Phylogenetics and evolution of *Trx SET* genes in fully sequenced land plants

Xinyu Zhu, Caoyi Chen, and Baohua Wang

Abstract: Plant Trx SET proteins are involved in H3K4 methylation and play a key role in plant floral development. Genes encoding Trx SET proteins constitute a multigene family in which the copy number varies among plant species and functional divergence appears to have occurred repeatedly. To investigate the evolutionary history of the *Trx SET* gene family, we made a comprehensive evolutionary analysis on this gene family from 13 major representatives of green plants. A novel clustering (here named as cpTrx clade), which included the III-1, III-2, and III-4 orthologous groups, previously resolved was identified. Our analysis showed that plant Trx proteins possessed a variety of domain organizations and gene structures among paralogs. Additional domains such as PHD, PWWP, and FYR were early integrated into primordial SET–PostSET domain organization of cpTrx clade. We suggested that the PostSET domain was lost in some members of III-4 orthologous group during the evolution of land plants. At least four classes of gene structures had been formed at the early evolutionary stage of land plants. Three intronless orphan *Trx SET* genes from the *Physcomitrella patens* (moss) were identified, and supposedly, their parental genes have been eliminated from the genome. The structural differences among evolutionary groups of plant *Trx SET* genes with different functions were described, contributing to the design of further experimental studies.

Key words: Trx SET genes, evolution, land plants, phylogenetics.

Résumé : Chez les plantes, les protéines SET Trx sont impliquées dans la méthylation H3K4 et jouent un rôle dans le développement floral. Les gènes codant pour les protéines SET Trx forment une famille multigénique dont le nombre de copies varie entre les espèces et une divergence fonctionnelle semble être survenue de façon répétée. Afin d'étudier l'évolution de la famille des gènes *SET Trx*, les auteurs ont réalisé une analyse évolutive exhaustive sur cette famille de gènes chez 13 espèces majeures représentatives au sein des plantes vertes. Un nouveau groupe (que les auteurs nomment clade cpTrx) incluant les groupes orthologues III-1, III-2 et III-4 a été identifié. Cette analyse montre que les protéines Trx chez les plantes présentent une grande diversité d'organisation des domaines et de structure des gènes parmi les paralogues. Des domaines additionnels tels que PHD, PWWP et FYR auraient été intégrés très tôt dans l'organisation primordiale SET–PostSET au sein du clade cpTrx. Les auteurs suggèrent que le domaine PostSET a été perdu chez certains membres du groupe d'orthologues III-4 au cours de l'évolution des plantes terrestres. Au moins quatre classes de structure génique existaient aux premiers stades de l'évolution des plantes terrestres. Trois gènes *Trx SET* orphelins sans introns chez le *Physcomitrella patens* (mousse) ont été identifiés et leurs gènes ancestraux auraient supposément été éliminés du génome. Les différences structurales entre groupes de gènes *Trx SET* ayant différentes fonctions chez les plantes sont décrites et cela facilitera la conception d'études futures.

Mots-clés : gènes *Trx SET*, évolution, plantes terrestres, phylogénétique.

[Traduit par la Rédaction]

Introduction

Currently, proteins containing the conserved SET domain (SM00317) can be found in organisms ranging from virus to all three domains of life (Bacteria, Archaea, and Eukaryota) (Alvarez-Venegas et al. 2006). The SET domain, ~130 amino acid residues, is the catalytic center of lysine methyltransferases, initially identified at the C-terminus of three regulatory factors (Su(var)3–9, E(z), and Trithorax) in *Drosophila* accounting for its name (Dorn et al. 1993; Jones and Gelbart 1993; Tschiersch et al. 1994; Stassen et al. 1995). In plants, Baumbusch et al. (2001) first identified 37 putative *Arabidopsis* SET genes and divided them into four distinct classes: (I) E(Z) homologues; (II) Trx homologues and related genes; (III) trx homologues and related genes; and (IV) Su(var) homologues and related genes. Subsequently, Springer et al. (2003) added 25 maize SET genes to those of 37 *Arabidopsis* and divided them into five classes based on phylogenetic re-

Received 3 December 2011. Accepted 17 February 2012. Published at www.nrcresearchpress.com/gen on 14 March 2012.

Paper handled by Associate Editor T. Bureau.

X. Zhu.* School of Life Sciences, Nantong University, Nantong 226019, China; State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China.
C. Chen* and B. Wang. School of Life Sciences, Nantong University, Nantong 226019, China.

Corresponding author: Xinyu Zhu (e-mail: zhuxinyu@ntu.edu.cn).

*These authors contributed equally to this work.

lationships and domain organizations. Recently, Ng et al. (2007) established two additional plant SET gene classes, i.e., class VI and VII; however, these two recent classes lack a typical SET domain either interrupted or truncated in the SET-I region of SET domain (Ying et al. 1999; Jenuwein and Allis 2001).

Previous reports (Springer et al. 2003; Ng et al. 2007) designated plant Trx SET-domain proteins as class III SET-domain proteins and demonstrated that this class of SET-domain proteins can be further divided into four orthologous clusters (labeled as III-1 to III-4). The *Arabidopsis ATX1* (*SDG27*) and *ATX2* (*SDG30*) SET genes, two homologs of *Drosophila Trx SET* genes, were first identified and documented for the presence of a conserved DAST domain (domain associated with SET in Trithorax) (Alvarez-Venegas and Avramova 2001), presently referred to as FYR domain (Phe-Tyr-rich) (SM00541 and SM00542) in the SMART (Letunic et al. 2006) database. The SET proteins of orthologous group III-1 (e.g., *Arabidopsis* SDG27 and maize SDG128) possess a similar arrangement of domains: one PWWP domain (PF00855), one FYR domain (SM00541 and SM00542), two PHD domains (SM00249), one SET domain, and one PostSET domain (SM00508). The PWWP domain is predicted to be involved in mediating protein–protein interactions that are regulators of cell growth and differentiation (Stec et al. 2000). The FYR domain is composed of an FYR C-terminal portion and an FYR N-terminal portion that often occurs near each other but can be separated (Letunic et al. 2006). PHD (plant homeodomain) finger (SM00249) is a C4HC3 zinc-finger-like motif found in nuclear proteins thought to be involved in epigenetics and chromatin-mediated transcriptional regulation. The SET proteins of the orthologous group III-2 (e.g., *Arabidopsis* SDG29 and maize SDG115) also possess a similar domain organization: one PWWP domain, three PHD domains, one SET domain, and one PostSET domain. In contrast to the above two subgroups, the SET proteins of the orthologous groups III-3 and III-4 (e.g., *Arabidopsis* SDG2 and SDG25, and maize SDG108 and SDG127) lack all additional highly conserved domains and only possess a C-terminally located SET domain.

In *Arabidopsis*, there are seven *Trx SET* genes, in which *SDG27 (ATX1)*, *SDG30 (ATX2)*, *SDG2 (ATXR3)*, and *SDG25 (ATXR7)* have been characterized functionally. ATX1 was the first protein confirmed to have H3K4 methyltransferase activity in plants, functionally involved in floral development through activating flower homeotic genes (Alvarez-Venegas et al. 2003; Alvarez-Venegas and Avramova 2005). A recent study (Saleh et al. 2008) revealed that the highly similar *Arabidopsis* Trx SET proteins, ATX1 and ATX2, possess the features of both partial redundancy and of functional divergence: both proteins methylate K4 of histone H3, but while ATX1 trimethylates it, ATX2 dimethylates it. *Arabidopsis* SDG2 (ATXR3) is the major and critical enzyme responsible for global genome-wide deposition of H3K4me3 and regulates gene expression and plant development (Berr et al. 2010; Guo et al. 2010); in contrast, mutants in other class III *Trx SET* genes (*ATX1/SDG27*, *ATX2/SDG30*, and *ATXR7/SDG25*) display locus-specific defects in H3K4me (Alvarez-Venegas and Avramova 2005; Saleh et al. 2008; Berr et al. 2009; Tamada et al. 2009). SDG25 is required for the proper levels of *FLOWERING LOCUS C* (*FLC*) expression (Berr et al. 2009); in loss-of-function mutation, SDG25 has an early flowering phenotype associated with suppression of *FLC* expression (Tamada et al. 2009). The yeast Sc-SET1 (AAB68867.1) that catalyzes histone H3 Lys-4 methylation (Briggs et al. 2001; Roguev et al. 2001) is closely related to the SDG25 SET protein.

