Plant Mol Biol (2010) 72:247–263
DOI 10.1007/s11103-009-9565-z

Genome-wide analysis of the chalcone synthase superfamily genes of *Physcomitrella patens*

P. K. Harshavardhan Koduri · Graeme S. Gordon · Elizabeth I. Barker · Che C. Colpitts · Neil W. Ashton · Dae-Yeon Suh

Received: 5 July 2009 / Accepted: 19 October 2009 / Published online: 31 October 2009

Abstract Enzymes of the chalcone synthase (CHS) superfamily catalyze the production of a variety of secondary metabolites in bacteria, fungi and plants. Some of these metabolites have played important roles during the early evolution of land plants by providing protection from various environmental assaults including UV irradiation. The genome of the moss, *Physcomitrella patens*, contains at least 17 putative *CHS* superfamily genes. Three of these genes (*PpCHS2b*, *PpCHS3* and *PpCHS5*) exist in multiple copies and all have corresponding ESTs. *PpCHS11* and probably also *PpCHS9* encode non-CHS enzymes, while *PpCHS10* appears to be an ortholog of plant genes encoding anther-specific CHS-like enzymes. It was inferred from the genomic locations of genes comprising it that the moss *CHS* superfamily expanded through tandem and segmental duplication events. Inferred exon–intron architectures and results from phylogenetic analysis of representative *CHS* superfamily genes of *P. patens* and other plants showed that intron gain and loss occurred several times during evolution of this gene superfamily. A high proportion of *P. patens CHS* genes (7 of 14 genes for which the full sequence is known and probably 3 additional genes) are intronless, prompting speculation that *CHS* gene duplication via retrotransposition has occurred at least twice in the moss lineage. Analyses of sequence similarities, catalytic motifs and EST data indicated that a surprisingly large number (as many as 13) of the moss *CHS* superfamily genes probably encode active CHS. EST distribution data and different light responsiveness observed with selected genes provide evidence for their differential regulation. Observed diversity within the moss *CHS* superfamily and amenability to gene manipulation make Physcomitrella a highly suitable model system for studying expansion and functional diversification of the plant *CHS* superfamily of genes.

Keywords *Physcomitrella patens* · Chalcone synthase superfamily · Multigene family · Enzyme evolution · Gene duplication · Retrotransposition · Gene regulation · *cis*-acting elements

Introduction

Chalcone synthase (CHS, E.C. 2.3.1.74) catalyzes the first committed step of flavonoid biosynthesis. CHS condenses a phenylpropanoid CoA ester (e.g., *p*-coumaroyl-CoA) with three acetate units from malonyl-CoA molecules, and cyclizes the resulting intermediate to produce a chalcone (e.g., naringenin chalcone), the precursor of diverse flavonoids (Fig. 1a). CHS is the representative member of the CHS superfamily, also known as type III polyketide synthases, and is found in all plant species. Enzymes of the CHS superfamily exhibit similarity in sequence, structure and general catalytic principles in that they are homodimers of 40–45 kDa subunits and all contain a Cys-His-Asn catalytic triad in the active site (Fig. 1b, c; Schröder 1997;

Electronic supplementary material The online version of this article (doi:10.1007/s11103-009-9565-z) contains supplementary material, which is available to authorized users.

P. K. H. Koduri · G. S. Gordon · C. C. Colpitts · D.-Y. Suh (✉)
Department of Chemistry and Biochemistry, University of Regina, 3737 Wascana Parkway, Regina, SK S4S 0A2, Canada
e-mail: suhdaey@uregina.ca

E. I. Barker · N. W. Ashton
Department of Biology, University of Regina, 3737 Wascana Parkway, Regina, SK S4S 0A2, Canada

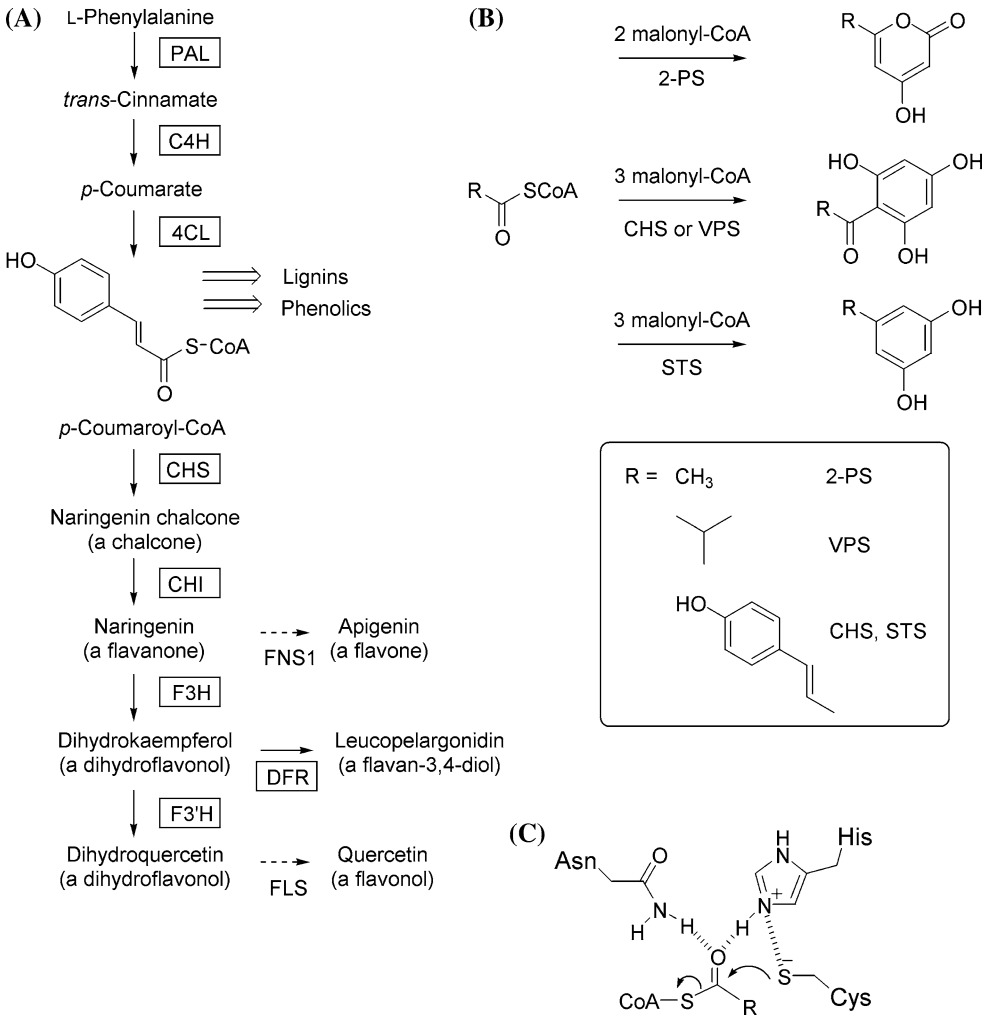


Fig. 1 **a** The phenylpropanoid pathway leading to flavonoids. *p*-Coumaroyl-CoA, the substrate for CHS, is also a substrate for the biosyntheses of lignins and phenolic compounds in plants. Genes encoding the enzymes marked with boxes are found in the *P. patens* genome, while no genes in the moss genome are annotated as *FNS1* or *FLS*. Full names of enzymes are: *PAL* phenylalanine-ammonia lyase, *C4H* cinnamate 4-hydroxylase, *4CL* 4-coumarate:CoA ligase, *CHI* chalcone isomerase, *F3H* flavanone 3-hydroxylase, *F3′H* flavonoid 3′-hydroxylase, *FNS1* flavone synthase, *DFR* dihydroflavanol 4-reductase, *FLS* flavonol synthase. **b** Representative enzymes of the CHS superfamily. 2-PS (2-pyrone synthase) condenses acetyl-CoA with two molecules of malonyl-CoA to produce triacetic acid lactone. While phlorisovalerophenone synthase (VPS) and CHS catalyze phloroglucinol ring formation to give phloroisovalerophenone and naringenin chalcone respectively, stilbene synthase (STS) catalyzes resorcinol ring formation to produce resveratrol. **c** Involvement of the catalytic CHN triad during the initial loading step of the CHS reaction. The nucleophilic Cys attacks the carbonyl carbon of the starter CoA substrate, while both His and Asn stabilize the developing oxyanion of the scissile thioester bond. For more details on the reaction mechanism, readers are referred to Austin and Noel (2003)

Austin and Noel 2003). Other members of the superfamily found in plants include stilbene synthase (STS; Austin et al. 2004), benzalacetone synthase (Abe et al. 2003), bibenzyl synthase (Han et al. 2006), biphenyl synthase (Liu et al. 2007), benzophenone synthase (Liu et al. 2003), 2-pyrone synthase (2-PS) (Jez et al. 2000a), pentaketide chromone synthase (Morita et al. 2007), octaketide synthase (Abe et al. 2005), and *p*-coumaroyltriacetic acid synthase (Akiyama et al. 1999). These enzymes differ from CHS in the choice of starter-CoA substrate, the number of condensation reactions catalyzed, and the mechanism by which the intermediate oligoketide is cyclized. For example, STS, found in grapes, peanuts and pine species, shows >65% amino acid sequence identity with CHS and also condenses *p*-coumaroyl-CoA with three molecules of malonyl-CoA. However, STS cyclizes the resulting tetraketide intermediate via an aldol-type reaction with a loss of CO_2 to yield resveratrol, a phytoalexin stilbene (Fig. 1b). In contrast, *Gerbera* 2-PS condenses acetyl-CoA with two molecules of malonyl-CoA to give a pyrone that is further converted to antifeedant glucoside metabolites (Eckermann et al. 1998). Therefore, enzymes of the CHS

superfamily are collectively responsible for biosyntheses of diverse natural products that play many roles in plants, including UV protection, antimicrobial defense, flower pigmentation, and pollen fertility. In recent years, members of the CHS superfamily have also been discovered in microorganisms (Gross et al. 2006). Alkylresorcinol synthase from *Azotobacter vinelandii* (Funa et al. 2006) and pentaketide resorcylic acid synthase from a fungus *Neurospora crassa* (Funa et al. 2007) are two recent additions to the growing list of non-plant members of the superfamily.

While flavonoids are ubiquitous in vascular plants, they have been found in only ~40% of liverworts and ~50% of mosses tested (Markham 1988; Iwashina 2000). Nevertheless, it has been proposed that flavonoids played a significant role in the early evolution of land plants, first as chemical messengers and then as UV filters (Stafford 1991). Thus, studies of CHS and related enzymes in nonvascular plants are required to obtain a clearer understanding of evolution of the flavonoid pathway in plants and of evolution of land plants themselves. Although many aspects of gene regulation and evolution of the CHS superfamily of seed plants have been studied, the literature contains little information on the superfamily from nonvascular plants, except that a gene presumably encoding a CHS was shown to be light-regulated in photoautotrophic cells of the liverwort, *Marchantia paleacea* (Harashima et al. 2004). Recently, a *CHS* gene (*PpCHS*) was cloned from the moss, *Physcomitrella patens*, and PpCHS was shown to exhibit enzymatic characteristics that are very similar to those of spermatophyte CHS enzymes. Thus, PpCHS prefers *p*-coumaroyl-CoA as the starter substrate, produces mostly pyrone derailment products from suboptimal substrates (e.g., hexanoyl-CoA), and shows comparable kinetic parameters to those of spermatophyte CHSs (Jiang et al. 2006).

Its informative evolutionary position and unparalleled amenability to gene targeting (Kamisugi et al. 2005, 2006) have made *P. patens* a popular and powerful model system for studying many plant phenomena. Furthermore, the data from several large-scale studies of the *P. patens* transcriptome and genome have been published (Nishiyama et al. 2003; Lang et al. 2005; Rensing et al. 2005, 2007; Quatrano et al. 2007). The 480 Mb haploid genome of the plant has been sequenced (Rensing et al. 2008), and the latest annotated version contains approximately 35,938 gene models distributed among 2,106 scaffolds, in turn composed of 19,136 contigs.

