Nucleic Acids Research, 2001, Vol. 29, No. 21 **4319–4333**

The *Arabidopsis thaliana* genome contains at least 29 active genes encoding SET domain proteins that can be assigned to four evolutionarily conserved classes

Lars O. Baumbusch, Tage Thorstensen, Veiko Krauss[1], Andreas Fischer[1], Kathrin Naumann[1], Reza Assalkhou, Ingo Schulz[1], Gunter Reuter[1] and Reidunn B. Aalen*

Division of Molecular Biology, Department of Biology, University of Oslo, PO Box 1031 Blindern, N-0315 Norway and [1]Institute of Genetics, Martin Luther University, D-06120 Halle, Germany

Received July 20, 2001; Revised and Accepted September 7, 2001

DDBJ/EMBL/GenBank accession nos[+]

ABSTRACT

SET domains are conserved amino acid motifs present in chromosomal proteins that function in epigenetic control of gene expression. These proteins can be divided into four classes as typified by their *Drosophila* members E(Z), TRX, ASH1 and SU(VAR)3-9. Homologs of all four classes have been identified in yeast and mammals, but not in plants. A BLASTP screening of the *Arabidopsis* genome identified 37 genes: three *E(z)* homologs, five *trx* homologs, four *ash1* homologs and 15 genes similar to *Su(var)3-9*. Seven genes were assigned as *trx*-related and three as *ash1*-related. Only four genes have been described previously. Our classification is based on the characteristics of the SET domains, cysteine-rich regions and additional conserved domains, including a novel YGD domain. RT–PCR analysis, cDNA cloning and matching ESTs show that at least 29 of the genes are active in diverse tissues. The high number of SET domain genes, possibly involved in epigenetic control of gene activity during plant development, can partly be explained by extensive genome duplication in *Arabidopsis*. Additionally, the lack of introns in the coding region of eight *SU(VAR)3-9* class genes indicates evolution of new genes by retrotransposition. The identification of putative nuclear localization signals and AT-hooks in many of the proteins supports an anticipated nuclear localization, which was demonstrated for selected proteins.

INTRODUCTION

Gene expression in eukaryotes depends on both intrinsic regulatory mechanisms, including enhancer–promoter interactions, and chromosomal context, including chromatin structure. Chromatin silencing mechanisms are involved in X chromosome inactivation, genomic imprinting, developmental control of homeotic genes, silencing of mating type loci in yeast and heterochromatin-induced gene silencing, known as position-effect variegation (PEV), in *Drosophila melanogaster* (1–3). In addition, segregation of chromosomes during cell division and telomere and centromere function is dependent on the correct higher order chromatin structure (see 4).

An understanding of the mechanisms governing modulation of chromatin structure is emerging from the identification of genes encoding proteins forming chromatin complexes. In *Drosophila*, the polycomb group (*PcG*) of genes maintains a repressive state, while the trithorax group (*trxG*) of genes preserves the activity of homeotic genes in appropriate segments throughout development (5). About 120 loci have been identified in which mutations enhance, *E(var)*, or suppress, *Su(var)*, PEV (6).

Chromatin-modulating proteins are thought to be involved in multimeric protein–protein interactions (7) and contain characteristic amino acid motifs, e.g. a chromo domain, PHD finger or SET domain (5,8,9). The SET domain, a 130–160 amino acid motif, is found in proteins that are members of PcG, trxG and SU(VAR) and was named after the genes *Su(var)3-9*, *Enhancer of zeste* [*E(z)*] and *trithorax* (*trx*).

Homologs of these three genes have been identified in yeast and mammals (see 5,10). The first plant genes identified encoding SET domain proteins were *E(Z)* homologs. The *CURLY LEAF* (*CLF*) gene of the model plant *Arabidopsis thaliana* is involved in the control of leaf and flower morphology and flowering time (11). *MEDEA* (*MEA*), alternatively called *FERTILIZATION INDEPENDENT SEED*

*To whom correspondence should be addressed. Tel: +47 22854437; Fax: +47 22854605; Email: reidunn.aalen@bio.uio.no
Present address:
Veiko Krauss, Department of Microbiology and Genetics, University of Leipzig, Johannisallee 21, D-04103 Leipzig, Germany

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

[+]AF344444–AF344452, AF394239, AY045576

DEVELOPMENT 1 (*FIS1*), is an inhibitor of endosperm development in the absence of fertilization (12,13) and is also implicated in imprinting of paternal genes (12,14). The CLF and MEA proteins, as well as the two mouse and human homologs of E(Z) share similar amino acid compositions over the length of the proteins and are particularly similar in a cysteine-rich region and the SET domain (15). The involvement in chromatin-dependent gene regulation (15,16), the influences on PEV (17) and decondensation of chromatin structure in some *E(z)* mutants, indicate that E(Z) class proteins play a major role in maintaining the integrity of chromosomes (18).

The protein encoded by *Su(var)3-9* and its yeast (CLR4), human (SUV39H1) and mouse (Suv39h1) homologs, presumably have a key function in heterochromatin packaging (4,8,19). The human and mouse SUVAR39 and CLR4 and the human G9a proteins have been shown to selectively transfer a methyl group to histone 3 (20–22). Mutations in the SET domain abolish methyltransferase activity. In addition to the SET domain, these proteins have a conserved chromo domain in the N-terminal part and a cysteine-rich region adjacent to the SET domain (4).

The *Drosophila* TRX protein and its human and mouse homologs (ALL-1/All-1, also called MLL or HRX) share high similarity in the C-terminal SET domain and in the central part of the protein, where PHD fingers and an extended PHD finger (ePHD) are found (23–25). These fingers have unique Cys-His-Cys patterns similar to zinc fingers and may be involved in protein–protein interactions (25,26). Recently, two *Arabidopsis trx* homologs, *ATX1* and *ATX2*, were identified and a new domain associated with SET in trithorax (DAST) class of proteins was described (27). A fourth *Drosophila* SET domain gene, *absent, small or homeotic discs 1* (*ash1*), can also be classified as a *trxG* gene, and its encoded protein contains a PHD finger. This is also the case for the human homolog huASH1 (28). The SET domain of the ASH1 class proteins is not localized in the C-terminus, but rather in the middle part of the protein (10,28,29).

The four classes of SET domain genes are evolutionarily conserved in the animal kingdom and they play important functions in epigenetic control of gene expression and chromatin packaging. Studies of (trans)gene silencing have given a clear indication of the importance of epigenetic control of gene expression in plants (30,31). Given the presence of *E(z)* class genes in plants, we expected that SET domain genes of the other classes would also be present. Since the complete sequence of the *Arabidopsis* genome is available (32), we chose to take a bioinformatics approach to identify such genes. In the present paper we show that *Arabidopsis* has more than 30 such genes and that they can be grouped into four distinct classes, based on the characteristics of the SET domains and cysteine-rich regions of E(Z), TRX, ASH1 and SU(VAR)3-9 and other associated domains. Our characterization of the expression patterns of these genes by RT–PCR indicates a wide, but spatially and temporally differential, distribution of their transcripts during plant development. At least 29 of the putative genes are expressed. Nuclear localization was demonstrated for selected proteins. The high number of genes and their diverse expression patterns may reflect a high complexity of epigenetic control of gene activity during plant development.

MATERIALS AND METHODS

Bioinformatics tools

Database searches were performed using BLASTP and TBLASTX (http://www.ncbi.nlm.nih.gov/BLAST/). Multiple alignments of protein sequences were done with the ClustalX program (http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/) and manually adjusted with the GeneDoc program (http://www.psc.edu/biomed/genedoc/). Proteins lacking vital conserved residues were excluded from the alignment. In cases where the annotations in the EMBL (http://www.ebi.ac.uk/Databases/index.html) and MIPS (http://mips.gsf.de/) databases deviated, gene predictions were controlled using GENSCAN (http://genome.dkfz-heidelberg.de/cgi-bin/GENSCAN/genscan.cgi), GeneMark (http://dixie.biology.gatech.edu/GeneMark/eukhmm.cgi) and Gene finder. Protein domains were identified using the programs RPS-BLAST and PSI-BLAST (NCBI), ProfileScan (http://www.isrec.isb-sib.ch/software/PFSCAN_form.html) and PROSEARCH (MIPS) searching the Pfam-A, Prosite profiles and Smart databases. BAC clone positions were determined using MapViewer (http://www.arabidopsis.org/servlets/mapper). Gene duplications were investigated using the MIPS Interactive Redundancy Viewer (http://mips.gsf.de/proj/thal/db/gv/rv/rv_frame.html).

RNA isolation and RT expression analyses

Total RNA was isolated using Trizol reagent (Gibco BRL) as described by the manufacturer. For the SUVH group, where most genes are intronless in the coding region, reverse transcription was carried out on total RNA using M-MLV reverse transcriptase (RT). First strand cDNA was made from ~1 µg total RNA from different *Arabidopsis* tissues (seeds, roots, leaves, stem, floral buds, inflorescences and green siliques) which had been treated with DNase I (Boehringer Mannheim). The RNA was incubated at 37°C for 1 h in 10 mM each dNTP, 100 pmol random hexamers (Promega) and 200 U M-MLV RT (Gibco BRL), in a total volume of 20 µl, followed by incubation in 0.2 mM NaOH for 1 h. After precipitation and dilution in 20 µl, 1 µl of the reaction was used for each PCR. PCR was carried out under standard conditions using 8 pmol of each gene-specific primer and 35 cycles of 95°C for 30 s, 50–62°C for 30 s and 72°C for 30–60 s in a Robocycler Gradient 96 (Stratagene). Products were separated on 1.0% agarose gels and revealed by ethidium bromide staining.

For genes of the other groups, mRNA was isolated from *Arabidopsis* tissues using magnetic oligo(dT) beads (GenoPrep mRNA beads; GenoVision, Norway) according to the manufacturer's instructions. The extracted mRNA, bound to the beads, was used for first strand cDNA synthesis with AMV RT. For each tissue a control reaction was run without RT. PCR using the control reaction as template would reveal DNA contamination in the mRNA. For RT–PCR, 5% of the generated first strand cDNA, and a corresponding amount from the control reaction, was used as template. Specific primers annealing to different exons were designed for each putative gene. RT–PCR products were sequenced using a MegaBACE 1000 sequencer.

For *SUVH4* a gene-specific PCR primer pair fitting the SET domain region was used to clone a genomic gene fragment. This fragment was used to screen a λZapII *Arabidopsis* cDNA library. Two nearly identical cDNA clones of ~1 kb length

were isolated. GENESCAN software and oligonucleotides for RT–PCR were used to lengthen the cDNA sequence up to a consensus start side immediately downstream of an in-frame stop.

Primer sequences used for the above purposes will be provided on request.

5′-RACE and 3′-RACE

The 5′-RACE was done using 5′RACE Kit version 2 (Life Technologies) according the protocol of the manufacturer with 200 ng poly(A)$^+$ RNA from *Arabidopsis* leaves, purified using a mRNA purification kit (Pharmacia). First strand cDNA synthesis was primed with a gene-specific primer. After second strand synthesis, the cDNA was amplified using an anchor oligo and gene-specific primers. 5′-RACE amplification was performed with a nested gene-specific primer.

The 3′-RACE was performed using 200 ng poly(A)$^+$ RNA from leaves and 200 U M-MLV RT (Gibco BRL). First strand cDNA synthesis was primed with a poly(T) primer with anchor sequences. After second strand synthesis, the cDNA was amplified using an anchor oligo and gene-specific primers. 3′-RACE amplification was performed with a nested gene-specific primer. Lists of used primers are available on request.

The resultant PCR products from both 5′- and 3′-RACE analysis were gel-eluted and directly sequenced using a cycle sequencing protocol (Perkin Elmer) and analyzed using an ABI377 sequencer.

