New Phytologist

Research

Molecular evolution and expression divergence of the *Populus* polygalacturonase supergene family shed light on the evolution of increasingly complex organs in plants

Zhi-Ling Yang[1,2], Hai-Jing Liu[1,2], Xiao-Ru Wang[3] and Qing-Yin Zeng[1]

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, 100093, China; [2]University of Chinese Academy of Sciences, Beijing, 100049, China; [3]Department of Ecology and Environmental Science, UPSC, Umeå University, SE-90187, Umeå, Sweden

Author for correspondence:
Qing-Yin Zeng
Tel: +86 10 62836440
Email: qingyin.zeng@ibcas.ac.cn

Received: *4 October 2012*
Accepted: *21 November 2012*

New Phytologist (2013) **197:** 1353–1365
doi: 10.1111/nph.12107

Key words: copy number variation, expression divergence, gene family, selective pressure, subcellular localization.

Summary

- Plant polygalacturonases (PGs) are involved in cell separation processes during many stages of plant development. Investigation into the diversification of this large gene family in land plants could shed light on the evolution of structural development.
- We conducted whole-genome annotation, molecular evolution and gene expression analyses of PG genes in five species of land plant: *Populus*, *Arabidopsis*, rice, *Selaginella* and *Physcomitrella*.
- We identified 75, 44, 16 and 11 PG genes from *Populus*, rice, *Selaginella* and *Physcomitrella* genomes, respectively, which were divided into three classes. We inferred rapid expansion of class I PG genes in *Populus*, *Arabidopsis* and rice, while copy numbers of classes II and III PG genes were relatively conserved in all five species. *Populus*, *Arabidopsis* and rice class I PG genes were under more relaxed selection constraints than class II PG genes, while this selective pressure divergence was not observed in *Selaginella* and *Physcomitrella* PG families. In addition, class I PG genes underwent marked expression divergence in *Populus*, rice and *Selaginella*.
- Our results suggest that PG gene expansion occurred after the divergence of the lycophytes and euphyllophytes, and this expansion was likely paralleled by the evolution of increasingly complex organs in land plants.

Introduction

The plant cell wall is involved in many essential biological processes, including mechanical support, regulation of cell expansion, the control of tissue cohesion, ion exchange and protection against pathogens (Popper, 2008; Wei *et al.*, 2009). Polysaccharides represent up to 95% of the plant cell wall mass, whereas cell wall proteins account only for 5–10% (Jamet *et al.*, 2008). The polysaccharides of plant cell walls are divided into three principal types: pectin, cellulose and hemicellulose (Yin *et al.*, 2010). Pectin is the most structurally complex family of polysaccharides, making up *c.* 35% of primary walls in dicots and nongraminaceous monocots and has functions in cell wall porosity, cell adhesion, plant defense and so on (Mohnen, 2008).

Plant cell walls are highly dynamic structures that undergo changes in composition and configuration in response to functional requirements (Prieto-Alcedo *et al.*, 2011). The pectin network is systematically disassembled during many stages of plant development, such as organ abscission, fruit ripening, and pod and anther dehiscence (Hadfield & Bennett, 1998; Ogawa *et al.*, 2009). The dismantling of the pectin network occurs through the action of pectin-degrading enzymes, including polygalacturonase (PG), pectate lyase and pectin methylesterase (Bosch & Hepler, 2005). Among these hydrolytic enzymes, the PGs belong to one of the largest hydrolase families (Kim *et al.*, 2006). PGs are important for pectin disassembly (Hadfield & Bennett, 1998; Munoz *et al.*, 1998). A recent study showed that the knockout of three *Arabidopsis* PGs led to the failure of pollen grain separation and silique and anther dehiscence (Ogawa *et al.*, 2009). Overexpression of a PG gene in transgenic apple (*Malus domestica* Borkh. cv Royal Gala) trees led to premature leaf shedding as a result of reduced cell adhesion in leaf abscission zones (Atkinson *et al.*, 2002). The functions of PGs are not restricted to plant developmental processes, but also include wound responses and host–parasite interactions (Orozco-Cardenas & Ryan, 2003). Thus, plant PG genes have extensive functional divergence.

Plant PGs are multifunctional proteins encoded by a large gene family. The *Arabidopsis* and rice (*Oryza sativa*) genomes contain 66 and 42 members, respectively, which are divided into three distinct groups (Kim *et al.*, 2006). To date, genome-wide analyses of the PG gene family have focused mainly on herbaceous annual plants, such as *Arabidopsis* and rice. The expansion and functional divergence of this large gene family in other land plants, however, have not been investigated. Thus, there is a lack of general understanding of the tempo and mode of PG gene family evolution in plants. Now with the completion of

lycophyte *Selaginella moellendorffii* (Banks *et al.*, 2011), bryophyte *Physcomitrella patens* (Rensing *et al.*, 2008), and *Populus trichocarpa* (Tuskan *et al.*, 2006) genome projects, a joint phylogenetic analysis of PG genes from bryophyte, lycophyte and angiosperm monocot (rice) and eudicot (*Populus, Arabidopsis*) will help us to understand the expansion and diversification of this large gene family in land plants. Bryophytes are the closest extant relatives of early land plants, which began to diverge *c.* 450 million yr ago (Rensing *et al.*, 2008). Lycophytes are early vascular plants with a dominant sporophyte generation (Banks *et al.*, 2011). Compared with lycophytes and bryophytes, the angiosperm body possesses more complex organ systems and structures, such as flowers and fruit. One of the major questions in plant evolution concerns the evolution of the body developmental program, which was modified through time so that sporophytes became larger and acquired the ability to branch, develop conducting tissues, and produce roots, leaves, seeds, and flowers (Bowman *et al.*, 2007). *Populus* represents a woody perennial plant, and is the most important tree model system of plant genomics currently available. The relatively close phylogenetic relationship between *Populus* and *Arabidopsis* in the eurosid clade of eudicotyledonous plants facilitates studies of gene family evolution in eudicots (Jansson & Douglas, 2007). In this study, we conducted genome-wide annotation of PG genes in the *Populus* genome. Through a comprehensive analysis of gene sequences, gene structures, molecular evolution, gene expression patterns under normal growth conditions and in response to stress treatments, and subcellular localization of PG proteins, we provided detailed characterization of the composition, expansion and expression divergence of the PG gene family in *Populus*. By further comparative analyses of this gene family in different land plant species, our study sheds light on the evolution of increasingly complex organs in plants.

Materials and Methods

Identification of PG genes from the *Populus* genome

To identify *Populus* PG gene family members, the *Populus* genome database version 2.2 (http://www.phytozome.net/) was searched using 66 *Arabidopsis* PG protein sequences in the TBLASTN program with default algorithm parameters. Analysis of the collected *Populus* PG candidates indicated that some sequences were partially misannotated during the automated genome annotation process. Thus, manual reannotation was performed to rectify incorrect start codon predictions, splicing errors, missed or extra exons. Previous studies showed that all PG proteins contained glycosyl hydrolase family 28 (GH28) domains (Kim *et al.*, 2006). In this study, all *Populus* PG candidates collected were primarily analyzed using the protein families database (Pfam) to confirm the presence of GH28 domains in their protein structures. To verify the intron/exon structures of *Populus* PGs, genes were amplified from *Populus* cDNA, cloned into *pEASY*-T3 vector (TransGen, Beijing, China) and sequenced in both directions. In this study, the cDNAs of 12 *Populus* PGs were not detected by PCR. The gene structures of these 12 *Populus* PGs were assumed to be identical to that of their closest relative on the phylogenetic tree. This approach was adapted from other studies (Meyers *et al.*, 2003).

Phylogenetic and molecular evolution analyses

Full-length PG protein sequences were aligned with the MUSCLE software (http://www.drive5.com/muscle/) and manually adjusted using BioEdit (Hall, 1999). Phylogenetic relationships were reconstructed using a maximum-likelihood (ML) procedure in PHYML software (Guindon & Gascuel, 2003) with the Jones, Taylor, and Thornton (JTT) amino acid substitution model. One-thousand bootstrap replicates were performed in each analysis.

The synonymous substitution (K_s) rates among all pairwise comparisons within *Populus* PG genes were calculated with the YN00 program in the PAML package (Yang, 2007). To evaluate variation in selective pressures between the *Populus* classes I and II PG genes, the branch models of CODEML in PAML were used to estimate ω ($= d_N/d_S$) under two assumptions: a one-ratio model that assumes the same ω ratio for two classes; and a two-ratio model in which the two classes were assigned to different ω ratios. To verify which of the models best fits the data, likelihood ratio tests (LRTs) were performed by comparing twice the difference in log-likelihood values between pairs of the models using a χ^2 distribution (Yang & Nielsen, 2000).

