JSE Journal of Systematics and Evolution

Journal of Systematics and Evolution 50 (3): 195–205 (2012) doi: 10.1111/j.1759-6831.2012.00187.x

Research Article

Evolution of the 4-coumarate:coenzyme A ligase (*4CL*) gene family: Conserved evolutionary pattern and two new gene classes in gymnosperms

[1,2]Hui GAO[§] [1]Dong-Mei GUO[§] [1,2]Wen-Juan LIU [1]Jin-Hua RAN* [1]Xiao-Quan WANG

[1](*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences,* Beijing 100093, China)
[2](*Graduate University of the Chinese Academy of Sciences,* Beijing 100049, China)

Abstract The 4-coumarate:coenzyme A ligase (4CL) is the branch point enzyme that channels the general phenylpropanoid metabolism into specific lignin and flavonoid biosynthesis branches. Genetic engineering experiments on the *4CL* gene have been carried out in many species, but the precise functions of different gene members are still unresolved. To investigate the evolutionary relationships and functional differentiation of the *4CL* gene family, we made a comprehensive evolutionary analysis of this gene family from 27 species representing the major lineages of land plants. The phylogenetic analysis indicates that both vascular and seed plant *4CL* genes form monophyletic groups, and that three and two *4CL* classes can be recognized in gymnosperms and angiosperms, respectively. The evolutionary rate and frequency of duplication of the *4CL* gene family are much more conserved than that of the *CAD/SAD* (cinnamyl/sinapyl alcohol dehydrogenase) gene family, which catalyzes the last step in monolignol biosynthesis. This may be due to different selective pressures on these genes whose products catalyze different steps in the biosynthesis pathway. In addition, we found two new major classes of *4CL* genes in gymnosperms.

Key words *4CL*, functional divergence, gene family evolution, lignin biosynthesis.

Lignin, occurring mainly in the secondarily thickened plant cell walls, is the second most abundant terrestrial biopolymer after cellulose. It is a complex aromatic heteropolymer composed mainly of three hydroxycinnamyl alcohol monomers named p-coumaryl (H), guaiacyl (G), and syringyl (S). Lignin plays a crucial role in plant adaptation to terrestrial environments by providing plants with mechanical support, facilitating water and solute transport, and playing roles in biotic and abiotic stress resistance. However, lignin also represents a major obstacle in paper pulping, forage digestibility, and processing of plant biomass to biofuels (Whetten et al., 1998; Boerjan et al., 2003; Simmons et al., 2010). To create plants that are more amenable to chemical degradation, many studies have focused on the genetic engineering of lignin amount and composition, especially of candidate genes involved in monolignol biosynthesis (Anterola & Lewis, 2002; Baucher et al., 2003; Vanholme et al., 2008; Simmons et al., 2010). The gene encoding the 4-coumarate:coenzyme A ligase (*4CL*), a central enzyme in monolignol biosynthesis (Knobloch & Hahlbrock, 1975, 1977; Peter & Neale, 2004), has attracted a lot of interest. For example, genetic engineering of *4CL* has been carried out in some important model and economic plants such as *Arabidopsis* (Lee et al., 1997; Yang et al., 2011), poplar (Hu et al., 1999; Li et al., 2003; Voelker et al., 2010), pine (Wagner et al., 2009), and tobacco (Kajita et al., 1997). However, some conclusions in these studies conflict with each other. For example, when *4CL* expression was downregulated, some studies showed that this has no negative effect on the growth of transgenic plants (Lee et al., 1997; Hu et al., 1999; Yang et al., 2011), whereas others indicated that lignin content/amount decreased following downregulation (Wagner et al., 2009; Voelker et al., 2010).

The 4CL enzyme catalyzes the activation of several hydroxycinnamic acids such as 4- hydroxycinnamic acid, caffeic acid, ferulic acid, 5-hydroxyferulic acid, and sinapic acid into their corresponding CoA esters, channeling the general phenylpropanoid metabolism into specific branch pathways, that is, lignin and flavonoid biosynthesis (Knobloch & Hahlbrock, 1975, 1977; Peter & Neale, 2004). The 4CL is encoded by a small gene family that belongs to the super gene family encoding adenylate-forming enzymes, and its homologs are mainly detected in land plants (Hu et al., 1998; Schneider et al., 2003; Wei & Wang, 2004; Silber et al., 2008; de Azevedo Souza et al., 2008). In *Arabidopsis* and poplar, different 4CL members often have distinct sub-

Received: 29 May 2011 Accepted: 22 February 2012
[§] These authors contributed equally to this work.
* Author for correspondence. E-mail: jinhua_ran@ibcas.ac.cn; Tel.: 86-10-62836114; Fax: 86-10-62590843.

strate specificities and expression patterns, suggesting diverse functions for potential isozymes (Hu et al., 1998; Ehlting et al., 1999; Harding et al., 2002; Hamberger & Hahlbrock, 2004; Costa et al., 2005). However, previous functional studies of the *4CL* gene mainly focused on enzymatic activity and gene expression, and gene-specific knockout experiments have not been carried out on this gene family in plants. Therefore, we still know little about the precise functions of different gene members, including how they affect lignin content and composition.

The reconstructed phylogeny of a gene family could shed light on its evolutionary history and functional differentiation. Although several phylogenetic analyses have been done on the *4CL* gene family, these have involved relatively limited taxon samplings. Some investigations suggested that the angiosperm *4CL* genes could be divided into two phylogenetically distinct classes (Class I and Class II), which may specialize in monolignol and flavonoid biosynthesis, respectively (Hu et al., 1998; Ehlting et al., 1999; Hamberger & Hahlbrock, 2004). For gymnosperms, the *4CL* gene was preliminarily investigated in the pine family (Zhang & Chiang, 1997; Wang et al., 2000; Wei & Wang, 2004). In other land plants, there is almost no information about the *4CL* gene evolution, although four members of the gene family from *Physcomitrella patens* (Hedw.) Bruch & Schimp ssp. *patens* were used in a phylogenetic analysis (Silber et al., 2008). Therefore, the evolutionary history of the *4CL* gene family is still poorly understood in land plants.

In order to investigate the evolutionary relationships and functional differentiation of the *4CL* gene family, we made a comprehensive phylogenetic analysis of this gene family from 27 species representing the major lineages of land plants, and here discuss its functional divergence. In addition, we report two new classes of *4CL* in gymnosperms.