Previous reports (Springer et al. 2003; Ng et al. 2007) did not resolve the relationships among four orthologous groups probably because of sampling limitations; in addition, the structural differences among these four evolutionary groups were not investigated, which will provide the solidity of phylogenetic classification and contribute to the design of further experimental studies. Here, we sampled from 13 representatives of land plants to investigate the phylogeny and evolution of plant *Trx SET* genes. This is the first analysis of these genes covering the range of land plants. We performed phylogenetic analysis using the conserved SET-domain region. On the basis of phylogenetic analyses, we tracked the evolution of domain organizations and gene structures of plant *Trx 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 literatures for functions of plant *Trx SET* genes. The results of our study would lay the foundation for the design of future experimental studies.

Materials and methods

Homologous *Trx SET* genes search

Thirteen plant species were selected for retrieving the genomic, cDNA, and protein sequences of *Trx SET* genes: the *Arabidopsis thaliana* and *Oryza sativa* data were downloaded from The Arabidopsis Information Resource (TAIR version 7.0) and The Institute for Genomic Research (TIGR version 5), respectively, according to the literatures (Baumbusch et al. 2001; Ng et al. 2007; Springer et al. 2003); the protein sequences of *Populus trichocarpa*, *Vitis vinifera*, and *Medicago truncatula* data were retrieved by tBLASTn searches from NCBI and JCVI (version 3) plant genome databases with default parameters, respectively; and *Zea mays*, *Sorghum bicolor*, *Selaginella moellendorfii* (lycophytes, pteridophytes), *Physcomitrella patens* (moss, bryophytes), and *Chlamydomonas reinhardtii* (green alga) data were retrieved from phytozome database as of April 2011 by tBLASTn search with default parameters. To better understand the evolutionary history of plant *Trx SET* genes in land plants, we also included Trx SET protein sequences from three other plant species, *Ricinus communis*, *Hordeum vulgare*, and *Pinus taeda* (Pinaceae, gymnosperm), by BLASTp search from NCBI protein database (nr) and TIGR plant gene indices (Lee et al. 2005). The protein sequences of SET domain region from five Trx SET proteins in *Arabidopsis* were used as the queries. The target sequences were selected when the pairwise amino acid identity between the queries and the targets was over 40% (Tian and Skolnick 2003). 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 repredicted using genomic scaffold sequences. For the true *Trx SET* genes, we expected to see their top hits in the candidates; for each candidate, we manually inspected their domain annotations in SMART (Letunic et al. 2006) and Pfam (Finn et al. 2006) platforms. A few candidates were found not to be *Trx SET* genes in the subsequent more rigorous domain organization and gene structure analyses and were removed. We also searched the dbEST databases in NCBI and plant gene indices in TIGR (Lee et al. 2005) for the above candidates using tBLASTn search, as EST data could provide some gene expression information.

Sequence alignment and phylogenetic analysis

Phylogenetic analyses were performed using protein sequences in SET domain regions. Alignments were generated using Clustal X (Thompson et al. 1997), followed by manual adjustment. PHYML (Guindon and Gascuel 2003) and MEGA 5 (Tamura et al. 2011) were used for maximum likelihood (ML) (Felsenstein 1981) and neighbor-joining (NJ) (Saitou and Nei 1987) analyses, respectively. For the ML method, the ProTest program (Abascal et al. 2005) was used for testing evolutionary model and optimizing parameters. For gene structure analyses, alignment was first generated at the amino acid level and then the corresponding codon alignment was constructed according to the protein sequence alignment using the PAL2NAL program (Suyama et al. 2006). Supports were estimated by nonparametric 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%; and strong, 86%–100%.

Analysis of gene structure

Gene structure was analyzed on the basis of phylogenetic analysis. Our analyses mainly focused on the SET domain regions because the regions outside of this domains are highly variable in plant Trx SET proteins. Intron–exon borders were determined by aligning the cDNA sequences to their respective genomic region with the spidey program (Wheelan et al. 2001), followed by manual inspection of the splice consensus signals. Intron phase was analyzed manually based on the intron–exon border information. The intron position information was obtained from nucleotide sequence alignment derived from the protein alignment. Intron positions that are apart, even by one base pair, were considered as nonidentical even if it cannot be excluded that they might have the same ancestor (Rogozin et al. 2000).

Results

Plant *Trx SET* genes

Arabidopsis thaliana and *O. sativa* contained seven and four full-length Trx SET protein sequences, respectively. To undertake an evolutionary analysis of *Trx SET* genes in land plants, seven other completely or nearly sequenced land-plant genomes and one algal genome were searched using seven Trx SET protein sequences of *A. thaliana* as the queries. By conducting BLASTp searches against NCBI and JCVI plant genome databases, we obtained four, four, and five Trx protein sequences from *P. trichocarpa* (Pt), *V. vinifera* (Vv), and *M. truncatula* (Mt), respectively; and by tBLASTn searches against the JGI genome database, we obtained five, four, five, three, and three Trx SET protein sequences from *Z. mays* (Zm), *S. bicolor* (Sb), *S. moellendorfii* (Sm), *P. patens* (Pp), and *C. reinhardtii* (Cr), respectively. Two cDNA sequences of *P. taeda* (Pta) were obtained from TIGR plant gene indices. In addition, two and five Trx SET protein sequences were also obtained by BLASTp search from NCBI protein database (nr) from *H. vulgare* (Hv) and *R. communis* (Rc), respectively. In total, 57 Trx SET protein sequences were collected from 13 species (Table 1), and the detailed information is provided in the Supplementary data,[1] Table S1.

Phylogenetic analysis

Alignment of the dataset from SET domains resulted in a matrix with a length of 121 sites (see Fig. S1), after removing ambiguous regions and autapomorphic insertions. The Jones-Taylor-Thorton (JTT) model (Jones et al. 1992) was selected as the best-fit evolutionary model under the AIC criterion (Kullback and Leibler 1951) with specific improvements (G (= 1.16) (Yang 1993) + I (= 0.04) (Sidow et al. 1992)). An ML analysis produced an optimal tree with a log likelihood score of –5071.46117. For the NJ method, we used the JTT model as well as simple models of amino acid replacement, such as *p*-distance (Nei and Kumar 2000) with pairwise deletion of gaps. The NJ analyses recovered trees with almost identical topologies and support values to those of ML analyses. Most of the differences between ML and NJ trees were distributed on extremely short branches (see Fig. S2). The midpoint rooting ML phylogenetic tree is presented in Fig. 1 with bootstrap percentages at the node of the branch.

Our analysis recovered all orthologous groups previously identified (Baumbusch et al. 2001; Springer et al. 2003; Ng et al. 2007). In the present investigation, we broadened these orthologous groups (group(s), hereafter) based on internal support (>90% BS) or conserved domain organization and gene structure (Figs. 1 and 2), thus resulting in the inclusion of more members in each group; we used the definition of groups previously identified in the current study mainly for the purpose of comparison, and it is possible that some groups we have designated as a single group or as a co-orthologous group (Sonnhammer and Koonin 2002) might actually represent multiple groups because of sampling limitations. Our tree showed that all members could be divided into two large clades with strong bootstrap supports. The smaller clade comprised only of the members of III-3 group; in contrast, the larger clade contains all remaining members, i.e., III-1, III-2, and III-4 groups plus one orphan clade, which was named cpTrx (core plant Trx proteins) clade in our analysis (Fig. 1). Within the cpTrx clade, the subclade containing III-1 and III-2 groups was weakly supported and named here as the Trx12 clade (Fig. 1), in which all members possess a characteristic PWWP and two or three PHD domains at the N-terminus and consistently possess PostSET domain at the C-terminus; in addition, within the Trx12 clade, all members of the III-1 group pos-

[1]Supplementary data are available with the article through the journal Web site (http://nrcresearchpress.com/doi/suppl/10.1139/g2012-012).