Expression of PG genes under different treatments

To investigate the expression patterns of the *Populus* PG genes under normal and abiotic stress conditions, seedlings of *Populus* were cultivated in potting soil at 25°C under 14 : 10 h light : dark coniditions in a growth chamber for 2 months before treatment. For salt and ABA treatments, seedlings were drenched and sprayed once a day using 150 mM NaCl solution for 1 wk and 100 μM ABA solution for 12 h, respectively. Drought stress was induced by withholding water for 2 wk. Cold stress was induced by growth at 4°C for 12 h. Each treatment consisted of three replicates. After treatment, total RNAs were isolated from roots, shoots, leaves, buds, phloem and leaf abscission zones (the narrow portion from the petiole to stem). In addition, total RNAs were isolated from the flower of *c.* 15-yr-old *Populus* trees. To investigate the expression patterns of the *P. patens* PG gene family, *Physcomitrella* plants were cultured in BCDATG solidified medium at 25°C under an 18 : 6 h light : dark cycle for *c.* 6 months. Total RNAs were isolated from rhizoids, stems and leaves, respectively. To investigate the expression patterns of *Selaginella* PG gene family, *Selaginella* plants were cultivated in potting soil for 2 yr. Total RNAs were isolated from root, stem and leaf tissues. Total RNAs isolated from *Populus*, *Physcomitrella* and *Selaginella* were treated with RNase-free DNase I (Promega, Madison, WI, USA) and reverse-transcribed into cDNA using an RNA PCR Kit (AMV) version 3.0 (TaKaRa, Dalian, China). Based on the multiple sequence alignment of PG gene sequences, specific PCR primers were designed (Supporting Information, Table S1). PCR conditions were optimized to consist of an initial

denaturation of 3 min at 94°C, followed by 35 cycles of 30 s at 94°C, 40 s at 65°C and 1 min at 72°C with a final extension of 3 min at 72°C. In all PCR analyses, the *Actin* gene was used as an internal control. PCR products from each sample were analyzed using a 1% agarose gel and were validated by DNA sequencing. Independent biological triplicates were used in all gene expression analyses.

To evaluate the significance of differences in the gene expression patterns of class I and II PGs, we conducted a multiple response permutation procedure (MRPP) test. MRPP is a nonparametric test and is flexible over unequal sample sizes and violation of normality assumptions. In the MRPP test, positive and negative detection of gene expression in the corresponding tissue were defined as 1 and 0, respectively. We used Euclidean distance to calculate dissimilarity and ran 10 000 permutations to calculate the *P* value.

Subcellular localization of *Populus* PGs

We selected nine *Populus* PG proteins and investigated their subcellular localizations by generating C-terminal green fluorescent protein (GFP) fusions. The nine full-length PG genes were subcloned into the modified pCAMBIA1302 transgenic vector (Fig. S1). The primers used to construct the PG subcellular localization vectors are listed in Table S2. Colonies containing the appropriate insert were identified by sequencing. The modified pCAMBIA1302 vectors containing PG sequences were transformed into *Agrobacterium tumefaciens* LBA4404. Cultures were infiltrated into epidermal cells of tobacco (*Nicotiana benthamiana*) leaves as described by Sparkes *et al.* (2006). Transformed tissues were harvested 3–5 d later and immediately observed under a confocal laser microscope (Olympus FV1000MPE, Tokyo, Japan). GFP fusion fluorescence was excited with a 488 nm laser. Localization in the epidermal cell wall was confirmed by plasmolysis. Plasmolysis was induced by infiltrating 0.3 g ml^{-1} sucrose into the leaves of tobacco using a needleless syringe.

Results

Sequence and structural characteristics of *Populus* PG genes

Seventy-five full-length genes encoding putative PG proteins were identified in the *Populus* genome (Table S3). All collected *Populus* PG candidates were analyzed primarily using Pfam to confirm the presence of GH28 domains in the protein structure. All 75 genes contained GH28 domains, confirming that these genes belong to the PG family. In addition to full-length PG genes, 21 partial PG fragments were identified in the *Populus* genome (Table S4). The length of these PG fragments ranged from 40 to 385 amino acids. Pfam analysis showed that these fragments contained partial GH28 domains. In this study, these PG fragments were considered to be pseudogenes.

Phylogenetic relationships among the 75 *Populus* PGs were reconstructed using an ML procedure. The phylogenetic tree showed that the PGs formed three distinct clades (blue, yellow and brown clades) with 100% bootstrap support (Fig. 1a). In this study, the PG genes in the blue, yellow and brown clades are termed classes I, II and III PGs, respectively. *Populus* classes I, II and III PG genes contained 60, 14 and one member, respectively.

Pairwise comparisons of the 75 full-length PG protein sequences revealed some notable features. All class I PGs showed > 23% pairwise sequence identity and all class II PGs showed > 35% pairwise sequence identity. A box plot showed that the protein sequence identities of class II PGs were higher than those of class I PGs (independent-sample *t*-test, $P < 0.0001$) (Fig. 2), indicating that the degree of sequence divergence among class I PGs was higher than that among class II PGs. In a pairwise protein sequence comparison, low sequence identities between the PG classes were found: < 24% between class I and class II PGs, < 18% between class I and class III PGs, and < 20% between class II and class III PGs. An independent sample *t*-test showed that the protein sequence identities within each class were higher than that between classes ($P < 0.0001$).

Highly variable gene structures were observed in *Populus* class I PGs (Fig. 1c). Among 60 class I PGs, two had two introns, 25 had three introns, nine had four introns, six had five introns, four had six introns, four had seven introns and 10 had eight introns. Compared with class I PGs, conservative gene structures were found among class II PG genes. Among 14 class II PGs, five had a five-intron/six-exon structure and nine had a four-intron/five-exon structure.

Expansion of the PG family in *Populus* and *Arabidopsis*

To investigate the extent of lineage-specific expansion of the PG genes in *Populus* and *Arabidopsis*, we performed a joint phylogenetic analysis of all *Populus* and *Arabidopsis* PGs. The 75 *Populus* and 66 *Arabidopsis* PGs fell into three distinct classes (Fig. 3). We identified the nodes that led to *Populus*- and *Arabidopsis*-specific clades (circles in Fig. 3). These nodes indicated the divergence point between *Populus* and *Arabidopsis*, and thus represented the most recent common ancestral genes before the split. Some PG genes might have been present in the most recent common ancestor (MRCA) of *Populus* and *Arabidopsis* but were later lost in either *Populus* or *Arabidopsis*. We found that five clades contained only *Populus* PG genes (Fig. 3, black arrows) and 10 clades contained only *Arabidopsis* PG genes (Fig. 3, red arrows), indicating that gene loss might have occurred in these clades. The number of clades indicated that there were at least 39 ancestral PG genes before the *Populus–Arabidopsis* split. In addition, four clades (Fig. 3; clades a14, a15, a16 and b6) had low bootstrap support (< 50%). If we assume these less well-supported nodes are correct, there were at least 43 ancestral PG genes before the *Populus–Arabidopsis* split.

Populus and *Arabidopsis* had 60 and 53 class I PG genes, respectively. There were at least 31 ancestral class I PG genes in the MRCA of *Populus* and *Arabidopsis* (Fig. 4). After the split, *Populus* and *Arabidopsis* have gained 36 and 26 genes and lost seven and four genes, respectively, resulting in the rapid