1 Material and methods

1.1 Plant materials used in sequencing and *4CL* gene identification

We obtained the *4CL* genes from 27 species that represent most of the major lineages of land plants, including a liverwort, a moss, a lycophyte, 12 gymnosperms, and 12 angiosperms (Table S1). All surveyed species except the gymnosperms and liverwort *Marchantia polymorpha* L. have available whole genome sequences. We carried out tBLASTn searches using public genome databases, using all known *Arabidopsis thaliana* (L.) Heynh. 4CL protein sequences (Hamberger & Hahlbrock, 2004) as queries. Sequences of two gymnosperms, *Picea glauca* (Moench) Voss and *Pinus taeda* L., and the liverwort *M. polymorpha* were retrieved from the expressed sequence tags database of GenBank. The target sequences were selected such that the pairwise amino acid identity between the queries and the targets was over 40% (Tian & Skolnick, 2003). Some highly divergent members were further excluded from the final phylogenetic analysis if they had a putative peroxisomal targeting signal of type 1 (PTS1) at the C termini, which has only been found in 4CL-like members (Schneider et al., 2005). The *4CL* gene sequences of the other 10 gymnosperm species were obtained by polymerase chain reaction (PCR) and cloning here (Table S2). Voucher specimens were deposited in the herbarium of the Institute of Botany, Chinese Academy of Sciences (PE) (Table S2).

1.2 DNA and RNA extraction, PCR and reverse transcription-PCR amplification, cloning, and sequencing

Most of the *4CL* sequences of the 10 gymnosperms were obtained from both genomic DNAs and cDNAs. Some sequences in *Thuja*, *Larix*, and *Ginkgo* could not be obtained easily from genomic DNA, and thus were amplified from cDNA. Genomic DNA extraction and PCR amplification followed the protocols of Ran et al. (2006) except that annealing was at 52–57 °C and the sequence extension lasted for 3 min. Protocols for total RNA extraction and purification, first-strand cDNA synthesis and RT-PCR, and cloning and sequencing are those of Ran et al. (2010). The primers used for PCR and sequencing, species names, gene names, gene accession numbers, and sources of genome databases are listed in Tables S1 and S2.

1.3 Phylogenetic analysis

Coding DNA sequences of the *4CL* genes were aligned using the program CLUSTALX version 2.0 (Thompson et al., 1997) and manually adjusted in BioEdit version 7.0.9 (Hall, 1999). The highly variable regions at the N-terminals and in exon 1 were removed. Substitution saturation at each codon position was checked using DAMBE version 5.1.1 (Xia & Xie, 2001), which indicated that the third codon positions were saturated. Therefore, only the first and second codon positions were used in phylogenetic reconstruction. The jModeltest 0.1.1 (Posada, 2008) program was used to determine the best-fit model for the nucleotide dataset, with GTR+G+I and TPM3uf+G+I indicated as the best models using the Akaike Information Criterion and Bayesian Information Criterion, respectively. We also used the deduced amino acid sequences for

phylogenetic analysis. The Prottest 3 (Darriba et al., 2011) program was used to determine the best-fit model for the amino acid data, and the JTT+G+I were indicated according to the Akaike Information Criterion and Bayesian Information Criterion. Maximum likelihood (ML) analyses were carried out using PhyML version 3.0 (Guindon & Gascuel, 2003), and the GTR+G+I and JTT+G+I models were used for the nucleotide and amino acid datasets, respectively, with a BIONJ tree as a starting point. The support values for nodes were estimated using 100 bootstrap replicates (Felsenstein, 1985). Bayesian inference analyses were carried out with the MrBayes 3.1.2 program (Huelsenbeck & Ronquist, 2001), and the parameter settings were nst = 6 and rates = invgamma for the nucleotide dataset (corresponding to GTR+G+I), and aamodelpr = fixed (jones) and rates = invgamma for the amino acid dataset (corresponding to JTT+G+I). Four chains of the Markov Chain Monte Carto were run, each for 1 000 000 generations, and were sampled every 100 generations. The first 300 samples for each run were discarded as burn-in. Phylogenetic inferences were based on those trees sampled after generation 30 000.

1.4 Southern blotting

Southern blotting was used to detect the number of *4CL* loci in *Ginkgo biloba* and *Cycas revoluta*. Genomic DNAs of the two species were extracted from leaves using the DNAsecure Plant Kit (Tiangen Biotech, Beijing, China) and quantified by the Nanodrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). Approximately 20 μg genomic DNA of each species was digested with different restriction enzymes (*Eco*RI, *Eco*RV, *Hind*III, and *Xba*I), separated on a 0.8% agarose gel, then transferred to a nylon Hybond N+ membrane using the VacuGene XL Vacuum Blotting System (Amersham Biosciences, Little Chalfont, UK). *Ginkgo 4CL1* (456 bp), *Ginkgo 4CL2* (456 bp), *Cycas 4CL1* (363 bp), and *Cycas 4CL2* (498 bp) sequences covering partial exon 1 of corresponding members from *G. biloba* and *C. revoluta* were used as the probes for Southern hybridization analysis. All probes were obtained by PCR amplification from species-specific clones and labeled with alkaline phosphatase using the Gene Images AlkPhos Direct Labeling and Detection System (GE Healthcare (formerly Amersham Biosciences), Piscataway, NJ, USA). The restriction enzymes used do not have recognition sites in the probe sequence, except that *Xba*I has a recognition site in *Ginkgo 4CL1* around the 356th nucleotide, *Eco*RV in *Ginkgo 4CL2* at the 32nd nucleotide, and *Eco*RI in *Cycas 4CL2* at the 483rd nucleotide. The membrane was hybridized and washed at 56 °C for all probes in both taxa (except for probe *Ginkgo 4CL2* in the case of *G. biloba*, where we used 72 °C for hybridization).

1.5 Testing for selection

An ML analysis was used to identify regions that may have been subject to diversifying selection, using both the Fitmodel program version 0.5.3 (Guindon et al., 2004) and the codeml program in 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 carried out for the land plant *4CL* genes. The branch-site, branch, and site models (codeml program) were carried out for four clades (angiosperm Class I (AI), angiosperm Class II (AII), gymnosperm Class I (GI), gymnosperm Class II (GII)) (see Fig. 1), separately, due to the highly variable sequences among them (Anisimova et al., 2001, 2002; Anisimova & Yang, 2007). The parameter settings in these analyses followed Guo et al. (2010).

