J Mol Evol (2010) 71:202–218
DOI 10.1007/s00239-010-9378-3

Evolution of the Cinnamyl/Sinapyl Alcohol Dehydrogenase (CAD/SAD) Gene Family: The Emergence of Real Lignin is Associated with the Origin of Bona Fide *CAD*

Dong-Mei Guo · Jin-Hua Ran · Xiao-Quan Wang

Received: 2 June 2010 / Accepted: 26 July 2010 / Published online: 19 August 2010

Abstract Lignin plays a vital role in plant adaptation to terrestrial environments. The cinnamyl alcohol dehydrogenase (CAD) catalyzes the last step in monolignol biosynthesis and might have contributed to the lignin diversity in plants. To investigate the evolutionary history and functional differentiation of the *CAD* gene family, we made a comprehensive evolutionary analysis of this gene family from 52 species, including bacteria, early eukaryotes and green plants. The phylogenetic analysis, together with gene structure and function, indicates that all members of land plants, except two of moss, could be divided into three classes. Members of Class I (bona fide *CAD*), generally accepted as the primary genes involved in the monolignol biosynthesis, are all from vascular plants, and form a robustly supported monophyletic group with the lycophyte *CADs* at the basal position. This class is also conserved in the predicted three-dimensional structure and the residues constituting the substrate-binding pocket of the proteins. Given that *Selaginella* has real lignin, the above evidence strongly suggests that the earliest occurrence of the bona fide *CAD* in the lycophyte could be directly correlated with the origin of lignin. Class II comprises members more similar to the aspen sinapyl alcohol dehydrogenase gene, and includes three groups corresponding to lycophyte, gymnosperm, and angiosperm. Class III is conserved in land plants. The three classes differ in patterns of evolution and expression, implying that functional divergence has occurred among them. Our study also supports the hypothesis of convergent evolution of lignin biosynthesis between red algae and vascular plants.

Keywords CAD/SAD · Gene family evolution · Convergent evolution · Functional divergence · Origin of lignin · Vascular plant · Red alga

Introduction

In the evolutionary history of flora, the most significant event that could be compared with the ‘‘Cambrian explosion’’ of marine faunas should be the origin and primary radiation of land plants (also known as embryophytes, comprising bryophytes, pteridophytes, and seed plants), which live primarily in terrestrial habitats (Bateman et al. 1998). The conquest of the land by plants might be a long process of slow adjustment to a new and inhospitable environment, which not only needs morphological innovations, but also requires numerous physiological and molecular adaptations, including metabolic pathways leading to lignins, flavonoids, cutins, plant hormones, and so on (Kenrick and Crane 1997; Waters 2003). The ability to synthesize lignin has been widely accepted as a key innovation in the evolutionary adaptation of plants from an aqueous to a gaseous environment, which could provide plants with mechanical support, water and solute transport, biotic, and abiotic stress resistance (Whetten et al. 1998;

Electronic supplementary material The online version of this article (doi:10.1007/s00239-010-9378-3) contains supplementary material, which is available to authorized users.

Dong-Mei Guo and Jin-Hua Ran contributed equally to the work.

D.-M. Guo · J.-H. Ran · X.-Q. Wang (✉)
State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, The Chinese Academy of Sciences, 20 Nanxincun, Xiangshan, Beijing 100093, China
e-mail: xiaoq_wang@ibcas.ac.cn

D.-M. Guo
Graduate University of the Chinese Academy of Sciences, Beijing 100039, China

Boerjan et al. 2003; Peter and Neale 2004; Vanholme et al. 2008). Lignins, deposited mainly in the secondarily thickened plant cell walls, are complex aromatic heteropolymers derived mainly from three hydroxycinnamyl alcohol monomers named *p*-coumaryl (H), guaiacyl (G), and syringyl (S). The variation of monomer content and composition results in the lignin diversity among different plant lineages (Whetten et al. 1998; Boerjan et al. 2003; Barceló et al. 2007; Vanholme et al. 2008).

The cinnamyl alcohol dehydrogenase (CAD) plays a vital role in the monolignol biosynthesis and catalyzes the last step in this pathway that reduces hydroxycinnamyl aldehydes into their corresponding alcohols (Gross et al. 1973; Mansell et al. 1974). The CAD is encoded by a multigene family and its homologs have been detected widely in bacteria and eukaryota except animals (e.g., Larroy et al. 2002; Kim et al. 2004; Mee et al. 2005; Tobias and Chow 2005; Barakat et al. 2009; Saballos et al. 2009). Tissue-specific expression analyses of *CADs* from *Arabidopsis* and *Populus* and enzyme kinetic characterization studies of *Arabidopsis* CADs showed that different members of this gene family could have distinct expression patterns and substrate specificities, implying that they might play diverse roles during plant development (Kim et al. 2004, 2007; Barakat et al. 2009). It is interesting that the CAD orthologs from gymnosperms have high specificity for coniferaldehyde and low affinity for sinapaldehyde, unlike those from angiosperms that show high affinity to the both substrates. This might have contributed to the great difference in lignin content and composition between gymnosperms and angiosperms (e.g., Kutsuki et al. 1982; Goffner et al. 1992; Galliano et al. 1993a, b; Hawkins and Boudet 1994). Has the *CAD* gene family diverged in structure and function in non-seed plants?

On the other hand, the sinapyl alcohol dehydrogenase (SAD), a homolog of CAD firstly reported in aspen (PtSAD:AAK58693), was suggested as a key enzyme involved in the sinapyl monolignol biosynthesis of angiosperms (Li et al. 2001). Recently, however, there is a hot debate about whether SAD has played a role in the formation of S lignin (e.g., Kim et al. 2004, 2007; Sibout et al. 2005; Stephens 2005; Goldie 2006; Youn et al. 2006; Barakat et al. 2009). To date, very little has been known about the functional diversification of the *CAD* gene family, except that bona fide *CADs*, such as *AthCAD4* (*AtCAD-C*:At3g19450) and *AthCAD5* (*AtCAD-D*: At4g34230) from *Arabidopsis*, are doubtlessly involved in the monolignol biosynthesis and conserved between gymnosperms and angiosperms (e.g., Sibout et al. 2005; Youn et al. 2006).

With the availability of genome sequence data, a couple of phylogenetic analyses have recently been performed on *CAD*/*SADs* from *Arabidopsis*, *Medicago*, *Oryza*, *Populus*, *Sorghum*, and *Vitis*, as well as some other species without whole genome sequences (e.g., Raes et al. 2003; Tobias and Chow 2005; Hamberger et al. 2007; Barakat et al. 2009; Saballos et al. 2009), indicating that *CAD*/*SAD*s could be divided into two to five clades (excluding some species-specific homologs). In particular, Barakat et al. (2009) constructed a phylogeny of this gene family using sequences from five angiosperms having whole genome sequence information and some gymnosperm ESTs as well as some other angiosperm *CAD*/*SAD*s, and classified the *CAD*/*SAD*s into three classes. Class I was weakly supported by a bootstrap value of 57%, but it included two robustly supported subclades, one comprising sequences from both gymnosperms and angiosperms and the other comprising only gymnosperm sequences. Classes II and III were angiosperm-specific, with <50 and 100% bootstrap support, respectively. Although the three-class classification seems to have outlined the evolutionary relationship of the *CAD*/*SAD*s in seed plants, the study of Barakat et al. (2009) did not include samples from basal land plants, and thus it could not investigate the origin and early diversification of the *CAD*/*SAD* gene family. Moreover, the evolutionary relationships between bona fide *CAD* and its homologs, as well as the correlations between evolutionary patterns of the *CAD* gene family and variations of lignin content and composition in different plant groups, are poorly understood.

Very recently, a comparative genomic analysis was performed for 14 plant and one fungal species to investigate evolution of the genes involved in the monolignol biosynthesis (Xu et al. 2009a). The authors suggested that the lignin biosynthesis pathway had been completely established in moss, in consideration to its harboring of all monolignol biosynthesis genes (except Ferulate 5-hydroxylase, F5H) and many more members of these gene families than green alga. However, this study did not explore the phylogenetic history of the *CAD* genes that are responsible for the last step of this pathway. Furthermore, as lignin was also reported in the red alga *Calliarthron cheilosporioides* (Martone et al. 2009), it was supposed that the lignin biosynthesis pathway might have existed before the divergence between green and red algae or have experienced convergent evolution between red algae and land plants (Xu et al. 2009a). It would be very interesting to investigate the distribution of the lignin biosynthesis genes such as *CAD* in red algae and other early eukaryotes.

A reliable phylogeny of a gene family is expected to give important clues for understanding its evolutionary history and functional differentiation. At the same time, comparative genomic analysis has been proved as a very efficient approach to retrieve the evolutionary histories of some important gene families (e.g., Nam et al. 2004; Bowman et al. 2007; Xu et al. 2009b). As more and more genomes have been sequenced, we could compare gene

families from different species thoroughly. In order to investigate the evolutionary history and functional differentiation of the *CAD*/*SAD* gene family and its correlation with the origin and evolution of lignin, here we made a comprehensive evolutionary analysis of this gene family from 52 species including bacteria, early eukaryotes and main lineages of green plants (also known as Viridiplantae, comprising green algae and land plants).

Materials and Methods

Plant Materials Used in the Sequencing and in the Identification of *CAD*/*SAD* Sequences

Eleven species of gymnosperms were sampled to clone the putative *SAD* gene (Supplementary Table S1). Voucher specimens are deposited in the herbarium of Institute of Botany, Chinese Academy of Sciences (PE). The DNA sequences determined in this study are deposited in GenBank under accession numbers HM185279–HM185296. Some additional *CAD*/*SAD* sequences of gymnosperms were retrieved from GenBank. The tBLASTn searches were performed in publicly available genome databases, using all known *Arabidopsis thaliana* CAD/SAD protein sequences (Kim et al. 2004) as queries. Besides plants, *CAD*/*SAD* sequences were also obtained from bacteria, fungi, and some other early eukaryotes (Supplementary Table S2). We also searched the EST databases of most of the above species from NCBI using the tBLASTn search, since EST data could provide some gene expression information. The target sequences were selected when the pairwise amino acid identity between the queries and the targets is over than 40% (Tian and Skolnick 2003). Finally, we obtained the *CAD*/*SAD* sequences from 52 species, 39 of which are green plants, including green alga (2 species), liverwort (1), moss (1), lycophyte (1), fern (2), gymnosperm (17), and angiosperm (15). All these species have available whole genome sequences except the liverwort, ferns, gymnosperms, and one angiosperm (*Nicotiana tabacum*).

