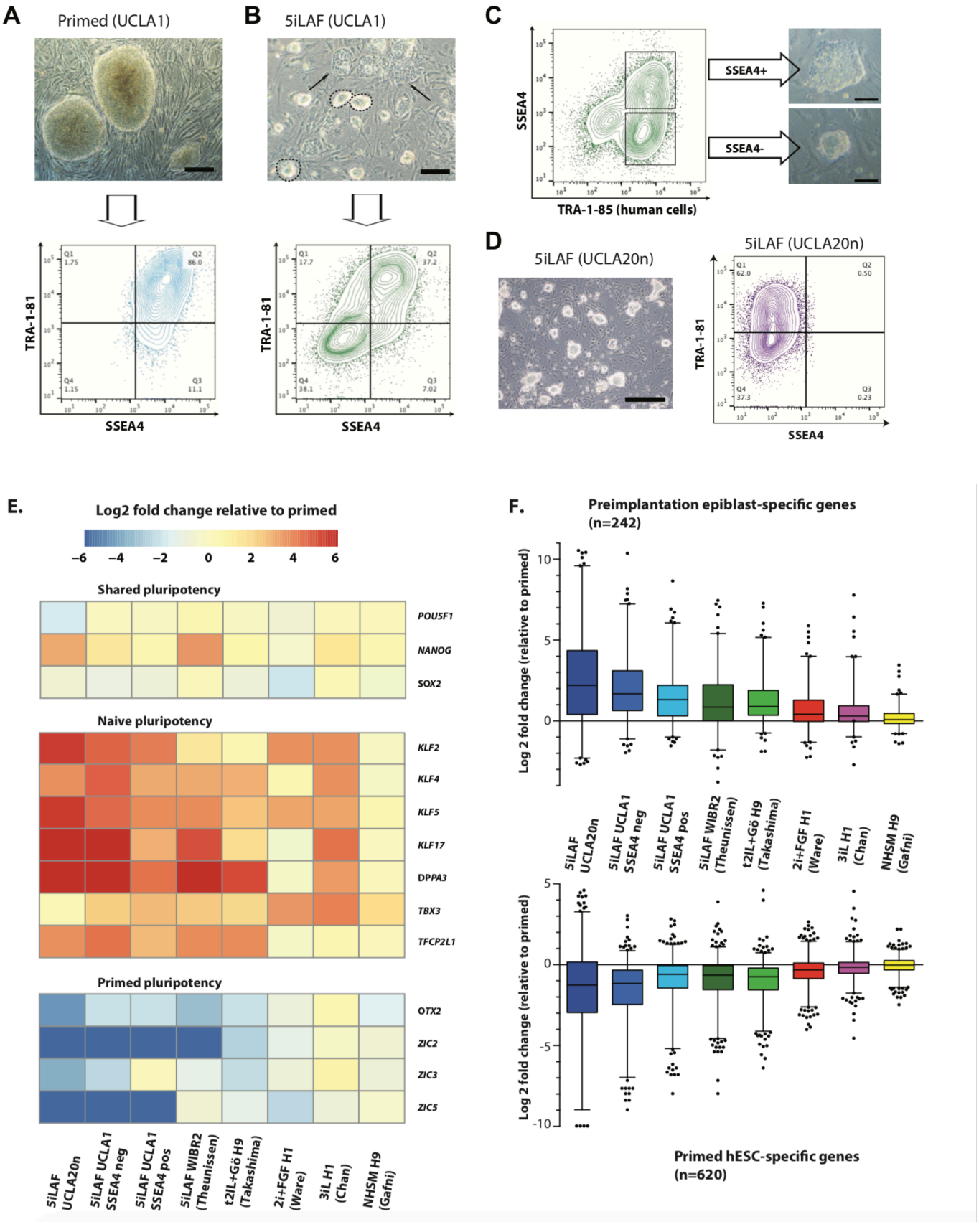

A
Primed (UCLA1)
B
5iLAF (UCLA1)
C
SSEA4
TRA-1-85 (human cells)
SSEA4+
SSEA4-
TRA-1-81
SSEA4
Q1 1.75
Q2 86.0
Q4 1.15
Q3 11.1
Q1 17.7
Q2 37.2
Q4 38.1
Q3 7.02
D
5iLAF (UCLA20n)
5iLAF (UCLA20n)
Q1 62.0
Q2 0.50
Q4 37.3
Q3 0.23
E.
Log2 fold change relative to primed
−6 −4 −2 0 2 4 6
Shared pluripotency
POU5F1
NANOG
SOX2
Naive pluripotency
KLF2
KLF4
KLF5
KLF17
DPPA3
TBX3
TFCP2L1
Primed pluripotency
OTX2
ZIC2
ZIC3
ZIC5
5iLAF UCLA20n
5iLAF UCLA1 SSEA4 neg
5iLAF UCLA1 SSEA4 pos
5iLAF WIBR2 (Theunissen)
t2iL+Gö H9 (Takashima)
2i+FGF H1 (Ware)
3iL H1 (Chan)
NHSM H9 (Gafni)
F.
Preimplantation epiblast-specific genes (n=242)
Log 2 fold change (relative to primed)
10
5
0
-5
-10
Primed hESC-specific genes (n=620)

