Divergent evolution of oxidosqualene cyclases in plants

Zheyong Xue[1], Lixin Duan[1], Dan Liu[1], Jie Guo[2], Song Ge[2], Jo Dicks[3], Paul ÓMáille[4], Anne Osbourn[4] and Xiaoquan Qi[1]

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

New Phytologist (2012) **193**: 1022–1038
doi: 10.1111/j.1469-8137.2011.03997.x

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/OsOSC11*) 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.

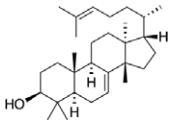
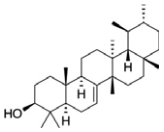
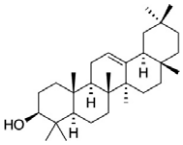
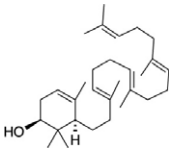
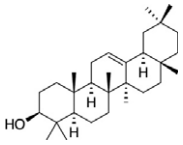
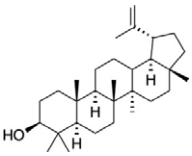
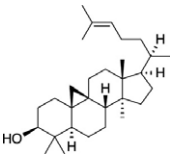
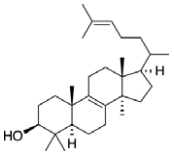
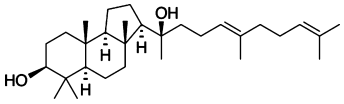
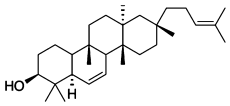
Introduction

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

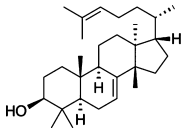
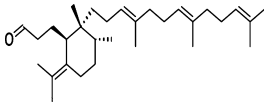
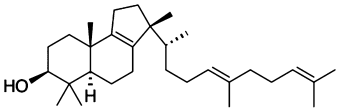
© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

Table 1 List of *Arabidopsis thaliana* oxidosqualene cyclases (OSCs) and their catalyzed products

Gene	Enzyme	Major product	Structure	Reference
At1g66960/LUP5	LUP5	Tirucalla-7,21-dien-3β-ol		Ebizuka *et al.*, 2003
At1g78500/PEN6	PEN6	Bauerenol		Ebizuka *et al.*, 2003
At1g78950	LUP4	β-amyrin		Shibuya *et al.*, 2009;
At1g78955/CAMS1	LUP3	Camelliol C		Kolesnikova *et al.*, 2007b
At1g78960/LUP2	LUP2	β-amyrin		Kushiro *et al.*, 2000 Husselstein-Muller *et al.*, 2001;
At1g78970/LUP1	LUP1	Lupeol, 3β,20-dihydroxylupane		Herrera *et al.*, 1998;
At2g07050/CAS	CAS1	Cycloartenol		Corey *et al.*, 1993
At3g45130/LSS1	LSS1	Lanosterol		Kolesnikova *et al.*, 2006;
At4g15340	PEN1	(3S,13R)-malabarica-17,21-dien-3, 14-diol (arabidiol)		Xiang *et al.*, 2006; Kolesnikova *et al.*, 2007a
At4g15370/BARS1/PEN2	PEN2	Baruol		Lodeiro *et al.*, 2007

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

Table 1 (Continued)

Gene	Enzyme	Major product	Structure	Reference
At5g36150/PEN3	PEN3	Tirucalla-7,24-dien-3-ol		Morlacchi *et al.*, 2009
At5g42600/MRN1	PEN5	Marneral		Xiong *et al.*, 2006
At5g48010/THA1	PEN4	Thalianol		Fazio *et al.*, 2004

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). Lanosterol synthases have recently been identified in several dicots, for example, *A. thaliana*, *Panax 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 annulations) (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

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

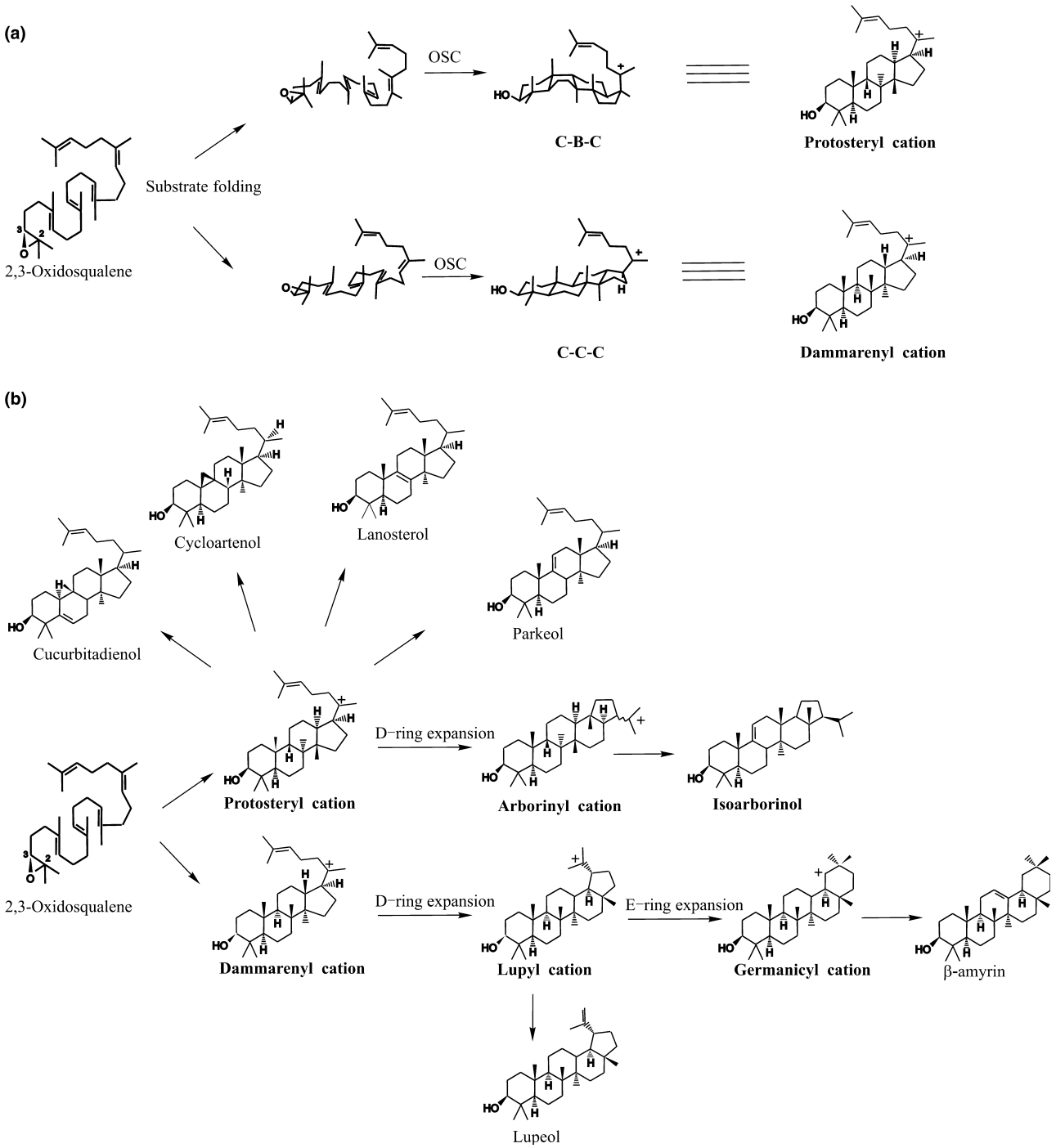


Fig. 1 Carbocation intermediates and products of 2,3-oxidosqualene tetracyclization by triterpene cyclases. (a) Substrate folding directs the cyclization of 2,3-oxidosqualene. The substrate 2,3-oxidosqualene adopts distinct folding patterns that, when directly cyclized by oxidosqualene cyclase (OSC) 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.