DNA and RNA Extraction, PCR and RT-PCR Amplification, Cloning and Sequencing

Genomic DNA was extracted from leaves using the modified cetyltrimethylammonium bromide (CTAB) method (Doyle and Doyle 1987; Rogers and Bendich 1988). Total RNA was prepared from fresh leaves using the Plant RNA Purification Reagent (Invitrogen, Carlsbad, CA), digested with RNase-free DNase I (Promega, Madison, USA), and then purified by Oligotex mRNA Mini kit (Qiagen, Hilden, Germany). First-strand cDNA was produced using Superscript III Reverse Transcriptase (Invitrogen) according to the manufacturer's instructions.

The primers used to amplify the *SAD* gene were designed based on the EST sequences that were obtained from the tBLASTn searches in the EST databases of gymnosperms in GenBank with the aspen SAD protein sequence (Li et al. 2001) as query. The PCR amplification was conducted in a Tpersonal Thermocycler or a T1 Thermocycler (Biometra, Goettingen, Germany) using the genomic DNA or cDNA as templates. The PCR products were purified using the TIANgel Midi Purification kit (Qiagen) and then cloned with the pGEM-T Easy Vector System II (Promega). For each species, 24 clones were screened by comparing restriction fragments of *Eco*RI or/and *Hinf*I. All distinct clones with correct insertion were sequenced in both directions. Sequencing reactions were performed with T7, SP6 and one internal primer SAD2F (5′-AGAGGTGAAGAATTTCGCTGTT) using the ABI Prism Bigdye Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems, Foster City, CA, USA). After precipitation in 95% EtOH and 10 M NH_4Ac (50:4 μl), the sequencing products were separated on a 96-capillary 3730XL DNA analyzer (Applied Biosystems). The species names, gene names, gene accession numbers, sources of genome databases, and primers are listed in Supplementary Tables S1 and S2.

Phylogenetic Reconstruction

Coding sequences of the *CAD*/*SAD* genes were aligned using the program Clustal X version 2.0 (Thompson et al. 1997) and manually adjusted in BioEdit version 7.0.9 (Hall 1999). After removing unalignable sequences at the N- and C-terminals, the poorly aligned positions in the alignment were further eliminated using the Gblocks server (http://molevol.cmima.csic.es/castresana/Gblocks_server.html). DAMBE version 5.1.1 (Xia and Xie 2001) was used to check for substitution saturation of each codon position and the results indicated that the third codon positions were saturated. Consequently, only the first and the second codon positions were used in the phylogenetic reconstruction. The jModeltest 0.1.1 (Posada 2008) was used to determine the best-fit model of nucleotide sequence evolution, and the GTR + I + G model was suggested as the best using both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). ProtTest version 2.4 (Abascal et al. 2005) was used to identify the best-fit model of amino acid evolution, and the WAG + I + G model was selected as the best using both AIC and BIC.

At first, the maximum-likelihood (ML) trees of the *CAD*/*SAD* genes from all the 52 species were constructed based on the deduced amino acid sequences (substitution model, WAG; bootstrap, 100; gamma distribution

parameter, estimated) and the first plus second codon positions (substitution model, GTR; bootstrap, 100; gamma distribution parameter, estimated), respectively, using PhyML version 2.4.4 (Guindon and Gascuel 2003) with the two members of bacteria as outgroups. The results indicated that both green and land plants are monophyletic (Supplementary Fig. S1). Therefore, we further conducted phylogenetic analyses with only the *CAD*/*SAD* genes from land plants using the two green algae as outgroups. Sequences from the liverwort and the fern *Adiantum capillus-veneris* were also excluded since the short partial EST sequences might mislead phylogenetic inference. Finally, the phylogenetic trees were constructed based on a land-plant *CAD*/*SAD* dataset comprising 35 land plant and two outgroup species. For this dataset, the best-fit model of nucleotide sequence evolution was the same as the former, while that of deduced amino acid sequence evolution was JTT + I + G. Bayesian and ML trees were constructed based on nucleotide sequence (excluding the third codon positions) with MrBayes version 3.1.2 (Huelsenbeck and Ronquist 2001) (nst = 6; rates = gamma; ngen = 1000000) and PhyML (substitution model, GTR; bootstrap, 100; gamma distribution parameter, estimated), respectively. Also, the phylogeny of the *CAD*/*SAD* genes was constructed based on the inferred amino acid sequences, using MrBayes (prset aamodelpr = mixed; ngen = 1000000) and PhyML (substitution model, JTT; bootstrap, 100; gamma distribution parameter, estimated), respectively.

Selection Test

To identify the genes that have been subject to diversifying selection, the maximum likelihood analysis was conducted using both the Fitmodel program version 0.5.3 (Guindon et al. 2004) and the codeml program in the PAML version 4.2b (Yang 1997, 2007). The Fitmodel program, which allows the site-specific selection process to vary along lineages of a phylogenetic tree, was performed for the land-plant *CAD*/*SAD* dataset after removing all pseudogenes as well as some low quality and short sequences. Models conducted in this analysis included M0, M3, M3 + S1, and M3 + S2. M0 assumes that all the sites have the same ω values ($\omega = d_N/d_S$; d_N, non-synonymous substitution rates; d_s, synonymous substitution rates), whereas M3 assumes three different ω values ($\omega 1 < \omega 2 < \omega 3$). If the switching rates between ω values ($\omega 1$ to $\omega 2$, $\omega 1$ to $\omega 3$, $\omega 2$ to $\omega 3$) are equally imposed, the model is designated with S1, otherwise with S2. The likelihood ratio tests (LRTs) were performed between each pair of models M0/M3, M3/M3 + S1, and M3 + S1/M3 + S2, using a χ^2 to calculate the significance of difference. The posterior probabilities (PPs) were estimated by Fitmodel for placing the $\omega 3$ value of a site on different branches of the phylogenetic tree, and were visualized for each codon position using BASS4 (Bayesian Analysis of Selected Sites) provided by J. Huelsenbeck.

The codeml program was performed for the four clades resolved in the phylogenetic analyses (bona fide *CAD*, Class III, angiosperm *SAD* and gymnosperm *SAD*), separately. The site models (model = 0; NSsites = 0, 1, 2, 3, 7, 8), which allow the ω ratio to vary among sites (Nielsen and Yang 1998), were performed for the four clades. The likelihood ratio tests (LRTs) using a χ^2 with d.f. were conducted to test positive selection between each of the pairs of models M0/M3, M1a/M2a, and M7/M8, respectively. The four clades were also tested with the branch models that allow the ω ratio to vary among branches in the phylogeny (Yang 1998) and the branch-site model (model A, test 2) (model = 2, NSsites = 2) that allows to detect positive selection at a small number of sites along a specific lineage (Zhang et al. 2005), respectively. To set the foreground branches, we chose some trunk branches leading to the major lineages of vascular plants (also known as tracheophytes comprising pteridophytes and seed plants, with lignified tissues for conducting water, minerals, and photosynthetic products) and the branches in which ancestral gene duplications have occurred. The LRTs were used to compare the null model (in the branch model test: model = 0, NSsites = 0; in the branch-site model test: fix omega = 1 and omega = 1 in codeml.ctl) and the alternative model (in the branch model test: model = 2, NSsites = 0; in the branch-site model test: fix omega = 0 in codeml.ctl).

Prediction of the Three-Dimensional Structure of the CAD/SAD Protein

We selected several CAD/SAD protein sequences representing four main clades of non-flowering plants (green algae, mosses, lycophytes, and gymnosperms) to predict their three-dimensional models. These sequences were submitted to the Swiss-Model homology modeling server (http://swissmodel.expasy.org/workspace/) and were analyzed using the automatic modeling mode (Arnold et al. 2006).

Results

Sequence Characterization

The *CAD*/*SAD* genes obtained from the 52 studied species are shown in Supplementary Table S2. The numbers of the genes we discovered in some species (e.g., *Physcomitrella*, *Selaginella*, *Oryza*, and *Populus*) were different from those reported in some previous studies (e.g., Barakat et al. 2009;

Xu et al. 2009a). Only one member was found in the genome sequence of some early eukaryotes such as the red alga *Cyanidioschyzon merolae* and the three protist species. The *CAD/SAD* gene coding sequences of the sampled plant species range from 1044 to 1131 bp in length. Most of the length variation occurred at the N- and C-terminals (Supplementary Table S3). Copy number of this gene family is very variable in green plants, from three in green alga (only one or two of them were used in the final analysis), and moss to eighteen in *Medicago*. Vascular plants have many more members than green alga and moss (Table 1). Moreover, there is great variation in the gene structure. This gene has seven to eight introns in green algae, and fewer than six introns in land plants, which is consistent with the previous finding that early eukaryotic genes are often very complex in structure (Roy and Gilbert 2005) (Fig. 1; Supplementary Table S3). The number of introns also varied among different members of the gene family, especially in moss (1, 4, and 5 introns in *Physcomitrella patens*) and angiosperm (1–5 introns). However, all members from the lycophyte *Selaginella moellendorffii* have five introns, and those from gymnosperms have four or five introns (Fig. 1; Supplementary Table S3).