Figure 4-2

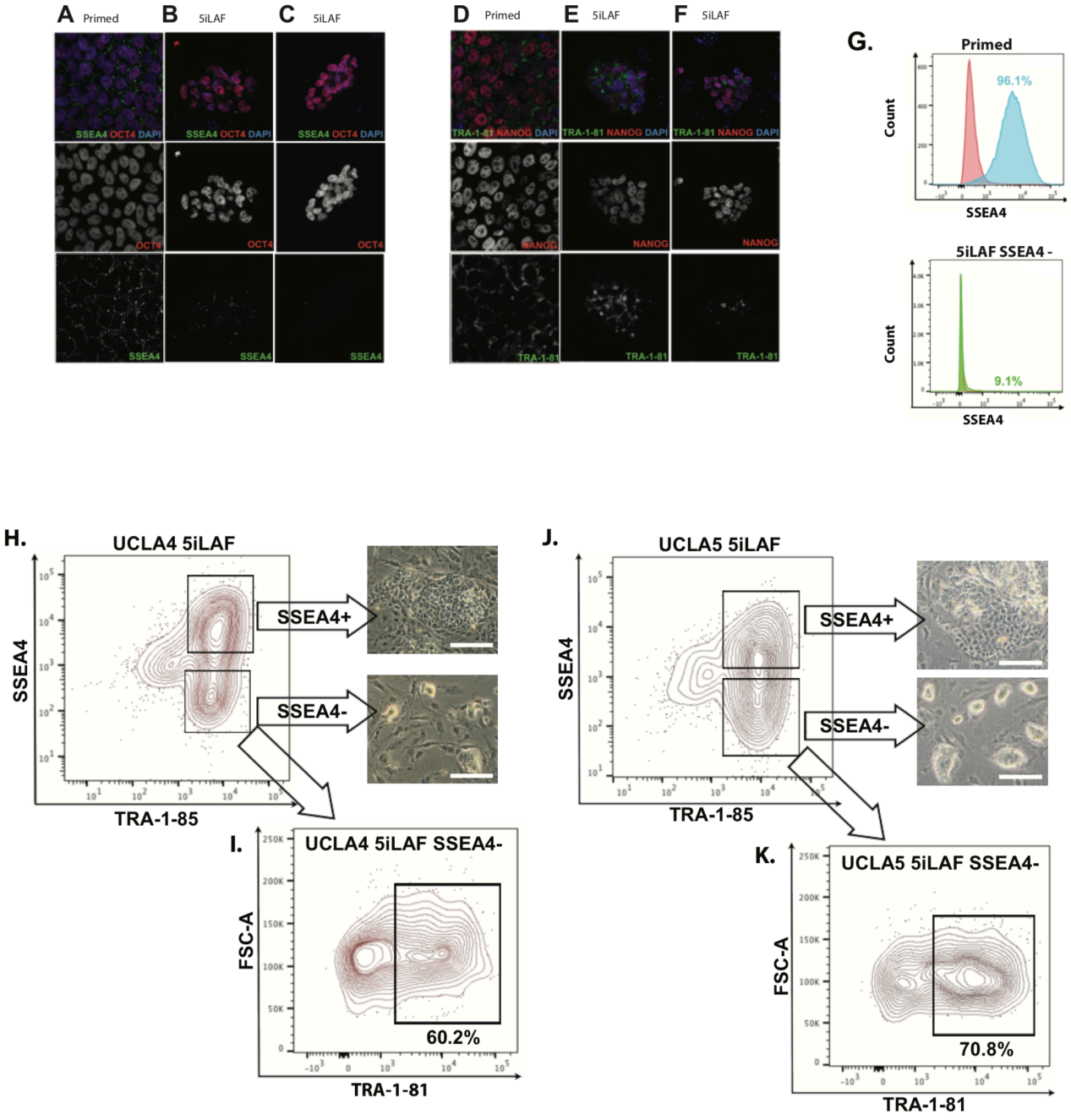

A Primed
B 5iLAF
C 5iLAF
D Primed
E 5iLAF
F 5iLAF
SSEA4 OCT4 DAPI
OCT4
SSEA4
TRA-1-81 NANOG DAPI
NANOG
TRA-1-81
G.
Primed
96.1%
Count
SSEA4
5iLAF SSEA4 -
9.1%
H.
UCLA4 5iLAF
SSEA4+
SSEA4-
SSEA4
TRA-1-85
I.
UCLA4 5iLAF SSEA4-
FSC-A
60.2%
TRA-1-81
J.
UCLA5 5iLAF
SSEA4+
SSEA4-
SSEA4
TRA-1-85
K.
UCLA5 5iLAF SSEA4-
FSC-A
70.8%
TRA-1-81

