BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Phylogenetics and evolution of *Su(var)3-9 SET* genes in land plants: rapid diversification in structure and function

Xinyu Zhu[1,2], Hong Ma[3,4], Zhiduan Chen[1*]

Abstract

Background: Plants contain numerous *Su(var)3-9* homologues (*SUVH*) and related (*SUVR*) genes, some of which await functional characterization. Although there have been studies on the evolution of plant *Su(var)3-9 SET* genes, a systematic evolutionary study including major land plant groups has not been reported. Large-scale phylogenetic and evolutionary analyses can help to elucidate the underlying molecular mechanisms and contribute to improve genome annotation.

Results: Putative orthologs of plant Su(var)3-9 SET protein sequences were retrieved from major representatives of land plants. A novel clustering that included most members analyzed, henceforth referred to as core *Su(var)3-9* homologues and related (*cSUVHR*) gene clade, was identified as well as all orthologous groups previously identified. Our analysis showed that plant Su(var)3-9 SET proteins possessed a variety of domain organizations, and can be classified into five types and ten subtypes. Plant *Su(var)3-9 SET* genes also exhibit a wide range of gene structures among different paralogs within a family, even in the regions encoding conserved PreSET and SET domains. We also found that the majority of SUVH members were intronless and formed three subclades within the SUVH clade.

Conclusions: A detailed phylogenetic analysis of the plant *Su(var)3-9 SET* genes was performed. A novel deep phylogenetic relationship including most plant *Su(var)3-9 SET* genes was identified. Additional domains such as SAR, ZnF_C2H2 and WIYLD were early integrated into primordial PreSET/SET/PostSET domain organization. At least three classes of gene structures had been formed before the divergence of *Physcomitrella patens* (moss) from other land plants. One or multiple retroposition events might have occurred among *SUVH* genes with the donor genes leading to the V-2 orthologous group. The structural differences among evolutionary groups of plant *Su(var)3-9 SET* genes with different functions were described, contributing to the design of further experimental studies.

Background

The SET domain (SM00317) is the catalytic center of lysine methyltransferases with a conserved sequence of ~130 amino acid residues, initially identified at the C- terminus of three regulatory factors (Su (var)3-9, E(z) and Trithorax) in *Drosophila* accounting for its name [1-4]. Currently, proteins containing the conserved SET domain can be found in organisms ranging from virus to all three domains of life (Bacteria, Archaea, and Eukaryota) [5]. In plants, Baumbusch et al. [6] first identified 37 putative *Arabidopsis* SET genes, and divided them into four distinct classes: (1) E(Z) homologues; (2) Ash1 homologues and related genes; (3) trx homologues and related genes; and (4) Su(var) homologues and related genes. Subsequently, Springer et al. [7] added 25 maize SET genes to those of 37 *Arabidopsis*, and divided them into five classes based on phylogenetic relationships and domain organization; among these, the Su(var) homologues and related genes were designated as class V. Recently, Ng et al. [8] established two additional plant SET-gene classes, i.e. class VI composed of the SET genes [9] and VII composed of the Putative RuBisCo genes [10]; however, these recent classes lack

* Correspondence: zhiduan@ibcas.ac.cn
[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
Full list of author information is available at the end of the article

typical SET domain, either interrupted in the SET-I region of SET domain or truncated.

Among the seven classes of plant SET genes, class V contains significantly more members relative to other classes and possess the PreSET domain (SM00468) in their proteins [7,8]; for example, from class I to VII, *Arabidopsis* contains 3, 5, 7, 2, 15, 5, and 9 members, respectively. Numerous copies in class V may complicate the evolutionary process of this class of plant SET genes. Previous studies [6,7,11] demonstrated that the class V SET proteins can be further divided into seven orthologous groups (V-1 to 7) and two major types (i.e. SUVH and SUVR) based on their phylogenetic relationship and domain organization. The SUVH proteins consist of orthologous groups V-1, 2, 3 and 5, and have an additional evolutionarily conserved SRA domain (SM00466) upstream of the PreSET domain. The SUVR proteins are composed of the remaining V-4, 6 and 7 orthologous groups and lack the SRA domain. Baumbusch et al. [6] and Springer et al. [7] noted that the majority of SUVH members in *Arabidopsis* and maize lacked introns, and supposed that these intronless SUVH members probably originated from ancient retrotransposition events.

In *Arabidopsis*, there are ten *SUVH* and five *SUVR* genes, in which five *SUVHs* and three *SUVRs* have been characterized functionally [[12], and references therein]. SUVH1, 2, 4, 5 and 6 have been shown to control heterochromatic silencing by the HMTase activity [13-17], and SUVR1, 2, 4 were mainly localized in the nucleolus or nuclear bodies, suggestive of involvement in regulation of rRNA expression [12]. In contrast to SUVH proteins, SUVR4 acts as a dimethyltransferase with preference for mono-methylated H3K9 as substrate, suggesting that SUVHs and SUVRs can act in concert in achieving various functional H3K9 methylation states. It has also been found that the SRA domain of the SUVH proteins may be involved in heterochromatin formation mediated by H3K9 methylation [16]. SUVRs, however, were once supposed to lack a shared N-terminal domain, although a novel conserved N-terminal domain, WIYLD (PF10440), was recently identified in a few members of the V-6 orthologous group, such as the *Arabidopsis* SUVR1, 2, and 4 [12].

Here, we sampled from ten representatives of land plants to investigate the phylogeny and evolution of plant *Su(var)3-9 SET* genes. This is the first analysis of these genes covering the range of land plants. We performed phylogenetic analysis using the combined datasets from the sequences of the conserved PreSET- and SET-domain regions to increase phylogenetic resolution. On the basis of phylogenetic analyses, we tracked the evolution of domain organizations and gene structures of plant *Su(var)3-9 SET* genes in land plants; in turn, these domain organizations and gene structures were used as synapomorphies (derived character states shared by two or more taxa/members) to confirm the phylogenetic relationships. Finally, we explored the relationships between evolutionary patterns and functional diversification by combining the phylogenetic results with available literature for functions of plant *Su(var)3-9 SET* genes; the results of our study would lay the foundation for the design of future experimental studies.

Results

Plant *Su(var)3-9 SET* genes

Arabidopsis thaliana and *Oryza sativa* contained 15 and 12 full-length Su(var)3-9 SET protein sequences, respectively. To undertake an evolutionary analysis of *Su(var) 3-9 SET* genes in land plants, three other completely sequencing plant genomes and one algal genome were searched using multiple representatives of Su(var)3-9 SET proteins in *Arabidopsis thaliana* as queries. By conducting tBLASTn searches against the JGI genome database, we obtained 16, 5, 7 and 1 Su(var)3-9 SET protein sequences from *Populus trichocarpa* (Pt), *Selaginella moellendorfii* (Sm), *Physcomitrella patens* (Pp) and *Chlamydomonas reinhardtii* (Cr), respectively. Seven cDNA sequences of *Pinus taeda* (Pta) were obtained from TIGR plant indices. In addition, 1, 2 and 7 Su(var)3-9 SET protein sequences were also obtained from *Nicotiana tabacum* (Nt), *Vitis vinifera* (Vv) and *Ricinus communis* (Rc), respectively. In total, 74 candidate SET sequences were collected from ten species, and the detailed information is provided in Additional file 1 and 2. Protein sequences lacking PreSET domain were not used for the further study even when they have very low *E* values in the BLAST searches. The *Arabidopsis* SDG11 (SUVH10) was also not used because it is likely a pseudogene [6].