Phylogenetic Analyses

Both Bayesian and ML analyses were performed on the nucleotide sequence (excluding the third codon positions) and deduced amino acid sequence datasets of the *CAD/SAD* gene family from land plants, respectively. The four phylogenetic trees generated are topologically identical except the positions of some branches with low bootstrap supports (Fig. 1, Supplementary Fig. S2). The ML tree of the amino acid sequences is shown in Fig. 1. The land plants are monophyletic with 100% bootstrap support (Fig. 1). Except two members of moss (*PpaCAD1* and *PpaCAD2*), all the *CAD/SAD*s of land plants could be divided into three classes based on the phylogenetic relationships as well as the gene structure and function (Fig. 1, Supplementary Fig. S3). For the convenience of discussion, the class names used in this study followed previous

Table 1 Information of the *CAD/SAD* gene family in the 18 green plant species with whole genome sequences analyzed in this study

Species	Copy number				Website	References
	Class I	Class II	Class III	Total		
Chlamydomonas reinhardtii	–	–	–	3	http://genome.jgi-psf.org/Chlre3/Chlre3.home.html	Merchant et al. (2007)
Volvox carteri f. *nagariensis*	–	–	–	3	http://genome.jgi-psf.org/Volca1/Volca1.home.html	–
Physcomitrella patens ssp patens	–	–	1	3	http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html	Rensing et al. (2008)
Selaginella moellendorffii	2	5	4	11	http://genome.jgi-psf.org/Selmo1/Selmo1.home.html	Wang et al. (2005)
Arabidopsis thaliana	2	6	1	9	http://www.ncbi.nlm.nih.gov	Kim et al. (2004)[a]
Brachypodium distachyon	1	6	1	8	http://www.phytozome.net	Vogel et al. (2010)
Carica papaya	2	11	1	14	http://www.ncbi.nlm.nih.gov	Ming et al. (2008)
Cucumis sativus	2	4	1	7	http://www.ncbi.nlm.nih.gov	Huang et al. (2009)
Glycine max	2	11	2	15	http://www.phytozome.net	Schmutz et al. (2010)
Manihot esculenta	3	12	1	16	http://www.phytozome.net	–
Mimulus guttatus	1	6	1	8	http://www.phytozome.net	–
Medicago truncatula	1	16	1	18	http://www.medicago.org/genome	Cannon et al. (2006)
Oryza sativa ssp *japonica*	1	9	2	12	http://www.ncbi.nlm.nih.gov	Goff et al. (2002)
Phoenix dactylifera	2	5	2	9	http://www.ncbi.nlm.nih.gov	–
Populus trichocarpa	1	11	5	17	http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html	Hamberger et al. (2007)[a]
Ricinus communis	2	5	1	8	http://castorbean.tigr.org/	–
Sorghum bicolor	1	8	5	14	http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html	Paterson et al. (2009)
Vitis vinifera PN40024	2	9	3	14	http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis	Jaillon et al. (2007)

[a] Sequences that were directly downloaded with gene accession numbers from the reference papers, while the others were obtained by the blast search

classifications for this gene family (e.g., Raes et al. 2003; Hamberger et al. 2007; Barakat et al. 2009; Saballos et al. 2009). Class I (bona fide CAD) comprising *AthCAD5* and its closest relatives is monophyletic (100% bootstrap support) and conserved in vascular plants, with the lycophyte at a basal position. Class II (SAD: members more similar to PtSAD than to AthCAD5) is not monophyletic, which includes three groups corresponding to lycophyte (S-SAD), gymnosperm (G-SAD), and angiosperm (A-SAD), respectively (Table 2, also see "Discussion"). Class III is conserved from moss to angiosperm (Fig. 1), and it consists of members that are a little more similar to PtSAD than to AthCAD5, but relationships among the members from seed plants are far from being resolved due to the high sequence similarity (Table 2). Moreover, the three classes of the *CAD*/*SAD* genes we designated were also recovered in the ML trees of all the 52 species constructed based on the deduced amino acid sequences and the first plus second codon positions, respectively (Supplementary Fig. S1). The only one member (*CmeCAD*) found in the red alga *Cyanidioschyzon merolae* is distantly related to the *CAD*/*SAD*s of green plants. The phylogenetic positions of the four EST sequences obtained from the liverwort *Marchantia polymorpha* are uncertain. It is interesting that the fern *Adiantum capillus-veneris* also harbors members from all of the three classes of the *CAD*/*SAD* genes (Supplementary Fig. S1).

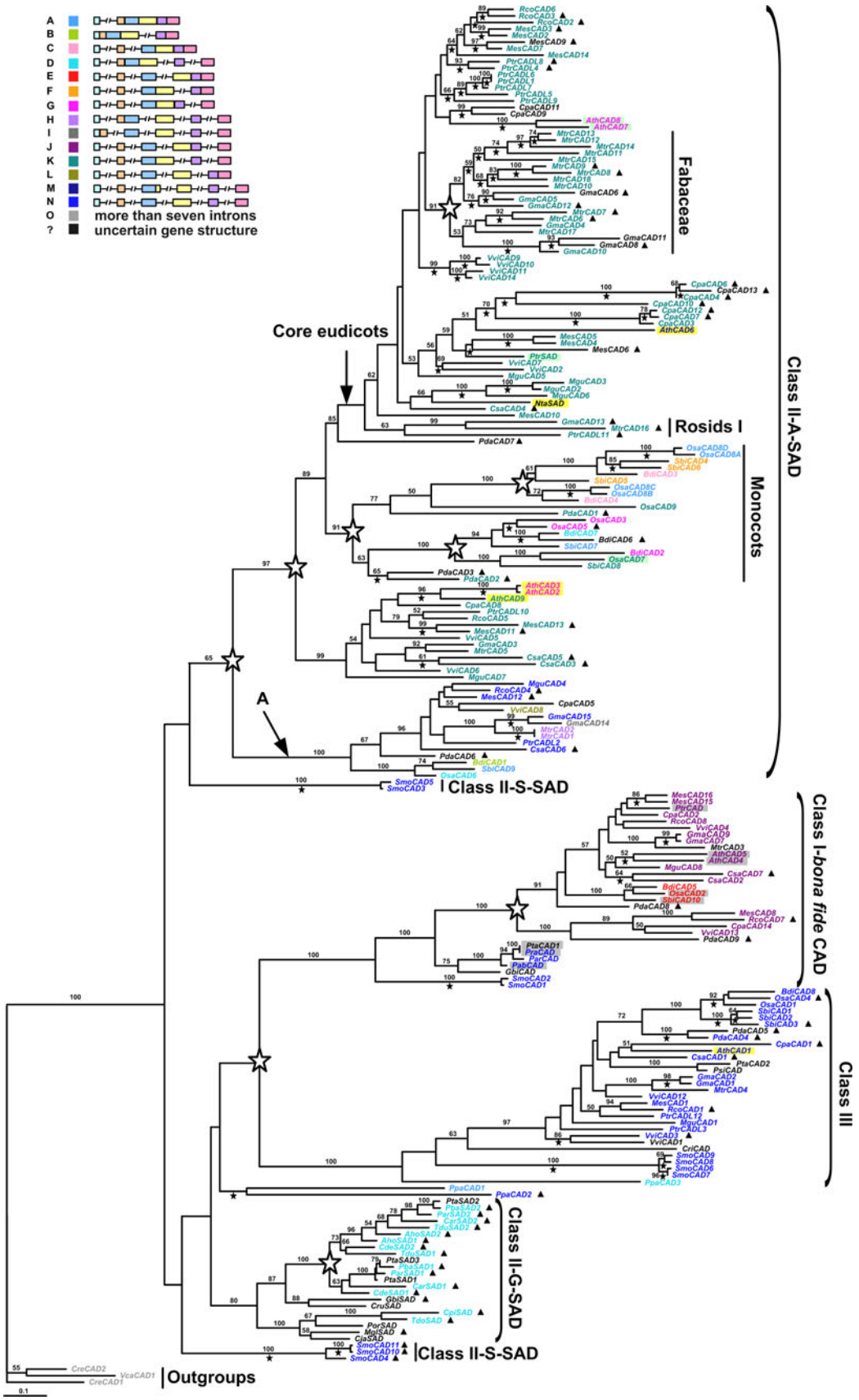
Duplication and extinction events, both ancient and recent, occurred frequently in the evolutionary history of the *CAD*/*SAD* gene family, especially in putative *SAD*s of angiosperm (A-SAD) (Fig. 1). In the A-SAD (Class II) clade, two duplications in the early evolution of angiosperms and one duplication in the common ancestor of monocots, as well as other duplications in the deep branches of some families such as Poaceae and Fabaceae, could have occurred. Additionally, many intraspecific duplications might have occurred, particularly in the core eudicots subclade. For instance, most members of this subclade are likely to have been generated by tandem repeat, including all members of *Arabidopsis* and some of the other species (e.g., Kim et al. 2004; Barakat et al. 2009; Saballos et al. 2009). At the same time, the putative *SAD* gene was frequently lost, such as the loss of members of *Arabidopsis* in subclade A and *Cucumis*, *Manihot*, and *Ricinus* in rosids I. In contrast, duplication and extinction events were much fewer in the other two classes. Except *Manihot esculenta*, no species has more than two copies of bona fide *CAD* (Class I), and most species has only one to three copies of Class III (both *Sorghum* and *Populus* have five copies of Class III, but some of them are too short to be included in the final phylogenetic analysis) (Table 1; Fig. 1). Nevertheless, one duplication in the early evolution of angiosperm bona fide *CAD* and subsequent loss of one copy in many species have occurred (Fig. 1).

Three-Dimensional Structure Modeling

The prediction of three-dimensional structure for CAD/SAD proteins indicated that AthCAD5 (PDB number: 2cf5A) was the most homologous template for SmoCAD1 and SmoCAD2 (Class I, bona fide CAD), while PtSAD (PDB number: 1yqdA) was the best template for CreCAD1 (green alga), PpaCAD1, PpaCAD3 (Class III of moss), SmoCAD3 and SmoCAD4 (lycophyte SADs, Class II), and PbaSAD1 (G-SAD, Class II) (Table 2). The similarity and identity of amino acid sequences between SmoCAD1 and AthCAD5 are as high as 77 and 62%, respectively (Table 2). Furthermore, among the twelve residues which constitute the proposed substrate-binding pocket (Youn et al. 2006), ten of SmoCAD1 and nine of SmoSAD2 are identical to those of AthCAD5, whereas there are six different residues between the two subclades of angiosperm bona fide CADs. Additionally, most of these residues are different from those of the members in Classes II and III (Supplementary Fig. S3).

