Abstract

Modification of the histone proteins that form the core around which chromosomal DNA is looped has profound epigenetic effects on the accessibility of the associated DNA for transcription, replication and repair. The SET domain is now recognized as generally having methyltransferase activity targeted to specific lysine residues of histone H3 or H4. There is considerable sequence conservation within the SET domain and within its flanking regions. Previous reviews have shown that SET proteins from Arabidopsis and maize fall into five classes according to their sequence and domain architectures. These classes generally reflect specificity for a particular substrate. SET proteins from rice were found to fall into similar groupings, strengthening the merit of the approach taken. Two additional classes, VI and VII, were established that include proteins with truncated/interrupted SET domains. Diverse mechanisms are involved in shaping the function and regulation of SET proteins. These include protein–protein interactions through both intra- and inter-molecular associations that are important in plant developmental processes, such as flowering time control and embryogenesis. Alternative splicing that can result in the generation of two to several different transcript isoforms is now known to be widespread. An exciting and tantalizing question is whether, or how, this alternative splicing affects gene function. For example, it is conceivable that one isoform may debilitate methyltransferase function whereas the other may enhance it, providing an opportunity for differential regulation. The review concludes with the speculation that modulation of SET protein function is mediated by antisense or sense–antisense RNA.

Keywords: Arabidopsis SET genes; Alternative splicing; Epigenetics; Histone methylation; Rice SET genes; SET domain classes

1. Introduction

In eukaryotes, chromosomal DNA is organized as chromatin, in which ~146-bp form almost 2 left-handed coils around an octamer of basic proteins comprised of a histone H3–H4 tetramer and two H2A–H2B dimers [1–3]. These fundamental units, nucleosomes, are linked together by ~50 bp of spacer DNA that is associated with histone H1 to yield a characteristic ‘beads-on-a-string’ structure that folds further to yield a very compact state in the nucleus [4,5]. Recently, it has become evident that fast, long range, reversible conformational fluctuations in nucleosomal structure, together with specific chemical modifications of the histones, play vital roles in eukaryotic gene regulation [6–8]. Covalent modifications of the N-terminal histone tails include acetylation, phosphorylation, methylation, ubiquitination and ADP-ribosylation [9–11]. These modifications form the basis of the histone code (or, possibly, codes) that regulate gene expression epigenetically through various mechanisms [9,12–14]. For example, methylation of histone H3K9 provides an epigenetic mark for the binding of the chromodomain (protein domain structure that binds methylated lysine) containing protein, HP1 (heterochromatin protein 1), that leads to heterochromatin formation and gene repression [15]. In contrast, acetylation of histones tends to decrease interaction between histones and DNA, and facilitates binding of bromodomain (protein domain structure that binds acetylated lysine) containing proteins, thereby promoting an open chromatin conformation suitable for transcriptional activation [16–18].
Histone methylation is linked to multiple developmental processes including heterochromatin formation, cell cycle regulation, transcriptional silencing and transcriptional activation [19–25]. At least six lysine residues on histone H3 (K4, K9, K27, K36, K79) and H4 (K20) are targeted by histone lysine methyltransferases (HKMTs) [13,24]. Except for H3K79 [26], any other lysine residues (K27, K36, K79) and H4 (K20) are targeted by histone lysine methyltransferases. Adding to the complexity of histone code, each lysine residue can be mono-, di- or trimethylated [27,28]. Moreover, unlike histone lysine acetylation which is generally associated with gene activation [29,30], histone methylation at specific lysine residues can lead to either gene activation or repression [10].

Insight into the nature and regulation of enzymes responsible for modifications of specific amino acid residues in the nucleosomal core histones is essential towards deciphering the histone code. Whilst mammalian SUV39H1 was the first enzyme to be shown to possess HKMT activity towards H3K9 [19], its homolog, Su(var)3–9, in Drosophila was the first to be identified in a genetic screen for a suppressor of position effect variegation (PEV) [31]. The phenomenon of PEV was discovered by H. J. Muller in 1930 when describing the rearrangement of the white color eye gene from euchromatin to heterochromatin in the chromosomal regions. These conformational changes result in silencing and cell-to-cell variation of gene expression that lead to the mosaic eye color phenotype in Drosophila. Subsequent identification of HP1, that binds methylated H3K9, provided a bridge connecting H3K9 methylation to heterochromatin formation and hence gene repression [15,32,33]. In Drosophila, transcriptional repression or activation of homeotic gene expression is mediated by two distinct and antagonistic groups of protein complexes. Repression of homeotic gene expression is mediated by a Polycomb group (PcG) protein complex, containing Enhancer of zeste [E(Z)] and extra sex combs (ESC), that targets H3K27 and H3K9 methylation [34,35]. In contrast, the trithorax group (trxG) proteins, ASH1 [36] and Trx [37], function in targeting H3K4 methylation [38–40] and hence in maintaining an active state of homeotic genes.