New Phytologist

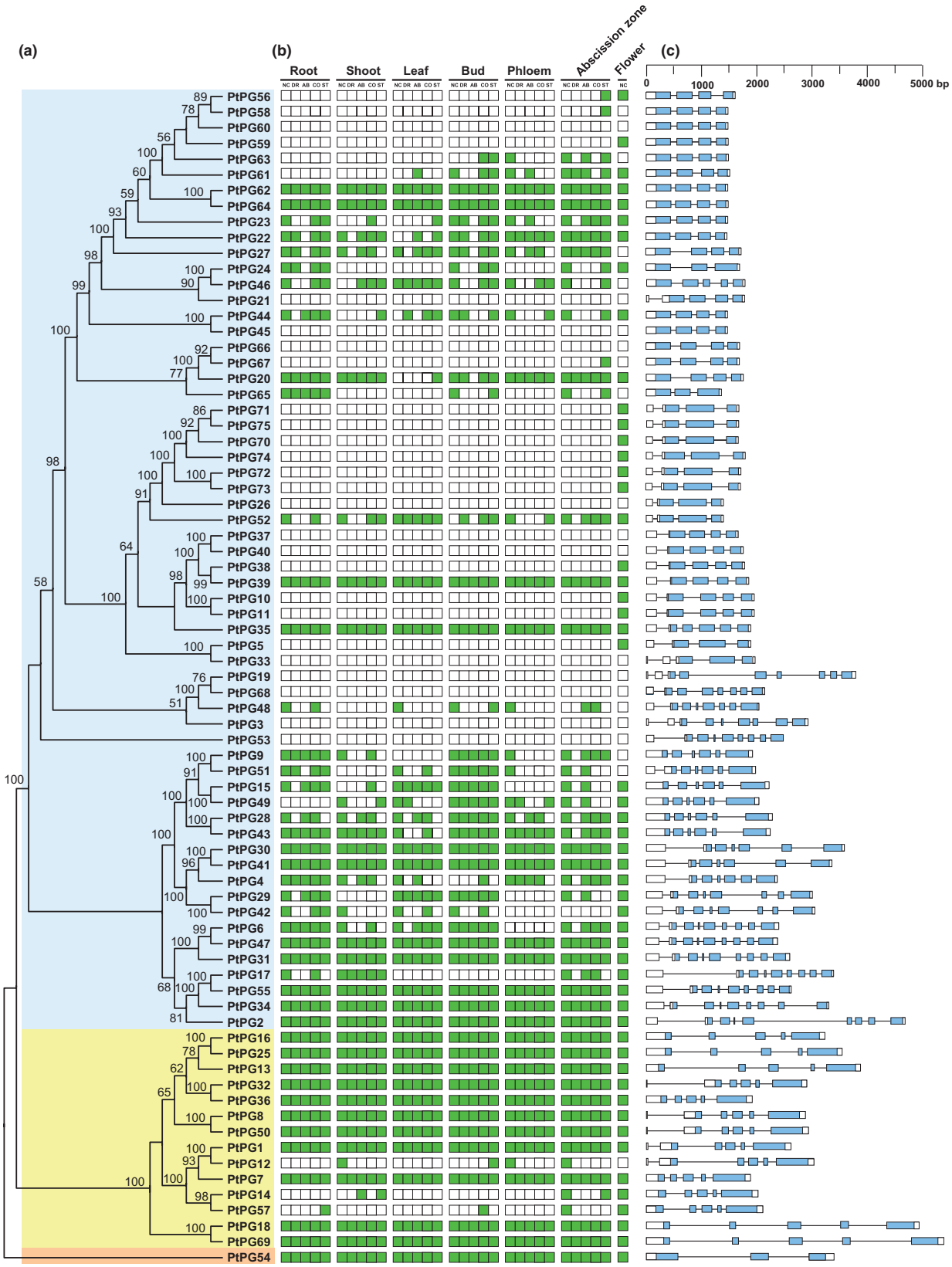

Fig. 1 Phylogenetic relationships among *Populus* polygalacturonases (PGs) (a), their expression patterns (b) and gene structures (c). Numbers on branches indicate the bootstrap percentage values calculated from 1000 replicates, and only values higher than 50% are shown. Classes I, II and III PGs are shaded blue, yellow and brown, respectively. In (b), the green box indicates positive detection of gene expression in the corresponding tissue under normal growth conditions (NC) and following drought (DR), abscisic acid (AB), cold (CO) and salt (ST) stress treatments. In (c), the GH28 domains are highlighted by blue boxes. Introns are shown as lines.

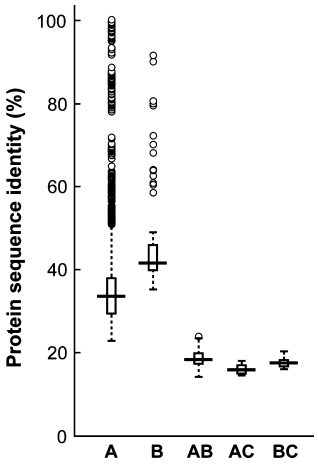

Fig. 2 Pairwise sequence identity of full-length *Populus* polygalacturonase (PG) proteins. A and B represent pairwise sequence identities of classes I and II PG proteins, respectively. AB, AC and BC represent pairwise sequence identities between classes I and II PG proteins, classes I and III PG proteins, and classes II and III PG proteins, respectively. The box plot shows the median (black line), interquartile range (box), and maximum and minimum scores (whiskers) of each data set. Outliers are shown as circles outside of the whiskers.

expansion of class I PG genes in these two plants. Clearly, the number of genes gained in the *Populus* lineage was greater than that in the *Arabidopsis* lineage.

Based on phylogenetic analysis, there were at least 11 ancestral class II PG genes in the MRCA of *Populus* and *Arabidopsis* (Fig. 4). Since the split, *Populus* and *Arabidopsis* have gained six and two genes and lost three and one gene, respectively, resulting in the conserved copy numbers of class II PG genes in the two species. Each *Populus* and *Arabidopsis* genome had one class III PG gene and neither gain nor loss of genes was observed in either species.

Duplication mechanisms accounting for the *Populus* PG family expansion

We examined the distribution of the PG genes on the *Populus* chromosomes. Seventy-three PG genes were localized on 19 *Populus* chromosomes (Fig. 5), whereas the other two were on one as-yet-unattributed scaffold fragment (Table S3). The distribution of the PG genes on the chromosomes appears to be non-random. Three clusters (clusters I, II and III) with high densities of PG genes were observed on chromosomes 7, 17 and 19. Four pairs of PG genes (*PtPG10*|*11*, *PtPG37*|*38*, *PtPG39*|*40*, and

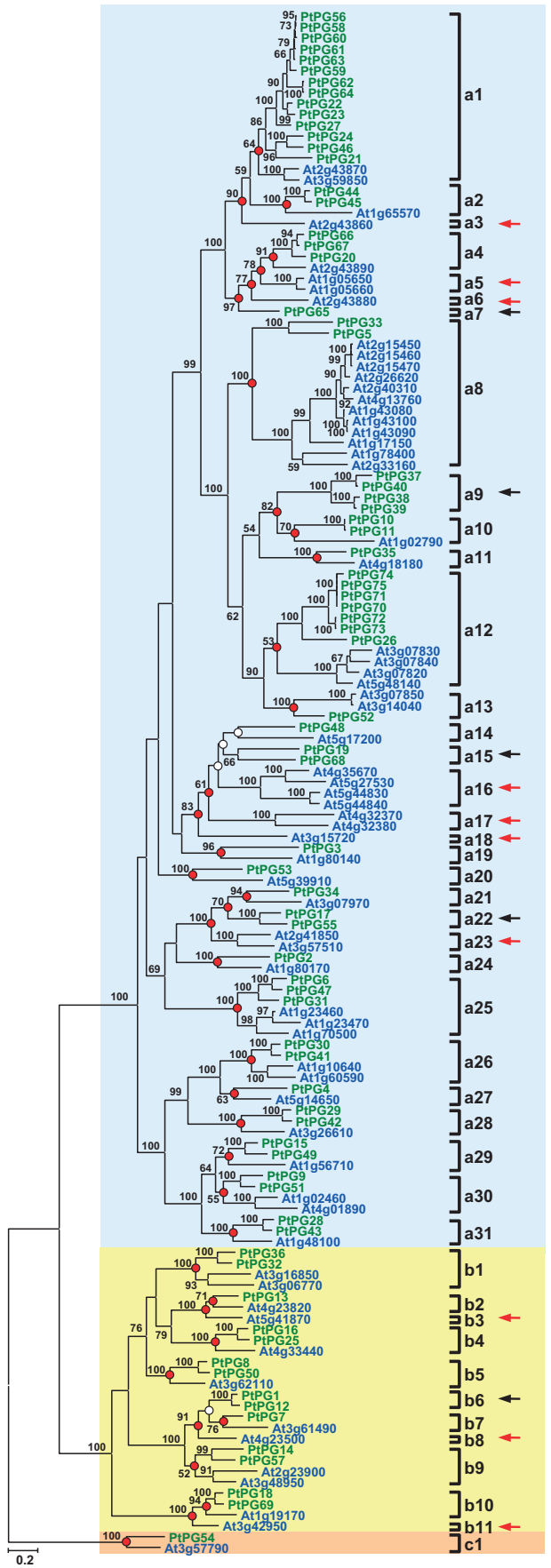

Fig. 3 Phylogenetic tree of *Populus* and *Arabidopsis* polygalacturonases (PGs). Numbers on branches indicate the bootstrap percentage values calculated from 1000 replicates, and only values higher than 50% are shown. Classes I, II and III PGs are shaded blue, yellow and brown, respectively. The nodes that represent the most recent common ancestral genes before the *Populus* and *Arabidopsis* split are indicated by red circles (bootstrap support > 50%) and white circles (bootstrap support < 50%). Clades that contain only *Populus* or *Arabidopsis* PGs are indicated by black or red arrows, respectively.

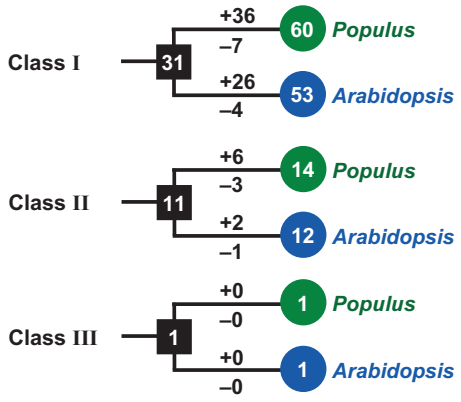

Fig. 4 The copy number changes of *Populus* and *Arabidopsis* polygalacturonase (PG) genes. Numbers in circles and rectangles represent the numbers of PG genes in extant and ancestral species, respectively. Numbers on branches with plus and minus symbols represent the numbers of gene gains and losses, respectively.

