doi:10.1038/nature12178

Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1

Julie A. Law[1]*†, Jiamu Du[2]*, Christopher J. Hale[1]*, Suhua Feng[1,3,4], Krzysztof Krajewski[5], Ana Marie S. Palanca[6], Brian D. Strahl[5], Dinshaw J. Patel[2] & Steven E. Jacobsen[1,3,4]

DNA methylation is an epigenetic modification that has critical roles in gene silencing, development and genome integrity. In *Arabidopsis*, DNA methylation is established by DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) and targeted by 24-nucleotide small interfering RNAs (siRNAs) through a pathway termed RNA-directed DNA methylation (RdDM)[1]. This pathway requires two plant-specific RNA polymerases: Pol-IV, which functions to initiate siRNA biogenesis, and Pol-V, which functions to generate scaffold transcripts that recruit downstream RdDM factors[1,2]. To understand the mechanisms controlling Pol-IV targeting we investigated the function of SAWADEE HOMEODOMAIN HOMOLOG 1 (SHH1)[3,4], a Pol-IV-interacting protein[3]. Here we show that SHH1 acts upstream in the RdDM pathway to enable siRNA production from a large subset of the most active RdDM targets, and that SHH1 is required for Pol-IV occupancy at these same loci. We also show that the SHH1 SAWADEE domain is a novel chromatin-binding module that adopts a unique tandem Tudor-like fold and functions as a dual lysine reader, probing for both unmethylated K4 and methylated K9 modifications on the histone 3 (H3) tail. Finally, we show that key residues within both lysine-binding pockets of SHH1 are required *in vivo* to maintain siRNA and DNA methylation levels as well as Pol-IV occupancy at RdDM targets, demonstrating a central role for methylated H3K9 binding in SHH1 function and providing the first insights into the mechanism of Pol-IV targeting. Given the parallels between methylation systems in plants and mammals[1,5], a further understanding of this early targeting step may aid our ability to control the expression of endogenous and newly introduced genes, which has broad implications for agriculture and gene therapy.

SHH1 was recently identified as a Pol-IV-interacting protein and shown to affect *de novo* DNA methylation[3]. To investigate the role of SHH1 in the RdDM pathway genome-wide, we generated siRNA profiles in wild-type Col plants, *shh1* mutant plants, and several other RdDM mutants for comparison. In wild-type plants approximately 12,500 siRNA clusters were defined, representing 84.2% of all uniquely mapping 24-nucleotide siRNAs. Consistent with previous findings, 81.4% of these siRNAs were Pol-IV-dependent[6,7] (Fig. 1a; *pol-iv* and *pol-v* mutants correspond to mutations in the *nrpd1* and *nrpe1* subunits of these polymerases, respectively). Analysis of the siRNA clusters reduced in *shh1* mutants demonstrated that SHH1 is a major regulator of siRNA levels, affecting 44% of Pol-IV-dependent clusters (Fig. 1b and Supplementary Fig. 1a). These *shh1*-affected clusters represent the majority of all 24-nucleotide siRNAs, as well as a majority of clusters reduced in two downstream RdDM mutants (*drm2* and *pol-v*) (Fig. 1b and Supplementary Fig. 1a). The overlap of the reduced siRNA clusters in these mutants formed four main subclasses (termed *pol-iv* only, *shh1*, *shh1/drm2/pol-v*, and *drm2/pol-v*; Fig. 1b), which were used for subsequent analyses. Interestingly, the clusters that depend solely on Pol-IV were more enriched in pericentromeric heterochromatin than those that also depend on SHH1, DRM2 and Pol-V (Fig. 1c and Supplementary Fig. 1b, c), indicating that different mechanisms may be controlling siRNA production in the euchromatic arms versus pericentromeric heterochromatin.

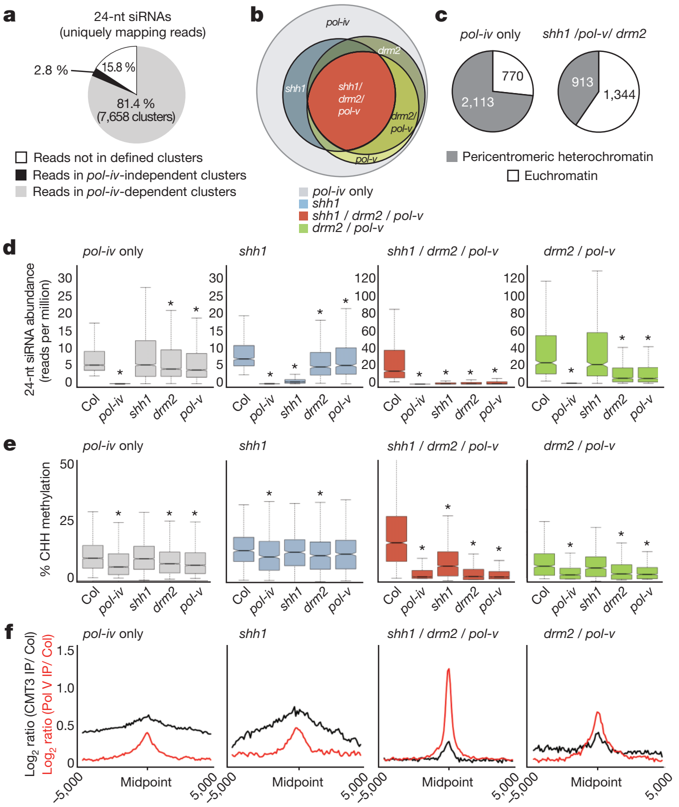

Figure 1 | Epigenetic profile of siRNA clusters affected in RdDM mutants. **a**, Pie chart showing the abundance of 24-nucleotide siRNA reads in wild-type (ecotype Col) sequencing libraries (5,967,213 uniquely mapping reads total). **b**, Schematic Venn diagram showing approximate relationships of 24-nucleotide siRNA clusters reduced in each genotype and the subclasses used for downstream analysis. **c**, Pie charts showing the chromosomal distribution (based on previously described definitions of pericentromeric heterochromatin and euchromatin[16]) of affected siRNA clusters in the indicated subclasses. **d**, **e**, Boxplots of siRNA and CHH methylation levels at the subclasses shown in **b** for various RdDM mutants (*indicates significant reduction; $P < 10^{-10}$ Mann–Whitney U test). **f**, Metaplots showing CMT3 and Pol-V enrichment at affected siRNA clusters (±5,000 base pairs (bp) from the siRNA cluster midpoint). IP, chromatin immunoprecipitation.