Figure 4-3

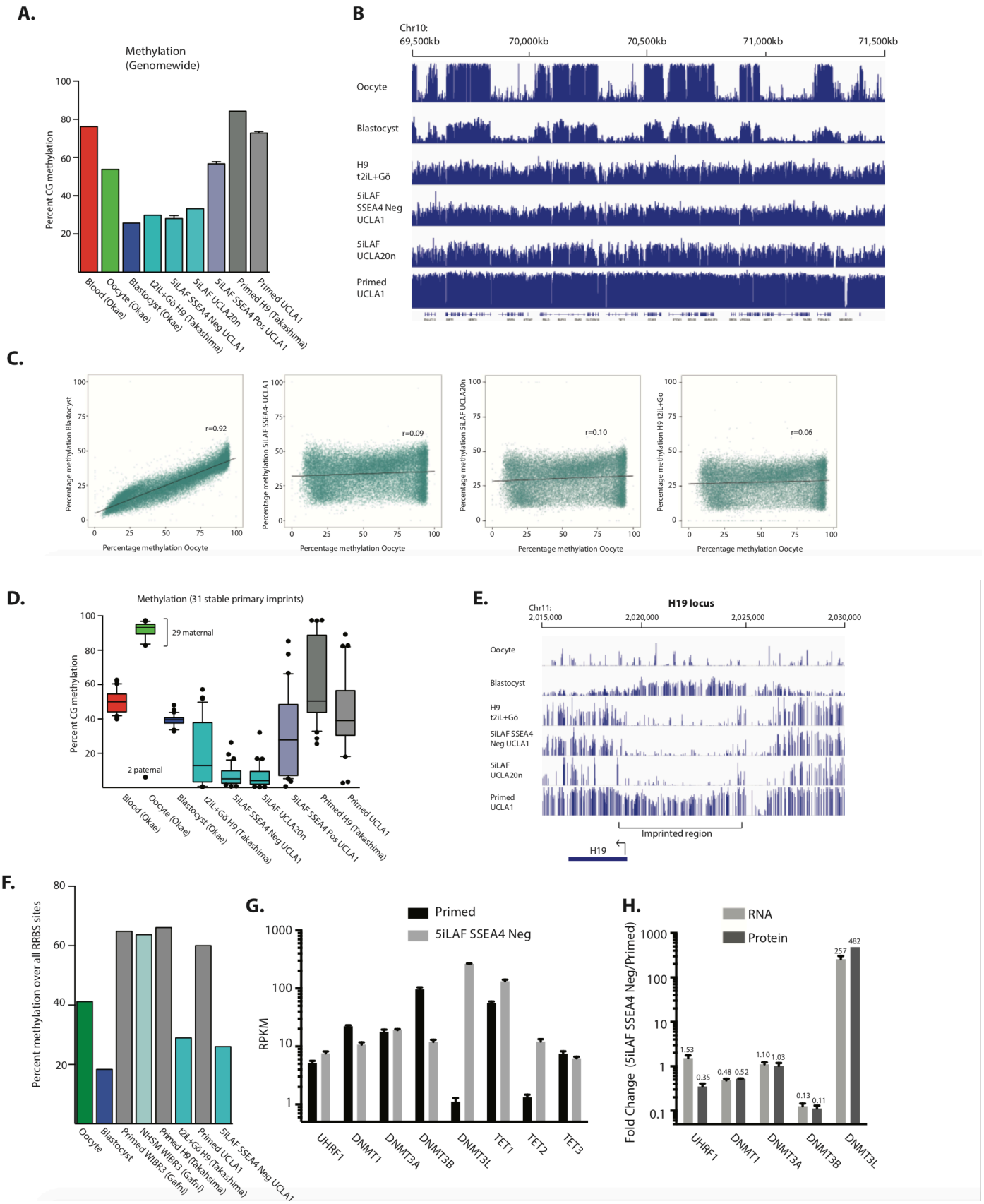
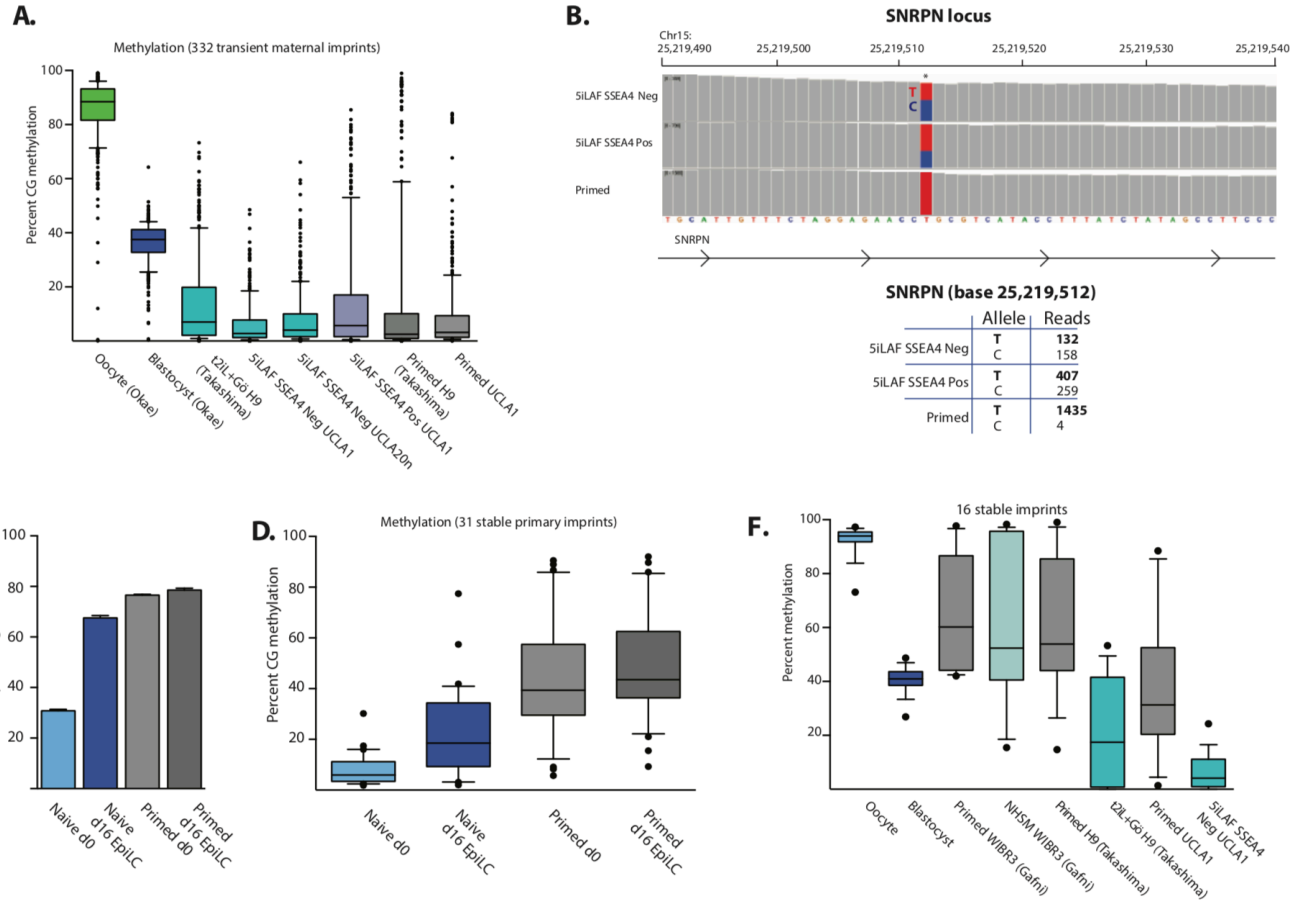