searched by blastn using sequences of *AsCS1* 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

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

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 pPICZAOsOSCs 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 OD_{600} = 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 OD_{600} = 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-trimethylsilyltrifluoroacetamide (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; Merch, 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 $CHCl_3$, filtered, and used directly for NMR. ^{1}H- and ^{13}C-NMR (600 Hz) were measured in $CDCl_3$ with trimethylsilate as an internal standard.

Metabolite extraction from plants and gas chromato graphy/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 Murashige 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 dipped 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

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

C. reinhardtii 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 Markov 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 nonsynonymous : 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 (Yang, 2007), to test the assumption that the foreground ω value of a specific branch was > 1 at these sites. The resulting likelihood ratio tests (LRTs) were performed at the 5% level in conjunction with a Bonferroni correction taking into account the number of branches tested.

Transposable elements prediction

We used RepeatMasker (Smit, AFA, Hubley, R. RepeatModeler Open-1.0; 2008–2010, http://www.repeatmasker.org) to annotate DNA repeats for rice and *S. bicolor* using the corresponding repeats databases, oryza_repeats.fa and sorghum_repeats.fa, respectively, as the references from The Institute for Genomic Research (TIGR) (ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats/). For *B. distachyon*, both of these repeat sequence sets were used as the references, as there was no species-specific repeat database available. Genes along with *c.* 10 kb intergenic sequence were used as input to the analysis.

Results

Identification of isoarborinol synthase and parkeol synthase in rice

For functional analysis of all rice OSCs, the heterologous expression strategy using *P. pastoris* yeast was applied in this study. *P. pastoris* yeast synthesizes ergosterol from 2,3-oxidosqualene via lanosterol and does not produce any other triterpenes. Heterologous expression of plant *OSC* genes in *P. pastoris* allows the expressed OSC enzymes to use endogenous 2,3-oxidosqualene to produce different triterpenes.

Transcript expression analysis by RT-PCR showed that seven out of the 12 predicted *OSCs* are expressed in different tissues of rice cv Zhonghua 11, a cultivar that is grown in the Beijing area (Fig. 2). The full-length cDNAs of these seven *OSCs* were cloned and expressed in *P. pastoris*. Metabolite analysis showed that Os11g08569/OsOSC7- and Os11g35710/OsOSC11-containg yeast cells produced and accumulated different additional compounds, respectively, compared with the empty vector (negative control) (Fig. 3). However, yeast transformants containing the other five *OSC* genes did not produce detectable additional compounds. 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 isoarborinol (Fig. 3d,f). The structures of purified parkeol and isoarborinol 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(CDCl3, 600M)δ: 0.65, 0.75, 0.82, 0.88, 0.99, 1.04, 1.60, 1.68(3H,8 × CH3), 3.20(1H, dd, J = 4.2, 12.0 Hz, 3α-H), 5.09(1H, m, H-24), 5.22 (1H, m, H-11); 13C-NMR(CDCl3,125M)δ: 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-10), 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), 130.92(C-25), 148.53(C-9). And the NMR data for isoarborinol (synthesized by Os11g35710) is as follows: 1H-NMR (CDCl3,600M)δ: 0.72, 0.73, 0.75, 0.77, 0.84, 0.88, 0.98, 1.03(3H,8 × CH3), 3.20(1H, dd, J = 4.2, 12.0Hz, 3α-H), 5.23(1H, d, J = 5.4 Hz, H-11); 13C-NMR(CDCl3,125M)δ: 14.00(C-28), 15.28(C-27), 15.62(C-23), 17.02(C-26),

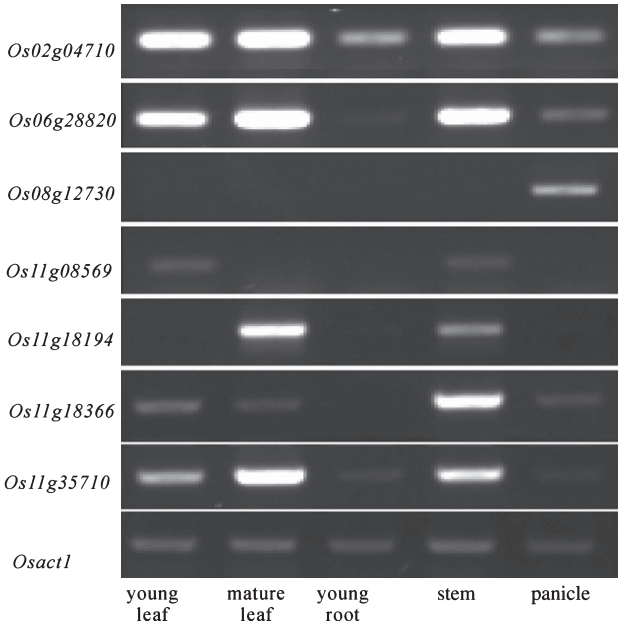


Fig. 2 Semiquantitative reverse transcription-polymerase chain reaction (RT-PCR) of rice oxidosqualene cyclase (*OSC*) genes from young leaf, mature leaf, young root, stem, and panicle of *Oryza sativa* ssp. *japonica* cv. Zhonghua11. The rice *Actin1* gene was used as the control.