Earlier, the *P. patens* genome was shown to contain as many as 19 putative genes that may encode enzymes of the CHS superfamily (Jiang et al. 2006). The presence of multiple *CHS* superfamily genes in plants is not uncommon and seems widespread across taxa. Multiple *CHS* superfamily genes have been found in, to name a few, a pteridophyte *Psilotum nudum* (Yamazaki et al. 2001), a *Pinus* species (Fliegmann et al. 1992), and many flowering plants including *Petunia* (Koes et al. 1989), *Ipomoea* (Durbin et al. 2000), and *Glycine max* (Wingender et al. 1989). Many of the moss *CHS* superfamily genes share greater than 85% sequence identity to *PpCHS* when their deduced amino acid sequences are compared (Jiang et al. 2006). More interestingly, the *P. patens* genome contains *PpCHS10*, a moss homolog of anther-specific *CHS*-like genes of spermatophytes (Jiang et al. 2008; Ageez et al. 2005), and *PpCHS11*, which is basal to plant *CHS* superfamily genes in phylogenetic trees (Jiang et al. 2008). This diversity of *P. patens CHS* superfamily genes makes the moss a particularly attractive organism for studying evolution and gene regulation of the *CHS* superfamily. Herein we report our findings on gene architecture, evolution, regulatory elements and responsiveness to light of this gene group in *P. patens*.

Materials and methods

Genomic sequence analysis

The *P. patens* genes of the *CHS* superfamily were identified using the nucleotide and amino acid sequences of PpCHS (Jiang et al. 2006) as BLASTn and BLASTp queries, respectively, against the United States Department of Energy's Joint Genome Institute's (JGI) Physcomitrella genome database (http://genome.jgi-psf.org//Phypa1_1/). The GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) database in the JGI site were also term-searched for automatically annotated *CHS*. Each sequence obtained from these searches was examined for similarity to known *CHS* genes, scrutinized for an encoded catalytic CHN triad (Fig. 1c) (Ferrer et al. 1999), the GFGPG loop (Suh et al. 2000a) and the conserved Arg residues (Fukuma et al. 2007), and then grouped according to both UTR and coding sequences (CDS). ESTs corresponding to each gene were located in the *P. patens* EST database (www.cosmoss.org) (Lang et al. 2005). Intron structure was analyzed manually based on amino acid sequence alignments and known EST sequences. Amino acid sequence identity was calculated using the GraphAlign program (Spalding and Lammers 2004).

Phylogenetic analysis

Full-length amino acid sequences were aligned with ClustalX (Thompson et al. 1997). Phylogenetic analysis with the Bayesian inference method was performed using the MrBayes program (v. 3.1.2) (Ronquist and Huelsenbeck 2003). In this study, the "mixed" amino acid model that allows switching between various fixed-rate amino acid

models was used. Markov chain Monte Carlo analysis was performed for 500,000 generations with four independent chains. The Markov chain was sampled every 100 generations. At the end of the run, the average standard deviation of split frequencies and the potential scale reduction factors for all the run parameters were used to confirm that the state of convergence was obtained. After all trees sampled during earlier generations were discarded, a consensus tree was constructed based on the remaining trees and displayed using MEGA3.1 (Kumar et al. 2004). The Neighbor-joining (NJ) trees were built with the Poisson correction model as implemented in the MEGA3.1 program. The confidence of the tree topology was assessed by a bootstrap set of 1,000 replicates.

Gene duplication analysis

The tandem and segmental duplication history of *CHS* superfamily genes in *P. patens* was inferred by scrutinizing the scaffold sequences comprising each *CHS* superfamily gene and the flanking regions extending 50 kb in each direction. To investigate whether duplication of *CHS* genes had been mediated by (retro)transposable elements, the *CHS* superfamily genes and flanking DNA were scanned for overlapping sequences resembling those of known (retro)transposon families. Detection of gene conversion tracts was attempted using Geneconv (Sawyer 1989) to analyze amino acid sequences aligned in Clustal X and trimmed at the N-terminus where the alignment was poor.

Culture, light treatment and cDNA synthesis

Gametophytes of the *pabB4* strain of *P. patens* were grown axenically and agar-free on solid medium with or without ammonium tartrate at ~21°C under continuous light supplied by Cool White fluorescent tubes (Westinghouse, Regina, SK, Canada) for approximately 3 weeks as described (Ashton et al. 1985). The Petri dishes were covered with clear resin sheets (Roscolux, No. 114; MacPhon Industries, Calgary, AB, Canada) to reduce evaporation from the growth medium. Photon flux at the surface of the medium was 36–70 μmol m^{-2} s^{-1}. Cultures were then grown under light or in darkness for an additional 3 days, or were exposed to light for 3 days after growing in darkness for 3 days. Tissues were harvested, chilled in liquid nitrogen and stored (−80°C).

Total mRNA was isolated following the protocol for the Straight A's mRNA Isolation System (Novagen). cDNA was synthesized using an oligo-dT primer and PowerScript reverse transcriptase (Clontech) according to the manufacturer's specifications. The cDNA mixture was quantified spectrophotometrically at 260 nm, where A_{260} of 1.0 was the equivalent of approximately 50 μg/mL.

Semi-quantitative PCR

PCR was performed using KOD Hot Start polymerase (Novagen) and gene specific primers (Table 1). The PCR program initially started with a 94°C denaturation for 3 min, followed by 5 cycles of 94°C for 15 s, 65°C for 30 s and 68°C for 2.5 min. The annealing temperature was lowered to 60°C for the next 5 cycles, lowered again to 55°C for another 5 cycles, and then to 50°C for the final 23 cycles. After 32, 34, 36 and 38 cycles of PCR, aliquots (10 μL) for amplicon quantification were removed and electrophoresed through agarose gels (1.2%) pre-stained with ethidium bromide. A 1 kb DNA ladder (Fermentas) was used to confirm expected sizes of the amplification products (Table 1). Digital images of the gels were acquired with a Canon A75 camera (No. 22 filter) and the intensity of each band corresponding to a respective gene was determined using Molecular Imaging software (v. 4.0, Kodak). The *P. patens* actin 3 gene (*Act3*) was used as the internal reference in each PCR experiment. The ratio between the intensity of the sample gene and that of *Act3* was calculated to normalize for initial variations in sample concentration and as a control for reaction efficiency. The *P. patens* gene coding for NADPH-protochlorophyllide oxidoreductase, a light-dependent enzyme (Fujita 1996), was used as a positive control for light responsiveness.

Gene and protein notation

Gene symbols are italicized, e.g., *PpCHS1*, whereas symbols for the corresponding proteins are non-italicized, e.g., PpCHS1.

Results

Classification of *P. patens CHS* superfamily genes

In the March, 2007 release of the genome sequence of *P. patens* (Rensing et al. 2008), we found 17 genes belonging to the *CHS* superfamily. The gene names are unchanged from those used in an earlier publication (Jiang et al. 2006) except for *PpCHS01* (formerly *PpCHS1b*) and *PpCHS13c* (formerly *PpCHS1d*) for which new sequence information necessitated changes (Table 2). Additionally, *PpCHS13a* has been renamed *PpCHS*. *PpCHS12*, found in scaffold 639 of the earlier version of genome sequence, has been removed as a suspected microbial contamination in the new genome sequence. As our attempt to clone *PpCHS12* from genomic DNA using gene-specific primers was unsuccessful, this gene was excluded from further analysis. The current *P. patens* genome sequence contains full sequences of 14 *CHS* superfamily genes and three

Table 1 Gene-specific primers used in semi-quantitative RT-PCR analysis

Gene	Primer sequence (5′→3′)	Amplicon (bp)
PpCHS1a, *PpCHS01*	F: TTGTCCCATGGCTTCTGCTGGGGATG	1,883
	R: GGATTGCTTCTCAAATAAACCATGTAATTCAAGG	
PpCHS01	F: TTGTCCCATGGCTTCTGCTGGGGATG	1,763
	R: CAAAATCTGTACCCATACCACACTTTGTCACG	
PpCHS, *PpCHS13c*	F: TTGTCCCATGGCTTCTGCTGGGGATG	1,740
	R: GCAAACGAACACATGGTCCGGAGATTC	
PpCHS2a, *PpCHS2b*	F: GGACTCAACATCCATTCCGAACCCAAACG	1,268
	R: CGAACAACGCCATTAGGACGCCGC	
PpCHS2c	F: GGACTCAACATCCATTCCGAACCCAAACG	1,371
	R: GCCACAACCTGGATGCTTTCTGACCC	
PpCHS3, *PpCHS5*	F: CAGCAATGGCACCGCGAGCCGG	1,264
	R: GCCCTGCTTCCCCCATTTACACGTTCG	
PpCHS6	F: CTGAGAGGTTTAGTGCCAGCGAGC	1,253
	R: GACAGAGACTTCACACAATGGGCATGCTTC	
PpCHS11	F: GAAATCAAAACCATGGCAGACTTGGGCACTGAAAGCAAC	1,330
	R: GCACCGAATTCCATCATGAACAAGGATGTCAG	
Actin[a]	F: ATGGTCGGTATGGGACAGAAGGACGCG	939
	R: CCACATCTGCTGGAACGTACTCAGCG	
POR[b]	F: CTCGTCGCCATGTCTACCTGCG	1,113
	R: GTTCCAGCTCCAGTATACACCAGACTTG	

F forward primer, *R* reverse primer

[a] Internal reference gene. *P. patens* actin 3 (AY382283)

[b] Light response control gene. *P. patens* NADPH-protochlorophyllide oxidoreductase B precursor (XM_001768941)

additional gene copies. Deduced amino acid sequences of the 14 *CHS* superfamily genes include the following amino acid signatures: the catalytic triad of Cys-His-Asn (Cys^{170}-His^{309}-Asn^{342}, numbering of PpCHS; Fig. 1c; Ferrer et al. 1999), the two active site Phe residues (Phe^{221} and Phe^{271}; Jez et al. 2000b, 2002), and the GFGPG loop (Suh et al. 2000a) (Supplementary data Fig. S1). In our previous study, one of the genes, *PpCHS* (*PpCHS13a* in Jiang et al. 2006), was shown to encode a naringenin chalcone synthase. While amino acid sequence identity between PpCHS and each of the other enzymes of the *P. patens* CHS superfamily varies from 99.5 to 31% (Table 2), identity between PpCHS and representative seed plant enzymes belonging to the CHS superfamily was <65%. Higher sequence identity (~70%) was observed between PpCHS and a stilbenecarboxylate synthase from the liverwort, *Marchantia polymorpha*, (AAW30010) and a chalcone synthase-like enzyme from the lycopod, *Huperzia serrata*, (ABI94386), consistent with liverworts and lycopods being more closely related to mosses than are spermatophytes.

The 1,194 bp CDS of *PpCHS01* differs from the CDSs of *PpCHS1a and PpCHS1c* at 11 nucleotide positions, resulting in two amino acid mismatches. *PpCHS3* and *PpCHS5* are very similar in sequence. Between them, there are six nucleotide and three amino acid mismatches in the CDS. These differences are not due to sequencing errors since there are corresponding ESTs for each gene (see below). There are three copies and two copies of *PpCHS3* and *PpCHS5*, respectively, indicating multiple recent gene duplication events.