Nuclear localization assay

For onion epidermis assays we used the vector pKEx4tr-G (33) containing the 35S* promoter (34) and a fusion between the N-terminal ORF of GUS and the full-length ORF of one of the tested proteins at the C-terminus. For plant GFP fusions, we used plasmid CD3-327 (kindly provided by ABRC Stock Center) as described (35) to amplify smRSGFP (GenBank accession no. U70495) by PCR using primers 5′-GGA TCC CGC ATG AGT AAA GGA GAA G-3′ and 5′-CTC GAG GAG CTC TTA TTT GTA TAG TTC ATC CAT GC-3′. The PCR product was digested with *Bam*HI and *Sst*I and cloned into the *Bam*HI and *Sst*I sites of the vector pKEx4tr-G to exchange the GUS ORF. The resulting vector was called pEKx-327.

GUS and GFP fusion constructs were transiently expressed in onion epidermal cells using a Vacuumbrand Heliumgun 461 essentially as described (36). Briefly, inner epidermal layers obtained from onions (purchased at a local market) were placed on MS basal medium with 30 g/l sucrose, 2% agar, 2.5 µg/ml amphotericin B (Sigma) and 5 µg/ml chloramphenicol. DNA-coated gold particles (1.6 µm gold; Bio-Rad) were briefly vortexed before bombardment. Purified plasmid DNA (0.8 µg) was bombarded onto each sample at a pressure of 700 kPa and a target distance of 9 cm. Petri dishes were sealed with parafilm and incubated for 4–24 h at 28°C in the dark. GUS activity was determined by histochemical staining at 37°C in X-glcUA solution (50 mM $NaPO_4$ pH 7.0, 0.5 mM sodium ferro/ferricyanide, 0.05% Triton X-100 and 3 mg/ml 5-bromo-4-chloro-3-indoxyl-β-D-glucuronic acid, cyclohexyl-ammonium salt; Duchefa). GFP fluorescence was determined by FITC-filtered visual inspection under a laser scanning microscope (Zeiss).

Accession numbers for cDNA sequences

Sequences submitted to GenBank can be found under the following accession nos: AF344444–AF344452 (SUVH1–SUVH9), AF394239 (SUVR1), AY045576 (SUVR2), AF408062 (SUVR4), AF401284 (ATX3), AY049754 (ATX4), AY049755 (ATX5), AF408061 (ATXR7), AF408059 (ASHH1), AF408060 (ASHH3).

RESULTS

Identification of 37 putative *Arabidopsis* genes encoding SET domain proteins

The SET domains from the proteins encoded by the *Drosophila* genes *E(z)*, *trx*, *Su(var)3-9* and *ash1* were used for BLASTP and TBLASTX searches against the *Arabidopsis* non-redundant sequence databases. Proteins encoded by putative genes recognized by the hits in the BLAST searches were identified from the annotations in the databases. In total 37 putative *Arabidopsis* SET domain protein-coding genes (*AtSET*) (Table 1) were found based on an *E* value inclusion threshold of <0.001 compared to one or more of the *Drosophila* SET domains.

The *AtSET* genes can be divided into four classes based on their SET domains

SET domains of putative *Arabidopsis* proteins were aligned with selected proteins from *Saccharomyces cervisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Homo sapiens* using the ClustalX program and manual adjustment with the GeneDoc program. Protein predictions were corrected on the basis of: (i) comparison of gene predictions generated by different programs (GENSCAN, GeneMark and Gene finder); (ii) comparison of duplicated genes (see below); (iii) analysis of protein domains encoded by predicted neighboring genes. In some cases, alignments indicated that exons had been overlooked in the annotations (see Table 1 and text below). These putative exons were added to the predicted proteins when confirmed by the alignments. Predicted exon–intron borders were checked against sequences of cDNAs, RT–PCR products or ESTs when available (see below and Table 1).

The majority of the putative *Arabidopsis* SET domain proteins could easily be fitted into the alignment. However, seven putative proteins contained only parts of the 130–160 amino acid long domain (Fig. 1). In addition, two domains (Fig. 1, ATXR5 and ATXR6) diverged substantially from all the others. A tree based on the alignment of 28 *Arabidopsis* SET domains and 12 such domains of proteins from other species was constructed by the neighbor joining method, using ClustalX (Fig. 2). Bootstrap values >60% are shown.

Three *Arabidopsis* proteins, MEDEA, CLF and EZA1, group together with *Drosophila* E(Z), its human counterpart EZH2 and *C.elegans* MES-2. The tree gives very good support (99.9%) for recognition of the E(Z)-like proteins of all species included, as a distinct group. The E(Z) group of *Arabidopsis* encompass two genes which are already known. The *MEDEA* gene is involved in inhibition of endosperm development in the absence of fertilization (12,13,14,37). Mutations in the *CLF* gene result in altered leaf morphology and also homeotic

Table 1. Putative genes encoding SET domain proteins in *Arabidopsis*

Genes	Protein Accession #	BAC	MIPS code	(chr)	(Mb)	Intr.	Intr. SET	Protein (# aa)	EST cDNA RT
1) E(Z) homologues									
MEA	AAC39446	T14P4.11c	At1g02580c	I	0.50	+	4	689	c
CLF	CAA71599	F26B6.3	At2g23380	II	10.23	+	4	902	c
EZA1	AAD09108	T10M13.3	At4g02020	IV	0.87	+	4	856	+++c
2) trx homologues									
ATX1	AAK01237	T9H9.15	At2g31630	II	13.69	+	5	1062	cr
		T9H9.16	At2g31640						
		T9H9.17	At2g31650						
ATX2	AAF29390a	T20M3.10	At1g05830	I	1.72	+	5	1051	++cr
ATX3	CAB71104a	F15G16.130	At3g61740	III	23.27	+	5	925	+++r
ATX4	CAB36760a	T13J8.20	At4g27910	IV	13.10	+	5	1026	++r
ATX5	BAA97320	MYN8.4	At5g53430	V	21.76	+	5	1040	++r
trx-related									
ATXR1	AAF87042	T24P13.13	At1g26760	I	9.26	1	d	969	+
ATXR2	BAB02844	MSD21.13	At3g21820	III	7.65	+	d	565	++
ATXR3	CAB10297a	ATFCA0	At4g15180	IV	7.25	+	d	2351	++
ATXR4	BAB11410	F15M7.14	At5g06610	V	1,98	+	d	626	+
	BAB11411b	F15M7.15	At5g06620						
ATXR5	CAB89351	F17I14.20	At5g09790	V	2.99	+	2	379	nd
ATXR6	BAB10399	MOP9.18	At5g24330	V	8.32	+	2	349	nd
ATXR7	BAB10481	MDH9.9	At5g42400	V	17.02	+	5	1421	r
3) ash1 homologues									
ASHH1	AAF04434	F28O16.8	At1g76710	I	28.52	+	3	528	+++r
ASHH2	AAC34358	T14N5.15	At1g77300	I	28.73	+	4	1767	++r
ASHH3	AAC23419	F6E13.28	At2g44150	II	18.42	+	4	344	++++r
ASHH4	CAB75815	F24G16.230	At3g59960	III	22.55	+	4	352	nd
ash1-related									
ASHR1	AAD03568	T13L16.8	At2g17900	II	8.02	+	d	447	nd
ASHR2	AAD10162	F3P11.24	At2g19640	II	8.69	0	d	398	+
ASHR3	CAA18207	F6I18.230	At4g30860	IV	14.21	+	3	477	nd
4) Su(var)3-9 homologues									
SUVH1	AAK28966	MUG13.20	At5g04940	V	1.43	0	0	670	+++c
SUVH2	AAK28967	F4P9.6	At2g33290	II	14.38	0	0	651	+c
SUVH3	AAK28968	F3N23.30	At1g73100	I	27.15	0	0	669	+++c
SUVH4	AAK28969	MAC12.7	At5g13960	V	4.52	+	3	624	c
SUVH5	AAK28970	T4C15.17	At2g35160	II	15.06	0	0	794	+c
SUVH6	AAK28971	T9I22.18	At2g22740	II	9.87	0	0	790	+c
SUVH7	AAK28972	F2H15.1	At1g17770	I	6.12	0	0	693	c
SUVH8	AAK28973	F27A10.5	At2g24770	II	10.79	0	0	755	?
SUVH9	AAK28974	T6G15.10	At4g13460	IV	7.02	0	0	650	c
SUVH10	AAC95167	T6P5.10	At2g05900	II	2.30	0	0	341	?
Su(var)3-9-related									
SUVR1	AAD10665	F21M11.1	At1g04050	I	0.94	+	5	734	c
SUVR2	AAK 9218	MRH10.10	At5g43990	V	17.77	+	5	718	c
SUVR3	AAF00642	F20H23.22	At3g03750	III	0.92	1	1	354	++r
SUVR4	AAF63769a	T27C4.2	At3g04380	III	1.16	+	4	477	r
SUVR5	AAC17089	F27L4.5	At2g23730	II	10.33	+	4	1326	?
	AAC17099	F27L4.6	At2g23740						
	AAC17088b	F27L4.7	At2g23750						

Genes were classified according to the similarity of the encoded proteins to the *Drosophila* proteins E(Z), TRX, ASH1 and SU(VAR)3-9 (see Results). Included in the table are EMBL accession numbers of the proteins, BAC clone annotation coordinates, MIPS chromosome localization codes and chromosomal positions based on TAIR map localizations. + under Intr denotes that a gene has more than one intron in the coding sequence. Intr SET shows the number of introns in the regions encoding the SET domains. This number is not given for genes with partially deleted SET domains (d). # aa denotes the number of amino acids in the respective proteins. Presence of matching database ESTs (each + denotes one accession), cDNAs (c) or RT–PCR products (r) is given in the last column. For five genes expression has not been determined by any method (nd). For three genes RT–PCR has not been successful (?). The cDNAs of five genes (*MEA*, *CLF*, *EZA1*, *ATX1* and *ATX2*) have been cloned previously by others (see text).

[a]Proteins have been modified based on alignment, prediction and/or sequencing analysis.

[b]Proteins have been joined and modified based on alignment, prediction and/or sequencing analysis.

[c]Annotated protein additionally covers nnn033.

alterations in flower development (11). The last member in this group is EZA1, for which the mRNA has been cloned (AAD09108).

The tree also gives solid support (93.9%) for the grouping of five *Arabidopsis* proteins in a separate class together with TRX of *Drosophila* and its human (HRX) and yeast (SET1) homologs (Fig. 2). Two of the genes encoding proteins of this class were recently described as *ARABIDOPSIS TRITHORAX 1* and *2* (27). The additional genes we have identified were therefore named *ATX3*, *ATX4* and *ATX5* (Table 1 and Fig. 1). We could also align the SET domain of one ARABIDOPSIS TRITHORAX-RELATED protein (ATXR7) which showed less overall similarity with the ATX proteins (see below). The SET domain of this protein is most similar to that of SET1.

Putative *Arabidopsis* proteins with SET domains more similar to the *Drosophila* ASH1 and SU(VAR)3-9 proteins were also identified (Fig. 2). Four proteins that group most closely together with ASH1 and its yeast homolog SET2 have the SET domain placed in the central region, as do other ASH1 class proteins. The genes encoding them were consequently named *ASH1 HOMOLOG 1* to *ASH1 HOMOLOG 4* (*ASHH1–ASHH4*; Table 1 and Fig. 1). A fifth protein has the SET domain at the C-terminus and was therefore named ASH1-RELATED (ASHR3; Fig. 1) (For ASHR1 and ASHR2 see below).