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

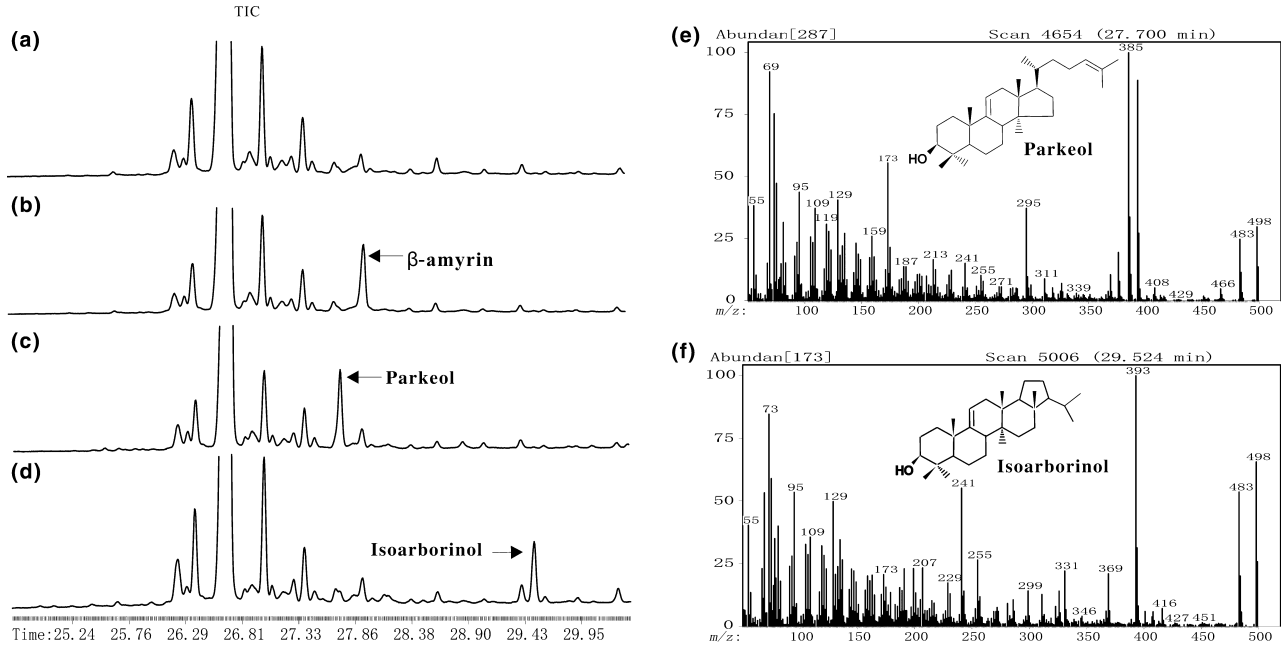


Fig. 3 Identification of rice parkeol and isoarborinol synthases by use of the yeast heterologous expression system. Gas chromatography chromatographs of the triterpene alcohol fraction of extracts of yeast cells transformed with empty vector or expression constructs (a) for AsbAS1 (b), Os11g08569 (c) or Os11g35710 (d) (TIC, total ion chromatogram). The mass spectrometry (MS) data of trimethylsilyl (TMS) ether for parkeol (e) at the retention time of 27.700 min in panel (c), and for isoarborinol (f) at the retention time of 29.524 min in panel (d).

20.17(C-19), 21.43(C-25), 22.13(C-29), 22.13(C-24), 22.70(C-7), 23.00(C-30), 26.68(C-6), 27.82(C-2), 28.22(C-20), 29.65(C-15), 30.77(C-22), 35.93(C-1), 36.01(C-12), 36.06(C-16), 36.78(C-13), 38.19(C-14), 39.07(C-10), 39.63(C-4), 40.97(C-8), 42.85(C-17), 52.08(C-18), 52.34(C-5), 59.65(C-21), 78.95(C-3), 114.32(C-11), 148.87(C-9).

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 *A. thaliana* OSCs make isoarborinol, we also tested the function of Os11g35710/OsOSC11 by overexpression in *A. thaliana*. 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: *C. 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 *A. thaliana* 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 *A. thaliana* 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 *C. reinhardtii* sequence was used as the outgroup. A ML phylogenetic tree containing 101 sequences was obtained using the GTR + Γ + 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