Selection Test

The LRT tests between nested models in the Fitmodel program suggested that the M3 + S2 model was significantly better than the other models for the land-plant *CAD*/*SAD* dataset. The switching rate between $\omega 2$ (moderate purifying selection) and $\omega 3$ (relaxed selection) ($R23 = 5.86$) was significantly higher than that between $\omega 1$ (strong purifying selection) and $\omega 2$ ($R12 = 0.89$), and that between $\omega 1$ and $\omega 3$ ($R13 = 0.21$) (Table 3). Using the M3 + S2 model, 14% of the codons were inferred to be under relaxed selection ($\omega 3 = 1.01$) (Table 3), of which some might have been under positive selection. For the 380 trees corresponding to all codon positions in the alignment, relaxed selection was only detected in some branches of 136 trees. Among the 377 branches across the gene tree, the number of sites under relaxed selection on each branch varied from 0 to 44, and branches with a large number of sites under relaxed selection were mainly detected in the A-SAD clade (Figs. 2, 3). Interestingly, the pattern of shift to relaxed selection at the twelve codon positions, which constitute the proposed substrate-binding pocket (Youn et al. 2006), was very different among the *CAD*/*SAD* gene classes. Relaxed selection was detected at 9, 5, 5, and 2 codon positions in A-SAD, G-SAD, Class III, and bona fide CAD, respectively (Fig. 4). Of the two codon positions under relaxed selection in bona fide CAD, one only occurred in a subclade of angiosperm (Fig. 4d), and the other in a subclade of core eudicots (Fig. 4e).


A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
?
more than seven introns
uncertain gene structure
Fabaceae
Core eudicots
Rosids I
Monocots
Class II-A-SAD
A
Class II-S-SAD
Class I-bona fide CAD
Class III
Class II-G-SAD
Class II-S-SAD
Outgroups

◀ **Fig. 1** Maximum-likelihood tree of the *CAD*/*SAD* genes constructed based on the amino acid (aa) sequences with two green algae as outgroups. *Numbers* above branches refer to bootstrap values higher than 50%. The genes without available EST sequences are indicated by *triangles*. The *big* and *small stars* denote the inferred early and recent duplication events, respectively. The genes with different structures are indicated in different *colors*, with their detailed structures shown in the diagrams at the *upper-left corner* (exons in *boxes* and introns in *lines*). The genes shaded in *different colors* have been functionally studied regarding their roles in lignin biosynthesis: *gray* undoubtedly involved, *yellow* not involved, *light green* debated. *A-SAD* putative *SAD* of angiosperm, *G-SAD* putative *SAD* of gymnosperm, *S-SAD* putative *SAD* of *Selaginella*, *Aho Abies holophylla*, *Ath Arabidopsis thaliana*, *Bdi Brachypodium distachyon*, *Car Cathaya argyrophylla*, *Cde Cedrus deodara*, *Cja Cryptomeria japonica*, *Cpa Carica papaya*, *Cpi Chamaecyparis pisifera*, *Cre Chlamydomonas reinhardtii*, *Cri Ceratopteris richardii*, *Cru Cycas rumphii*, *Csa Cucumis sativus*, *Gbi Ginkgo biloba*, *Gma Glycine max*, *Mes Manihot esculenta*, *Mgl Metasequoia glyptostroboides*, *Mgu Mimulus guttatus*, *Mtr Medicago truncatula*, *Nta Nicotiana tabacum*, *Osa Oryza sativa*, *Pab Picea abies*, *Par Pinus armandii*, *Pba Pinus banksiana*, *Pda Phoenix dactylifera*, *Por Platycladus orientalis*, *Ppa Physcomitrella patens*, *Psi Picea sitchensis*, *Pra Pinus radiata*, *Pta Pinus taeda*, *Ptr Populus trichocarpa*, *Rco Ricinus communis*, *Sbi Sorghum bicolor*, *Smo Selaginella moellendorffii*, *Tdo Thujopsis dolabrata*, *Tdu Tsuga dumosa*, *Vca Volvox carteri*, *Vvi Vitis vinifera*. Gene names and identifiers are shown in Supplementary Table S2

The branch-site test showed that most sites of the *CAD*/*SAD* genes experienced purifying selection ($\omega < 1$), but some branches could have experienced neutral ($\omega = 1$) or positive selection ($\omega > 1$) at a few sites (Supplementary Fig. S4; Table S4). In Class I (bona fide CAD) and Class III, positive selection was detected from most of the trunk branches leading to the major lineages of vascular plants. In A-SAD of Class II, positive selection was detected on branches where ancestral gene duplications occurred. Only two inner branches of the G-SAD (Class II) were suggested to have been positively selected (significance at 5%) (Supplementary Fig. S4; Table S4). However, the results of the branch model tests indicated that all branches, except one of bona fide CAD, one of G-SAD and three of A-SAD, were under purifying selection after removing branches with d_S below 0.005 that could lead to uncertain estimates (Palmé et al. 2009) (Supplementary Fig. S4). The site model tests also indicated that no sites were under significant positive selection except one of G-SAD (344 S) and one of bona fide CAD (97 N).

Discussion

Phylogenetic History of the *CAD*/*SAD* Gene Family in Plants

Plant *CAD*/*SAD*s have attracted the most interest due to their important roles in monolignol biosynthesis (e.g., Galliano et al. 1993a; Brill et al. 1999; Li et al. 2001; Kim et al. 2004, 2007; Tobias and Chow 2005; Hamberger et al. 2007; Barakat et al. 2009; Saballos et al. 2009), although, *CAD*/*SAD* homologs have been found in bacteria and fungi (e.g., Larroy et al. 2002; Valencia et al. 2004; Mee et al. 2005), as well as some protists (Peacock et al. 2007). Most previous studies focused on functions of the *CAD*/*SAD* genes, while evolution of the *CAD*/*SAD* gene family and its correlation with the origin and the evolution of lignin remains unresolved, although, there have been some phylogenetic analyses of this gene family from different species (e.g., Raes et al. 2003; Tobias and Chow 2005; Hamberger et al. 2007; Barakat et al. 2009; Saballos et al. 2009; Ma 2010).

Table 2 The similarity and identity between the inferred protein sequences of the *CAD*/*SAD* genes and the automated models (AthCAD5 and PtSAD) in predicted three-dimensional structure

Protein	AthCAD5 (%)		PtSAD (%)		Automated model
	Identity	Similarity	Identity	Similarity	
PpaCAD1	47.1	67.8	54.6	70.1	PtSAD
PpaCAD3	43.5	59.9	46.3	62.0	PtSAD
SmoCAD1	62.0	77.0	54.0	54.0	AthCAD5
SmoCAD3	48.5	67.8	55.9	55.9	PtSAD
SmoCAD8	44.1	63.7	48.8	66.7	PtSAD
CreCAD1	47.0	67.0	55.5	71.6	PtSAD
PbaSAD1	51.8	63.9	63.3	73.5	PtSAD

In this study, the evolutionary history of the *CAD*/*SAD* gene family was reconstructed based on a sampling of 39 species representing most major lineages of green plants (chlorophyte, bryophyte, lycophyte, fern, gymnosperm, and angiosperm). All phylogenetic analyses, using the amino acid sequences or the first plus second codon positions, consistently indicate that land-plant *CAD*/*SAD*s form a monophyletic group, in which three classes could be recognized when gene structure and function are also considered. Class I, i.e., bona fide CAD, is conserved in vascular plants. Class II comprises three groups, i.e., A-SAD, G-SAD, and S-SAD. Class III is conserved in land plants (Fig. 1, Supplementary Fig. S1). The three classes designated in our study are different from those recognized in Barakat et al. (2009), although, some similarities can be found between the two classifications.

Our Class I corresponds to one subclade of Barakat et al.'s Class I, and includes *CAD* genes not only from seed plants but also from the lycophyte *Selaginella* and the fern *Adiantum capillus-veneris*. The functions of our Class I *CAD* genes would be very conserved given their highly conserved substrate-binding sites, as proposed in Youn et al. (2006), and evolution under strong purifying selection (Figs. 3, 4, Supplementary Fig. S3), as well as the results of previous functional studies on some members of this class

(Fig. 1). The other subclade of Barakat et al.'s Class I, represented by sequences from only gymnosperms, was recognized by us as G-SAD, a lineage of Class II. Members of this lineage are more similar to *PtSAD* (Class II) than to *AthCAD5* (Class I) in sequence similarity, three-dimensional structure of protein, and substrate-binding sites (Table 2; Supplementary Fig. S3). Our Class II contains members from lycophyte (S-SAD), fern (AcaCAD1), gymnosperm (G-SAD), and angiosperm (A-SAD), but Barakat et al.'s Class II is angiosperm-specific. Frequent gene duplications and losses have occurred in our Class II, especially in A-SAD (Fig. 1). Genes of this class might have been under much more relaxed selection than the other two classes (Fig. 3), showing diverse expression patterns and enzyme activities (e.g., Kim et al. 2007). Most branches with a number of sites under relaxed selection (PP $>$ 0.9) belong to clades A-SAD and G-SAD (data not shown), and most of the twelve putative substrate-binding sites are relatively variable and under more relaxed selection in Class II or its members (Fig. 4, Supplementary Fig. S3). Therefore, we grouped all the putative *SAD*s into Class II due to their differences from Class I to Class III genes in evolution and function, although, they are not monophyletic. In addition, different from Barakat et al.'s Class III that is also angiosperm-specific; our Class III includes members from liverwort, moss, lycophyte, fern, gymnosperm, and angiosperm. The Class III genes are highly conserved in sequence and gene structure, and differ from the other two classes in sequence similarity (Table 2), substrate-binding sites (Supplementary Fig. S3), and enzyme activity (Kim et al. 2004). All the above evidence, together with the number of *CAD*/*SAD* genes found in each species (Supplementary Table S2), indicates that the radiation of the *CAD*/*SAD* gene family occurred at least before the divergence of vascular plants rather than in the early ancestry of angiosperms as suggested by Barakat et al. (2009). The EST database of the fern *Ceratopteris richardii* is very small, and thus only one member of this gene family was found in the species (Fig. 1; Supplementary Table S2).

