Divergent evolution of oxidosqualene cyclases in plants

Zheyong Xue1, Lixin Duan1, Dan Liu1, Jie Guo2, Song Ge2, Jo Dicks3, Paul ÔMáille4, Anne Osbourn4 and Xiaoquan Qi1

1Key Laboratory of Plant Molecular Physiology, Institute of Botany, Chinese Academy of Sciences, Nanxincun 20, Fragrant Hill, Beijing 100093, China; 2State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Nanxincun 20, Fragrant Hill, Beijing 100093, China; 3Department of Computational and Systems Biology, John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK; 4Department of Metabolic Biology, John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK

Author for correspondence:
Xiaoquan Qi
Tel: +86 10 62836671
Email: xqi@ibcas.ac.cn

Received: 16 September 2011
Accepted: 30 October 2011


Key words: metabolic diversity, oxidosqualene cyclase (OSC), sterols, tandem duplication, triterpenes.

Summary

• Triterpenes are one of the largest classes of plant metabolites and have important functions. A diverse array of triterpenoid skeletons are synthesized via the isoprenoid pathway by enzymatic cyclization of 2,3-oxidosqualene. The genomes of the lower plants Chlamydomonas reinhardtii and moss (Physcomitrella patens) contain just one oxidosqualene cyclase (OSC) gene (for sterol biosynthesis), whereas the genomes of higher plants contain nine to 16 OSC genes.
• Here we carry out functional analysis of rice OSCs and rigorous phylogenetic analysis of 96 OSCs from higher plants, including Arabidopsis thaliana, Oryza sativa, Sorghum bicolor and Brachypodium distachyon.
• The functional analysis identified an amino acid sequence for isoarborinol synthase (OsIAS) (encoded by Os11g35710/OsOSC1) in rice. Our phylogenetic analysis suggests that expansion of OSC members in higher plants has occurred mainly through tandem duplication followed by positive selection and diversifying evolution, and consolidated the previous suggestion that dicot triterpene synthases have been derived from an ancestral lanosterol synthase instead of directly from their cycloartenol synthases.
• The phylogenetic trees are consistent with the reaction mechanisms of the protosteryl and dammarenyl cations which parent a wide variety of triterpene skeletal types, allowing us to predict the functions of the uncharacterized OSCs.

Introduction

Triterpenes are one of the most diverse groups of plant metabolites, and nearly 200 distinct skeletons have been identified (Xu et al., 2004). Glycosylated triterpenes (saponins) have a diverse range of properties, including beneficial or detrimental effects on human health, antinutritional effects, sweetness and bitterness (Haralampidis et al., 2002; Augustin et al., 2011; Osbourn et al., 2011). Triterpenes, like sterols, are synthesized via the 30-carbon intermediate 2,3-oxidosqualene, which is cyclized by members of the oxidosqualene cyclase (OSC) family (Phillips et al., 2006; Abe, 2007). The first plant OSC to be cloned was Arabidopsis thaliana cycloartenol synthase (CAS1). This OSC was cloned using a heterologous expression strategy in which an A. thaliana cDNA library was introduced into yeast and the resulting transformants screened for the ability to convert oxidosqualene to cycloartenol (Corey et al., 1993). These experiments were facilitated by the use of a yeast mutant that was unable to synthesize lanosterol (LS, the fungal cyclization product of 2,3-oxidosqualene) and so accumulated 2,3-oxidosqualene. Although cycloartenol is the primary route to sterol synthesis in plants, it has recently been found that A. thaliana also possesses a LS that contributes to phytosterol biosynthesis (Kolesnikova et al., 2006; Suzuki et al., 2006; Ohyama et al., 2009). The other 11 members of the A. thaliana OSC gene family produce a diverse array of different triterpene skeletons (over 40 in total) (Phillips et al., 2006; Abe, 2007; Morlacchi et al., 2009). Thus a remarkable amount of chemical diversity is derived from a single substrate 2,3-oxidosqualene. Over the last 16 yr, OSCs have been characterized from a diverse range of plant species. The 13 A. thaliana OSCs and their major cyclization products are summarized in Table 1.

Metabolic diversification may originate through the generation of new enzymes by gene duplication, mutation and fixation, and/or by extending (or switching) the function of existing genes/enzymes (Pichersky & Gang, 2000). Gene duplication and subsequent recruitment of the duplicates for establishment of new functions (neofunctionalization) comprise a major mechanism of pathway evolution (Ober, 2005). For example, type II chalcone isomerase (CHI) enzymes which catalyze the formation of 5-deoxyflavanone most probably originated from tandem duplication of type I CHI genes during legume evolution (Shimada et al., 2003). In another case, retrotransposon-mediated duplication of CYP98A3, a cytochrome P450 (CYP450) gene involved in lignin monomer biosynthesis, led to the realization of
<table>
<thead>
<tr>
<th>Gene</th>
<th>Enzyme</th>
<th>Major product</th>
<th>Structure</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>At1g66960/LUP5</td>
<td>LUP5</td>
<td>Tirucalla-7,21-dien-3β-ol</td>
<td><img src="image1.png" alt="Structure" /></td>
<td>Ebizuka et al., 2003</td>
</tr>
<tr>
<td>At1g78500/PEN6</td>
<td>PEN6</td>
<td>Bauerenol</td>
<td><img src="image2.png" alt="Structure" /></td>
<td>Ebizuka et al., 2003</td>
</tr>
<tr>
<td>At1g78950</td>
<td>LUP4</td>
<td>β-amyrin</td>
<td><img src="image3.png" alt="Structure" /></td>
<td>Shibuya et al., 2009;</td>
</tr>
<tr>
<td>At1g78955/CAMS1</td>
<td>LUP3</td>
<td>Camelliol C</td>
<td><img src="image4.png" alt="Structure" /></td>
<td>Kolesnikova et al., 2007b</td>
</tr>
<tr>
<td>At1g78960/LUP2</td>
<td>LUP2</td>
<td>β-amyrin</td>
<td><img src="image5.png" alt="Structure" /></td>
<td>Kushiro et al., 2000; Husselstein-Muller et al., 2001;</td>
</tr>
<tr>
<td>At1g78970/LUP1</td>
<td>LUP1</td>
<td>Lupeol, 3β,20-dihydroxyupane</td>
<td><img src="image6.png" alt="Structure" /></td>
<td>Herrera et al., 1998;</td>
</tr>
<tr>
<td>At2g07050/CAS</td>
<td>CAS1</td>
<td>Cycloartenol</td>
<td><img src="image7.png" alt="Structure" /></td>
<td>Corey et al., 1993</td>
</tr>
<tr>
<td>At3g45130/LSS1</td>
<td>LSS1</td>
<td>Lanosterol</td>
<td><img src="image8.png" alt="Structure" /></td>
<td>Kolesnikova et al., 2006;</td>
</tr>
<tr>
<td>At4g15340</td>
<td>PEN1</td>
<td>(3S,13R)-malabarica-17,21-dien-3,14-diol (arabidiol)</td>
<td><img src="image9.png" alt="Structure" /></td>
<td>Xiang et al., 2006; Kolesnikova et al., 2007a</td>
</tr>
<tr>
<td>At4g15370/BARS1/PEN2</td>
<td>PEN2</td>
<td>Baruol</td>
<td><img src="image10.png" alt="Structure" /></td>
<td>Lodeiro et al., 2007</td>
</tr>
</tbody>
</table>
a novel phenolic pathway in Brassicaceae (Matsuno et al., 2009). Families of genes for enzymes implicated in plant secondary metabolism (e.g. cytochrome P450s, glycosyltransferases, acyltransferases, prenyltransferases) have commonly expanded, and the different members have acquired new functions by shifting or broadening substrate and/or product specificity (Vogt & Jones, 2000; Suzuki et al., 2004; Matsuno et al., 2009).