Identification of the SET domain followed from the observation [41] that the Drosophila E(Z) PcG protein includes a region with high sequence similarity to two previously identified trxG proteins: Drosophila Trx [37] and human acute lymphoblastic leukemia 1 (ALL-1) [42,43]. The presence of this conserved region in these two groups of proteins with opposing functions (maintaining gene repression by PcG and activation by trxG proteins) led to the proposition [41] that it may comprise a domain that interacts with common nucleic acid or protein targets for gene regulation. Later, Su(var)3–9 was found to contain the C-terminal domain that is also shared by E(Z) and Trx [31]. The SET domain is named after these three founding Drosophila proteins: Suppressor of variegation 3–9 [Su(var)3–9], Enhancer of zeste [E(Z)], and Trithorax (Trx) [31]. The discovery of the evolutionarily conserved SET domain, present in the above-mentioned HKMTs, was an exciting step towards a comprehensive understanding of epigenetic regulation of gene expression through histone methylation. Here, we summarize current knowledge concerning plant SET proteins and compare the known functions of orthologous SET proteins of plants and other organisms. In addition to the existing classification of SET proteins in Arabidopsis and maize [22,44], we have included 9 additional Arabidopsis SET proteins and 34 rice SET proteins from Plant Chromatin Database (ChromDB; http://www.chromdb.org). We also comment on recent evidence that alternative splicing may be relatively common for plant SET genes and consider the possibility that it has a functional role in regulating histone methylation. In addition, the possible role of antisense RNA in expression of Arabidopsis SET genes will be discussed.

2. Classification and functions of plant SET proteins

Proteins containing the conserved SET domain can be found in organisms ranging from virus to all three domains of life (Bacteria, Archaea, and Eukayota). At the time of writing, 1026 entries of SET proteins are cataloged in the Pfam sequence alignment database [45]. According to annotation from both Pfam [45] and ChromDB, at least 47, 37 and 35 SET proteins are present in Arabidopsis, rice and maize respectively. In plants, Arabidopsis SET (AtSET) proteins are the best annotated and characterized. Baumbusch et al. [44] were the first to classify 37 putative Arabidopsis SET genes into four distinct classes: (1) Enhancer of zeste [E(z)] homologs; (2) Ash1 homologs and related; (3) trithorax (trx) homologs and related; and (4) Suppressor of variegation [Su(var)] homologs and related. Subsequently, Springer et al. [22] classified 32 Arabidopsis and 22 maize SET proteins into 19 orthology groups distributed among five classes according to their phylogenetic relationships and domain organization.

We have added SET genes from rice to those of Arabidopsis and maize based on their sequence homology and phylogenetic relationships (Table 1). Except for the addition of two classes (VI and VII), this analysis is consistent with the classification by Springer et al. [22]. Although much remains to be experimentally verified, in general, it appears that the resulting groupings reflect the substrate specificities of the members.

2.1. Class I: enhancer of zeste [E(Z)] homologs (H3K27)

In Drosophila, homeotic gene regulation (via H3K27 methylation) is mediated by the polycomb repressive complex 2 (PRC2) that contains the PcG proteins, E(Z), ESC, P55 and Su(z)12 [34,35,46–48]. As reviewed in Gutton and Berger [49], homologs corresponding to these PcG proteins are also present in Arabidopsis PcG complexes that function in mediating various aspects of plant development. In Arabidopsis, CURLY LEAF (CLF), SWINGER (SWN) and MEDEA (MEA) constitute the E(Z) homologs of SET domain-containing PcG proteins (Table 1). CLF was the first identified plant SET domain-containing protein within this class [50]. In Arabidopsis, CLF controls leaf and flower morphology as well as flowering time through repression of the floral homeotic genes.
AGAMOUS (AG) and SHOOTMERISTEMLESS (STM) [51,52]. Its close relative, SWN, acts redundantly in mediating tri-methylation of histone H3K27 at both AG and STM loci [52,53]. MEA plays a role in maintaining its own imprinting during endosperm development and methylation of H3K27 is associated with MEA silencing during vegetative development and male gametogenesis [54–56]. These recent findings provide further evidence that the E(Z) group of SET domain-containing proteins is primarily associated with H3K27 methylation [22] and that tri-methylation of H3K27 is present at specific repressed loci at different stages of plant development. In Drosophila, an additional polycomb repressive complex, PRC1, was found to impede chromatin remodeling mediated by hSWI/SNF [57,58].

Although no complex corresponding to PRC1 has been identified in plants, the presence of a family of multiple PcG members in Arabidopsis suggested that they may constitute alternative functional PRC2 complexes [49,53]. For example, Makarevich et al. [59] recently showed that at least two different MEA, CLF or SWN-containing PcG complexes are involved in histone H3K27 tri-methylation at FHERES1 (PHE1) or FUSCA3 (FUS3) loci during seed development. In addition, vegetative expression of PHE1 and FUS3 was upregulated only in the swn clf double mutant, suggesting their redundant role in constituting PRC2, and hence their participation in different PRC2 complexes, thereby repressing PHE1 [59].

A conserved role of SET proteins in plant development is reinforced by the finding that overexpression of the SET domain of the rice CLF homolog, OsSET (SDG711), affects vegetative development of transgenic Arabidopsis [60]. In addition to the PcG complex that contains ASHR3 and SDG736, an additional PHD (plant homeodomain) element is located upstream of the ASH domain (Fig. 1). Therefore, it is possible that during the evolution of these SET domain-containing proteins, an additional domain was acquired to enable additional functions.

Although biochemical characterization of proteins within this class is not yet available, two lines of evidence suggest that this class of proteins is involved in H3K36 methylation. ASH1 is the Arabidopsis homolog of yeast Set2 [44] which has been shown to exhibit H3K36 methyltransferase activity [64]. In addition, genetic studies revealed that ash2 mutants show an early flowering phenotype that is accompanied by a decrease in H3K36 methylation at the FLOWERING LOCUS C (FLC) [65].