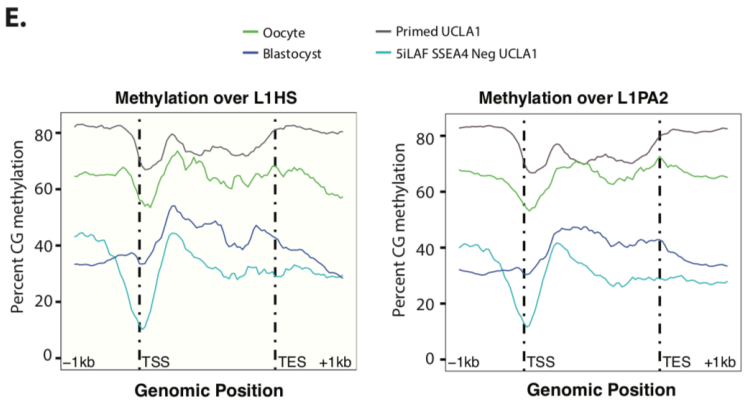
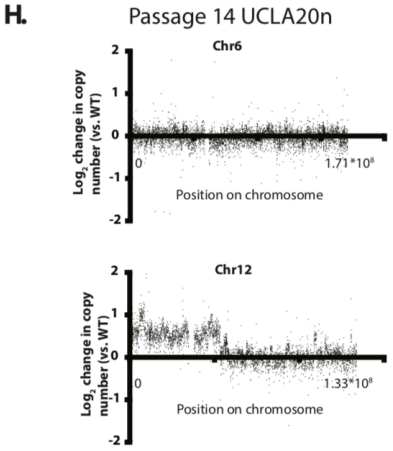