Evolutionary Mode and Functional Divergence of the *CAD*/*SAD* Gene Family

The birth-and-death evolution might be the most popular evolutionary mode of multigene family (Nei and Rooney 2005). However, there is a great variation in the rate of gene duplication and loss among different gene families or even different clusters of the same gene family (e.g., Nam et al. 2004; Xu et al. 2009b). Gene families encoding ancient "conserved biological functions," such as DNA metabolism, nuclease activity, and RNA binding, often have conserved copy numbers and gene structure whereas the copy number and structure might be dramatically variable in those gene families encoding transcription factors, protein kinases, and ribosomal proteins (Nei 2007; Freeling 2009). Moreover, a strong correlation between evolutionary pattern and gene function might exist among different gene families or even within the same gene family. That is, genes with conserved functions tend to experience strong purifying selection and little or no change in copy number and structure, while those with specific functions usually experience rapid duplication and relaxed selection (e.g., Nam et al. 2004; Nei 2007; Yang 2007; Freeling 2009; Xu et al. 2009b). In this study, we found a great variation in the rate of gene duplication and loss among the three classes of *CAD*/*SAD* genes. Class I (bona fide *CAD*) comprises members from lycophyte to angiosperm (Fig. 1), and the substrate-binding pockets of this class members (e.g., AthCAD5) are significantly different and smaller when compared to that of Class II (e.g., PtSAD), with highly conserved residues (Bomati and Noel 2005;

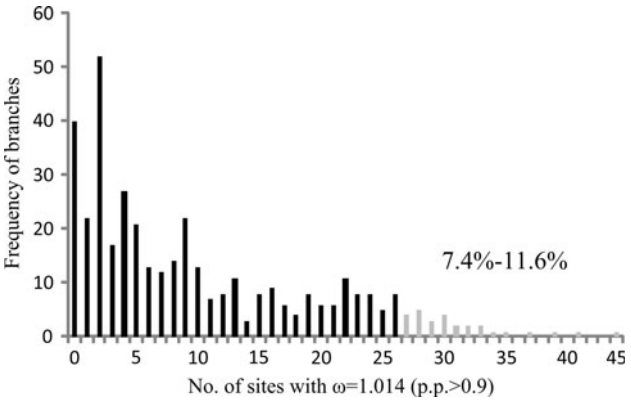
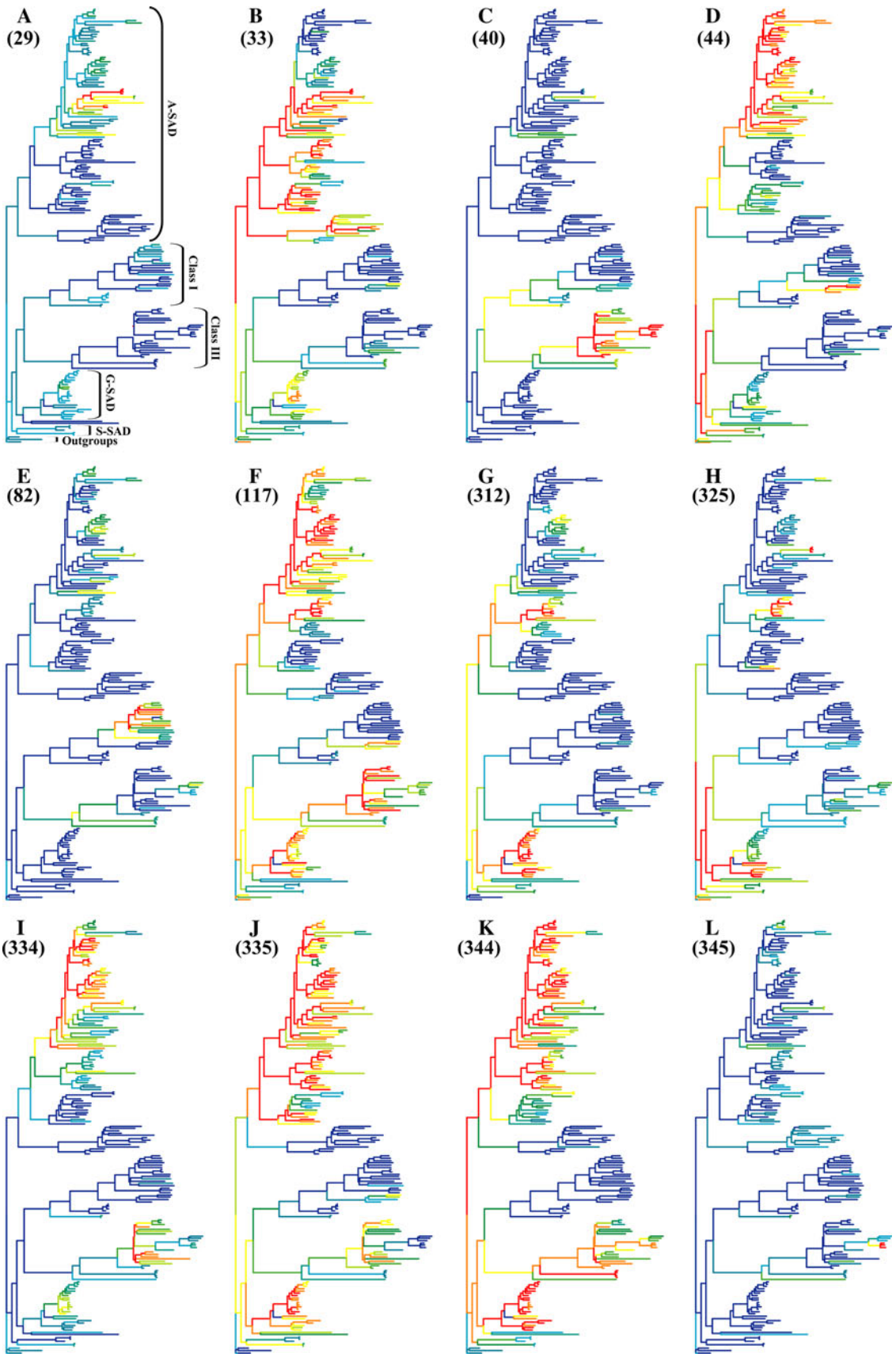


Fig. 2 Distribution of branches with different numbers of sites under relaxed selection across the *CAD*/*SAD* gene tree. *Gray bars* refer to branches with the number of sites under relaxed selection in the 95th percentile. The *number* on the panel represents the percentage of sites under relaxed selection in the whole alignment

Table 3 Results of the LRT test of the models in the Fitmodel program for the land-plant *CAD*/*SAD* gene sequence data

	M0	M3	M3 + S1	M3 + S2
ln *L*	−89310.68	−87076.92	−85956.6	−85820.37
$\omega 1\ \omega 2\ \omega 3$	0.16	0.03 0.16 0.41	0.00 0.21 0.85	0.01 0.12 1.01
*p*1 *p*2 *p*3	1.00	0.35 0.42 0.24	0.54 0.32 0.14	0.47 0.39 0.14
*R*12 *R*13 *R*23			1.71 1.71 1.71	0.89 0.21 5.86

Fig. 3 Shifts of selection across all sites on branches of the *CAD*/*SAD* gene tree. Branches under relaxed selection (*gray bars* in Fig. 2) are shown in *bold lines*


A
(29)
A-SAD
Class I
Class III
G-SAD
S-SAD
Outgroups
B
(33)
C
(40)
D
(44)
E
(82)
F
(117)
G
(312)
H
(325)
I
(334)
J
(335)
K
(344)
L
(345)

◀**Fig. 4** Shifts of selection at the twelve residues (**a**–**l**) which constitute the proposed substrate-binding pocket of bona fide CAD (Youn et al. 2006) on the *CAD*/*SAD* gene tree. Branches under relaxed selection are shown in *red* (PP greater than 90% for placing a branch in the $\omega 3$ selection class). Branches under strong purifying selection are shown in *blue* (PP lower than 20%). Branches shown in the *rest colors* are under moderate purifying selection (PP from 21 to 89%). Numbers in (**a**–**l**) correspond to the positions of the twelve residues in the whole alignment

Youn et al. 2006). Except *Manihot esculenta*, no more than two Class I members were detected in the same species, which might be due to the conserved function that they perform in the lignin formation. Class III genes are also highly conserved in sequence and structure, with all members having six exons except *PpaCAD3* (5 exons), *PsiCAD*, and *PtaCAD2* (uncertain) (Fig. 1), and thus this class might also have a conserved function. Interestingly, Class II includes three subgroups without clustering together, and shows a remarkable variation in gene copy number (5–16) and the number of exon (2–6) (Table 1; Fig. 1). All the four modes of gene duplication (tetraploid, segmental, transpositional, and tandem) (Freeling 2009) seem to have occurred in this class (e.g., Tobias and Chow 2005; Barakat et al. 2009). Moreover, rapid accumulation of tandemly arrayed gene duplicates usually induced by an environmental stressor (Demuth and Hahn 2009) was detected in this class, suggesting that these genes could be defense-related or play roles in plant adaptation.

The maximum likelihood analysis is widely used to identify genes that have been subject to diversifying selection, especially using Fitmodel (Guindon et al. 2004) and PAML (Yang 1997, 2007). In this study, both the Fitmodel program and the codeml program in PAML were performed to detect selective constraints on the evolution of the *CAD*/*SAD* genes. Both the results showed that most branches and sites were under purifying selection. However, some discrepancies also exist between results of the two analyses. First, unlike PAML, the Fitmodel analysis did not detect any significant positive selection (Table 3, Supplementary Table S4), although, it is possible that some sites placed in the $\omega 3$ class have actually experienced positive selection (Shan et al. 2009). Secondly, greatly relaxed selection has occurred on many branches of the A-SAD clade according to the Fitmodel analysis, but it was not detected by the PAML analysis (Fig. 3; Supplementary Fig. S4, Table S4). The branch or branch-site model in PAML assumes that the variation pattern of selective constraint is different between the foreground and the background branches (Yang 1998; Zhang et al. 2005), so it is very difficult to specify the foreground branch when the functions and evolutionary histories of the genes under study are poorly understood (Guindon et al. 2004; Anisimova and Yang, 2007 and references cited therein), such as the A-SAD and G-SAD genes in this study. In contrast, the Fitmodel program does not require a priori knowledge of lineages evolving under positive selection (Guindon et al. 2004). In addition, the total branch length, defined as number of nucleotide substitutions per codon, is higher than 55 in the A-SAD clade, which also suggests that sequences of this clade are too divergent to be analyzed with PAML (Anisimova et al. 2001, 2002; Anisimova and Yang 2007) (Supplementary Table S4). Finally, in Class I (bona fide CAD) and Class III, the branch-site test indicates that most of the trunk branches leading to the major lineages of vascular plants have been under positive selection, while the Fitmodel analysis suggests purifying selection in these two classes (Fig. 3, Supplementary Fig. S4). This difference might be due to that the very variable mode and strength of selection on the background branches has lead to unreliable inferences for the foreground branches in PAML or that the Fitmodel has less statistical power than the branch-site analysis (Shan et al. 2009). However, the fact that the distribution of sites under positive or relaxed selection is more regular in the Fitmodel than in the branch-site analysis (Table S4) indicates that the Fitmodel program is more suitable for our dataset. Moreover, the reliability of the branch-site model is still hotly debated (e.g., Zhang et al. 2005; Nozawa et al. 2009; Yang et al. 2009). Therefore, the following discussion about the evolution and functional diversification of the *CAD*/*SAD* gene family is based on the results of the Fitmodel analysis.