The remaining *Arabidopsis* SET domains are most closely related to SU(VAR)3-9 and its human (SUV39H) and *S.pombe* (CLR4) homologs. We have called 10 of the encoding genes *SU(VAR)3-9 HOMOLOGS* (*SUVH1–SUVH10*; Table 1 and Fig. 1). The SUVH proteins are clustered in the tree and have a common additional domain (see below). There are high bootstrap values for branches within this group, SUVH1, SUVH3, SUVH7 and SUVH8 (99.6%); SUVH5 and SUVH6 (89.3%); and SUVH2 and SUVH9 (100%). *SUVH10* seems to be a copy of the *SUVH* class that has suffered an internal deletion that removed a part of the region encoding the SET domain.

The proteins encoded by the remaining *SU(VAR)3-9-RELATED* (*SUVR1–SUVR5*) genes are more diverse, with a separate branch for SUVR1, SUVR2 and SUVR4. The SET domains of these three proteins are most similar to that of the human G9a protein (Figs 2 and 3). SUVR4 seems most closely related to SUVR1 and appears to have been generated after a deletion resulting in the removal of nearly 290 amino acids, but leaving the C-terminus, including the SET domain and surrounding cysteine-rich regions, intact.

The relationship between the putative proteins with truncated or deviating SET domains and the *Arabidopsis* SET domain classes were analyzed separately. Six were most similar to the domain of the ATX group and the putative encoding genes were therefore called *ATXR1–ATXR6* (Table 1 and Fig. 1). Two proteins have a truncated domain most equal to the ASHH group and the putative genes were named *ASHR1* and *ASHR2* (Table 1 and Fig. 1).

Distinct cysteine-rich domains are present in the four classes of AtSET proteins

After assignment of the identified proteins to different classes based on their SET domains, characteristics of the other parts of the proteins were investigated by comparison to the homologs of other species and searches in the conserved domain databases Smart, Pfam-A and Profile prosite.

In addition to distinct differences in the amino acid sequence of the SET domain, the four classes of proteins have other significant characteristics (10; Fig. 3). Proteins of the E(Z)

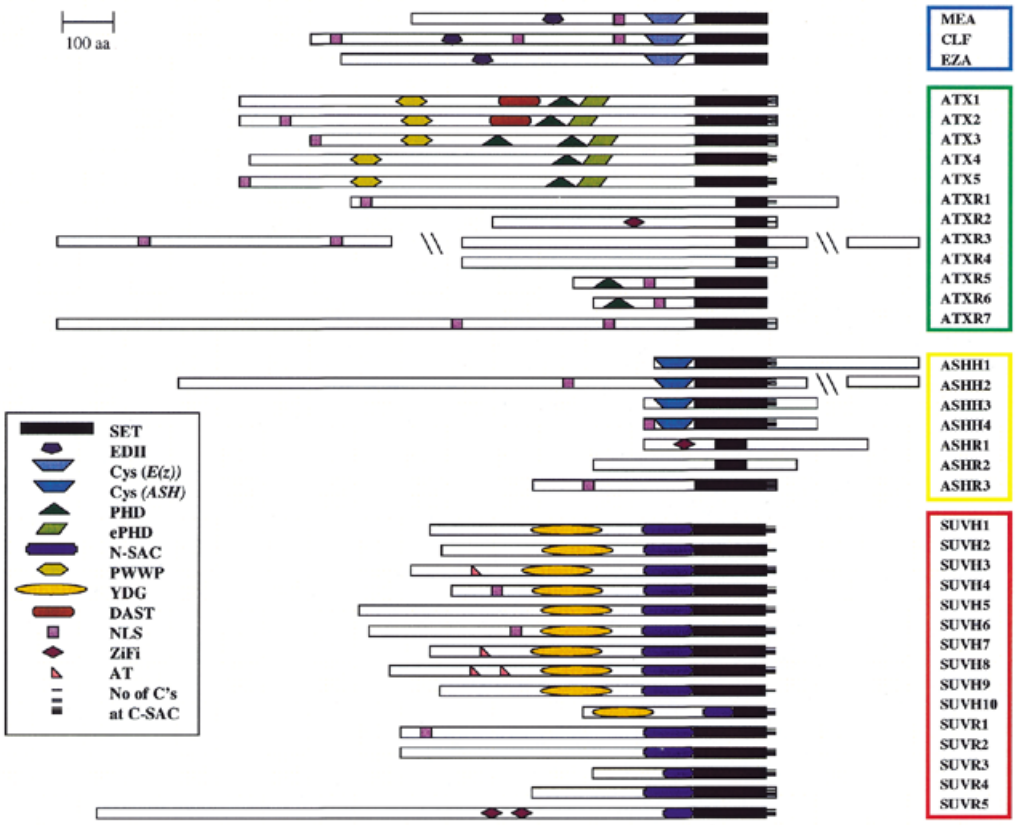


Figure 1. Structure of *Arabidopsis* SET domain proteins. Protein sequences obtained from annotations in the EMBL and MIPS databases, adjusted by ESTs, sequences of RT–PCR products and cDNAs, were analyzed for conserved domains (see Materials and Methods). Lengths of proteins and position of domains are shown to scale except when indicated by \\. SET, SET domain; EDII, E(Z) domain II; Cys (E(Z)), cysteine-rich region found in E(Z) class proteins; Cys (ASH), cysteine-rich region found in ASH1 class proteins; PHD, PHD finger; ePHD, extended PHD finger; N-SAC, N-terminal part of SET-associated cysteine-rich (SAC) region; PWWP, PWWP domain; YDG, YDG domain; NLS, bipartite nuclear localization signal; ZiFi, zinc finger; AT, AT-hook; one, two or three horizontal lines indicate the number of cysteines in the C-terminal SAC. The Cys-rich domain of ASHH3 is not significant according to the domain searches, but aligns well with the Cys-rich domains of the other ASHH proteins.

class have a region with 16–18 cysteine residues spaced in a given pattern in front of the C-terminal SET domain. Proteins of the SU(VAR)3-9 class have a SET domain-associated cysteine-rich region (SAC) with seven to eight cysteines in certain positions in front of the SET domain (N-SAC) and three C-terminal cysteines in the pattern $CXC(X)_4C$ (C-SAC) after the SET domain. The C-SAC is also found in TRX class proteins, which lack a cysteine-rich region N-terminal to the SET domain. ASH1 class proteins have, in contrast to the other three classes, the SET domain centrally placed. Their SET domains are preceded by a cysteine-rich region and followed by the C-SAC pattern. The number and spacing of the cysteine residues in the N-terminal cysteine region differ from that of the E(Z) C-rich region and also from the N-SAC.

The similarities between the E(Z), MEA and CLF proteins have been recognized previously (11,12). In addition to the SET domain, these proteins have a C-rich stretch and the so-called domain II in common with the E(Z) proteins of other organisms (Figs 1 and 3). The C-rich region is also present in the protein encoded by *EZA1* (Figs 1 and 3).

All the ATX proteins have a complete C-SAC motif, while the ATXR proteins lack at least one of the C-terminal cysteines (Figs 1 and 3). All four ASHH proteins have the C-SAC motif and the cysteine-rich domain conforming to ASH1 class proteins (Figs 1 and 3). The ASHR proteins lack the C-rich regions with the exception of the C-SAC motif of ASHR3 (Figs 1 and 3).

A complete SAC domain is present in 10 SUVH and SUVR proteins (Figs 1 and 3). Truncated N-SAC regions are present in SUVR3, SUVR5 and SUVH10, while SUVH2 and SUVH9 only have one C-terminal cysteine residue.

A novel domain called the YDG domain is present in SUVH proteins

Alignments of the N-terminal part of the SUVH proteins and the use of domain-finder programs revealed that these 10 proteins contain a conserved domain also found in non-SET proteins containing a RING finger motif (in mammals and *Arabidopsis*) or an HNH nuclease motif (in the bacteria *Deinococcus radiodurans*) (Fig. 4A). We have chosen to call the 150–170 amino acid long region the YDG domain because of a characteristic YDG motif. Further characteristics of the domain are the conservation of up to 13 evenly spaced glycine residues and a VRV(I/V)RG motif (Fig. 4A).

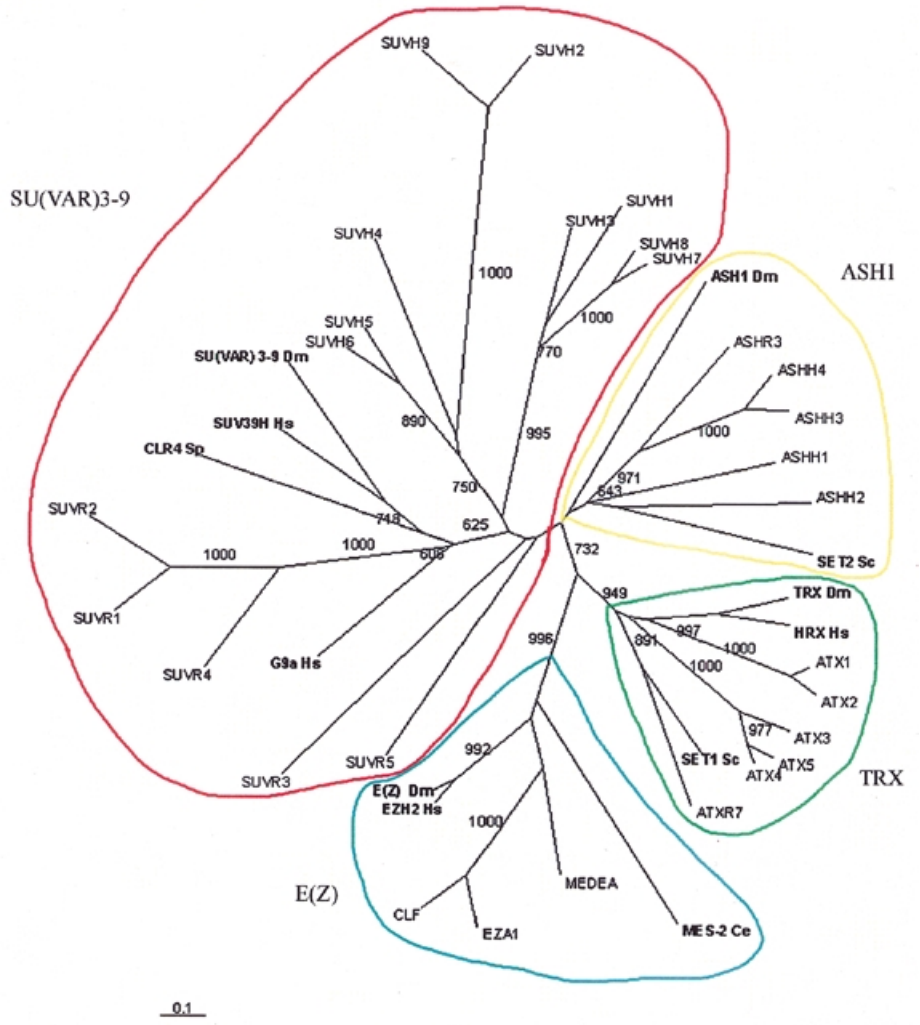


Figure 2. Relationship between SET domain proteins of *Arabidopsis* and other organisms. The tree was constructed using the ClustalX program based on alignments of SET domains by ClustalX and manual adjustment. Figures indicate bootstrap values (1000 = 100%). Values >60% are shown. E(Z), *Drosophila* E(Z), P42124; EZH2, human E(Z) homolog 2, Q15910; MES-2, *C.elegans* maternal effect sterile 2 E(Z) homolog, AAC27125.1; TRX, *Drosophila* TRX, P20659; HRX, human TRX homolog, Q03164; SET1, *S.cerevisiae* TRX homolog, NP_011987.1; ASH1, *Drosophila* ASH1, AAF49140.2; SET2, *S.cerevisiae* ASH1 homolog, YJL168c; SU VAR 3-9, *Drosophila* SU(VAR)3-9, P45975; SUV39H, human SU(VAR)3-9 homolog, AAF06805.1; G9a, human SET domain protein, NP_006700; CLR4, *S.pombe* SU(VAR)3-9 homolog, T43700.