Phylogenetic analysis

Alignment of the combined dataset from PreSET and SET domains resulted in a matrix with length of 228 sites after removing ambiguous regions and autapomorphic insertions (see Additional file 3). The WAG model [18] was selected as the best-fit evolutionary model under the AIC criterion [19] with specific improvements (+G [20]; +F [21]). A maximum-likelihood (ML) analysis produced an optimal tree with an InL score of -20557.59. The NJ analyses recovered trees with almost identical topologies and support values to those of ML analyses. Most of differences between ML and NJ trees were distributed on extremely short branches (see Additional file 4). The ML tree is presented in Figure 1 with bootstrap percentages at the node of the branch.

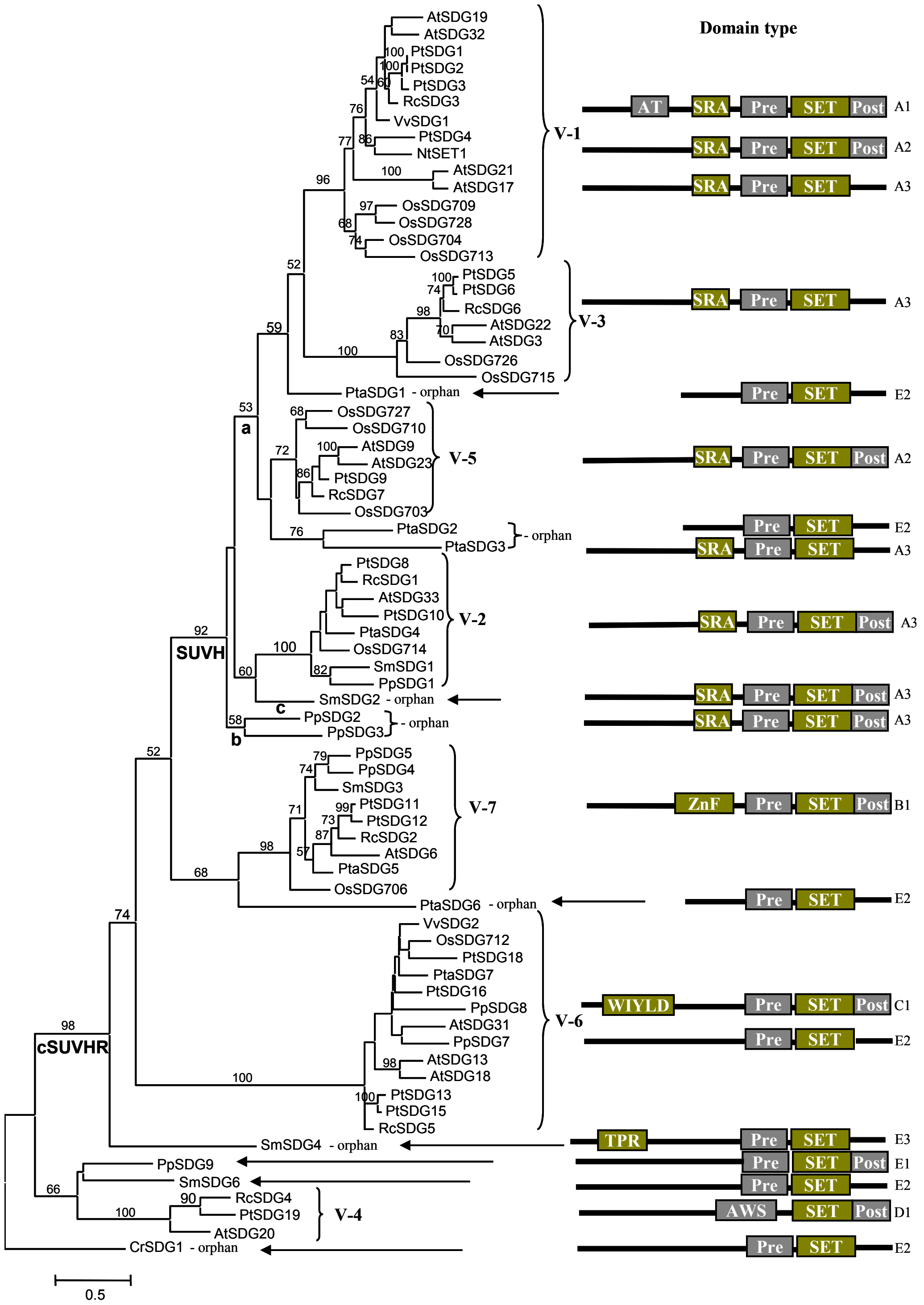

Figure 1 An ML phylogenetic tree of plant Su(var)3-9 SET proteins. The numbers above branches are bootstrap percentages >50, and those below are the clade name. The lowercase letter "a, b, c" represent three intronless clades. The name of Su(var)3-9 SET protein sequences is formed through species abbreviation plus SDG (SET-domain protein group) numbering. Species abbreviation: At, *Arabidopsis thaliana*; Os, *Oryza sativa*; Pt, *Populus trichocarpa*; Nt, *Nicotiana tabacum*; Vv, *Vitis vinifera*; Rc, *Ricinus communis*; Pta, *Pinus taeda*; Sm, *Selaginella moellendorfii*; Pp, *Physcomitrella patens*; Cr, *Chlamydomonas reinhardtii*. The SDG numbering for *Arabidopsis thaliana* and *Oryza sativa* are from ChromDB (http://www.chromdb.org/), and these of other species are numbered in this study. Domain type (see Table 1) within each corresponding clade is depicted on the right. Domain abbreviations: AT, AT_hook; Pre, PreSET; Post, PostSET; ZnF, ZnF_C2H2; TPR, TPR_1.