PpCHS7 contains an in-frame stop codon (TGA) 366 nt downstream from the start codon, indicating that it might be a pseudogene. However, the stop codon is found at a position where Trp (TGG) is highly conserved in other enzymes of the CHS superfamily (Fig. S1). Thus, it is possible that the observed stop codon reflects a sequencing error and *PpCHS7* is a functional gene. There are two ESTs derived from a cDNA clone (pphb3m11) in the cosmoss cDNA database that match the sequence of *PpCHS7*, although neither EST includes the sequence in question. Another notable aspect of *PpCHS7* is that it encodes a Tyr residue in place of the Cys^{170} that is essential for catalysis and absolutely conserved in the CHS superfamily (Lanz et al. 1991). Hence, the sequence surrounding Cys^{170} in PpCHS7 is MMCQTGYFGGA, while in other proteins (PpCHS, PpCHS2, etc.) it is MMYQTGCFGGA (Fig. S1). Since the Cys and Tyr codons differ by a single nucleotide, here too we cannot exclude the possibility that a sequencing error has occurred. The high sequence identity of PpCHS7 (90%) and PpCHS suggests that PpCHS7, if functional, is also a CHS. However, the intriguing possibility remains that PpCHS7 is a novel enzyme with different active site architecture.

PpCHS9 is not well represented in the existing EST databases. Of the two ESTs that correspond to this gene, one (PP015028177R) matches the 5′-UTR (from −162 to

Table 2 *CHS* superfamily genes from *P. patens*

Gene[a]	GenBank accession no.	JGI protein ID	Gene location Scaffold:base	Sequence identity to *PpCHS* (%)	Remarks
PpCHS01	XP_001756277	104998	22:12268	99.5	Differs at 2 a.a. and 3′-UTR from *PpCHS1a*
PpCHS1a	XP_001785666	201011	500:75157	99.2	No intron. Differs at 3 a.a. from *PpCHS*
PpCHS1c	XP_001784853	100508	426:238637	99.2	Differs at 3′-UTR from *PpCHS1a*
PpCHS2a	XP_001785363	101257	463:99838	94.0	One intron splits Cys^{72}
PpCHS2b.1	XP_001785368	101260	463:129624	94.0	Differs at 5′-UTR from *PpCHS2a*
PpCHS2c	XP_001751424	110814	1:81181	94.0	Differs at 5′- and 3′-UTRs from *PpCHS2a*
PpCHS3.1	XP_001781464	149692	303:69378	89.6	No intron
PpCHS3.2	XP_001781433	149682	303:79319	89.6	Another copy of *PpCHS3.1*
PpCHS3.3	XP_001781001	149180	292:1050	89.6	Another copy of *PpCHS3.1*
PpCHS4	XP_001759926	122336	39:1774363	91.6	No intron
PpCHS5.1	XP_001783453	152430	365:131783	90.1	No intron
PpCHS5.2	XP_001783444	98737	365:151777	90.1	Another copy of *PpCHS5.1*
PpCHS6	XP_001777888	63283	228:276349	85.5	One intron splits Cys^{76}
PpCHS7	XP_001784819	109184	425:82874	89.7	No intron
PpCHS9	XP_001757076	118540	25:2490141	42.1	One intron splits Cys^{77}
PpCHS10	XP_001781520	149790	304:419812	38.2	Two introns split Cys^{88} and Gly^{333}
PpCHS11	ABU87504	56368	34:1066567	30.9	Two introns split Cys^{135} and Arg^{403}
Putative CHS superfamily genes of which full sequences are not available					
PpCHS	ABB84527		410:292865	100	49 bp 5′-UTR, 847 bp cds
PpCHS13b	XM_001755075	38810	16:2859768	(100)	506 bp 5′-UTR, 369 bp cds
PpCHS13c			16:2867184		3′-UTR sequence
PpCHS2b.2	XM_001765408	129458	76:1642269		950 bp, another copy of *PpCHS2b.1*
Pseudogenes					
PpCHS8	XP_001756967	56017	25:2487652		One intron splits Asn^{2} and Lys^{3} (?)
PpCHSpg1	XP_001785366	155379	463:76478		Identical to *PpCHS2a* in 5′-UTR and intron sequences. 16 nt deletion causing frame-shift
PpCHSpg2	XP_001756020	68833	20:2009335		Two nonsense mutations, no intron.

[a] Gene names are according to Jiang et al. (2006) except for *PpCHS13a*, *PpCHS1b* and *PpCHS1d*, which have been renamed *PpCHS*, *PpCHS01* and *PpCHS13c*, respectively

−746) of *PpCHS9* and the other (PP020054304R) covers 229 nt of exon 1 and 116 nt of intron. This is the only EST, among the >500 ESTs analyzed in this study, that contains a portion of intronic sequence. A plausible explanation is that PP020054304R was produced by alternative splicing, although a stop codon exists immediately downstream from the exon–intron splice site. Alternatively, the pre-mRNA may have been incompletely processed. If PpCHS9 is functional, its lower sequence identity with PpCHS (42%) could indicate that it is a non-CHS enzyme, although this has yet to be established.

PpCHS10 displays a substantially higher amino acid sequence similarity (50–60%) to spermatophyte anther-specific CHS-like enzymes (ASCL) than to any other members of the *P. patens* CHS superfamily (30–40%). ASCLs are specifically expressed in anthers and comprise a monophyletic clade in phylogenetic reconstructions with sequences from various angiosperm and gymnosperm representatives of the CHS superfamily (Ageez et al. 2005; Jiang et al. 2008).

The sequence of *PpCHS11* contains four in-frame candidate translation initiation codons (Fig. S2). Based on similarity to the initiation consensus sequence (caA(A/C) aATGGCg) in plants (Lütcke et al. 1987; Joshi et al. 1997) and comparison of predicted amino acid sequences, obtained using the four potential start codons, with those of other *P. patens* CHS polypeptides, the first (most 5′) and third ATG are the most plausible candidates for the real start codon.

More *CHS* superfamily genes were recognized although their full-length sequences are not available. A partial sequence (896 bp) of *PpCHS* is located on scaffold 410, while two additional partial *CHS* superfamily sequences are found on scaffold 16. One of them, *PpCHS13b*, is 875 bp long and its sequence matches perfectly the *PpCHS* sequence except for two nucleotides in the CDS. The other,

PpCHS13c, is 1,142 bp long and contains 228 bp of 3′-UTR. A BLASTn search of the cosmoss (www.cosmoss.org) EST database (pp1206_fil) with the *PpCHS13c* sequence as query yielded six ESTs containing CDSs that are identical to the *PpCHS* sequence. Thus, *PpCHS13c* probably has the same CDS as *PpCHS* while differing from *PpCHS* in the 3′-UTR. A partial sequence of a second copy of *PpCHS2b* is located on scaffold 76 (*PpCHS2b.2*). Its 950 bp-long sequence, including a 258 bp intron, is identical to the sequence of *PpCHS2b.1* found on scaffold 463.

No EST sequence corresponding to *PpCHS8* was found, suggesting either that *PpCHS8* is not expressed under the standard culture conditions employed to obtain tissues for cDNA preparation or that it is a pseudogene. One candidate start codon, if correct, would give an intronless gene encoding a polypeptide of 359 amino acids, ~40 amino acids shorter than other enzymes of the *P. patens* CHS superfamily. However, upstream from this putative start codon, the well conserved sequence, G(I/V)GTA(V/N), can be discerned in MNKGRSAEGPAVILSIGTAVPPYVHE, derived by virtual translation assuming the presence of a phase-0 intron after the codon for asparagine (N) and thus inserted three nucleotides downstream from the ATG encoding the N-terminal methionine (Fig. S3), suggesting that this is the true start codon. This intron position is unique among known *CHS* superfamily genes, in which the intron is typically found at a highly conserved position ~180 nucleotides downstream from the start codon, suggesting that the putative intron in *PpCHS8*, if real, was acquired relatively recently.

Another *CHS* superfamily gene, *PpCHSpg1*, is found close to *PpCHS2a* on the same scaffold. The sequence of *PpCHSpg1* is identical to the *PpCHS2a* sequence in the 5′-UTR and the intron, but differs in the 3′-UTR and more importantly in the CDS. The *PpCHSpg1* CDS contains a 15 nt deletion as compared to the *PpCHS2a* CDS, causing a frame-shift (Fig. S4). A sequence error (deletion of a nucleotide) could be responsible for the frame-shift. However, no EST encompassing the deleted region has been found, giving further credence to the suggestion that *PpCHSpg1* is a pseudogene. Also present in the genome is another putative pseudogene (*PpCHSpg2*) with no matching EST in the databases. Virtual translation revealed that it contains two in-frame stop codons, presumably generated by nonsense mutations that degraded the meaningful CDS of a functional precursor of *PpCHSpg2* (Fig. S5).

Intron analysis

Most plant *CHS* superfamily genes have one intron that splits the Cys in the consensus sequence of (K/Q)R(M/I)C(D/E)KS (Harashima et al. 2004). Of the 14 genes whose full genome sequences are available, five genes, *PpCHS2a*, *PpCHS2b*, *PpCHS2c*, *PpCHS6*, and *PpCHS9*, each contain one intron at the conserved Cys (Fig. 2). In contrast, *PpCHS10* and *PpCHS11* have two introns each; one at the conserved Cys and the other at different sites closer to their C-termini (Figs. 2a, S1). The remaining seven genes (*PpCHS01*, *1a*, *1c*, *3*, *4*, *5*, *7*) are intronless (Table 2). The available genomic sequences of *PpCHS* and *PpCHS13b* lack the intron at the conserved splice site of Cys, and these two genes as well as another closely related gene, *PpCHS13c*, are most probably intronless.

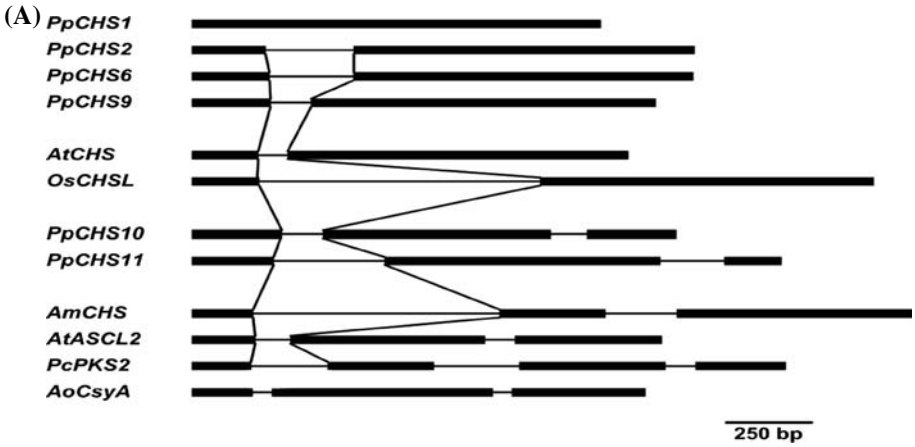
In the seven *P. patens* genes, the conserved intron at Cys is a phase-1 intron (T/GT), as is the case in all plant *CHS* superfamily genes whose gene structures are known. While the second intron in *PpCHS10* is also a phase-1 intron (G/GT), the second intron in *PpCHS11* is a phase-2 intron (AG/A; Fig. 2b). Conforming to the GT-AG rule, all of the introns begin with the nucleotides 'GT' and end with the nucleotides, 'AG.' Intron length varies from 107 nt (the second intron of *PpCH10*) to 325 nt (the first intron of *PpCHS11*). No other intronic sequence similarities were noticed. Blast searches using the intronic sequences against GenBank and the *P. patens* genome assembly (v1.1) yielded no significant hits.

Phylogenetic analysis

The NJ tree (Fig. 3a), built using amino acid sequences derived by virtual translation of full length CDSs of exclusively *P. patens* genes, and the Bayesian-inferred tree (Fig. 3b), which incorporates representatives from a broad range of plants and the cyanobacterium, *Synechococcus* sp., reveal an identical arrangement of *P. patens* sequences common to both trees with similar branch lengths. Furthermore, a previous phylogenetic study (Jiang et al. 2008) involving other enzymes of the thiolase superfamily had indicated that (a) the *Synechococcus* gene (*SyPKS*) is basally positioned relative to all genes belonging to the plant *CHS* superfamily and (b) the *P. patens* gene *PpCHS11*, which encodes a non-CHS protein (see below), forms a sister group to the rest of the plant *CHS* superfamily genes.