A.
Methylation
(Genomewide)
Percent CG methylation
Blood (Okae)
Oocyte (Okae)
Blastocyst (Okae)
t2iL+Gö H9 (Takashima)
5iLAF SSEA4 Neg UCLA1
5iLAF UCLA20n
5iLAF SSEA4 Pos UCLA1
Primed H9 (Takashima)
Primed UCLA1
B
Chr10:
69,500kb
70,000kb
70,500kb
71,000kb
71,500kb
Oocyte
Blastocyst
H9
t2iL+Gö
5iLAF
SSEA4 Neg
UCLA1
5iLAF
UCLA20n
Primed
UCLA1
C.
Percentage methylation Blastocyst
r=0.92
Percentage methylation 5iLAF SSEA4- UCLA1
r=0.09
Percentage methylation 5iLAF UCLA20n
r=0.10
Percentage methylation H9 t2iL+Go
r=0.06
Percentage methylation Oocyte
D.
Methylation (31 stable primary imprints)
Percent CG methylation
29 maternal
2 paternal
E.
H19 locus
Chr11:
2,015,000
2,020,000
2,025,000
2,030,000
Imprinted region
H19
F.
Percent methylation over all RRBS sites
Oocyte
Blastocyst
Primed WIBR3 (Gafni)
NHSM WIBR3 (Gafni)
Primed H9 (Takahsima)
t2iL+Go H9 (Takahsima)
Primed UCLA1
5iLAF SSEA4 Neg UCLA1
G.
Primed
5iLAF SSEA4 Neg
RPKM
UHRF1
DNMT1
DNMT3A
DNMT3B
DNMT3L
TET1
TET2
TET3
H.
RNA
Protein
Fold Change (5iLAF SSEA4 Neg/Primed)
1.53
0.35
0.48
0.52
1.10
1.03
0.13
0.11
257
482