3.3. Class III: the trithorax homologs and related proteins (H3K4)

In contrast to PcG proteins, trxG proteins are positive regulators of homeotic gene expression [66,67]. In Drosophila, Trx functions through the trithorax acetyltransferase complex (TAC1) in mediating homeotic gene expression [68]. H3K4 methyltransferase activity of Trx was only recently demonstrated experimentally [40]. In contrast, Set1, the Trx homolog in yeast, was the first and only H3K4 methyltransferase identified in yeast [69] and it is important for normal cell growth and silencing of rDNA [70]. Similar to Trx, Set1 functions in a large protein complex, COMPASS (Complex Proteins Associated with Set1), in mediating histone methylation [71,72]. In Arabidopsis, it has been demonstrated that H3K4 methylation is associated with a permissive state of gene expression during plant development [73,74]. ATX1 (subgroup III-1; Table 1) was the first protein confirmed to have H3K4 methyltransferase activity reported by Springer et al. [22], this class of proteins can be grouped into several orthology groups, based on the position of the SET domain. As shown in Table 1, the rice homologs can be classified into corresponding subgroups.

Arabidopsis ASH1H1, maize SDG102 and rice SDG708 contain an N-terminally located SET domain while ASHH3, ASHH4, SDG110 and SDG724 possess a more centrally located SET domain. Compared to other proteins within this class, ASHH2 (1759-aar) is significantly larger and it, like rice SDG725 (2133-aar), can be placed in a distinct orthology group. In addition to the presence of canonical domains (AWS, SET, post-SET), ASHH2 contains a CW (cysteine and tryptophan conserved) domain (as predicted by Pfam-A) that can be found in at least five other protein families in higher plants [63]. As for the orthology group that contains ASHR3 and SDG736, an additional PHD (plant homeodomain) element is located upstream of the AWS domain (Fig. 1). Therefore, it is possible that during the evolution of these SET domain-containing proteins, an additional domain was acquired to enable additional functions.

Notes to Table 1:

a SET protein from Arabidopsis, maize and rice were classified into seven classes. Classification of Arabidopsis and maize Class I to V was reported in Springer et al. [22]. Class VI and VII include SET proteins with truncated or interrupted SET domains.

b Different orthology groups of Class I to V SET proteins defined by Springer et al. [22].

c Arabidopsis SET proteins listed in Pfam database, common name, ChromDB name, locus number and amino acid length are indicated.

d Maize SET proteins listed in ChromDB. Asterisk denotes length of partial protein sequence and corresponding accession number of EST sequence.

e Rice SET proteins from Chrom DB were queried against Arabidopsis protein database using NCBI BLAST and their domain architecture were determined using NCBI conserved domains database. Phylogenetic relationship between homologous proteins were determined by Vector NTI software (Invitrogen) and the proteins are grouped into different orthology groups corresponding to Arabidopsis and maize SET proteins based on both sequence homology and domain architecture. Locus number and amino acid length are indicated. Rice SDG701, SDG732 and SDG738 were not grouped.
activity in plants and it is involved in floral development through activating flower homeotic genes [75,76].

Chromatin immunoprecipitation experiments showed that tri-methylation of H3K4 is associated with active transcription from the seed-specific promoter, *phaseolin (phas)* through the activity of a family of B3 domain-containing transcription factors, including FUS3 (D.W-K Ng and T.C. Hall, unpublished data) and *PvALF* (*Phaseolus vulgaris* AB13-like factor).

### Table 1

<table>
<thead>
<tr>
<th>Class</th>
<th>Gp</th>
<th><em>Arabidopsis thaliana</em></th>
<th><em>Zea mays</em></th>
<th><em>Oryza sativa</em> spp. <em>Japonica</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1</td>
<td>MEA (SDG5; At1g02580; 689aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>CLF (SDG1; At2g23380; 902aar)</td>
<td>MEZ1 (AF443596; 931aar)</td>
<td>SDG711 (Os06g16390; 896aar)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>SWN (SDG10; At4g02020; 856aar)</td>
<td>MEZ2 (AF443597; 894aar), MEZ3 (AF443598; 895aar)</td>
<td>SDG718 (Os03g19480; 895aar)</td>
</tr>
<tr>
<td>II</td>
<td>1</td>
<td>ASHH3 (SDG7; At2g44150; 363aar)</td>
<td>SDG110 (AF545814; 342aar)</td>
<td>SDG724 (Os09g13740; 340aar)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>ASHH4 (SDG24; At3g59960; 352aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>ASHR1 (SDG37; At3g59960; 352aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>ASHR3 (SDG4; At4g30860; 497aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>VI</td>
<td>1</td>
<td>ATX1 (SDG27; At2g31650; 1062aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>ATX3 (SDG14; At3g61740; 902aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>ATX4 (SDG16; At4g27910; 912aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>ATX7 (SDG25; At2g42400; 1421aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>IV</td>
<td>1</td>
<td>SUVH1 (SDG32; At3g04940; 670aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>SUVH4 (SDG33; At5g13960; 624aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>SUVH6 (SDG29; At5g53430; 1043aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>SUVH8 (SDG21; At2g24740; 755aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>V</td>
<td>1</td>
<td>SUVH3 (SDG19; At1g73100; 669aar)</td>
<td>SDG105 (AY093419; 678aar), SDG11 (AY187718; 486aar)</td>
<td>SDG728 (Os05g11700; 672aar)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>SUVH7 (SDG17; At1g17770; 693aar)</td>
<td>SDG113 (AF545813; 766aar)</td>
<td>SDG709 (Os01g59620; 736aar), SDG733 (Os11g03700; 663aar), SDG734 (Os12g03460; 663aar)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>SUVH8 (SDG21; At2g24740; 755aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>SUVH9 (SDG22; At2g13460; 650aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>SUVH10 (SDG34; At5g24330; 349aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>VI</td>
<td>1</td>
<td>SUVH1 (SDG31; At1g04050; 630aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>SUVR2 (SDG18; At5g43990; 717aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>SUVR4 (SDG31; At1g04380; 492aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>SUVR5 (SDG6; At2g23740; 203aar)</td>
<td>SDG107 (DV544186; 148aar)*, SDG117 (AY187719; 1198aar), SDG11 (AY187718; 486aar)</td>
<td>SDG706 (Os02g40770; 741aar)</td>
</tr>
<tr>
<td>VII</td>
<td>1</td>
<td>SDG40 (At3g17240; 491aar)</td>
<td>SDG131 (DV522881; 131aar)<em>, SDG131 (DV522881; 131aar)</em></td>
<td>SDG716 (Os03g49730; 403aar)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Putative RuBiCo SSMT (At3g07670; 504aar)</td>
<td>SDG130 (AA175997; 410aar)</td>
<td>SDG740 (Os08g10470; 392aar)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>At2g18850 (543aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>At5g14260 (514aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>RuBiCo LSMT (At1g14030; 482aar)</td>
<td>SDG122 (DV253779; 534aar)</td>
<td>SDG739 (Os03g72660; 536aar)</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>At1g01920 (572aar)</td>
<td>SDG140 (DV535249; 217aar)*, SDG123 (AY117296; 303aar)</td>
<td>SDG722 (Os04g53706; 517aar)</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>At1g24610 (476aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>At3g53080 (463aar)</td>
<td>SDG127 (DV520485; 2351aar)</td>
<td>SDG741 (Os11g27060; 298aar)</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>At3g56570 (531aar)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>At4g20130 (483aar)</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
[77–79]. In addition, an increase in H3K4me3 at the phas chromatin was detected within 1 h upon phas activation, illustrating the dynamic aspects of chromatin remodeling in mediating the activation of this seed-specific promoter [79]. Moreover, LEC1 (LEAFY COTYLEDON1; [80]) and FUS3 (transcription factors involved in embryo development) are up-regulated in the clf swm mutant [59]. Therefore, further studies on the roles of the various classes SET proteins in modulating epigenetic regulation of gene expression during seed development are likely to provide new insight to developmental processes.