[1]Department of Molecular, Cell and Developmental Biology, University of California at Los Angeles, Los Angeles, California 90095, USA. [2]Structural Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [3]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, Los Angeles, California 90095, USA. [4]Howard Hughes Medical Institute, University of California at Los Angeles, Los Angeles, California 90095, USA. [5]Department of Biochemistry & Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [6]Plant Molecular and Cellular Biology, Salk Institute, La Jolla, California 92037, USA. †Present address: Plant Molecular and Cellular Biology, Salk Institute, La Jolla, California 92037, USA.
*These authors contributed equally to this work.

In *shh1* mutants, siRNA levels at SHH1-dependent clusters (*shh1* and *shh1/drm2/pol-v* subclasses) are reduced to nearly zero, whereas siRNA levels at SHH1-independent clusters experienced little to no change (Fig. 1d). These results demonstrate that SHH1 is a locus-specific RdDM component that has strong effects at a large subset of RdDM loci. Notably, the two downstream RdDM mutants (*drm2* and *pol-v*) have the strongest effect on siRNAs levels at clusters that also require SHH1 (*shh1/drm2/pol-v* subclass), and these same clusters are among the highest siRNA-producing clusters in the genome (Fig. 1d, e and Supplementary Fig. 1d, e). Together, these findings indicate that SHH1, and the downstream RdDM mutants, converge to control siRNA levels at the most active sites of RdDM.

Using whole-genome bisulphite sequencing (BS-seq), we assessed DNA methylation levels at the loci showing reduced siRNA levels and found that, consistent with its interaction with Pol-IV, SHH1 is an upstream RdDM component—*shh1* mutants only affect DNA methylation at sites where siRNA levels are reduced (Fig. 1e and Supplementary Fig. 2). Furthermore, the residual siRNAs present in *shh1* mutants seem to target some methylation (Supplementary Fig. 2b), as predicted for an upstream RdDM component. This is in contrast to the downstream mutants, *drm2* and *pol-v*, which reduced DNA methylation to nearly *pol-iv* levels even at sites that retain siRNAs (Fig. 1e), presumably due to an inability of these mutants to use siRNAs to target DNA methylation.

At loci corresponding to the *shh1/drm2/pol-v* and *drm2/pol-v* subclasses of siRNA clusters, the observed losses of siRNAs were accompanied with a correspondingly large loss of DNA methylation (Fig. 1e and Supplementary Fig. 2). However, at the *pol-iv only* and *shh1* subclasses, large losses of siRNAs were accompanied by relatively little DNA methylation loss. A likely explanation for this finding is that other DNA methylation pathways are active at sites corresponding to the *pol-iv only* and *shh1* siRNA clusters. In addition to the RdDM pathway, DNA methylation in *Arabidopsis* is controlled by two maintenance methyltransferase pathways[1]: the DNA METHYLTRANSFERASE 1 (MET1) pathway, which acts to maintain CG methylation, and the CHROMOMETHYLTRANSFERASE 3 (CMT3) pathway, which acts along with several H3K9 histone methyltransferases to maintain CHG and some CHH methylation[8]. Consistent with this explanation we found, using a previously published CMT3 chromatin immunoprecipitation coupled with high throughput sequencing (ChIP-seq) data set[9], that the *pol-iv* only and *shh1* subclasses of reduced siRNA clusters had the highest levels of CMT3 occupancy (Fig. 1f), indicating that CMT3 is able to maintain DNA methylation at nearly wild-type levels at these loci. In contrast, the *shh1/drm2/pol-v* and *drm2/pol-v* subclasses, which show marked DNA methylation losses in RdDM mutants, had lower levels of CMT3 enrichment (Fig. 1f) and are more highly and precisely enriched for the Pol-V polymerase[10] (Fig. 1f and Supplementary Fig. 2c), indicating that they are primarily targeted by the RdDM pathway.

To test the hypothesis that the siRNA losses observed in *shh1* mutants are due to a lack of Pol-IV targeting, we determined the genome-wide profile of Pol-IV occupancy in wild-type and *shh1* mutant backgrounds via ChIP-seq experiments using a Flag-tagged version of the largest Pol-IV subunit, NRPD1[3]. Consistent with our profile of Pol-IV-dependent siRNA clusters (Supplementary Fig. 1b), Pol-IV was broadly enriched at pericentromeric heterochromatin (Supplementary Fig. 3a) and at the defined subclasses of siRNA clusters (Fig. 2 and Supplementary Fig. 3b). In the *shh1* mutant background, Pol-IV levels were markedly reduced or eliminated specifically at *shh1*-dependent siRNA clusters (Fig. 2 and Supplementary Fig. 3c), further supporting the biological relevance of our ChIP-seq profile and confirming that the reduced-siRNA phenotype of *shh1* mutants is due to altered Pol-IV chromatin association. At *shh1*-independent siRNA clusters, Pol-IV levels, like siRNA levels, were not reduced in *shh1* mutants (Fig. 2 and Supplementary Fig. 3c), indicating that Pol-IV targeting to these loci requires an alternative mechanism.

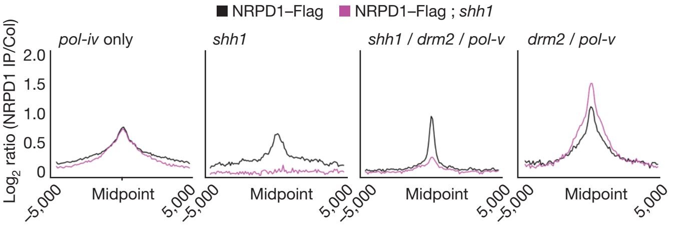

Figure 2 | Pol-IV levels at defined siRNA clusters. Metaplots of Pol-IV enrichment over the defined siRNA clusters in the indicated genetic backgrounds. Metaplots extend ±5,000 bp from the midpoint of the siRNA cluster.