PWWP domains, PHD fingers and extended PHD fingers are found in ATX proteins

In all the ATX proteins the PWWP motif, first identified in the human protein WHSC1 (38), was found (Fig. 4C). WHSC1 is most closely related to the ASH1 class of SET domain proteins, but we did not identify this domain in any of the *Arabidopsis* ASHH or ASHR proteins. The PWWP domain is present in a diverse groups of nuclear proteins (38) and typically has conserved PWWP residues. The first proline residue is present in three of the five ATX proteins, but none of them contain the first of the two tryptophans. The most conserved motifs are GDΦΦWXK (where Φ are hydrophobic residues), WPAΦΦΦD and VXFFG (Fig. 4C).

In four of the five ATX proteins, as well as in ATXR5 and ATXR6, we identified amino acid motifs similar to the PHD finger (Figs 1 and 4B) found in the *Drosophila* and mammalian TRX/HXR proteins and a number of other nuclear proteins (26). The characteristic C_4-H-C_3 pattern is present once or twice in the *Arabidopsis* proteins. In the ATX proteins the PHD fingers are situated about midway between the PWWP motif and the SET domain.

Finally, the ePHD motif (25) was found in all the ATX proteins, positioned just after the PHD finger (Fig. 4D). The second half of this motif resembles a PHD finger (compare conserved cysteine and histidine residues).

The DAST motif recently identified in ATX1, ATX2, TRX and HRX (27) was not found in the other ATX proteins.

Putative nuclear localization signals and AT-hooks are found in many AtSET proteins

The domain-finder programs recognized putative bipartite nuclear localization signals (NLSs; see Fig. 1) in MEA and CLF, but not in EZA1. Among the ATX and ATXR proteins one or two such NLSs were identified in all proteins but four (ATX1, ATX4, ATXR2 and ATXR4). This signal was also

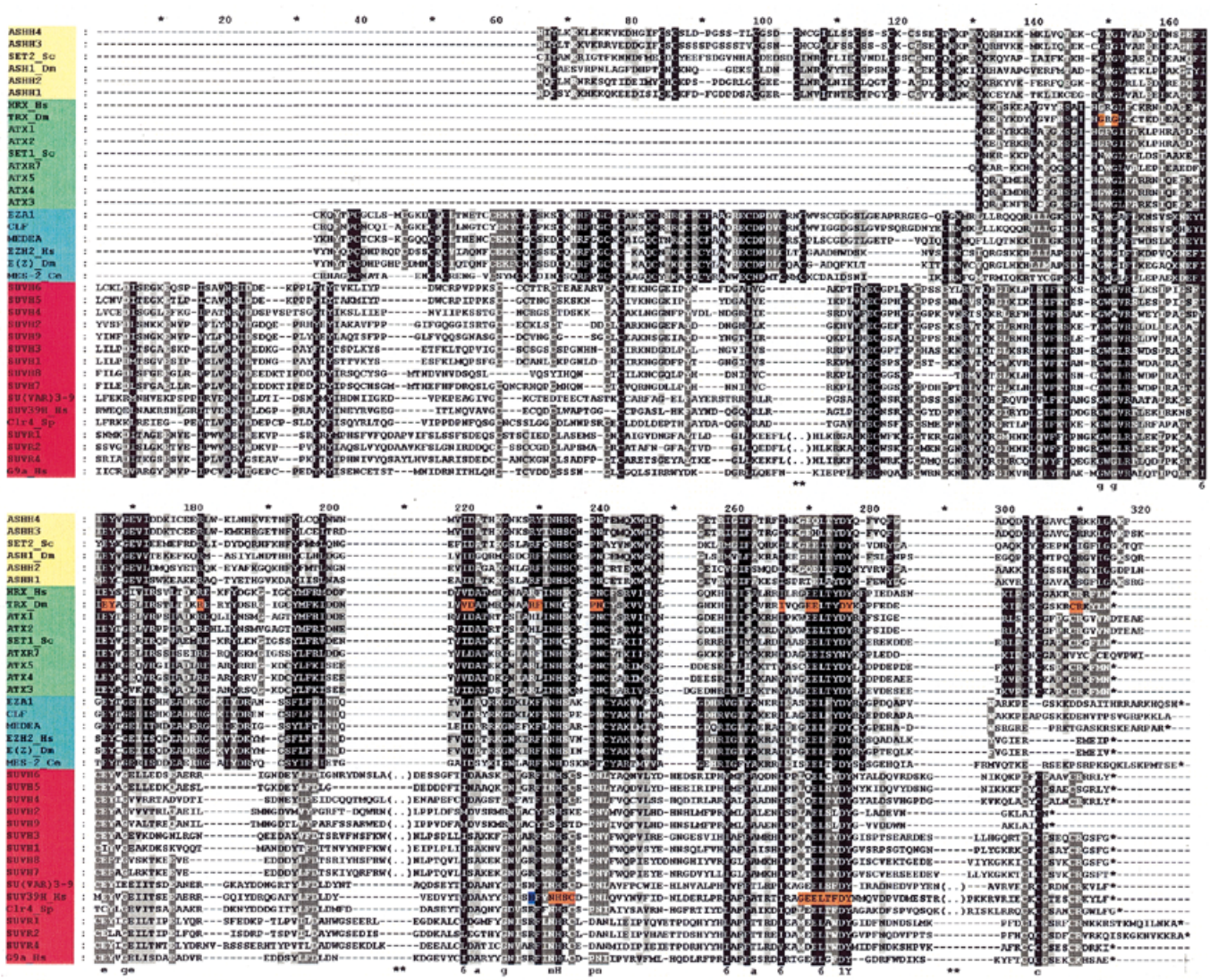


Figure 3. Alignment of SET domains and flanking cysteine-rich regions of the four classes of SET domain proteins. The SET domains of all proteins are perfectly aligned from the GWG motif (positions 149–151), while the cysteine-rich domains N-terminal to the SET domains are aligned within each group to show class characteristics in this region. The TRX class lacks such a region. Note also the C-SAC motif from position 300, which is lacking in the E(Z) class. The degree of conservation is distinguished at four levels (100, 80 and 60% and not conserved), where 100% has the darkest shade of gray. The upper and lower case letters in the consensus line indicate 100 and 80% conservation within all groups, respectively. Numbers in the consensus line represent conserved similarity groups as defined by the Blosum 62 scoring table. Yellow, ASH1 class proteins; green, TRX class proteins; blue, E(Z) class proteins; red, SU(VAR)3-9 class proteins; orange, residues that when mutated abolish self-association and the SNR1 interaction of the TRX SET domain and loss of HMTase acitivity of SUV39H (20,47); dark blue, H residue which when changed to R increases HMTase activity of SUV39H (20). (..) indicates that short stretches of non-conserved amino acids were omitted from sequences in the SU(VAR)3-9 class, in regions marked ** below the alignment, so as to fit the figure on one page.

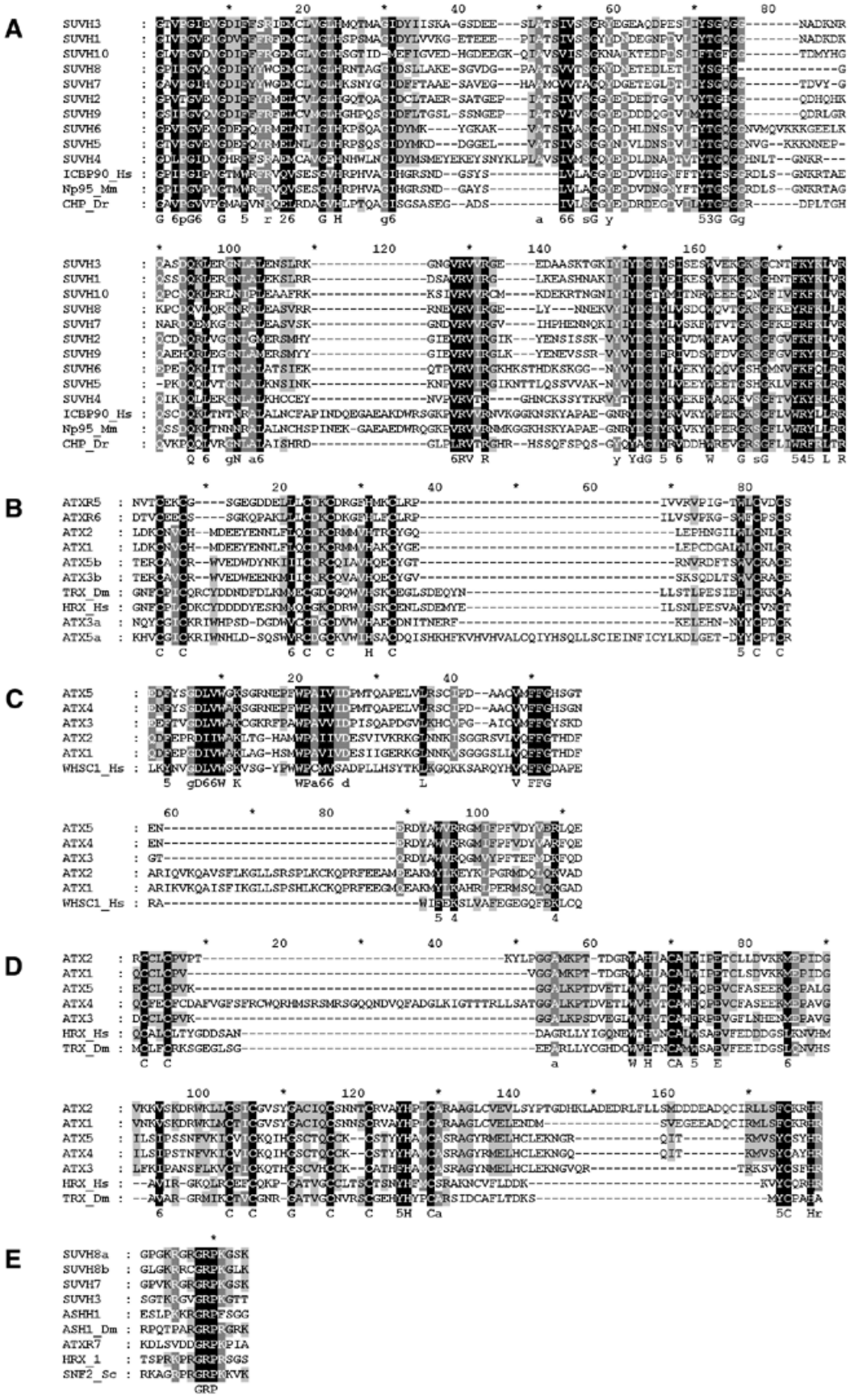
found in three proteins of the ASH groups (ASHH2, ASHH4 and ASHR3) and three proteins in the SUV groups (SUVH6, SUVH4 and SUVR1). We cannot exclude the possibility that other types of NLSs are present in the other proteins.

Putative AT-hooks, which mediate protein binding to the minor groove of AT-rich tracts in DNA (39), were identified in three of the SUVH proteins, in ASHH1 and in ATXR7 (Figs 1 and 4E). This motif has a characteristic GRP core.

Pairs of *AtSET* genes are found in large genomic duplications

The chromosomal positions of the 37 putative genes encoding SET domains are spread over all five *Arabidopsis* chromosomes (Table 1). The MIPS Interactive Redundancy Viewer was used to investigate whether any of these genes were positioned in duplicated regions of the genome. Five likely gene pairs were found: *MEA* and *EZA1* seem to be part of a large duplication between chromosomes I and IV; *ATX1* and *ATX2* belong to a duplication between chromosomes I and II; *ATX4* and *ATX5* are found in a duplication on chromosomes IV and V; *ASHH3* and *ASHH4* are found in a duplication on chromosomes II and III; and *SUVH3* and *SUVH7* are found in a duplication on different regions on chromosome I (Table 1). In addition, *SUVR1* is in an area on chromosome I that shares duplicated regions with the area on chromosome V where *SUVR2* is positioned.