According to the Fitmodel analysis, the lineages with a large number of sites under relaxed selection include the ancestral branch of land plants as well as many deep internal branches in the A-SAD clade (Class II), implying that the function of Class II genes might not be very conserved, which is consistent with previous experimental studies (e.g., Kim et al. 2004, 2007; Stephens 2005; Goldie 2006; Barakat et al. 2009). In contrast, stronger purifying selection has acted on Classes I and III, suggesting conserved functions of the two classes (Fig. 3). Additionally, the selective constraints on the twelve residues, which constitute the proposed substrate-binding pocket of bona fide CAD (Youn et al. 2006), are very different among the three classes of *CAD*/*SAD* genes. Only two residues were detected to be under relaxed selection in Class I (Fig. 4d, e), of which some members of gymnosperms and angiosperms have been confirmed to be involved in the monolignol biosynthesis (Fig. 1). Unlike in Class I, much more residues in Classes II and III have been under relaxed selection (Fig. 4), indicating that the two classes may have obtained other functions. As discussed above, the three classes of *CAD*/*SAD* genes differ in evolutionary mode, conserved in Classes I and III but variable in Class II, which might associate with the functional divergence among them.

The Occurrence of Bona Fide *CAD* is Very Likely Correlated with the Origin of Lignin

Lignin plays a vital role in plant adaptation to terrestrial environments and represents a major obstacle in paper pulping, forage digestibility, and processing of plant biomass to biofuels (Vanholme et al. 2008). Of great importance is to investigate the molecular basis for the lignin biosynthesis. It has been clear that monolignol biosynthesis is one of the two branches derived from the general phenylpropanoid pathway (Ferrer et al. 2008). The CADs catalyze the last step in monolignol biosynthesis (Gross et al. 1973; Mansell et al. 1974). Most researchers agreed that Class I (bona fide CAD) is essential for the biosynthesis of monolignols, since bona fide *CAD* mutants in several plants have distinct phenotypes and different lignin contents and compositions when compared to the wild-type, such as *cad-n1* of *Pinus taeda* (MacKay et al. 1997; Ralph et al. 1997; Gill et al. 2003), *cad-c cad-d* double mutant of *Arabidopsis thaliana* (Sibout et al. 2005), *gh2* of *Oryza sativa* (Zhang et al. 2006), and *midrib6* of *Sorghum bicolor* (Saballos et al. 2009).

This study shows that the bona fide *CAD* exists in all main lineages of vascular plants, including lycophyte, fern, gymnosperm, and angiosperm, and represents a monophyletic clade (Fig. 1, Supplementary Fig. S1). The bona fide *CADs* we recognized are also supported by the prediction of the three-dimensional structure of protein, the identity of amino acid residues in the proposed substrate-binding sites and the strong purifying selection on these genes (Table 2; Fig. 3, Supplementary Fig. S3). Therefore, it is very likely that the lycophyte has the earliest bona fide *CAD* involved in the monolignol biosynthesis. This inference is consistent with the fact that *Selaginella* has real lignin (Weng et al. 2008). In bryophytes, only some lignin-related compounds were detected (Ligrone et al. 2008), indicating that the monolignol biosynthesis pathway might have not been completely established in the earliest land plant due to the lack of bona fide *CAD*.

Previous studies showed that some upstream genes of the monolignol biosynthesis pathway are highly conserved from bryophytes to angiosperms, such as the genes encoding the phenylalanine ammonia lyase (PAL) (Emiliani et al. 2009), cinnamate 4-hydroxylase (C4H), and *p*-coumaroyl shikimate/quinate 3′-hydroxylase (C3′H) (Weng et al. 2008). Also, three of the four gene members encoding 4-coumarate: CoA ligase (4CL) from the moss *Physcomitrella patens* were found to have substrate activity similar to *Arabidopsis* 4CLs, although, the evolutionary history of the *4CL* gene is still ambiguous (Silber et al. 2008; Souza et al. 2008). Xu et al. (2009a) suggested that the monolignol biosynthesis pathway had been completely established in early land plants based on the fact that all the nine gene families involved in the pathway occur in moss and have many more members in moss than in green alga, and even considered moss as the turning point—the host of "original lignin". However, it is very common that a specific character appeared much later than the origin of the gene families responsible for it. For example, the main types of the MADS-box genes that control diverse developmental processes in flowering plants, in particular the development of flower, originated before the divergence of seed plants, and its homologs have been reported in green alga, moss, and so on (Becker and Theissen 2003; Riese et al. 2005; Riaño-Pachón et al. 2008), but the flower first appeared in angiosperms. In addition, the sudden expansion of the monolignol biosynthesis gene family members in moss might not directly cause the origin of lignin. Some of the upstream genes with number expansion, such as *PAL*, *C4H*, and *4CL*, also participate in the flavonoid biosynthesis, the other branch of the general phenylpropanoid pathway (Ferrer et al. 2008). Moreover, some gene copies reported in Xu et al. (2009a) might be family-like members. For instance, 11 and 26 *4CLs*, respectively, were detected in *Physcomitrella* and *Selaginella* in their study, but only four and eight, respectively, were proved to be the real *4CL*s in these two species by Silber et al. (2008) and our unpublished data. Nevertheless, it cannot be completely ruled out that the bryophyte had lost the Class I *CAD* genes or some of its extant members of the *CAD*/*SAD* gene family might play some roles in the monolignol biosynthesis. Further studies are needed to attribute specific functions to these members of bryophytes.

One may argue that lignin has been found in the red alga *Calliarthron cheilosporioides* (Martone et al. 2009). However, although, the genome sequence of this species is unavailable right now, only one *CAD*/*SAD* homolog (*CmeCAD*) was found in the whole genome sequence of another red alga *Cyanidioschyzon merolae*, and it is phylogenetically distantly related to the *CAD*/*SAD*s of green plants (Supplementary Fig. S1). In addition, the blast search did not find significant homologs of the other monolignol biosynthesis gene families from the red alga *C. merolae* (unpublished data). The above information strongly supports the hypothesis of convergent evolution of the lignin biosynthesis between red alga and vascular plants (Martone et al. 2009). All the above evidence strongly suggests that the occurrence of bona fide *CAD*, the primary gene for the monolignol biosynthesis in vascular plants, is very likely directly correlated with the origin of lignin.

Up to now, the functions of Classes II and III still remain controversial. Though several studies indicated that some members of Class II [e.g., *PtSAD*, *AthCAD7*, *8*, *OsaCAD7* (*FC1*)] might have played important roles in lignin biosynthesis (Li et al. 2001, 2009; Kim et al. 2007; Barakat et al. 2009), it has also been reported that the Class

II genes could be involved in plant stress resistance such as *AthCAD7* (*AtELI3-1*), *AthCAD8* (*AtELI3-2*), *NtaSAD* (e.g., Somssich et al. 1996; Kim et al. 2004; Stephens 2005; Goldie 2006). Especially, the lignin content and composition did not change when expression of the tobacco *SAD* (*NtaSAD*) was suppressed, suggesting that SAD might not play a major role in lignin biosynthesis (Stephens 2005; Goldie 2006). In addition, some investigations showed that different members of Class II have different expression patterns and enzyme substrate specificities (Kim et al. 2004, 2007; Barakat et al. 2009), which may imply that the function of this class is not as conserved as that of bona fide *CAD*. Nevertheless, the many lineage-specific duplications and extinctions in Class II (Fig. 1) as well as the higher substrate versatility of the proteins caused by the larger substrate-binding pocket, such as in PtSAD (Bomati and Noel 2005; Youn et al. 2006), seem to suggest that this class of genes have had relatively specialized functions, such as stress resistance (Somssich et al. 1996; Kim et al. 2004; Stephens 2005; Goldie 2006; Xu et al. 2009b). The Class III genes are strongly supported as a monophyletic clade, and comprise members from the five main clades of land plants (bryophyte, lycophyte, fern, gymnosperm, and angiosperm) (Fig. 1, Supplementary Fig. S1). Functional studies indicated that some members of this class lack detectable CAD catalytic activities in vitro but express widely (Kim et al. 2004, 2007; Barakat et al. 2009). Additionally, the twelve residues in the proposed substrate-binding sites (Youn et al. 2006) of these class members are distinct from those of both Class I and II members (Supplementary Fig. S3). Hence, Class III might not be responsible for the lignin biosynthesis. Nevertheless, it is very likely that Class III has important functions during plant development and is necessary for the adaptation of land plants, given its ancient origin and high conservation, like *PAL*, *C4H*, and *C3′H*, and wide expression. Further studies are needed to attribute specific functions to this class.

Acknowledgments The authors thank the handling editor and the two anonymous reviewers for their insightful comments and suggestions on the manuscript, Dr. Qing-Yin Zeng for his helpful advice on the three-dimensional structure modeling of protein and Dr. Hong-Yan Shan for her kind help in the Fitmodel analysis. We also thank Ms. Wan-Qing Jin for her assistance in the DNA sequencing. This study was supported by the National Natural Science Foundation of China (Grant Nos. 30730010, 30990240, 30425028) and the 100-Talent Project of the Chinese Academy of Sciences.