The NJ phylogram suggests progressive evolution with substantial late (moss lineage-specific) expansion of the *CHS* superfamily (Fig. 3a). *PpCHS9* and the putative pseudogene *PpCHS8* appear to have diverged relatively early in that order while *PpCHS6* diverged later prior to the relatively recent evolution of the gene clusters. One cluster contains *PpCHS4*, *PpCHS3* and *PpCH5* and the other cluster is comprised of *PpCHS2*, *PpCHS*, *PpCHS1*, *PpCHS01* and *PpCHS7*.

Next, the evolutionary relationship of the *P. patens* CHS superfamily and other plant enzymes was studied. Selected for phylogenetic analysis were the *P. patens* *CHS*



(B)

Gene	First intron	Second intron	Third intron
PpCHS10	CTCT GTGA (Cys)	CCAG GTGG (Gly)	
PpCHS11	ATCT GTAA (Cys)	GCAG ATCT (Arg)	
AmCHS	ATGT GCGA (Cys)	ACAG GGTT (Gln|Gly)	
AthASCL2	CTCT GCAA (Cys)	CCTG AAAC (Glu)	
PcPKS2	ATTT GTGA (Cys)	CTAG GATG (Gly)	CCTG GTGG (Gly)
Proto-splice site consensus	CAG A A G		

(C)

PpCHS10	-KLRRLCETTTV ^{93}-	-MLYMLGCYGGA ^{196}-	-QSFIPGTNKTI ^{282}-
PpCHS11	-FVDRICKASGI ^{140}-	-FVSFLGCHGGV ^{244}-	-STIIPNTSKSI ^{330}-
AmCHS	-KFKRMCDKSNI ^{65}-	-MMYQQGCFAGG ^{168}-	-QTLLPDSHGAI ^{254}-
AthASCL2	-KLTRLCKTTTV ^{67}-	-LLYFVGCSGGV ^{170}-	-QNFLPETEKTI ^{256}-
PcPKS2	-KFKRICENSRI ^{63}-	-MFYHLGCYAGG ^{166}-	-QTIVPESDGAV ^{252}-
PpCHS10	-WAVHPGGPAIL ^{338}-	-KLQTARQVLRD ^{361}-	
PpCHS11	-WAAHPGGKPIL ^{385}-	-QLENSRSVLEN ^{408}-	
AmCHS	-WIAHPGGPAIL ^{310}-	-KLRSTRQVLSE ^{333}-	
AthASCL2	-WAVHPGGPAIL ^{309}-	-KLSPSRRALMD ^{337}-	
PcPKS2	-WITHPGGPAIL ^{308}-	-KLKATRQVLND ^{331}-	

Fig. 2 **a** Exon–intron architecture of *P. patens CHS* superfamily genes. The schematic diagram depicts architectures of representative *CHS* genes from *P. patens*, *Antirrhinum majus*, *A. thaliana*, *Polygonum cuspidatum* and *Oryza sativa* plus the related *csyA* gene from *Aspergillus oryzae*. Black rectangles and lines represent exons and introns respectively. Vertical lines delineate introns at a conserved position. Included in the figure are *AtCHS* (*A. thaliana* CHS, Y18602), *OsCHSL* (*Oryza sativa* ASCL, NM_001071426), *AmCHS* (*Antirrhinum majus* CHS, X03710), *AthASCL2* (*A. thaliana* ASCL2, NM_119651), *PcPKS2* (*Polygonum cuspidatum* PKS2, EU305678), and *Aspergillus oryzae csyA* (AB206758). **b** Nucleotide sequences surrounding splice sites in multi-intron *CHS* superfamily genes are shown as well as the consensus sequence for the proto-splice site. Amino acid residues and codons (*underlined*) that are split by introns are also shown. **c** Amino acid sequences flanking the intron splice sites (*shaded*) in multi-intron *CHS* superfamily genes are shown. The catalytic Cys and His that are absolutely conserved in the superfamily are *underlined*

superfamily genes, all known genes encoding ASCL enzymes, plus representative *CHS* genes of the same species from which *ASCL*s were taken or, if not available, *CHS*s from closely related species. Excluding *PpCHS11*, shown in a previous study (Jiang et al. 2008) to be sister to the plant *CHS* superfamily, the plant genes are distributed in two sister groups in a Bayesian-inferred phylogram (Fig. 3b), one of them comprised exclusively of *ASCL*s. As expected based on sequence similarity, *PpCHS10* is included in the *ASCL* gene cluster (Fig. 3b). In the other sister group, *PpCHS9* and *PpCHS8* are basal and sister, respectively to the rest of the genes. The remaining *P. patens* genes form an exclusive sister group to the *CHS* genes from other plant taxa.

Gene duplication

Currently, the *P. patens* genome sequences are assembled in contigs and scaffolds. Half of the assembly is contained in 111 scaffolds, each of which is at least 1.3 Mb in length. Of the 24 *CHS* superfamily genes, pseudogenes and partial

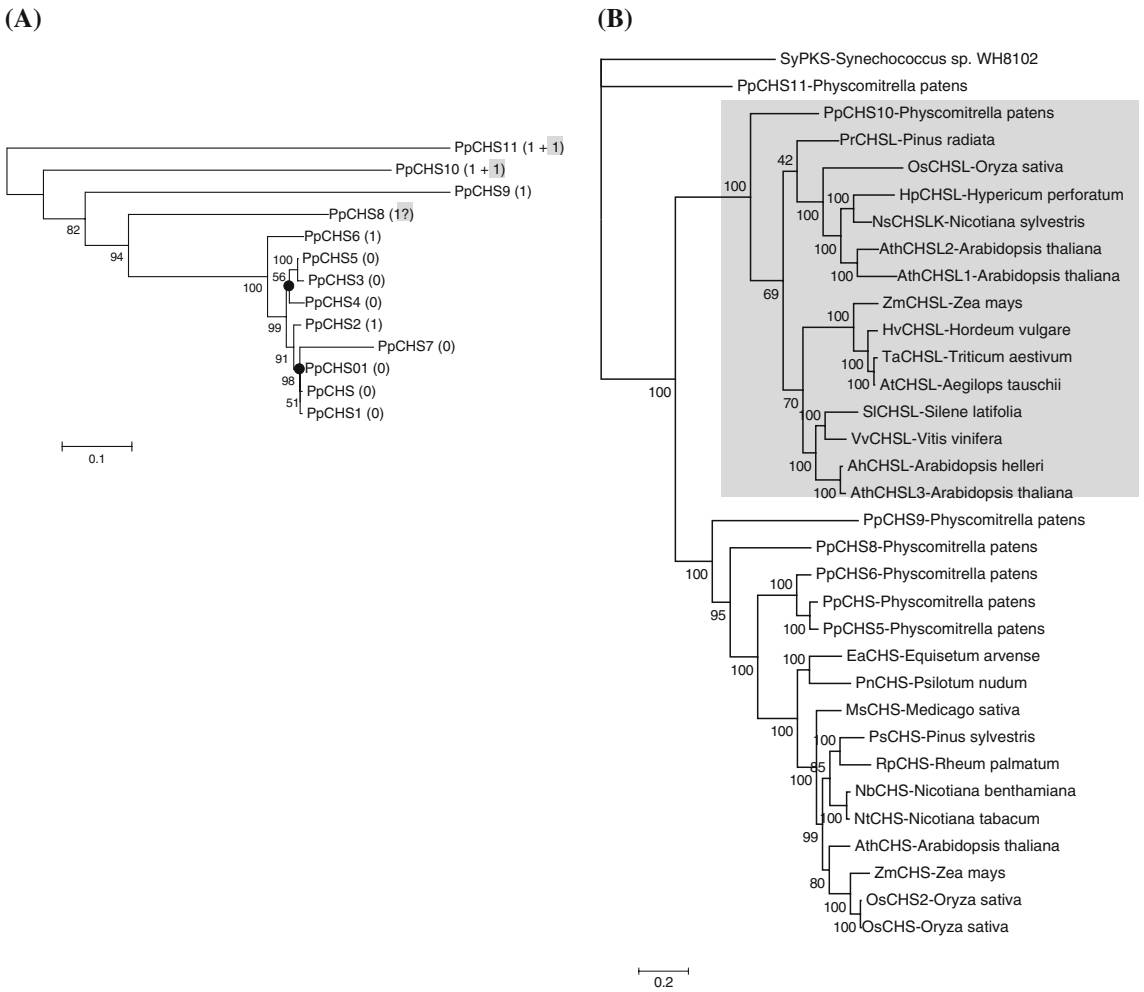


Fig. 3 Phylogenetic reconstructions of the *CHS* superfamily. **a** Neighbor-joining phylogram of the *P. patens* CHS superfamily, including *PpCHS8* but excluding other pseudogenes. *Numbers* above branches indicate bootstrap support for each clade. *Numbers* in *parentheses* refer to introns at a conserved position (*unshaded*) or at non-conserved positions (*shaded*). *Filled circles* represent postulated intronless ancestral genes, which may have arisen by retrotransposition. **b** Bayesian-inferred phylogram of representative *CHS* superfamily genes from *P. patens* (a moss), *Equisetum arvense* (an arthrophyte), *Psilotum nudum* (a psilophyte), *Pinus radiata* and *sylvestris* (gymnosperms) and various angiosperm taxa. Numbers above branches are posterior probabilities; branch lengths are proportional to expected numbers of amino acid substitutions per site. Amino acid sequences obtained by virtual translation of complete CDSs were aligned using Clustal X after protruding N- and C-ends were trimmed to be flush with the sequence of *Medicago sativa* CHS2 (MsCHS). At the end of the analysis, the average standard deviation of split frequencies reached 0.0076, and the potential scale reduction factors approached 1.000 for all run parameters. All trees sampled before 150,000 generations were discarded (burn-in = 1,500), a consensus tree was constructed from the remaining trees. All known anther-specific CHS-like enzymes (*shaded*) as well as a cyanobacterial CHS-like enzyme from *Synechococcus* sp. WH8102 (SyPKS) were included in the analysis. Accession numbers of the enzymes included in the analysis are: SyPKS (NP_897086), PrCHS (AAB 80804), OsCHSL (AAL59036), HpCHSL (ABP98922), NsCHSL (CAA74847), AthCHSL2 (NM_119651), AthCHSL1 (NM_100085), ZmCHSL (AY105247), HvCHSL (AAV49989), TaCHSL (CAJ 15412), AtCHSL (CAJ13966), AhCHSL (AAZ23686), AthCHSL3 (AAM63363), VvCHSL (CAO47307), SlCHSL (BAE80096), EaCHS (AB030004), PnCHS (AB022682), NbCHS (ABN80439), NtCHS (AAK49457), MsCHS (P30074), PsCHS (X60754), RpCHS (DQ205352), ZmCHS (CAA42763), OsCHS2 (BAA19186), OsCHS (BAB39764), AthCHS (AAB3581)

gene sequences in the *P. patens* genome, 11 gene sequences (four pairs and one triplet) are clustered on five scaffolds. Additionally, pairs of scaffolds contain similar clusters of one or more *CHS* superfamily genes along with genes from other gene families.

PpCHS is linked tail-to-tail to a putative phenylalanine ammonia-lyase gene (*PAL*) on scaffold 410 while *PpCHS1a* and *PpCHS01* are located next to two more putative *PAL* genes in head-to-head orientation on scaffolds 500 and 22, respectively (Fig. 4a). *PpCHS13b* and *PpCHS13c* are tandemly arranged on scaffold 16 in head-to-tail orientation with an intervening sequencing gap of 6.5 kb (Fig. 4b). *PpCHS8* and *PpCHS9* are tandemly linked on scaffold 25 in tail-to-tail orientation with only 1.4 kb separating their stop codons (Fig. 4c).