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

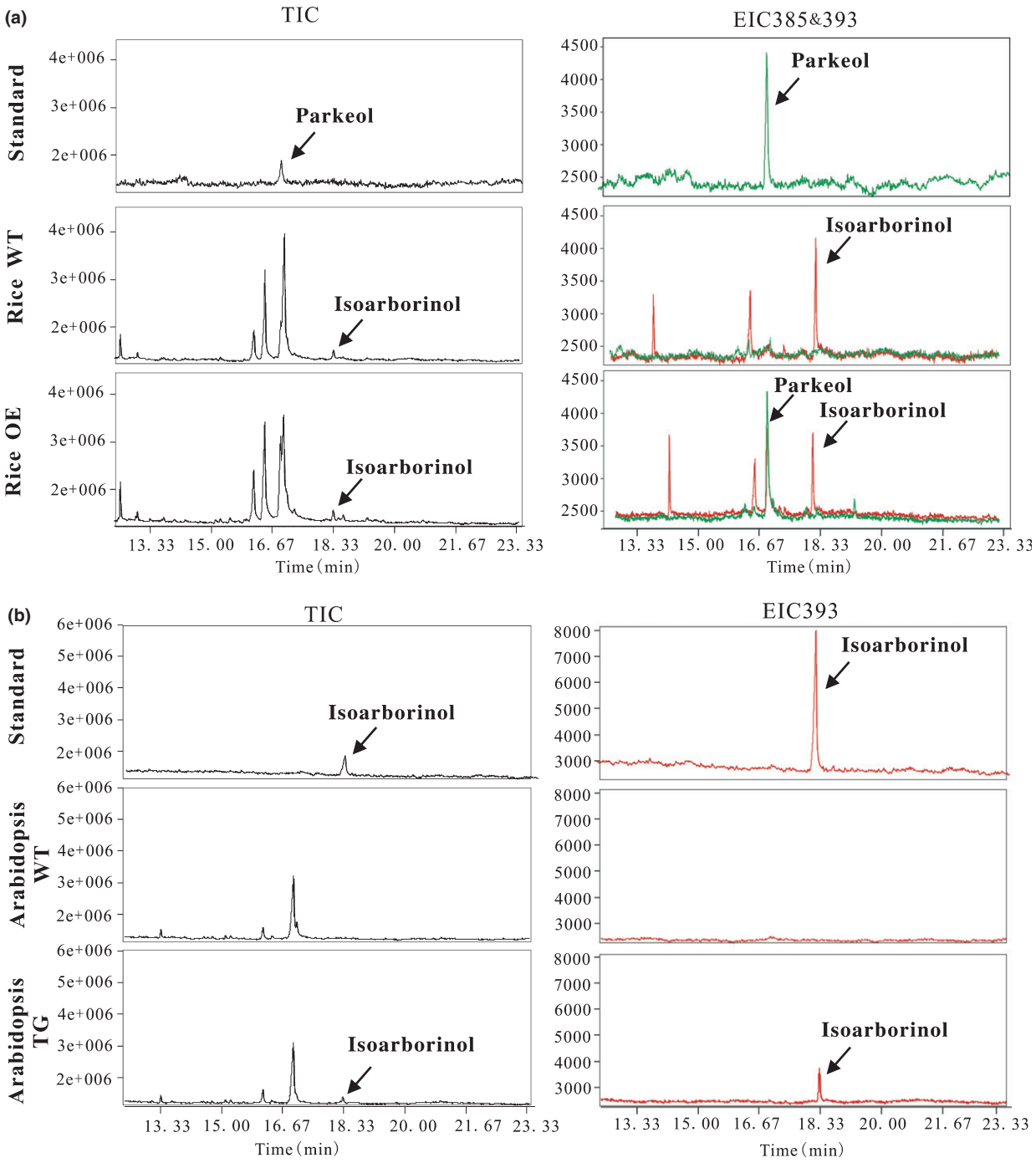


Fig. 4 Analysis of functions of rice parkeol and isoarborinol synthases in plants. (a) Gas chromatography (GC) chromatographs of the triterpene alcohol fraction of parkeol (standard), the extract of leaves of overexpressing Os11g08569 transgenic rice plants (Rice OE) and the wildtype rice plants (Rice WT) at the adult plant stage. (b) GC chromatographs of the triterpene alcohol fraction of isoarborinol (standard), the extract of leaves of overexpressing Os11g35710 transgenic *Arabidopsis thaliana* plants (Arabidopsis TG) and the wildtype *A. thaliana* plant (Arabidopsis WT). TIC, total ion chromatogram. EIC385&393, extracted ion chromatograms at *m/z* 385 (green) and 393 (red), respectively.

rice isoarborinol synthase defined in this study. 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

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

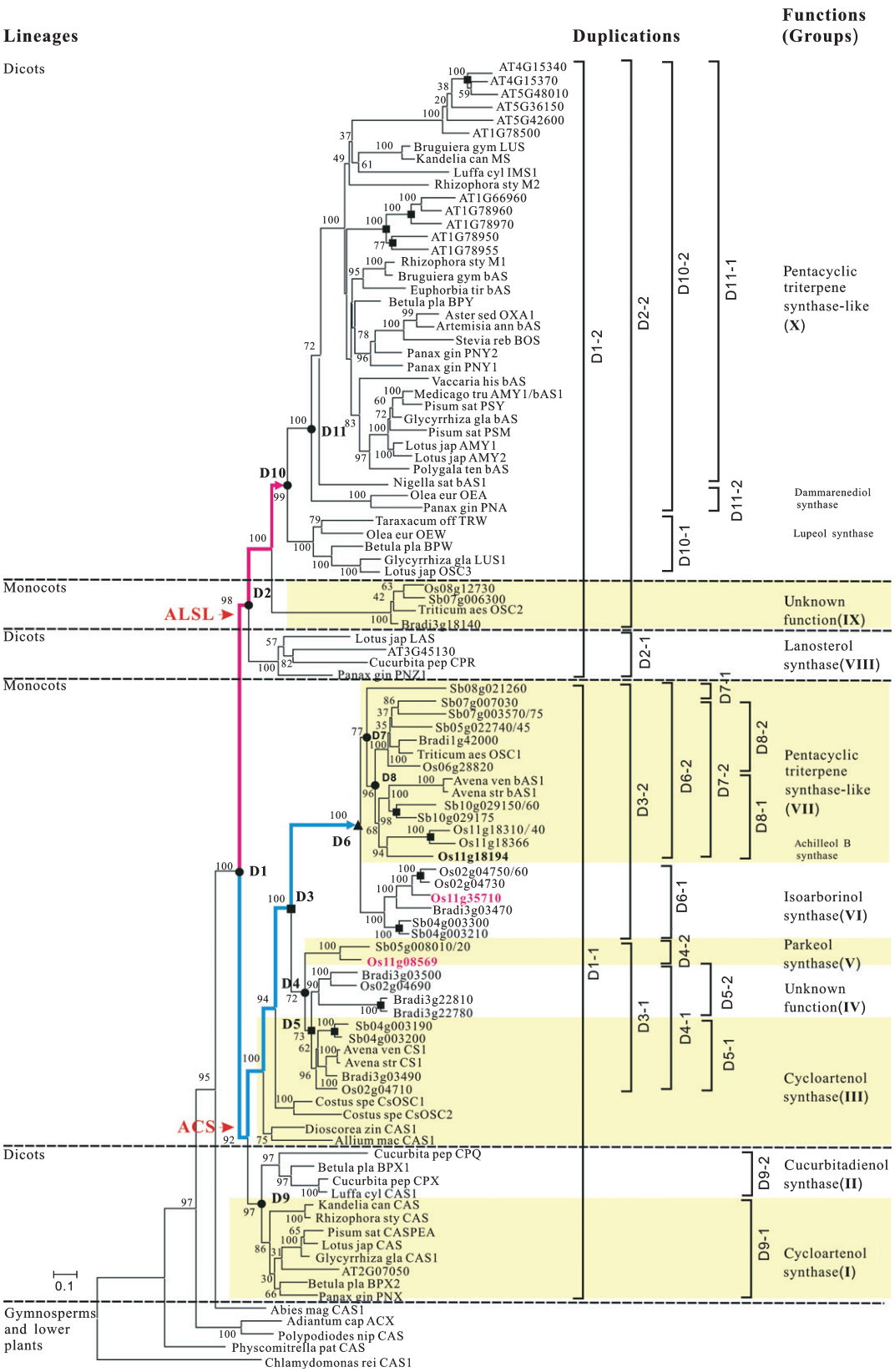


Fig. 5 Phylogenetic analysis of the coding sequence of the oxidosqualene cyclase (OSC) family from higher and lower plants. A maximum likelihood (ML) tree was constructed with RAxML using the GTR + Γ + I model with 100 ML runs and 500 bootstraps. Broken lines are used to separate monocots and dicots. 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.

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

events that occurred during *OSC* 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 isoarborinol 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/CAMS1*, *At1g78960/LUP2* and *At1g78970/LUP1*, is likely to have arisen by three tandem duplication events (Fig. 5). Another tandem duplicate gene pair, *At4g15340* and *At4g15370*, encoding arabidiol synthase and baruol 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