The previous analysis of the complete rice (Oryza sativa L. ssp. japonica cv Nipponbare) genome sequence identified 12 predicted OSC genes (Inagaki et al., 2011). One of these (Os02g04710/OsOSC2) encodes cycloartenol synthase (CS), while a further two (Os11g08569/OsOSC7 and Os11g18194/OsOSC8) have been shown to synthesize the triterpenes, parkeol and achilleol B, respectively, in Saccharomyces cerevisiae GIL77 (Ito et al., 2011). In addition, automated whole-genome annotation of the Sorghum bicolor and Brachypodium distachyon genomes (Paterson et al., 2009) indicate a number of OSC genes of unknown function in these species. However, the genomes of the lower plants Chlamydomonas reinhardtii and moss (Physcomitrella patens) each contain only one predicted OSC gene which is likely to be required for sterol biosynthesis (Merchant et al., 2007; Desmond & Gribaldo, 2009). It is generally believed that plant OSC genes involved in triterpene biosynthesis are derived directly or indirectly from an ancient CS gene required for essential plant sterol biosynthesis (Sawai et al., 2006). Lanoster synthases have recently been identified in several dicots, for example, A. thaliana, Penax ginseng (Kolesnikova et al., 2006; Suzuki et al., 2006) and Lotus japonicus (Sawai et al., 2006). Their biological significance is not fully understood, but in A. thaliana LS may perform a minor role in sterol biosynthesis (Ohyama et al., 2009). It has been proposed that plant LSs are likely to have diverged from the ancestral CS, based on an analysis of a limited number of plant OSCs (Sawai et al., 2006). Phillips et al. (2006) divided the plant OSCs into two groups based on the nature of their presumed catalytic intermediates, the protosteryl and dammarenyl cations. Both cations originate from the same tetracyclization reaction mechanism (initial cyclization forms a 6-6-5 tricycle, followed by ring expansion and D-ring annihilations) (Corey et al., 1995; Corey & Cheng, 1996; Jenson & Jorgensen, 1997; Hess, 2002), while starting from different folds of the 2,3-oxidosqualene substrate (Fig. 1a). The protosteryl and dammarenyl cations are centrally important, as these intermediates are the parents of a wide variety of triterpene skeletal types (Fig. 1b). The resulting cations, in turn, possess distinct stereochemistry and ring configurations. For example, the protosteryl cation adopts the chair-boat-chair (C-B-C) configuration and is an intermediate leading to cycloartenol, lanosterol, parkeol and cucurbitadienol tetracyclic triterpene skeletons (6-6-6-5). Isoarborinol is an unusual pentacyclic triterpene (6-6-6-6-5) clearly derived from an additional D-ring expansion of the protosteryl cation, based on the C-B-C configuration. Most pentacyclic triterpene skeletons, however, are derived from the dammarenyl cation by D-ring expansion to form lupeol or further E-ring expansion to form β-amyrin (Xu et al., 2004).

Despite these efforts, the origin and the evolution of OSCs in plants, especially the variable triterpene cyclases, are largely unclear. In order to address this, we have assembled and analyzed predicted/characterized OSC sequences from plants for which there is a well-annotated genome sequence (O. sativa, S. bicolor, B. distachyon and A. thaliana) and for functionally characterized OSCs from other plant species and have carried out a comprehensive analysis of phylogeny, genome-wide gene duplication events and codon substitutions in order to reconstruct the probable expansion and functional diversification of the OSC family in higher plants. We have further carried out functional analysis of rice OSCs and have discovered a new enzyme, isoarborinol synthase. Our analyses provide new insights into the likely origin and evolutionary basis for metabolic diversity in plant triterpenes.

Materials and Methods

Gene annotation

Two databases, Phytozome Version 6.0 (http://www.phytozome.net) and the BrachyBlast portal (http://www.brachypodium.org), were
searched by blastn using sequences of *AcCS1* and *AsbAS1* from *Avena strigosa* (Haralampidis *et al.*, 2001) to identify *OSC* genes for *Sorghum bicolor* (L.) and *Brachypodium distachyon* (L.), respectively. Annotation of the 12 predicted rice *OSC* genes was based on our previous analysis of the rice genome (Inagaki *et al.*, 2011).

Where limited transcript data were available, intron-exon patterns of the *S. bicolor* and *B. distachyon* genes were predicted using the National Center for Biotechnology Information (NCBI) tblastn tool. Manual annotation was performed for some mis-annotated genes. *OSC* genes from other species were downloaded from

---

**Fig. 1** Carbocation intermediates and products of 2,3-oxidosqualene tetracyclization by triterpene cyclases. (a) Substrate folding directs the cyclization of 2,3-oxidosqualene. The substrate 2,3-oxidosqualene adopts distinct folding patterns that, when directly cyclized by oxidosqualene cyclase (OSC) enzymes, produce stereochemically distinct cation products. In the case of tetracyclization, these reactions produce the protosteryl and dammarenyl cations. (b) A single substrate, 2,3-oxidosqualene (2,3-OS) is cyclized by OSC enzymes into numerous triterpene skeletal types with varying numbers of rings and stereochemistry. The tetracyclization reaction mechanism is prominent and gives rise to the protosteryl and dammarenyl cations, which in turn parents the formation of other cations through further rearrangements, most notably further ring expansions to produce a spectrum of derived products.
NCBI’s GenBank database according to their gene names or by blasting the homologous gene sequences.