As shown in Fig. 1, in addition to the SET domain, Class III proteins bear several highly conserved protein domains (PWPP, FYRN, FYRC) and the PHD domain (often found in chromatin-associated proteins [81,82]). While all these conserved domains can be located in proteins within subgroup III-1 (ATX1 and ATX2), subgroup III-2 proteins (ATX3, ATX4 and ATX5) lack the FYRN and FYRC domains. In contrast to other subgroups of proteins within the same class, ATXR3, ATXR7 (subgroups III-3 and -4) and their corresponding rice homologs (SDG701 and SDG717) lack all of the highly conserved domains, and they possess only a C-terminally located SET domain and a post-SET domain. Although biochemical characterization is lacking for the other members of the plant Trx class, the presence of various highly conserved domains within this class of proteins may suggest diverse functions for these SET domain-containing proteins. In fact, Alvarez-Venegas et al. [83] showed that ATX1 interacts with phosphatidylinositol 5-phosphate (PI5P) through its PHD finger. PI5P is a component of the cell lipid pool and acts as secondary messenger mediating several plant processes, such as hyperosmotic stress response and membrane trafficking, via lipid signal transduction [84]. Such interaction between PHD finger and PI5P has also been demonstrated in animals [85]. These findings support the idea that SET domain proteins may act as transducers that bind signaling molecules such as PI5P, modifying enzymatic activity either by complete deactivation or by changing substrate specificity.

2.4. Class IV: proteins with a SET and a PHD domain

This class of proteins is characterized based on the presence of a PHD finger and a C-terminally located SET domain (Fig. 1). Corresponding rice homologs, SDG720 and SDG730, can be located in the ChromDB in addition to the Arabidopsis (ATXR5 and ATXR6) and maize (SDG139 and SDG129) members of this class. In contrast to Class III proteins, Springer et al. [22] suggested that Arabidopsis ATXR5 and ATXR6 arose from an independent duplication of Arabidopsis SET genes and that certain conserved amino acid residues within the SET domain important for HKMT activity are different. Nevertheless, the presence of the PHD finger and recent findings that this domain can bind methylated H3K4 [86,87], suggested that this class of proteins, like those in Class III, is closely related to the trithorax class of SET-domain containing proteins. Although little is known about the function of this class of proteins, recent data suggested that ATXR5 and ATXR6 may contribute to cell cycle

Fig. 1. Domain architecture of plant SET proteins. Representative schematic diagram (right) showing overall domain composition of various orthology groups of SET proteins from different classes (left). Diagrams for Arabidopsis and maize SET proteins are described in detail by Springer et al. [22]. Abbreviations: EZD, E(Z) domain; SANT, SWI3, ADA2, N-CoR and TFIIB′′′ DNA-binding domain; CXC, cysteine-rich region; AWS, associated with SET domain; PHD, plant homeodomain; Zf-CW, a zinc finger with conserved Cys and Trp residues; PWPP, domain named after a conserved Pro–Trp–Trp–Pro motif; FYRN, F/Y-rich N-terminus; FYRC, F/Y-rich C-terminus; YDG, conserved Tyr–Asp–Gly motif.
regulation or DNA replication through interactions with proliferating cell nuclear antigen (PCNA) [88].