In all cases, members of gene pairs belonged to the same class of SET domain genes. The encoded proteins were


A
SUVH3
SUVH1
SUVH10
SUVH8
SUVH7
SUVH2
SUVH9
SUVH6
SUVH5
SUVH4
ICBP90_Hs
Np95_Mm
CHP_Dr
B
ATXR5
ATXR6
ATX2
ATX1
ATX5b
ATX3b
TRX_Dm
HRX_Hs
ATX3a
ATX5a
C
ATX5
ATX4
ATX3
ATX2
ATX1
WHSC1_Hs
D
ATX2
ATX1
ATX5
ATX4
ATX3
HRX_Hs
TRX_Dm
E
SUVH8a
SUVH8b
SUVH7
SUVH3
ASHH1
ASH1_Dm
ATXR7
HRX_1
SNF2_Sc

compared pair-wise and could be aligned along their total lengths (data not shown). The positions of annotated exons and introns in gene pairs were also very similar. Twelve introns of 16 were in identical positions in *MEA/EZA1*, 20 of 23 in *ATX1/ATX2*, all 20 introns in *ATX4/ATX5* and eight of the 11 and 10 introns in *ASHH3* and *ASHH4*, respectively.

The majority of SUVH group ORFs are intronless

The number of annotated introns in all the *Arabidopsis* proteins and their positions in the SET domain were compared. In the majority, numerous introns are present and up to five were found within the SET domain (Table 1). In the *Arabidopsis* E(Z) class genes the positions of these introns are conserved. However, intron positions differ from those in the E(Z) proteins of other species (data not shown). For the other classes, identical intron positions are only found between closely related pairs of genes (cf. above).

In contrast to the majority of genes, *ATXR1* and *SUVR3* contain one intron only, which for *SUVR3* is found in the SET domain-encoded region. Among the 10 *SUVH* genes all but *SUVH4* have intronless ORFs (Table 1). In contrast, the *SUVH4* gene contains 13 introns.

The *AtSET* genes are active

For each of the putative *AtSET* genes different databases were examined for the presence of matching ESTs and cDNA sequences (Table 1). As mentioned above, the cDNAs from the genes in the *E(z)* class and recently also two *ATX* genes have been cloned by others (11,12,27). RT–PCR and cDNA cloning were used to verify expression of additional *AtSET* genes. cDNAs for *SUVH1* and *SUVH5* (Fig. 5A), and *SUVH7* and *SUVH9* (not shown) confirmed these genes as being intronless and showed that an annotated intron in the genomic region corresponding to *SUVH7* (F2H15.1) is not spliced out. This results in an ORF encoding a protein containing the C-SAC motif but is shorter (693 amino acids) than the annotated gene (954 amino acids).

For *SUVH2*, 5′-RACE and RT–PCR showed no intron in the leader sequence and the putative ORF, but an intron of 83 bp in the trailer of the transcript (Fig. 5A). *SUVH3* has matching ESTs (AA728521, AI998299 and T04123) which together with RT–PCR could be extended to an almost complete cDNA containing the expected intronless ORF. However, in the leader sequence of the *SUVH3* gene there are two introns of 464 and of 111 bp (Fig. 5A).

The *SUVH4* transcript, identified from two λZAP cDNA clones and by RT–PCR, consists of 2.1 kb and sequencing confirmed the presence of 13 introns (Fig. 5A). Expression of *SUVH1*, *SUVH2*, *SUVH3*, *SUVH5* and *SUVH6* is supported by the presence of matching ESTs in the databases generated from rosette leaves, roots and/or developing seeds (Table 2). Expression of five *SUVH* genes was detected by RT–PCR in seeds, roots, leaves, stems, flowers and/or siliques (Fig. 5A and Table 2). Only *SUVH1* seems to be expressed in roots. We did not succeed in RT–PCR amplification of *SUVH8* and *SUVH10* in any tissues tested. Analysis of the DNA sequence upstream of the annotated *SUVH10* gene indicates that this is a gene that has been inactivated by mutations. The database protein sequence of SUVH10 (T6P5.10) starts just inside the YDG domain (see Fig. 4A). However, this domain would be completely contained in SUVH10 if 1 nt was inserted 33 nt before the start codon. This would lengthen the putative ORF of about 279 nt (including a full YDG domain).

Primers designed to investigate whether *SUVR1*, *SUVR2*, *SUVR3* and *SUV4* were expressed, successfully amplified RT–PCR products that were shorter than their genomic counterparts due to the presence of introns (Fig. 5B). Expression of *SUVR3* is further confirmed by corresponding ESTs. An additional intron in the C-terminal part of the *SUVR2* gene changes the amino acids between the C-SAC and the stop codon, as compared to the annotated protein sequence (MRH10.10). Sequence analysis of *SUVR4* revealed the omission of an exon in the C-terminal region of the annotated protein (T27C4.2). This exon contains the C-SAC motif and renders the protein 477 amino acids long, not 424 as annotated. Two sets of *SUVR5* primers were designed, but none of these produced any RT–PCR product. This putative gene is annotated either as one large gene (see Fig. 1) or three separate genes of which one contains only the SET domain (see Table 1).

In the ATX group, ESTs matching four genes have been cloned from developing seeds and aerial organs (Table 2). EST sequences are also found which correspond to the four *ATXR* genes with truncated SET domains (Table 1). These are from inflorescences and aerial organs (Table 2). Expression of the *ATX1*, *ATX2*, *ATX3*, *ATX4*, *ATX5* and *ATXR7* genes was confirmed by amplification of their central parts by RT–PCR using mRNA from different organs (Fig. 5B and Table 2). Our RT–PCR products and RACE verified that the annotations of ATX1 and ATX2 are not in agreement with the cDNA sequences, as noted recently (27). In the EMBL database, the region encoding the *ATX1* transcript is annotated as three separate genes (T9H9.15, T9H9.16 and T9H9.17). The *ATX2* gene (T20M3.10) is annotated with two additional exons at the C-terminus, resulting in an amino acid extension after the C-SAC which would not agree with the notion that TRX class proteins have the C-SAC at their C-terminus. For ATX3, sequence analyses revealed the presence of a GWG motif at the beginning of the SET domain, seven additional amino acids in the ePHD domain and an exon encoding 25 amino acids in the SET domain, in contrast to the annotated protein (F15G16.130). Two exons missing in *ATX4* were revealed by comparison to the matching EST AV524242. The annotated ATX4 protein (T13J8.20) terminates just after the NHSC motif in the SET domain. The additional two exons (T13J8.30) extend the C-terminal end of the protein so as to give a complete SET domain and C-SAC.

Figure 4. (Opposite) Alignment of domains found in SET domain proteins. (**A**) YDG domain. Note that the first six amino acids (GLVPGV) of SUVH10 are from another reading frame, followed by 11 amino acids (DVGDIFFFRGE) from the same frame as the annotated ORF (T6P5.10). HsICBP90, *Homo sapiens in vitro* CCAAT-binding protein 90, AAF28469.1; MmNp95, *Mus musculus* nuclear binding protein 95, AAK55743.1; DrCHP, *Deinococcus radiodurans* conserved hypothetical protein, AAC28190. (**B**) PHD fingers. (**C**) PWWP domain. WHSC1, human WHSC1 protein, DD19343. (**D**) ePHD fingers. (**E**) AT-hooks. For shadings and consensus line see Figure 3.

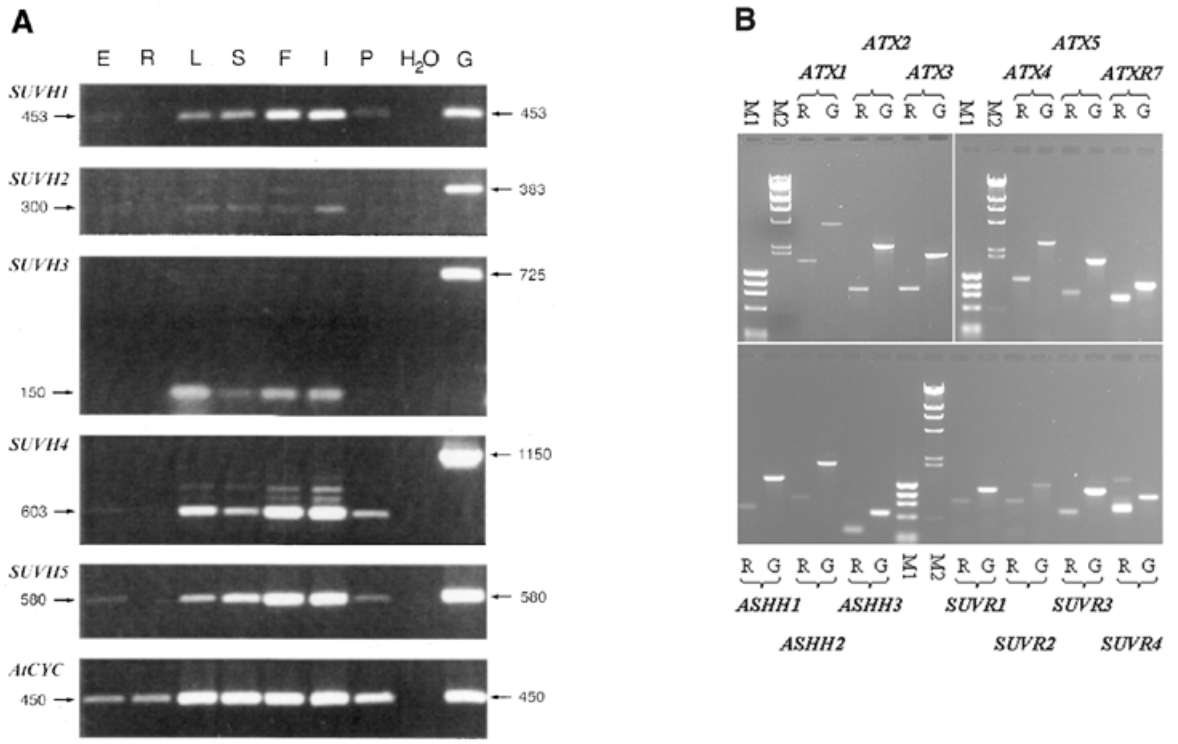


Figure 5. RT–PCR expression analyses. (**A**) Agarose gels stained with ethidium bromide showing cDNA fragments of *SUVH1*, *SUVH2*, *SUVH3*, *SUVH4*, *SUVH5* and *AtCyclophilin* (positive control; 63) amplified by RT–PCR using gene-specific primers. RT–PCR reactions were performed on DNase I-treated total RNA isolated from seeds (E), roots (R), leaves (L), stems (S), floral buds (F), inflorescences (I) and green siliques (P). A negative (H_2O) and a positive (genomic DNA, G) control reaction are shown to the right of the RT–PCR reactions. The PCR fragment sizes are given on both sides in bp. Note the intronless fragments of *SUVH1* and *SUVH5*. The PCR primers for *SUVH2* and *SUVH3* were designed to amplify their 3′- and 5′-UTR, respectively, where introns are found in the genomic sequences (see text). (**B**) Agarose gels showing RT–PCR fragments (R) of selected *ATX*, *ATXR*, *ASHH* and *SUVR* mRNAs, amplified by gene-specific primers. RT–PCR reactions were performed on mRNA isolated from floral buds using magnetic oligo(dT) beads. Note that each genomic fragment (G) is longer than the corresponding RT–PCR fragment obtained with the same primers due to the presence of introns. Size markers are ΦX174 DNA digested with *Hae*III and λ DNA digested with *Hin*dIII.