Our analysis recovered all orthologous groups previously identified [6,7,11] (Figure 1). In the present investigation, we broadened these orthologous groups (group (s), hereafter) based on internal support (>95% BS) or conserved domain organization and gene structure (Table 1 and Figure 2), thus resulting in the inclusion of more members in each group. In the current study, we used the definition of groups previously identified [6,7,11] mainly for the purpose of comparison, and it is possible that some groups we have designated as a single group might actually represent multiple groups because of sampling limitations. Our tree showed that all members could be divided into two clades when the member of *Chlamydomonas reinhardtii* (green alga) was designed as the outgroup. The smaller clade was moderately supported, including V-4 group and other two members; it is worth noting that the V-4 group only contained angiosperm members excluding rice (monocot); the larger clade was strongly supported, including all the remaining members, which was named as cSUVHR (core Su(var)3-9 homologues &related genes) clade in our analysis (Figure 1). Within the cSUVHR clade, the subclade including V-1, 2, 3, and 5 groups was strongly supported and named here as the SUVH clade (Figure 1) because all members possess a characteristic SRA domain at the N terminus [7]; this result was consistent with a previous hypothesis that all *SUVH* genes had a common ancestor [7,11]. The V-7 group within the cSUVHR clade appears to be sister to the SUVH clade, but only with low support. The V-6 group within the cSUVHR clade was placed at the basal position that did not appear to have a clear relationship with other groups.

Within the SUVH clade, the V-1, V-3 and V-5 groups plus several orphan members (PtaSDG1, 2 and 3) formed a subclade with low support and only seed plant members; in contrast, the V-2 group was strongly supported and contained all representative land plants, with usually one copy in each species (Figure 1); the SUVH clade also contained several other orphan members (SmSDG2, PpSDG2 and PpSDG3), and their relationships with other members of this clade were uncertain (Figure 1). The V-6 group contained members from both seed plants and moss, but not *Selaginella moellendorfii* (fern), and was characterized by the WIYLD domain at the N-terminal region of protein sequences [12]. The V-7 group contained members from each major land plant groups with one or several copies in each species, and possessed a characteristic ZnF_C2H2 domain (SM00355) [22] at the N-terminal region of protein sequences. An orphan member SmSDG4 possessed a unique TPR_1 domain (PF00515) at the N-terminal regions of its protein sequences.

Domain organization

To trace their evolutionary history in land plants, we predicted the domain organization of candidate Su(var)3-9 SET proteins. The candidate proteins could be classified into five types (groups) and ten subtypes (subgroups) based on their domain organization, with the major differences lying in their N-terminal regions. Type A contained a characteristic SRA domain at the N-terminus (Table 1), which was identified as the YDG_SRA domain (PF02182) in the Pfam platform [6]. The subtype A1, which only existed in V-1 group, had an additional N-terminal domain, AT_hook (SM00348), a small DNA-binding motif that functions in the transcription regulation of genes containing or in close proximity to AT-rich regions [23,24]. In contrast, the subtypes A2 and A3 were broadly distributed in V-1, 2, and 3 and V-5 groups. It was also worth noting that all members in V-3 group lack the PostSET domain (SM00508) at their C-terminal regions. Type B contains one or more ZnF_C2H2 domain(s) at its N-terminus (Table 2) and was only distributed in the V-7 group.

Table 1 The domain organizations of plant Su(var)3-9 SET proteins

Type	Subtype	Domain architectures	Species										Distribution
			At	Os	Pt	Nt	Rc	Vv	Pta	Sm	Pp	Cr	
A	1	AT_hook-SRA-PreSET-SET-PostSET	+	+	-	-	-	-	-	-	-	-	V1
	2	XXX-SRA-PreSET-SET-PostSET	+	+	+	+	+	+	+	+	+	-	V1,V2,V5
	3	XXX-SRA-PreSET-SET-XXX	+	+	+	-	+	-	+	-	-	-	V1,V2,V3
B	1	ZnF_C2H2-PreSET-SET-PostSET	+	+	+	-	+	-	-	+	+	-	V7
C	1	WIYLD-PreSET-SET-PostSET	+	+	+	-	-	-	-	-	+	-	V6
	2	WIYLD-PreSET-SET-XXX	+	-	+	-	+	+	-	-	-	-	V6
D	1	AWS-SET-PostSET	+	-	+	-	+	-	-	-	-	-	V4
E	1	XXX-PreSET-SET-PostSET	-	-	+	-	-	-	-	-	+	-	Orphan
	2	XXX-PreSET-SET-XXX	-	-	+	-	+	-	+	-	+	+	V6, Orphan
	3	TPR-PreSET-SET-XXX	-	-	-	-	-	-	-	+	-	-	Orphan

Plus sign (+) and minus sign (-) indicate presence and absence of a subtype, respectively; domain and species abbreviations are listed in figure 1; XXX indicates the protein sequence regions without predicted domain.