2.5. Class V: suppressor of variegation [Su(var)] homologs and relatives (H3K9)

This large class of proteins is characterized by the presence of pre-SET, SET and post-SET domains (Fig. 1). In addition to the already classified Arabidopsis and maize Su(var) proteins, corresponding rice homologs are sorted in Table 1 into the various orthologous groups. In all the Su(var) homologs (subgroups V-1, -2 and -5), an additional conserved YDG (conserved Tyr–Asp–Gly motif) domain is present upstream of the pre-SET domain.

Members of this class are likely to be involved in heterochromatin formation mediated by H3K9 methylation [33,89,90]. In Arabidopsis, SUVH4 or KRYPTONITE (KYP), the best studied member of this class, represses expression of the floral homeotic gene, SUPERMAN (SUP), through its H3K9 methyltransferase activity [90]. Naumann et al. [91] reported that the N-terminus and YDG domain in SUVH2 are involved in directing DNA methylation and subsequent histone methylation at its target locus so that heterochromatic methylation marks are affected in suvh2 mutant plants. Recently, Ebbs et al. [92,93] showed that there is a hierarchical and locus-specific control of H3K9 di-methylation as well as non-CG DNA methylation by SUVH4, SUVH5 and SUVH6 HMT in Arabidopsis. The involvement of this class of proteins in heterochromatin formation was demonstrated for NtSET1 in tobacco [94], and reinforced by the subsequent finding that NtSET1 interacts with LHP1 (Arabidopsis homolog of HP1) [95]. In addition, Yu et al. [95] showed that NtSET1 contains both H3K9 and H3K27 methylation activities in vitro and that overexpression of NtSET1 leads to an increase in H3K9 di-methylation in tobacco BY2 cells. Overall, the highly conserved, plant-specific distribution of domains (YDG, Pre-SET and SET) found in subgroups V-1, -2 and -5 (Fig. 1) of proteins suggests that they share a recent, plant-specific common ancestor [61].

Unlike the Su(var) homologs, all the Su(var)-related homologs (subgroups V-4, -6 and -7) lack the YDG domain (Fig. 1). However, Thorstensen et al. [96] have identified a novel conserved plant specific N-terminal domain, WIYLD, that is present in Arabidopsis SUVR1, SUVR2, SUVR4 as well as in other SUVR-homologs of other plant species. Interestingly, the authors also showed that recombinant SUVR4 acts as a di-methyltransferase with preference for mono-methylated H3K9 as substrate. Therefore, this finding may suggest that the SUVH and SUVR proteins can act in concert in achieving various functional H3K9 methylation states that will eventually lead to DNA methylation in a locus-specific manner [97].

2.6. Class VI: proteins with an interrupted SET domain

Proteins from both Classes VI and VII lack the diverse domains that can be found in other classes of SET proteins and the SET domain is either truncated or interrupted in the SET-I region (Fig. 1). Class VI consists of two ASH-related proteins (ASHR1 and ASH2R) and three Trx-related (ATXR1, ATXR2 and ATXR4) proteins. Homologous SET genes can be found in maize (SDG122, SDG123, SDG130 and SDG140) and rice (SDG716, SDG740, SDG739, SDG722 and SDG741). This suggests that this class of proteins existed before the divergence of monocot and dicot plants. The SET domain of ASH1 is interrupted by a Zf-MYND [zinc finger MYND (Myeloid, Nervy, and DEAF-1)] domain. Brown et al. [98] recently reported that Smyd2, the human homolog of Arabidopsis ASH1, restricts cell-cycle progression through di-methylation of H3K36. Although not yet established in plants, it is likely that, like its human homolog, ASH1 will be shown to have H3K36 methyltransferase activity.

2.7. Class VII: RBCMT and other SET-related proteins

While the majority of SET domain-containing proteins possess histone-methylating activity, Class VII proteins methylate non-histone proteins (Table 1). Rubisco large (RBLSMT) and small (RBSSM) subunit methyltransferases (RBCMTs) exemplify SET-related proteins that methylate non-histone targets [99–101]. While the N- and C-termini of these proteins bear high similarity to those of other SET domain proteins (Fig. 1), the SET-I region has only weak similarity to, and is longer than that of canonical SET proteins. Nevertheless, the structural features within the SET-I region essential for substrate-specific methylation are retained [100,102]. We include Arabidopsis SDG40, two unnamed proteins (corresponding to At2g18850 and At5g14260) and rice SDG731 in this class as they show some similarity to RBSSMT. The remaining five uncharacterized Arabidopsis proteins were classified into this class as they contain domain architecture resembling that of RBLSMT.

3. Regulation of plant SET proteins

3.1. Regulation of SET protein functions through protein–protein interactions

At least some SET proteins are capable of self-association [103–105]. This property suggests that the lysine-methyltransferase activity, including varying degrees of methylation (mono, di- or tri-methylation), substrate specificity and localization can be subjected to regulation by intra- and inter-molecular interactions with other proteins. Indeed, modulation of the SET domain-mediated lysine-methyltransferase activity by interacting proteins has been demonstrated for yeast SET1-COMPASS [106,107] and for human MLL1 trithorax complex [108].