ESTs matching *ASHH1*, *ASHH2* and *ASHH3* have been found in developing seeds (Table 2). RT–PCR using mRNA from different tissues (Fig. 5B and Table 2) confirmed that these genes are active.

Nuclear localization

To show that the gene products of the SET domain coding transcripts are nuclear proteins, as already shown for all functionally described SET domain proteins to date (see for example 4,10), transient expression assays were used (34) (see Materials and Methods). Constructs containing an in-frame fusion of the GUS gene, or a red-shifted GFP gene, and a cDNA encoding *CLF*, *SUVH1*, *SUVH2* or *SUVH3* in a transient expression vector, were shot into the inner epidermis of onions using a particle gun. Whereas the GUS protein alone is not localized to the nucleus, all of the fusion protein variants became concentrated in the nucleus (data not shown). The transiently expressed GFP fusions confirmed that nuclear transport was not an artifact of the test system: in contrast to GFP alone, all proteins became concentrated in the nucleus (Fig. 6A–E). *Drosophila* SU(VAR)3-9 also showed nuclear localization in onion cells (Fig. 6F).

DISCUSSION

Arabidopsis has at least 29 expressed SET domain genes

Our search has identified 30 putative genes in the *Arabidopsis* genome containing a conserved SET domain and seven with truncated SET domains (Fig. 1). At least 29 of these are expressed genes, as demonstrated by the cloning of cDNAs for 15 genes, RT–PCR products for 11 genes and/or the presence of matching ESTs for 18 genes (Table 1). We failed to amplify RT–PCR products for three potential genes, *SUVH8*, *SUVR5* and *SUVH10*. Despite the long open reading frames, and for SUVH8 intactness of all conserved domains, these genes may represent pseudogenes. Alternatively, their expression may be very low or restricted to untested tissues and developmental stages.

SUVR5 and *SUVH10* seem to have suffered deletions affecting the regions encoding the N-SAC and SET domains (Fig. 3). Other genes encoding divergent proteins are not pseudogenes (Fig. 5 and Table 2). *SUVR3* mRNA is present in buds and seeds and *SUVR4* is expressed in all tissues tested. The presence of ESTs for the *ATXR1–ATXR4* and *ASHR2* genes confirm their active status (Table 1).

Table 2. Expression pattern of genes encoding SET domain proteins in *Arabidopsis*

	roots	rosettes	cauline	stems	buds	flowers	siliques	seeds
ATX1	+	+	+	+	+	+	+	+
ATX2	+	E1		+	+	+	E+	+
ATX3		E1		+	+	+	E+	+
ATX4		E1+			+	+	E	+
ATX5		E2+	+	+	+	+	E	+
ATXR1	?	?	?	?	E3	E3	?	?
ATXR2	?	?	?	?	E3	E3	?	?
ATXR3	?	E1	?	?	?	?	E	?
ATXR4	?	E, E2	?	?	?	?	?	?
ATXR7				+	+		?	+
ASHH1	+				E+	E+	E	
ASHH2	+				+		E	
ASHH3	+		+	+	+	+	E	+
ASHR2	?	E1	?	?	?	?	?	?
SUVH1	+	+	?	+	+	+	+	+
SUVH2		+	?	+	+	+		
SUVH3		+	?	+	+	+		
SUVH4		+	?	+	+	+	+	+
SUVH5		+	?	+	+	+	+	+
SUVH6	?	?	E	?	?	?	?	?
SUVH7	?	+	?	?	?	?	?	?
SUVH9	?	+	?	?	?	?	?	?
SUVR1	?	+	?	?	+	+	?	+
SUVR2	?		?	?	+	+	?	+
SUVR3					E3+	E3	E	+
SUVR4	+	+	+	+	+	+	?	+

The table shows in gray shading tissues in which there is evidence of expression of the respective genes. + denotes tissues were RT–PCR products were amplified using gene-specific primers (this study and for ATX1 and ATX2 also Alvarez-Venegas and Avramova; 27). E denotes tissues from which matching ESTs have been cloned (E1, above ground organs from 2–6-week old plants; E2, seedlings; E3, inflorescence). ? indicates tissues where expression has not been determined.

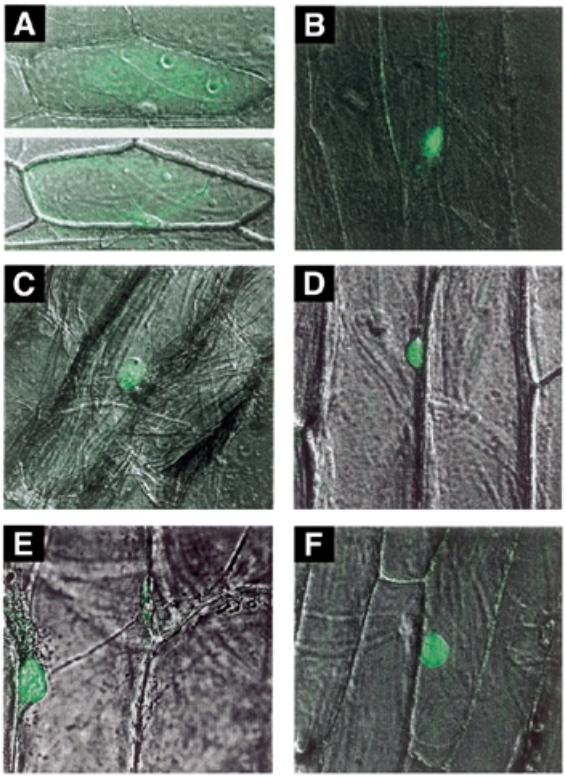


Figure 6. Nuclear localization of SET domain proteins in onion epidermis transient expression assay with the plant GFP reporter system. Histochemical localization of GFP activity following bombardment of onion epidermal cell layers with DNA constructs expressing either GFP alone (**A**), a fusion of CLF to GFP (**B**), a fusion of SUVH1 to GFP (**C**), a fusion of SUVH2 to GFP (**D**), a fusion of SUVH3 to GFP (**E**) and a fusion of *Drosophila* SU(VAR)3-9 to GFP (**F**) is shown. GFP fluorescence was revealed 2 h after bombardment utilizing Nomarski optics.

Arabidopsis has a remarkably high number of active genes encoding SET domain proteins compared to the number known in other organisms.

The *AtSET* genes are expressed in a diversity of tissues

In situ hybridization studies have shown *CLF* to be expressed in leaves, vasculature and meristem, as well as in buds and flowers (11). *In situ* and promoter–GUS expression studies revealed *MEA* expression restricted to the female gametophyte and young developing seeds (13,14). We have used RT–PCR to gain a first overview of the expression patterns of the novel SET domain genes analyzed. This method does not allow a quantitative comparison of expression levels between tissues, but demonstrates that each of the 20 active genes tested is expressed in more than one tissue, that all are expressed in floral buds and the majority in flowers and seeds (Fig. 5 and Table 2). For only a few genes were RT–PCR products generated from roots. A confirmation of this general expression pattern is that the matching ESTs were derived from cDNA libraries made from inflorescences, developing seeds and aerial organs (Table 2).

The AtSET proteins can be assigned to evolutionarily conserved classes

We have investigated the relationship between the AtSET proteins and their homologs in other organisms. The tree based on alignment of SET domains (Fig. 2) indicates that these genes fall into the four classes which were previously recognized in *Drosophila*. In *Arabidopsis* we found three *E(z)* class genes, five *trx* homologs, four *ash1* homologs and 15 genes similar to *Su(var)3-9*. In addition, seven genes have been assigned as *trx*-related and three as *ash1*-related (Fig. 1).

The significant similarity between the SET domain proteins of *Arabidopsis* and their counterparts in other organisms is evident not only from the SET domains but also from the upstream cysteine-rich regions in the E(Z), ASH1 and SU(VAR) classes (Fig. 3). The C-SAC is found downstream of the SET domain in all ATX, ASHH, SUVH and SUVR proteins, as in their animal and yeast counterparts, with the exceptions of SUVH2 and SUVH9 (Figs 2 and 3).

The classification is further substantiated by the internal class similarities in the N-terminal parts of the *Arabidopsis* proteins, notably the common domain II in the E(Z) proteins, the PHD fingers and PWWP and ePHD motifs in the ATX proteins, and the YGD domain in the SUVH proteins, (Figs 2 and 4).

Two different mechanisms may have contributed to the high number of SET domain genes in *Arabidopsis*

Analyses of sequenced genomes have shown that *Arabidopsis* in general has more gene families with many members compared to other organisms. In *Arabidopsis*, 37.4% of the gene families have more than five members (32). It is assumed that *Arabidopsis* had a tetraploid ancestor and that segments have been lost gradually, resulting in the present day genome where duplicated regions encompass 60% (32).

We found five pairs of SET domain genes localized in five different large genomic duplications (Table 1). The members of the duplicated pairs are highly similar along their whole lengths and the majority of the introns and their positions are conserved. The members of the two ATX pairs and the ASHH pair group closely together when aligning SET domains of all the AtSET proteins (Fig. 2). Although *EZA1* and *MEA* reside in duplicated regions, judged from amino acid similarity EZA1 is more closely related to CLF than to MEA. Therefore, *CLF* is likely to be a more recent duplication of *EZA1*. Likewise, *SUVH3* and *SUVH7* may represent a pair of a large duplication from which new genes have been generated, since SUVH3 seems more closely related to SUVH1 and SUVH7 is more closely related to SUVH8 (Fig. 2). It is remarkable that the significant nucleotide sequence similarity between *SUVH7* and *SUVH8* continues upstream and downstream of the coding regions of both ORFs. This homology also exceeds the 23 nt long poly(A) stretch which is found 73 bp downstream of the stop codon of the *SUVH8* ORF. The hypothesis that *SUVR1* and *SUVR2* have arisen from a duplication event is likewise substantiated by the presence of highly similar ORFs upstream of both genes. The multiplicity of *AtSET* genes is not necessarily surprising, since gene duplications occur at a relatively high rate (40), especially in angiosperms (see for instance 41).

The majority of the *AtSET* genes contain a large number of introns, both in the SET domain itself and the rest of the gene (Table 1). In sharp contrast, all the *SUVH* genes are intronless in their coding regions, except for *SUVH4*. This suggests at least one retrotransposition event of an *SUVH4*-like gene transcript during evolution of the *SUVH* gene group. A promoter must accidentally have been recruited or evolved 5′ of this retrotransposition(s). Interestingly, introns are found in the transcribed but non-coding regions of the *SUVH2* and *SUVH3* genes, as is evident from the different sizes of the RT–PCR and genomic fragments generated by the same primers (Fig. 5A). The intron-containing UTRs of *SUVH2* and *SUVH3* may stem from the insertion point rather that from the reverse transcribed mRNA. Another possibility could be that these introns have evolved or have been inserted after generation of the new genes.

In mammals many cases of intronless processed genes have been reported, including the human splicing factor gene Srp46 (which was generated by retrotranspositon from the PR264/SC35 splicing factor gene; 42), several genes encoding proteins with RING and C (3) zinc finger motifs (43), the gene for the testes-specific poly(A) polymerase mPAPT (44), the HMGN4 protein encoding gene (45), two 1-Cys peroxiredoxin encoding genes of mouse (46), the mouse *U2af1-rs1* gene (47) and the genes for the RBM12 and αCP-1 RNA-binding proteins (48,49). The human *αCP-1* gene was generated by retrotransposition of a fully processed *αCP-2* mRNA before mammalian radiation. Stringent structural conservation and ubiquitous tissue expression of *αCP-1* indicates its rapid recruitment to a distinct cellular function (49). Retrotransposition has also been reported for glycerol kinase-related genes (50). The human glycerol kinase gene family consists of an X-linked locus and several X-linked and autosomal intronless genes. The intronless genes comprise both functional genes and pseudogenes. Similarly, the mammalian CDYL gene transcript has been reverse transcribed and inserted into the simian Y chromosome, resulting, after amplification, in the two testis-specifically expressed CDY1 and CDY2 genes (51). It could also be shown that some of the expressed 5S rRNA genes in the mouse and rat were derived from retrotransposition of 5S rRNA transcripts (52).