In addition to assessing the levels of Pol-IV enrichment over the affected siRNA cluster subclasses, we also defined 928 reproducible, high confidence Pol-IV peaks using multiple ChIP-seq data sets. These peaks were enriched for siRNAs and DNA methylation (Supplementary Fig. 4a) and preferentially overlapped with the high siRNA-producing *shh1/drm2/pol-v* or *drm2/pol-v* clusters as compared to the *pol-iv* only and *shh1* clusters ($P < 2.2 \times 10^{-16}$, Fisher's exact test), indicating that the ChIP procedure is preferentially identifying sites where Pol-IV is most active. At the 928 defined Pol-IV peaks, we observed a variable level of SHH1 dependency and divided the peaks into three categories, SHH1-independent, SHH1-dependent and SHH1-enhanced (Supplementary Fig. 4b). In *shh1* mutants, DNA methylation and siRNA levels were reduced at the SHH1-dependent sites and, to a lesser extent, at sites defined as SHH1-independent (Supplementary Fig. 4c, d). However, siRNA and Pol-IV levels were increased at SHH1-enhanced sites in *shh1* mutants, indicating a redistribution of Pol-IV to these sites in *shh1* mutants (Supplementary Fig. 4b, c). Notably, these SHH1-enhanced sites are unique amongst the Pol-IV peaks as they have very low levels of Pol-V enrichment (Supplementary Fig. 4b), which could explain the correspondingly low levels of CHH methylation observed at these sites in wild-type plants (Supplementary Fig. 4d). Together with our analysis of SHH1-dependent siRNA clusters, these findings demonstrate that SHH1 plays a critical role in facilitating Pol-IV–chromatin association at a subset of the most active sites of RdDM.

To gain insight into the mechanism through which SHH1 facilitates Pol-IV targeting, we investigated the function of its previously uncharacterized SAWADEE domain[11]. Because there are precedents for cross talk between DNA methylation and histone modifications[1,12], we tested the ability of the SAWADEE domain to bind modified histone tails using an Active Motif-modified peptide array. This assay revealed that the SAWADEE domain has a preference for H3K9 methylation, but is also influenced by the methylation status of the H3K4 residue, with only unmodified or H3K4me1 modifications being tolerated (Supplementary Fig. 5a). To confirm these results, isothermal calorimetry (ITC) experiments were conducted using modified histone tail peptides (Fig. 3a, b and Supplementary Table 1). These analyses revealed that the SAWADEE domain is quite unique in its ability to bind all three H3K9 methylation states (me1, me2 and me3) with very similar affinity, dissociation constant (K_d) ≈ 2 μM, which is approximately 17-fold stronger than that observed using unmodified H3 peptides (Fig. 3a and Supplementary Table 1). ITC experiments also confirmed that although the SAWADEE domain will bind H3K9me2 peptides that contain H3K4me1 modifications, the presence of H3K4me2 or H3K4me3 modifications resulted in reduced binding affinity (Supplementary Table 1). Finally, ITC experiments using modified peptides corresponding to other known methylated lysine residues on the amino-terminal tails of the core histone proteins confirmed the specificity of the SHH1 SAWADEE domain for H3K9 methylation (Fig. 3b and Supplementary Table 1).

The anti-correlated effects of H3K9 and H3K4 methylation on SHH1 binding are reflective of genome profiling studies in *Arabidopsis* showing that the distribution of H3K9 methylation is anti-correlated with H3K4 methylation[13]. Consistent with these studies and the observed *in vitro*

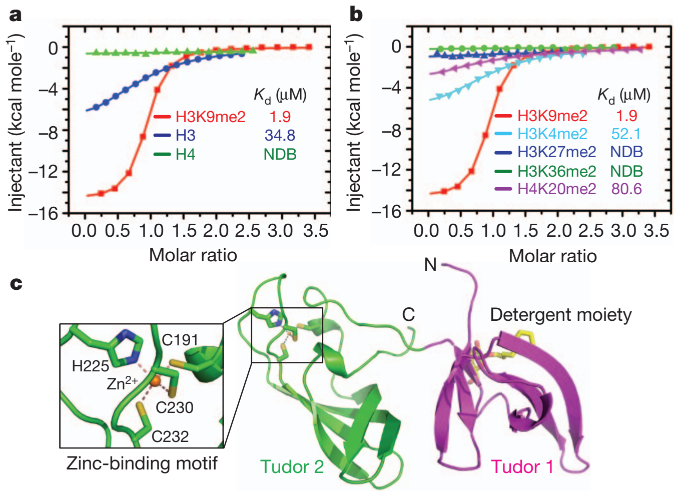

Figure 3 | The SHH1 SAWADEE domain recognizes H3K9 methylation and adopts a unique tandem Tudor domain-like fold. **a**, **b**, ITC-based measurements of the SAWADEE domain binding to the modified or unmodified histone peptides as indicated. K_d values are listed. NDB means no detectable binding. **c**, The overall structure of the SHH1 SAWADEE domain in the free form. The zinc-binding motif is shown as an enlarged ball-and-stick model, highlighting the details of the metal coordination. A bound detergent molecule, 4-cyclohexyl-1-butyl-β-D-maltoside moiety from the crystallization condition, is shown in a stick representation.

binding specificity of the SHH1 SAWADEE domain to H3K9 methylation, SHH1-dependent Pol-IV ChIP-seq peaks are enriched for H3K9me2 (Supplementary Fig. 5b) and depleted for H3K4 methylation (Supplementary Fig. 5c). Together, these binding studies demonstrate that the SAWADEE domain is a novel chromatin-binding module that probes both the K4 and K9 positions of the H3 tail and specifically binds repressive H3K9 methyl-modifications.

To determine the mode of methyl-lysine recognition by the SHH1 SAWADEE domain, crystal structures of this domain either in the free state or in complex with modified H3 tails were solved (Supplementary Tables 2 and 3, Fig. 3c and Supplementary Fig. 6a). In the free state, the SHH1 SAWADEE domain adopts a tandem Tudor domain-like fold that contains a unique zinc-binding motif located within the Tudor 2 subdomain (Fig. 3c). The overall structure of the SAWADEE domain resembles the UHRF1 tandem Tudor domain with an root mean squared deviation (r.m.s.d.) of 2.3 Å (Supplementary Fig. 6b) despite only sharing 11.8% sequence identity (Supplementary Fig. 7)[14,15]. This finding demonstrates that, although the sequence of the SAWADEE domain is plant-specific, its fold is highly conserved in eukaryotic organisms.

The structures of the SHH1 SAWADEE domain in complexes with H3K9me1, H3K9me2 and H3K9me3 peptides were also solved and all three peptides were bound in a similar manner (Supplementary Table 3). Given the known role of the H3K9me2 modification in gene silencing genome-wide in plants[16], we focused on the 2.70 Å structure solved with an H3(1–15)K9me2 peptide (Fig. 4a and Supplementary Fig. 8a). This peptide binds in a groove between the two Tudor subdomains, forming contacts with both subdomains (Fig. 4a, b and Supplementary Fig. 8b, c). Interestingly, there is no significant conformational change in the SAWADEE domain upon ligand binding (Supplementary Fig. 9a), which differs from the situation for UHRF1 (ref. 15).