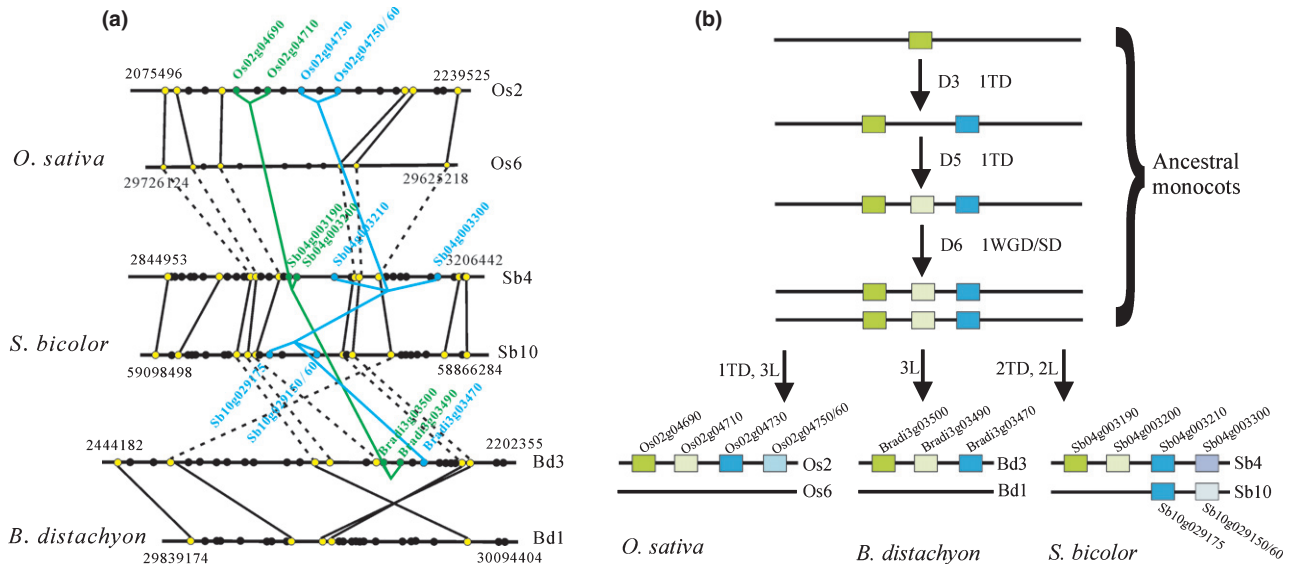


Fig. 6 (a) Collinearity of duplicated blocks of a key oxidosqualene cyclase (OSC) region in rice, *Sorghum bicolor* and *Brachypodium distachyon*. The green and blue dots signify *OSC* genes involved in sterol and triterpene pathways, respectively. The yellow and black dots signify anchored genes and 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).

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

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.

Transposon-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 retrotransponsable 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 ($\log_e L$ = −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 ($\log_e L$ = −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 multi-function 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 section. Interestingly, significant positive section is detected in only one of the two sister branches after gene duplication events in six cases (Fig. 7; branches *a*, d_s, *e*, *h*, j_s, *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_s). 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

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

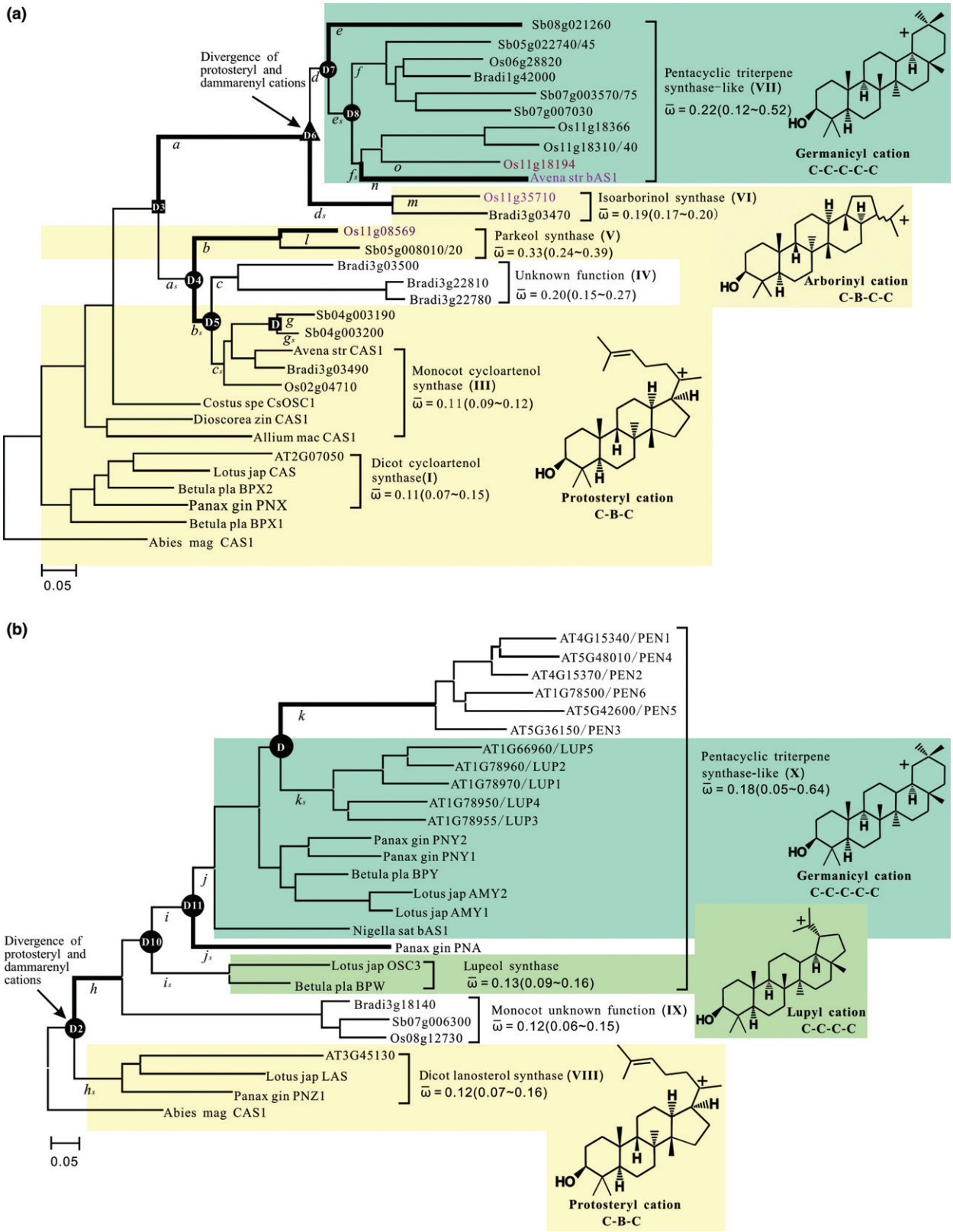


Fig. 7 Phylogenetic trees of cycloartenol synthase (CS)-derived (a) and lanosterol synthase (LS)-derived (b) oxidosqualene cyclases (OSCs). The 10 thick branches (a, b, b_s, d_s, e, h, j_s, k, l and n; see Table 3) indicate branches or genes evolving under positive selection with significant statistical support at $P < 0.01$. Average estimates of the nonsynonymous : synonymous substitution (dN/dS) ratio, ω, 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.]

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

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

Lineages	Models	$\text{Log}_e L$	$2\Delta L$[a]
CS tree	M0 (one ratio)	−40 070.62	–
	M1 (free ratio)	−39 881.44	378.36**
LS tree	M0(one ratio)	−35 919.33	–
	M1(free ratio)	−35 797.20	244.26**

[a]$2\Delta L$ is twice the log-likelihood difference between models M1 and M0.
**χ^2 test indicates the difference at the highly significant level of $P < 0.01$.