PtPG44/*45*) were arranged in tandem repeats in four positions on chromosomes 2 and 10. Chromosomes 4, 11, 12, 13 and 15 each harbored only one PG gene.

Whole-genome analysis showed a recent whole-genome duplication event (salicoid duplication, 60–65 million yr ago) that affected *c.* 92% of the *Populus* genome (Tuskan *et al.*, 2006). Paralogous segments created by this whole-genome duplication event were identified in previous analyses of the *Populus* genome (Tuskan *et al.*, 2006). Fourteen duplicate pairs, including *PtPG1*/*12*, *PtPG6*/*47*, *PtPG8*/*50*, *PtPG9*/*51*, *PtPG14*/*57*, *PtPG15*/*49*, *PtPG16*/*25*, *PtPG17*/*55*, *PtPG18*/*69*, *PtPG19*/*68*, *PtPG28*/*43*, *PtPG29*/*42*, *PtPG30*/*41* and *PtPG32*/*36*, are each located in a pair of paralogous blocks and can be considered to be direct results of the salicoid duplication event (Fig. 5). Similarly, PG gene clusters I and II correspond to paralogous blocks created by the salicoid duplication. The phylogenetic tree showed that the 15 PG genes in clusters I and II were grouped together (Fig. 1a), indicating that these genes descended from a single common ancestor.

Based on the phylogenetic tree and the positions of these 15 PG genes, we attempted to reconstruct the expansion history of the two clusters. The most parsimonious scenario for gene duplication, loss and rearrangement is presented in Fig. S2. It is likely that six ancestral genes created by three rounds of tandem duplications existed in the *Populus* genome before salicoid whole-genome duplication. After whole-genome duplication, one round of tandem duplication, loss and rearrangement took place in cluster I, while cluster II might have undergone complex tandem duplication, loss and rearrangement events.

Rapid expansion of class I PG genes in angiosperm plants

The newly available plant genomes (e.g. the lycophyte *Selaginella moellendorffii* and the bryophyte *Physcomitrella patens*) allow new insight into the evolution of the PG gene family in land plants. In this study, we identified 16 and 11 PG genes from the *Selaginella* and *Physcomitrella* genomes (Table S5), respectively. A previous study showed that rice and *Arabidopsis* possess 42 and 66 PG genes, respectively (Kim *et al.*, 2006). In this study, based on the latest version of the rice genome sequence (version 7, released on 31 October 2011), we identified 44 PG genes in the rice genome. Using the ML method, we reconstructed the phylogenetic relationships among 212 PGs from *Physcomitrella*, *Selaginella*, rice, *Arabidopsis* and *Populus*. The phylogenetic tree showed that all the PGs were grouped into three distinct classes (classes I, II and III) with 100% bootstrap support (Fig. 6).

The marked difference in PG gene family size among bryophyte, lycophyte and angiosperms (Fig. 7) suggests that PG gene expansion occurred after the divergence of the lycophytes and euphyllophytes. Owing to the lack of genome information for gymnosperms and monilophytes, the tempo of PG gene family evolution in land plants cannot yet be precisely defined. We found highly conserved copy numbers of the class III PGs in the each of the five species: each *Physcomitrella* and *Selaginella* genome had two class III PG genes, and each of rice, *Arabidopsis* and *Populus* had only a single class III PG gene. Similar to class III PGs, conserved copy numbers of the class II PGs in the five species were also observed: *Physcomitrella* and *Selaginella* had six and five copies, respectively, and rice, *Arabidopsis* and *Populus* had 11, 12 and 14 copies, respectively. However, pronounced copy number variations of class I PG genes were observed in these plant species: *Physcomitrella* and *Selaginella* contained three and nine genes, respectively, while rice, *Arabidopsis* and *Populus* had 32, 53 and 60 genes, respectively. Thus, the expansion of the PG gene family in angiosperm plants is mainly the result of the rapid expansion of class I PG genes.

Divergence of selective pressure between the class I and class II PG genes

Compared with class II PGs, Class I PGs showed rapid expansion in *Populus*, *Arabidopsis* and rice. To infer the influence of selection on the expansion of class I PGs, we determined if there was a significant change in the selective pressure between the two classes using ML codon models in PAML software. Two assumptions were tested: a one-ratio model that assumes the same ω ($= d_N/d_S$) ratio for the two classes and a two-ratio model in which the two classes were assigned to a different ω ratio (Table 1). For *Populus* PG genes, the log-likelihood values under the one-ratio and two-ratio models were $\log_e L = -36156.1315$ and -36142.9111, respectively. The LRT showed that the two-ratio model rejected the null model (one-ratio model), indicating that the selective pressure differed significantly between the two classes ($P < 0.0001$). Under the two-ratio model, the ω values for *Populus* classes I and II were 0.1879 and 0.1245, respectively, indicating that the *Populus* class I PG genes were under more relaxed selection constraints than the class II PG genes. Similar to *Populus*, class I PG genes in *Arabidopsis* and rice also had higher ω values than that of their class II PG genes. However, *Selaginella* class I PG genes showed similar selective pressure to class II PG genes, and

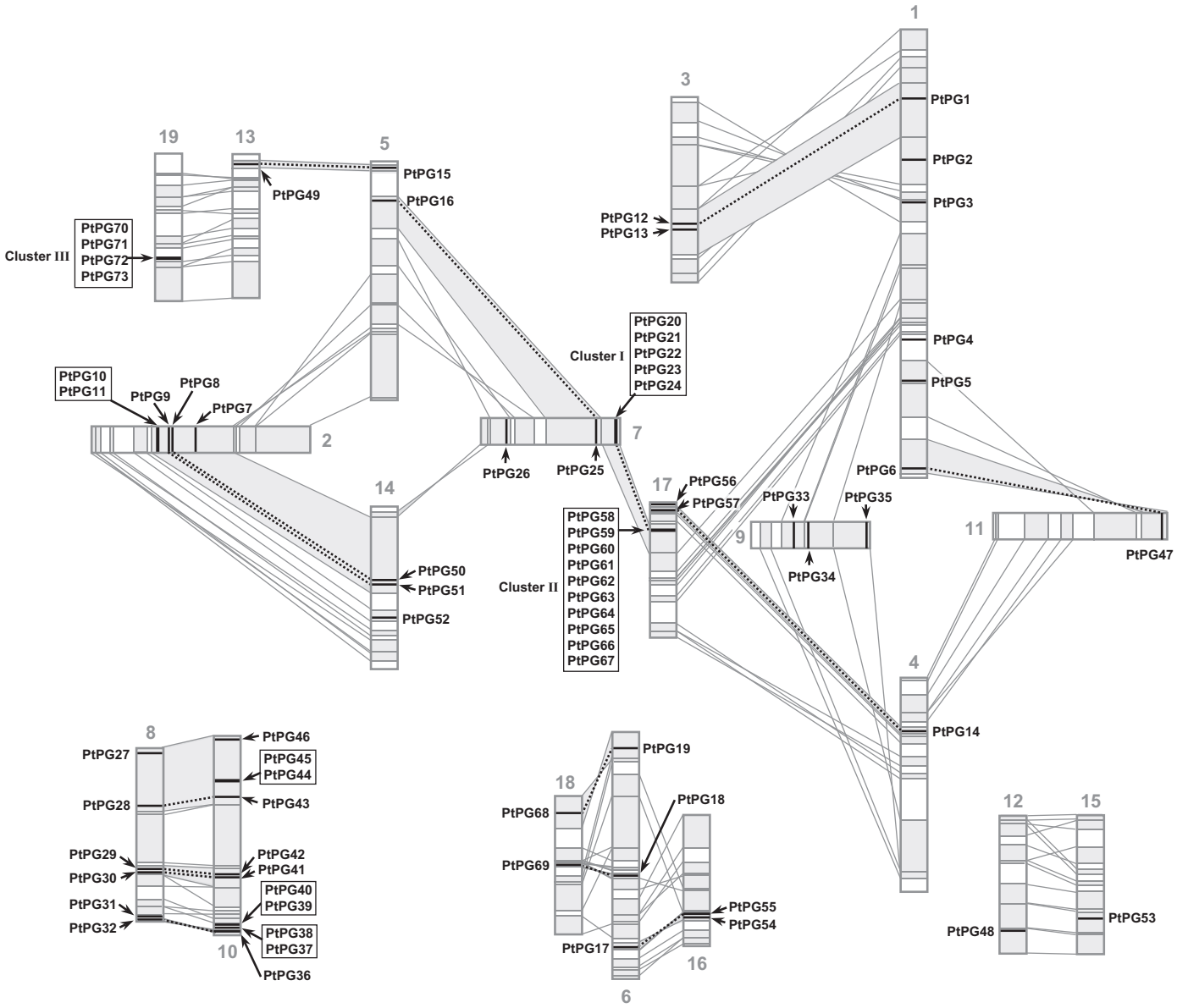

Fig. 5 Genomic localization of *Populus* polygalacturonase (PG) genes. Schematic view of chromosome reorganization by the most recent whole-genome duplication in *Populus* (adapted from Tuskan *et al.*, 2006, with permission from AAAS). Regions that are assumed to correspond to homologous genome blocks are shaded gray and connected by lines. Paralogous PG genes and clusters are indicated by dashed lines within the gray-shaded trapezoids.