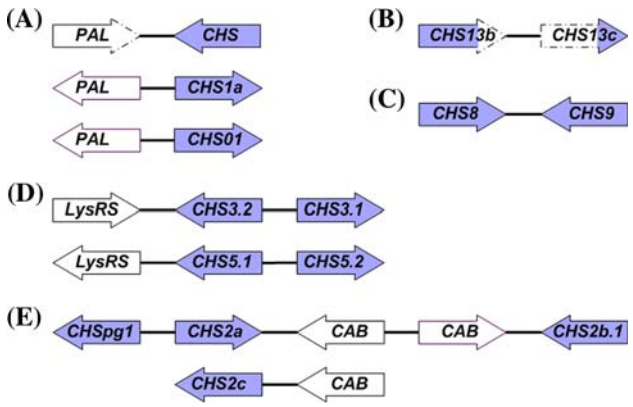


Fig. 4 Segmental and tandem duplication of *P. patens CHS* superfamily genes. *Diagram* (not to scale) shows scaffold locations of selected *P. patens CHS* genes and linked non-homologous genes in the *P. patens* genome. *Arrows* pointing in the 5′→3′ direction represent *CHS* genes (*shaded*), phenylalanine ammonia-lyase (*PAL*), lysyl tRNA-like genes (*LysRS*) and chlorophyll a/b binding protein (*CAB*) genes. *Dotted lines* indicate that the gene sequence is incomplete at the 5′- or 3′-end. Horizontal black lines represent intergenic regions

PpCHS3.1 and *PpCHS3.2* are found in head-to-head orientation with their start codons separated by ~8.8 kb, linked to a sequence similar to a lysyl-tRNA synthetase gene from *Arabidopsis* ~14.6 kb downstream of *PpCHS3.2* (Fig. 4d). Similarly, *PpCHS5.1* and *PpCHS5.2* are linked head-to-head with ~18.9 kb between their start codons and in turn are linked to a second lysyl-tRNA synthetase-like sequence ~19.6 kb downstream of *PpCHS5.1*. *PpCHS2a* and *PpCHS2b.1*, each linked tail-to-tail to a gene encoding a chlorophyll a/b binding protein (CAB), form tandem gene pairs in a cluster including the related pseudogene, *PpCHSpg1* spanning ~54.3 kb on scaffold 463 (Fig. 4e). *PpCHS2c* is located downstream of a third CAB gene in head-to-tail orientation on scaffold 1, with ~18.9 kb separating the start codon of *PpCHS2c* from the stop codon of the CAB gene.

Sequences related to (retro)transposable elements were found in the regions flanking *CHS* genes but did not overlap the *CHS* genes. Geneconv did not detect any evidence of gene conversion between *CHS* genes.

Expression analysis

The abundance of ESTs corresponding to each gene was examined by performing BLASTn with CDS and UTR sequences. The cosmoss pp1206_fil EST database contains 371,536 filtered sequences from different *P. patens* EST libraries derived from various tissues or mixtures of tissues from regenerated protoplasts, protonemata, gametophores, archegonia and sporophytes. The database also contains ESTs of unknown origin. Certain ESTs could not be assigned to a specific gene, since some genes share mostly identical sequences (e.g., *PpCHS1a* and *PpCHS1c*) and only partial sequences are known for a few other genes (e.g., *PpCHS13b* and *PpCHS13c*), For example, a total of 23 ESTs could only be assigned to *PpCHS1a* or *PpCHS1c*, but not specifically to either gene. Nonetheless, as shown in Fig. 5, it is clear that each gene is expressed at different levels in different tissues. While *PpCHS6* and *PpCHS01* are widely expressed in all the tissues represented by libraries, *PpCHS2a*, *PpCHS3* and *PpCHS5* are disproportionately expressed in the upper half portions of gametophores (ppls). ESTs corresponding to *PpCHS13c* are derived either from regenerated protoplasts or chloronemata, or from the mixture of chloronemata, caulonemata and rhizoid-like protonemata.

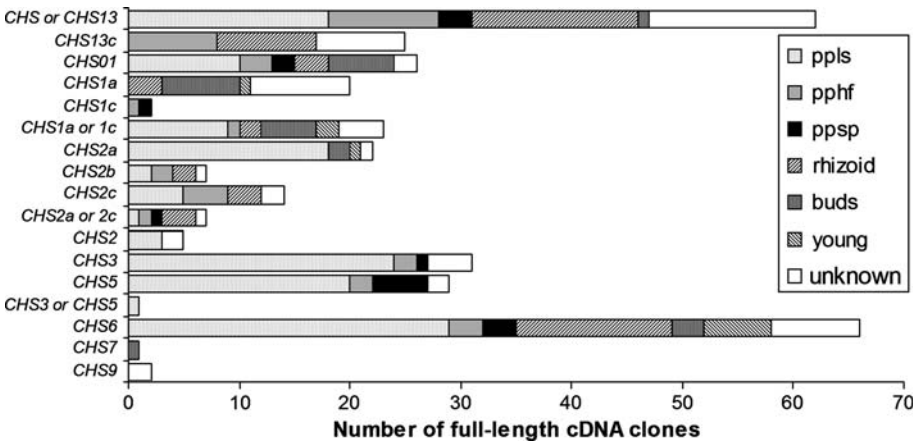


Fig. 5 Abundances of full-length cDNA clones of *CHS* superfamily genes in the *P. patens* EST database (www.cosmoss.org). The cosmoss database (pp1206_fil) contains 371,536 filtered sequences from different EST libraries. Counted for each gene are corresponding ESTs derived from upper halves of gametophores (**ppls**), regenerated protoplasts or chloronemata (**pphf**), sporophytes (embryos) with surrounding archegonia (**ppsp**), a mixture of chloronemata, caulonemata and rhizoid-like protonemata (**rhizoid**), a mixture of chloronemata, caulonemata and malformed buds (**buds**), and chloronemata and young gametophores (**young**). ESTs with unknown origin are grouped as "**unknown**."

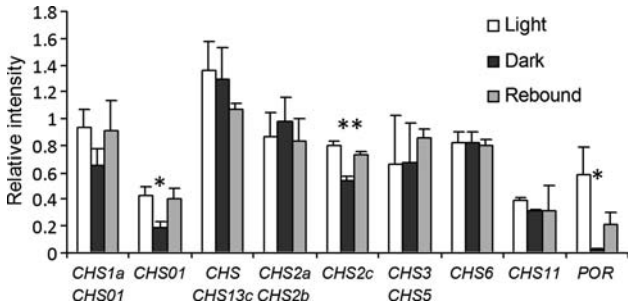


Fig. 6 Light responsiveness of *P. patens CHS* superfamily genes. Total mRNA was extracted from gametophores after different light treatments: 3 days under light (Light), 3 days in darkness (Dark), and 3 days in darkness followed by another 3 days under light (Rebound). cDNA was synthesized and amplified with gene-specific primers during 32 cycles of PCR. The expression level of each gene under each light treatment was determined by comparing the band intensity after agarose gel electrophoresis. The ratio between intensity of the sample gene and that of the internal reference gene (*actin 3*) was calculated to remove variations among different experiments. NADPH-protochlorophyllide oxidoreductase (*POR*) was used as a positive control for light responsiveness. Means and standard errors of at least three independent experiments are reported. * $P < 0.01$; ** $P < 0.001$ (ANOVA)

ESTs corresponding to *PpCHS4* are found only in the cDNA library (ppgs) derived from green sporophytes (e.g., GenBank accession no. DC957761, etc.), but not in the cosmoss database, indicating that this gene is exclusively expressed in moss sporophytes. There are more than 2500 ESTs corresponding to *PpCHS11* probably due to cloning artifacts, and the expression pattern of *PpCHS11* could not be studied.

Light responsiveness

To determine the changes in expression levels of *P. patens CHS* genes in response to light, semi-quantitative RT-PCR analysis was performed. Gametophytic tissues were collected after 3 days under white light (Light in Fig. 6) and after 3 days in the dark (Dark). Also, plant tissues were exposed to light for 3 days after having been kept in darkness for 3 days (Rebound). Due to sequence similarity, all of the genes could not be analyzed individually. Figure 6 shows that *PpCHS01* and *PpCHS2c* exhibited light responsiveness. The expression level of these two genes significantly decreased in dark and recovered, upon re-exposure to light, to the level observed under light (Fig. 6). A noticeable but not significant ($P = 0.17$) change observed in the level of expression of both *PpCHS1a* and *PpCHS01* might be due to the light responsiveness of *PpCHS01*. *POR*, which encodes NADPH-protochlorophyllide oxidoreductase and is regulated by light (Fujita 1996), responded to light to a greater extent under the same conditions.

Discussion

The CHS superfamily in *P. patens*

The complete complement of genes in *P. patens* belonging to the *CHS* superfamily comprises 17 genes, for 14 of which the complete CDS is known, plus 3 putative pseudogenes and multiple gene fragments. Three of these genes exist in multiple copies (*PpCHS2b*, *PpCHS3* and *PpCHS5*) and all are expressed under laboratory conditions. In most cases what enzymes are encoded by these genes is unknown. Only one of the genes, *PpCHS,* has been demonstrated to encode a *bona fide* CHS (Jiang et al. 2006). Lower sequence identity (<42%) of *PpCHS9*, *PpCHS10* and *PpCHS11* to *PpCHS* suggests that they encode non-CHS enzymes, although this has yet to be demonstrated. However, consistent with this contention, PpCHS11 was shown to produce long-chain (C_{12}–C_{18}) alkylpyrones and related compounds in vitro (unpublished data) and in this respect is similar to a pentaketide resorcylic acid synthase from the fungus *N. crassa* (Funa et al. 2007). Moreover, in a previous study (Jiang et al. 2008), phylogenetic reconstruction demonstrated that *PpCHS11* is positioned in a sister group separate from the rest of the *CHS* superfamily and it was proposed that PpCHS11 is an extant enzyme that may resemble more closely their most recent common ancestor than do other members of the plant CHS superfamily.

Among the *P. patens* genes, *PpCHS10* is unique in being orthologous to seed plant *ASCLs* (Fig. 3b), raising the exciting possibility that its function is similar to theirs. Spermatophyte ASCLs, vital for male fertility, are expressed in the tapetum during the early stages of microspore/pollen development and probably play a role in the biosynthesis of exine (outer wall of pollen grains) (Atanassov et al. 1998; Wu et al. 2008). A recent study showed that two *Arabidopsis* ASCLs (AthCHSL1 and AthCHSL2 in Fig. 3b) sequentially condense a long-chain (up to C_{20}) acyl-CoA with malonyl-CoA molecules to produce alkylpyrones in vitro (Mizuuchi et al. 2008). Our experiments indicated that PpCHS10 also functions in vitro as a long-chain alkylpyrone synthase (unpublished data). *PpCHS10* is not expressed in moss gametophytes but ESTs corresponding to *PpCHS10* are found in a *P. patens* sporophytic cDNA library derived from sporophytes harvested at different stages (thus containing developing spores) (S. Rensing, U. of Freiburg, personal communication). A case can be made that moss spores and pollen are homologous structures since both result from meiotic divisions, are protected by a multilayered coat containing sporopollenin (Domínguez et al. 1999) and germinate giving rise initially to a single filamentous apical cell or pollen tube which grows one-dimensionally. Therefore, PpCHS10 is

potentially a valuable tool for studying the roles of ASCLs in spore/pollen formation and seed plant male fertility.