**Transcript analysis**

The expression patterns of the rice OSC genes were determined by reverse transcription-polymerase chain reaction (RT-PCR) analysis. The TRIZol reagent (Invitrogen) was used according to the manufacturer’s instructions to extract total RNAs of shoots, roots, and panicles of rice (*O. sativa* L. ssp. *japonica*) cv Zhonghua11. For each sample, 2 μg RNA were used to synthesize the first strand of cDNA by using a SuperScript III reverse transcriptase (Invitrogen) according to the manufacturer’s instructions. About 1/50 of the first-strand cDNA generated was used as a template for PCR in a reaction volume of 20 μl with the ExTaq DNA polymerase (Takara, Dalian, Liaoning, China). PCR was performed with the following cycling profile: 94°C for 2 min, 2530 cycles at 94°C for 30 s, 55°C for 30 s, 72°C for 30 s, and 72°C for 10 min. Five microliters of the PCR product was separated in a 1% agarose gel and stained with ethidium bromide for visualization. The rice *Actin1* gene (Yamanouchi et al., 2002) was used as an internal control for RT-PCR. For each *OSC* gene, 25 or 30 cycles were used for PCR, depending on the expression levels of different genes. All RT-PCRs were carried out three times independently in separate experiments with different reverse-transcribed templates.

**Functional analysis of rice OSCs in yeast**

The coding sequence of each *OsOSC* gene was obtained from different tissues of Zhonghua11. The amplified products were cloned into pGEM-T easy vector (Promega) and sequenced from both ends, and were subcloned into the expression vector pPICZA (Invitrogen) to place the OsOSC open reading frame (ORF) under the control of the methanol-inducible promoter, 5′-AOX (pPICZAOSOSCs). *Pichia pastoris* wildtype strain X33 was transformed with pPICZAOOSCs and pPICZA using the electroporation according to the manufacturer’s instructions. X33s harboring *OsOSC* genes were grown at 30°C in minimal glycerol medium (MGY, 1.34% yeast nitrogen base, 1% glycerol, 4 × 10⁻⁵% biotin) to OD₆₀₀ = 2~6. The cells were collected by centrifugation, resuspended in minimal methanol medium (MM,1.34% yeast nitrogen base, 4 × 10⁻⁵% biotin, 0.5% methanol) to OD₆₀₀ = 1.0 and incubated at 30°C for 72 h, adding methanol every 24 h to maintain its concentration at 0.5%. Cells were finally collected and every 25 ml MM medium disrupted with 2 ml 20% KOH/50% EtOH (1/1, v/v). The refluxed products were extracted twice with 2 ml hexane, and combined with both hexane solutions to obtain the crude extract. The extracts were either directly derivatized using N-Methyl-N-trimethylsilyl trifluoroacetamide (MSTFA) and analyzed by GC-MS as described in our previous study (Qi et al., 2006) or further purified by thin layer chromatography (TLC) (20 × 20 cm, silica gel, 0.5 mm; Merck, Darmstadt, Germany). TLC plates were developed using a sandwich technique and ethoxyethane as the mobile phase, then stained with primuline and viewed under UV light. Bands for compounds of interest were removed from the plates, extracted with CHCl₃, filtered, and used directly for NMR. ¹H- and ¹³C-NMR (600 Hz) were measured in CDCl₃ with trimethylsilylate as an internal standard.

**Metabolite extraction from plants and gas chromatography/time-of-flight mass spectrometry (GC/TOF-MS) analysis**

Metabolite was extracted from lyophilized rice leaves (100–500 mg) using a method described previously (Field & Osbourn, 2008). The crude products and 13 μg standards of parkeol and isoarborinol were derivatized with MSTFA, then analyzed on a LECO Pegasus® IV (GC/TOF-MS). The GC was fitted with an Agilent DB-5MS column (29.5 m × 250 μm internal diameter, 0.25 μm film). The inlet, transfer line, and ion source temperatures were set at 290, 280, and 200°C, respectively, and an oven temperature program from 80°C (2 min) to 270°C (2 min) at 20°C min⁻¹, followed by 270°C to 320°C (5 min) at 5°C min⁻¹ was used. The flow rate of the carrier gas (helium) was 1.5 ml min⁻¹. Splitless injections (1 μl) were used and mass spectral data in the range m/z 70–550 were acquired.

**cDNA cloning and transformation of rice and Arabidopsis thaliana**

The parkeol and isoarborinol synthase (*Os11g08569 and Os11g35710*) coding sequences were amplified from Zhonghua11 leaf cDNA with Phusion polymerase (New England Biolabs Inc., Beverly, Massachusetts, USA) and cloned into pDONR221 (Invitrogen). The entry clone was confirmed by sequencing, recombined (with Invitrogen LR GATEWAY recombinase) into the plant expression vector pH7WG2D under the control of the 35S promoter. The resulting construct was transferred into *Agrobacterium tumefaciens* (strain EHA105) and used to transform rice calli induced from mature embryos of rice cv Zhonghua11. Transgenic calli were selected on Murahige and Skoog (MS) medium containing 50 mg l⁻¹ hygromycin B (Roche). Hygromycin-resistant plants regenerated from calli were transplanted into soil and grown in a glasshouse or in local paddy fields. For *Arabidopsis thaliana* (L.) Heynh, *Agrobacterium tumefaciens* (strain EHA105) harboring *Os11g35710* in pB2WG7 was dapped on the flowers of wildtype *A. thaliana* (Col-0) and 41 *Os11g35710*-overexpressing transgenic plants were obtained.

**Phylogenetic tree construction and molecular evolution analyses**

Multiple alignment of OSC protein sequences was performed with Muscle 3.6 (Edgar, 2004) and refined manually. The protein matrix was transformed into a cDNA matrix with the aa2DNA script (https://homes.bio.psu.edu/people/faculty/nei/Software/aa2dna/aa2dna.zip). Maximum likelihood (ML) phylogenetic trees were constructed from the cDNA alignment with the software RAxML (Windows version 7.0.4, Stamatakis, 2006) using the GTR + Γ + I substitution model and with the
C. rei*hard*ii CS as an outgroup. We performed 100 ML runs
and 500 bootstraps, and bootstrapped trees were mapped on
to the ML run exhibiting the highest likelihood. To confirm the
topology of the phylogenetic tree, a Bayesian phylogenetic tree
was also estimated under the GTR + Γ + I substitution model.
The MrBayes3.1.2 software (Windows version 3.1.2; Ronquist &
Huelsenbeck, 2003) was used, with 10 000 000 generations
performed with a sampling every 10 000 generations by the Mar-
kov chain Monte Carlo method.

