The Plant Journal (2005) **44**, 409–419 doi: 10.1111/j.1365-313X.2005.02540.x

Use of genomic history to improve phylogeny and understanding of births and deaths in a gene family

Javier Sampedro[1], Yi Lee[2], Robert E. Carey[1], Claude dePamphilis[1] and Daniel J. Cosgrove[1,*]
[1]*Department of Biology, 208 Mueller Lab, Pennsylvania State University, University Park, PA 16802, USA, and*
[2]*Department of Industrial Plant, Chungbuk National University, 12 Gaesin-dong Hungduk-ku, Cheongju 361-763, Korea*

Received 15 July 2005; accepted 25 July 2005.
*For correspondence (fax +1 814 865 9131; e-mail dcosgrove@psu.edu).

Summary

Polyploidy events have played an important role in the evolution of angiosperm genomes. Here, we demonstrate how genomic histories can increase phylogenetic resolution in a gene family, specifically the expansin superfamily of cell wall proteins. There are 36 expansins in Arabidopsis and 58 in rice. Traditional sequence-based phylogenetic trees yield poor resolution below the family level. To improve upon these analyses, we searched for gene colinearity (microsynteny) between Arabidopsis and rice genomic segments containing expansin genes. Multiple rounds of genome duplication and extensive gene loss have obscured synteny. However, by simultaneously aligning groups of up to 10 potentially orthologous segments from the two species, we traced the history of 49 out of 63 expansin-containing segments back to the ancestor of monocots and eudicots. Our results indicate that this ancestor had 15–17 expansin genes, each ancestral to an extant clade. Some clades have strikingly different growth patterns in the rice and Arabidopsis lineages, with more than half of all rice expansins arising from two ancestral genes. Segmental duplications, most of them part of polyploidy events, account for 12 out of 21 new expansin genes in Arabidopsis and 16 out of 44 in rice. Tandem duplications explain most of the rest. We were also able to estimate a minimum of 28 gene deaths in the Arabidopsis lineage and nine in rice. This analysis greatly clarifies expansin evolution since the last common ancestor of monocots and eudicots and the method should be broadly applicable to many other gene families.

Keywords: expansins, comparative genomics of rice and Arabidopsis, gene family phylogeny, gene birth and death, expansin evolution, genome duplication.

Introduction

Large duplicated chromosomal segments within the Arabidopsis and rice genomes have been taken as evidence of multiple polyploidy events, up to three in Arabidopsis (Bowers *et al.*, 2003; Simillion *et al.*, 2002) and at least one in rice (Paterson *et al.*, 2004). Limited synteny between Arabidopsis and rice has also been reported (Devos *et al.*, 1999; Salse *et al.*, 2002; Simillion *et al.*, 2004; Vandepoele *et al.*, 2002). This complex genomic history has profound implications for our understanding of gene family evolution (Cannon *et al.*, 2004), implications that have yet to be fully exploited in phylogenetic analysis. By linking gene duplications to polyploidy events or other segmental duplications, we can resolve uncertainties that remain after traditional sequence-based phylogenetic analysis and in addition shed light on the processes of gene birth and death. This enriched understanding of the evolution of gene families can also provide a better framework for an analysis of gene functions and their evolution through time. This methodology should be widely applicable because the Arabidopsis genome includes at least 780 gene families with more than five members and there are at least 824 in rice (Horan *et al.*, 2005). Furthermore, a recent estimate attributes to genome duplication events 59% of the new and surviving genes generated in the Arabidopsis lineage since the oldest detected polyploidy (Maere *et al.*, 2005).

In this study, we examined the evolution of the expansin superfamily in rice and Arabidopsis. Expansins were originally discovered as non-enzymic proteins that promote cell wall loosening (Li *et al.*, 1993; McQueen-Mason *et al.*, 1992) during cell growth and other developmental processes such

as fruit ripening and pollination (Cosgrove, 2000). On the basis of sequence divergence, four expansin families are currently recognized, named α-expansin (EXPA), β-expansin (EXPB), expansin-like A (EXLA) and expansin-like B (EXLB) (Kende *et al.*, 2004). Members of the EXPA and EXPB families are known to have wall-loosening activity (Cho and Kende, 1997; Cosgrove *et al.*, 1997; McQueen-Mason *et al.*, 1992), whereas the other two families have been identified only from sequence homology, without protein function analysis (Lee *et al.*, 2001; Li *et al.*, 2002). Expansins often exhibit cell-specific expression patterns (Cho and Cosgrove, 2000, 2002; Cho and Kende, 1998; Gray-Mitsumune *et al.*, 2004; Zenoni *et al.*, 2004) and such specific expression suggests that the evolutionary diversification of plant cell types may have involved a parallel duplication and specialization of expansin genes (Cosgrove, 2000). The independent growth and evolution of the expansin superfamily in the monocot and eudicot lineages presents an obstacle for attempts to extrapolate gene function from one lineage to the other, but understanding when and how this growth took place is a necessary first step.