Since sequence identity of non-CHS enzymes to CHS from the same plant is typically <70%, the enzymes of the *P. patens* CHS superfamily that show sequence identity of ~90% to PpCHS are probably also CHSs (Table 2). *PpCHS6*, which is 86% identical to *PpCHS* at the amino acid level, is a sister to the remaining moss *CHS* genes and thus may be most directly related to the gene encoding the ancestral enzyme of the *P. patens* CHS group (Fig. 3a). Ancestral enzymes of the plant CHS superfamily may have employed simple acyl-CoA esters as substrates before the metabolic pathway furnishing phenylpropanoid-CoA esters (e.g., *p*-coumaroyl-CoA) evolved (Jiang et al. 2008). In this context, PpCHS6, if not a CHS, may be a non-CHS such as phlorisovalerophenone synthase (VPS, Paniego et al. 1999) that differs from CHS only in the choice of starter-CoA and accepts isovaleryl-CoA (Fig. 1b).

The presence of a *CHS* multigene family is common in seed plants. *Ipomoea*, *Petunia*, and soybean species have all been shown to have at least 6–10 *CHS* genes. However, the fact that the moss has as many as 17 *CHS* genes and gene copies (assuming that those genes with ~90% amino acid sequence identity code for CHS) is unexpected and puzzling. Chalcones, the products of CHS, are considered minor products in seed plants. CHS functions together with other downstream enzymes in the flavonoid pathway to produce secondary metabolites with diverse functions such as UV protection, antimicrobial defense and flower pigmentation (Fig. 1a). Therefore, multiple *CHS* genes in spermatophytes are under differential temporal and spatial regulation to serve the various branches of the pathway. In contrast, mosses do not have such a well-developed flavonoid pathway and produce fewer flavonoids (Markham 1988). Flavones, biflavones, aurones, isoflavones, and 3-deoxyanthocyanins have been found in various moss species (Basile et al. 2003; Brinkmeier et al. 1999; Geiger and Markham 1992). However, to the best of our knowledge, no phytochemical study has been carried out on the flavonoid profile of *P. patens*.

The *P. patens* genome contains two genes for chalcone isomerase (CHI) and five genes for flavanone 3-hydroxylase (F3H). CHI converts chalcone to flavanone in the second step of flavonoid pathway, and F3H converts flavanone to dihydroflavonol in the third step. Flavone synthase that also utilizes flavanone to produce flavone seems to be absent in *P. patens* (Fig. 1a). Putative *CHS* genes outnumber these downstream genes and, thus, maintenance of gene stoichiometry does not explain the retention of the observed number of *CHS* genes in *P. patens*. It is possible that, in the moss, chalcones and/or their downstream flavonoid products (albeit of simpler profile) play various physiological roles that are performed by more sophisticated metabolites in seed plants. In this case, the multiple *CHS* genes are probably differentially regulated in response to varying metabolic needs in different parts of the plant at different developmental stages. This argument is supported by the EST distribution data showing divergent expression patterns of the *CHS* genes (Fig. 5). Some groups of genes share the same CDS but have dissimilar UTRs (e.g., *PpCHS2a*, *PpCHS2b* and *PpCHS2c*), suggesting that these genes are differently regulated. The 5′-UTR of *PpCHS3.1* diverges from those of *PpCHS3.2* and *PpCHS3.3* beyond 409 nt from the translation initiation site, although the three genes are classified as gene copies in this study. Thus, *PpCHS3.1* may also be subjected to different regulation from *PpCHS3.2* and *PpCHS3.3*. At least one example of different expression of such closely related genes was found in this study (see below).

Intron analysis

Most plant *CHS* superfamily genes contain a single phase-1 intron in the codon for the conserved Cys (T/GT or T/GC) in the consensus sequence of (K/Q)R(M/I)C(D/E)KS. The *P. patens CHS* gene family exhibits an unusual collection of genes in terms of exon–intron architectures in that there are intronless genes (*PpCHS*, *PpCHS01*, *PpCHS1*, *PpCHS3*, *PpCHS4*, *PpCHS5*, *PpCHS7*, and *PpCHS13*) and genes with two introns (*PpCHS10* and *PpCHS11*) as well as genes with one intron (*PpCHS2*, *PpCHS6*, and *PpCHS9*). While no other intronless *CHS* superfamily gene has been discovered from other plants, a few genes with more than two introns have been reported. Benzalacetone synthase (PcPKS2) from *Polygonum cuspidatum* is encoded by a three intron gene (Ma et al. 2009). The genes coding for *Antirrhinum majus* CHS (*AmCHS*, Sommer and Saedler 1986) as well as *Arabidopsis* ASCL (*AthASCL2*) contain two introns each (Fig. 2). All of these multi-intron genes share the first intron at the conserved Cys. Interestingly, fungal *CHS* superfamily genes differ from their plant counterparts in intron structure. Most fungal genes also have two introns. For example, the *Aspergillus oryzae csyA* gene (Seshime et al. 2005) has two intron sites at Ala59 (GC/T) and Gly274 (G/GA) (Fig. 2a). While the second intron splice site is conserved among the fungal genes, the first intron site is conserved in its position only (data not shown). Although the size of the first and second exons in the plant and fungal genes are similar, there is no apparent relatedness between plant and fungal intron splice sites. Taken together, these findings suggest that the acquisition of the first intron at the Cys in plant genes occurred at an early stage of plant evolution before the divergence of bryophytes and other lineages of plants, independently of the fungal lineage.

Unlike the first intron that is conserved in the plant multi-intron genes, the second and third introns are not

shared by all. However, their positions do not seem to be random but instead are situated within the codons for conserved amino acid residues (Fig. 2b, c). *PpCHS10* and *PcPKS2* share the same intron splice site within the codon for a highly conserved Gly, two amino acids downstream from the catalytic His (Jez and Noel 2000; Suh et al. 2000b). The second introns of *AmCHS* and *PcPKS2* occur at the codon for another conserved Gly adjacent to the catalytic Cys (Ferrer et al. 1999), although the *AmCHS* intron splice site (phase-0) occurs one nucleotide 5′ from the *PcPKS2* intron splice site (phase-1). The second *PpCHS11* intron splits the codon for an Arg residue, which has also shown to be highly conserved (Fukuma et al. 2007). The second *AthASCL2* intron is within the codon for an amino acid residue (Glu in AthASCL2) that is typically either Glu (GAA or GAG), Asp (GAT or GAC), Asn (AAT or AAC), or Gly (GGN) in plant CHS enzymes. The number of multi-intron genes is small and it may be premature to draw any conclusions regarding the intron evolution; nevertheless, two alternative explanations are proposed. (a) The five multi-intron plant *CHS* superfamily genes may have gained additional introns relatively recently and independently. In this case, our data provide some support for the controversial proto-splice site hypothesis (Dipp and Newman 1989; Long and Rosenberg 2000), which postulates that intron gain occurs in the consensus exon/intron 5′ splice junction of (C/A)AG|R, where '|'represents the intron insertion site. According to the hypothesis, such sequences may originate and persist in the absence of introns because of coding constraints and may serve as "hot spots" for intron gain (Dipp and Newman 1989). A few plant gene families including grass catalases have been presented as supporting evidence for intron gain at proto-splice sites (Frugoli et al. 1998 and references therein). As shown in Fig. 2b, the exonic nucleotide sequences flanking the second and third introns of the multi-intron *CHS* superfamily genes agree well with the consensus sequence of the proposed proto-splice site. (b) An alternative proposal contends that the ancestral plant *CHS* gene had two or more introns, of which all but the intron at the conserved Cys have been lost in most lineages. Sequential intron loss has been observed in a few plant gene families such as terpene synthase genes (Trapp and Croteau 2001) and legumin genes (Häger et al. 1996). In these gene families, more evolved genes lack one or more introns that are present in primitive genes, which differs from the situation observed in the *CHS* superfamily genes. *PpCHS11*, the gene proposed to most closely resemble a common ancestor of plant *CHS* superfamily genes (Jiang et al. 2008), has a unique second intron. Furthermore, *PpCHS10* and *AthASCL2*, both of which belong to the *ASCL* clade (Fig. 3b), do not share the same second intron. The fact that the insertion sites of *AmCHS* and *PcPKS2* differ by one nucleotide also provides an argument against the ancient origin of the second intron of these genes. Distinguishing which alternative is correct requires discovery and evaluation of more multi-intron genes in this superfamily.

Phylogenetic analysis, gene structure and sequence analysis and evolution of the *CHS* superfamily

The importance of gene duplication in evolution is well documented (Zhang 2003). Many genes encoding enzymes of plant secondary metabolism have been recruited following gene duplication. Gene families that have evolved via this mechanism include methyltransferases and terpene synthases (Ober 2005 and references therein).

Careful consideration and comparison of the topologies of trees shown in Fig. 3, as well as of others published elsewhere (Jiang et al. 2008), allow the following inferences concerning the contributions of gene duplication to evolution of the plant *CHS* superfamily. (1) Gene duplications giving rise to the progenitor of *PpCHS11* and the ancestor of all other *CHS* superfamily genes and giving rise to the progenitor of all *ASCL*s including *PpCHS10* and the ancestor of the remaining *CHS* superfamily genes preceded the separation of lineages leading, respectively to mosses and tracheophytes that occurred ~400 mya (Bateman et al. 1998). (2) Tracheophyte orthologs of *PpCHS11* were lost during evolution of the vascular plant lineage. (3) The genome of *P. patens* contains an ortholog, *PpCHS10*, of seed plant *ASCL*s, all of which were derived from the same ancestral gene. (4) Further expansion of the plant *ASCL* and *CHS* families has occurred after the separation of moss and tracheophyte lineages.

Tree topologies, gene architectural similarities and low Ks values indicate that some genes have evolved recently. Thus, intron loss and subsequent duplication have generated 10 intronless genes and their copies in *P. patens* (*PpCHS*, *PpCHS01*, *1a*, *1c*, *3*, *4*, *5*, *7*, *13b*, and *13c*). This is in marked contrast to other plants in which no intronless *CHS* gene has been discovered to date. Those sequences that have been acquired relatively recently, showing ~90% amino acid sequence identity to PpCHS, form two separate clusters in the NJ tree (Fig. 3a). *PpCHS4*, *PpCHS3* and *PpCHS5* form one cluster and the rest form the other. It is plausible that the observed clustering represents two independent evolutionary lineages. Whether evolution of these lineages has been accompanied by functional divergence requires investigation. Thus, at least one of the enzymes belonging to the *PpCHS4* cluster needs to be characterized. A parsimonious mechanism that plausibly accounts for the evolution of both clusters of intronless moss *CHS* genes involved two sequential retrotranspositions. One such event gave rise to an intronless

progenitor (indicated by a filled circle in Fig. 3a) of the *PpCHS4*, *PpCHS3*, *PpCHS5* gene cluster, while the other gave rise to another intronless progenitor (also indicated by a filled circle in Fig. 3a) of the *PpCHS1*, *PpCHS*, *PpCHS01* cluster. Subsequent tandem and segmental duplications of these progenitor genes and their derivatives generated the two intronless clusters (see below for further discussion).

The detection of several peaks of retrotransposon activity in the *P. patens* genome during the last 10 my (Rensing et al. 2008) lends support to this hypothesis. However, there is no consistent association of intronless CHS genes and retrotransposon sequences in the genome and there was no sign of reverse transcribed polyA sequences downstream from their stop codons. It should be noted that a Bayesian-inferred phylogram (not shown), similarly built using the *P. patens* genes, had similar overall tree topology and branch lengths to those in the NJ tree (Fig. 3a). One difference was that the intronless *PpCHS4* was sister to the other seven genes comprising the two recently evolved clusters, requiring a third retrotransposition to be invoked.

Phylogenetic reconstruction, linkage analysis and comparison of gene architectures indicate both tandem and segmental duplications have contributed to expansion of the *P. patens CHS* superfamily. *PpCHS8* and *PpCHS9*, although distinctly dissimilar in sequence, are linked in tandem on scaffold 25 in tail-to-tail orientation with only 1.4 kb separating their stop codons, suggesting they were produced by an ancient gene duplication (Fig. 4c). By contrast, a recent tandem duplication probably accounts for *PpCHS13b* and *PpCHS13c*, although the possibility that *PpCHS13b* and *PpCHS13c* are portions of a single gene erroneously separated by misassembly cannot be dismissed (Fig. 4b).