2 Results

2.1 Sequence characterization

The *4CL* genes obtained from the 27 studied species are listed in Table S1. Only three to nine members were detected in each species with available whole genome sequences (Table 1), which is much less than that reported in Xu et al. (2009). The *4CL* coding sequences range from 1557 to 1764 bp, and the most variable regions are located at the N-terminals and in exon 1 (data not shown). The structure of the *4CL* gene is relatively conserved, and most members have six exons and five introns. The fourth intron is maintained in all members of land plants, although several independent intron loss or gain events likely occurred in individual clades or members (Fig. 2). Three important intron loss and gain events occurred in the evolution of the *4CL* gene family, including the gain of intron 3 and intron 5 in vascular plants, the loss of intron 3 in AI, and an additional intron gain in one subclade of monocots in AI. Interestingly, single introns were inserted into different locations of exon 1 of *Ppa4CL1*, *4*, *Smo4CL2*, *8*, *Mtr4CL1*, *Sbi4CL5*, *Csa4CL3*, and the rest monocots of AI (Figs. 1, 2).

2.2 Phylogenetic analysis

The ML tree generated from the first and second codons of the *4CL* genes indicates that both vascular and seed plant *4CL* genes form monophyletic groups with 77% and 96% bootstrap support, respectively (Fig. 1). Members from angiosperm and gymnosperm could be further divided into two (AI and AII) and three (GI, GII,

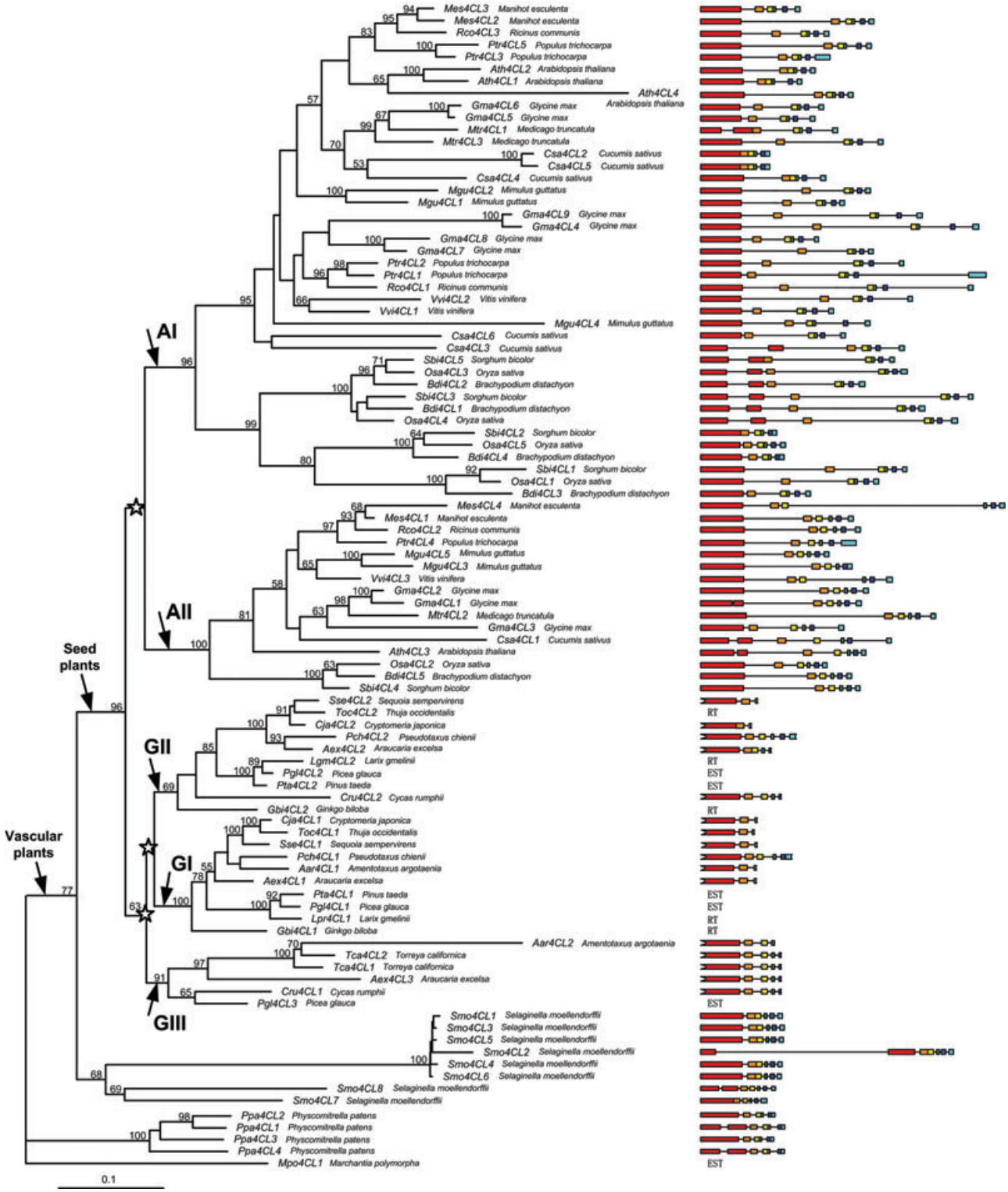

Fig. 1. Maximum likelihood tree of the 4-coumarate:coenzyme A ligase (*4CL*) genes constructed based on the nucleotide sequence (excluding the third codon position) with *Marchantia polymorpha* (*Mpo4CL*) used as an outgroup. Numbers above branches indicate bootstrap values higher than 50%. The stars denote several major inferred duplication events that occurred in the ancestors of gymnosperms or angiosperms. The diagrams on the right show the structure of each gene, including exons (boxes) and introns (lines). Gene names and identifiers are shown in Table S1. AI, angiosperm Class I; AII, angiosperm Class II; EST, expressed sequence tags retrieved from NCBI; GI, gymnosperm Class I; GII, gymnosperm Class II; GIII, gymnosperm Class III; RT, sequences amplified from cDNA.