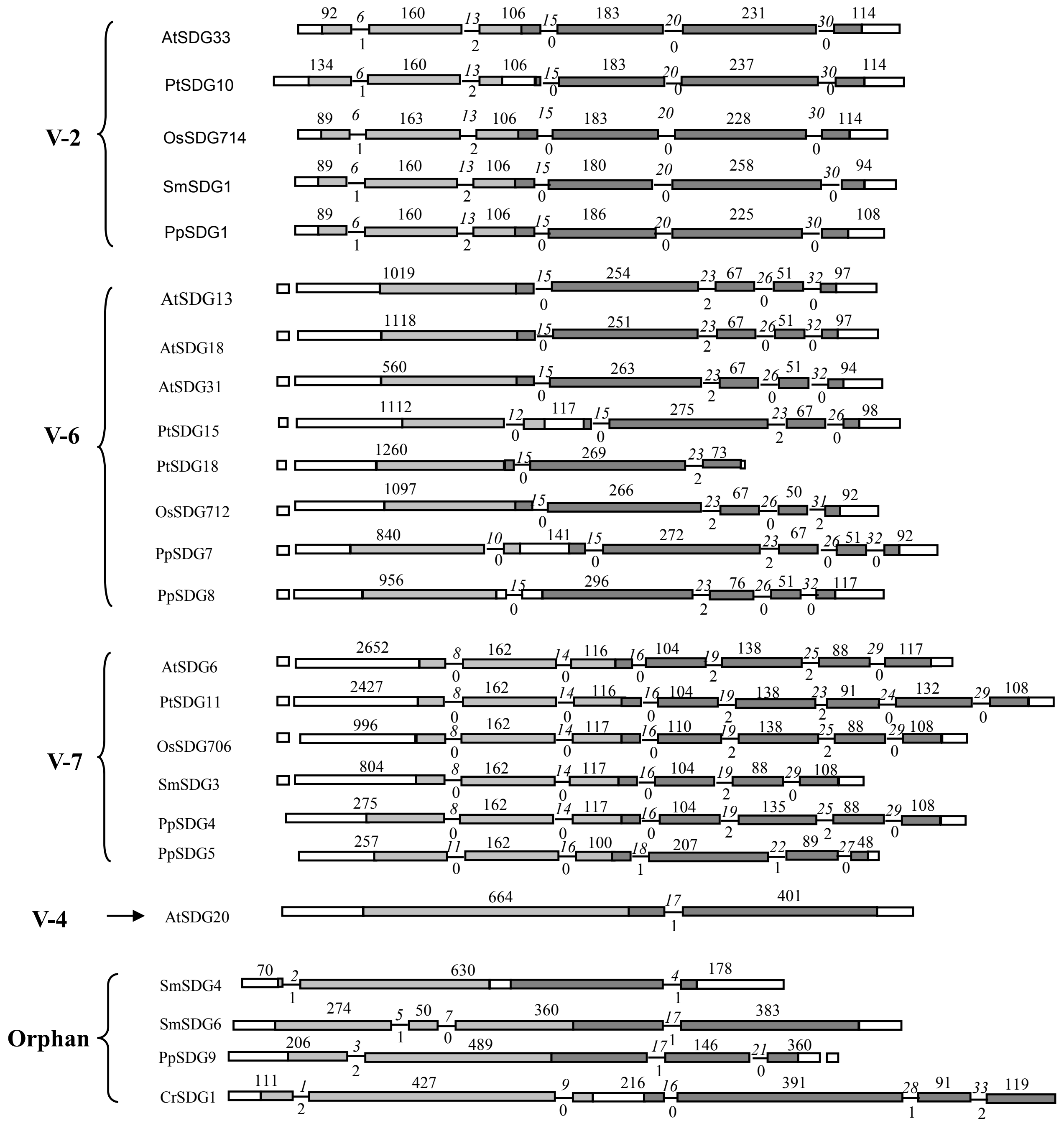
V-2
AtSDG33 92 6 160 13 106 15 183 20 231 30 114 1 2 0 0 0
PtSDG10 134 6 160 13 106 15 183 20 237 30 114 1 2 0 0 0
OsSDG714 89 6 163 13 106 15 183 20 228 30 114 1 2 0 0 0
SmSDG1 89 6 160 13 106 15 180 20 258 30 94 1 2 0 0 0
PpSDG1 89 6 160 13 106 15 186 20 225 30 108 1 2 0 0 0
V-6
AtSDG13 1019 15 254 23 67 26 51 32 97 0 2 0 0
AtSDG18 1118 15 251 23 67 26 51 32 97 0 2 0 0
AtSDG31 560 15 263 23 67 26 51 32 94 0 2 0 0
PtSDG15 1112 12 117 15 275 23 67 26 98 0 0 2 0
PtSDG18 1260 15 269 23 73 0 2
OsSDG712 1097 15 266 23 67 26 50 31 92 0 2 0 2
PpSDG7 840 10 141 15 272 23 67 26 51 32 92 0 0 2 0 0
PpSDG8 956 15 296 23 76 26 51 32 117 0 2 0 0
V-7
AtSDG6 2652 8 162 14 116 16 104 19 138 25 88 29 117 0 0 0 2 2 0
PtSDG11 2427 8 162 14 116 16 104 19 138 23 91 24 132 29 108 0 0 0 2 2 0 0
OsSDG706 996 8 162 14 117 16 110 19 138 25 88 29 108 0 0 0 2 2 0
SmSDG3 804 8 162 14 117 16 104 19 88 29 108 0 0 0 2 0
PpSDG4 275 8 162 14 117 16 104 19 135 25 88 29 108 0 0 0 2 2 0
PpSDG5 257 11 162 16 100 18 207 22 89 27 48 0 0 1 1 0
V-4
AtSDG20 664 17 401 1
Orphan
SmSDG4 70 2 630 4 178 1 1
SmSDG6 274 5 50 7 360 17 383 1 0 1
PpSDG9 206 3 489 17 146 21 360 2 1 0
CrSDG1 111 1 427 9 216 16 391 28 91 33 119 2 0 0 1 2

Figure 2 Schematic representation of the *Su(var)3-9 SET* gene structures in PreSET and SET domain regions. Boxes represent exons and lines represent introns. Length of exons are roughly at scale but introns not. Dark gray regions encode for SET domains; light gray regions encode for PreSET domains. The numbers above boxes are the length of exons (bp). The numbers above lines are the intron positions (see Additional file 5 for the alignment of mRNA sequences) and those below are the intron phases.

The ZnF_C2H2 domain is one type of the C2H2-type zinc fingers (Znf), very common DNA-binding motifs found extensively in eukaryotic and prokaryotic transcription factors [25,26]. Type C contains one WIYLD domain at its N-terminus and was only found in the V-6 group (Table 2). Type D lacked a typical PreSET domain and contains instead the AWS domain (SM00570) (Associated With SET), a subdomain of PreSET domain. This domain organization might have arisen recently because it was only found in angiosperms in the present study. The AWS domain was often found in association with the SET domain, suggesting a

Table 2 Phase and number of introns in plant *Su(var)3-9 SET* genes

	No. of introns in each phase (%)				
Clade (no. of genes)	0	1	2	Total no. of introns	Mean no. of introns per gene
V-2 (5)	43	14	14	71	14
V-4 (1)	0	1	0	1	1
V-6 (8)	35	12	12	59	7.6
V-7 (6)	39	9	9	57	9.5
Orphan (4)	8	7	5	20	5
Total (24)	125(61)	43(20)	40(19)	208	8.7

role in methylation of lysine residues in histones and other proteins [27]. Type E refers to remaining domain organizations that were mostly from orphan members, either lacking identifiable N-terminal domains or having unique N-terminal domains; subtype E1 and E2 might be the ancestral domain organization due to their extensive distribution in eukaryotes (data not shown).

Gene structure

In the present study, the structures of only 24 plant *Su(var)3-9 SET* genes (see Additional file 5 and 6) were analyzed due to the lack of the corresponding genomic sequences in other *Su(var)3-9 SET* genes. We found that the number of intron is highly variable in plant *Su(var)3-9 SET* genes, ranging from 0 in V-1, -3, and -5 groups to 20 in SmSDG1. A total of 208 introns were present in 24 analyzed genes, an average of 8.7 introns per gene; the average number of introns per gene also varied among groups, ranging from 7.6 in V-6 group to 14 in V-2 group (Table 2). Among the 208 introns, 125 (61%) were in phase 0, 43 (20%) in phase 1, and 40 (19%) in phase 2 (Table 2), similar to the previous reports of 57.3% for phase 0, 21.5% for phase 1, and 21.2% for phase 2 in 21,570 rice genes [28]. To trace the evolutionary pattern of gene structure, the current study mainly focused on the most conserved PreSET and SET domain regions. Figure 2 presents the gene structures of these two regions. Our result showed that at least three classes of gene structures (i.e. V-2, V-6 and V-7 groups) were formed probably through frequent inron loss and gain before the divergence of *Physcomitrella patens* from other land plants. In these three groups, the ancestral gene structures might be similar to PpSDG1, PpSDG8 and PpSDG4 (Figure 2), respectively. The sequence similarity between introns was not analyzed because their lengths were highly variable. Within the V-2 group, all introns maintained identical phases and positions, indicating a high degree of structural conservation during the evolution of land plants. In contrast, the V-6 and -7 groups were less conserved; for example, in V-6 the intron sliding occurred in the last intron (position 31 of OsSDG712) (see Additional file 5). Also in V-7 PpSDG5 had only one conserved intron (position 16) compared to other genes. AtSDG20, SmSDG6 and PpSDG9 had a common intron (position17), together with the low support on the relationship among them in phylogenetic tree, suggesting that V-4 group might have a common ancestor with these two orphan members.