Table 1. Information on the *Trx* gene family in the 13 plant species studied.

Index	Abbr.	Clade	Species	Copy No.	Web site	Reference
1	At	Eudicots	*Arabidopsis thaliana*	7	http://www.arabidopsis.org/index.jsp	Arabidopsis Genome Initiative 2000
3	Vv	Eudicots	*Vitis vinifera*	4	http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html	Jaillon et al. 2007
4	Pt	Eudicots	*Populus trichocarpa*	4	http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html	Tuskan et al. 2006
5	Rc	Eudicots	*Ricinus communis*	5	http://blast.ncbi.nlm.nih.gov/Blast.cgi	Thakur et al. 2011
6	Mt	Eudicots	*Medicago truncatula*	5	http://www.jcvi.org/	Cannon et al. 2006
2	Os	Monocots	*Oryza sativa*	4	http://www.jcvi.org/	Goff et al. 2002
7	Zm	Monocots	*Zea mays*	5	http://www.phytozome.net/search.php	Chan et al. 2006
8	Sb	Monocots	*Sorghum bicolor*	4	http://www.phytozome.net/search.php	Paterson et al. 2009
9	Hv	Monocots	*Hordeum vulgare*	2	http://blast.ncbi.nlm.nih.gov/Blast.cgi	—
10	Pta	Gymnosperm	*Pinus taeda*	2	http://compbio.dfci.harvard.edu/tgi/	Lee et al. 2005
11	Sm	Lycophytes	*Selaginella moellendorfii*	5	http://www.phytozome.net/search.php	Wang et al. 2005
12	Pp	Moss	*Physcomitrella patens*	7	http://www.phytozome.net/search.php	Rensing et al. 2008
13	Cr	Green algae	*Chlamydomonas reinhardtii*	3	http://www.phytozome.net/search.php	Merchant et al. 2007

sess a characteristic FYR domain at the N-terminus. The III-1 and III-2 groups each contains one green alga member, indicating that they might have been established at the stage of single cell green algae. The orphan clade (PpSDG8 plus ZmSDG106) is nested in Trx12 clade, but its relationship with III-1 and III-2 groups is currently unclear. The III-4 group contains most of the land plant representatives usually with one copy in each species, and their protein sequences have no characteristic domains at the N-terminus, with or without PostSET domain at the C-terminus (Fig. 1). The III-3 group contains all land plant representatives usually with one copy in each species and one green alga member (CrSDG3401), indicating that this group might have been established at the stage of single cell green algae; similar to the III-4 group, the protein sequences of the III-3 group has no characteristic domains at the N-terminus, but consistently lacking PostSET domain at the C-terminus (Fig. 1).

Domain organization

To trace the evolutionary history of *Trx SET* genes in land plants, we predicted the domain organization of Trx SET proteins. The III-1 group possessed a conserved arrangement of domains, i.e., one PWWP domain, one FYR domain (named DAST by Alvarez-Venegas and Avramova 2001), two PHD domains, and one PostSET domain (Fig. 1). The FYR domain is composed of an FYR C-terminal portion and an FYR N-terminal portion that often occurs near each other but can be separated. There were some changes in several members of the III-1 group (Fig. 1): VvSDG1 and HvSDG1 (A-3 type) contain one unique domain, i.e., TUDOR domain (SM00333), a domain of unknown function present in several RNA-binding proteins; SmSDG2 lacks two PHD domains (A-4); PpSDG1512 lacks PostSET domain (A-6 type); green alga member CrSDG3404 also contains one unique domain (A-7 type), i.e., Agenet (SM00743), with possible chromatin-associated functions; and SbSDG1 lacks all other domains except SET and PostSET (A-8 type). The III-2 group usually contains one PWWP domain, three PHD domains, and one PostSET domain in addition to the SET domain (Fig. 1). Similarly, there were some changes found in several members outside this conserved domain organization: the RcSDG2 and RcSDG3 (B-2 type) each has an additional SAND domain; the SbSDG2 and ZmSDG115 (B-3 type) each has an additional PostSET domain. The two members of orphan clade, PpSDG8 and ZmSDG106 (C-1 type), lack PWWP and FYR domains compared with other members of Trx12 clade (Fig. 1).

The III-4 group usually lack additional domains except a C-terminally located SET domain, with or without PostSET domain. MtSDG5 and PpSDG1502 (D-2 and D-4) each contains one additional GYF domain (SM00444) near the N-terminus of their protein sequence, which is involved in binding Pro-rich regions of other proteins (Freund et al. 1999). SmSDG5 (D-5) also contains one additional ZnF_C2HC domain (SM00343) near the N-terminus of its protein sequence, which is involved in RNA binding or single-stranded DNA binding in eukaryotic proteins (Clay and Nelson 2005). Similarly, the III-3 group also lack additional domains except the conserved SET domain, but in contrast to the III-4 group, this group consistently lacks PostSET domain and possesses

Fig. 1. The midpoint-rooted ML tree of plant Trx SET proteins. The numbers above branches are bootstrap percentages >50, and those below are the clade name. The lowercase letters a, b, and c represent three intronless orphan sequences. The name of Trx SET protein sequences is formed through species abbreviation plus SDG (SET-domain protein group) numbering; species abbreviations are listed in Table 1. Type(s) of domain organization within each clade are depicted on the right. Domain abbreviations: Post, PostSET; ZnF, Znf_C2H2; Bromo, bromodomain; AT, AT_hook; AG, Agenet; LBR, LBR_tudor. The genes without available EST sequences are indicated by triangles. The stars denote the inferred early duplication events.