Figure 4-4







G.

	Origin	Karyotype
UCLA1 5iLAF p13	Derived under primed conditions and reverted to naive in 5iLAF. Grown in 5iLAF 13 passages	G-band karyotyping. 18/20 cells 46,XX. 2/20 non-clonal abberations.
UCLA1 5iLAF p24	Derived under primed conditions and reverted to naive in 5iLAF. Grown in 5iLAF 24 passages	G-band karyotyping. 18/20 cells abnormal female karyotype with 13/20 = 46, XX,add(8)(q24.1). 30% polyploid based on n=70 metaphases
UCLA19n	Derived from blastocyst in 5iLAF. Grown in 5iLAF 15 passages.	G-band karyotyping. 70% polyploid.
UCLA20n	Derived from blastocyst in 5iLAF. Grown in 5iLAF 14 passages.	Array comparative genomic hybridization Trisomy over regions of Chr3, Chr7, 12 and 20.

REFERENCES

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Buecker, C., Srinivasan, R., Wu, Z., Calo, E., Acampora, D., Faial, T., Simeone, A., Tan, M., Swigut, T., and Wysocka, J. (2014). Reorganization of enhancer patterns in transition from naive to primed pluripotency. Cell Stem Cell *14*, 838–853.

Butler, M.G. (2009). Genomic imprinting disorders in humans: a mini-review. J. Assist. Reprod. Genet. *26*, 477–486.

Chan, Y.-S., Göke, J., Ng, J.-H., Lu, X., Gonzales, K.A.U., Tan, C.-P., Tng, W.-Q., Hong, Z.-Z., Lim, Y.-S., and Ng, H.-H. (2013). Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. Cell Stem Cell *13*, 663–675.

Diaz Perez, S.V., Kim, R., Li, Z., Marquez, V.E., Patel, S., Plath, K., and Clark, A.T. (2012). Derivation of new human embryonic stem cell lines reveals rapid epigenetic progression in vitro that can be prevented by chemical modification of chromatin. Hum. Mol. Genet. *21*, 751–764.