Most members in the SUVH clade were intronless except for the V-2 group. Previous studies found that most *Arabidopsis* genes in this clade were intronless, and suggested that these intronless members may have originated from one or a few retroposition events, followed by tandem duplication [6]. If this hypothesis is correct, the donor genes of the retroposition might also be in the V-2 group, because the descendent retrogene and the donor genes should cluster together in the phylogenetic tree just as in Figure 1. In the SUVH clade, intronless members formed three independent subclades, each with weak BP support values (see a, b and c branches in Figure 1). If multiple retroposition events occurred in donor gene lineage, the donor gene lineage would cluster with these retrogenes arranged paraphyletically in phylogenetic tree (Figure 1). Owing to the low support values in the current data, we are still unable to determine whether these three intronless branches originated independently or had a common ancestor.

Discussion

The presence of gene families is one of the characteristics of eukaryotes [29,30]. Since the genes within families are initially redundant in molecular function, they likely have undergone evolutionary selection processes, and eventually formed multiple orthologous groups to carry out different functions [31,32]. The current research first presented the phylogeny and evolution of plant *Su(var)3-9 SET* gene family in land plants. Our analyses identified a novel phylogenetic relationship, that is, the cSUVHR clade that includes most members analyzed except for the V-4 group and a few orphan members (Figure 1). In addition, our results support the following evolutionary scenario of this gene family: multiple gene duplications had occurred independently before the split of *Physcomitrella patens* (moss) from other land plants, and since then each of orthologs experienced molecular divergence by mutations, domain acquisition and gene structure changes, resulted in different orthologous groups. We suggested that the SAR, ZnF_C2H2 and WIYLD domains were early integrated into primordial PreSET/SET/PostSET domain organization to form different evolutionary groups (Figure 3) because the type A, B and C domain organization in Table 1 were all found in *Physcomitrella patens*. In contrast to previous reports [7,8], our analyses showed that the PostSET domain was present in most plant Su(var)3-9 SET proteins, but not in the V-3 group. In the light of the parsimony rule of evolution, we propose that the ancestral plant Su(var)3-9 SET proteins might have possessed the PostSET domain, which was lost in some members during the subsequent evolution (Figure 3).

The plant *Su(var)3-9* SET gene family exhibits a large diversity of gene structures, even in the conserved PreSET and SET domains (Figure 2), implying frequent gain and/or loss of introns during evolution [33]. For plant *Su(var)3-9* SET gene family, such frequent gain/loss of introns might have occurred during early evolution of land plants because the gene structure of the V-2, 6 and 7 groups had appeared before the divergence of *Physcomitrella patens* (Figure 2). Because introns might have regulatory functions [34,35], the gain or loss of introns may have contributed to functional divergence between paralogs, such as subfunctionalization, either directly by introducing regulatory differences or by facilitating exon shuffling. In our study, the V-2 group demonstrated strict conservation of gene structure, indicating that this group may have evolved under high selective pressures and is functionally important; in contrast, V-6 and V-7 may have evolved under relatively relaxed selective pressures (Figure 2). As in previous studies [33,36,37], our study showed that the shared variations in gene structure can be used for the classification of paralogous genes into different evolutionary groups (V-2, 6 and 7); accordingly we further suggest that V-4 group be expanded to include two orphan members, SmSDG6 and PpSDG9, because these two genes have a common intron (position17) with AtSDG20 of V-4 group (Figure 2).

In the SUVH clade (Figure 1), the majority of the genes, except the V-2 group, were intronless. We suppose that their last common ancestor might possess introns and the intronless genes (V-1, 3, 5 groups) originated from the lineages leading to the V-2 group by retroposition because all 24 genes analyzed including from basal evolutionary groups (V-4, 6 and 7) and an outgroup gene (CrSDG1) from *Chlamydomonas reinhardtii* (green alga), also possessed introns (Figure 2). Many retrogenes have been identified in plant gene families [31,38-41]. It is generally believed that most retrogenes become non-functional because they lack the regulatory elements required for expression [42]. However, several recent studies have demonstrated that functional genes can occasionally be generated from retrogenes and that these processed genes take on a non-redundant functional role [38-40]. In the plant *Su(var)3-9 SET* gene family, transcripts of many intronless *Arabidopsis*, *Oryza* and maize genes have been detected in RT-PCR and/or microarray analyses [6,7,43], suggesting that some retrogenes of the *Su(var)3-9 SET* family might have gained regulatory elements and became functional.

We found that *Arabidopsis Su(var)3-9 SET* genes of different groups or clades have different functions (Table 3), suggesting that they interact with the different substrates. SUVH4 (SDG33) (also known as KYP [13]) and SUVH2 [16,44] play major roles in *Arabidopsis* histone H3K9 methylation modification; in contrast, the loss of SUVH1 (SDG32) [16] and SUVH5 (SDG9) [17] or SUVH6 (SDG23) [45] results in only minor reductions in global H3K9 methylation levels. SUVR1 (SDG13), SUVR2 (SDG18) and SUVR4 (SDG31) proteins have been studied

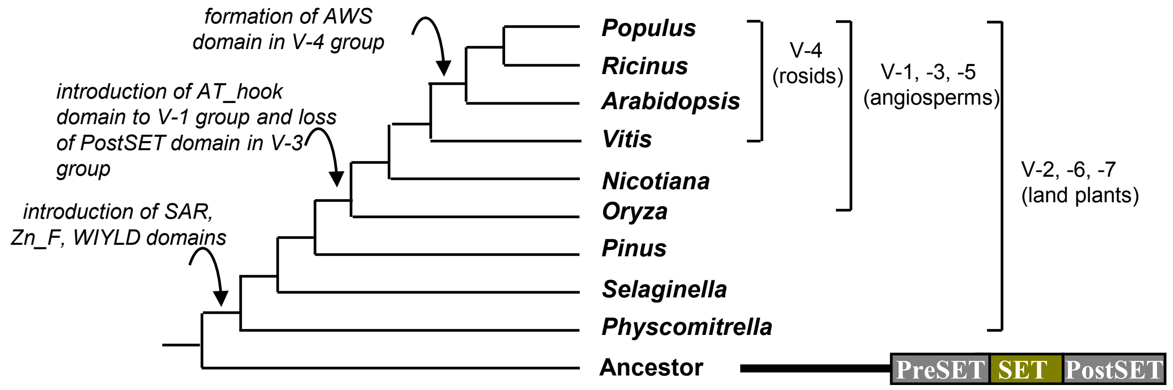
formation of AWS domain in V-4 group
introduction of AT_hook domain to V-1 group and loss of PostSET domain in V-3 group
introduction of SAR, Zn_F, WIYLD domains
Populus
Ricinus
Arabidopsis
Vitis
Nicotiana
Oryza
Pinus
Selaginella
Physcomitrella
Ancestor
V-4 (rosids)
V-1, -3, -5 (angiosperms)
V-2, -6, -7 (land plants)
PreSET
SET
PostSET

Figure 3 The proposed evolutionary events of the domain structures of Su(var)3-9 SET proteins in land plant lineages. Proposed evolutionary events are in italic.