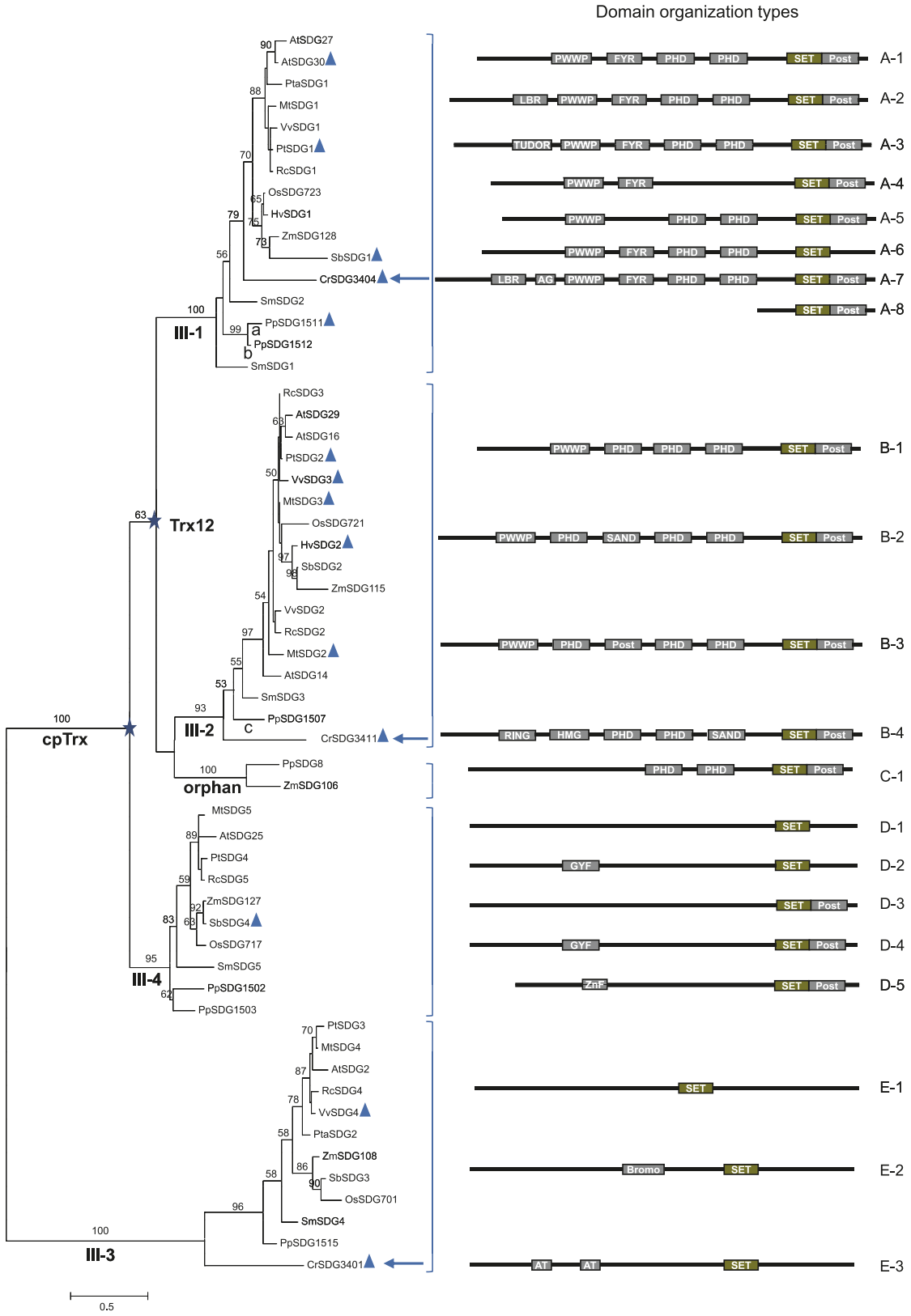


Fig. 2. Schematic representation of the *Trx SET* gene structures in SET domain regions. Boxes represent exons and lines represent introns. Length of exons are roughly at scale but introns are not. Dark gray regions encode for SET domains. The numbers above boxes are the length of exons (bp). The numbers above lines are the intron positions (see Fig. S3 for the alignment of mRNA sequences) and those below are the intron phases. The highlighted columns indicate the intron positions shared by III-1, III-2, and III-4 groups and orphan clade.

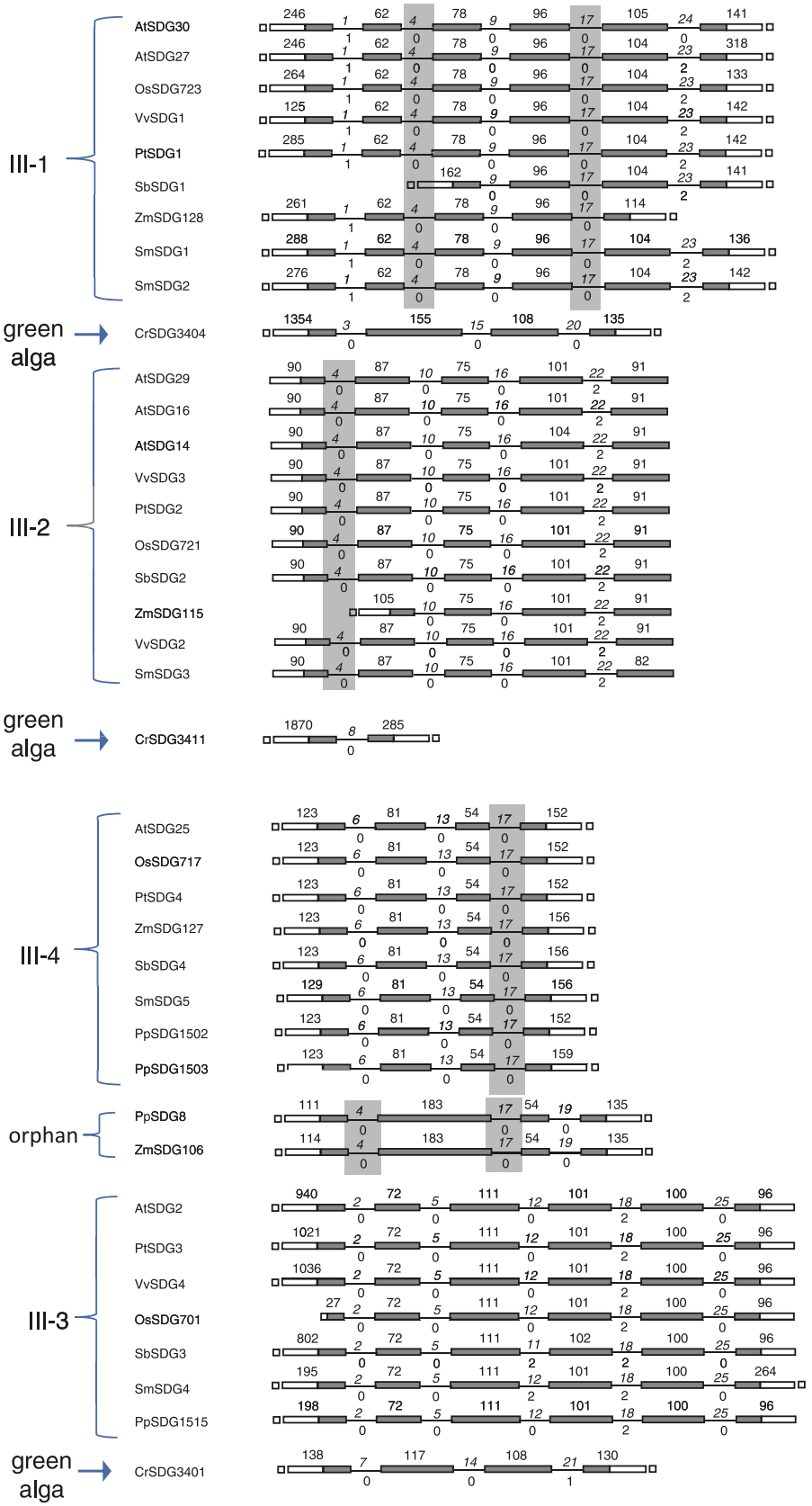


Table 2. Phase and number of introns in plant *Trx SET* genes.

	No. of introns in each phase (%)				
Clade (no. of genes)	0	1	2	Total no. of introns	Mean no. of introns per gene
III-1 (10)	106	61	39	206	20.6
III-2 (11)	122	29	71	222	20.2
III-3 (8)	98	21	24	143	17.9
III-4 (8)	61	13	12	86	10.5
Orphan (2)	22	7	4	33	16.5
Total (39)	409 (59)	131 (19)	150 (22)	690	17.7

a relatively centrally located SET domain. In this group, ZmSDG108 (E-2 type) contains one unique Bromodomain (PF00439), which may be involved in protein–protein interactions and play a role in assembly or activity of multicomponent complexes involved in transcriptional activation (Owen et al. 2000). The green alga member CrSDG3401 (E-3 type) contains two additional AT_hook domains (SM00348), a small DNA-binding motif that functions in the transcription regulation of genes containing or in close proximity to AT-rich regions (Reeves and Nissen 1990).