All these examples demonstrate that functional processed genes can occasionally be generated from retrotransposed fully processed mRNA transcripts and that these processed genes can take on a non-redundant essential functional role. The indications of genomic duplications of *SUVH* genes discussed above suggest a similar route of evolution: for these genes one or a few retrotransposition events were followed by conventional duplication. Comparison of gene number and order in related species, e.g. *Brassica napus*, may reveal how and when the *SUVH* genes evolved.

AtSET proteins can be found in the nucleus

SET domain proteins of other species have a nuclear localization (4,10). The human homolog to ASH1, huASH1, has in addition been localized to tight junctions between cells (28). The amino acid sequences of the AtSET proteins give some indications of nuclear localization. First, sequences with significant similarity to bipartite NLSs were found in 16 putative proteins (Fig. 1). Secondly, AT-hooks, which promote protein binding to the minor groove of AT-rich tracts of DNA (53), were found in five proteins.

Using protein–marker fusion constructs, we have investigated the localization of one protein (CLF) with three putative NLSs, two (SUVH1 and SUVH2) without recognizable bipartite NLSs and one with an AT-hook (SUVH3). For all four proteins, as well as *Drosophila* SU(VAR)3-9, nuclear localization was demonstrated (Fig. 6). Clearly, the absence of a bipartite NLS, as recognized by the protein domain databases, is not a strong prediction for non-nuclear localization. In addition to bipartite NLSs, single and multiple continuous as well as double and multiple bipartite NLSs have been recognized in plants (54).

E(z) is required continuously through development in order to maintain homeotic gene repression in *Drosophila* (55). MEA is suggested to be involved in the repression of one or more seed development genes and in maintenance of the repressed state (13). Similarly, the *CLF* gene is required to repress expression of *AGAMOUS* in hypocotyls, cotyledons, leaves and inflorescence stems and petals (11). The demonstrated nuclear localization of CLF would be a prerequisite for direct involvement in repression of gene expression.

Domains in ATX and ASHH proteins are possibly involved in protein–protein interactions

The *trx* and *ash1* genes of *Drosophila* are involved in the transcriptional memory mechanisms that maintain homeotic gene activity in appropriate segments throughout development (5). The ATX proteins share a PHD finger and the ePHD with TRX and HRX. However, in the latter proteins the ePHD motif is separated from the classical PHD fingers. The position of the ePHD adjacent to a PHD finger is more similar to the situation in the HRX partner genes AF10, AF17 and CBP (25). The amino acid sequences of their ePHDs have extensive similarity to those of the ATX proteins. The ePHD of AF10 has the ability to mediate oligomerization (25).

The SET domains of TRX and ASH1 have been shown to self-associate (56). In both cases mutation of the GXG residues

(Fig. 3, positions 149–151) to VXV resulted in abolition of self-association. The two G residues are conserved in all ASHH proteins. The second of these G residues is conserved in all ATX but ATX3, while the first is not conserved in ATX3 and ATX5, nor in SET1 of yeast, which is known to not self-associate (56). The conserved arginine in the RΦINHSC motif (Fig. 3, positions 229–235) is replaced by a histidine in ATX1 and ATX2. Other positions necessary for self-association are all conserved in the ATX and ASHH1 homologs (Fig. 3). In TRX these amino acids are also needed for interaction with SNR1 (56), a constituent of the SWI–SNF complex, which acts to remodel chromatin and thereby activate transcription. The conservation of these amino acids indicates the capability of the *Arabidopsis* proteins for similar interactions. The SNR1 interaction of TRX is not shared with ASH1, E(Z) and SU(VAR)3-9 class proteins.

The SUVH and SUVR SET domains are similar to such domains conferring histone 3 methyltransferase activity

The SET domain of yeast CLR4, the mammalian SU(VAR)3-9 homologs and human G9a protein have H3-specific histone methyltransferase (HMTase) activity (20–22). Such activity has not been demonstrated for E(Z) and TRX, which lack the C-SAC and N-SAC domain, respectively. There are marked differences between the conserved sequences of the SET domains of the four classes (Fig. 3, positions 198–337). Mutational analyses of the mammalian SU(VAR)3-9 homolog proteins demonstrated that the HΦΦNHSC (Fig. 3, positions 229–235) and GEELTFDY (positions 269–276) motifs were crucial for HMTase activity (20). The cysteine residue in the NHSC motif is not conserved in the E(Z) class proteins (Fig. 3). In this respect, two of the SUVH proteins (SUVH2 and SUVH9) also deviate. Furthermore, these two proteins lack two of three cysteines in the C-SAC, in contrast to the other SUVH and SUVR proteins. While mutation of H233 or C235 abolished HMTase activity, the activity increased when H229 was changed to arginine (20). Interestingly, all the *Arabidopsis* SUVH and SUVR proteins have an arginine in this position.

While the amino acid sequences of the SET domain and the SAC regions of the SUVH and SUVR proteins are similar to proteins with H3-specific HMTase activity, the *Arabidopsis* proteins lack an N-terminal chromo domain present in the *Drosophila* and mammalian homologs. In other organisms the YDG motif present in all the SUVH proteins is found in proteins in combination with different domains connected with an irreversible enzymatic activity [SET domain, histone methylation (20); RING domain, ubiquitination (57); HNH domain, non-specific endonuclease reaction (58)]. Thus, the YDG domain may be a target binding domain of functional similarity to the chromo domain (specific binding on methylated histone H3; 59) or the bromo domain (specific binding on acetylated histone H4; 60), which are alternatively found in SET domain or RING finger proteins, respectively.

In *Drosophila*, SU(VAR)3-9 is involved in the establishment of heterochromatin (8). About one third of the genome of the fruit fly is heterochromatin (61). In contrast, the *Arabidopsis* genome has only heterochromatic regions around the centromeres and two chromomeres, which are prerequisites for chromosome condensation on chromosomes 4 and 5 (32). In mammals, the SU(VAR) homologs are associated with the constitutive heterochromatin of the centromeres and play a role in chromosome segregation and mitotic progression (62). The YDG domain and the presence of AT-hooks in *Arabidopsis* SUVH proteins suggest that these proteins are not necessarily associated with heterochomatin.

What are the biological functions of AtSET proteins?

SET domain proteins have been implicated in developmental processes in *Drosophila*, mammals and *Arabidopsis*. It is therefore reasonable to postulate that many of the newly discovered SET domain proteins in *Arabidopsis* also serve functions in development. The high numbers of such proteins in *Arabidopsis*, compared to other organisms, seems to be a result of extensive duplications in the genome and occasional retrotransposition events. The high numbers may also reflect the differences between animal and plant development: during a plants life, meristems go through transitions to allow inflorescences to develop and, thereafter, flowers with reproductive organs to form. In animals, however, cells giving rise to somatic and reproductive organs segregate early in development. It is interesting to note that our expression analysis showed that all tested genes were expressed in buds and the majority in flowers and seeds, organs where important developmental processes take place. One possibility is that the AtSET proteins have different functions in the same cells. Alternatively, the different paralogs are committed to related functions in different cells and tissues.

We suggest that epigenetic control of gene programs is needed to maintain meristem and organ identity during the different stages of plant development and to aid developmental transitions in response to genetic programs and environmental influences. The four classes of SET domain genes may play crucial roles in such control.

ACKNOWLEDGEMENTS

We thank Kjetill S. Jakobsen for inspiring and motivating the bioinformatics approach to this project. We thank the *Arabidopsis* Biological Resource Center (DNA stock center) for the smRSGFP clone. We thank M. Loebler for the use of and instruction on the particle gun. This work was supported by grants from SFB363 of the Deutsche Forschungsgemeinschaft. V.K. was a post-doctoral fellow of the Deutsche Forschungsgemeinschaft (DFG-Re911/1-3).

REFERENCES

1. Henikoff,S. (1996) Position-effect variegation in *Drosophila*. In Russo,V.E.A. (ed.), *Epigenetic Mechanisms of Gene Regulation*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 319–334.
2. Wallrath,L. (1998) Unfolding the mysteries of heterochromatin. *Curr. Opin. Genet. Dev.*, **8**, 147–153.
3. Weiler,K.S. and Wakimoto,B.T. (1995) Heterochromatin and gene expression in *Drosophila*. *Annu. Rev. Genet.*, **29**, 577–605.
4. Aagaard,L., Laible,G., Selenko,P., Schmid,M., Dorn,R., Schotta,G., Kuhfittig,S., Wolf,A., Lebersorger,A., Singh,P.B. *et al.* (1999) Functional mammalian homologs of the *Drosophila* PEV-modifier *Su(var)3-9* encode centromere-associated proteins which complex with the heterochromatin component M31. *EMBO J.*, **18**, 1923–1938.
5. van Lohuizen,M. (1998) Functional analysis of mouse *polycomb* group genes. *Cell. Mol. Life Sci.*, **54**, 71–79.
6. Reuter,G. and Spierer,P. (1992) Position effect variegation and chromatin proteins. *Bioessays*, **14**, 605–612.
7. Pirrotta,V. (1997) Chromatin-silencing mechanisms in *Drosophila* maintain patterns of gene expression. *Trends Genet.*, **13**, 314–318.