Three very similar *CHS* genes located on different scaffolds, *PpCHS*, *PpCHS1a* and *PpCHS01*, are linked to *PAL* genes that, taken pair-wise, are 97.1–98.5% identical at the nucleotide level supporting the occurrence of two relatively recent segmental duplication events (Fig. 4a). Segmental duplication of *PpCHS3* and a putative lysyl-tRNA synthetase-like gene that produced *PpCHS5* and a second putative lysyl-tRNA synthetase-like gene or vice versa is another recent event since *PpCHS3* and *PpCHS5* are 99.5% identical at the nucleotide level and the lysyl-tRNA synthetase-like genes are 97.7% identical at the nucleotide level. Even more recently, both *CHS* genes duplicated in tandem, since all three copies of *PpCHS3* exhibit identical coding sequences as do both copies of *PpCHS5*. Similarly, three genes (*PpCHS2a*, *PpCHS2b.1* and *PpCHS2c*) are products of tandem and segmental duplication events that probably occurred very recently since even the intronic sequences are identical. Furthermore, each possible pair of the three *CAB* genes linked to *PpCHS2* differ by only 1 or 2 nucleotides out of the 1,220 nt that comprise the CDS and the single intron.

Nonfunctionalization does not appear to be the fate of the majority of the recently duplicated genes. One exception is *PpCHSpg1*, which is highly similar to the *PpCHS2* genes (*2a*, *2b.1*, *2b.2* and *2c*). Thus, only one of five copies of *PpCHS2* has degenerated into a pseudogene. The existence of several copies of each of *PpCHS1*, *PpCHS2*, *PpCHS3*, *PpCHS5* and *PpCHS13* in which the coding sequences of the copies are identical, suggests that neofunctionalization of gene copies has not occurred. However, it is notable that the copies of *PpCHS1* and *PpCHS2* differ in their UTRs, indicating that the copies may be differentially regulated and that subfunctionalization may have occurred. Subfunctionalization may be a transitional step preceding neofunctionalization (Ober 2005) and it is possible that neofunctionalization may be the eventual fate of some copies of these genes. Retention of redundant genes may be required for production of sufficient protein product of some genes. Dosage balance of genes encoding proteins that form complexes has also been suggested as a selective advantage for retention of redundant gene copies. However, gene copies that have been retained for a long period of time would be expected to contain synonymous substitutions and, possibly, non-synonymous substitutions in codons corresponding to non-critical amino acid residues. It is possible, therefore, that redundancy of *CHS* superfamily genes in *P. patens* is, at least in some cases, a temporary condition following a recent burst of duplications.

Linkage of *PpCHS1a*, *PpCHS01*, and *PpCHS* to *PAL* genes (Fig. 4a) could indicate that these genes, which act in the same pathway (Fig. 1a), are co-regulated. However, although a flavonol reductase/cinnamoyl-CoA reductase gene was found approximately 47 kb downstream from *PpCHS4*, no other flavonoid pathway gene was found within 50 kb of a *CHS* gene, indicating that physical proximity to other genes in the flavonoid pathway is not a general feature of *CHS* genes in *P. patens*. The significance, if any, of the linkage of *CHS* genes with *CAB* and *LysRS* genes in the moss genome is unknown.

Overrepresentation of metabolic genes in the moss genome has been noted previously and attributed to preferential retention of their paralogs after an ancient genome duplication that occurred between 30 and 60 mya (Rensing et al. 2007). This study adds *CHS* superfamily genes to the list of metabolic genes that are overrepresented. This may have provided mosses with the variability and plasticity of secondary metabolism that would have allowed them to adapt more readily to environmental changes.

Gene expression and light responsiveness

Since flavonoids (UV protection), isoflavonoids (*Rhizobium* nodulation), anthocyanins (flower color), and antimicrobial phytoalexins are all derived from polyketides synthesized by the members of the CHS superfamily, elucidating the properties of these enzymes and discovering how the genes encoding them are regulated have been important within the applied sciences of agriculture and horticulture. Thus, it is not surprising that the *CHS* promoter is one of the best studied plant promoters. Spatio-temporal patterns of synthesis vary substantially among the flavonoids and many *cis*-acting elements have been identified that play roles in the regulation of the seed plant *CHS* genes. For examples, the G-box (with the consensus sequence of CACGTG) is involved in UV-light induced expression of *CHS* genes in spermatophytes (Staiger et al. 1989). The H-box (ACCTAC) is known to play a role in feed-forward activation of the *CHS* promoter by *p*-coumarate, an upstream intermediate in the phenylpropanoid-pathway (Fig. 1a; Loake et al. 1992). The W-box (CTGACC/T) is a *cis*-acting element involved in pathogen- or elicitor-induction of *CHS* and other genes involved in plant defense (Trognitz et al. 2002). Some *CHS* promoters contain multiple *cis*-elements. *Arabidopsis thaliana CHS* (M20308) and *Pueraria lobata CHS* (D63855) genes each contain a G-box and an H-box in their promoters, while the *Phaseolus vulgaris CHS15* (X59469) promoter has a G-box and two H-box elements. Several R2R3-Myb and bHLH type *trans*-acting factors have been found to bind to these *cis*-acting elements and regulate *CHS* gene expression (Jin and Martin 1999; Qian et al. 2007; Cominelli et al. 2008). Thus, differential interactions among *cis*-acting elements and *trans*-acting factors lead to the complex expression patterns of the *CHS* genes documented in seed plants (Hartmann et al. 2005).

Consistent with the central role of plant flavonoids in UV protection, *CHS* is highly regulated by light (Wingender et al. 1989). The *P. patens CHS* multigene family provides an opportunity to study the regulation of *CHS* genes in a non-seed plant, about which little is known. We have studied in particular the expression of representative genes of the *P. patens CHS* superfamily in response to light. In an experiment using a proven light responsive *POR* gene as a positive control, *PpCHS01* and *PpCHS2c* exhibited light responsiveness (Fig. 6). By scrutinizing upstream DNA sequences, we found that both *PpCHS01* and *PpCHS2c* have the G-box core sequence (ACGT) at −411 and −471 bp upstream from their ATG codons, respectively. Since the same sequence is also found in the 5′-UTRs of *PpCHS1a, PpCHS2b.1, PpCHS10* and *PpCHS11*, the functional significance of the ACGT sequence in the light responsiveness of *PpCHS01* and *PpCHS2c* is unclear. Further study should provide insight into the evolution of the roles of *cis*-acting elements in the regulation of *CHS* genes. It is still notable that *PpCHS2a* and *PpCHS2c* showed different responses to light, since their 5′-UTR sequences are identical up to −249 bp and must had been diverged very recently. This provides an example of differentially expressed duplicated genes (subfunctionalization).

Other notable putative *cis*-elements are also found in *P. patens CHS* genes. A W-box-like sequence (TTGACC) is found in the 5′-UTRs of *PpCHS3.1* (at −935 bp from the start codon), *PpCHS3.2* and *PpCHS3.3* (−610), *PpCHS5.1* (−1,000 and −930), *PpCHS6* (−341), *PpCHS9* (−670), and *PpCHS10* (−661). Some of these genes were not responsive to light and we speculate that some of them may be involved in plant defense. Many genes with anther-specific expression possess, in their promoter regions, a unique sequence called an anther-box (Höfig et al. 2003). The 5′-UTR of *PpCHS10* contains an 18 bp sequence (TAGAGAATGCTTGAAATC at −1149) that appears to be homologous to the anther box. This provides further support for the contention that *PpCHS10* is the moss ortholog of seed plant *ASCLs*.

In conclusion, the combination of amenability to genetic manipulation and its diversity of *CHS* superfamily genes makes Physcomitrella an exciting and unique model system for further studies of *CHS* superfamily gene functions, regulation and evolution.

Acknowledgments This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the University of Regina. E. I. B. and C. C. C. are recipients of NSERC postgraduate scholarships (PGS-D and CGS-M, respectively).

References

Abe I, Sano Y, Takahashi Y, Noguchi H (2003) Site-directed mutagenesis of benzalacetone synthase. The role of the Phe215 in plant type III polyketide synthases. J Biol Chem 278:25218–25226

Abe I, Oguro S, Utsumi Y, Sano Y, Noguchi H (2005) Engineered biosynthesis of plant polyketides: chain length control in an octaketide-producing plant type III polyketide synthase. J Am Chem Soc 127:12709–12716

Ageez A, Kazama Y, Sugiyama R, Kawano S (2005) Male-fertility genes expressed in male flower buds of *Silene latifolia* include homologs of anther-specific genes. Genes Genet Syst 80:403–413

Akiyama T, Shibuya M, Liu HM, Ebizuka Y (1999) *p*-Coumaroyltriacetic acid synthase, a new homologue of chalcone synthase, from *Hydrangea macrophylla* var. *thunbergii*. Eur J Biochem 263:834–839

Ashton NW, Schulze A, Hall P, Bandurski RS (1985) Estimation of indole-3-acetic acid in gametophytes of the moss, *Physcomitrella patens*. Planta 164:142–144

Atanassov I, Russinova E, Antonov L, Atanassov A (1998) Expression of an anther-specific chalcone synthase-like gene is

correlated with uninucleate microspore development in *Nicotiana sylvestris*. Plant Mol Biol 38:1169–1178

Austin MB, Noel JP (2003) The chalcone synthase superfamily of type III polyketide synthases. Nat Prod Rep 20:79–110

Austin MB, Bowman ME, Ferrer JL, Schroder J, Noel JP (2004) An aldol switch discovered in stilbene synthases mediates cyclization specificity of type III polyketide synthases. Chem Biol 11:1179–1194

Basile A, Sorbo S, Lopez-Saez JA, Cobianchi RC (2003) Effects of seven pure flavonoids from mosses on germination and growth of *Tortula muralis* HEDW (Bryophyta) and *Raphanus sativus* L (Magnoliophyta). Phytochemistry 62:1145–1151

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

Brinkmeier E, Geiger H, Zinsmeister HD (1999) Biflavonoids and 4, 2′-epoxy-3-phenylcoumarins from the moss Mnium hornum. Phytochemistry 52:297–302

Cominelli E, Gusmaroli G, Allegra D, Galbiati M, Wade HK, Jenkins GI, Tonelli C (2008) Expression analysis of anthocyanin regulatory genes in response to different light qualities in *Arabidopsis thaliana*. J Plant Physiol 165:886–894

Dipp NJ, Newman AJ (1989) Evidence that introns arose at proto-splice site. EMBO J 8:2015–2021

Domínguez E, Mercado JA, Quesada MA, Heredia A (1999) Pollen sporopollenin: degradation and structural elucidation. Sex Plant Reprod 12:171–178

Durbin ML, McCaig B, Clegg MT (2000) Molecular evolution of the chalcone synthase multigene family in the morning glory genome. Plant Mol Biol 42:79–92

Eckermann S, Schröder G, Schmidt J, Strack D, Edrada RA, Helariutta Y, Elomaa P, Kotilainen M, Kilpeläinen I, Proksch P, Teeri TH, Schröder J (1998) New pathway to polyketides in plants. Nature 396:387–390

Ferrer JL, Jez JM, Bowman ME, Dixon RA, Noel JP (1999) Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis. Nat Struct Biol 6:775–784

Fliegmann J, Schröder G, Schanz S, Britsch L, Schröder J (1992) Molecular analysis of chalcone and dihydropinosylvin synthase from Scots pine (*Pinus sylvestris*), and differential regulation of these and related enzyme activities in stressed plants. Plant Mol Biol 18:489–503

Frugoli JA, McPeek MA, Thomas TL, McClung CR (1998) Intron loss and gain during evolution of the catalase gene family in angiosperms. Genetics 149:355–365