Gafni, O., Weinberger, L., Mansour, A.A., Manor, Y.S., Chomsky, E., Ben-Yosef, D., Kalma, Y., Viukov, S., Maza, I., Zviran, A., et al. (2013). Derivation of novel human ground state naive pluripotent stem cells. Nature *504*, 282–286.

Gkountela, S., Zhang, K.X., Shafiq, T.A., Liao, W.-W., Hargan-Calvopiña, J., Chen, P.-Y., and Clark, A.T. (2015). DNA Demethylation Dynamics in the Human Prenatal Germline. Cell *161*, 1425–1436.

Haaf, T. (1995). The effects of 5-azacytidine and 5-azadeoxycytidine on chromosome structure and function: implications for methylation-associated cellular processes. Pharmacol. Ther. *65*, 19–46.

Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S., and Saitou, M. (2011). Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. Cell *146*, 519–532.

Holm, T.M., Jackson-Grusby, L., Brambrink, T., Yamada, Y., Rideout, W.M., and Jaenisch, R. (2005). Global loss of imprinting leads to widespread tumorigenesis in adult mice. Cancer Cell *8*, 275–285.

Huang, K., Maruyama, T., and Fan, G. (2014). The naive state of human pluripotent stem cells: a synthesis of stem cell and preimplantation embryo transcriptome analyses. Cell Stem Cell *15*, 410–415.

Leitch, H.G., McEwen, K.R., Turp, A., Encheva, V., Carroll, T., Grabole, N., Mansfield, W., Nashun, B., Knezovich, J.G., Smith, A., et al. (2013). Naive pluripotency is associated with global DNA hypomethylation. Nat. Struct. Mol. Biol. *20*, 311–316.

Liao, J., Karnik, R., Gu, H., Ziller, M.J., Clement, K., Tsankov, A.M., Akopian, V., Gifford, C.A., Donaghey, J., Galonska, C., et al. (2015). Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. Nature Genetics *47*, 469–478.

Nichols, J., and Smith, A. (2009). Naive and primed pluripotent states. Cell Stem Cell *4*, 487–492.

Okae, H., Chiba, H., Hiura, H., Hamada, H., Sato, A., Utsunomiya, T., Kikuchi, H., Yoshida, H., Tanaka, A., Suyama, M., et al. (2014). Genome-wide analysis of DNA methylation dynamics during early human development. PLoS Genet. *10*, e1004868.

Oliveros-Etter, M., Li, Z., Nee, K., Hosohama, L., Hargan-Calvopiña, J., Lee, S.A., Joti, P., Yu, J., and Clark, A.T. (2015). PGC Reversion to Pluripotency Involves Erasure of DNA Methylation from Imprinting Control Centers followed by Locus-Specific Re-methylation. Stem Cell Reports *5*, 337–349.

Rugg-Gunn, P.J., Ferguson-Smith, A.C., and Pedersen, R.A. (2007). Status of genomic imprinting in human embryonic stem cells as revealed by a large cohort of independently derived and maintained lines. Hum. Mol. Genet. *16 Spec No. 2*, R243–R251.

Smith, Z.D., Chan, M.M., Humm, K.C., Karnik, R., Mekhoubad, S., Regev, A., Eggan, K., and Meissner, A. (2014). DNA methylation dynamics of the human preimplantation embryo. Nature *511*, 611–615.

Smith, Z.D., Chan, M.M., Mikkelsen, T.S., Gu, H., Gnirke, A., Regev, A., and Meissner, A. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature *484*, 339–344.

Tada, T., Tada, M., Hilton, K., Barton, S.C., Sado, T., Takagi, N., and Surani, M.A. (1998). Epigenotype switching of imprintable loci in embryonic germ cells. Dev. Genes Evol. *207*, 551–561.

Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W., et al. (2014). Resetting transcription factor control circuitry toward ground-state pluripotency in human. Cell *158*, 1254–1269.

Tang, F., Barbacioru, C., Nordman, E., Bao, S., Lee, C., Wang, X., Tuch, B.B., Heard, E., Lao, K., and Surani, M.A. (2011). Deterministic and stochastic allele specific gene expression in single mouse blastomeres. PLoS ONE *6*, e21208.

Tang, W.W.C., Kobayashi, T., Irie, N., Dietmann, S., and Surani, M.A. (2016). Specification and epigenetic programming of the human germ line. Nature Reviews Genetics *17*, 585–600.

Theunissen, T.W., Powell, B.E., Wang, H., Mitalipova, M., Faddah, D.A., Reddy, J., Fan, Z.P., Maetzel, D., Ganz, K., Shi, L., et al. (2014). Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. Cell Stem Cell *15*, 523.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Ware, C.B., Nelson, A.M., Mecham, B., Hesson, J., Zhou, W., Jonlin, E.C., Jimenez-Caliani, A.J., Deng, X., Cavanaugh, C., Cook, S., et al. (2014). Derivation of naive human embryonic stem cells. Proc. Natl. Acad. Sci. U.S.a. *111*, 4484–4489.

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics *10*, 232.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. *20*, 1131–1139.

Yang, S.-H., Kalkan, T., Morissroe, C., Marks, H., Stunnenberg, H., Smith, A., and Sharrocks, A.D. (2014). Otx2 and Oct4 drive early enhancer activation during embryonic stem cell transition from naive pluripotency. Cell Rep *7*, 1968–1981.

CHAPTER 5

Concluding Remarks

In this dissertation, I first described a novel technique modified from GRO-seq which we applied to study the characteristics of the nascent non-coding Pol V transcripts in RNA-directed DNA methylation (RdDM) pathway. We have identified unexpected co-transcriptional small RNA guided slicing of Pol V transcripts. With this modified GRO-seq protocol, it is possible to study and uncover features of non-coding RNAs in other organisms which may shed light on the mechanism and function of the 'dark matter' of genome. Also, we discovered that the co-transcriptional slicing of Pol V transcripts depends on SPT5L while *spt5l* mutant only show a partial loss of CHH methylation at RdDM sites. Future studies on the mysterious function of SPT5L and Pol V transcripts slicing in RdDM likely to shed additional light on the mechanisms of DNA methylation control.

In chapter 2 and 3, I described tools we developed for epigenome engineering including DNA methylation and demethylation targeting. We tethered an artificial zinc finger (ZF) to various proteins in RdDM as well as human demethylase TET1 and showed their capability to establish or erase DNA methylation at target loci. With this tool, we also identified thousands of off target sites. When we target Pol V associated protein (DMS3) to those thousands of off target sites, we showed more than 90% of successful recruitment of Pol V to those off target sites while only about 5% off target sites get hypermethylation. In addition, when we tried to co-target DMS3 (Pol V associated) and Pol IV to those off target sites, we found more than 20% of sites get methylated. This methylation targeting enhancement suggested in order to successful target DNA methylation in *Arabidopsis*, targeting combinatory effector proteins simultaneously might be a potential approach. With the advancement of CRISPR/dCas9 system, future studies on target DNA methylation with CRIPSR/dCas9 may also lead to a more specific epigenome engineering approach.

In chapter 4, I described the methylation landscape in human embryonic stem cells (hESCs). We profiled whole genome DNA methylation landscape in primed and naive hESCs and discovered that the hypomethylated naive hESCs lost the 'memory' of DNA methylation pattern over imprinting regions from oocyte. Since the loss of stable primary imprints is potentially serious in human pluripotent stem cell research, future studies in developing novel naive hESCs culture conditions maintaining epigenetic memory *in vivo* may be critical.