In this study, we show how gene colinearity (microsynteny) within and between the genomes of Arabidopsis and rice provides a much clearer picture of the family evolution than is possible by sequence-based phylogenetic analysis alone. An integrated analysis allowed us to reconstruct the expansin superfamily as it likely existed in the last common ancestor of monocots and eudicots and to propose an account of gene births and deaths along the separate lineages of rice and Arabidopsis.

Results

A complex superfamily

The published genome of *Arabidopsis thaliana* ecotype Columbia (The Arabidopsis Genome Initiative, 2000) includes 36 genes belonging to the expansin superfamily. For rice, we limited our analysis to the map-based sequence of *Oryza sativa* L. cv. Nipponbare (*Japonica* cultivar group; Sasaki and Burr, 2000). This genome contains 58 expansin genes, including two pairs of identical genes (*OsEXPA23a*/*OsEXPA23b* and *OsEXPB1a*/*OsEXPB1b*).

In rice and Arabidopsis, the four expansin families are of comparable size, with the notable exception that EXPB genes are three times more numerous in rice (19 versus six genes). EXPA is the largest family, 34 genes in rice and 26 in Arabidopsis, while EXLB is the smallest, with a single gene in each species. The EXLA family has three members in Arabidopsis and four in rice. For individual gene annotations, see the expansin web site at http://www.bio.psu.edu/expansins.

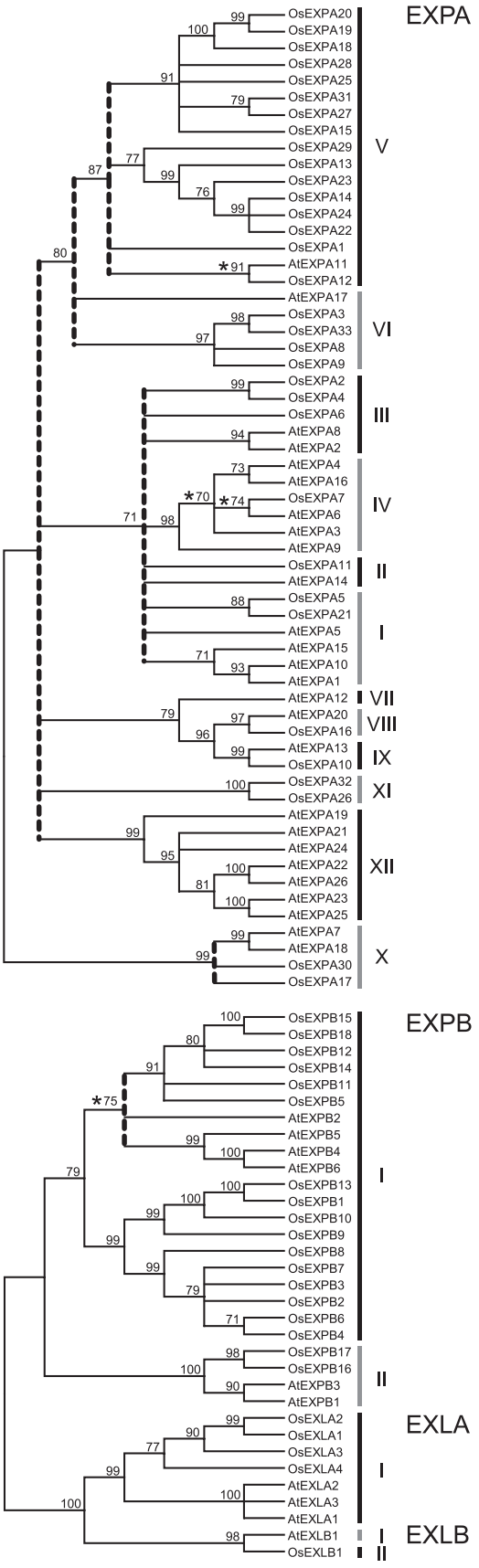
Combining rice and Arabidopsis protein sequences, we obtained six phylogenetic trees using neighbor-joining, parsimony and Bayesian analyses (Figure S1). EXPA sequences were analyzed independently because of the presence of family-specific insertions and deletions. We found it impossible to define orthologous groups of genes with any certainty, due to poor support at key nodes (see example in Figure 1) and contradictory results using different methods or a different set of sequences. A recent analysis (Li *et al.*, 2003) noted similar difficulties.