8. Tschiersch,B., Hofmann,A., Krauss,V., Dorn,R., Korge,G. and Reuter,G. (1994) The protein encoded by the *Drosophila* position-effect variegation suppressor gene *Su(var)3-9* combines domains of antagonistic regulators of homeotic gene complexes. *EMBO J.*, **13**, 3822–3831.
9. Pirrotta,V. (1998) Polycombing the genome: PcG, trxG and chromatin silencing. *Cell*, **93**, 333–336.
10. Jenuwein,T., Laible,G., Dorn,R. and Reuter,G. (1998) SET domain proteins modulate chromatin domains in eu- and heterochromatin. *Cell. Mol. Life Sci.*, **54**, 80–93.
11. Goodrich,J., Puangsomlee,P., Martin,M., Long,D., Meyerowitz,E.M. and Coupland,G. (1997) A *polycomb*-group gene regulates homeotic gene expression in *Arabidopsis*. *Nature*, **386**, 44–51.
12. Grossniklaus,U., Vielle Calzada,J.P., Hoeppner,M.A. and Gagliano,W.B. (1998) Maternal control of embryogenesis by *MEDEA*, a *Polycomb* group gene in *Arabidopsis*. *Science*, **280**, 446–450.
13. Luo,M., Bilodeau,P., Dennis,E.S., Peacock,W.J. and Chaudhury,A. (2000) Expression and parent-of-origin effects for *FIS2*, *MEA* and *FIE* in the endosperm and embryo of developing *Arabidopsis* seeds. *Proc. Natl Acad. Sci. USA*, **97**, 10637–10642.
14. Vielle Calzada,J.P., Thomas,J., Spillane,C., Coluccio,A., Hoeppner,M.A. and Grossniklaus,U. (1999) Maintenance of genomic imprinting at the *Arabidopsis MEDEA* locus requires zygotic DDM1 activity. *Genes Dev.*, **13**, 2971–2982.
15. Laible,G., Wolf,A., Dorn,R., Reuter,G., Nislow,C., Lebersorger,A., Popkin,D., Pillus,L. and Jenuwein,T. (1997) Mammalian homologs of the *Polycomb*-group gene *enhancer of zeste* mediate gene silencing in *Drosophila* heterochromatin and at *S.cerevisiae* telomeres. *EMBO J.*, **16**, 3219–3232.
16. Lajeunesse,D. and Shearn,A. (1996) E(z): a *Polycomb* group gene or a *trithorax* group gene? *Development*, **122**, 2189–2197.
17. Laible,G., Haynes,A.R., Lebersorger,A.O.C.D., Mattei,M.G., Denny,P., Brown,S.D.M. and Jenuwein,T. (1999) The murine *polycomb*-group genes *Ezh*1 and *Ezh*2 map close to *Hox* gene clusters on mouse chromosomes 11 and 6. *Mamm. Genome*, **10**, 311–314.
18. Rastelli,L., Chan,C.S. and Pirrotta,V. (1993) Related chromosome binding sites for zeste, suppressors of zeste and polycomb group proteins in *Drosophila* and their dependence on enhancer of zeste function. *EMBO J.*, **12**, 1513–1522.
19. Ivanova,A.V., Bonaduce,M.J., Ivanov,S.V. and Klar,A.J.S. (1998) The chromo and SET domains of the Clr4 protein are essential for silencing in fission yeast. *Nature Genet.*, **19**, 192–195.
20. Rea,S., Eisenhaber,F., O'Carroll,D., Strahl,B.D., Sun,Z.W., Schmid,M., Opravil,S., Mechtler,K., Ponting,C.P., Allis,C.D. *et al.* (2000) Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature*, **406**, 593–599.
21. Tachibana,M., Sugimoto,K., Fukushima,T. and Shinkai,Y. (2001) SET-domain containing protein, G9a, is a novel lysine-preferring mammalian histone methylthansferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3. *J. Biol. Chem.*, **276**, 25309–25317.
22. Nakayama,J., Rice,J.C., Strahl,B.D., Allis,C.D. and Grewal,S.I. (2001) Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science*, **292**, 110–113.
23. Sedkov,Y., Tillib,S., Mizrokhi,L. and Mazo,A. (1994) The *bithorax* complex is regulated by *trithorax* earlier during *Drosophila* embryogenesis than is the *Antennapedia* complex, correlating with a *bithorax*-like expression pattern of distinct early *trithorax* transcripts. *Development*, **120**, 1907–1917.
24. Stassen,M.J., Bailey,D., Nelson,S., Chinwalla,V. and Harte,P.J. (1995) The *Drosophila* trithorax proteins contain a novel variant of the nuclear receptor type DNA binding domain and an ancient conserved motif found in other chromosomal proteins. *Mech. Dev.*, **52**, 209–223.
25. Linder,B., Newman,R., Jones,L.K., Debernardi,S., Young,B.D., Freemont,P., Verrijzer,C.P. and Saha,V. (2000) Biochemical analyses of the AF10 protein: the extended LAP/PHD-finger mediates oligomerisation. *J. Mol. Biol.*, **299**, 369–378.
26. Aasland,R., Gibson,T.B. and Stewart,A.F. (1995) The PHD finger: implications for chromatin-mediated transcriptional regulation. *Trends Biochem. Sci.*, **20**, 56–59.
27. Alvarez-Venegas,R. and Avramova,Z. (2001) Two *Arabidopsis* homologs of the animal *trithorax* genes: a new structural domain is a signature feature of the *trithorax* gene family. *Gene*, **271**, 215–221.
28. Nakamura,T., Blechman,J., Tada,S., Rozovskaia,T., Itoyama,T., Bullrich,F., Mazo,A., Croce,C.M., Geiger,B. and Canaani,E. (2000) huASH1 protein, a putative transcription factor encoded by a human homolog of the *Drosophila ash1* gene, localizes to both nuclei and cell-cell tight junctions. *Proc. Natl Acad. Sci. USA*, **97**, 7284–7289.
29. Tripoulas,N., Lajeunesse,D., Gildea,J. and Shearn,A. (1996) The *Drosophila ash1* gene product, which is localized at specific sites on polytene chromosomes, contains a SET domain and a PHD finger. *Genetics*, **143**, 913–928.
30. Habu,Y., Kakutani,T. and Paszkowski,J. (2001) Epigenetic developmental mechanisms in plants: molecules and targets of plant epigenetic regulation. *Curr. Opin. Genet. Dev.*, **11**, 215–220.
31. Paszkowski,J. and Whitham,S.A. (2001) Gene silencing and DNA methylation processes. *Curr. Opin. Plant Biol.*, **4**, 123–129.
32. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
33. Van Den Ackerveken,G., Marois,E. and Bonas,U. (1996) Recognition of the bacterial avirulence protein AvrBs3 occurs inside the host plant cell. *Cell*, **87**, 1307–1316.
34. Mindrinos,M., Katagiri,F., Yu,G.L. and Ausubel,F.M. (1994) The *A. thaliana* disease resistance gene RPS2 encodes a protein containing a nucleotide-binding site and leucine-rich repeats. *Cell*, **78**, 1089–1099.
35. Davis,S.J. and Vierstra,R.D. (1998) Soluble, highly fluorescent variants of green fluorescent protein (GFP) for use in higher plants. *Plant Mol. Biol.*, **36**, 521–528.
36. Varagona,M.J., Schmidt,R.J. and Raikhel,N.V. (1992) Nuclear localization signal(s) required for nuclear targeting of the maize regulatory protein Opaque-2. *Plant Cell*, **4**, 1213–1227.
37. Kiyosue,T., Ohad,N., Yadegari,R., Hannon,M., Dinneny,J., Wells,D., Katz,A., Margossian,L., Harada,J.J., Goldberg,R.B. *et al.* (1999) Control of fertilization-independent endosperm development by the MEDEA *polycomb* gene in *Arabidopsis*. *Proc. Natl Acad. Sci. USA*, **96**, 4186–4191.
38. Stec,I., Nagl,S.B., van Ommen,G.J.B. and den Dunnen,J.T. (2000) The PWWP domain: a potential protein-protein interaction domain in nuclear proteins influencing differentiation? *FEBS Lett.*, **473**, 1–5.
39. Chuang,R.Y. and Kelly,T.J. (1999) The fission yeast homolog of Orc4p binds to replication origin DNA via multiple AT-hooks. *Proc. Natl Acad. Sci. USA*, **96**, 2656–2661.
40. Lynch,M. and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.
41. Clegg,M.T., Cummings,M.P. and Durbin,M.L. (1997) The evolution of plant nuclear genes. *Proc. Natl Acad. Sci. USA*, **94**, 7791–7798.
42. Soret,J., Gattoni,R., Guyon,C., Sureau,A., Popielarz,M., Le,R.E., Dumon,S., Apiou,F., Dutrillaux,B., Voss,H. *et al.* (1998) Characterization of SRp46, a novel human SR splicing factor encoded by a PR264/SC35 retropseudogene. *Mol. Cell. Biol.*, **18**, 4924–4934.
43. Gray,T.A., Hernandez,L., Carey,A.H., Schaldach,M.A., Smithwick,M.J., Rus,K., Marshall-Graves,.J.A., Stewart,C.L. and Nicholls,R.D. (2000) The ancient source of a distinct gene family encoding proteins featuring RING and C3H zinc-finger motifs with abundant expression in developing brain and nervous system. *Genomics*, **66**, 76–86.
44. Le,Y.J., Kim,H., Chung,J.H. and Lee,Y. (2001) Testis-specific expression of an intronless gene encoding a human poly(A) polymerase. *Mol. Cell*, **11**, 379–385.
45. Birger,Y., Ito,Y., West,K.L., Landsman,D. and Bustin,M. (2001) HMGN4, a newly discovered nucleosome-binding protein encoded by an intronless gene. *DNA Cell Biol.*, **20**, 257–264.
46. Lee,T.H., Yu,S.L., Kim,S.U., Lee,K.K., Rhee,S.G. and Yu,D.Y. (1999) Characterization of mouse peroxiredoxin I genomic DNA and its expression. *Gene*, **239**, 243–250.
47. Nabetani,A., Hatada,I., Morisaki,H., Oshimura,M. and Mukai,T. (1997) Mouse U2af1-rs1 is a neomorphic imprinted gene. *Mol. Cell. Biol.*, **17**, 789–798.
48. Stover,C., Gradl,G., Jentsch,I., Speicher,M.R., Wieser,R. and Schwaeble,W. (2001) cDNA cloning, chromosome assignment and genomic structure of a human gene encoding a novel member of the RBM family. *Cytogenet. Cell Genet.*, **92**, 225–230.
49. Makeyev,A.V., Chkheidze,A.N. and Liebhaber,S.A. (1999) A set of highly conserved RNA-binding proteins, alphaCP-1 and alphaCP-2, implicated in mRNA stabilization, are coexpressed from an intronless gene and its intron-containing paralog. *J. Biol. Chem.*, **274**, 24849–24857.
50. Pan,Y., Decker,W.K., Huq,A.H.H.M. and Craigen,W.J. (1999) Retrotransposition of glycerol kinase-related genes from the X chromosome to autosomes: functional and evolutionary aspects. *Genomics*, **59**, 282–290.

51. Lahn,B.T. and Page,D.C. (1999) Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. *Nature Genet.*, **21**, 429–433.
52. Drouin,G. (2000) Expressed retrotransposed 5S rRNA genes in the mouse and rat genomes. *Genome*, **43**, 213–215.
53. Aravind,L. and Landsman,D. (1998) AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic Acids Res.*, **26**, 4413–4421.
54. Liu,L.S., White,M.J. and MacRae,T.H. (1999) Transcription factors and their genes in higher plants—functional domains, evolution and regulation. *Eur. J. Biochem.*, **262**, 247–257.
55. Jones,C.A., Ng,J., Peterson,A.J., Morgan,K., Simon,J. and Jones,R.S. (1998) The *Drosophila* esc and E(z) proteins are direct partners in polycomb group-mediated repression. *Mol. Cell. Biol.*, **18**, 2825–2834.
56. Rozovskaia,T., Rozenblatt-Rosen,O., Sedkov,Y., Burakov,D., Yano,T., Nakamura,T., Petruck,S., Ben-Simchon,L., Croce,C.M., Mazo,A. *et al.* (2000) Self-association of the SET domains of human ALL-1 and of *Drosophila* TRITHORAX and ASH1 proteins. *Oncogene*, **19**, 351–357.
57. Ruffner,H., Joazeiro,C.A.P., Hemmati,D., Hunter,T. and Verma,I.M. (2001) Cancer-predisposing mutations within the RING domain of BRCA1: loss of ubiquitin protein ligase activity and protection from radiation hypersensitivity. *Proc. Natl Acad. Sci. USA*, **98**, 5134–5139.
58. Kleanthous,C., Kuhlmann,U.C., Pommer,A.J., Ferguson,N., Radford,S.E., Moore,G.R., James,R. and Hemmings,A.M. (1999) Structural and mechanistic basis of immunity toward endonuclease colicins. *Nature Struct. Biol.*, **6**, 243–252.
59. Lachner,M., O'Carroll,N., Rea,S., Mechtler,K. and Jenuwein,T. (2001) Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature*, **410**, 116–120.
60. Jacobson,R.H., Ladurner,A.G., King,D.S. and Tjian,R. (2000) Structure and function of a human TAFII250 double bromodomain module. *Science*, **288**, 1422–1425.
61. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
62. Aagaard,L., Schmid,M., Warburton,P. and Jenuwein,T. (2000) Mitotic phosphorylation of SUV39H1, a novel component of active centromeres, coincides with transient accumulation at mammalian centromeres. *J. Cell Sci.*, **113**, 817–829.
63. Lippuner,V., Chou,I.T., Scott,S.V., Ettinger,W.F., Theg,S.M. and Gasser,C.S. (1994) Cloning and characterization of chloroplast and cytosolic forms of cyclophilin from *Arabidopsis thaliana*. *J. Biol. Chem.*, **269**, 7863–7868.