For molecular evolution analysis, genes from the CS and LS
groups were separated and used to construct CS-derived and LS-
derived phylogenetic trees, respectively, using the program
PHYML (Guindon & Gascuel, 2003) under the GTR + Γ + I
nucleotide substitution models. To evaluate variation in selection
pressures over these two OSC phylogenies, the free ratio model
of CODEML within the PAML4 software package (Yang, 2007)
was used to estimate lineage specific rates of the nonsynony-
mous : synonymous substitution (dn/ds) ratio, ω. To detect
whether positive selection had acted at some amino acid sites
within specific lineages, a branch-site analysis was also performed
comparing the nearly neutral model (M1a) with the Model A
branches tested.

Transposable elements prediction

We used RepeatMasker (Smit, AFA, Hubley, R. RepeatModeler
Open-1.0: 2008–2010, http://www.repeatmasker.org) to anno-
tate DNA repeats for rice and other species. Genes along with c. 10 kb intergenic sequence
were cloned and expressed in P. pastoris. Metabolite analysis showed that
Os11g08569/OsOSC7- and Os11g35710/OsOSC11-containing yeast cells produced and accumulated different additional com-
ounds, respectively, compared with the empty vector (negative
control) (Fig. 3). However, yeast transformants containing the
other five OSC genes did not produce detectable additional com-
ounds. Approx. 2 mg of the compounds produced by Os11g08569/OsOSC7 and Os11g35710/OsOSC11 were separated and purified. GC-MS analyses indicated that Os11g08569/
OsOSC7 produces parkeol (Fig. 3c,e) in P. pastoris X33, while
Os11g35710/OsOSC11 produces isorarbinol (Fig. 3d,f). The
structures of purified parkeol and isorarbinol from cell extracts
of P. pastoris were confirmed by NMR and by comparison with
mass spectral fragmentation profiles (Supporting Information, Fig. S1a,b) (Hanisch et al., 2003; Pearson et al., 2003).

The NMR data for parkeol (synthesized by Os11g08569) is as
follows: 1H-NMR(CDC13, 600M): δ: 0.65, 0.75, 0.82, 0.88,
0.99, 1.04, 1.60, 1.68(3H, CH3), 3.20(1H, dd, J = 4.2,
12.0 Hz, 3x-H), 5.09(1H, m, H-24), 5.22 (1H, m, H-11); 13C-NMR(CDC13,125M): 14.64(C-18), 15.65(C-30),
17.67(C-26), 18.57(C-21), 18.91(C-28), 21.38(C-6), 22.26(C-
19), 25.04(C-23), 25.73(C-27), 27.83(C-2), 28.09(C-7),
28.24(C-16), 28.28(C-29), 38.87(C-15), 35.14(C-12),
35.67(C-20), 36.12(C-1), 37.28(C-22), 39.11(C-4), 39.39(C-
28), 41.82(C-8), 44.31(C-13), 47.16(C-14), 50.78(C-17),
52.51(C-5), 78.92(C-3), 114.98(C-11), 125.12(C-24),
128.24(C-16), 130.92(C-25), 148.53(C-9). And the NMR data for isorarbinol
(synthesized by Os11g35710) is as follows: 1H-NMR(CDC13,600M): δ: 0.72, 0.73, 0.75, 0.77, 0.84, 0.88, 0.98,
1.03(3H, CH3), 3.20(1H, dd, J = 4.2, 12.0Hz, 3x-H),
5.23(1H, d, J = 5.4 Hz, H-11); 13C-NMR(CDC13,125M): 14.00(C-28),
15.28(C-27), 15.62(C-23), 17.02(C-26),
...
Os11g08569/OsOSC7 is expressed at low levels in mature rice leaves, while Os11g35710/OsOSC11 is expressed strongly in mature leaves (Fig. 2). To establish whether Os11g08569/OsOSC7 and Os11g35710/OsOSC11 produce parkeol and isoarborinol, respectively, in plants, transgenic rice plants overexpressing each of these two OSCs were generated. GC/TOF-MS analysis of extracts from mature leaves of the wildtype rice cv Zhonghua11 and transgenic rice plants overexpressing Os11g08569/OsOSC7 revealed the presence of parkeol in transgenic plants and abundant isoarborinol in the wildtype (Figs 4a, S1f). Since none of the 13 Arabidopsis OSCs make isoarborinol, we also tested the function of Os11g35710/OsOSC11 by overexpression in Arabidopsis. In comparison with wildtype plants, the transgenic plants produce an additional compound (Fig. 4b), which was confirmed as isoarborinol by GC/TOF-MS analysis (Fig. S1g). Thus Pichia expression experiments together with these in planta tests of function allowed us to conclude that Os11g08569/OsOSC7 is indeed a parkeol synthase (OsPS1) and that Os11g35710/OsOSC11 encodes isoarborinol synthase (OsIAS1).

Expansion and functional diversification of the OSC gene family in higher plants

A single OSC gene predicted to encode CS was identified from each of the genomes of the following lower plant species: Chlamydomonas reinhardtii (green alga), Physcomitrella patens ssp. patens (moss), Adiantum capillus-veneris (fern) and Polypodiodes niponica (fern). There are 12 predicted OSC genes in the rice genome (Fig S2a). Manual annotation of OSC genes based on the whole-genome sequences of S. bicolor and B. distachyon (angiosperms) revealed that there are 16 and nine predicted OSC genes in these two genomes, respectively (Table S1, Fig. S2b,c). These data and the fact that there are 13 functional OSC genes in the Arabidopsis genome clearly demonstrate that there has been a large increase in OSC gene members in higher plant genomes.

To predict the functions of the expanded OSC members in these three Poaceae species, 53 OSCs with known functions, 13 functionally defined Arabidopsis OSCs (Table 1), plus predicted full-length OSCs from rice (11 OSCs), S. bicolor (12 OSCs) and B. distachyon (seven OSCs), and five CAS members from lower plants were assembled for phylogenetic analysis. The Chlamydomonas sequence was used as the outgroup. A ML phylogenetic tree containing 101 sequences was obtained using the GTR + I substitution model (Fig. 5). A Bayesian phylogenetic tree exhibited a very similar topological structure to this ML tree (Fig. S3). This phylogenetic analysis allowed us to classify the 96 OSCs from higher plants into 10 groups (groups I–X) based on their product specificity and higher rank phylogeny (dicots vs monocots) (Fig. 5). In dicots, OSCs were grouped into CSs (I), cucurbitadienol synthases (II), LSs (VIII) and a pentacyclic triterpene synthase-like group (X). Five more groups of OSCs were defined in monocots in addition to the CS group (III) (Fig. 5). One of these was defined as being of unknown function (IV), while another contained parkeol synthases (V), including the rice parkeol synthase characterized in this study and in Ito et al. (2011). A third group (VI) contains the...
Most OSC members from the Poaceae species belong to a pentacyclic triterpene synthase-like group (VII) and are predicted to produce variable triterpene skeletons. Interestingly, a group of unknown function (IX) that contains four monocot sequences (one from each of the four species analyzed here), is closely related to the dicot pentacyclic triterpene synthase-like group (X) and LS group (VIII) (Fig. 5).