Gene structure

In the present study, the structures of only 42 plant *Trx SET* genes were analyzed because of the lack of corresponding genomic sequences in other *Trx SET* genes. We found that three *Physcomitrella* genes, i.e., *PpSDG1511* and *PpSDG1512* in III-1 group and *PpSDG1507* in III-2 group (Fig. 1), were intronless; these three intronless genes did not form independent clades during evolution as in previous reports (Fablet et al. 2009; Zhu et al. 2011), instead, were mixed with other genes with introns as an orphan member. In 39 *Trx SET* genes with introns, the number of introns highly variable, ranging from 3 in *SmSDG5* and *ZmSDG115* to 26 in *VvSDG4*; a total of 690 introns were present in the *Trx SET* genes with introns, a mean of 17.7 introns per gene; the mean number of introns per gene also varied among groups, ranging from 10.5 in III-4 group to 20.6 in III-1 group (Table 2). Among the 690 introns, 409 (59%) were in phase 0, 131 (19%) in phase 1, and 150 (22%) 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 (Lin et al. 2006).

To trace the evolutionary pattern of gene structure, the current study mainly focused on the most conserved SET domain regions. Our result showed that at least four classes of conserved gene structures (i.e., III-1 to III-4 groups, see Fig. 2) were formed probably through frequent intron loss and gain at the early evolutionary stage of land plants; among these, the III-1 and III-2 groups occurred before the divergence of *S. moellendorfii* from other land plants, and the III-3 and III-4 groups occurred before the divergence of *P. patens* from other land plants. In these four classes of conserved gene structures, the ancestral gene structures might be similar to *SmSDG2*, *SmSDG3*, *PpSDG 1503*, and *PpSDG1515* (Fig. 2), respectively. Almost all introns in these four classes of gene structures have maintained identical phases and positions (Fig. 2), only two intron sliding events occurred in AtSDG30 (position 24) and SbSDG3 (position 11), respectively (see Fig. 2 and Fig. S3), indicating a high degree of structural conservation of these *Trx SET* genes in land plants. We found that the intron position 4 was shared by the III-1 and III-2 groups and the orphan clade (Fig. 2), and the intron position 17 was shared by the III-1 and III-4 groups and the orphan clade (Fig. 2); the existence of these two shared intron positions suggested that the three classes of gene structures in cpTrx clade may originate from ancient gene duplication events that occurred at the early evolutionary stage of land plants (Fig. 1). The gene structures in three green alga genes, *CrSDG3404*, *CrSDG3411*, and *CrSDG3401*, were unique compared with other analyzed *Trx SET* genes, although they possessed similar domain organizations with III-1, III-2, and III-4 groups, respectively (Fig. 1). The sequence similarity between introns was not analyzed in the present study because their lengths were highly variable.

Discussion

The presence of gene families is one of the characteristics of eukaryotes (Li et al. 2001; Horan et al. 2005). As 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 (Boudet et al. 2001; Lespinet et al. 2002). The current research first presented the phylogeny and evolution of the plant *Trx SET* gene family in land plants. Our analyses identified a novel phylogenetic relationship, that is, the cpTrx clade that includes most members analyzed except for the III-3 group (Fig. 1). In addition, our results support the following evolutionary scenario of this gene family: at least two gene duplications had occurred independently before the divergence of *P. patens* (moss) from other land plants, and since then each paralog has experienced molecular divergence by mutations, domain acquisition, and gene structure changes, resulting in different orthologous groups in cpTrx clade. We suggested that the PHD, PWWP, and FYR domains were early integrated into primordial SET–PostSET domain organization (data not shown) to form III-1 and III-2 groups (Fig. 3), as these three domains were found in *C. reinhardtii* (green alga) in III-1 and III-2 groups. In contrast to previous reports (Baumbusch et al. 2001; Springeret al. 2003; Ng et al. 2007), our analyses showed that the PostSET domain was present in most members of cpTrx clade, but not in some members of III-4 group. In the light of the parsimony rule of evolution, we propose that the ancestral plant cpTrx clade might have possessed the PostSET domain, which was lost in some members of III-4 group during the subsequent evolution. In the cpTrx clade (see Fig. 1), the III-1 and III-2 groups have a closer relation-

Fig. 3. The proposed evolutionary events of *Trx SET* genes in the lineages of green plants. Proposed evolutionary events are in italic.

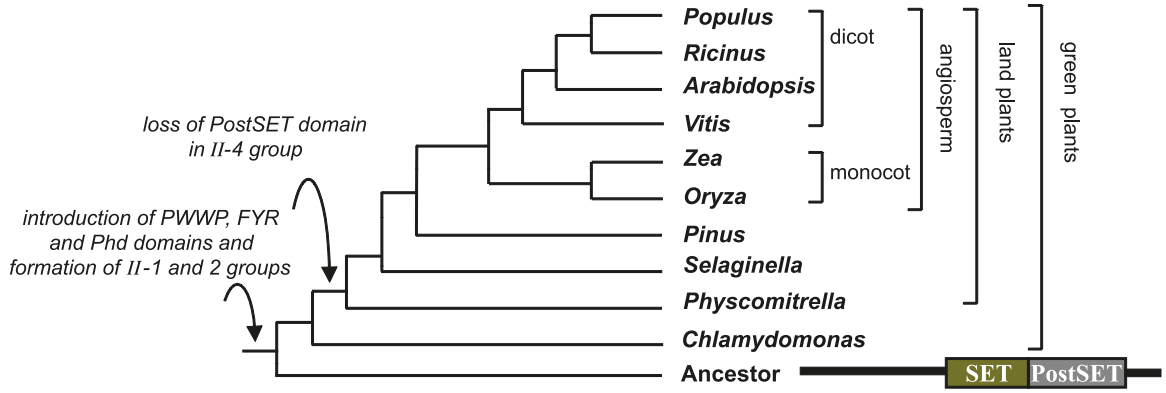
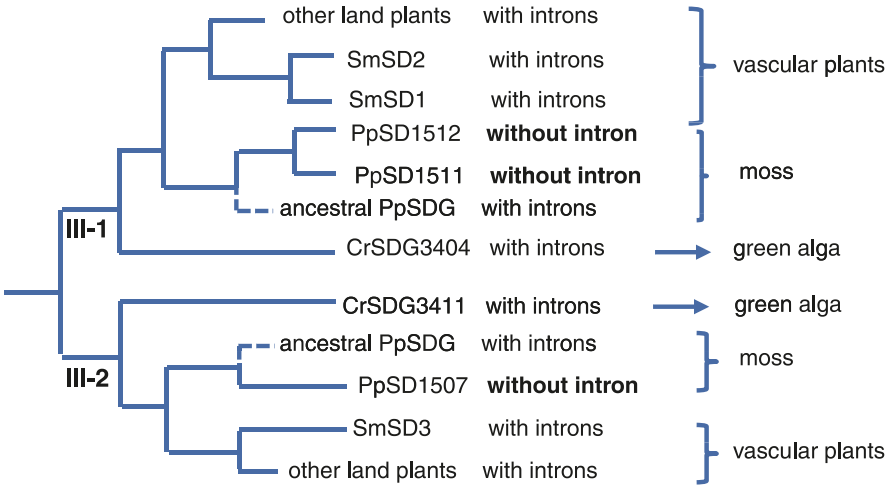


Fig. 4. The γ type of tree topology showing the existence of orphan retroposons in III-1 and III-2 orthologous groups. The branches using the dash lines indicate the hypothetical ancestral sequences, which had been eliminated from the genome.



ship based on the bootstrap supports and domain organizations, and the two groups should also have originated earlier than the III-4 group because they contain green algae that does not appear in the III- 4 group. Thus, the III-4 group may have originated from the lineage of III-1 or III-2 groups, and more likely from the III-1 group owing to the shared intron position 17 (see Fig. 2).