Physcomitrella class II PG genes were under more relaxed selection constraints than class I PG genes.

Expression divergence of *Populus* PG genes under normal growth condition and abiotic stress

The expression patterns of *Populus* PG genes were investigated by PCR under normal growth conditions and in response to abiotic stress treatments (drought, ABA, cold and salt). We examined the expression of all 75 PG genes in seven tissues, including roots, shoots, leaves, buds, phloem, leaf abscission zones and flowers (Fig. 1b). Except for *PtPG12*, *14* and *57*, which were selectively expressed either in a specific tissue and/or in response to a specific treatment, all class II PG genes were expressed in all tissues under all growth conditions. However, substantially higher variation in expression patterns was found among the class I members than among the class II members (Fig. 1b). Among the 60 class I PG genes, 11 (*PtPG2*, *30*, *31*, *34*, *35*, *39*, *41*, *47*, *55*, *62* and *64*) were expressed in all tissues under all growth conditions, while 12 (*PtPG3*, *19*, *21*, *26*, *33*, *37*, *40*, *45*, *53*, *60*, *66* and *68*) were not expressed in any tissue or in response to any treatment applied in this study. The other 37 class I PG genes were selectively expressed either in a specific tissue and/or in response to a specific treatment. Among the selectively expressed class I PG genes, 11 (*PtPG5*, *10*, *11*, *38*, *59*, *70*, *71*, *72*, *73*, *74* and *75*) were expressed only in flower tissue, suggesting specific roles for

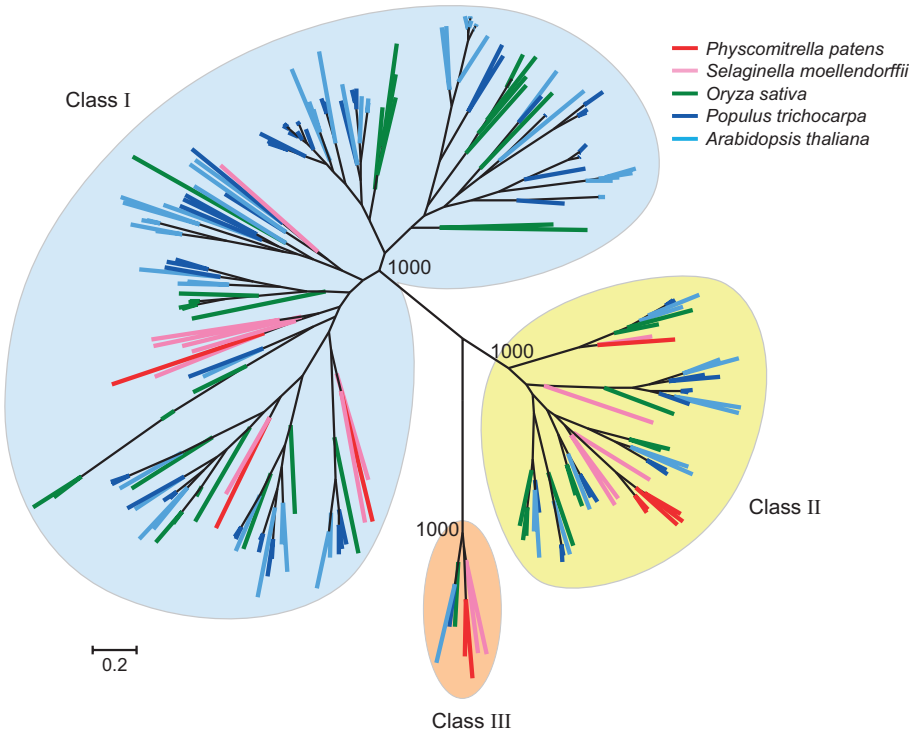

Fig. 6 Phylogenetic tree of the 212 polygalacturonases (PGs) from five land plant species. Classes I, II and III PGs are shaded blue, yellow and brown, respectively. The tree was constructed using the maximum-likelihood (ML) procedure. Numbers at the internal branches leading to the three PG classes indicate the bootstrap support from 1000 replicates.

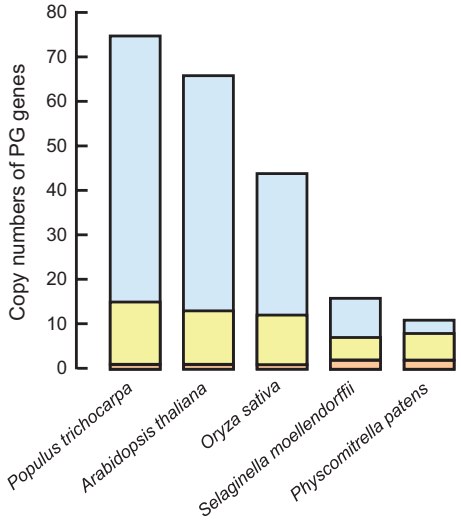

Fig. 7 Comparison of the copy numbers of polygalacturonase (PG) genes in five land plant species. The blue, yellow and brown boxes represent classes I, II and III PGs, respectively.

these genes in flower development, and two (*PtPG58* and *PtPG67*) were specifically expressed in the leaf abscission zone under salt stress, suggesting a possible function in leaf abscission under this condition. An MRPP test showed a significant difference in gene expression patterns between *Populus* classes I and II PGs (MRPP test, $P < 0.0001$).

With regard to the tandem-arrayed PG clusters I and II, marked expression divergence was observed among the members. Among the 10 PGs in cluster II, two genes (*PtPG62* and *64*) were expressed in all tissues, two (*PtPG60* and *66*) were not expressed in any tissue examined, while the other six (*PtPG58*, *59*, *61*, *63*, *65* and *67*) showed restricted tissue-specific expression patterns under normal and/or stress conditions (Fig. 1b). Among the five PG genes in cluster I, *PtPG21* was not expressed in any tissue examined, while the other four genes (*PtPG20*, *22*, *23* and *24*) showed overlapping but distinct expression profiles. For example, *PtPG20* was expressed in shoot tissue under drought stress, but *PtPG22*, *23* and *24* were not; *PtPG22* was expressed in leaf tissue under ABA treatment, while the other three genes, *PtPG20*, *23* and *24*, were not. For the tandem-arrayed PG cluster III, four genes in this cluster showed similar expression patterns, expressed only in flower tissue.

Effects of the duplication mechanism on *Populus* PG gene expression

It was considered that more closely related genes would tend to have similar expression patterns. To test this, we investigated the expression patterns of recently duplicated gene pairs. Twenty-five recently duplicated gene pairs were identified at the terminal nodes of the phylogenetic tree shown in Fig. 1(a) (Table 2). Six categories of expression patterns were observed among these duplicate gene pairs. In the first category, both duplicates were expressed in all tissues under normal growth and stress treatments (AA model in Table 2), suggesting that the duplicate genes may have similar functions. Six duplicate gene pairs (*PtPG8*/*50*, *18*/*69*, *32*/*36*, *62*/*64*, *16*/*25* and *30*/*41*) showed this expression pattern. In the second category, one copy was selectively

Table 1 Summary statistics for detection of selection using branch-specific models of PAML

Tree	Model	Estimates of parameters	$\text{Log}_e L$	$2\Delta l$	P
Tree 1 (*Populus trichocarpa*)	One ratio	$\omega = 0.1727$ for classes I and II PGs	−36156.1315		
	Two ratios	$\omega_1 = 0.1879$ for class I PGs	−36142.9111	26.4408	< 0.0001
		$\omega_0 = 0.1245$ for class II PGs			
Tree 2 (*Arabidopsis thaliana*)	One ratio	$\omega = 0.1286$ for classes I and II PGs	−33806.1316		
	Two ratios	$\omega_1 = 0.1540$ for class I PGs	−33774.6197	63.0238	< 0.0001
		$\omega_0 = 0.0619$ for class II PGs			
Tree 3 (*Oryza sativa*)	One ratio	$\omega = 0.2172$ for classes I and II PGs	−7381.8815		
	Two ratios	$\omega_1 = 0.2462$ for class I PGs	−7377.3946	8.9738	< 0.003
		$\omega_0 = 0.1613$ for class II PGs			
Tree 4 (*Selaginella moellendorffii*)	One ratio	$\omega = 0.0065$ for classes I and II PGs	−15583.3174		
	Two ratios	$\omega_1 = 0.0065$ for class I PGs	−15583.2243	0.1862	> 0.67
		$\omega_0 = 0.0120$ for class II PGs			
Tree 5 (*Physcomitrella patens*)	One ratio	$\omega = 0.1220$ for classes I and II PGs	−9626.8095		
	Two ratios	$\omega_1 = 0.0053$ for class I PGs	−9617.7324	18.1542	< 0.0001
		$\omega_0 = 0.1260$ for class II PGs			