Within the SHH1 SAWADEE domain, there are two pockets that form key intermolecular interactions with the unmodified K4 and the K9me2 side chains of the bound peptide (Fig. 4c, d). The unmodified H3K4 side chain inserts into an interfacial pocket formed by residues from both Tudor subdomains. In this pocket, the K4 side chain is stabilized via intermolecular hydrogen bonds and electrostatic interactions with the side chains of Glu 130 and Asp 141 (Fig. 4c). The H3K9me2 side chain inserts into a hydrophobic aromatic cage in the Tudor 1 subdomain (Fig. 4d) where it is stabilized by cation-π interactions in a manner similar to those reported previously for methylated lysine-binding modules[17]. The SAWADEE complexes with H3K9me3 and H3K9me1 peptides also position the methylated lysines within the same aromatic cage (Supplementary Fig. 10). The ability of the SAWADEE domain to bind equally against all three H3K9 methylation states can be well explained by structural observations: The methylated lysine recognition aromatic cage can accommodate both H3K9me2 and H3K9me3 side chains through common hydrophobic interactions, resulting in a lack of discrimination between these two methylation states. In the H3K9me1 complex, although the lower lysine methylation state has a decreased hydrophobic interaction with the aromatic cage, the side chain of His 169 undergoes a small but significant conformational change in order to hydrogen bond with the K9me1 ammonium proton, thereby contributing to the recovery of the binding affinity (Supplementary Fig. 10). This lack of specificity for the state of K9 methylation is in contrast with the higher level of methylation specificity observed for the tandem Tudor domain of UHRF1, which has a slightly wider aromatic cage binding pocket (Supplementary Fig. 9b). Thus our structural analysis indicates how very subtle changes in the tandem Tudor domain fold can result in a fine tuning of methyl-lysine specificity.

Consistent with our peptide-binding studies (Supplementary Table 1), we were also able to solve a structure of the SAWADEE domain in a complex with an H3(1–15)K4me1K9me1 peptide (Supplementary Table 3). Overall, this structure resembles the structure with the H3K9me2 peptide, with the K4me1 accommodated within the same K4 binding pocket. However, the methyl group forms a stabilizing hydrophobic interaction with Leu 201 in place of the hydrogen bond that is formed between the unmethylated K4 and the Glu 130 side chain (Fig. 4e). Because this K4 binding pocket is relatively closed and narrow, higher methylation states of K4 would probably introduce steric conflicts and/or disrupt all the hydrogen bonding interactions, explaining the observed decreases in binding affinity (Supplementary Table 1).

To test the biological significance of methyl-H3K9 binding activity observed for the SHH1 SAWADEE domain, we generated point mutations within the two lysine-binding pockets as well as the zinc-binding motif and tested their effect on DNA methylation, siRNA levels and Pol-IV recruitment *in vivo*. These point mutations were engineered into an SHH1–3×Myc–BLRP (biotin ligase recognition peptide) construct and transformed into an *shh1* mutant background. DNA methylation levels were assessed at a well-characterized locus, *MEA-ISR*, by Southern blotting (Supplementary Fig. 11a) and genome-wide by BS-seq experiments (Fig. 4f and Supplementary Fig. 11c–e). Addition of a wild-type SHH1–3×Myc–BLRP transgene restored DNA methylation, but constructs harbouring mutations within the H3K9 or the H3K4 pockets were unable to fully complement the methylation defect observed in the *shh1* mutant (Fig. 4f and Supplementary Fig. 11c–e) despite being expressed at levels comparable to the wild-type SHH1–3×Myc–BLRP protein (Supplementary Fig. 11a). In line with a canonical role for the zinc-binding motif in protein structure and/or stability, mutations in the zinc coordinating residues resulted in nearly undetectable levels of protein (Supplementary Fig. 11b) and thus were not characterized further.

Similar to the *shh1* null mutant, the DNA methylation defects in the SHH1 lysine binding pocket mutants were most pronounced in the *shh1/drm2/pol-v* subclass of affected siRNA clusters (Fig. 4f and Supplementary Fig. 11c–e). Consistent with their positions and predicted contributions to the binding affinity of the SHH1 SAWADEE domain, the F162AF165A and the D141A mutants show stronger DNA methylation defects (Fig. 4f). Assessment of siRNA levels in these lysine binding pocket mutants via siRNA-seq experiments revealed a similar pattern of defects (Fig. 4g and Supplementary Fig. 11f). Finally, to determine whether the observed losses of siRNAs and DNA methylation reflect a defect in Pol-IV activity at chromatin, Pol-IV ChIP experiments were conducted in the SAWADEE domain point mutant backgrounds. All