The role of tandem duplication in the expansion of the OSC gene family

The availability of the whole-genome sequences of *A. thaliana* and the three Poaceae species, rice, *S. bicolor* and *B. distachyon*, provides an opportunity to investigate the evolutionary history of the OSC gene family in plants and to predict the duplication
Fig. 5 Phylogenetic analysis of the coding sequence of the oxidosqualene cyclase (OSC) family from higher and lower plants. A maximum likelihood (ML) tree was constructed with RAxML using the GTR + I + I model with 100 ML runs and 500 bootstraps. Broken lines are used to separate monocots and dicots. D1–D11 indicate the gene duplication events. Black squares indicate the tandem duplications, the triangle indicates the segmental or whole-genome duplication, and the dots indicate unknown types of duplications. The bold red and blue lines show the evolutionary paths of triterpene synthases in dicots and monocots, respectively. ALSL, ancestral lanosterol synthase-like; ACS, ancestral cycloartenol synthase.
events that occurred duringOSC gene family evolution. One duplication event (D1) for which there exists high bootstrap support (Fig. 5) must have occurred before the divergence of dicots and monocots, which occurred c. 140 million yr ago (mya; Moore et al., 2007; Jiao et al., 2011), so giving rise to two ancient OSC genes, the ancestral cycloartenol synthase (ACS) gene and the ancestral LS-like (ALSL) gene. These ancestral genes then provided the foundation for the two distinct groups, D1-1 and D1-2 (Fig. 5). This duplication event may have been the result of whole-genome duplication, tandem gene duplication or other types of duplication. We were unable to distinguish between these possibilities. After the divergence of monocots from dicots, the ACS gene was duplicated many times, leading to the expansion of OSC genes in monocot species, whereas only one duplication event is evident in Betulaceae species of dicots (Fig. 5, D9-2). Another ancient duplication event (D2, Fig. 5) is proposed for the ALSL gene before the divergence of monocots from dicots. The original LS gene was maintained in many dicot species, while the duplicated gene is likely to have been the origin of most of the dicot triterpene synthase genes. The function of the genes within monocot group IX, closely related to the dicot LSs (VIII), is currently unclear. Our experiments in which we expressed rice Os08g12730 and 6 additional rice OSC genes in S. cerevisiae suggest that these seven rice OSCs are unable to rescue the Gil77 strain, which is deficient in lanosterol synthesis (Fig. S4). However, we cannot eliminate the possibility that the Os08g12730-containing group (IX) contains LSs. Our phylogenetic analysis indicates that it is also possible that the original LS gene was lost in monocots and that the current group is derived from a duplicated gene. The dicot triterpene synthases, including lupeol, dammarenediol and β-amyrin synthase, may have originated from the ALSL gene via three successive gene duplication events (D2, D10 and D11, Fig. 5). These data suggest that the dicot triterpene synthases are not directly derived from ACS, but rather arose via duplication of ALSL, as previously postulated by Sawai et al. (2006).

It is interesting to note that of the 13 A. thaliana OSC genes, the 11 triterpene synthase genes are grouped into one functional group (X). Furthermore, 20 out of 36 Poaceae OSC genes were assigned either to the pentacyclic triterpene synthase-like group (VII) (Fig. 5), based on the characterized β-amyrin synthase from Avena species (Haralampidis et al., 2001; Qi et al., 2004), or to the rice isorborniol synthase group (VI) characterized in this study. These data indicate that a major expansion of the OSC gene family has occurred after the divergence of monocots and dicots. Considering together the gene family phylogeny (Fig. 5) with the genomic distributions of its constituent genes, some predictions can be made about key duplication events underpinning aspects of this expansion. For example, in A. thaliana, a tandem cluster on chromosome 1 containing four homologous OSC genes (with c. 85% similarity), At1g78950, At1g78955/CAMSI, At1g78960/LUP2 and At1g78970/LUP1, is likely to have arisen by three tandem duplication events (Fig. 5). Another tandem duplicate gene pair, Ar4g15340 and Ar4g15370, encoding arabidol synthase and baroul synthase, respectively (Xiang et al., 2006; Lodeiro et al., 2007), is located on A. thaliana chromosome 4 (Fig. 5). In monocots, syntenic genomic regions containing four rice, three B. distachyon and six S. bicolor genes (Fig. 6a) indicate three shared duplication events (D3, D5 and D6) plus three lineage-specific tandem duplications and up to eight gene losses whose lineage dependency is currently unclear (Fig. 6b). Indeed most triterpene synthase genes in the Poaceae family appear to have arisen from Cs genes by the D3 gene duplication event, which caused the divergence of the 20 triterpene synthase genes (D3-2) from the 12 Cs genes and other closely related genes that form group D3-1 (Fig. 5). The D3 duplication event is highly likely to have

Fig. 6 (a) Collinearity of duplicated blocks of a key oxidosqualene cyclase (OSC) region in rice, Sorghum bicolor and Brachypodium distachyon. The green and blue dots signify OSC genes involved in sterol and triterpene pathways, respectively. The yellow and black dots signify anchored genes and other genes in the relevant genome. (b) The likely sequence of duplication and loss events that led to the genomic distribution of genes seen in (a).
been a tandem duplication that occurred during the ancient Poaceae genome before the ρ whole-genome duplication (WGD), which was estimated to have occurred between 117 and 50 mya (Kellogg, 2001; Gaut, 2002; Yu et al., 2005; Lescot et al., 2008; Salse et al., 2008; Jiao et al., 2011). The subsequent D5 duplication event can also be seen to be a tandem duplication while the D6 event is most likely to be the ρ whole-genome duplication itself or a segmental duplication. Using the same strategy we are not currently able to define the duplication events for D4, D7 and D8. The genes derived from these duplication events were not included in segmental blocks and also were not clustered on the same chromosome regions (Figs S5, S6). The addition of future genome data, as they become available, may serve to define these events. However, in total 11 tandem duplication events and one whole-genome/segmental duplication event could be defined by our rigorous genome and phylogenetic analyses.