All trees are shown in Fig. S5.

expressed either in a specific tissue and/or in response to a specific treatment, while the other was not detected in any tissue type under any growth condition (SN model in Table 2), suggesting that one duplicate gene may have become a pseudogene or evolved a new function not identified in this study. Three duplicate gene pairs (*PtPG5*/*33*, *44*/*45* and *66*/*67*) belonged to this category. In the third category, one copy of each duplicate pair was expressed in all tissues under all growth conditions, while the other was expressed only following a specific treatment and/or in a specific tissue (AS model in Table 2). Four duplicate gene pairs (*PtPG1*/*12*, *6*/*47*, *17*/*55* and *38*/*39*) showed this expression pattern. In the fourth category, both duplicates were expressed only in flower tissue (EF model in Table 2). Three duplicate gene pairs (*PtPG10*/*11*, *71*/*75* and *72*/*73*) showed this expression pattern. In the fifth category, both duplicates exhibited selective expression in the tissues examined, but the expression patterns of the two duplicates differed (SE model in Table 2). Seven duplicate gene pairs (*PtPG9*/*51*, *14*/*57*, *15*/*49*, *24*/*46*, *28*/*43*, *29*/*42* and *56*/*58*) belonged to this category. In the last category, neither duplicate was expressed in any of the tested tissues (NN model in Table 2). Two duplicate gene pairs (*PtPG19*/*68* and *37*/*40*) showed this expression pattern. Among the 25 recently duplicated gene pairs, 14 (56%) duplicate gene pairs showed divergent expression patterns, indicating that the expression patterns of closely related genes have not always been similar.

To explore the relationship between the synonymous substitution rate (K_s) and the expression profile, K_s among all pairwise comparisons of the 75 *Populus* PG genes were calculated. Expression divergences between two duplicate genes were represented by the Hamming distance. In this study, the Hamming distance between duplicate genes was defined as the number of tissues in which one duplicate gene was expressed while the other was not. For example, if one duplicate gene was expressed in all 31 tissues examined in this study, while another duplicate gene was not expressed in these tissues, then the Hamming distance between the two duplicate genes was 31; if two duplicate genes were expressed in all tissues examined, or not expressed, the Hamming distance between the two duplicate genes was 0. Linear regression

Table 2 Divergence between paralogous polygalacturonase (PG) gene pairs in *Populus*

No.	Gene1	Gene2	K_a	K_s	K_a/K_s	Gene expression
1(W)	*PtPG8*	*PtPG50*	0.0611	0.1759	0.35	AA
2(W)	*PtPG18*	*PtPG69*	0.0442	0.2471	0.18	AA
3(W)	*PtPG32*	*PtPG36*	0.0531	0.2034	0.26	AA
4(T)	*PtPG62*	*PtPG64*	0.0101	0.0247	0.41	AA
5(W)	*PtPG16*	*PtPG25*	0.0602	0.1991	0.30	AA
6(W)	*PtPG30*	*PtPG41*	0.0535	0.2638	0.20	AA
7(O)	*PtPG5*	*PtPG33*	0.2885	0.8280	0.35	SN
8(T)	*PtPG44*	*PtPG45*	0.0319	0.1241	0.26	SN
9(T)	*PtPG66*	*PtPG67*	0.0307	0.0691	0.44	SN
10(W)	*PtPG1*	*PtPG12*	0.0474	0.2084	0.23	AS
11(W)	*PtPG17*	*PtPG55*	0.1137	0.2493	0.46	AS
12(O)	*PtPG38*	*PtPG39*	0.0218	0.0984	0.22	AS
13(W)	*PtPG6*	*PtPG47*	0.0543	0.2517	0.22	AS
14(W)	*PtPG19*	*PtPG68*	0.3559	0.9746	0.37	NN
15(O)	*PtPG37*	*PtPG40*	0.0511	0.1307	0.39	NN
16(T)	*PtPG10*	*PtPG11*	0.0047	0.0053	0.88	EF
17(O)	*PtPG71*	*PtPG75*	NA	0.0109	0.00	EF
18(T)	*PtPG72*	*PtPG73*	0.0033	0.0074	0.45	EF
19(W)	*PtPG9*	*PtPG51*	0.0541	0.2282	0.24	SE
20(W)	*PtPG14*	*PtPG57*	0.1615	0.4831	0.33	SE
21(W)	*PtPG15*	*PtPG49*	0.0673	0.3432	0.20	SE
22(O)	*PtPG24*	*PtPG46*	0.0927	0.5780	0.16	SE
23(W)	*PtPG28*	*PtPG43*	0.0508	0.2384	0.21	SE
24(W)	*PtPG29*	*PtPG42*	0.0479	0.2960	0.16	SE
25(T)	*PtPG56*	*PtPG58*	0.0011	0.0035	0.32	SE

The synonymous (K_s) and nonsynonymous (K_a) substitution rates between gene pairs were calculated by the K-Estimator program (Comeron, 1999). Gene pairs created by tandem duplication (T), whole-genome duplication (W) or other duplication (O) events are indicated in the first column of the table. Gene expression patterns were categorized into six classes (see text).

analysis was carried out between the K_s value and the Hamming distance. The dot plot showed there was not significant correlation between the K_s value and the gene expression divergence ($R = 0.028$, $P > 0.145$) (Fig. S3).

Expression patterns of PG genes in *Physcomitrella*, *Selaginella* and rice

The expression patterns of *Populus* class I PG genes were distinct from those of class II PG genes. To investigate whether other land plants had similar gene expression patterns to *Populus* PGs, we examined the expression of PG genes in *Physcomitrella*, *Selaginella* and rice. The expression patterns of all 11 *Physcomitrella* and 16 *Selaginella* PG genes in different tissues were investigated using PCR (Fig. 8). The *Physcomitrella* genome contains three class I, six class II and two class III PGs. Except for one class I PG gene (*PpPG9*), which was not expressed in any of the tissues examined, the other 10 genes were expressed in all tissues examined. The expression patterns of classes I and II PG genes did not differ significantly (MRPP test, $P = 1$).

The *Selaginella* genome contains nine class I, five class II and two class III PGs. All five class II PGs were expressed in all tissues examined, while substantially greater variation in expression pattern was found among the nine class I PGs. Among these, two (*SmPG11* and *14*) were expressed in all tissues examined, three (*SmPG7*, *8* and *15*) were selectively expressed in specific tissues and four (*SmPG5, 9, 10* and *16*) were not expressed in any of the tissues examined. A MRPP test showed a significant difference in the gene expression pattern between *Selaginella* classes I and II PGs (MRPP test, $P < 0.02$).

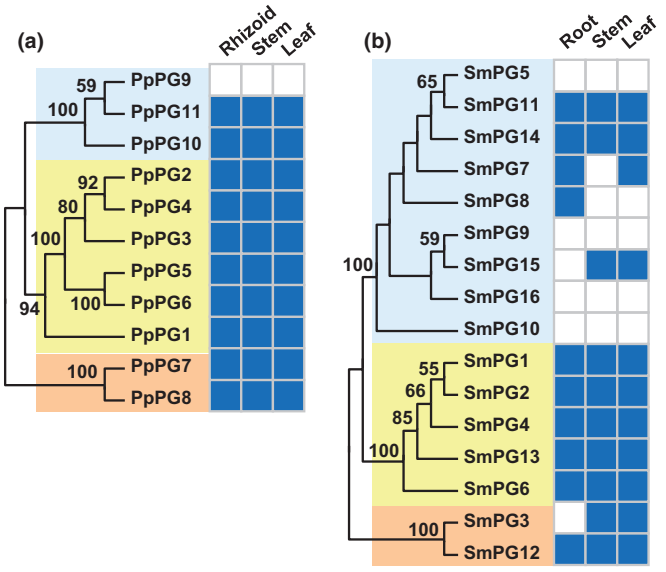

Fig. 8 Expression patterns of polygalacturonase (PG) genes in *Physcomitrella* (a) and *Selaginella* (b). Numbers on branches indicate the bootstrap percentage values calculated from 1000 replicates, and only values higher than 50% are shown. Classes I, II and III PGs are shaded blue, yellow and brown, respectively. The blue box indicates positive detection of gene expression in the corresponding tissue under normal growth conditions.