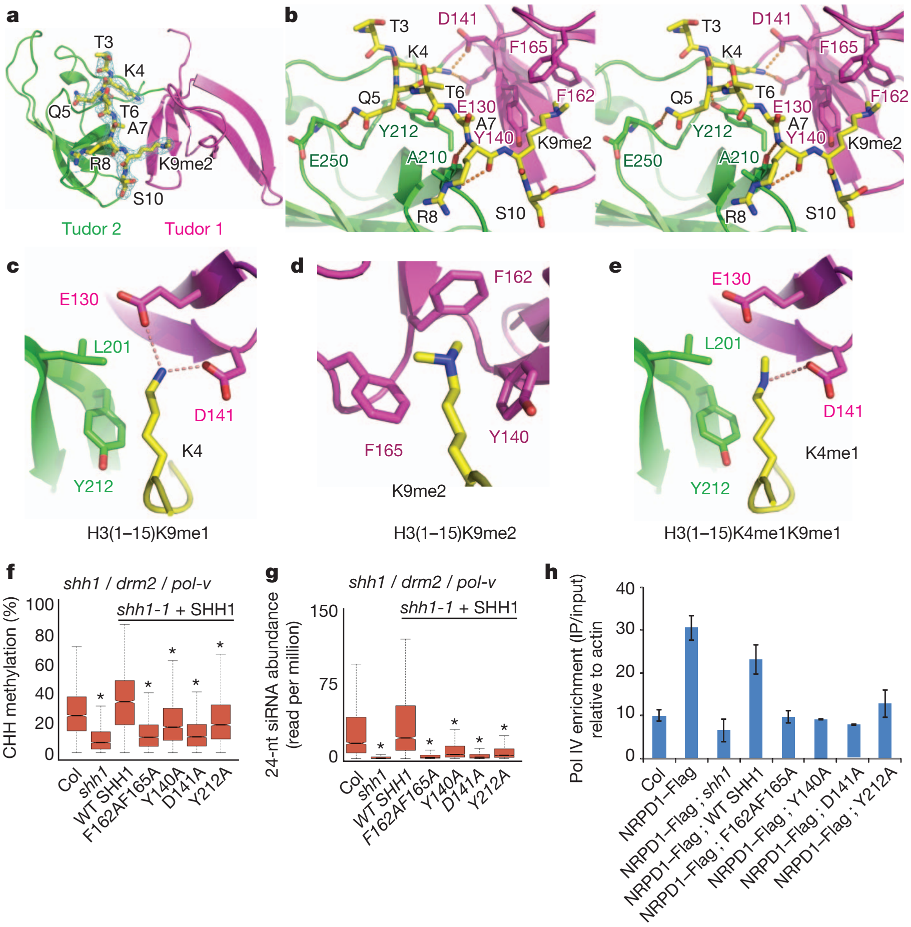

Figure 4 | Structural basis for recognition of H3(1–15)K9me2 peptide by the SHH1 SAWADEE domain and the functional impact of mutations of residues lining the K4 and K9me2 pockets. a, Overall structure of the H3(1–15)K9me2–SAWADEE complex with the SAWADEE domain as a ribbon diagram and the peptide as a stick representation . The simulated annealing composite omit map at 1σ level of the bound peptide is also shown. **b**, Stereo view highlighting the intermolecular interactions between the SAWADEE domain and the bound peptide. Intermolecular hydrogen-bonding interactions are designated by dashed red lines. **c**–**e**, Close-up views of H3 lysine residues (**c**, H3(1–15)K9me1; **d**, H3(1–15)K9me2; **e**, H3(1–15)K4me1K9me1) in their respective binding pockets. **f**, **g**, Boxplots of genome-wide percentage CHH methylation and siRNA levels in wild-type, *shh1* mutants and *shh1* mutants transformed with *SHH1* constructs (*shh1* + SHH1) that encode wild-type SHH1 or K9 (F162AF165A and Y140A) or K4 (D141A and Y212A) binding pocket mutants. **h**, qPCR of Pol-IV enrichment in the backgrounds described in **f** at a defined Pol-IV binding site. Bars are the average of two biological replicates normalized to input and actin levels (± standard error).

four point mutants displayed reduced levels of Pol-IV occupancy in two biological replicates (Fig. 4h). In addition, co-immunoprecipitation experiments revealed that the SAWADEE domain point mutants were still able to interact with Pol-IV (Supplementary Fig. 11g), demonstrating the interaction between SHH1 with the Pol-IV complex is not dependent on its H3K9me binding activity. Together, these findings show that residues within both the K4 and K9 binding pockets are critical for SHH1 function *in vivo* and demonstrate a central role for methyl-H3K9 binding by SHH1 at the level of Pol-IV association with chromatin.

The finding that the H3K4 binding pocket is critical for SHH1 function *in vivo* was unexpected considering that the SHH1 SAWADEE does not bind H3K4 methylation in the absence of H3K9 methylation, and that the addition of a methyl group to K4 does not impart additional binding affinity (Supplementary Table 1). One hypothesis to explain these *in vivo* findings is that the mere presence of a lysine at the position five residues back from the methylated H3K9 residue is necessary for SAWADEE domain binding. Indeed, such dual lysine reading could serve to help ensure that the SAWADEE domain only binds lysine methylation when it is present at the K9 position of the H3 tail as opposed to a methylated lysine at a different position on the H3 tail, especially the H3K27 position which has similar ARKS sequence context as H3K9 but a Thr 22 at five residues back. To test this hypothesis, ITC experiments were conducted using H3 tails harbouring an H3K4A mutation with or without the presence of the H3K9me2 modification. Indeed, the SAWADEE domain binds the H3K4AK9me2 peptide with approximately 30-fold weaker affinity than the H3K9me2 peptide (Supplementary Table 1). Furthermore, the SHH1 SAWADEE domain binds the H3K4A peptide with weaker affinity than the wild-type H3 tail (Supplementary Table 1), demonstrating that the K4 residue is contributing to binding independent of the methylation status of the K9 residue.

Together, these *in vivo* and *in vitro* analyses demonstrate that the SHH1 SAWADEE domain is probing the H3 tail at both the K4 and K9 positions and is quite selective for the combination of histone modifications present at transposons and other repetitive DNA elements, namely unmodified H3K4 and methylated H3K9. Although H3K9 methylation is anti-correlated with H3K4 methylation genome-wide[13], the aversion of the SAWADEE domain to higher order H3K4 methylation could serve to allow transcription, which is correlated with H3K4 methylation, to overcome DNA methylation and associated repressive H3K9 methyl modifications in a developmental or locus-specific manner. Likewise, the specificity of the SAWADEE domain could inhibit siRNA generation at body methylated genes which contain CG methylation and H3K4 methyl-modifications, but lack CHG and CHH methylation as well as siRNAs[13,18,19].

In summary, we demonstrate that SHH1 is a novel chromatin-binding protein that functions to enable Pol-IV recruitment and/or stability at the most actively targeted genomic loci to promote siRNA biogenesis. The finding that SHH1 binds to repressive histone modifications, together with the observation that SHH1 is required for Pol-IV chromatin association at a similar set of loci as downstream RdDM mutants, could explain the previously observed self-reinforcing loop in which downstream RdDM mutants are required for the production of full levels of siRNAs from a subset of genomic loci[20–23]. Indeed, it has been shown that downstream RdDM mutants can cause a reduction of both DNA methylation and H3K9 methylation at RdDM loci[24], suggesting that the loss of siRNAs in these mutants may be due to the associated loss of the appropriate chromatin marks necessary for SHH1 binding.

METHODS SUMMARY

Materials and methods for histone peptide array, ITC binding, crystallization, structure determination and analysis, plant lines, and genomic data analysis are described in detail in the Methods. Read statistics for all genomics analyses are listed in Supplementary Table 4, and the defined small RNA clusters and Pol-IV peaks are listed in Supplementary Tables 5 and 6, respectively.

Full Methods and any associated references are available in the online version of the paper.

Received 2 September 2012; accepted 12 April 2013.
Published online 1 May 2013.

1. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Rev. Genet.* **11,** 204–220 (2010).
2. Haag, J. R. & Pikaard, C. S. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nature Rev. Mol. Cell Biol.* **12,** 483–492 (2011).
3. Law, J. A., Vashisht, A. A., Wohlschlegel, J. A. & Jacobsen, S. E. SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. *PLoS Genet.* **7,** e1002195 (2011).
4. Liu, J. *et al.* An atypical component of RNA-directed DNA methylation machinery has both DNA methylation-dependent and -independent roles in locus-specific transcriptional gene silencing. *Cell Res.* **21,** 1691–1700 (2011).
5. Olovnikov, I., Aravin, A. A. & Fejes Toth, K. Small RNA in the nucleus: the RNA-chromatin ping-pong. *Curr. Opin. Genet. Dev.* **22,** 164–171 (2012).
6. Mosher, R. A., Schwach, F., Studholme, D. & Baulcombe, D. C. PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proc. Natl Acad. Sci. USA* **105,** 3145–3150 (2008).
7. Zhang, X., Henderson, I. R., Lu, C., Green, P. J. & Jacobsen, S. E. Role of RNA polymerase IV in plant small RNA metabolism. *Proc. Natl Acad. Sci. USA* **104,** 4536–4541 (2007).
8. Cao, X. *et al.* Role of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation. *Curr. Biol.* **13,** 2212–2217 (2003).
9. Du, J. *et al.* Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell* **151,** 167–180 (2012).
10. Zhong, X. *et al.* DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. *Nature Struct. Mol. Biol.* 19**,** 870–875 CrossRef (2012).
11. Mukherjee, K., Brocchieri, L. & Burglin, T. R. A comprehensive classification and evolutionary analysis of plant homeobox genes. *Mol. Biol. Evol.* **26,** 2775–2794 (2009).
12. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Rev. Genet.* **10,** 295–304 (2009).
13. Zhang, X., Bernatavichute, Y. V., Cokus, S., Pellegrini, M. & Jacobsen, S. E. Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis* thaliana. *Genome Biol.* **10,** R62 (2009).
14. Holm, L. & Rosenstrom, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38,** W545–W549 (2010).
15. Nady, N. *et al.* Recognition of multivalent histone states associated with heterochromatin by UHRF1 protein. *J. Biol. Chem.* **286,** 24300–24311 (2011).
16. Bernatavichute, Y. V., Zhang, X., Cokus, S., Pellegrini, M. & Jacobsen, S. E. Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in *Arabidopsis thaliana*. *PLoS ONE* **3,** e3156 (2008).
17. Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D. & Patel, D. J. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nature Struct. Mol. Biol.* **14,** 1025–1040 (2007).
18. Zhang, X. *et al.* Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis. Cell* **126,** 1189–1201 (2006).
19. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452,** 215–219 (2008).
20. Zilberman, D. *et al.* Role of *Arabidopsis* ARGONAUTE4 in RNA-directed DNA methylation triggered by inverted repeats. *Curr. Biol.* **14,** 1214–1220 (2004).
21. Xie, Z. *et al.* Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol.* **2,** e104 (2004).
22. Li, C. F. *et al.* An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in *Arabidopsis thaliana*. *Cell* **126,** 93–106 (2006).
23. Pontes, O. *et al.* The *Arabidopsis* chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell* **126,** 79–92 (2006).
24. Zilberman, D., Cao, X. & Jacobsen, S. E. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299,** 716–719 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to the staff of beamlines 24ID-C/E at the Argonne National Laboratory and X29A at the Brookhaven National Laboratory for support in diffraction data collection, to M. Akhavan and S. Fuchs for technical assistance, and to the UCLA BSCRC BioSequencing Core Facility for DNA sequencing. This work was supported by funds from the Abby Rockefeller Mauze Trust and Maloris and STARR Foundations to D.J.P. Work in the Jacobsen laboratory was supported by NIH grant GM60398. C.J.H. is supported by the Damon Runyon postdoctoral fellowship. S.F. is a Special Fellow of the Leukemia & Lymphoma Society. B.D.S. is supported by NIH grant GM85394. S.E.J. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions J.A.L., J.D., C.J.H., S.F. and A.M.S.P. conducted the experiments, K.K. synthesized modified peptides, B.D.S., D.J.P. and S.E.J. directed the research, and J.A.L., J.D., C.J. H., D.J.P. and S.E.J. wrote the manuscript.

Author Information Coordinates and structure factors have been deposited in the RCSB Protein Data Bank with the accession codes: 4IUP for the Se–SAWADEE (L200M, L218M) complex and 4IUQ for the wild-type SAWADEE domain in the free state, 4IUR for the H3(1–15)K9me3–SAWADEE complex, 4IUT for the H3(1–15)K9me2–SAWADEE complex, 4IUU for the H3(1–15)K9me1–SAWADEE complex, and 4IUV for the H3(1–15)K4me1K9me1–SAWADEE complex. The genomics data are submitted to GEO under the accession number GSE45368. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.E.J. (jacobsen@ucla.edu) or D.J.P. (pateld@mskcc.org).

METHODS

ChIP-seq, BS-seq and siRNA-seq library construction and sequencing. The first replicate of ChIP-seq libraries (NRPD1–Flag and Col) was generated using the Ovation Ultralow IL Multiplex System (NuGEN) whereas the second replicate (NRPD1–Flag, NRPD1–Flag ; *shh1*, and Col) was generated using the Ovation Ultralow DR Multiplex System (NuGEN). Both sets of ChIP-seq libraries used 18 cycles for the library amplification step. BS-seq libraries were generated using the pre-methylated adapter method as described previously[25]. siRNA-seq libraries were generated using the small RNA TruSeq kit (Illumina) following the manufacturer instructions with the exception that 15 cycles were used during the amplification step. The wild-type (Col) and *nrpe1* BS-seq libraries used in this study were previously published[10] and were subsequently reanalyzed for this study. All libraries were sequenced using the HiSeq 2000 platform following manufacturer instructions (Illumina) at a length of 50 bp. Read statistics are listed in Supplementary Table 4.

Mapping and processing of reads. Sequenced reads were base-called using the standard Illumina pipeline. For ChIP-seq and BS-seq libraries, only full 50 nucleotides reads were retained, whereas for siRNA-seq libraries, reads had adapter sequence trimmed and were retained if they were between 18 and 28 nucleotides in length. For ChIP-seq and siRNA-seq libraries, reads were mapped to the *Arabidopsis* genome (TAIR8, http://www.arabidopsis.org) with Bowtie[26] and only perfect matches that mapped uniquely to the genome were retained for further analysis although the total number of mapping reads, unique and non-unique, were used when normalizing the siRNA-seq libraries to total number of reads per library. For BS-seq libraries, reads were mapped using the BSseeker wrapper for Bowtie[27]. For ChIP-seq and BS-seq, identical reads were collapsed into one read, whereas for siRNA-seq identical reads were retained. For methylation analysis, percent methylation was calculated as previously reported[19] with the unmethylated chloroplast genome serving as the measure of non-bisulphite converted background methylation. For the second replicate of ChIP-seq, there was a large disparity of resultant reads for the NRPD1–Flag and NRPD1–Flag ; *shh1* libraries, so the NRPD1–Flag and Col libraries were sampled down to match the read total of the smaller library (the NRPD1–Flag ; *shh1* library).