Fujita Y (1996) Protochlorophyllide reduction: a key step in the greening of plants. Plant Cell Physiol 37:411–421

Fukuma K, Neuls ED, Ryberg JM, Suh D-Y, Sankawa U (2007) Mutational analysis of conserved outer sphere arginine residues of chalcone synthase. J Biochem 142:731–739

Funa N, Ozawa H, Hirata A, Horinouchi S (2006) Phenolic lipid synthesis by type III polyketide synthases is essential for cyst formation in *Azotobacter vinelandii*. Proc Natl Acad Sci USA 103:6356–6361

Funa N, Awakawa T, Horinouchi S (2007) Pentaketide resorcylic acid synthesis by type III polyketide synthase from *Neurospora crassa*. J Biol Chem 282:14476–14481

Geiger H, Markham KR (1992) Campylopusaurone, an auronoflavanone biflavonoid from the mosses *Campylopus clavatus* and *Campylopus holomitrium*. Phytochemistry 31:4325–4328

Gross F, Luniak N, Perlova O, Gaitatzis N, Jenke-Kodama H, Gerth K, Gottschalk D, Dittmann E, Muller R (2006) Bacterial type III polyketide synthases: phylogenetic analysis and potential for the production of novel secondary metabolites by heterologous expression in pseudomonads. Arch Microbiol 185:28–38

Häger KP, Müller B, Wind C, Erbach S, Fischer H (1996) Evolution of legumin genes: loss of an ancestral intron at the beginning of angiosperm diversification. FEBS Lett 387:94–98

Han Y-Y, Ming F, Wang W, Wang J-W, Ye M-M, Shen D-L (2006) Molecular evolution and functional specialization of chalcone synthase superfamily from *Phalaenopsis* Orchid. Genetica 128:429–438

Harashima S, Takano H, Ono K, Takio S (2004) Chalcone synthase-like gene in the liverwort, *Marchantia paleacea* var. *diptera*. Plant Cell Rep 23:167–173

Hartmann U, Sagasser M, Mehrtens F, Stracke R, Weisshaar B (2005) Differential combinatorial interactions of *cis*-acting elements recognized by R2R3-MYB, BZIP, and BHLH factors control light-responsive and tissue-specific activation of phenylpropanoid biosynthesis genes. Plant Mol Biol 57:155–171

Höfig KP, Moyle RL, Putterill J, Walter C (2003) Expression analysis of four *Pinus radiata* male cone promoters in the heterogeneous host *Arabidopsis*. Planta 217:858–867

Iwashina T (2000) The structure and distribution of the flavonoids in plants. J Plant Res 113:287–299

Jez JM, Noel JP (2000) Mechanism of chalcone synthase. p*K*a of the catalytic cysteine and the role of the conserved histidine in a plant polyketide synthase. J Biol Chem 275:39640–39646

Jez JM, Austin MB, Ferrer J, Bowman ME, Schröder J, Noel JP (2000a) Structural control of polyketide formation in plant-specific polyketide synthases. Chem Biol 7:919–930

Jez JM, Ferrer JL, Bowman ME, Dixon RA, Noel JP (2000b) Dissection of malonyl-coenzyme a decarboxylation from polyketide formation in the reaction mechanism of a plant polyketide synthase. Biochemistry 39:890–902

Jez JM, Bowman ME, Noel JP (2002) Expanding the biosynthetic repertoire of plant type III polyketide synthases by altering starter molecule specificity. Proc Natl Acad Sci USA 99:5319–5324

Jiang C, Schommer CK, Kim SY, Suh D-Y (2006) Cloning and characterization of chalcone synthase from the moss, *Physcomitrella patens*. Phytochemistry 67:2531–2540

Jiang C, Kim SY, Suh D-Y (2008) Divergent evolution of the thiolase superfamily and chalcone synthase family. Mol Phylogenet Evol 49:691–701

Jin H, Martin C (1999) Multifunctionality and diversity within the plant MYB-gene family. Plant Mol Biol 41:577–585

Joshi CP, Zhou H, Huang X, Chiang VL (1997) Context sequences of translation initiation codon in plants. Plant Mol Biol 35:993–1001

Kamisugi Y, Cuming AC, Cove DJ (2005) Parameters determining the efficiency of gene targeting in the moss *Physcomitrella patens*. Nucleic Acids Res 33:e173

Kamisugi Y, Schlink K, Rensing SA, Schween G, von Stackelberg M, Cuming AC, Reski R, Cove DJ (2006) The mechanism of gene targeting in *Physcomitrella patens*: homologous recombination, concatenation and multiple integration. Nucleic Acids Res 34:6205–6214

Koes RE, Spelt CE, Mol JNM (1989) The chalcone synthase multigene family of Petunia hybrida (V30): differential, light-regulated expression during flower development and UV light induction. Plant Mol Biol 12:213–225

Kumar S, Tamura K, Nei M (2004) *MEGA3*: Intergrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 5:150–163

Lang D, Eisinger J, Reski R, Rensing S (2005) Representation and high-quality annotation of the *Physcomitrella patens* transcriptome demonstrates a high proportion of proteins involved in metabolism in mosses. Plant Biol 7:238–250

Lanz T, Tropf S, Marner F-J, Schröder J, Schröder G (1991) The role of cysteines in polyketide synthases. Site-directed mutagenesis

of resveratrol and chalcone synthases, two key enzymes in different plant-specific pathways. J Biol Chem 266:9971–9976

Liu B, Falkenstein-Paul H, Schmidt W, Beerhues L (2003) Benzophenone synthase and chalcone synthase from *Hypericum androsaemum* cell cultures: cDNA cloning, functional expression, and site-directed mutagenesis of two polyketide synthases. Plant J 34:847–855

Liu B, Raeth T, Beuerle T, Beerhues L (2007) Biphenyl synthase, a novel type III polyketide synthase. Planta 225:1495–1503

Loake GJ, Faktor O, Lamb CJ, Dixon RA (1992) Combination of H-box [CCTACC(N)7CT] and G-box (CACGTG) *cis* elements is necessary for feed-forward stimulation of a chalcone synthase promoter by the phenylpropanoid-pathway intermediate *p*-coumaric acid. Proc Natl Acad Sci USA 89:9230–9234

Long M, Rosenberg C (2000) Testing the "proto-splice sites" model of intron origin: evidence from analysis of intron phase correlations. Mol Biol Evol 17:1789–1796

Lütcke HA, Chow KC, Mickel FS, Moss KA, Kern HF, Scheele GA (1987) Selection of AUG codons differs in plants and animals. EMBO J 6:43–48

Ma L-Q, Pang X-B, Shen H-Y, Pu GB, Wang HH, Lei CY, Wang H, Li GF, Liu BY, Ye HC (2009) A novel type III polyketide synthase encoded by a three-intron gene from *Polygonum cuspidatum*. Planta 229:457–469

Markham KR (1988) Distribution of flavonoids in the lower plants and its evolutionary significance. In: Harborne JB (ed) The flavonoids. Chapman and Hall, London, pp 427–468

Mizuuchi Y, Shimokawa Y, Wanibuchi K, Noguchi H, Abe I (2008) Structure function analysis of novel type III polyketide synthases from *Arabidopsis thaliana*. Biol Pharm Bull 31:2205–2210

Morita H, Kondo S, Oguro S, Noguchi H, Sugio S, Abe I, Kohno T (2007) Structural insight into chain-length control and product specificity of pentaketide chromone synthase from *Aloe arborescens*. Chem Biol 14:359–369

Nishiyama T, Fujita T, Shin-I T, Seki M, Nishide H, Uchiyama I, Kamiya A, Carninci P, Hayashizaki Y, Shinozaki K, Kohara Y, Hasebe M (2003) Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: implication for land plant evolution. Proc Natl Acad Sci USA 100:8007–8012

Ober D (2005) Seeing double: gene duplication and diversification in plant secondary metabolism. Trends Plant Sci 10:444–449

Paniego NB, Zuurbier KW, Fung SY, van der Heijden R, Scheffer JJ, Verpoorte R (1999) Phlorisovalerophenone synthase, a novel polyketide synthase from hop (*Humulus lupulus* L.) cones. Eur J Biochem 262:612–616

Qian W, Tan G, Liu H, He S, Gao Y, An C (2007) Identification of a bHLH-type G-box binding factor and its regulation activity with G-box and Box I elements of the *PsCHS1* promoter. Plant Cell Rep 26:85–93

Quatrano RS, McDaniel SF, Khandelwal A, Perroud PF, Cove DJ (2007) *Physcomitrella patens*: mosses enter the genomic age. Curr Opin Plant Biol 10:182–189

Rensing SA, Fritzowsky D, Lang D, Reski R (2005) Protein encoding genes in an ancient plant: analysis of codon usage, retained genes and splice sites in a moss, *Physcomitrella patens*. BMC Genomics 6:43

Rensing SA, Ick J, Fawcett JA, Lang D, Zimmer A, Van de Peer Y, Reski R (2007) An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. BMC Evol Biol 7:130–139

Rensing SA, Lang D, Zimmer AD et al (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science 319:64–69

Ronquist F, Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574

Sawyer SA (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538

Schröder J (1997) A family of plant-specific polyketide synthases: facts and predictions. Trends Plant Sci 2:373–378

Seshime Y, Juvvadi PR, Fujii I, Kitamoto K (2005) Discovery of a novel superfamily of type III polyketide synthases in *Aspergillus oryzae*. Biochem Biophys Res Commun 331:253–260

Sommer H, Saedler H (1986) Structure of the chalcone synthase gene of *Antirrhinum majus*. Mol Gen Genet 202:429–434

Spalding JB, Lammers PJ (2004) BLAST Filter and GraphAlign: rule-based formation and analysis of sets of related DNA and protein sequences. Nucleic Acids Res 32:W26–W32

Stafford HA (1991) Flavonoid evolution: an enzymic approach. Plant Physiol 96:680–685

Staiger D, Kaulen H, Schell J (1989) A CACGTG motif of the *Antirrhinum majus* chalcone synthase promoter is recognized by an evolutionarily conserved nuclear protein. Proc Natl Acad Sci USA 86:6930–6934

Suh D-Y, Fukuma K, Kagami J, Yamazaki Y, Shibuya M, Ebizuka Y, Sankawa U (2000a) Identification of amino acid residues important in the cyclization reactions of chalcone and stilbene synthases. Biochem J 350:229–235

Suh D-Y, Kagami J, Fukuma K, Sankawa U (2000b) Evidence for catalytic cysteine-histidine dyad in chalcone synthase. Biochem Biophys Res Commun 275:725–730

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

Trapp SC, Croteau RB (2001) Genomic organization of plant terpene synthases and molecular evolutionary implications. Genetics 158:811–832

Trognitz F, Manosalva P, Gysin R, Niñio-Liu D, Simon R, del Herrera MR, Trognitz B, Ghislain M, Nelson R (2002) Plant defense genes associated with quantitative resistance to potato late blight in *Solanum phureja* × dihaploid *S. tuberosum* hybrids. Mol Plant Microbe Interact 15:587–597

Wingender R, Röhrig H, Höricke C, Wing D, Schell J (1989) Differential regulation of soybean chalcone synthase genes in plant defence, symbiosis and upon environmental stimuli. Mol Gen Genet 218:315–322

Wu S, O'Leary SJ, Gleddie S, Eudes F, Laroche A, Robert LS (2008) A chalcone synthase-like gene is highly expressed in the tapetum of both wheat (*Triticum aestivum* L.) and triticale (×*Triticosecale* Wittmack). Plant Cell Rep 27:1441–1449

Yamazaki Y, Suh D-Y, Sitthithaworn W, Ishiguro K, Kobayashi Y, Shibuya M, Ebizuka Y, Sankawa U (2001) Diverse chalcone synthase superfamily enzymes from the most primitive vascular plant, *Psilotum nudum*. Planta 214:75–84

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298