Transposition-based gene duplication has been proposed as one of the mechanisms of gene family expansion (Jiang et al., 2004; Hoen et al., 2006; Xiao et al., 2008; Elrouby & Bureau, 2010). Our transposable elements analysis in OSC gene-containing regions in the rice, S. bicolor and B. distachyon genomes have revealed that three classes/families of retrotransposable elements (LRT/Gypsy, LRT/copia and LINE/L1) and six classes/families of DNA transposable elements (DNA/Tourist, DNA/TcMar-stowaway, DNA/En-Spm, DNA/hAT-Ac, DNA/MuDR and SINE) have been distributed in the OSC gene regions of the three genomes (Tables S2, S3, S4). The DNA/MuDR, LRT/Gypsy and DNA/TcMar-Stowaway elements predominate with the high score weight among those elements. For example, a 7901 bp DNA fragment insertion in Os02g04750/60 is a Mutator-like element which could encode a transposase. However, our analysis revealed no evidence to indicate that any of the rice OSC genes were likely to have arisen by transposon-based duplication. Gene structure analysis (Fig. S2) further indicated that none of the OSC genes in Poaceae were likely to be transduplicates, which normally have reduced numbers of introns relative to the progenitor gene.

These results indicate that tandem duplication has contributed greatly to the expansion of the OSC gene family in the genomes of both dicots and monocots, while whole-genome duplication or segmental duplication has made only a limited contribution and no transposon-based duplicates have been discovered to date.

Positive selection drives one duplicate to evolve at accelerated rates to acquire a new function following gene duplication

Phylogenetic trees for the CS- (CS tree) and LS-derived (LS tree) groups were reconstructed separately (see Fig. 7a,b, respectively) for adaptive molecular evolutionary analysis of the plant OSCs using the PAML software. Likelihood ratio tests revealed that log-likelihood values (log_ L = -39 881.44 and -35 797.20 for CS and LS trees, respectively) under the free ratio model (M1) were significantly higher (P < 0.001) than those (log_ L = -40 070.62 and -35 919.33 for CS and LS trees, respectively) under the one ratio model (M0) in both groups (Table 2). These results indicate that the free ratio model (M1; where dN/dS ratios, ω, may vary between branches) fits both the CS- and LS-derived datasets better than the one ratio model (M0, where ω is fixed), suggesting that members of the OSC family experienced varied selection pressures during their expansion. Differential evolutionary rates were also observed in glutathione S-transferase gene family (Chi et al., 2011).

Indeed, a large variation in lineage-specific estimates of ω, as indicated in Fig. 7 for the duplication or functional groups, was observed among OSC family members. The average dN/dS ratios of the dicot CS genes (I) and monocot CS genes (II) were found to be 0.12 and 0.10, respectively (Fig. 7a), and for the dicot LS genes (VIII), lupeol synthase genes and monocot unknown-function genes (IX) to be 0.12, 0.13 and 0.12, respectively (Fig. 7b). These small variations and very low average dN/dS ratios in each group (see Fig. 7) reveal that the amino acid sequences of the CS, LS and lupeol synthase gene members and the monocot unknown-function gene members have been largely constrained by strong purifying selection. By contrast, the relatively higher and more variable average dN/dS ratios of the dicot pentacyclic triterpene synthase-like genes (X), including β-amyrin, lupeol and multifunction synthase genes (0.17 (0.05–0.64)) (Fig. 7b) and Poaceae predicted pentacyclic triterpene synthase genes (VII) (0.21 (0.12–0.52), parkeol synthase genes (V) (0.33 (0.24–0.39)) and unknown function group (IV) (0.20 (0.15–0.27)) (Fig. 7a) suggest that most triterpene synthase genes for both dicots and monocots may have been under more relaxed selective constraints.

The dN/dS ratios (ω) of the seven pairs of branches (Fig. 7a, a to g) of Poaceae triterpene synthase genes and four pairs of branches (Fig. 7b, h to k) of dicot triterpene synthase genes derived from duplication events (Fig. 7 marked with D) were estimated using branch-site models along with four other branches (Fig. 7a, l to o) leading to key extant genes (Table 3). Among the 26 branches that were analyzed, nine branches were under highly significant positive selection. Interestingly, significant positive selection is detected in only one of the two sister branches after gene duplication events in six cases (Fig. 7; branches a, d, e, h, j, k and with all significant branches marked with thick lines), indicating that one duplicate may have been free to acquire a new function while the other duplicate maintained the original function under purifying selection.

The functional evolution of plant OSC genes

Subsequent to the D3 tandem duplication event, the rice isoarborinol synthase gene (Os11g35710) can be seen to have evolved during a long period of relaxed selection (Fig. 7a; branches a and d). The oat (Avena strigosa) β-amyrin synthase gene has experienced two significant periods of relaxed selection (Fig. 7a, branches a and n), while the rice achilleol B synthase (Os11g18194) has experienced one significant period of relaxed selection (Fig. 7a; branches a) since the D3 event. The rice
Fig. 7 Phylogenetic trees of cycloartenol synthase (CS)-derived (a) and lanosterol synthase (LS)-derived (b) oxidosqualene cyclases (OSCs). The 10 thick branches (a, b, d, e, f, h, j, k, l, and n; see Table 3) indicate branches or genes evolving under positive selection with significant statistical support at $P < 0.01$. Average estimates of the nonsynonymous : synonymous substitution ($dN/dS$) ratio, $\omega$, are shown to the right of each function group. The conformation of products and the corresponding intermediate cations are shown on the right side of the figure. D1–D11 indicate the gene duplication events. Black squares indicate the tandem duplications, the triangle indicates the segmental or whole-genome duplication, and the dots indicate unknown types of duplications. [Correction added after online publication 25 January 2012: a new version of Fig. 7 is inserted here, to correct errors noted in the Early View version of this article.]
parkeol synthase gene (Os11g08569) experienced a period of relaxed selection after the D4 duplication (Fig. 7a; branches b and l). Clearly, all four triterpene synthases have been able to gain new functions as a consequence of exploiting periods of relaxed selection following duplication events.

Given the distribution of triterpene synthase activities across our phylogenetic trees, the dammarenyl-derived triterpene synthases arose early from the ALSL enzyme by the D2 duplication event before the divergence of dicots and monocots in the LS-tree (Fig. 7b), while appearing only more recently in monocot lineages after the D6 duplication in the CS tree (Fig. 7a).

Our results reveal that the parkeol synthase gene is more similar to the CS gene than are the isoarborinol synthase and β-amyrin synthase genes. This is consistent with the reaction mechanism where parkeol and cycloartenol derive from a common protosteryl cation, while isoarborinol and β-amyrin require additional ring expansion mechanisms (Fig. 1) (Xu et al., 2004). Therefore, we expect that uncharacterized OSCs will produce either protosteryl and dammarenyl-cation-derived triterpenes based on their phylogenetic lineages as indicated in Figs 5 and 7.