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

		Model A (branch-site)				M1 (free ratio)
Lineages	Branches	$\text{Log}_e L$	$2\Delta L$[a]	p_2[b]	ω_2[c]	ω[d]
CS tree	*a*	−39 570.13	71.96**	0.15	2.88	0.46
	a_s	−39 606.11	0.00	0.00	1.00	0.08
	b	−39 583.46	45.30**	0.13	3.13	0.39
	b_s	−39 597.71	16.8**	0.01	122.26	0.20
	c	−39 603.72	4.78	0.09	2.29	0.20
	c_s[e]	−39 598.38	15.46	0.10	68.18	∞
	d	−39 605.30	1.62	0.45	1.00	0.12
	d_s	−39 597.87	16.48**	0.02	47.32	0.17
	e	−39 592.80	26.62**	0.07	2.55	0.18
	e_s	−39 599.29	13.64	0.07	22.57	0.52
	f	−39 603.21	5.8	0.01	43.35	0.16
	f_s	−39 605.67	0.88	0.23	1.00	0.28
	g	−39 603.35	5.52	0.05	2.15	0.29
	g_s	−39 606.11	0.00	0.00	1.00	0.09
	l	−39 588.76	34.7**	0.12	2.90	0.35
	m	−39 600.18	11.88	0.02	8.30	0.19
	n	−39 598.45	15.32**	0.03	9.70	0.17
	o	−39 602.09	8.04	0.03	2.49	0.16
LS tree	*h*	−35 741.01	36.90**	0.10	40.31	0.43
	h_s	−35 756.88	5.16	0.02	12.08	0.07
	i[e]	−35 725.99	66.94	0.06	∞	0.41
	i_s	−35753.87	11.18	0.06	3.05	0.16
	j[e]	−35 737.98	42.96	0.08	∞	219.72
	j_s	−35 732.95	53.02**	0.09	7.17	0.18
	k	−35 745.91	27.1**	0.11	1.63	0.21
	k_s	−35 757.01	4.90	0.01	83.30	0.05

[a]$2\Delta L$ 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 −35 759.46 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]χ^2 test was not applied because of the infinite value of ω_2 or ω.
∞The dN/dS value was estimated to be 999.00.
**The difference at the highly significant level of $P < 0.01$ (the Bonferroni correction was used, where $P < 0.01/18$, $2\Delta L > 14.99$ and $P < 0.01/8$, $2\Delta L > 13.36$ for the CS tree and the LS tree, respectively).

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

myxobacterium (*Stigmatella aurantiaca*) and LS (parkeol) in planctomycete (*Gemmata obscuriglobus*) (Bode *et al.*, 2003; Pearson *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) suggested that hopanoid and steroid cyclases diverged from a common ancestor instead of the previous assumption that hopanoid biosynthesis was an evolutionary predecessor to steroid biosynthesis in the ancient life forms. Extensive phylogenetic analysis based on 5288 putative triterpene cyclase homologues in the publicly available databases revealed that a few sequences from above three bacterial species grouped with a set of OSCs from eukaryotic 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 lateral 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 & lkuse, 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 perspective, 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 biosynthesis 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 presumed 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 tandem duplications, single gene transposition-duplications, or combinations 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 agreement with the observation that gene families involved in the biosynthesis of secondary metabolites tend to arise by gene duplication, forming tandem clusters within the plant genome (Ober, 2005). *OSC* genes have been lost in most of the species we analyzed here after segmental duplication or whole-genome duplication. 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 metabolic 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 (Parniske *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 comparison 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 duplication 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, oat β-amyrin synthase (AsbAS1) 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* OSCs 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 & Osbourn, 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

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

Abe I. 2007. Enzymatic synthesis of cyclic triterpenes. *Natural Products Reports* **24**: 1311–1331.

Albrecht P, Ourisson G. 1969. Triterpene alcohol isolation from oil shale. *Science* **163**: 1192–1193.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

Augustin JM, Kuzina V, Andersen SB, Bak S. 2011. Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochemistry* **72**: 435–457.

Bishop JG, Dean AM, Mitchell-Olds T. 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. *Proceedings of the National Academy of Sciences, USA* **97**: 5322–5327.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Research* **13**: 137–144.

Bode HB, Zeggel B, Silakowski B, Wenzel SC, Reichenbach H, Muller R. 2003. Steroid biosynthesis in prokaryotes: identification of myxobacterial steroids and cloning of the first bacterial 2,3(S)-oxidosqualene cyclase from the myxobacterium *Stigmatella aurantiaca*. *Molecular Microbiology* **47**: 471–481.

Chi YH, Cheng YS, Vanitha J, Kumar N, Ramamoorthy R, Ramachandran S, Jiang S-Y. 2011. Expansion mechanisms and functional divergence of the glutathione s-transferase family in sorghum and other higher plants. *DNA Research* **18**: 1–16.

Corey EJ, Cheng HM. 1996. Conversion of a C_{20} 2,3-oxidosqualene analog to tricyclic structures with a five-membered C-ring by lanosterol synthase. Further evidence for a C-ring expansion step in sterol biosynthesis. *Tetrahedron Letters* **37**: 2709–2712.

Corey EJ, Matsuda SP, Bartel B. 1993. Isolation of an *Arabidopsis thaliana* gene encoding cycloartenol synthase by functional expression in a yeast mutant lacking lanosterol synthase by the use of a chromatographic screen. *Proceedings of the National Academy of Sciences, USA* **90**: 11628–11632.

Corey EJ, Virgil SC, Cheng H, Baker CH, Matsuda SPT, Singh V, Sarshar S. 1995. New insights regarding the cyclization pathway for sterol biosynthesis from (S)-2,3-oxidosqualene. *Journal of the American Chemical Society* **117**: 11819–11820.

Desmond E, Gribaldo S. 2009. Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature. *Genome Biology and Evolution* **1**: 364–381.

Ebizuka Y, Katsube Y, Tsutsumi T, Kushiro T, Shibuya M. 2003. Functional genomics approach to the study of triterpene biosynthesis. *Pure and Applied Chemistry* **75**: 369–374.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.

Elrouby N, Bureau TE. 2010. *Bs1*, a new chimeric gene formed by retrotransposon-mediated exon shuffling in maize. *Plant Physiology* **153**: 1413–1424.

Fazio GC, Xu R, Matsuda SP. 2004. Genome mining to identify new plant triterpenoids. *Journal of the American Chemical Society* **126**: 5678–5679.

Field B, Fiston-Lavierb A-S, Kemena A, Geislera K, Quesnevillec H, Osbourna AE. 2011. Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proceedings of the National Academy of Sciences, USA* **108**: 16116–16121.

Field B, Osbourn AE. 2008. Metabolic diversification–independent assembly of operon-like gene clusters in different plants. *Science* **320**: 543–547.

Fischer WW, Pearson A. 2007. Hypotheses for the origin and early evolution of triterpenoid cyclases. *Geobiology* **5**: 19–34.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology* **60**: 433–453.