The plant *Trx SET* gene family exhibits a large diversity of gene structures, even in the conserved SET domain (Fig. 2), implying frequent gain and loss of introns during evolution (Park et al. 2008). For the plant *Trx* SET gene family, such frequent gain and loss of introns might have occurred at the early evolutionary stage of land plants because the gene structure of all groups had appeared before the divergence of *P. patens* (moss) or *S. moellendorfii* (fern) (Fig. 2). As introns might have regulatory functions (Fu et al. 1995; Lynch and Conery 2000), the gain and 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, all groups demonstrated strict conservation of gene structure (Fig. 2), indicating that these groups may have evolved under high selective pressure and are functionally important. As in previous studies (Wattler et al. 1998; Trapp and Croteau 2001), our study showed that the shared variations in gene structure can be used for the clustering of paralogous genes, and accordingly we further suggest that the III-1, III-2, and III-4 groups and an orphan clade can be grouped into a larger clade, i.e., cpTrx clade (Fig. 1) based on the shared intron positioin 4 and 17 (Fig. 2).

In the cpTrx clade (Fig. 1), there were three intronless genes, among these, two in III-1 group (*PpSDG1511* and *PpSDG1512*) and one in III-2 group (*PpSDG1507*). We characterized these three orphan single exonic genes as retrogenes by comparing them with the green algal orthologs (*CrSDG3404* and *CrSDG3411*) within the same orthologous group, which contain 21 and 19 introns, respectively (data not shown); the underlying assumption is that the possibility of about 20 introns lost in *P. patens* and then gained again in vascular plants within the same orthologous group is nil, and such a situation can only be explained as a result of the reverse transcription process (Roy et al. 2003). The existence of orphan retrogenes (Fig. 4) was termed as γ type of tree topology (Fablet et al. 2009); it seemed to imply that such retrogenes had some adaptive advantages over their parental genes that were gradually eliminated from the genome, leaving those orphan genes alone as suggested in a previous report (Zhang et al. 2005). It is generally believed that most retrogenes become nonfunctional because they lack the regulatory elements required for expression (Graur and Li 2000).

Table 3. Orthologous groups and functions of Arabidopsis *Trx SET* genes.

Orthologous group	Gene name	Function	Reference
III-1	*SDG27* (*ATX1*), *SDG30* (*ATX2*)	SDG27 functions as an activator of floral homeotic genes, with histone H3K4 methyltransferase; in loss-of-function mutation, SDG27 leads to relatively mild and pleiotropic phenotypes. ADG27 and SDG30 possess the features of both partial redundancy and of functional divergence. Both proteins methylate K4 of histone H3, but while ATX1 trimethylates it, ATX2 dimethylates it.	Alvarez-Venegas et al. 2003; Alvarez-Venegas and Avramova 2005; Saleh et al. 2008
III-2	*SDG14* (*ATX3*), *SDG16* (*ATX4*), *SDG29* (*ATX5*)	Unknown	None
III-3	*SDG2* (*ATXR3*)	SDG2 is required for global H3K4me3 deposition; in loss-of-function mutation, SDG2 leads to severe and pleiotropic phenotypes as well as the misregulation of a large number of genes.	Berr et al. 2010; Guo et al. 2010
III-4	*SDG25* (*ATXR7*)	SDG25 is required for the proper levels of *FLOWERING LOCUS C* (*FLC*) expression; in loss-of-function mutation, SDG25 has an early flowering phenotype and associated with suppression of *FLC* expression.	Berr et al. 2009; Tamada et al. 2009

However, several recent studies have demonstrated that functional genes can occasionally be generated from processed genes and that these retrogenes take on a non-redundant functional role (Kong et al. 2004; Benovoy and Drouin 2006; Wang et al. 2006). As an indication of functionality, the EST evidences of the orphan retrogenes (*PpSDG1512* and *PpSDG1507*) actively transcribed have been validated by searching EST databases (Fig. 1 and Table S1). It is conceivable that with more genomic sequences available, more functional orphan retrogenes will be identified.

Lecharny et al. (2003) grouped gene families into three categories with respect to the retention of introns among genes: families of genes with a conserved structure, families of genes with almost no introns, and families of genes with an unconserved structure among paralogs but with a conserved structure within orthologs. This should be modified as we have shown that the structures of *Trx SET* genes are unconserved both among paralogs and within orthologs. Intron indels, gene retroposition events, and gene duplications were the three main mechanistic forces for *Trx SET* gene family diversity. Molecular clock analyses revealed that green algae and land plants appeared about 700 and 500 mya ago, respectively (Zimmer et al. 2007), and thus nonconservation of gene structures of *Trx SET* genes between green algae and land plants might imply that the ancient gene structures of land plants were established between 500 and 700 mya.

We found that plant *Trx SET* genes of different groups or clades have different functions (Table 3), suggesting that they interact with different substrates. Previous reports showed that mutants in *Arabidopsis Trx SET* genes from cpTrx clade (*ATX1*/*SDG27*, *ATX2*/*SDG30*, and *ATXR7*/*SDG25*) display locus-specific defects in H3K4me, whereas *SDG2* (*ATXR3*) from III-3 group leads to severe and pleiotropic phenotypes as well as the misregulation of a large number of genes (Guo et al. 2010; Berr et al. 2009, 2010; Tamada et al. 2009; Saleh et al. 2008; Alvarez-Venegas and Avramova 2005); these findings conform to the result of present phylogenetic analyses. The functional characterization of *Arabidopsis* Trx SET proteins from III-2 group are lacking so far; however, the highly conserved domain organization different from other orthologous groups suggest different functions for this group of proteins. As the structures of green algal *CrSDG3404*, *CrSDG3411*, and *CrSDG3401* genes did not correspond with those of other *Trx SET* genes of land plant within respective groups, we suggest that functional divergence might have occurred between the single cell green alga and land plants.

In conclusion, our study provides a novel phylogenetic relationship, which includes most members analyzed except for the III-3 group. Some new insights into the evolution of the plant *Trx SET* gene family in land plants were obtained. We found that the PostSET is not a common domain in plant cpTrx SET proteins (Fig. 1); it might be an ancestral characteristic but was lost in some members of III-4 group during the evolution. We propose that the PWWP, FYR, and PHD domains were integrated into primordial domain organization (SET–PostSET) before the emergence of land plants and resulted in function differentiation. Plant *Trx SET* genes exhibit a diversity of structures, even in the conserved SET domain region. At least four classes of gene structures had appeared before the divergence of *P. patens* (moss) or *S. moellendorfii*

(fern) from other land plants through frequent intron loss and gain. We identified three intronless orphan members that probably originated from retroposition events; their parental genes with introns might be lost during evolution. Our results also revealed the structural differences among evolutionary groups of plant *Trx SET* genes, indicating that the functions of *Trx SET* genes of the III-2 group may differ from other *Trx SET* genes.

Acknowledgements

This research was financially supported by State Key Laboratory of Systematic and Evolutionary Botany (Grant No. LSEB2011-10), Natural Science Fund for Colleges and Universities in Jiangsu Province (Grant No. 09KJB180006, 10KJB210004), and National Natural Science Foundation of China (Grant No. NSFC 31000729).

References

Abascal, F., Zardoya, R., and Posada, D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics, **21**(9): 2104–2105. doi:10.1093/bioinformatics/bti263. PMID:15647292.

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**(2): 215–221. doi:10.1016/S0378-1119(01)00524-8. PMID:11418242.

Alvarez-Venegas, R., and Avramova, Z. 2005. Methylation patterns of histone H3 Lys 4, Lys 9 and Lys 27 in transcriptionally active and inactive *Arabidopsis* genes and in *atx1* mutants. Nucleic Acids Res. **33**(16): 5199–5207. doi:10.1093/nar/gki830. PMID: 16157865.

Alvarez-Venegas, R., Pien, S., Sadder, M., Witmer, X., Grossniklaus, U., and Avramova, Z. 2003. ATX-1, an *Arabidopsis* homolog of trithorax, activates flower homeotic genes. Curr. Biol. **13**(8): 627–637. doi:10.1016/S0960-9822(03)00243-4. PMID:12699618.

Alvarez-Venegas, R., Sadder, M., Tikhonov, A., and Avramova, Z. 2006. Origin of the bacterial SET domain genes: vertical or horizontal? Mol. Biol. Evol. **24**(2): 482–497. doi:10.1093/molbev/msl184. PMID:17148507.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, **408**(6814): 796–815. doi:10.1038/35048692. PMID:11130711.