Table 3 Orthology groups and functions of *Arabidopsis Su(var)3-9 SET* genes

Orghology groups		Gene name	Function(s)	Refs
SUVH	V-1	*Suvh1,3,7,8,10*	heterochromatic silencing (minor roles); monomethyltransferase.	[16]
	V-2	*Suvh4*	heterochromatic silencing (major roles); monodimethyltransferase.	[13,14]
	V-3	*Suvh2,9*	heterochromatic silencing (major roles); monodimethyltransferase.	[16,44]
	V-5	*Suvh5,6*	heterochromatic silencing (minor roles); monodimethyltransferase.	[17,45]
SUVR	V-6	*Suvr1,2,4*	nucleolus; repressor of rDNA gene clusters; dimethyltransferase.	[12]
	V-7	*Suvr5*	unknow function.	none
	V-4	*Suvr3*	unknow function.	none

in detail [12]. They are localized to the nucleolus or noncondensed nuclear bodies, which differs from SUVH proteins localizing to heterochromatin region. *In vitro* SUVR4 acts as efficient dimethyltransferase specifically adding the second methyl group to monomethylated H3K9; in contrast, *in vitro* SUVH4 (SDG33), SUVH5 (SDG9) and SUVH6 (SDG23) proteins are very efficient monomethyltransferases but moderately efficient dimethyltransferase [14,17]. The localization of the SUVR proteins suggests that these proteins are not involved in heterochromatic gene silencing, and may function as a repressor of rDNA gene clusters in the decondensed part of the nucleolus. The *SUVR5* (SDG6) and *SUVR3* (SDG20) genes were the only *Arabidopsis* representative of *Su(var)3-9 SET* genes in the V-7 and V-4 groups (Figure 1), respectively, and their functions are unknown and will need to be investigated in the future.

Conclusions

Our study provides novel phylogenetic relationship and new insights into the evolution of plant *Su(var)3-9 SET* gene family in land plants, which includes most members analyzed except for the V-4 group and a few orphan members. We found that the PostSET is not a common domain in plant Su(var)3-9 SET proteins; it might be an ancestral characteristic of this gene family, which was lost in some members during the evolution. We propose that the SAR, ZnF_C2H2 and WIYLD domains were integrated into primordial domain organization, PreSET/SET/PostSET, during the early evolution of land plant and resulted in evolutionary differentiation. Plant *Su(var)3-9 SET* genes exhibit a diversity of structures, even in the conserved PreSET and SET domain regions. At least three classes of gene structures in the V-2, V-6 and V-7 groups had appeared before the divergence of *Physcomitrella patens* from other land plants through frequent inron loss and gain. In the SUVH clade, the majority of the members were intronless retrogenes, probably originated from the ancestral genes leading to V-2 group with introns. Our results revealed the structural differences among evolutionary groups of plant *Su(var)3-9 SET* genes with different functions, and further predicted that the function of *Arabidopsis SUVR5* (SDG6) and *SUVR3* (SDG20) genes belonging to the V-7 and V-4 groups, respectively, are different from other *Arabidopsis Su(var)3-9 SET* genes.

Methods

Homologous Su(var)3-9 SET proteins search

Six completely sequenced plant genomes were selected for retrieving the Su(var)3-9 SET protein sequences. The protein sequences of *Arabidopsis thaliana* and *Oryza sativa* were obtained from the literature [6,8,11]; the protein sequences of *Populus trichocarpa* (angiosperm), *Selaginella moellendorfii* (fern), *Physcomitrella patens* (moss), and *Chlamydomonas reinhardtii* (green alga) were retrieved from JGI genome database (http://genome.jgi-psf.org) by tBLASTn search with default parameters (E value = 1e-5). To better understand the evolutionary history of plant class V SET genes in land plants, we also included Su(var)3-9 SET protein sequences from other plant species, including incompletely sequenced *Nicotiana tabacum* (angiosperm), *Vitis vinifera* (angiosperm), *Ricinus communis* (angiosperm) and *Pinus taeda* (gymnosperm), either by BLASTp from NCBI protein database (nr) or from TIGR EST databases [46]. The protein sequences of SET and PreSET domain regions from 7 Su(var)3-9 SET proteins in *Arabidopsis* were used as the queries. In the JGI database, if alternative splicing was present in the gene model, only the longest transcript was selected, and if truncated SET proteins were found, their gene models will be re-predicted using genomic scaffold sequences. Protein domains were predicted by SMART [47] and Pfam [48] platforms and the sequences possessing PreSET and SET domains are regarded as the candidate Su(var)3-9 SET proteins.

Sequence alignment and phylogenetic analysis

We used the protein sequences of PreSET and SET regions to construct a combined dataset. Alignments of these two regions were first generated independently at the amino acid level using Clustal X [49], followed by manual adjustment, and then the combined matrix of protein sequences was constructed for 73 plant *Su(var) 3-9 SET* genes. The corresponding codon alignment was also constructed according to the protein sequence alignment using the PAL2NAL program [50] for gene structure analyses. Phylogenetic analyses were performed using protein sequences. PHYML [51] and MEGA 3.1[52] were used for ML [53] and NJ [54] analyses, respectively. For the ML method, the ProTest [55] program was used for testing evolutionary model and optimizing parameters. For the NJ method, we used the Jones-Taylor-Thorton +Γmodel as well as simple models of amino acid replacement, such as p-distance [56] with pairwise deletion of gaps. Supports were estimated by non-parametric bootstrap using 1000 replicates for the NJ tree and 500 replicates for the ML tree. In this paper, we used the following descriptions and ranges in the text for describing bootstrap support: weak, 50-75%; moderate, 76-85%; strong, 86-100%.

Analysis of gene structure

Gene structure was analyzed on the basis of phylogenetic analysis. Our analyses mainly focused on the Pre-SET and SET domain regions because the regions outside of these two domains are highly variable in plant Su(var)3-9 SET proteins. Intron-exon borders were determined by aligning the cDNA sequences to their respective genomic region with the spidey program [57] followed by manual inspection of the splice consensus signals. Intron phase was analyzed manually based on the intron-exon border information: phase 0 designated introns between codons, phase 1 designated introns between the first and second bases of a codon, and phase 2 designated introns between the second and third bases of a codon. The intron position information was obtained from nucleotide sequence alignments derived from the protein alignments. Intron positions that are apart even by one base pair were considered as non-identical even if it cannot be excluded that they might have the same ancestor [58].