Regulation of histone methylation by SET domain-containing methyltransferases via their interaction with other proteins has been demonstrated in yeast and mammalian systems [48,72,109]. In plants, the regulation of SET protein-mediated histone methylation through protein–protein interactions is poorly understood. However, the involvement of E(Z) class proteins in Pcg complexes that contain various non-SET proteins provides an avenue to investigate the role, if any, of
interacting proteins in regulating the methyltransferase activity. In Arabidopsis, the involvement of MEA, CLF and SWN in developmental processes may be linked to their interactions with other non-SET proteins, as mutations in the interacting partners display phenotypes indicative of affecting similar developmental pathways. During seed development, MEA is part of a PcG complex that includes several other regulators of early seed development such as FERTILIZATION-INDEPENDENT ENDOSPERM (FIE) and FERTILIZATION-INDEPENDENT SEED2 (FIS2) [110–112]. FIE is a WD40 domain-containing protein that shows similarity to the Drosophila EXTRA SEX COMB (ESC). FIS2 is a Zn-finger protein similar to Su(Z)12 [35,47]. Physical interactions between MEA and FIE [113,114]), and between MEA (or SWN) and FIS2 [53,115] have been demonstrated using yeast-two-hybrid assays. Therefore, based on the similarities in the mutant phenotypes and their requirement for the repression of PHE1 during endosperm development [59], complex formation is a prerequisite for MEA or SWN to function as HMTs in planta.

Recently, Kohler et al. [116] showed that Arabidopsis MULTICOPY SUPPRESSOR OF IRA 1 (MSI1) interacts with FIE and is required for seed development. Like FIE, MSI1 is a component of a putative complex containing FIE and CLF [53]. Based on the similarity of mutant phenotypes, EMF2 (a WD40 domain-containing protein) is thought to be the plant homolog of the fly and mammalian ESC-E(Z) complexes. Interestingly, FIE also physically interacts with CLF [51]. Based on the similarity of mutant phenotypes, EMF2 (a zinc-finger protein) has been proposed to be an additional component of a putative complex containing FIE and CLF [53].

In addition to its methyltransferase activity, the substrate specificity of a particular SET protein may be mediated by protein interactions. For example, the WDR5 (a WD40 domain-containing protein) component of MLL1 trithorax complex has been shown to confer substrate specificity by interacting with di-methylated H3K4 [118,119]. By analogy, the importance of the WD40 domain-containing protein, FIE, to overall plant development can be explained by its association with MEA and CLF in the regulation of sporophytic and gametophytic stages, respectively [51,120]. It can be speculated that FIE may regulate HMT activity by maintaining complex integrity and (or) by providing substrate specificity similar to that observed for WDR5 in the human MLL1 complex. The presence of an additional WD40 domain protein, MSI1, in the MEA–FIE–FIS2–MSI1 complex may confer the ability to recognize two different methylated lysine residues on histones.

3.2. Alternative splicing of Arabidopsis SET genes

The presence of introns in eukaryotic RNAs confers multiple degrees of flexibility in gene expression via various molecular mechanisms, such as developmental and cell-specific regulation of transcription [121], generation of non-coding RNAs [122], and alternative splicing [123]. Among these mechanisms, alternative splicing is of special interest in that it can produce two or more forms of mature mRNA from the same precursor transcript and hence generate transcripts bearing diverse information.

In general, alternative splicing in non-translated regions can have an impact on protein expression levels [124]. Conversely, in coding regions, it may alter protein structure and function. In extreme cases, alternative splicing can lead to the production of two proteins with antagonistic functions [125]. It is estimated that 60–80% of human genes undergo alternative splicing [126], contributing to human genome complexity. A smaller percentage of plant genes, ~22% in Arabidopsis and 10% in rice [127], also experience alternative splicing. It is possible that these numbers will increase in the future, as more plant cDNAs or ESTs are characterized. Any statement about alternative splicing events should take into consideration the fact that they may be the result of experimental artifacts such as sequencing errors, splicing errors, and RNA degradation events [127,128]. However, despite these possibilities, some alternative splicing events, especially those that are evolutionarily conserved, may be biologically important and contribute to organismal fitness [129]. Expression of alternatively spliced transcripts can be tissue specific [130] and can be differentially regulated by physical [131] or chemical [132] stimuli, arguing for the involvement of multiple regulatory systems in the expression of individual isoforms.

Alternative splicing in the coding region of Arabidopsis SET genes may generate proteins with distinct functions. Such effects have been reported for Drosophila Su(var)3–9, which can express two distinct transcripts (2.4 kb and 2.0 kb), encoding two proteins that share similarity only for the first 80 amino acid residues (aar). Whereas the 2.4 kb transcript is translated to yield a SET domain-containing HMT, the 2.0 kb transcript encodes the γ subunit of eukaryotic translation initiation factor 2 [133]. Conversely, protein isoforms from alternative splicing in coding regions may have redundant functions if the conserved domains are not affected. An example is zebrafish SET gene, SmyD1. Inactivation of either isoform of this gene results in no morphological phenotype whereas debilitation of both isoforms have severe effects on myofibril organization [134]. The possibility also exists that SET protein isoforms resulting from alternative splicing assemble into higher order protein complexes. Homodimerization or heterodimerization has been reported for the virus SET protein sSET [135], human G9a, and GLP [136], and Drosophila ASH1 and Trx [105]. If plant SET proteins form dimers, alternative splicing of SET genes could generate proteins forming various homo- and hetero-dimers having distinct biological activities.