Frickey T, Kannenberg E. 2009. Phylogenetic analysis of the triterpene cyclase protein family in prokaryotes and eukaryotes suggests bidirectional lateral gene transfer. *Environmental Microbiology* **11**: 1224–1241.

Gaut BS. 2002. Evolutionary dynamics of grass genomes. *New Phytologist* **154**: 15–28.

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

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiology* **148**: 993–1003.

Hanisch S, Ariztegui D, Puttmann W. 2003. The biomarker record of Lake Albano, central Italy - implications for Holocene aquatic system response to environmental change. *Organic Geochemistry* **34**: 1223–1235.

Haralampidis K, Bryan G, Qi X, Papadopoulou K, Bakht S, Melton R, Osbourn A. 2001. A new class of oxidosqualene cyclases directs synthesis of antimicrobial phytoprotectants in monocots. *Proceedings of the National Academy of Sciences, USA* **98**: 13431–13436.

Haralampidis K, Trojanowska M, Osbourn A. 2002. Biosynthesis of triterpenoid saponins in plants. *Advances in Biochemical Engineering/ Biotechnology* **75**: 31–49.

Hauke V, Adam P, Trendel JM, Albrecht P, Schwark L, Vliex M, Hagemann H, Puttmann W. 1995. Isoarborinol through geological times: evidence for its presence in the Permian and Triassic. *Organic Geochemistry* **23**: 91–93.

Herrera JB, Bartel B, Wilson WK, Matsuda SP. 1998. Cloning and characterization of the *Arabidopsis thaliana* lupeol synthase gene. *Phytochemistry* **49**: 1905–1911.

Hess BA. 2002. Concomitant C-ring expansion and D-ring formation in lanosterol niosynthesis from squalene without violation of Markovnikov's Rule. *Journal of the American Chemical Society* **124**: 10286–10287.

Hoen DR, Park KC, Elrouby N, Yu Z, Mohabir N, Cowan RK, Bureau TE. 2006. Transposon-mediated expansion and diversification of a family of ULP-like genes. *Molecular Biology and Evolution* **23**: 1254–1268.

Husselstein-Muller T, Schaller H, Benveniste P. 2001. Molecular cloning and expression in yeast of 2,3-oxidosqualene-triterpenoid cyclases from *Arabidopsis thaliana*. *Plant Molecular Biology* **45**: 75–92.

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

Inagaki Y-S, Etherington G, Geisler K, Field B, Dokarry M, Ikeda K, Mutsukado Y, Dicks J, Osbourn A. 2011. Investigation of the potential for triterpene synthesis in rice through genome mining and metabolic engineering. *New Phytologist* **191**: 432–448.

Ito R, Mori K, Hashimoto I, Nakano C, Sato T, Hoshino T. 2011. Triterpene cyclases from *Oryza sativa* L.: cycloartenol, parkeol and achilleol B synthases. *Organic Letters* **13**: 2678–2681.

Jaffé R, Hausmann KB. 1995. Origin and early diagenesis of arborinone/isoarborinol in sediments of a highly productive freshwater lake. *Organic Geochemistry* **22**: 231–235.

Jenson C, Jorgensen WL. 1997. Computational investigations of carbenium ion reactions relevant to sterol biosynthesis. *Journal of the American Chemical Society* **119**: 10846–10854.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**: 569–573.

Jiao YN, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang HY, Soltis PS *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97–100.

Kellogg EA. 2001. Evolutionary history of the grasses. *Plant Physiology* **125**: 1198–1205.

Kolesnikova MD, Obermeyer AC, Wilson WK, Lynch DA, Xiong Q, Matsuda SP. 2007a. Stereochemistry of water addition in triterpene synthesis: the structure of arabidiol. *Organic Letters* **9**: 2183–2186.

Kolesnikova MD, Wilson WK, Lynch DA, Obermeyer AC, Matsuda SP. 2007b. *Arabidopsis* camelliol C synthase evolved from enzymes that make pentacycles. *Organic Letters* **9**: 5223–5226.

Kolesnikova MD, Xiong Q, Lodeiro S, Hua L, Matsuda SP. 2006. Lanosterol biosynthesis in plants. *Archives of Biochemistry and Biophysics* **447**: 87–95.

Kovalchuk I, Kovalchuk O, Kalck V, Boyko V, Filkowski J, Heinlein M, Hohn B. 2003. Pathogen induced systemic plant signal triggers DNA rearrangements. *Nature* **423**: 760–762.

Kushiro T, Shibuya M, Masuda K, Ebizuka Y. 2000. A novel multifunctional triterpene synthase from *Arabidopsis thaliana*. *Tetrahedron Letter* **41**: 7705–7710.

Lamb DC, Jackson CJ, Warrilow AG, Manning NJ, Kelly DE, Kelly SL. 2007. Lanosterol biosynthesis in the prokaryote *Methylococcus capsulatus*: insight into the evolution of sterol biosynthesis. *Molecular Biology and Evolution* **24**: 1714–1721.

Leister D. 2004. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends in Genetics* **20**: 116–122.

Lescot M, Piffanelli P, Ciampi AY, Ruiz M, Blanc G, Mack JL, Silva FR, Santos CMR, Hont AD, Garsmeur O *et al.* 2008. Insights into the *Musa* genome: Syntenic relationships to rice and between Musa species. *BMC Genomics* **9**: 58.

Lodeiro S, Xiong QB, Wilson WK, Kolesnikova MD, Onak CS, Matsuda SP. 2007. An oxidosqualene cyclase makes numerous products by diverse mechanisms: a challenge to prevailing concepts of triterpene biosynthesis. *Journal of the American Chemical Society* **129**: 11213–11222.

Lucht JM, Mauch-Mani B, Steiner HY, Metraux JP, Ryals J, Hohn B. 2002. Pathogen stress increases somatic recombination frequency in Arabidopsis. *Nature Genetics* **30**: 311–314.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences, USA* **102**: 5454–5459.

Matsuno M, Compagnon V, Schoch GA, Schmitt M, Debayle D, Bassard JE, Pollet B, Hehn A, Heintz D, Ullmann P *et al.* 2009. Evolution of a novel phenolic pathway for pollen development. *Science* **325**: 1688–1692.

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Laylin LKF, Drouard LM *et al.* 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**: 245–250.

Michelmore RW, Meyers BC. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Research* **8**: 1113–1130.

Mondragon-Palomino M, Gaut BS. 2005. Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana*. *Molecular Biology and Evolution* **22**: 2444–2456.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences, USA* **104**: 19363–19368.

Morlacchi P, Wilson WK, Xiong Q, Bhaduri A, Sttivend D, Kolesnikova MD, Matsuda SP. 2009. Product profile of PEN3: the last unexamined oxidosqualene cyclase in *Arabidopsis thaliana*. *Organic Letters* **11**: 2627–2630.