Table 1 4-Coumarate:coenzyme A ligase (*4CL*) gene family in 14 land plant species with whole genome sequences analyzed in this study

Taxon	Copy number of *4CL*			Website	Reference
	Class I	Class II	Total		
Physcomitrella patens (Hedw.) Bruch & Schimp ssp. *patens*	_	_	4	http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html	(Rensing et al., 2008)
Selaginella moellendorffii Hieron.	_	_	8	http://genome.jgi-psf.org/Selmo1/Selmo1.home.html	(Wang et al., 2005)
Arabidopsis thaliana (L.) Heynh	3	1	4	http://www.ncbi.nlm.nih.gov/	(Ehlting et al., 1999)*
Brachypodium distachyon (L.) P. Beauv.	4	1	5	http://www.phytozome.net	(Vogel, 2010)
Cucumis sativus L.	5	1	6	http://www.ncbi.nlm.nih.gov/	(Huang et al., 2009)
Glycine max (L.) Merr.	6	3	9	http://www.phytozome.net	(Schmutz et al., 2010)
Manihot esculenta Crantz	2	2	4	http://www.phytozome.net	_
Mimulus guttatus Fischer ex DC.	3	2	5	http://www.phytozome.net	_
Medicago truncatula Gaertn.	2	1	3	http://www.medicagohapmap.org/?genome	(Cannon et al., 2006)
Oryza sativa L. ssp. *japonica*	4	1	5	http://www.ncbi.nlm.nih.gov/	(Hamberger et al., 2007)*
Populus trichocarpa Torr. & Gray	4	1	5	http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html	(Hamberger et al., 2007)
Ricinus communis L.	2	1	3	http://castorbean.tigr.org/	_
Sorghum bicolor (L.) Moench	4	1	5	http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html	(Paterson et al., 2009)
Vitis vinifera L.	2	1	3	http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/	(Jaillon et al., 2007)

*Sequences downloaded directly with gene accession numbers as in the references. –, Gene not present.

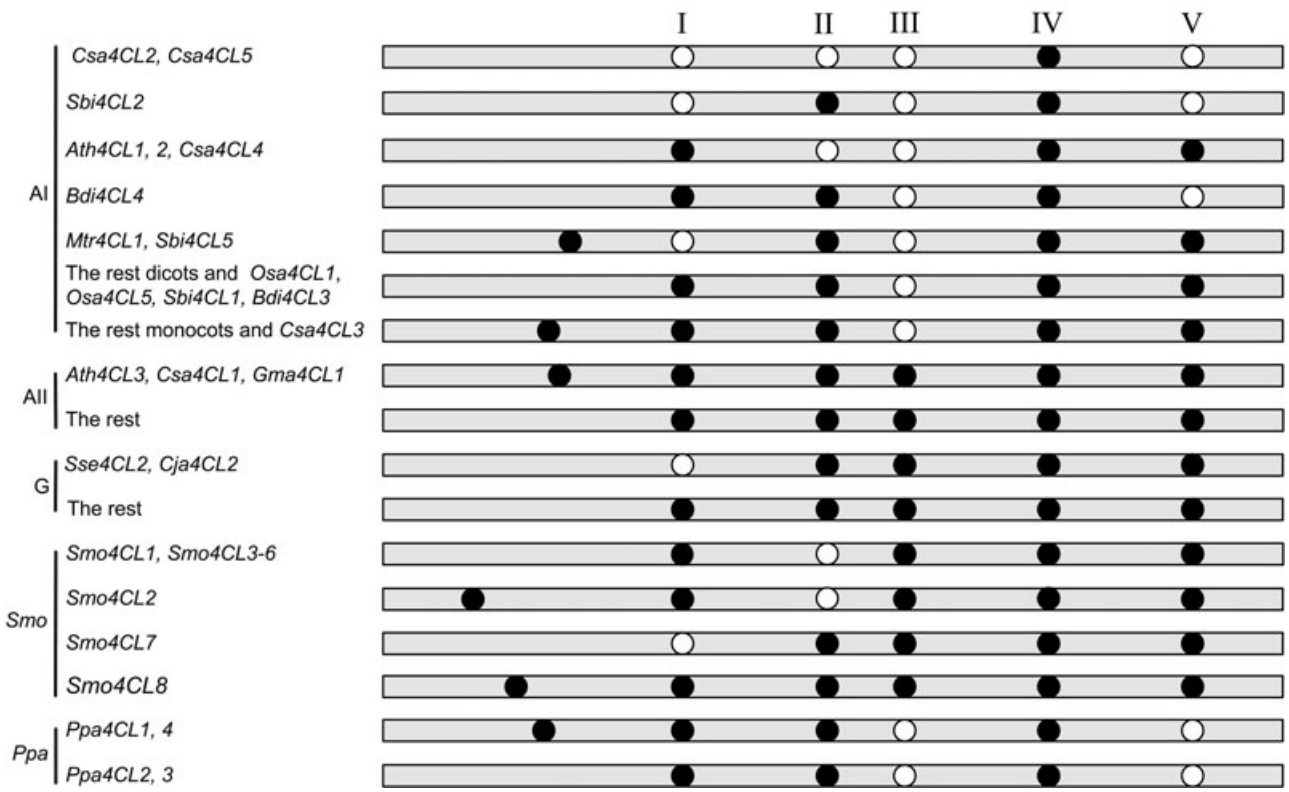

Fig. 2. Locations of the 4-coumarate:coenzyme A ligase (*4CL*) gene introns in terrestrial plants. Intron absence and presence are represented by white and black circles, respectively. I–V indicate the number of introns. AI, angiosperm Class I; AII, angiosperm Class II; G, gymnosperm; *Smo*, *Selaginella moellendorffi*; *Ppa*, *Physcomitrella patens.*

and GIII) clades, respectively (Fig. 1). All angiosperm species recovered from whole genomes, except *Manihot esculenta* Crantz, had more members from AI than from AII (Table 1). Most gymnosperms recovered using PCR had members of both GI and GII, whereas GIII members were only detected in four conifer species. We recovered members of all the three classes from *Araucaria excelsa* (Lamb.) R. Br. and *Picea glauca* (Fig. 1; Table S3). The protein sequence identity between any two *4CL* classes of a gymnosperm ranges from 62% to 78%, which is similar to that between AI and AII (Hu et al., 1998). The Bayesian trees and the ML tree generated from the deduced amino acid sequences are nearly identical to Fig. 1 in topology except for some branches with low bootstrap support (data not shown).

Both ancient and recent duplications occurred in the evolutionary history of the *4CL* gene family. Duplication events might have occurred in the ancestors of angiosperms and gymnosperms, which gave rise to the two angiosperm classes and the three gymnosperm classes. Three duplications might also have occurred in the ancestor of Poaceae. Additionally, there were possibly many interspecific duplications, leading to multiple versions of the AI gene recovered in several eudicots, such as *Arabidopsis* (*Ath4CL1*, *4CL2*, and *4CL4*), *Glycine* (*Gma4CL5*, *4CL6*, *4CL7*, and *4CL8*), and *Populus* (*Ptr4CL3* and *4CL5*) (Fig. 1). Mapping the location of other duplications (and extinctions) within each major clade (e.g., using the method described in Page & Charleston, 1997; Wehe et al., 2008) may require substantially greater taxon density, so we did not attempt this here.