The expression patterns of rice PGs were investigated using RNA-seq data from the Michigan State University (MSU) Rice Genome Annotation (http://rice.plantbiology.msu.edu) databases (Fig. S4). With the exception of *LOC_Os05 g50960*, all rice class II PG genes were expressed in all tissues examined in this study. However, compared with class II PG genes, rice class I PGs showed extensive expression divergences. Among the 32 class I PGs, one (*LOC_Os05 g14150*) was specifically expressed in pistil tissue, two (*LOC_Os03 g11760* and *LOC_Os06 g40880*) in inflorescence tissue, five (*LOC_Os06 g35320*, *LOC_Os06 g 35370*, LOC_Os06 g35300, *LOC_Os06 g40890* and *LOC_Os 01 g33300*) in the anther, pistil and inflorescence, while five (*LOC_Os06 g16810*, *LOC_Os07 g10680*, *LOC_Os07 g10740*, *LOC_Os06 g31270* and *LOC_Os11 g14410*) were not expressed in any of tissues examined. The expression patterns of rice classes I and II PG genes differed significantly (MRPP test, $P < 0.0001$).

Subcellular localization of *Populus* PG proteins

Polygalacturonase proteins are involved in the degradation of pectin, which are considered to be localized in the plant cell walls. In this study, we first predicted the subcellular localization of *Populus* PG proteins using TargetP 1.1 Server (http://www.cbs.dtu.dk/services/TargetP). TargetP 1.1 was used to predict the subcellular location of eukaryotic proteins. The location assignment is based on the predicted presence of any of the N-terminal presequences: chloroplast transit peptide (cTP), mitochondrial targeting peptide (mTP) or secretory pathway signal peptide (SP) (Emanuelsson *et al.*, 2007). Among the 75 *Populus* PGs, 66 proteins were predicted to contain SP, indicating these proteins might be localized to the cell wall, three (PtPG1, 12 and 54) to contain mTP, one (PtPG43) to contain cTP and the other five (PtPG16, 18, 21, 25 and 69) to contain none of the three signal peptides. To examine the accuracy of the prediction of subcellular locations of *Populus* PGs, we selected nine PG proteins and investigated their subcellular localization by generating C-terminal GFP fusions and visualization by confocal microscopy after transient expression of the fusions in *N. benthamiana*. Among the nine proteins from the three PG classes, five (PtPG22, 28, 34, 38 and 47) were predicted to contain SP, two (PtPG1 and 54) to contain mTP and the other two (PtPG18 and 69) were predicted not to contain any of the three signal peptides. Confocal microscopy analysis showed that the fluorescent signal for the nine fusion proteins was only detected at the peripheral regions of the *N. benthamiana* epidermal cells, suggesting a plasma membrane and/or cell wall localization. To distinguish the plasma membrane and/or cell wall localization, the epidermal cells were plasmolyzed by treatment with 0.3 g ml^{-1} sucrose. After plasmolysis, the fluorescent signal was retained in the cell wall, as illustrated by PtPG22, 18 and 54–GFP (Fig. 9). Thus, these nine PG–GFP fusion proteins were localized to cell wall.

Discussion

Polygalacturonases were first identified over 35 yr ago and have been suggested to play important roles in the disassembly of

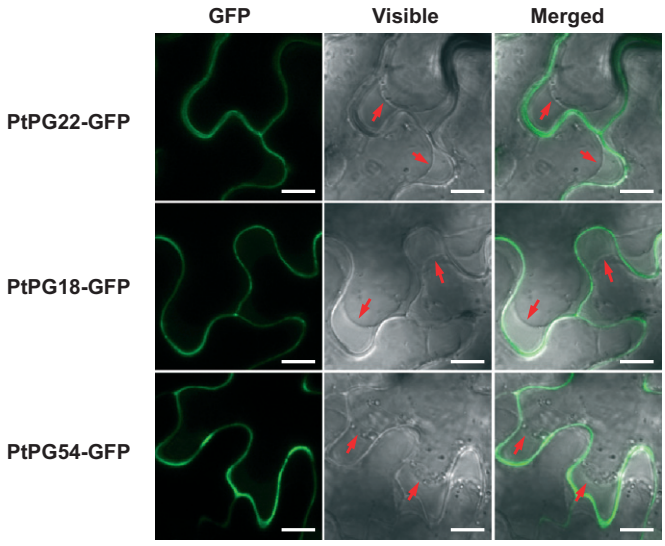

Fig. 9 Subcellular localizations of *Populus* polygalacturonase (PG)–green fluorescent protein (GFP) fusion proteins. GFP signal (green) was detected using confocal laser-scanning microscopy. The plasma membrane separated from the cell wall after plasmolysis is indicated by red arrows. Bars, 10 μm.

pectin that accompanies many stages of plant development, particularly in various cell separation processes (Hadfield & Bennett, 1998). PGs belong to a large gene family with > 44 members in the rice, *Arabidopsis* and *Populus* genomes. A previous study showed that *Arabidopsis* and rice PGs were divided into three groups (Kim *et al.*, 2006). In this study, we found that the 212 PGs from the *Physcomitrella*, *Selaginella*, rice, *Arabidopsis* and *Populus* genomes fell into three distinct groups (Fig. 6). In addition, the protein sequence identities within each group of *Populus* PGs were higher than those between groups (Fig. 2). Thus, in this study, the PG genes in the three groups were classified as three classes (classes I, II and III).

Compared with classes II and III PGs, rapid expansion of class I PG genes were observed in angiosperms (Fig. 7). Why did this rapid expansion of class I PG genes occur in angiosperms? One possible explanation is functional requirement. Class I PGs play important roles in plant developmental processes, such as organ abscission, pod and anther dehiscence, pollen grain maturation and fruit softening (Atkinson *et al.*, 2002; Fabi *et al.*, 2009). Compared with bryophytes and lycophytes, angiosperms possess more complex organ systems and structures, such as flowers and fruit. New organ systems and structures might require more PGs to maintain their biological functions. The expansion of class I PGs might provide new raw materials for specific expression of PG genes in the abscission zone of these new organs or structures. If this assumption is correct, angiosperm class I PGs may be under either relaxed selection constraints or positive selection, which allows new duplicate genes to diverge into new specific expression patterns or functions. In fact, the *Populus*, *Arabidopsis* and rice class I PG genes are under more relaxed selection constraints than the class II genes. The class I PG genes in these three species showed extensive expression divergence. In particular, some class I PG genes in these three species were expressed in specific organs, such as flowers (Figs 1b, S4) (Kim *et al.*, 2006). Thus, functional requirements could explain the rapid expansion of class I PG genes in angiosperms.

The rapid expansion of the PG gene family in angiosperm plants is largely the result of the expansion of class I PG genes. Among the 60 *Populus* class I PG genes, 20 (33%) were created by tandem duplication and eight (13%) by the salicoid whole-genome duplication event, indicating that tandem duplication was the major factor responsible for the rapid expansion of the class I PG genes in *Populus*. This is similar to other gene families, such as *Populus* glutathione S-transferase (GST) (Lan *et al.*, 2009) and late embryogenesis abundant protein family (Lan *et al.*, 2012). In addition, we found 21 PG fragments in the *Populus* genome (Table S4), indicating that a large number of PG duplicates were lost. Our previous study of the *Populus* GST supergene family also identified a large number of GST fragments in the *Populus* genome (Lan *et al.*, 2009). Thus, our data suggest that gene loss might be a general evolutionary patterns of large gene family evolution (see also Zou *et al.*, 2009). This gene loss was considered an important mechanism in the generation of genome diversity among eukaryotic species (Li *et al.*, 2005a).

Although a large number of PG genes were lost, many duplicates have been retained in the *Populus* genome, bringing us to a central issue in the evolution of duplicate genes: why have so many duplicate genes been retained? At present, the neofunctionalization or subfunctionalization models are frequently invoked to explain the retention of duplicate genes (Li *et al.*, 2005b). In the neofunctionalization model, one duplicate copy accumulates beneficial mutations and acquires a new function, whereas another duplicate copy retains the original function. In the subfunctionalization model, each duplicate copy partitions the ancestral gene function. Neofunctionalization or subfunctionalization of duplicate genes is associated with the processes of tissue expression divergence (Ganko *et al.*, 2007). In addition, the expansions of large gene family are often associated with high levels of tissue-specific expression divergence (Lan *et al.*, 2009; Huerta-Cepas *et al.*, 2011). Among the 60 *Populus* class I PGs, 11 were specifically expressed in flowers. The flower-specific expressions of class I PG genes were also observed in rice and *Arabidopsis* (Fig. S4) (Kim *et al.*, 2006). Compared with bryophytes and lycophytes, flowers are a key innovation in angiosperms. The class I PG genes expressed specifically in flowers might acquire new functions related to flower development. Thus, the evolutionary fates of these class I PG genes might be neofunctionalization and could contribute to the retention of these class I PG genes. Among the 25 recently duplicated gene pairs, the six duplicate gene pairs (*PtPG9/51*, *14/57*, *15/49*, *24/46*, *28/43* and *29/42*) have overlapping but distinct expression profiles, indicating typical subfunctionalization. In the tandem-arrayed PG clusters I and II, we found that some PG genes were expressed in all tissues, some had restricted tissue-specific expression patterns under normal and/or stress conditions, while others were not expressed in any tissue examined. This markedly divergent expression pattern indicates that multiple

evolutionary fates occurred among the PG genes in clusters I and II, which may contribute to the retention of these duplicate genes.