Additional material

Additional file 1: Plant *Su(var)3-9 SET* homologues and related genes surveyed. The MS excel file provides sampling information of plant *Su(var)3-9* SET homologues and related genes in ten plant species.

Additional file 2: 73 protein sequences used in this study. A txt file gives all protein sequences with fasta format used for phylogenetic analyses.

Additional file 3: Alignment of 73 protein sequences. A txt file provides an alignment of 73 protein sequences with 228 sites.

Additional file 4: NJ tree with branch lengths. A single NJ tree with branch length proportional to the amount of change. The numbers above branches are bootstrap percentage >50. JTT model was used.

Additional file 5: Alignment of 24 mRNA sequences with intron position information. The string of dots indicates the alignment region not containing any intron position information. The red arrows denote the positions of introns.

Additional file 6: Genomic DNA sequences and corresponding mRNA sequences. A zip file provides the 24 genomic DNA sequences and corresponding mRNA sequences used for gene structure analysis with fasta format containing a concise annotation.

Acknowledgements
The authors thank Hongzhi Kong and Wei Wang for comments and suggestions. This research was financially supported by National Basic Research Program of China (973 project no. 2007CB411600), National Natural Science Foundation of China (NSFC 40830209), and Natural Science Fund for Colleges and Universities in Jiangsu Province (09KJB180006).

Author details
[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China. [2]School of Life Sciences, Nantong University, Nantong 226019, China. [3]State Key Laboratory of Genetic Engineering, School of Life Sciences, Institute of Plant Biology, Center for Evolutionary Biology, Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China. [4]Department of Biology, the Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park PA 16802, USA.

Authors' contributions

ZDC, HM and XYZ designed this study. XYZ carried out data searches and analyses, and drafted this manuscript. HM, and ZDC revised several versions with input from all the authors. All authors have read and approved the final manuscript.

Received: 30 July 2010 Accepted: 9 March 2011
Published: 9 March 2011

References

1. Dorn R, Krauss V, Reuter G, Saumweber H: **The enhancer of position-effect variegation of Drosophila, E(var)3-93D, codes for a chromatin protein containing a conserved domain common to several transcriptional regulators.** *Proceedings of the National Academy of Sciences of the United States of America* 1993, **90(23)**:11376-11380.
2. Jones RS, Gelbart WM: **The Drosophila Polycomb-group gene Enhancer of zeste contains a region with sequence similarity to trithorax.** *Molecular and Cellular Biology* 1993, **13(10)**:6357-6366.
3. Tschiersch B, Hofmann A, Krauss V, Dorn R, Korge G, Reuter G: **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 Journal* 1994, **13(16)**:3822-3831.
4. Stassen MJ, Bailey D, Nelson S, Chinwalla V, Harte PJ: **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.** *Mechanisms of Development* 1995, **52(2-3)**:209-223.
5. Alvarez-Venegas R, Sadder M, Tikhonov A, Avramova Z: **Origin of the Bacterial SET Domain Genes: Vertical or Horizontal?** *Molecular Biology and Evolution* 2006, **24(2)**:482-97.
6. Baumbusch LO, Thorstensen T, Krauss V, Fischer A, Naumann K, Assalkhou R, Schulz I, Reuter G, Aalen RB: **The Arabidopsis thaliana genome contains at least 29 active genes encoding SET domain proteins that can be assigned to four evolutionarily conserved classes.** *Nucleic Acids Research* 2001, **29(21)**:4319-4333.
7. Springer NM, Napoli CA, Selinger DA, Pandey R, Cone KC, Chandler VL, Kaeppler HF, Kaeppler SM: **Comparative analysis of SET domain proteins in maize and Arabidopsis reveals multiple duplications preceding the divergence of monocots and dicots.** *Plant Physiology* 2003, **132(2)**:907-925.
8. Ng DW, Wang T, Chandrasekharan MB, Aramayo R, Kertbundit S, Hall TC: **Plant SET domain-containing proteins: structure, function and regulation.** *Biochimica et Biophysica Acta* 2007, **1769(5-6)**:316-329.
9. Jenuwein T, Allis CD: **Translating the histone code.** *Science* 2001, **293(5532)**:1074-1080.
10. Ying Z, Mulligan RM, Janney N, Houtz RL: **Rubisco small and large subunit N-methyltransferases. Bi- and mono-functional methyltransferases that methylate the small and large subunits of Rubisco.** *Journal of Biological Chemistry* 1999, **274(51)**:36750-36756.