**Discussion**

The sterol pathways may originate from ancestral bacteria, as OSCs have been identified in several bacteria, for example, LS in proteobacterium (*Methylococcus capsulatus*); CS and LS in

---

**Table 2** Likelihood ratio test of evolutionary models for the cycloartenol synthase-derived group (CS tree) and the lanosterol synthase-derived group (LS tree)

<table>
<thead>
<tr>
<th>Lineages</th>
<th>Models</th>
<th>Log_e L</th>
<th>2ΔL^a</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS tree</td>
<td>M0 (one ratio)</td>
<td>−40 070.62</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>M1 (free ratio)</td>
<td>−39 881.44</td>
<td>378.36**</td>
</tr>
<tr>
<td>LS tree</td>
<td>M0(one ratio)</td>
<td>−35 919.33</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>M1(free ratio)</td>
<td>−35 797.20</td>
<td>244.26**</td>
</tr>
</tbody>
</table>

^a^2ΔL is twice the log-likelihood difference between models M1 and M0. **p** indicates the difference at the highly significant level of ^P_< 0.01.

**Table 3** Summary of statistics for detection of positive selection for cycloartenol synthase-derived group (CS tree) and lanosterol synthase-derived groups (LS tree)

<table>
<thead>
<tr>
<th>Lineages</th>
<th>Branches</th>
<th>Model A (branch-site)</th>
<th>Model A (branch-site)</th>
<th>M1 (free ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Log_e L</td>
<td>2ΔL^a</td>
<td>p^2^</td>
</tr>
<tr>
<td>CS tree</td>
<td>a</td>
<td>−39 570.13</td>
<td>71.96**</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>a, b</td>
<td>−39 606.11</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>−39 597.71</td>
<td>16.8**</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>c, c, d</td>
<td>−39 598.38</td>
<td>15.46</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>e</td>
<td>−39 605.30</td>
<td>1.62</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>f</td>
<td>−39 597.87</td>
<td>16.48**</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>g</td>
<td>−39 592.80</td>
<td>5.8</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>h</td>
<td>−39 599.29</td>
<td>13.64</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>i</td>
<td>−39 603.21</td>
<td>5.8</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>j</td>
<td>−39 592.90</td>
<td>26.62**</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>k</td>
<td>−39 599.76</td>
<td>34.7**</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>−39 600.18</td>
<td>11.88</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>−39 598.45</td>
<td>15.32**</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>o</td>
<td>−39 602.09</td>
<td>8.04</td>
<td>0.03</td>
</tr>
<tr>
<td>LS tree</td>
<td>h</td>
<td>−35 741.01</td>
<td>36.90**</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>i</td>
<td>−35 756.88</td>
<td>5.16</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>j</td>
<td>−35 725.99</td>
<td>66.94</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>k</td>
<td>−35 745.91</td>
<td>42.96</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>−35 732.95</td>
<td>53.02**</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>−35 737.98</td>
<td>27.1**</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>−35 757.01</td>
<td>4.90</td>
<td>0.01</td>
</tr>
</tbody>
</table>

^a^2ΔL is twice the log-likelihood difference between Ma and M1a, where under M1a (nearly neutral model) log_e L was estimated to be −39 606.11 for the CS tree and −35 799.46 for the LS tree.

^b^The proportion of sites evolving under positive selection.

^c^Nonsynonymous : synonymous substitution (dN/dS) ratio of site classes 2a and 2b.

^d^dN/dS ratio estimated under free ratio model (M1).

^e^2ΔL was not applied because of the infinite value of o^2^ or o.

^f^The dN/dS value was estimated to be 999.00.

**p** The difference at the highly significant level of ^P_< 0.01 (the Bonferroni correction was used, where ^P_< 0.01/18, 2ΔL > 14.99 and ^P_ < 0.01/8, 2ΔL > 13.36 for the CS tree and the LS tree, respectively).
myxobacterium (Stigmatella aurantiaca) and LS (parkeol) in
planctomycete (Gemmata obscuriglobus) (Bode et al., 2003; Pear-
son et al., 2003; Lamb et al., 2007; Nakano et al., 2007),
although hopane cyclase is the dominant form in most bacteria.
Recent comprehensive analysis (Fischer & Pearson, 2007) sug-
gested that hopanoid and steroid cyclases diverged from a com-
mon ancestor instead of the previous assumption that hopanoid
biosynthesis was an evolutionary predecessor to steroid biosyn-
thesis in the ancient life forms. Extensive phylogenetic analysis
based on 5288 putative triterpene cyclase homologues in the pub-
licly available databases revealed that a few sequences from above
three bacterial species grouped with a set of OSCs from eukary-
otic species, while a small group of sequences from seven fungal
species and a sequence from the fern Adiantum grouped with a
cluster of bacterial squalene cyclases, suggesting bidirectional lat-
eral gene transfer among the prokaryotes and eukaryotes (Frickey
& Kannenberg, 2009). However, our phylogenetic analysis
(Fig. 5) and analysis of gene structure of the OSC genes from the
four higher plant species with well-annotated genome sequence
(Fig. S2) do not give any evidence of lateral gene transfer from
prokaryotes.

Isoarborinol was first isolated from several families of higher
plants in the 1960s (e.g. Rutaceae Vorbrüggen et al., 1963;
Poaceae Nishimoto et al., 1968; Ohmoto & Ikuse, 1970). It was
also frequently identified in exceptional abundance in some
ancient immature and contemporary sediments which were dated
back to Permian or Triassic periods (299–200 mya) (e.g. Albrecht
& Ourisson, 1969; Hauke et al., 1995; Jaffé & Hausmann,
1995), proposing that isoarborinol and arborinol must originate
from microorganisms such as aerobic bacteria or algae (Hauke
et al., 1995) during early evolution. By re-analysis of numerous
sedimentary records of the hopanes, steranes and other triterpenes,
and the crystal structures and amino acid sequences of triterpene
cyclases using a combined phylogenetic and biochemical perspec-
tive, Fischer & Pearson (2007) suggested that malabaricanoids
would be the most ancient polycyclic triterpenoids, and hopanoid
and steroid cyclases diverged from a common ancestor.
Isoarborinol synthase was predicted to be one of the phylogenetic
intermediates between the primitive squalene-bacteriohopanoid
cyclase and the lanosterol/cycloartenol-producing epoxysqualene
cyclases (Ourisson et al., 1982; Fischer & Pearson, 2007). It was
generally believed that isoarborinols in the ancient sediments were
derived from as-yet-unknown microbial sources (Ourisson et al.,
1982; Fischer & Pearson, 2007), but until now isoarborinol
cyclase has not been reported in any microorganism. Here we have
identified an amino acid sequence encoding isoarborinol biosyn-
thesis from rice (Poaceae). Our phylogenetic analysis clearly
showed that the identified monocot isoarborinol synthase clade
(VI) (Fig. 5) was derived recently from monocot ACS through
independent convergent evolution in comparison with the pre-
sumed ancient isoarborinol synthase (Fischer & Pearson, 2007)
from microorganisms in the period 299–200 mya (Permian or
Triassic periods).