2.3 Southern blot analysis

When the *Ginkgo 4CL1* probe was used, two clear signals (bands) were detected in *G. biloba* and one to four weak signals were found in *Cycas revoluta* in each of the restriction enzyme digests (Fig. 3: a). In contrast, only one band was detected in each digestion for both species when the *Ginkgo 4CL2* probe was used (Fig. 3: b). We also blotted the genomic DNA of *C. revoluta* with two *Cycas* probes (*Cycas 4CL1* and *Cycas 4CL2*), and detected one to five bands in each lane (Fig. S1).

2.4 Test for selection

The selection test using the Fitmodel program showed that none of the sites and branches was under positive or relaxed selection ($PP>0.9$). In addition, the PAML analysis did not detect any positively selected site or branch in this gene family (data not shown).

3 Discussion

3.1 Evolution of the *4CL* gene family

As a gene encoding a branch point enzyme in the phenylpropanoid metabolism, the *4CL* gene family has drawn substantial interest from plant biologists due to its potential use in plant lignin genetic engineering (Lee et al., 1997; Hu et al., 1999; Wagner et al., 2009; Voelker et al., 2010; Yang et al., 2011). However, most previous studies focused on the functions of the *4CL* genes, especially concerning the impact of antisense-mediated downregulation of *4CL* expression on plant development, and lignin content and composition. The evolution of the *4CL* genes is still poorly understood. Initially, the *4CL* gene family was classified into two classes based on data from some angiosperms (Ehlting et al., 1999; Hamberger & Hahlbrock, 2004). However, a *4CL* gene of the gymnosperm *Pinus taeda* was found to be located outside classes I and II (Hamberger et al., 2007), and the encoded enzyme was shown as having broad substrate specificity in mixed substrate assays (Harding et al., 2002; de Azevedo Souza et al., 2008). Therefore, Hamberger et al. (2007) inferred that the angiosperm classes I and II might have evolved after the angiosperm–gymnosperm split and that the apparent diversification within Class II could be attributed to recent independent duplication events in angiosperm lineages. In accordance with earlier studies (Ehlting et al., 1999; Lindermayr et al., 2002; Hamberger & Hahlbrock, 2004), Silber et al. (2008) also divided the angiosperm *4CL* genes into two groups. They found that the four moss *4CL* genes formed a highly supported monophyletic group distinct from the higher plant *4CL*s, but the members from moss and gymnosperm formed a clade sister to the angiosperm Class I. The phylogenetic analysis by Silber et al. (2008) is helpful to our understanding of the evolutionary scenario of the *4CL* gene family, but it suffered from a very limited taxon sampling.

In the present study, we analyzed the *4CL* genes from 27 species representing most major lineages of land plants. The *4CL* gene copies we obtained from each species with available whole genome sequences (Table 1) are much fewer than that reported in Xu et al. (2009). This might be caused by the inclusion of some *4CL*-like genes in Xu et al. (2009). For example, they reported 13 *4CL* genes in *Arabidopsis*, but only four of them are true *4CL* genes (Ehlting et al., 1999). Based on the phylogeny of the *4CL* gene family, we found that all members from the moss *Physcomitrella patens* form a monophyletic group sister to the vascular plant clade (Fig. 1), in which seed plant, gymnosperm, and angiosperm *4CL* genes each form monophyletic groups, respectively. In addition, our results indicate that at least

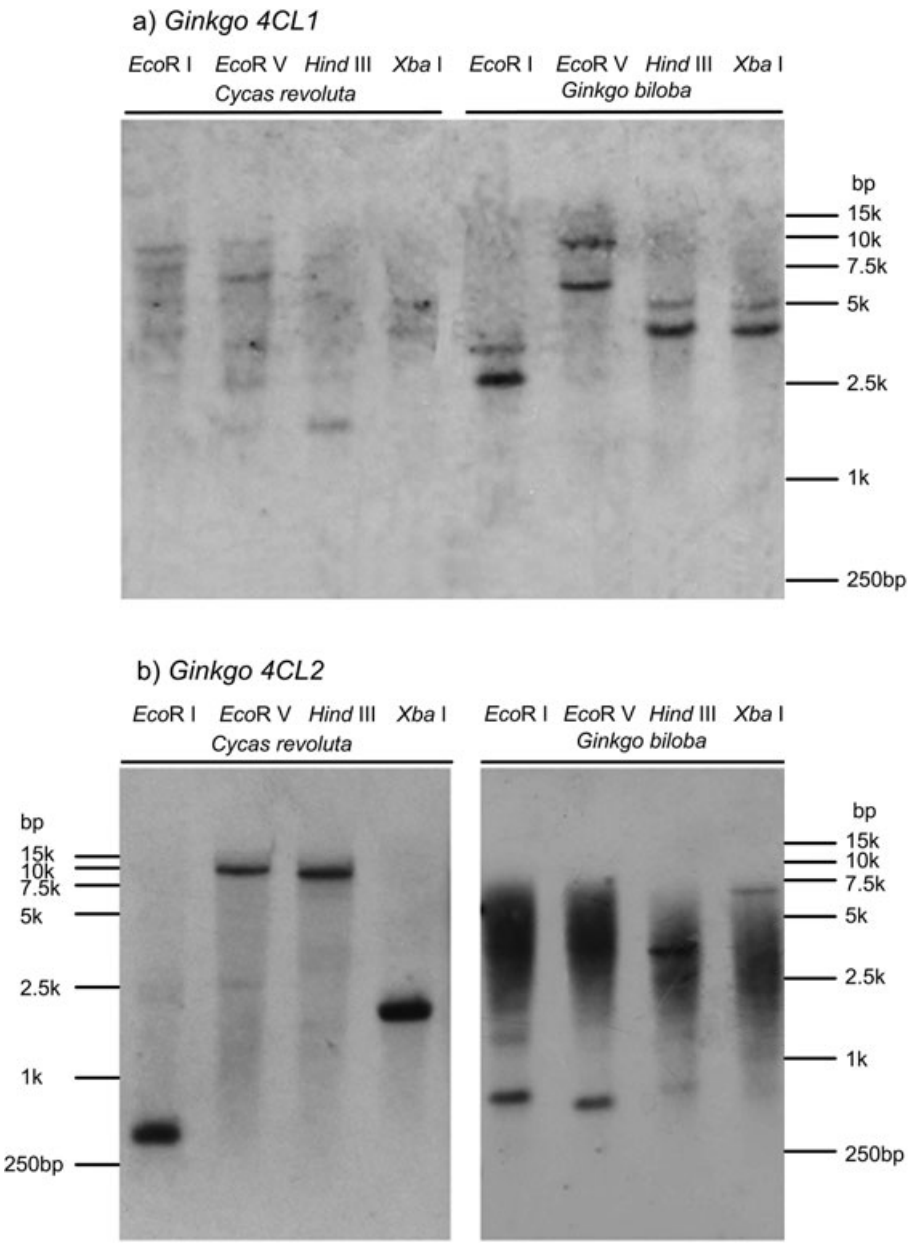

Fig. 3. Southern blot hybridization of the genomic DNAs of *Ginkgo biloba* and *Cycas revoluta* with the *Ginkgo 4CL1* (a) and the *Ginkgo 4CL2* (b) probes. Numbers on the right indicate DNA molecular weight marker.

three and two major classes could be further recognized in gymnosperm and angiosperm *4CL* genes, respectively (Fig. 1). The designation of AI and AII is also consistent with the loss of the third intron in AI (Fig. 2).