References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105

Anisimova M, Yang ZH (2007) Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol 24:1219–1228

Anisimova M, Bielawski JP, Yang ZH (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18:1585–1592

Anisimova M, Bielawski JP, Yang ZH (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950–958

Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modeling. Bioinformatics 22:195–201

Barakat A, Bagniewska-Zadworna A, Choi A, Plakkat U, DiLoreto DS, Yellanki P, Carlson JE (2009) The cinnamyl alcohol dehydrogenase gene family in *Populus*: phylogeny, organization, and expression. BMC Plant Biol 9:26

Barceló A, Ros Gómez Ros LV, Carrasco AE (2007) Looking for syringyl peroxidases. Trends Plant Sci 12:486–491

Bateman RM, Crane PR, Dimichele WA, Kenrick PR, Rowe NP, Speck T, Stein WE (1998) Early evolution of land plants: phylogeny, physiology, and ecology of the primary terrestrial radiation. Annu Rev Ecol Syst 29:263–292

Becker A, Theissen G (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. Mol Phylogenet Evol 29:464–489

Boerjan W, Ralph J, Baucher M (2003) Lignin biosynthesis. Annu Rev Ecol Syst 54:519–546

Bomati EK, Noel JP (2005) Structural and kinetic basis for substrate selectivity in *Populus tremuloides* sinapyl alcohol dehydrogenase. Plant Cell 17:1598–1611

Bowman JL, Floyd SK, Sakakibara K (2007) Green genes-comparative genomics of the green branch of life. Cell 129:229–234

Brill EM, Abrahams S, Hayes CM, Jenkins CLD, Watson JM (1999) Molecular characterisation and expression of a wound-inducible cDNA encoding a novel cinnamyl-alcohol dehydrogenase enzyme in lucerne (*Medicago sativa* L.). Plant Mol Biol 41:279–291

Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang XH, Mudge J, Vasdewani J, Schiex T, Spannagl M, Monaghan E, Nicholson C, Humphray SJ, Schoof H, Mayer KFX, Rogers J, Quétier F, Oldroyd Gm E, Debellé F, Cookm DR, Retzel EF, Roe BA, Town CD, Tabata S, Van de Peer Y, Young ND (2006) Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. Proc Natl Acad Sci USA 103:14959–14964

Demuth JP, Hahn MW (2009) The life and death of gene families. Bioessays 31:29–39

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15

Emiliani G, Fondi M, Fani R, Gribaldo S (2009) A horizontal gene transfer at the origin of phenylpropanoid metabolism: a key adaptation of plants to land. Biol Direct 4:7

Ferrer JL, Austin MB, Stewart C Jr, Noel JP (2008) Structure and function of enzymes involved in the biosynthesis of phenylpropanoids. Plant Physiol Biochem 46:356–370

Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome segmental, or by transposition. Annu Rev Plant Biol 60:433–453

Galliano H, Cabane M, Eckerskorn C, Lottspeich F, Sandermann H Jr, Ernst D (1993a) Molecular-cloning. sequence-analysis and elicitor-/ozone-induced accumulation of cinnamyl alcohol dehydrogenase from Norway spruce (*Picea abies* L.). Plant Mol Biol 23:145–156

Galliano H, Heller W, Sandermann H Jr (1993b) Ozone induction and purification of spruce cinnamyl alcohol dehydrogenase. Phytochemistry 32:557–563

Gill GP, Brown GR, Neale DB (2003) A sequence mutation in the cinnamyl alcohol dehydrogenase gene associated with altered lignification in loblolly pine. Plant Biotechnol J 1:253–258

Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M, Glazebrook J, Sessions A, Oelle P, Varma H, Hadley D, Hutchinson D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong JP, Miguel T, Paszkowski U, Zhang SP, Colbert M, Sun WL, Chen LL, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu YS, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science 296:92–100

Goffner D, Joffroy I, Grima-Pettenati J, Halpin C, Knight ME, Schuch W, Boudet AM (1992) Purification and characterization of isoforms of cinnamyl alcohol dehydrogenase from *Eucalyptus xylem*. Planta 188:48–53

Goldie AS (2006) Elucidating the role of sinapyl alcohol dehydrogenase in tobacco. University of Dundee, Scotland

Gross GG, Stöckigt J, Mansell RL, Zenk MH (1973) Three novel enzymes involved in the reduction of ferulic acid to coniferyl alcohol in higher plants: ferulate: CoA ligase, feruloyl-CoA reductase and coniferyl alcohol oxidoreductase. FEBS Lett 31:283–286

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP (2004) Modeling the site-specific variation of selection patterns along lineages. Proc Natl Acad Sci USA 101:12957–12962

Hall T (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acid Symp Ser 41:95–98

Hamberger B, Ellis M, Friedmann M, Souza CDA, Barbazuk B, Douglas CJ (2007) Genome-wide analyses of phenylpropanoid-related genes in *Populus trichocarpa*, *Arabidopsis thaliana*, and *Oryza sativa*: the Populus lignin toolbox and conservation and diversification of angiosperm gene families. Can J Bot 85:1182–1201

Hawkins SW, Boudet AM (1994) Purification and characterization of cinnamyl alcohol dehydrogenase isoforms from the periderm of *Eucalyptus Gunnii* Hook. Plant Physiol 104:75–84

Huang SW, Li RQ, Zhang ZH, Li L, Gu XF, Fan W, Lucas WJ, Wang XW, Xie BY, Ni PX, Ren YY, Zhu HM, Li J, Lin K, Jin WW, Fei ZJ, Li GC, Staub J, Kilian A, van der Vossen EAG, Wu Y, Guo J, He J, Jia ZQ, Ren Y, Tian G, Lu Y, Ruan J, Qian WB, Wang MW, Huang QF, Li B, Xuan ZL, Cao JJ, Asan Wu ZG, Zhang JB, Cai QL, Bai YQ, Zhao BW, Han YH, Li Y, Li XF, Wang SH, Shi QX, Liu SQ, Cho WK, Kim JY, Xu Y, Heller-Uszynska K, Miao H, Cheng ZC, Zhang SP, Wu J, Yang YH, Kang HX, Li M, Liang HQ, Ren XL, Shi ZB, Wen M, Jian M, Yang HL, Zhang GJ, Yang ZT, Chen R, Liu SF, Li JW, Ma LJ, Liu H, Zhou Y, Zhao J, Fang XD, Li GQ, Fang L, Li YR, Liu DY, Zheng HK, Zhang Y, Qin N, Li Z, Yang GH, Yang S, Bolund L, Kristiansen K, Zheng HC, Li SC, Zhang XQ, Yang HM, Wang J, Sun RF, Zhang BX, Jiang SZ, Wang J, Du YC, Li SG (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41:1275–1281

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quetier F, Wincker P (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Kenrick P, Crane PR (1997) The origin and early evolution of plants on land. Nature 389:33–39

Kim SJ, Kim MR, Bedgar DL, Moinuddin SGA, Cardenas CL, Davin LB, Kang CL, Lewis NG (2004) Functional reclassification of the putative cinnamyl alcohol dehydrogenase multigene family in *Arabidopsis*. Proc Natl Acad Sci USA 101:1455–1460

Kim SJ, Kim KW, Cho MH, Franceschi VR, Davin LB, Lewis NG (2007) Expression of cinnamyl alcohol dehydrogenases and their putative homologues during *Arabidopsis thaliana* growth and development: lessons for database annotations? Phytochemistry 68:1957–1974

Kutsuki H, Shimada M, Higuchi T (1982) Regulatory role of cinnamyl alcohol dehydrogenase in the formation of guaiacyl and syringyl lignins. Phytochemistry 21:19–23

Larroy C, Parés X, Bīosca JA (2002) Characterization of a *Saccharomyces cerevisiae* NADP(H)-dependent alcohol dehydrogenase (ADHVII), a member of the cinnamyl alcohol dehydrogenase family. Eur J Biochem 269:5738–5745

Li LG, Cheng XF, Leshkevich J, Umezawa T, Harding SA, Chiang VL (2001) The last step of syringyl monolignol biosynthesis in angiosperms is regulated by a novel gene encoding sinapyl alcohol dehydrogenase. Plant Cell 13:1567–1585

Li XJ, Yang Y, Yao JL, Chen GX, Li XH, Zhang QF, Wu CY (2009) *FLEXIBLE CULM 1* encoding a cinnamyl-alcohol dehydrogenase controls culm mechanical strength in rice. Plant Mol Biol 69:685–697

Ligrone R, Carafa A, Duckett JG, Renzaglia KS, Ruel K (2008) Immunocytochemical detection of lignin-related epitopes walls in bryophytes and the charalean alga *Nitella*. Plant Syst Evol 270:257–272

Ma QH (2010) Functional analysis of a cinnamyl alcohol dehydrogenase involved in lignin biosynthesis in wheat. J Exp Bot. doi: 10.1093/jxb/erq107

MacKay JJ, ÓMalley DM, Presnell T, Booker FL, Campbell MM, Whetten RW, Sederoff RR (1997) Inheritance, gene expression, and lignin characterization in a mutant pine deficient in cinnamyl alcohol dehydrogenase. Proc Natl Acad Sci USA 94:8255–8260

Mansell RLG, Gross GG, Stöckigt J, Franke H, Zenk MH (1974) Purification and properties of cinnamyl alcohol dehydrogenase from higher plants involved in lignin biosynthesis. Phytochemistry 13:2427–2435

Martone PT, Estevez JM, Lu F, Ruel K, Denny MW, Somerville C, Ralph J (2009) Discovery of lignin in seaweed reveals convergent evolution of cell-wall architecture. Curr Biol 19:169–175

Mee B, Kelleher D, Frias J, Malone R, Tipton KF, Henehan GTM, Windle HJ (2005) Characterization of cinnamyl alcohol dehydrogenase of *Helicobacter pylori:* an aldehyde dismutating enzyme. FEBS J 272:1255–1264

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Marechal-Drouard L, Marshall WF, Qu LH, Nelson DR, Sanderfoot AA, Spalding MH, Kapitonov VV, Ren Q, Ferris P, Lindquist E, Shapiro H, Lucas SM, Grimwood J, Schmutz J, Cardol P, Cerutti H, Chanfreau G, Chen CL, Cognat V, Croft MT, Dent R, Dutcher S, Fernandez E, Fukuzawa H, Gonzalez-Ballester D, Gonzalez-Halphen D, Hallmann A, Hanikenne M, Hippler M, Inwood W, Jabbari K, Kalanon M, Kuras R, Lefebvre PA,