Scanning all 47 Arabidopsis SET genes against the Alternative-Splicing-in-Arabidopsis database (ASIP) [127], the
Alternative splicing of Arabidopsis SET genes could have important consequences in regard to protein functions. For 3 \((ATX2, SUVH1, At5g14260)\) of the 18 alternatively spliced SET genes, examination of ESTs in the ASIP database revealed that splicing differences occurred in the 3′ UTs (Table 2). The 3′ or 5′ UTs are known to contain elements that can control protein stability [138], protein translation efficiency [139], and spatial [140] or subcellular distribution of proteins [141]. Casas-Mollano et al. [142] showed that an intron located in the 5′ UTR of the Arabidopsis \(SUVH3\) regulates both constitutive and alternative splicing affecting the SET domain. In the case of \(At1g01920\), ~40-aar are added/removed from the SET-N motif high levels of expression. Therefore, alternative splicing events observed in \(ATX2, SUVH1\) and \(At5g14260\) may affect the localization or abundance of transcripts or encoded proteins, rather than causing a change in the structure of encoded proteins.

Alternative splicing for the other 15 Arabidopsis SET genes is predicted by conceptual translation to occur at the protein-encoding region of transcripts (Table 2). For 10 SET genes, alternative splicing events appear to affect the start/stop codon or a region of the protein that lies outside of a conserved domain of sequence similarity. However, the other 5 \(SET\) genes \((At1g01920, At5g35080, At5g17240, ATXR5\) and \(SUVR3\)\) have alternative splicing affecting the SET domain. In the case of \(At1g01920\), ~40-aar are added/removed from the SET-N motif.

### Table 2

<table>
<thead>
<tr>
<th>Class</th>
<th>SET gene (locus)</th>
<th>Alternative splicing patterns</th>
<th>Possible effects on proteins</th>
<th>Antisense</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-2</td>
<td>ASHR3 ((At4g30860))</td>
<td>IntronR</td>
<td>Use a different stop codon</td>
<td>–</td>
</tr>
<tr>
<td>II-3</td>
<td>ASHH1 ((At1g76710))</td>
<td>AltA</td>
<td>No effect on domain structure</td>
<td>–</td>
</tr>
<tr>
<td>II-4</td>
<td>ASHH2 ((At1g77300))</td>
<td>AltA; ExonS</td>
<td>No effect on domain structure</td>
<td>Y</td>
</tr>
<tr>
<td>III-1</td>
<td>ATX1 ((At2g31650))</td>
<td>AltD</td>
<td>No effect on domain structure</td>
<td>–</td>
</tr>
<tr>
<td>III-1</td>
<td>ATX2 ((At1g39830))</td>
<td>AltA</td>
<td>Affect 3′UTR</td>
<td>–</td>
</tr>
<tr>
<td>III-2</td>
<td>ATX3 ((At3g17470))</td>
<td>IntronR</td>
<td>Use a different stop codon</td>
<td>–</td>
</tr>
<tr>
<td>IV</td>
<td>ATXR5 ((At5g09790))</td>
<td>AltA; ExonS</td>
<td>Affect SET domain</td>
<td>Y</td>
</tr>
<tr>
<td>V-1</td>
<td>SUVH1 ((At5g49490))</td>
<td>IntronR</td>
<td>Affect 3′UTR</td>
<td>Y</td>
</tr>
<tr>
<td>V-1</td>
<td>SUVH3 ((At1g37110))</td>
<td>–</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td>V-1</td>
<td>SUVH8 ((At2g24740))</td>
<td>–</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td>V-2</td>
<td>SUVH6 ((At2g22740))</td>
<td>–</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td>V-4</td>
<td>SUVR3 ((At3g03750))</td>
<td>AltP</td>
<td>Affect SET domain</td>
<td>–</td>
</tr>
<tr>
<td>V-5</td>
<td>SUVHS ((At2g35160))</td>
<td>AltA</td>
<td>Use a different start codon</td>
<td>–</td>
</tr>
<tr>
<td>V-6</td>
<td>SUVR1 ((At1g04050))</td>
<td>IntronR</td>
<td>Use a different start codon</td>
<td>–</td>
</tr>
<tr>
<td>V-6</td>
<td>SUVR2 ((At5g43990))</td>
<td>AltA; AltD</td>
<td>Use a different start codon</td>
<td>–</td>
</tr>
<tr>
<td>V-6</td>
<td>SUVR4 ((At3g04380))</td>
<td>IntronR</td>
<td>No effect on domain structure</td>
<td>–</td>
</tr>
<tr>
<td>VI</td>
<td>ASHR2 ((At2g19640))</td>
<td>IntronR</td>
<td>No effect on domain structure</td>
<td>–</td>
</tr>
<tr>
<td>VI</td>
<td>ATXR1 ((At1g26760))</td>
<td>–</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td>VII</td>
<td>SDG40 ((At5g17240))</td>
<td>IntronR</td>
<td>Affect SET domain</td>
<td>–</td>
</tr>
<tr>
<td>VII</td>
<td>At5g14260</td>
<td>AltD; IntronR</td>
<td>Affect 3′UTR</td>
<td>–</td>
</tr>
<tr>
<td>VII</td>
<td>At1g01920</td>
<td>AltA; IntronR</td>
<td>Affect SET domain</td>
<td>–</td>
</tr>
<tr>
<td>VII</td>
<td>At1g24610</td>
<td>–</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td>VII</td>
<td>At3g55080</td>
<td>AltA; ExonS</td>
<td>Affect SET domain</td>
<td>–</td>
</tr>
</tbody>
</table>

**a** Orthology groups of Arabidopsis SET proteins listed in Table 1.

**b** Arabidopsis SET genes, common name and/or locus number are indicated.

**c** Abbreviations for alternative splicing mechanisms described in Fig. 2.

**d** The possible effects of alternative splicing on proteins according to the conceptual translation of alternatively spliced transcripts.