It has also been reported that expression divergence between duplicate genes is significantly correlated with their K_s value: when the K_s is relatively small, the duplicated genes tend to have more similar expression patterns (Gu *et al.*, 2002). However, no such correlation was identified in the *Populus* PG gene family (Fig. S3). The *Arabidopsis* PG gene family did not show this correlation either (Kim *et al.*, 2006). The lack of this correlation might be attributable to the small number of genes examined, resulting in a lack of statistical power. However, this correlation was also not observed upon examination of the expression divergences of 503 duplicate gene pairs in *Arabidopsis* (Ganko *et al.*, 2007). Thus, there is a need to further study the relationship between expression divergence and synonymous substitutions of duplicate genes.

The duplicate genes created by large segmental duplication tend to maintain similar expression patterns, while those of genes created by nonsegmental duplications rapidly diverge (Ganko *et al.*, 2007). One possible mechanism of this pattern is that large-scale segmental duplication results in the duplication of multiple genes together with their promoter and/or enhancer elements (Casneuf *et al.*, 2006), while nonsegmental duplications, such as tandem duplication, may disrupt the regulatory regions of target genes, resulting in large variations in their expression responses. In this study, 14 segmental duplicate gene pairs and six tandem duplicate gene pairs were identified in the *Populus* PG gene family. Among the 14 segmental duplicate gene pairs, eight (57%) showed expression divergences between two duplicate genes. Among the six tandem duplicate gene pairs, three (50%) shared similar expression patterns. Thus, this study did not observe the significant effects resulting from the duplication mechanism on gene expression divergence.

Casneuf *et al.* (2006) reported a strong bias in the divergence of gene expression towards gene function: genes that are involved in signal transduction, hydrolase activity and response to external stimuli appear to have diverged very quickly after duplication, while those involved in nucleic acid and protein metabolisms appear to have diverged slowly after duplication. Because class I PGs have typical hydrolase activity, according to the report of Casneuf *et al.* (2006), the expression of class I PG duplicate genes should diversify after gene duplications. Indeed, we found rapid divergence in the expression of class I PG genes in *Populus* and rice. However, among the 36 class II PG genes from *Populus*, rice, *Selaginella* and *Physcomitrella*, except for five PG genes, the other 31 PG genes were expressed in all the tissues examined (Figs 1b, 8, S4), indicating that the function of class II PGs might be conserved in land plants. The maintenance of similar expression patterns in class II PGs in land plants might be the result of functional requirements.

By exploring the currently available genome information in land plants, our comparative evolutionary analysis coupled with expression assays provided new insights into the evolution of the PG gene family in land plants. The expansion of PG genes seems to be correlated with the evolution of increasingly complex organs in plants. Further sampling of ferns and gymnosperms would help us to achieve a better understanding of whether the expansion of the PG genes predated the split of gymnosperms and angiosperms, and of the functional significance of PG gene family evolution in plants.

Acknowledgements

This study was supported by grants from the National Basic Research Program of China (2009CB119104) and Vetenskapsrådet, Sweden.

References

Atkinson RG, Schroder R, Hallett IC, Cohen D, MacRae EA. 2002. Overexpression of polygalacturonase in transgenic apple trees leads to a range of novel phenotypes involving changes in cell adhesion. *Plant Physiology* **129**: 122–133.

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, dePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, *et al.* 2011. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**: 960–963.

Bosch M, Hepler PK. 2005. Pectin methylesterases and pectin dynamics in pollen tubes. *Plant Cell* **17**: 3219–3226.

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

Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biology* **7**: R13.

Comeron JM. 1999. K-estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. *Bioinformatics* **15**: 763–764.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using targetP, signalP and related tools. *Nature Protocols* **2**: 953–971.

Fabi JP, Cordenunsi BR, Seymour GB, Lajolo FM, do Nascimento JR. 2009. Molecular cloning and characterization of a ripening-induced polygalacturonase related to papaya fruit softening. *Plant Physiology and Biochemistry* **47**: 1075–1081.

Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in Arabidopsis. *Molecular Biology and Evolution* **24**: 2298–2309.

Gu Z, Nicolae D, Lu HH, Li WH. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics* **18**: 609–613.

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

Hadfield KA, Bennett AB. 1998. Polygalacturonases: many genes in search of a function. *Plant Physiology* **117**: 337–343.

Hall TA. 1999. Bioedit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt. *Nucleic Acids Symposium Series* **41**: 95–98.

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldon T. 2011. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. *Briefings in Bioinformatics* **12**: 442–448.

Jamet E, Albenne C, Boudart G, Irshad M, Canut H, Pont-Lezica R. 2008. Recent advances in plant cell wall proteomics. *Proteomics* **8**: 893–908.

Jansson S, Douglas CJ. 2007. *Populus*: a model system for plant biology. *Annual Review of Plant Biology* **58**: 435–458.

Kim J, Shiu SH, Thoma S, Li WH, Patterson SE. 2006. Patterns of expansion and expression divergence in the plant polygalacturonase gene family. *Genome Biology* **7**: R87.

Lan T, Gao J, Zeng QY. 2012. Genome-wide analysis of the LEA (late embryogenesis abundant) protein gene family in *Populus trichocarpa*. *Tree Genetics and Genomes*. doi:10.1007/s11295-11012-10551-11292.

Lan T, Yang ZL, Yang X, Liu YJ, Wang XR, Zeng QY. 2009. Extensive functional diversification of the *Populus* glutathione S-transferase supergene family. *Plant Cell* **21**: 3749–3766.

Li HM, Rotter D, Bonos SA, Meyer WA, Belanger FC. 2005a. Identification of a gene in the process of being lost from the genus *Agrostis*. *Plant Physiology* **138**: 2386–2395.

Li WH, Yang J, Gu X. 2005b. Expression divergence between duplicate genes. *Trends in Genetics* **21**: 602–607.

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. 2003. Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *Plant Cell* **15**: 809–834.

Mohnen D. 2008. Pectin structure and biosynthesis. *Current Opinion in Plant Biology* **11**: 266–277.

Munoz JA, Coronado C, Perez-Hormaeche J, Kondorosi A, Ratet P, Palomares AJ. 1998. *MsPG3*, a *Medicago sativa* polygalacturonase gene expressed during the alfalfa–*Rhizobium meliloti* interaction. *Proceedings of the National Academy of Sciences, USA* **95**: 9687–9692.

Ogawa M, Kay P, Wilson S, Swain SM. 2009. ARABIDOPSIS DEHISCENCE ZONE POLYGALACTURONASE1 (ADPG1), ADPG2, and QUARTET2 are polygalacturonases required for cell separation during reproductive development in *Arabidopsis*. *Plant Cell* **21**: 216–233.

Orozco-Cardenas ML, Ryan CA. 2003. Polygalacturonase β-subunit antisense gene expression in tomato plants leads to a progressive enhanced wound response and necrosis in leaves and abscission of developing flowers. *Plant Physiology* **133**: 693–701.

Popper ZA. 2008. Evolution and diversity of green plant cell walls. *Current Opinion in Plant Biology* **11**: 286–292.

Prieto-Alcedo M, Veiga-Crespo P, Poza M, Coronado C, Zarra I, Villa T. 2011. Expression of a yeast polygalacturonase gene in *Arabidopsis thaliana*. *Biologia Plantarum* **55**: 349–352.

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, *et al.* 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319**: 64–69.

Sparkes IA, Runions J, Kearns A, Hawes C. 2006. Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. *Nature Protocols* **1**: 2019–2025.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–1604.

Wei H, Xu Q, Taylor LE, Baker JO, Tucker MP, Ding SY. 2009. Natural paradigms of plant cell wall degradation. *Current Opinion in Biotechnology* **20**: 330–338.

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

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**: 32–43.

Yin Y, Chen H, Hahn MG, Mohnen D, Xu Y. 2010. Evolution and function of the plant cell wall synthesis-related glycosyltransferase family 8. *Plant Physiology* **153**: 1729–1746.

Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH. 2009. Evolutionary and expression signatures of pseudogenes in *Arabidopsis* and rice. *Plant Physiology* **151**: 3–15.

Supporting Information

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

Fig. S1 Modified pCAMBIA1302 vector.

Fig. S2 Hypothetical evolutionary histories of the *Populus* PGs in clusters I and II.

Fig. S3 The relationship between the K_s value and the expression divergence among *Populus* PG genes.

Fig. S4 Phylogenetic relationship among rice PGs and their expression patterns.

Fig. S5 Phylogenetic trees used for molecular evolution analyses.

Table S1 Primers used to detect the expression of PG genes

Table S2 Primers used to construct the *Populus* PG subcellular localization vectors

Table S3 Full-length PG genes identified from the *Populus* genome

Table S4 PG fragments identified from the *Populus* genome

Table S5 Full-length PG protein sequences

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.