Lemaire SD, Lobanov AV, Lohr M, Manuell A, Meier I, Mets L, Mittag M, Mittelmeier T, Moroney JV, Moseley J, Napoli C, Nedelcu AM, Niyogi K, Novoselov SV, Paulsen IT, Pazour G, Purton S, Ral JP, Riano-Pachon DM, Riekhof W, Rymarquis L, Schroda M, Stern D, Umen J, Willows R, Wilson N, Zimmer SL, Allmer J, Balk J, Bisova K, Chen CJ, Elias M, Gendler K, Hauser C, Lamb MR, Ledford H, Long JC, Minagawa J, Page MD, Pan J, Pootakham W, Roje S, Rose A, Stahlberg E, Terauchi AM, Yang P, Ball S, Bowler C, Dieckmann CL, Gladyshev VN, Green P, Jorgensen R, Mayfield S, Mueller-Roeber B, Rajamani S, Sayre RT, Brokstein P, Dubchak I, Goodstein D, Hornick L, Huang YW, Jhaveri J, Luo Y, Martinez D, Ngau WC, Otillar B, Poliakov A, Porter A, Szajkowski L, Werner G, Zhou K, Grigoriev IV, Rokhsar DS, Grossman AR (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318:245–250

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KLT, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang M-L, Zhu YJ, Schatz M, Nagarajan N, Acob RA, Guan P, Blas A, Wai CM, Ackerman CM, Ren Y, Liu C, Wang J, Wang J, Na J-K, Shakirov EV, Haas B, Thimmapuram J, Nelson D, Wang X, Bowers JE, Gschwend AR, Delcher AL, Singh R, Suzuki JY, Tripathi S, Neupane K, Wei H, Irikura B, Paidi M, Jiang N, Zhang W, Presting G, Windsor A, Navajas-Perez R, Torres MJ, Feltus FA, Porter B, Li Y, Burroughs AM, Luo M-C, Liu L, Christopher DA, Mount SM, Moore PH, Sugimura T, Jiang J, Schuler MA, Friedman V, Mitchell-Olds T, Shippen DE, dePamphilis CW, Palmer JD, Freeling M, Paterson AH, Gonsalves D, Wang L, Alam M (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452:991–996

Nam J, Kim J, Lee S, An G, Ma H, Nei M (2004) Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc Natl Acad Sci USA 101:1910–1915

Nei M (2007) The new mutation theory of phenotypic evolution. Proc Natl Acad Sci USA 104:12235–12242

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121–152

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Nozawa M, Suzuki Y, Nei M (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci USA 106:6700–6705

Palmé AE, Pyhäjärvi T, Wachowiak W, Savolainen O (2009) Selection on Nuclear Genes in a *Pinus* Phylogeny. Mol Biol Evol 26:893–905

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Lyons E, Maher C, Martis M, Narechania A, Penning B, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, ur-Rahman M, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, Kerhornou A, Ivens A, Fraser A, Rajandream MA, Carver T, Norbertczak H, Chillingworth T, Hance Z, Jagels K, Moule S, Ormond D, Rutter S, Squares R, Whitehead S, Rabbinowitsch E, Arrowsmith C, White B, Thurston S, Bringaud F, Baldauf SL, Faulconbridge A, Jeffares D, Depledge DP, Oyola SO, Hilley JD, Brito LO, Tosi LRO, Barrell B, Cruz AK, Mottram JC, Smith DF, Berriman M (2007) Comparative genomic analysis of three Leishmania species that cause diverse human disease. Nat Genet 39:839–847

Peter G, Neale D (2004) Molecular basis for the evolution of xylem lignification. Curr Opin Plant Biol 7:737–742

Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25:1253–1256

Raes J, Rohde A, Christensen JH, Van de Peer Y, Boerjan W (2003) Genome-wide characterization of the lignification toolbox in *Arabidopsis*. Plant Physiol 133:1051–1071

Ralph J, MacKay JJ, Hatfield RD, OMalley DM, Whetten RW, Sederoff RR (1997) Abnormal lignin in a loblolly pine mutant. Science 277:235–239

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, Tanahashi T, Sakakibara K, Fujita T, Oishi K, Shin-I T, Kuroki Y, Toyoda A, Suzuki Y, Hashimoto S, Yamaguchi K, Sugano S, Kohara Y, Fujiyama A, Anterola A, Aoki S, Ashton N, Barbazuk WB, Barker E, Bennetzen JL, Blankenship R, Cho SH, Dutcher SK, Estelle M, Fawcett JA, Gundlach H, Hanada K, Heyl A, Hicks KA, Hughes J, Lohr M, Mayer K, Melkozernov A, Murata T, Nelson DR, Pils B, Prigge M, Reiss B, Renner T, Rombauts S, Rushton PJ, Sanderfoot A, Schween G, Shiu SH, Stueber K, Theodoulou FL, Tu H, Van de Peer Y, Verrier PJ, Waters E, Wood A, Yang LX, Cove D, Cuming AC, Hasebe M, Lucas S, Mishler BD, Reski R, Grigoriev IV, Quatrano RS, Boore JL (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science 319:64–69

Riaño-Pachón DM, Corréa LGG, Trejos-Espinosa R, Mueller-Roeber B (2008) Green transcription factors: a *Chlamydomonas* overview. Genetics 179:31–39

Riese M, Faigl W, Quodt V, Verelst A, Matthesm A, Saedlerm H, Münsterm T (2005) Isolation and characterization of new MIKC*-type MADS-box genes from the moss *Physcomitrella patens*. Plant Biol 7:307–314

Rogers SO, Bendich AJ (1988) Extraction of DNA from plant tissues. Plant Mol Biol (Manual) A6:1–10

Roy SW, Gilbert W (2005) Complex early genes. Proc Natl Acad Sci USA 102:1986–1991

Saballos A, Ejeta G, Sanchez E, Kang C, Vermerris W (2009) A genomewide analysis of the cinnamyl alcohol dehydrogenase family in *Sorghum* [*Sorghum bicolor* (L.) Moench] identifies *SbCAD2* as the *Brown midrib6* gene. Genetics 181:783–795

Schmutz J, Cannon SB, Schlueter J, Ma JX, Mitros T, Nelson W, Hyten DL, Song QJ, Thelen JJ, Cheng JL, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu SQ, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du JC, Tian ZX, Zhu LC, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Shan H-Y, Zahn L, Guindon S, Wall PK, Kong H-Z, Ma H, dePamphilis CW, Leebens-Mack J (2009) Evolution of plant MADS Box transcription factors: evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications. Mol Biol Evol 26:2229–2244

Sibout R, Eudes A, Mouille G, Pollet B, Lapierre C, Jouanin L, Séguin A (2005) *CINNAMYL ALCOHOL DEHYDROGENASE-C* and *-D* are the primary genes involved in lignin biosynthesis in the floral stem of *Arabidopsis*. Plant Cell 17:2059–2076

Silber MV, Meimberg H, Ebel J (2008) Identification of a 4-coumarate:CoA ligase gene family in the moss, *Physcomitrella patens*. Phytochemistry 69:2449–2456

Somssich IE, Wernert P, Kiedrowski S, Hahlbrock K (1996) *Arabidopsis thaliana* defense-related protein ELI3 is an aromatic alcohol:$NADP^+$ oxidoreductase. Proc Natl Acad Sci USA 93:14199–14203

Souza CD, Barbazuk B, Ralph SG, Bohlmann J, Hamberger B, Douglas CJ (2008) Genome-wide analysis of a land plant-specific *acyl:coenzymeA synthetase* (*ACS*) gene family in *Arabidopsis*, poplar, rice and *Physcomitrella*. New Phytol 179:987–1003

Stephens J (2005) Investigating the role of sinapyl alcohol dehydrogenase in lignin biosynthesis. University of Dundee, Scotland

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl Acids Res 25:4876–4882

Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333:863–882

Tobias CM, Chow EK (2005) Structure of the cinnamyl-alcohol dehydrogenase gene family in rice and promoter activity of a member associated with lignification. Planta 220:678–688

Valencia E, Larroy C, Ochoa WF, Parés X, Fita I, Biosca JA (2004) *Apo* and *Holo* structures of an NADP(H)-dependent cinnamyl alcohol dehydrogenase from *Saccharomyces cerevisiae*. J Mol Biol 341:1049–1062

Vanholme R, Morreel K, Ralph J, Boerjan W (2008) Lignin engineering. Curr Opin Plant Biol 11:278–285

Vogel JP (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768

Wang WM, Tanurdzic M, Luo MZ, Sisneros N, Kim HR, Weng JK, Kudrna D, Mueller C, Arumuganathan K, Carlson J, Chapple C, de Pamphilis C, Mandoli D, Tomkins J, Wing RA, Banks JA (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

Waters ER (2003) Molecular adaptation and the origin of land plants. Mol Phylogenet Evol 29:456–463

Weng JK, Li X, Stout J, Chapple C (2008) Independent origins of syringyl lignin in vascular plants. Proc Natl Acad Sci USA 105:7887–7892

Whetten RW, MacKay JJ, Sederoff RR (1998) Recent advances in understanding lignin biosynthesis. Annu Rev Plant Phys 49:585–609

Xia X, Xie Z (2001) DAMBE: Software package for data analysis in molecular biology and evolution. J Hered 92:371–373

Xu GX, Ma H, Nei M, Kong HZ (2009a) Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. Proc Natl Acad Sci USA 106:835–840

Xu Z, Zhang D, Hu J, Zhou X, Ye X, Reichel KL, Stewart NR, Syrenne RD, Yang X, Gao P, Shi W, Doeppke C, Sykes RW, Burris JN, Bozell JJ, Cheng MZ, Hayes DG, Labbe N, Davis M, Stewart CN Jr, Yuan JS (2009b) Comparative genome analysis of lignin biosynthesis gene families across the plant kingdom. BMC Bioinformatics 10(Suppl 11):S3

Yang ZH (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang ZH (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yang ZH (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Yang ZH, Nielsen R, Goldman N (2009) In defense of statistical methods for detecting positive selection. Proc Natl Acad Sci USA 106:E95

Youn B, Camacho R, Moinuddin SGA, Lee C, Davin LB, Lewis NG, Kang CH (2006) Crystal structures and catalytic mechanism of the *Arabidopsis* cinnamyl alcohol dehydrogenases AtCAD5 and AtCAD4. Org Biomol Chem 4:1687–1697

Zhang JZ, Nielsen R, Yang ZH (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472–2479

Zhang KW, Qian Q, Huang ZJ, Wang YQ, Li M, Hong LL, Zeng DL, Gu MH, Chu CC, Cheng ZK (2006) *GOLD HULL AND INTERNODE2* encodes a primarily multifunctional cinnamyl-alcohol dehydrogenase in rice. Plant Physiol 140:972–983