Previous evolutionary studies of different genes within a synthetic/metabolic pathway have reported that upstream genes (Rausher et al., 1999; Lu & Rausher, 2003; Riley et al., 2003; Cork & Purugganan, 2004; Ramsay et al., 2009) and genes encoding enzymes at biochemical pathway branch points (Eanes, 1999; Whitt et al., 2002; Flowers et al., 2007) or enzymes catalyzing multiple steps (Yang et al., 2009) usually have the lowest evolutionary rates due to strong purifying selection. By comparing the *4CL* gene family in the pathway upstream and branch point and the *CAD/SAD* gene family in the pathway downstream that we studied earlier (Guo et al., 2010), we found that the evolutionary patterns of the two gene families are very different. First, major evolutionary and functional divergence of the *4CL* gene family might have occurred through

neo- or sub-functionalization after the divergence of seed plants rather than at a very early stage of the land plant evolution, as suggested by Silber et al. (2008). In contrast, the radiation of the *CAD/SAD* gene family occurred at least before the divergence of vascular plants (Guo et al., 2010). Second, there is variation in the nine residues that constitute the putative substrate binding pocket (Hu et al., 2010) among conspecific or non-conspecific members of each angiosperm and gymnosperm *4CL* class (Table S4), suggesting that *4CL* might also differ from *CAD* in its functional differentiation pattern. Finally, no positive or relaxed selection was detected in the evolution of the *4CL* gene family (data not shown), whereas some amino acid sites and branches of the *CAD/SAD* gene family have been under relaxed selection. This could imply that the *4CL* gene in the pathway upstream and branch point has experienced stronger selective constraints than the *CAD/SAD* gene in the pathway downstream.

Previous functional studies of the plant *4CL* genes indicated that some members of AI and GI play major roles in lignin biosynthesis, such as *Ath4CL1* and *Ath4CL2* (Ehlting et al., 1999), *Ath4CL4* (Hamberger & Hahlbrock, 2004), *Ptr4CL3* (orthologous to *Pt4CL1*) (Hu et al., 1998), and *Pta4CL1* (Zhang & Chiang, 1997), whereas some members of AII might play a role in flavonoid biosynthesis, such as *Ath4CL3* (Ehlting et al., 1999) and *Ptr4CL4* (orthologous to *Pt4CL2*) (Hu et al., 1998) (Fig. 1). However, these studies mainly focused on enzyme activity and gene expression, and gene-specific knockout mutations have never been carried out. In addition, there are many controversial results in the previous studies. Therefore, more studies are needed to uncover the precise functions of each *4CL* gene, including how they affect lignin content and composition. This may then allow us to engineer plants that are more amenable to chemical degradation in the future.

3.2 Two new *4CL* classes in gymnosperms

In angiosperms with available whole genome sequences, the *4CL* gene family comprises three to nine members, which can be distinctly classified into two major classes (Table 1, Fig. 1). In gymnosperms, although some studies have shown that there are more than one *4CL* copies in the same species (Zhang & Chiang, 1997; Wang et al., 2000; Wei & Wang, 2004), those reported sequences share high sequence similarity and all belong to GI designated in the present study. This might imply that the *4CL* gene family, at least GI, has expanded in gymnosperms. Surprisingly, we found two new *4CL* classes in gymnosperms (GII and GIII). Members of GII were detected in many gymnosperms including *Ginkgo* and nine genera of four coniferous families (Araucariaceae, Cupressaceae s.l., Pinaceae, and Taxaceae). *Cycas* might also harbor members of GI and GII according to the results of Southern blotting (Fig. 3), although we did not recover GI members from this group. We detected more than one band in each lane in the Southern blot using genomic DNA of *C. revoluta* with two *Cycas* probes (*Cycas 4CL1* and *Cycas 4CL2*), which might suggest that the numbers of copies shown in Fig. S1 are underestimates in this species and likely in other taxa. The two classes might have originated from one ancient gene duplication that occurred in the early evolution of gymnosperms. Members of GIII were found only in four species, which represent four families of conifers, that is, Pinaceae, and three conifer II families (Araucariaceae, Cephalotaxaceae, and Taxaceae). This could be due to the loss of GIII members in the other gymnosperm species or PCR bias, which needs to be examined in future studies. Also, more work is needed to attribute specific functions to the different *4CL* classes of gymnosperms, especially GII and GIII.

Acknowledgements We thank Dr. Zu-Yu YANG for her help with the Southern blotting, Ms. Wan-Qing JIN for her assistance in the DNA sequencing, and two anonymous reviewers for their insightful comments and suggestions. This work was supported by the Chinese Academy of Sciences (The 100-Talent Project, and Grant No. KSCX2-EW-J-1) and the National Natural Science Foundation of China (Grant Nos. 30990240 and 30425028).

References

Anisimova M, Yang ZH. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Molecular Biology and Evolution 24: 1219–1228.

Anisimova M, Bielawski JP, Yang ZH. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Molecular Biology and Evolution 18: 1585–1592.

Anisimova M, Bielawski JP, Yang ZH. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Molecular Biology and Evolution 19: 950–958.

Anterola AM, Lewis NG. 2002. Trends in lignin modification: A comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. Phytochemistry 61: 221–294.

Baucher M, Halpin C, Petit-Conil M, Boerjan W. 2003. Lignin: Genetic engineering and impact on pulping. Critical Reviews in Biochemistry and Molecular Biology 38: 305–350.

Boerjan W, Ralph J, Baucher M. 2003. Lignin biosynthesis. Annual Review of Plant Biology 54: 519–546.

Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang X-H, Mudge J, Vasdewani J, Schiex T, Spannagl M, Monaghan E, Nicholson C, Humphray SJ, Schoof H, Mayer KFX, Rogers J, Quétier F, Oldroyd GE, 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. Proceedings of the National Academy of Sciences USA 103: 14959–14964.