11. Zhao Z, Shen WH: **Plants contain a high number of proteins showing sequence similarity to the animal SUV39H family of histone methyltransferases.** *Annals of the New York Academy of Sciences* 2004, **1030**:661-669.
12. Thorstensen T, Fischer A, Sandvik SV, Johnsen SS, Grini PE, Reuter G, Aalen RB: **The Arabidopsis SUVR4 protein is a nucleolar histone methyltransferase with preference for monomethylated H3K9.** *Nucleic Acids Research* 2006, **34(19)**:5461-5470.
13. Jackson JP, Lindroth AM, Cao X, Jacobsen SE: **Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase.** *Nature* 2002, **416(6880)**:556-560.
14. Jackson JP, Johnson L, Jasencakova Z, Zhang X, PerezBurgos L, Singh PB, Cheng X, Schubert I, Jenuwein T, Jacobsen SE: **Dimethylation of histone H3 lysine 9 is a critical mark for DNA methylation and gene silencing in Arabidopsis thaliana.** *Chromosoma* 2004, **112(6)**:308-315.
15. Jasencakova Z, Soppe WJ, Meister A, Gernand D, Turner BM, Schubert I: **Histone modifications in Arabidopsis- high methylation of H3 lysine 9 is dispensable for constitutive heterochromatin.** *Plant Journal* 2003, **33(3)**:471-480.
16. Naumann K, Fischer A, Hofmann I, Krauss V, Phalke S, Irmler K, Hause G, Aurich AC, Dorn R, Jenuwein T, Reuter G: **Pivotal role of AtSUVH2 in heterochromatic histone methylation and gene silencing in Arabidopsis.** *EMBO Journal* 2005, **24(7)**:1418-1429.
17. Ebbs ML, Bender J: **Locus-specific control of DNA methylation by the Arabidopsis SUVH5 histone methyltransferase.** *Plant Cell* 2006, **18(5)**:1166-1176.
18. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.** *Molecular Biology and Evolution* 2001, **18(5)**:691-699.
19. Kullback S, Leibler RA: **On Information and Sufficiency.** *Annals of Mathematical Statistics* 1951, **22(1)**:79-86.
20. Yang Z: **Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites.** *Molecular Biology and Evolution* 1993, **10(6)**:1396-1401.
21. Cao Y, Adachi J, Janke A, Paabo S, Hasegawa M: **Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene.** *Journal of Molecular Evolution* 1994, **39(5)**:519-527.
22. Englbrecht CC, Schoof H, Bohm S: **Conservation, diversification and expansion of C2H2 zinc finger proteins in the Arabidopsis thaliana genome.** *BMC Genomics* 2004, **5(1)**:39.
23. Reeves R, Nissen MS: **The AT-DNA-binding domain of mammalian high mobility group I chromosomal proteins. A novel peptide motif for recognizing DNA structure.** *Journal of Biological Chemistry* 1990, **265(15)**:8573-8582.
24. Friedmann M, Holth LT, Zoghbi HY, Reeves R: **Organization, inducible-expression and chromosome localization of the human HMG-I(Y) nonhistone protein gene.** *Nucleic Acids Research* 1993, **21(18)**:4259-4267.
25. Bouhouche N, Syvanen M, Kado CI: **The origin of prokaryotic C2H2 zinc finger regulators.** *Trends in Microbiology* 2000, **8(2)**:77-81.
26. Wolfe SA, Nekludova L, Pabo CO: **DNA recognition by Cys2His2 zinc finger proteins.** *Annual Review of Biophysics and Biomolecular Structure* 2000, **29**:183-212.
27. Doerks T, Copley RR, Schultz J, Ponting CP, Bork P: **Systematic identification of novel protein domain families associated with nuclear functions.** *Genome Research* 2002, **12**:47-56.
28. Lin H, Zhu W, Silva JC, Gu X, Buell CR: **Intron gain and loss in segmentally duplicated genes in rice.** *Genome Biology* 2006, **7(5)**:R41.
29. Li WH, Gu Z, Wang H, Nekrutenko A: **Evolutionary analyses of the human genome.** *Nature* 2001, **409(6822)**:847-849.
30. Horan K, Lauricha J, Bailey-Serres J, Raikhel N, Girke T: **Genome cluster database. A sequence family analysis platform for Arabidopsis and rice.** *Plant Physiology* 2005, **138(1)**:47-54.
31. Boudet N, Aubourg S, Toffano-Nioche C, Kreis M, Lecharny A: **Evolution of intron/exon structure of DEAD helicase family genes in Arabidopsis, Caenorhabditis, and Drosophila.** *Genome Research* 2001, **11(12)**:2101-2114.
32. Lespinet O, Wolf YI, Koonin EV, Aravind L: **The role of lineage-specific gene family expansion in the evolution of eukaryotes.** *Genome Research* 2002, **12(7)**:1048-1059.
33. Park KC, Kwon SJ, Kim PH, Bureau T, Kim NS: **Gene structure dynamics and divergence of the polygalacturonase gene family of plants and fungus.** *Genome* 2008, **51(1)**:30-40.
34. Fu H, Kim SY, Park WD: **High-level tuber expression and sucrose inducibility of a potato Sus4 sucrose synthase gene require 5′ and 3′ flanking sequences and the leader intron.** *Plant Cell* 1995, **7(9)**:1387-1394.
35. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290(5494)**:1151-1155.
36. Wattler S, Russ A, Evans M, Nehls M: **A combined analysis of genomic and primary protein structure defines the phylogenetic relationship of new members if the T-box family.** *Genomics* 1998, **48(1)**:24-33.
37. Trapp SC, Croteau RB: **Genomic organization of plant terpene synthases and molecular evolutionary implications.** *Genetics* 2001, **158(2)**:811-832.
38. Benovoy D, Drouin G: **Processed pseudogenes, processed genes, and spontaneous mutations in the Arabidopsis genome.** *Journal of Molecular Evolution* 2006, **62(5)**:511-522.
39. Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, Lu Z, Wong GK, Long M, Wang J: **High rate of chimeric gene origination by retroposition in plant genomes.** *Plant Cell* 2006, **18(8)**:1791-1802.
40. Kong H, Leebens-Mack J, Ni W, dePamphilis CW, Ma H: **Highly heterogeneous rates of evolution in the SKP1 gene family in plants and animals: functional and evolutionary implications.** *Molecular Biology and Evolution* 2004, **21(1)**:117-128.
41. Kong H, Landherr LL, Frohlich MW, Leebens-Mack J, Ma H, dePamphilis CW: **Patterns of gene duplication in the plant SKP1 gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth.** *Plant Journal* 2007, **50(5)**:873-885.
42. Graur D, Li W: **Fundamentals of Molecular Evolution.** Sunderland, MA: Sinauer Associates; 2000.
43. Shen WH: **NtSET1, a member of a newly identified subgroup of plant SET-domain-containing proteins, is chromatin-associated and its ectopic overexpression inhibits tobacco plant growth.** *Plant Journal* 2001, **28(4)**:371-383.
44. Fischer A, Hofmann I, Naumann K, Reuter G: **Heterochromatin proteins and the control of heterochromatic gene silencing in Arabidopsis.** *Journal of Plant Physiology* 2006, **163(3)**:358-368.
45. Ebbs ML, Bartee L, Bender J: **H3 lysine 9 methylation is maintained on a transcribed inverted repeat by combined action of SUVH6 and SUVH4 methyltransferases.** *Molecular and Cellular Biology* 2005, **25(23)**:10507-10515.
46. Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J: **The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes.** *Nucleic Acids Research* 2005, , **33 Database:** D71-74.
47. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P: **SMART 5: domains in the context of genomes and networks.** *Nucleic Acids Research* 2006, , **34 Database:** D257-260.
48. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A: **Pfam: clans, web tools and services.** *Nucleic Acids Research* 2006, , **34 Database:** D247-251.
49. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Research* 1997, **25(24)**:4876-4882.
50. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.** *Nucleic Acids Research* 2006, , **34 Web Server:** W609-612.
51. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Systematic Biology* 2003, **52(5)**:696-704.
52. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.** *Briefings in Bioinformatics* 2004, **5(2)**:150-163.
53. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *Journal of Molecular Evolution* 1981, **17(6)**:368-376.
54. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Molecular Biology and Evolution* 1987, **4(4)**:406-425.
55. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21(9)**:2104-2105.

56. Nei M, Kumar S: **Molecular Evolution and Phylogenetics.** Oxford: Oxford University Press; 2000.
57. Wheelan SJ, Church DM, Ostell JM: **Spidey: a tool for mRNA-to-genomic alignments.** *Genome Research* 2001, **11(11)**:1952-1957.
58. Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV: **Analysis of evolution of exon-intron structure of eukaryotic genes.** *Briefings in Bioinformatics* 2005, **6(2)**:118-134.

doi:10.1186/1471-2148-11-63
Cite this article as: Zhu *et al.*: Phylogenetics and evolution of *Su(var)3-9 SET* genes in land plants: rapid diversification in structure and function. *BMC Evolutionary Biology* 2011 **11**:63.