Our analysis suggests that OSCs from higher plants have arisen
from an ancient CS (Fig. 5). An increase in the number of
members of a gene family may be attributable to whole-genome
duplication events, small-scale segmental duplications, local tan-
dem duplications, single gene transposition-duplications, or com-
binations of these possibilities (Freeling, 2009). The phylogenetic
genome-wide duplication and codon substitution analyses in this
study showed that local tandem gene duplication has contributed
greatly to the expansion of the OSC gene family. This is in agree-
ment with the observation that gene families involved in the
biosynthesis of secondary metabolites tend to arise by gene dupli-
cation, forming tandem clusters within the plant genome (Ober,
2005). OSC genes have been lost in most of the species we ana-
yzed here after segmental duplication or whole-genome duplica-
tion. This finding is consistent with the high loss rate of
duplicates and the tendency for selective retention of only those
genes with high expression levels and more conserved functions
after whole-genome duplication in A. thaliana (Simillion et
al., 2002; Blanc et al., 2003; Wu & Qi, 2010). The preferential
retention of tandem repeats and the under-retention of segmental
duplicates or whole-genome duplicates within the OSC gene
family can best be explained by the dosage-sensitive relationship
in the gene balance hypothesis (Freeling, 2009). In brief, this
hypothesis presumes that after long-term evolution, ‘connected
genes’ of multi-component complexes (such as genes in the meta-
abolic pathways) in the present genomes have been in an optimum
balance state and changes of the individual genes in the complex
would display dosage sensitivity, resulting in out-of-balance
phenotypes which have disadvantages in fitness. OSC genes,
especially new tandemly duplicated triterpene synthase genes,
may be less well connected with other genes, so facilitating
exploitation of new functions.

Tandem duplication, which has been estimated to be the
source of 1435% of all duplicated genes in the plant genomes,
has contributed significantly to the expansion of plant gene
families (Arabidopsis Genome Initiative, 2000; Zhang &
Gaut, 2003; Rizzon et al., 2006; Paterson et al., 2009; Schnable
et al., 2009). Many previous studies (Parmiske et al., 1997;
Michelmore & Meyers, 1998; Lucht et al., 2002; Kovalchuk
et al., 2003; Leister, 2004; Shiu et al., 2004; Maere et al., 2005;
Mondragon-Palomino & Gaut, 2005; Rizzon et al., 2006) have
demonstrated that tandem duplication tends to be associated
with biotic and abiotic stresses. A recent study involving compari-
son of gene family expansion among four land plant species
(Arabidopsis, poplar, rice and the moss Physcomitrella patens)
revealed that gene families that have expanded via tandem dupli-
cation tend to be involved in responses to environmental stimuli,
while those that expanded via nontandem mechanisms tended to
have intracellular regulatory roles (Hanada et al., 2008). Indeed,
the α-amyrin synthase (AshA1) catalyzes the first step in a
biosynthetic pathway for the synthesis of defense compounds in
oat (Papadopoulou et al., 1999; Haralampidis et al., 2001; Qi
et al., 2004). Although catalytic functions of all 13 Arabidopsis
OSC s have been characterized by yeast expression experiments
(Morlacchi et al., 2009) and two of them (At5g48010/THAS
and At5g42600/MRN1) have been analyzed in planta (Field &
Osborn, 2008; Field et al., 2011, in press), the biological
roles of the 12 OSCs (except for CS) are still unclear. Future
experiments involving functional analysis of rice OSC-derived

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust
www.newphytologist.com
pathways will address the biological roles of the rice parkeol synthase, isoarborinol synthase and other OSCs. The OSC genes in higher plants have experienced repeated cycles of gene duplications and divergence in a lineage-specific expansion pattern (Bishop et al., 2000). The codon substitution analysis based on the branch-site model in this study has revealed that OSC genes are likely to multiply through tandem gene duplication, with positive selection driving one duplicate to evolve preferentially via nonsynonymous mutation to acquire a new function and with the other tending to retain its original function after gene duplication. Interestingly, dicot triterpene synthases were derived from an ALSL enzyme instead of directly from their CSs. LS (Figs 5, 7b, group VIII) in higher plants evolved before the divergence of monocots and dicots, and still maintains its function in dicots, indicating that LS has played an important role in dicots. Indeed, biosynthesis of phytosterols in dicots (e.g. sitosterol, campesterol and stigmasterol) occurs mainly through cycloartenol, further supplemented by the lanosterol-derived sterol pathway (Ohyama et al., 2009). Monocot-specific OSCs for lanosterol biosynthesis have not been identified and whether another sterol pathway exists in monocots has yet to be determined.

Acknowledgements

We thank Hongyan Shan for technical assistance, Hongzhi Kong and Manyan Long for valuable comments. This work was supported by funding from the 973 Program (2007CB108800) and NNSF (30900114, 30670167 & 30990242) of China to Z.X., L.D. Z.W., J.G., S.G., and X.Q., and from BBSRC (UK) to J.D., P.O. and A.O.

References


Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 The mass spectrometry (MS) data of trimethylsilyl (TMS) ether for parkeol and isoarborinol.

Fig. S2 Exon-intron structures of oxidosqualene cyclase (OSC) genes from rice.

Fig. S3 Bayesian phylogenetic tree constructed under the GTR + I + I substitution model using the MrBayes3.1.2 software.

Fig. S4 Complementary experiment of rice oxidosqualene cyclase (OSC) genes in Saccharomyces cerevisiae Gil77.

Fig. S5 Segmentally duplicated blocks of oxidosqualene cyclases (OSC) regions in Arabidopsis thaliana and rice.

Fig. S6 Segmentally duplicated blocks of oxidosqualene cyclases (OSC) regions in Brachypodium distachyon and Sorghum bicolor.

Table S1 List of plant oxidosqualene cyclases (OSCs) involved in this paper.

Table S2 Transposable elements in oxidosqualene cyclase (OSC) gene regions in the rice genome.

Table S3 Transposable elements in oxidosqualene cyclase (OSC) gene regions in the Sorghum bicolor genome.

Table S4 Transposable elements in oxidosqualene cyclase (OSC) gene regions in the Brachypodium distachyon genome.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the New Phytologist Central Office.