Cork JM, Purugganan MD. 2004. The evolution of molecular genetic pathways and networks. Bioessays 26: 479–484.

Costa MA, Bedgar DL, Moinuddin SGA, Kim KW, Cardenas CL, Cochrane FC, Shockey JM, Helms GL, Amakura Y, Takahashi H, Milhollan JK, Davin LB, Browse J, Lewis NG. 2005. Characterization *in vitro* and *in vivo* of the putative multigene 4-coumarate:CoA ligase network in *Arabidopsis*: Syringyl lignin and sinapate/sinapyl alcohol derivative formation. Phytochemistry 66: 2072–2091.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: Fast selection of best-fit models of protein evolution. Bioinformatics 27: 1164–1165.

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

Eanes WF. 1999. Analysis of selection on enzyme polymorphisms. Annual Review of Ecology and Systematics 30: 301–326.

Ehlting J, Buttner D, Wang Q, Douglas CJ, Somssich IE, Kombrink E. 1999. Three 4-coumarate:coenzyme A ligases in *Arabidopsis thaliana* represent two evolutionarily divergent classes in angiosperms. The Plant Journal 19: 9–20.

Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783–791.

Flowers JM, Sezgin E, Kumagai S, Duvernell DD, Matzkin LM, Schmidt PS, Eanes WF. 2007. Adaptive evolution of metabolic pathways in *Drosophila*. Molecular Biology and Evolution 24: 1347–1354.

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

Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP. 2004. Modeling the site-specific variation of selection patterns along lineages. Proceedings of the National Academy of Sciences USA 101: 12957–12962.

Guo D-M, Ran J-H, Wang X-Q. 2010. 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*. Journal of Molecular Evolution 71: 202–218.

Hall T. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95–98.

Hamberger B, Hahlbrock K. 2004. The 4-coumarate:CoA ligase gene family in *Arabidopsis thaliana* comprises one rare, sinapate-activating and three commonly occurring isoenzymes. Proceedings of the National Academy of Sciences USA 101: 2209–2214.

Hamberger B, Ellis M, Friedmann M, de Azevedo Souza C, 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. Canadian Journal of Botany 85: 1182–1201.

Harding SA, Leshkevich J, Chiang VL, Tsai CJ. 2002. Differential substrate inhibition couples kinetically distinct 4-coumarate:coenzyme A ligases with spatially distinct metabolic roles in quaking aspen. Plant Physiology 128: 428–438.

Hu W-J, Kawaoka A, Tsai CJ, Lung J-H, Osakabe K, Ebinuma H, Chiang VL. 1998. Compartmentalized expression of two structurally and functionally distinct 4-coumarate:CoA ligase genes in aspen (*Populus tremuloides*). Proceedings of the National Academy of Sciences USA 95: 5407–5412.

Hu W-J, Harding SA, Lung J, Popko JL, Ralph J, Stokke DD, Tsai CJ, Chiang VL. 1999. Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. Nature Biotechnology 17: 808–812.

Hu Y-L, Gai Y, Yin L, Wang X-X, Feng C-Y, Feng L, Li D-F, Jiang X-N, Wang D-C. 2010. Crystal structures of a *Populus tomentosa* 4-coumarate:CoA ligase shed light on its enzymatic mechanisms. The Plant Cell 22: 3093–3104.

Huang S-W, Li R-Q, Zhang Z-H, Li L, Gu X-F, Fan W, Lucas WJ, Wang X-W, Xie B-Y, Ni P-X, Ren Y-Y, Zhu H-M, Li J, Lin K, Jin W-W, Fei Z-J, Li G-C, Staub J, Kilian A, van der Vossen EAG, Wu Y, Guo J, He J, Jia Z-Q, Ren Y, Tian G, Lu Y, Ruan J, Qian W-B, Wang M-W, Huang Q-F, Li B, Xuan Z-L, Cao J-J, Asan, Wu Z-G, Zhang J-B, Cai Q-L, Bai YQ, Zhao B-W, Han Y-H, Li Y, Li X-F, Wang S-H, Shi Q-X, Liu S-Q, Cho WK, Kim JY, Xu Y, Heller-Uszynska K, Miao H, Cheng Z-C, Zhang S-P, Wu J, Yang Y-H, Kang H-X, Li M, Liang H-Q, Ren X-L, Shi Z-B, Wen M, Jian M, Yang H-L, Zhang G-J, Yang Z-T, Chen R, Liu S-F, Li J-W, Ma L-J, Liu H, Zhou Y, Zhao J, Fang X-D, Li G-Q, Fang L, Li Y-R, Liu D-Y, Zheng H-K, Zhang Y, Qin N, Li Z, Yang G-H, Yang S, Bolund L, Kristiansen K, Zheng H-C, Li S-C, Zhang X-Q, Yang H-M, Wang J, Sun R-F, Zhang B-X, Jiang S-Z, Wang J, Du Y-C, Li S-G. 2009. The genome of the cucumber, *Cucumis sativus* L. Nature Genetics 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.

Kajita S, Hishiyama S, Tomimura Y, Katayama Y, Omori S. 1997. Structural characterization of modified lignin in transgenic tobacco plants in which the activity of 4-coumarate:coenzyme A ligase is depressed. Plant Physiology 114: 871–879.

Knobloch KH, Hahlbrock K. 1975. Isoenzymes of ρ-Coumarate:CoA ligase from cell-suspension cultures of

Glycine max. European Journal of Biochemistry 52: 311–320.

Knobloch KH, Hahlbrock K. 1977. 4-coumarate:CoA ligase from cell suspension cultures of *Petroselinum hortense* Hoffm.: Partial purification, substrate specificity, and further properties. Archives of Biochemistry and Biophysics 184: 237–248.

Lee D, Meyer K, Chapple C, Douglas CJ. 1997. Antisense suppression of 4-coumarate:coenzyme A ligase activity in *Arabidopsis* leads to altered lignin subunit composition. The Plant Cell 9: 1985–1998.

Li L, Zhou Y-H, Cheng X-F, Sun J-Y, Marita JM, Ralph J, Chiang VL. 2003. Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. Proceedings of the National Academy of Sciences USA 100: 4939–4944.

Lindermayr C, Möllers B, Fliegmann J, Uhlmann A, Lottspeich F, Meimberg H, Ebel J. 2002. Divergent members of a soybean (*Glycine max* L.) 4-coumarate:coenzyme A ligase gene family. Primary structures, catalytic properties, and differential expression. European Journal of Biochemistry 269: 1304–1315.

Lu Y-Q, Rausher MD. 2003. Evolutionary rate variation in anthocyanin pathway genes. Molecular Biology and Evolution 20: 1844–1853.