Baumbusch, L.O., Thorstensen, T., Krauss, V., Fischer, A., Naumann, K., Assalkhou, R., et al. 2001. 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 Res. **29**(21): 4319–4333. doi:10.1093/nar/29.21.4319. PMID:11691919.

Benovoy, D., and Drouin, G. 2006. Processed pseudogenes, processed genes, and spontaneous mutations in the *Arabidopsis* genome. J. Mol. Evol. **62**(5): 511–522. doi:10.1007/s00239-005-0045-z. PMID:16612535.

Berr, A., Xu, L., Gao, J., Cognat, V., Steinmetz, A., Dong, A., and Shen, W.H. 2009. *SET DOMAIN GROUP25* encodes a histone methyltransferase and is involved in *FLOWERING LOCUS C* activation and repression of flowering. Plant Physiol. **151**(3): 1476–1485. doi:10.1104/pp.109.143941. PMID:19726574.

Berr, A., McCallum, E.J., Menard, R., Meyer, D., Fuchs, J., Dong, A., and Shen, W.H. 2010. *Arabidopsis SET DOMAIN GROUP2* is required for H3K4 trimethylation and is crucial for both sporophyte and gametophyte development. Plant Cell, **22**(10): 3232–3248. doi:10.1105/tpc.110.079962. PMID:21037105.

Boudet, N., Aubourg, S., Toffano-Nioche, C., Kreis, M., and Lecharny, A. 2001. Evolution of intron/exon structure of DEAD helicase family genes in *Arabidopsis*, *Caenorhabditis*, and *Drosophila*. Genome Res. **11**(12): 2101–2114. doi:10.1101/gr.200801. PMID:11731501.

Briggs, S.D., Bryk, M., Strahl, B.D., Cheung, W.L., Davie, J.K., Dent, S.Y., et al. 2001. Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in *Saccharomyces cerevisiae*. Genes Dev. **15**(24): 3286–3295. doi:10.1101/gad.940201. PMID:11751634.

Cannon, S.B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., et al. 2006. Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. Proc. Natl. Acad. Sci. U.S.A. **103**(40): 14959–14964. doi:10.1073/pnas.0603228103. PMID:17003129.

Chan, A.P., Pertea, G., Cheung, F., Lee, D., Zheng, L., Whitelaw, C., et al. 2006. The TIGR Maize Database. Nucleic Acids Res. **34** (Database issue): D771–D776. doi:10.1093/nar/gkj072. PMID: 16381977.

Clay, N.K., and Nelson, T. 2005. The recessive epigenetic swellmap mutation affects the expression of two step II splicing factors required for the transcription of the cell proliferation gene *STRUWWELPETER* and for the timing of cell cycle arrest in the *Arabidopsis* leaf. Plant Cell, **17**(7): 1994–2008. doi:10.1105/tpc.105.032771. PMID:15937226.

Dorn, R., Krauss, V., Reuter, G., and Saumweber, H. 1993. 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. Proc. Natl. Acad. Sci. U.S.A. **90**(23): 11376–11380. doi:10.1073/pnas.90.23.11376. PMID:8248257.

Fablet, M., Bueno, M., Potrzebowski, L., and Kaessmann, H. 2009. Evolutionary origin and functions of retrogene introns. Mol. Biol. Evol. **26**(9): 2147–2156. doi:10.1093/molbev/msp125. PMID: 19553367.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**(6): 368–376. doi:10.1007/BF01734359. PMID:7288891.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., et al. 2006. Pfam: clans, web tools and services. Nucleic Acids Res. **34**(Database issue): D247–D251. doi:10.1093/nar/gkj149. PMID:16381856.

Freund, C., Dötsch, V., Nishizawa, K., Reinherz, E.L., and Wagner, G. 1999. The GYF domain is a novel structural fold that is involved in lymphoid signaling through proline-rich sequences. Nat. Struct. Biol. **6**(7): 656–660. doi:10.1038/10712. PMID: 10404223.

Fu, H., Kim, S.Y., and Park, W.D. 1995. 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, **7**(9): 1387–1394. doi:10.1105/tpc.7.9.1387. PMID:8589623.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science, **296**(5565): 92–100. doi:10.1126/science.1068275. PMID:11935018.

Graur, D., and Li, W. 2000. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, Mass.

Guindon, S., and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**(5): 696–704. doi:10.1080/10635150390235520. PMID:14530136.

Guo, L., Yu, Y., Law, J.A., and Zhang, X. 2010. SET DOMAIN GROUP2 is the major histone H3 lysine 4 trimethyltransferase in *Arabidopsis*. Proc. Natl. Acad. Sci. U.S.A. **107**(43): 18557–18562. doi:10.1073/pnas.1010478107. PMID:20937886.

Horan, K., Lauricha, J., Bailey-Serres, J., Raikhel, N., and Girke, T. 2005. Genome cluster database. A sequence family analysis platform for *Arabidopsis* and rice. Plant Physiol. **138**(1): 47–54. doi:10.1104/pp.104.059048. PMID:15888677.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al.French-Italian Public Consortium for Grapevine Genome Characterization. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, **449** (7161): 463–467. doi:10.1038/nature06148. PMID:17721507.

Jenuwein, T., and Allis, C.D. 2001. Translating the histone code. Science, **293**(5532): 1074–1080. doi:10.1126/science.1063127. PMID:11498575.

Jones, R.S., and Gelbart, W.M. 1993. The *Drosophila* Polycomb-group gene *Enhancer of zeste* contains a region with sequence similarity to *trithorax*. Mol. Cell. Biol. **13**(10): 6357–6366. doi:10.1128/MCB.13.10.6357 . PMID:8413234.

Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**(3): 275–282. PMID:1633570.

Kong, H., Leebens-Mack, J., Ni, W., dePamphilis, C.W., and Ma, H. 2004. Highly heterogeneous rates of evolution in the *SKP1* gene family in plants and animals: functional and evolutionary implications. Mol. Biol. Evol. **21**(1): 117–128. doi:10.1093/molbev/msh001. PMID:14595103.

Kullback, S., and Leibler, R.A. 1951. On information and sufficiency. Ann. Math. Stat. **22**(1): 79–86. doi:10.1214/aoms/1177729694.

Lecharny, A., Boudet, N., Gy, I., Aubourg, S., and Kreis, M. 2003. Introns in, introns out in plant gene families: a genomic approach of the dynamics of gene structure. J. Struct. Funct. Genomics, **3**(1–4): 111–116. doi:10.1023/A:1022614001371. PMID:12836690.

Lee, Y., Tsai, J., Sunkara, S., Karamycheva, S., Pertea, G., Sultana, R., et al. 2005. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res. **33**(Database issue): D71–D74. doi:10.1093/nar/gki064. PMID:15608288.

Lespinet, O., Wolf, Y.I., Koonin, E.V., and Aravind, L. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res. **12**(7): 1048–1059. doi:10.1101/gr.174302. PMID:12097341.

Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., and Bork, P. 2006. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res. **34**(Database issue): D257–D260. doi:10.1093/nar/gkj079. PMID:16381859.

Li, W.H., Gu, Z., Wang, H., and Nekrutenko, A. 2001. Evolutionary analyses of the human genome. Nature, **409**(6822): 847–849. doi:10.1038/35057039. PMID:11237007.

Lin, H., Zhu, W., Silva, J.C., Gu, X., and Buell, C.R. 2006. Intron gain and loss in segmentally duplicated genes in rice. Genome Biol. **7**(5): R41. doi:10.1186/gb-2006-7-5-r41. PMID:16719932.

Lynch, M., and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. Science, **290**(5494): 1151–1155. doi:10.1126/science.290.5494.1151. PMID:11073452.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., et al. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science, **318**(5848): 245–250. doi:10.1126/science.1143609. PMID:17932292.