Nakano C, Motegi A, Sato T, Onodera M, Hoshino T. 2007. Sterol biosynthesis by a prokaryote: first *in vitro* identification of the genes encoding squalene epoxidase and lanosterol synthase from *Methylococcus capsulatus*. *Bioscience Biotechnology and Biochemistry* **71**: 2543–2550.

Nishimoto K, Ito M, Natori S, Ohmoto T. 1968. The structures of arundoin, cylindrin and fernenol. *Tetrahedron* **24**: 735–752.

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

Ohmoto T, Ikuse M. 1970. Triterpenoids of the *Gramineae*. *Phytochemistry* **9**: 2137–2148.

Ohyama K, Suzuki M, Kikuchi J, Saito K, Muranaka T. 2009. Dual biosynthetic pathways to phytosterol via cycloartenol and lanosterol in *Arabidopsis*. *Proceedings of the National Academy of Sciences, USA* **106**: 725–730.

Osbourn A, Goss RJM, Field RA. 2011. The saponins – polar isoprenoids with important and diverse biological activities. *Natural Products Reports* **28**: 1261–1268.

Ourisson G, Albrecht P, Rohmer M. 1982. Predictive microbial biochemistry–from molecular fossils to procaryotic membranes. *Trends in Biochemical Sciences* **7**: 236–239.

Papadopoulou K, Melton RE, Leggett M, Daniels MJ, Osbourn AE. 1999. Compromised disease resistance in saponin-deficient plants. *Proceedings of the National Academy of Sciences, USA* **96**: 12923–12928.

Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones DA, Harrison K, WulffBrande BH, Jones JDG. 1997. Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of Tomato. *Cell* **91**: 821–832.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A *et al.* 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**: 551–556.

Pearson A, Budin M, Brocks JJ. 2003. Phylogenetic and biochemical evidence for sterol synthesis in the bacterium *Gemmata obscuriglobus*. *Proceedings of the National Academy of Sciences, USA* **100**: 15352–15357.

Phillips DR, Rasbery JM, Bartel B, Matsuda SP. 2006. Biosynthetic diversity in plant triterpene cyclization. *Current Opinion in Plant Biology* **9**: 305–314.

Pichersky E, Gang DR. 2000. Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends in Plant Science* **5**: 439–445.

Qi X, Bakht S, Leggett M, Maxwell C, Melton R, Osbourn A. 2004. A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proceedings of the National Academy of Sciences, USA* **101**: 8233–8238.

Qi X, Bakht S, Qin B, Leggett M, Hemmings A, Mellon F, Eagles J, Reichhart DW, Schaller H, Lesot A *et al.* 2006. A different function for a member of an ancient and highly conserved cytochrome P450 family: from essential sterols to plant defense. *Proceedings of the National Academy of Sciences, USA* **103**: 18848–18853.

Rizzon C, Ponger L, Gaut BS. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Computational Biology* **2**: e115.

Ronquist F, Huelsenbeck JP. 2003. MRBAYES3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C. 2008. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**: 11–24.

Sawai S, Akashi T, Sakurai N, Suzuki H, Shibata D, Ayabe S, Aoki T. 2006. Plant lanosterol synthase: divergence of the sterol and triterpene biosynthetic pathways in eukaryotes. *Plant Cell Physiology* **47**: 673–677.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA *et al.* 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112–1115.

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust

Shibuya M, Katsube Y, Otsuka M, Zhang H, Tansakul P, Xiang T, Ebizuka Y. 2009. Identification of a product specific beta-amyrin synthase from *Arabidopsis thaliana*. *Plant Physiology Biochemistry* **47**: 26–30.

Shimada N, Aoki T, Sato S, Nakamura Y, Tabata S, Ayabe S. 2003. A cluster of genes encodes the two types of chalcone isomerase involved in the biosynthesis of general flavonoids and legume-specific 5-deoxy(iso)flavonoids in *Lotus japonicus*. *Plant Physiology* **131**: 941–951.

Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KF, Li WH. 2004. Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. *Plant Cell* **16**: 1220–1234.

Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* **99**: 13627–13632.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analysis with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Suzuki H, Sawada S, Watanabe K, Nagae S, Yamaguchi MA, Nakayama T, Nishino T. 2004. Identification and characterization of a novel anthocyanin malonyltransferase from scarlet sage (*Salvia splendens*) flowers: an enzyme that is phylogenetically separated from other anthocyanin acyltransferases. *Plant Journal* **38**: 994–1003.

Suzuki M, Xiang T, Ohyama K, Seki H, Saito K, Muranaka T, Hayashi H, Katsube YJ, Kushiro T, Shibuya M *et al.* 2006. Lanosterol synthase in dicotyledonous plants. *Plant Cell Physiology* **47**: 565–571.

Vogt T, Jones P. 2000. Glycosyltransferases in plant natural product synthesis: characterization of a supergene family. *Trends in Plant Science* **5**: 380–386.

Vorbrüggen H, Pakrashi SC, Djerassi C. 1963. Arborinol, ein neuer Triterpen-Typus. *Annalen Der Chemie-Justus Liebig* **668**: 57–76.

Wu X, Qi X. 2010. Genes encoding hub and bottleneck enzymes of the *Arabidopsis* metabolic network preferentially retain homeologs through whole genome duplication. *BMC Evolutionary Biology* **10**: 145.

Xiang T, Shibuya M, Katsube Y, Tsutsumi T, Otsuka M, Zhang H, Masuda K, Ebizuka Y. 2006. A new triterpene synthase from *Arabidopsis thaliana* produces a tricyclic triterpene with two hydroxyl groups. *Organic Lettters* **13**: 2835–2838.

Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E. 2008. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* **319**: 1527–1530.

Xiong Q, Wilson WK, Matsuda SP. 2006. An *Arabidopsis* oxidosqualene cyclase catalyzes iridal skeleton formation by Grob fragmentation. *Angewandte Chemie International Edition* **45**: 1285–1288.

Xu R, Fazio GC, Matsuda SP. 2004. On the origins of triterpenoid skeletal diversity. *Phytochemistry* **65**: 261–291.

Yamanouchi U, Yano M, Lin HX, Ashikari M, Yamada K. 2002. A rice spotted leaf gene, *spl7*, encodes a heat shock transcription factor protein. *Proceedings of the National Academy of Sciences, USA* **99**: 7530–7535.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**: 1586–1591.

Yu J, Wang J, Lin W, Li SG, Li H, Zhou J, Ni PX, Dong W, Hu SN, Zeng CQ *et al.* 2005. The Genomes of *Oryza sativa*: a history of duplications. *PLoS Biology* **3**: e38.

Zhang LQ, Gaut BS. 2003. Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? *Genome Research* **13**: 2533–2540.

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 + Γ + I substitution model using the MrBayes3.1.2 software.

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

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

Fig. S6 Segmentally duplicated blocks of oxidosqualene cyclases (*OSCs*) regions in *Brachypodium distachyo*n 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 distachyo*n 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.

© 2011 The Authors
New Phytologist © 2011 New Phytologist Trust