Page RD, Charleston MA. 1997. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. Molecular Phylogenetics and Evolution 7: 231–240.

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, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature 457: 551–556.

Peter G, Neale D. 2004. Molecular basis for the evolution of xylem lignification. Current Opinion in Plant Biology 7: 737–742.

Posada D. 2008. jModelTest: Phylogenetic model averaging. Molecular Biology and Evolution 25: 1253–1256.

Ramsay H, Rieseberg LH, Ritland K. 2009. The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. Molecular Biology and Evolution 26: 1045–1053.

Ran J-H, Gao H, Wang X-Q. 2010. Fast evolution of the retroprocessed mitochondrial *rps*3 gene in Conifer II and further evidence for the phylogeny of gymnosperms. Molecular Phylogenetics and Evolution 54: 136–149.

Ran J-H, Wei X-X, Wang X-Q. 2006. Molecular phylogeny and biogeography of *Picea* (Pinaceae): Implications for phylogeographical studies using cytoplasmic haplotypes. Molecular Phylogenetics and Evolution 41: 405–419.

Rausher MD, Miller RE, Tiffin P. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Molecular Biology and Evolution 16: 266–274.

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.

Riley RM, Jin W, Gibson G. 2003. Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in *Drosophila*. Molecular Ecology 12: 1315–1323.

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.

Schneider K, Hovel K, Witzel K, Hamberger B, Schomburg D, Kombrink E, Stuible HP. 2003. The substrate specificity-determining amino acid code of 4-coumarate:CoA ligase. Proceedings of the National Academy of Sciences USA 100: 8601–8606.

Schneider K, Kienow L, Schmelzer E, Colby T, Bartsch M, Miersch O, Wasternack C, Kombrink E, Stuible HP. 2005. A new type of peroxisomal acyl-coenzyme A synthetase from *Arabidopsis thaliana* has the catalytic capacity to activate biosynthetic precursors of jasmonic acid. The Journal of Biological Chemistry 280: 13962–13972.

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.

Simmons BA, Logué D, Ralph J. 2010. Advances in modifying lignin for enhanced biofuel production. Current Opinion in Plant Biology 13: 313–320.

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. Nucleic Acids Research 25: 4876–4882.

Tian W, Skolnick J. 2003. How well is enzyme function conserved as a function of pairwise sequence identity? Journal of Molecular Biology 333: 863–882.

Vanholme R, Morreel K, Ralph J, Boerjan W. 2008. Lignin engineering. Current Opinion in Plant Biology 11: 278–285.

Voelker SL, Lachenbruch B, Meinzer FC, Jourdes M, Ki CY, Patten AM, Davin LB, Lewis NG, Tuskan GA, Gunter L, Decker SR, Selig MJ, Sykes R, Himmel ME, Kitin P, Shevchenko O, Strauss SH. 2010. Antisense

down-regulation of 4CL expression alters lignification, tree growth, and saccharification potential of field-grown poplar. Plant Physiology 154: 874–886.
Vogel JP. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463: 763–768.
Wagner A, Donaldson L, Kim H, Phillips L, Flint H, Steward D, Torr K, Koch G, Schmitt U, Ralph J. 2009. Suppression of 4-coumarate-CoA ligase in the coniferous gymnosperm *Pinus radiata*. Plant Physiology 149: 370–383.
Wang W-M, Tanurdzic M, Luo MZ, Sisneros N, Kim HR, Weng J-K, 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 Biology 5: 10.
Wang X-Q, Tank DC, Sang T. 2000. Phylogeny and divergence times in Pinaceae: Evidence from three genomes. Molecular Biology and Evolution 17: 773–781.
Wehe A, Bansal MS, Burleigh JG, Eulenstein O. 2008. DupTree: A program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics 24: 1540–1541.
Wei X-X, Wang X-Q. 2004. Evolution of 4-coumarate:coenzyme A ligase (*4CL*) gene and divergence of *Larix* (Pinaceae). Molecular Phylogenetics and Evolution 31: 542–553.
Whetten RW, MacKay JJ, Sederoff RR. 1998. Recent advances in understanding lignin biosynthesis. Annual Review of Plant Physiology and Plant Molecular Biology 49: 585–609.
Whitt SR, Wilson LM, Tenaillon MI, Gaut BS, Buckler IV ES. 2002. Genetic diversity and selection in the maize starch pathway. Proceedings of the National Academy of Sciences USA 99: 12959–12962.
Xia X, Xie Z. 2001. DAMBE: Software package for data analysis in molecular biology and evolution. Journal of Heredity 92: 371–373.
Xu Z-Y, Zhang D-D, Hu J, Zhou X, Ye X, Reichel KL, Stewart NR, Syrenne RD, Yang X-H, Gao P, Shi WB, Doeppke C, Sykes RW, Burris JN, Bozell JJ, Cheng Z-M, Hayes DG, Labbe N, Davis M, Stewart CN, Yuan J-S. 2009. Comparative genome analysis of lignin biosynthesis gene families across the plant kingdom. BMC Bioinformatics 10: 15.
Yang J, Chen F, Yu O, Beachy RN. 2011. Controlled silencing of 4-coumarate:CoA ligase alters lignocellulose composition without affecting stem growth. Plant Physiology and Biochemistry 49: 103–109.
Yang Y-H, Zhang F-M, Ge S. 2009. Evolutionary rate patterns of the gibberellin pathway genes. BMC Evolutionary Biology 9: 206.
Yang Z-H. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences 13: 555–556.
Yang Z-H. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586–1591.
Zhang X-H, Chiang VL. 1997. Molecular cloning of 4-coumarate:coenzyme A ligase in loblolly pine and the roles of this enzyme in the biosynthesis of lignin in compression wood. Plant Physiology 113: 65–74.

Supplementary Materials

The following supplementary materials are available online for this article:

Fig. S1. Southern blot hybridization of the genomic DNAs of *Cycas revoluta* with the *Cycas 4CL1* (a) and the *Cycas 4CL2* (b) probes at 56 °C. Numbers on the right indicate DNA molecular weight marker.
Table S1 Sources of the 4-coumarate:coenzyme A ligase (*4CL*) genes we analyzed, their accession numbers and expressed sequence tag homologs.
Table S2 Species used in the 4-coumarate:coenzyme A ligase (*4CL*) gene amplification and the primers used.
Table S3 Information on the 4-coumarate:coenzyme A ligase (*4CL*) gene family from 12 included gymnosperm species.
Table S4 Variation in the nine residues that constitute the putative substrate binding pocket of the 4-coumarate:coenzyme A ligase (*4CL*) gene.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.