DNA methylation analysis. For assessment of DNA methylation at siRNA clusters, only those clusters with at least one cytosine in the respective class being assayed (CG, CHG or CHH), were considered. For calculating significance levels of methylation change via the Mann–Whitney *U* test of methylation levels for clusters within the different subclasses (Fig. 1e) the number of clusters within each subclass was down sampled to the smallest subclass (the *drm2/nrpe1* subclass) to allow for comparable significance values between subclasses.

Identification of siRNA clusters. Small RNA clusters (Supplementary Table 5) in the *Arabidopsis* genome were defined in a manner similar to a previously published approach[28]. In brief, the genome was divided into 200-bp bins, and the average coverage per bin of non-identical siRNA reads was calculated in two technical replicates of our wild-type (Col) library. This average was used to assay the significance of the number of non-identical reads at a given bin in wild-type plants, assuming a Poisson distribution of such counts. In the R environment a Poisson exact test was carried out for each bin, and bins with a *P*-value less than 10^{-5} in each wild-type technical replicate were considered as clusters.

Once clusters were defined, comparisons between read counts, including identical reads, were carried out for each mutant and the wild-type (Col) library using a Fisher's exact test. Resultant *P*-values were Benjamini–Hochberg adjusted to estimate false discovery rates (FDRs), and clusters reduced in a mutant background at a $\text{FDR} < 10^{-10}$ were then considered to be dependent on the wild-type function of the mutant protein (Supplementary Table 5). For boxplot analysis of siRNA levels, the first technical replicate of the Col library was used as representative of Col siRNA levels. For calculating significance levels of siRNA change via the Mann–Whitney *U* test of siRNA levels for clusters within the different genotypic subclasses (Fig. 1d) the number of clusters within each subclass was down sampled to the smallest subclass (the *drm2/nrpe1* subclass) to allow for comparable significance values between subclasses.

Identification of NRPD1 peaks. The R package BayesPeak[29,30] was used to identify regions of Pol-IV enrichment in a NRPD1–Flag ChIP-seq library as compared to a paired Col ChIP-seq control library done in parallel. Only high scoring peaks ($\text{PP} > 0.999$) identified in both NRPD1–Flag ChIP-seq replicates (928 peaks) were retained for further analysis (Supplementary Table 6). For the purposes of assaying overlap of Pol-IV peaks with siRNA clusters, 'overlap' is called when more than 1 bp of a peak overlaps with a locus.

To classify peaks as SHH1-dependent, -independent or -enhanced, read counts over Pol-IV peaks were compared between the NRPD1–Flag and NRPD1–Flag; *shh1* ChIP-seq libraries, and significance was assessed using Fisher's exact test. Resultant P values were Benjamini–Hochberg adjusted to estimate FDRs. Peaks with a loss of NRPD1 signal in the *shh1* library at a $\text{FDR} < 0.001$ were considered SHH1-dependent. Similarly, peaks that gained signal in *shh1* at a $\text{FDR} < 0.001$ were considered SHH1-enhanced. Peaks that fell into neither of these categories were considered SHH1-independent.

Protein preparation. The gene encoding the SAWADEE domain of *Arabidopsis thaliana* SHH1 (residues 125–258) was cloned into a self-modified vector, which fuses a hexa-histidine tag plus a yeast sumo tag onto the N terminus of the target gene. The plasmid was transformed into the *Escherichia coli* strain BL21 (DE3) RIL (Stratagene). The cells were grown at 37 °C until the $D_{600\,\text{nm}}$ reached 0.8 and then the media was cooled to 20 °C and 0.2 mM IPTG was added to induce protein expression overnight. The recombinant expressed protein was first purified using a HisTrap FF column (GE Healthcare). The hexa-histidine-sumo tag was cleaved by the Ulp1 protease and removed by passing through a second HisTrap FF column. The pooled target protein was further purified using a Q FastFlow column and a Hiload Superdex G200 16/60 column (GE Healthcare) with buffer (150 mM NaCl, 20 mM Tris pH 8.0, and 5 mM dithiothreitol (DTT)). To prepare the Se-methionine-substituted protein, Leu 200 and Leu 218 of the SAWADEE domain were mutated to methionine using a QuikChange Site Directed Mutagenesis kit (Stratagene). The Se-methionine-substituted SAWADEE protein was expressed in M9 medium supplemented with amino acids Lys, Thr, Phe, Leu, Ile, Val and Se-Met, and purified using the same protocol as the wild-type protein. Peptides were synthesized by the Tufts University peptide synthesis facility or by K. Krajewski.

Crystallization. Crystallization of the SAWADEE domain was conducted at 4 °C using the sitting drop vapour diffusion method by mixing 1 μl of protein sample at a concentration of 5 mg ml^{-1} and 1 μl of reservoir solution (0.2 M NH_4F and 20% PEG 3350), which was equilibrated against a 0.4 ml reservoir. 4-cyclohexyl-1-butyl-β-D-maltoside (CYMAL-4, Hampton Research) was added in the drop with a final concentration of 7.6 mM as an additive, which resulted in considerable improvement in crystal quality. Thin plate-shaped crystals appeared within 2 days. To generate crystals of complexes of SAWADEE domain with modified H3 peptides (H3(1–15)K9me3, H3(1–15)K9me2, H3(1–15)K9me1 and H3(1–15)K4me1K9me1), the SAWADEE domain was mixed with peptides at a molar ratio of 1:2 at 4 °C for 1 h. The crystals of the different complexes were grown under the same conditions as described for free SAWADEE protein. All the crystals were soaked into a reservoir solution supplemented with 20% glycerol for 2 min. The crystals were then mounted on a nylon loop for diffraction data collection. The diffraction data from the native SAWADEE protein and its Se-methionine-substituted counterpart were collected at the NE-CAT beamline 24ID-C, Advanced Photon Source (APS), Argonne National Laboratory, Chicago, USA, at the zinc peak and selenium peak, respectively. The data of the complex of the H3K9me3 peptide bound to the SAWADEE domain were collected at beamline X29A, National Synchrotron Light Source (NSLS) at Brookhaven National Laboratory, New York, USA. The data on the SAWADEE domain in complex with H3K9me2, H3K9me1 and H3K4me1K9me1 peptides were collected at APS 24ID-E. All the crystallographic data were processed with the HKL2000 program[31]. The statistics of the diffraction data are summarized in Supplementary Tables 2 and 3.