**e** “Y” denotes SET gene to which naturally occurring antisense transcripts can be located in ASIP [127] and Arabidopsis Cis-NAT Pairs Database [148].
of the encoded protein. The splicing event of At3g55080 seems to directly regenerate or eliminate the SET domain coding region. In At5g17240, a 79-nt intron is retained in one isoform, resulting in a frameshift near the beginning of the SET domain coding region. Similarly, a 59-nt exon deletion from one isoform of ATXR5 is likely to cause a frameshift at the SET domain coding region. In the case of SUVR3, the consequences resulting from alternative splicing may be dramatic. We have found that RNAi-mediated debilitation of SUVR3 in transgenic Arabidopsis results in aberrant flower formation and other abnormalities (Wang, T. et al., unpublished data). During analysis of these plants, we found that two different SUVR3 transcripts were present. Sequences identical to those of the transcripts are present in cDNA and EST libraries [143,144], each comprising two exons and one intron (Fig. 3). It became evident that these RNAs encoded SUVR3 isoforms, with the total coding sequence length of isoform SUVR3α being 48-nt shorter than that in SUVR3β. This results in a 16-aar difference between the two isoforms, with a loss of 3 invariant and 3 conserved residues from SUVR3α, suggesting that the SUVR3 isoforms may differ in their ability to methylate histone H3. If so, alternative splicing could be seen to have a significant regulatory role in regard to chromatin remodeling.

3.3. Anti-sense RNA expression of Arabidopsis SET genes

Our understanding of anti-sense transcription in all organisms is very limited. In mouse, genome-wide analysis of transcripts suggests that antisense transcription is widespread and not restricted to imprinted genes, as previously postulated [145]. In Arabidopsis, the ratio of expression of the sense (S) to the antisense (AS) transcripts fluctuates markedly among different tissues during development [146]. In addition, sense–antisense transcripts (SAT) tend to be poly(A)-negative and nuclear localized in Arabidopsis, as in mouse [147].

Naturally occurring AS transcripts (NATs) corresponding to nine Arabidopsis SET genes were found in the ASIP [127] and Arabidopsis Cis-NAT Pairs Databases [148] (Table 2). This opens the possibility that the expression of the sense transcript for some of these genes might be under the control of the anti-sense message. Sense–antisense association would result in double-stranded RNA (dsRNA), a natural substrate for a Dicer ribonuclease [149]. The small RNAs produced by this reaction could direct the formation of a silent chromatin state, thus silencing the expression of the homologous genomic region.

NATs can be found in both prokaryotes and eukaryotes. In eukaryotic cells, such RNAs are known to be mostly involved in genomic imprinting [150] and X chromosome inactivation [151]. The discovery that AS transcription is more widespread than previously thought challenges this view and suggests that dsRNA may play a more general role (or perhaps a wider role) in regulation of gene expression. Therefore, it would be quite informative to determine if NATs affect the expression and chromatin state of one or more AtSET genes and then to elucidate the mechanism by which NATs act at these loci. NATs could facilitate a sensitive, rapid and dynamic control over At-SET expression under different environmental cues. Determination of the spatial and temporal expression pattern of these NATs and their corresponding genes will provide some clues to
the role of AtSET NATs in regulating plant development. Alternatively, NATs could assist in the permanent imprinting of AtSET genes. Except for MEA, a self-regulated imprinting SET gene [54–56] devoid of antisense transcript, no AtSET has been reported to be imprinted. Therefore, the nine AtSET genes for which AS transcript information is listed in Table 2 could be additional targets for genomic imprinting.

4. Future directions

The extensive genome duplication observed in plants appears to provide them with the plasticity necessary to successfully adapt to the various environmental conditions they encounter as sessile organisms. This is best exemplified by the diversification of genes involved in plant secondary metabolism through gene duplication [152,153]. Such gene duplication events offer a great opportunity for studying gene evolution [154,155]. In particular, studying the diverse functions of families of paralogous plant SET genes in plant development is likely to provide new insight into the molecular and functional evolution of proteins containing the SET and other auxiliary domains. In addition, although considerable attention has been paid to the SET domain proteins as histone methyltransferases, their activity as protein methyltransferases capable of modifying non-histone substrates and modulating their functions is still poorly understood. The paucity of knowledge is not restricted to plant biology, as only a few instances have been reported even in yeast and mammalian systems. For example, methylation of p53 by mammalian SET9 has been shown to cause the retention of p53 in the nucleus and increase the stability of an p53-dependent gene expression [156]. SET9 has also been demonstrated to methylate TAF10, a general transcription factor, in mammalian cells [25]. SET9 is capable of modifying non-histone substrates and modulating their functions in various cellular processes other than transcription. For instance, SET1 a histone H3 Lys4 methyltransferase has also been demonstrated to methylate TAF10, a general transcription factor, in mammalian cells [25]. SET9 is capable of modifying non-histone substrates and modulating their functions in various cellular processes other than transcription.

Acknowledgements

This work was supported in part by NSF grants MCB 0110477 and MCB 0346681 to TCH and U.S. Public Health Service Grant GM58770 to R.A. We thank Xiangyu Shi for help and discussion.

References

D.W.-K Ng et al. / Biochimica et Biophysica Acta 1769 (2007) 316–329


D.W. Ng, M.B. Chandrasekharan, T.C. Hall, Ordered histone modifications are associated with transcriptional poising and activation of the phaseolin promoter, PLoS One 3 (2008) e315–e326.


V. Krauss, G. Reuter, Two genes become one: the genes encoding heterochromatin protein Su(var)3–9 and translation initiation factor subunit elf-2gamma are joined to a dicistronic unit in holometabolic insects, Genetics 156 (2000) 1157–1167.