Nei, M., and Kumar, S. 2000. Molecular evolution and phylogenetics. Oxford University Press, Oxford, U.K.

Ng, D.W., Wang, T., Chandrasekharan, M.B., Aramayo, R., Kertbundit, S., and Hall, T.C. 2007. Plant SET domain-containing proteins: structure, function and regulation. Biochim. Biophys. Acta, **1769**(5–6): 316–329. doi:10.1016/j.bbaexp.2007.04.003. PMID:17512990.

Owen, D.J., Ornaghi, P., Yang, J.-C., Lowe, N., Evans, P.R., Ballario, P., et al. 2000. The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase gcn5p. EMBO J. **19**(22): 6141–6149. doi:10.1093/emboj/19.22.6141. PMID:11080160.

Park, K.-C., Kwon, S.-J., Kim, P.-H., Bureau, T., and Kim, N.-S. 2008. Gene structure dynamics and divergence of the polygalacturonase gene family of plants and fungus. Genome, **51**(1): 30–40. doi:10.1139/G07-093. PMID:18356937.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature, **457**(7229): 551–556. doi:10.1038/nature07723. PMID:19189423.

Reeves, R., and Nissen, M.S. 1990. The A.T-DNA-binding domain of mammalian high mobility group I chromosomal proteins. A novel peptide motif for recognizing DNA structure. J. Biol. Chem. **265** (15): 8573–8582. PMID:1692833.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., et al. 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science, **319**(5859): 64–69. doi:10.1126/science.1150646. PMID:18079367.

Rogozin, I.B., Lyons-Weiler, J., and Koonin, E.V. 2000. Intron sliding in conserved gene families. Trends Genet. **16**(10): 430–432. doi:10.1016/S0168-9525(00)02096-5. PMID:11050324.

Roguev, A., Schaft, D., Shevchenko, A., Pijnappel, W.W., Wilm, M., Aasland, R., and Stewart, A.F. 2001. The *Saccharomyces cerevisiae* Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. EMBO J. **20**(24): 7137–7148. doi:10.1093/emboj/20.24.7137. PMID:11742990.

Roy, S.W., Fedorov, A., and Gilbert, W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc. Natl. Acad. Sci. U.S.A. **100**(12): 7158–7162. doi:10.1073/pnas.1232297100. PMID:12777620.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4** (4): 406–425. PMID:3447015.

Saleh, A., Alvarez-Venegas, R., Yilmaz, M., Le, O., Hou, G., Sadder, M., et al. 2008. The highly similar *Arabidopsis* homologs of Trithorax ATX1 and ATX2 encode proteins with divergent biochemical functions. Plant Cell, **20**(3): 568–579. doi:10.1105/tpc.107.056614. PMID:18375658.

Sidow, A., Nguyen, T., and Speed, T.P. 1992. Estimating the fraction of invariable codons with a capture–recapture method. J. Mol. Evol. **35**(3): 253–260. doi:10.1007/BF00178601. PMID:1518092.

Sonnhammer, E.L., and Koonin, E.V. 2002. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet. **18** (12): 619–620. doi:10.1016/S0168-9525(02)02793-2. PMID:12446146.

Springer, N.M., Napoli, C.A., Selinger, D.A., Pandey, R., Cone, K.C., Chandler, V.L., et al. 2003. Comparative analysis of SET domain proteins in maize and *Arabidopsis* reveals multiple duplications preceding the divergence of monocots and dicots. Plant Physiol. **132**(2): 907–925. doi:10.1104/pp.102.013722. PMID:12805620.

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**(2–3): 209–223. doi:10.1016/0925-4773(95)00402-M. PMID:8541210.

Stec, I., Nagl, S.B., van Ommen, G.J., and den Dunnen, J.T. 2000. The PWWP domain: a potential protein–protein interaction domain in nuclear proteins influencing differentiation? FEBS

Lett. **473**(1): 1–5. doi:10.1016/S0014-5793(00)01449-6. PMID: 10802047.

Suyama, M., Torrents, D., and Bork, P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. **34**(Web Server issue): W609–W612. doi:10.1093/nar/gkl315. PMID:16845082.

Tamada, Y., Yun, J.Y., Woo, S.C., and Amasino, R.M. 2009. *ARABIDOPSIS TRITHORAX-RELATED7* is required for methylation of lysine 4 of histone H3 and for transcriptional activation of *FLOWERING LOCUS C*. Plant Cell, **21**(10): 3257–3269. doi:10.1105/tpc.109.070060. PMID:19855050.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28**(10): 2731–2739. doi:10.1093/molbev/msr121. PMID:21546353.

Thakur, S., Jha, S., and Chattoo, B.B. 2011. CastorDB: a comprehensive knowledge base for *Ricinus communis*. BMC Res. Notes, **4**: 356. doi:10.1186/1756-0500-4-356. PMID: 21914200.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**(24): 4876–4882. doi:10.1093/nar/25.24.4876. PMID:9396791.

Tian, W., and Skolnick, J. 2003. How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol. **333**(4): 863–882. doi:10.1016/j.jmb.2003.08.057. PMID: 14568541.

Trapp, S.C., and Croteau, R.B. 2001. Genomic organization of plant terpene synthases and molecular evolutionary implications. Genetics, **158**(2): 811–832. PMID:11404343.

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**(16): 3822–3831. PMID:7915232.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science, **313**(5793): 1596–1604. doi:10.1126/science.1128691. PMID:16973872.

Wang, W., Tanurdzic, M., Luo, M., Sisneros, N., Kim, H.R., Weng, J.K., et al. 2005. Construction of a bacterial artificial chromosome library from the spikemoss *Selaginella moellendorffii*: a new resource for plant comparative genomics. BMC Plant Biol. **5**: 10. doi:10.1186/1471-2229-5-10. PMID:15955246.

Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., Cai, Z., et al. 2006. High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell, **18**(8): 1791–1802. doi:10.1105/tpc.106.041905. PMID:16829590.

Wattler, S., Russ, A., Evans, M., and Nehls, M. 1998. A combined analysis of genomic and primary protein structure defines the phylogenetic relationship of new members if the T-box family. Genomics, **48**(1): 24–33. doi:10.1006/geno.1997.5150. PMID: 9503012.

Wheelan, S.J., Church, D.M., and Ostell, J.M. 2001. Spidey: a tool for mRNA-to-genomic alignments. Genome Res. **11**(11): 1952–1957. PMID:11691860.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**(6): 1396–1401. PMID:8277861.

Ying, Z., Mulligan, R.M., Janney, N., and Houtz, R.L. 1999. Rubisco small and large subunit *N*-methyltransferases. Bi- and mono-functional methyltransferases that methylate the small and large subunits of Rubisco. J. Biol. Chem. **274**(51): 36 750 – 36 756. doi:10.1074/jbc.274.51.36750. PMID:10593982.

Zhang, Y., Wu, Y., Liu, Y., and Han, B. 2005. Computational identification of 69 retroposons in *Arabidopsis*. Plant Physiol. **138**(2): 935–948. doi:10.1104/pp.105.060244. PMID:15923328.

Zhu, X., Ma, H., and Chen, Z. 2011. Phylogenetics and evolution of *Su(var)3-9 SET* genes in land plants: rapid diversification in structure and function. BMC Evol. Biol. **11**: 63. doi:10.1186/1471-2148-11-63. PMID:21388541.

Zimmer, A., Lang, D., Richardt, S., Frank, W., Reski, R., and Rensing, S.A. 2007. Dating the early evolution of plants: detection and molecular clock analyses of orthologs. Mol. Genet. Genomics, **278**(4): 393–402. doi:10.1007/s00438-007-0257-6. PMID: 17593393.

This article has been cited by:

1. Yuzhu Chen, Jun Cao. 2014. Comparative genomic analysis of the Sm gene family in rice and maize. *Gene* . [CrossRef]