Structure determination and refinement. The structure of the selenomethionine-substituted SAWADEE domain was solved using the single-wavelength anomalous dispersion (SAD) method as implemented in the Phenix program[32]. The model building was carried out using the Coot program[33] and structural refinement using the Phenix program[32]. The structure of the wild type SAWADEE domain in the free state was solved using the molecular replacement method using the Phenix program[32]. Zn^{2+} ions were identified and further confirmed by anomalous signal scattering. All the structures of SAWADEE domain in complexes with different modified H3 peptides were solved using the molecular replacement method with the same protocol as the native protein. The peptides showed clear electron density and were properly built with residues from Thr 3 to Ser 10 for H3(1–15)K9me3/2/1 and from Thr 3 to Thr 11 for H3(1–15)K4me1K9me1. Throughout the refinement, a free *R* factor was calculated using 5% random chosen reflections. The stereochemistry of the structural models were analysed using the Procheck program[34]. The refinement and structure statistics are shown in Supplementary Tables 2 and 3. All the molecular graphics were generated with the Pymol program (DeLano Scientific LLC).

Isothermal titration calorimetry. The protein samples were not stable at room temperature. Thus, all the binding experiments were performed on a Microcal calorimeter ITC 200 instrument at 6 °C. First, protein samples were dialysed overnight against a buffer of 100 mM NaCl, 2 mM β-mercaptoethanol and 20 mM HEPES, pH 7.5, at 4 °C. Then the protein samples were diluted and the lyophilized peptides were dissolved with the same buffer. The titration was performed according to standard protocol and the data were fit using the Origin 7.0 program with a 1:1 binding model. Thermodynamic parameters for complex formation are listed in Supplementary Table 1.

Modified peptide array binding. A glutathione *S*-transferase-conjugated SAWADEE domain (GST–SHH1, amino acids 125–258) construct was generated in the pENTR/

TEV/D plasmid (Invitrogen), recombined into the pDEST 15 plasmid (Invitrogen) and transformed into the Rosetta 2 (DE3) bacterial cell line (Novagen). Protein expression was induced by the addition of 500 µl of 1 M IPTG per 500 ml at $D_{600\,\mathrm{nm}}$ of 0.6 and cultures were grown at 16 °C overnight. At the time of induction the media was supplemented with 500 µl of 500 mM $ZnSO_4$. The GST fusion protein was then purified as described in ref. 34 and dialysed into storage buffer (50 mM Tris, pH 6.8, 300 mM NaCl, 40% glycerol, 2 mM DTT, 0.1% Triton X-100). The purified GST–SHH1 (125–258) protein was used to probe a MODified Histone Peptide Array (Active Motif) under the following conditions: The array was blocked at 25 °C for 45 min in a 5% milk 1× TBS solution, washed three times in a 1× TBS-T solution at 25 °C for 5 min, and then probed overnight at 4 °C with the GST–SHH1 SAWADEE domain protein at a concentration of 6.5 µg ml^{-1} in binding buffer (50 mM HEPES, pH 7.5, 50 mM NaCl, 5% glycerol, 0.4 mg ml^{-1} BSA, 2 mM DTT). The array was then washed three times as above, and probed an HRP conjugated GST antibody at a 1:5,000 dilution at 25 °C for 1 h. The array then washed as detailed above and developed using an ECL Plus kit (GE Healthcare).

Plant lines, site-directed mutagenesis, southern and western blotting. The various previously characterized *Arabidopsis* RdDM mutant alleles, the complementing SHH1–3×Myc–BLRP transgenic plant line, and the *pSHH1::SHH1–3×Myc–BLRP* construct used are as described in ref. 3. The *pol-iv* and *pol-v* mutants correspond to mutations in the *nrpd1* and *nrpe1* subunits of these polymerases, respectively. The structure-based mutations were generated in the *pSHH1::SHH1–3×Myc–BLRP* construct using a QuikChange Site Directed Mutagenesis kit (Stratagene) and were transformed into the *shh1-1* mutant background via the floral dip method. siRNA-seq and ChIP-seq experiments in the Col and RdDM mutant lines were conducted using floral tissue and BS-seq experiments were conducted using 10-day-old seedlings. Southern and western blotting experiments were conducted using tissue from the same individual plant lines in the T_1 generation and using previously described probes[35] and antibodies[36]. The siRNA-seq and BS-seq experiments in the SAWADEE domain point mutant lines were conducted using floral tissue or 10-day-old seedlings, respectively, from T_3 plants homozygous for the various *pSHH1::SHH1–3×Myc–BLRP* transgenes. The Pol-IV ChIP experiments and co-immunoprecipitation experiments in the various SAWADEE domain point mutant backgrounds were conducted using floral tissue from F_1 plants that were homozygous for the *shh1* mutant allele.

25. Feng, S., Rubbi, L., Jacobsen, S. E. & Pellegrini, M. Determining DNA methylation profiles using sequencing. *Methods Mol. Biol.* **733,** 223–238 (2011).
26. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
27. Chen, P. Y., Cokus, S. J. & Pellegrini, M. B. S. Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* **11,** 203 (2010).
28. Heisel, S. E. *et al.* Characterization of unique small RNA populations from rice grain. *PLoS ONE* **3,** e2871 (2008).
29. Spyrou, C., Stark, R., Lynch, A. G. & Tavare, S. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformatics* **10,** 299 (2009).
30. Cairns, J. *et al.* BayesPeak–an R package for analysing ChIP-seq data. *Bioinformatics* **27,** 713–714 (2011).
31. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
32. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
33. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
34. Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. Procheck: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26,** 283–291 (1993).
35. Johnson, L. M., Law, J. A., Khattar, A., Henderson, I. R. & Jacobsen, S. E. SRA-domain proteins required for DRM2-mediated de novo DNA methylation. *PLoS Genet.* **4,** e1000280 (2008).
36. Law, J. A. *et al.* A protein complex required for polymerase V transcripts and RNA- directed DNA methylation in *Arabidopsis*. *Curr. Biol.* **20,** 951–956 (2010).