Position-based phylogeny

In addition to sequence information, the assembled genomes of Arabidopsis and rice provide positional information. The 36 expansin genes of Arabidopsis are found in 29 non-adjacent genomic locations (three of these locations contain tandems). In rice, the 58 expansins appear in 36 locations, including 10 tandems. As described below, we used positional information, in combination with dates of segment divergence, to refine the phylogeny of rice and Arabidopsis expansins (summarized in Figure 2).

For Arabidopsis, we consulted two analyses of segmental duplications that identify the individual genes involved in these events. The first study dated the duplications by calculating *Ks* (synonymous substitution rate) between gene pairs (Simillion *et al.*, 2002). A second study dated them in relation to speciation events (Bowers *et al.*, 2003). In both cases, the results were interpreted as evidence for three rounds of polyploidy. A recent analysis of duplication rates also supports this hypothesis (Maere *et al.*, 2005). In this work, we follow Bowers *et al.* (2003) in referring to these events as α,β and γ, with α being the most recent and γ being the oldest. A polyploidy event in rice, predating the divergence of the grasses, has also been proposed (Paterson *et al.*, 2004) and we will refer to it as event ρ. These three genome-wide studies allowed us to identify paralogous relations between different expansin-containing segments (that is, segments created by duplications within a genome, without speciation involved). Segmental duplications in Arabidopsis can be linked to 12 surviving expansin gene duplications and rice event ρ seem seems responsible for another 10 (Figure 2).

By microsynteny analysis, we were able to go further and assemble 49 of the expansin-containing genomic segments from rice and Arabidopsis into 12 groups (see Experimental procedures). We propose that all the segments within a group descended from a single expansin-containing segment in the genome of the last common ancestor of monocots and eudicots, and thus refer to them as orthologous groups. They all contain at least one rice and one Arabidopsis segment with an expansin gene. A simplified synteny diagram for one of them is shown as Figure 3(a) (for full diagrams with gene identifications and BLASTP results see Figure S2). A total of 68 expansins (23 from Arabidopsis, 45 from rice) are present in the 12 orthologous groups of



segments (Figure 2). Rice–Arabidopsis microsynteny also allowed us to identify seven rice segmental duplications not previously described.

Each orthologous group of segments includes between four and 43 orthologous groups of genes (including expansins) with representation in both species (average 20). These groups are shown connected by blue lines in Figure 3(a). Most of them (68–98% in different segment groups) include an Arabidopsis gene that is the 'best hit' in a BLASTP search of the entire Arabidopsis genome for one of the rice genes in the same orthologous group (yellow symbols in Figure 3a). Searches were done with protein sequences (see Experimental procedures). These results support the orthology of the segment groups used in this study. The smallest group of orthologous segments has just four groups of orthologous genes (three including best hits), but it contains an Arabidopsis segment and a rice segment whose one-to-one synteny had been previously shown to be statistically significant (Salse *et al.*, 2002). The second smallest group has eight genes and seven best hits.

Alternative scenarios

We used segmental duplication dates from the literature (Bowers *et al.*, 2003; Paterson *et al.*, 2004; Simillion *et al.*, 2002) to construct cladograms for each orthologous group of segments. An example can be seen on the right side of Figure 3(a) (all cladograms can be found in Figure S2; for the duplication dates on which they are based see Supplementary Text). Both published analyses of the Arabidopsis genome are in general agreement as to the relative dating of individual segmental duplication events. The separation in time between the polyploidy events in the Arabidopsis lineage is also large enough that the assignment of individual segmental duplications to each of them is mostly straightforward. However, some duplicated segments in Arabidopsis appear in tandem and seem to be the result of small-scale events independent of whole genome duplications (yellow dots in Figure 2). We used expansin phylogenetic trees and parsimony considerations in addition to segmental duplication dates from the literature to establish their relative position with respect to the three genome duplications. In rice, all segmental duplications that are not linked to event ρ seem to be older, according to expansin phylogenetic trees.

Figure 1. Neighbor-joining cladogram of Arabidopsis and rice expansins. Clades of orthologous genes, as determined in the integrated trees, are also indicated on the right with alternating black and gray bars and numbered as in Figure 2. Family names are shown next to the first gene of the family. Bootstrap values are shown above well-supported nodes. Thick dashed lines are poorly resolved areas that affect orthology. Branches with support below 70 have been collapsed. An asterisk indicates branches that were rejected in the integrated trees and that affect orthology.

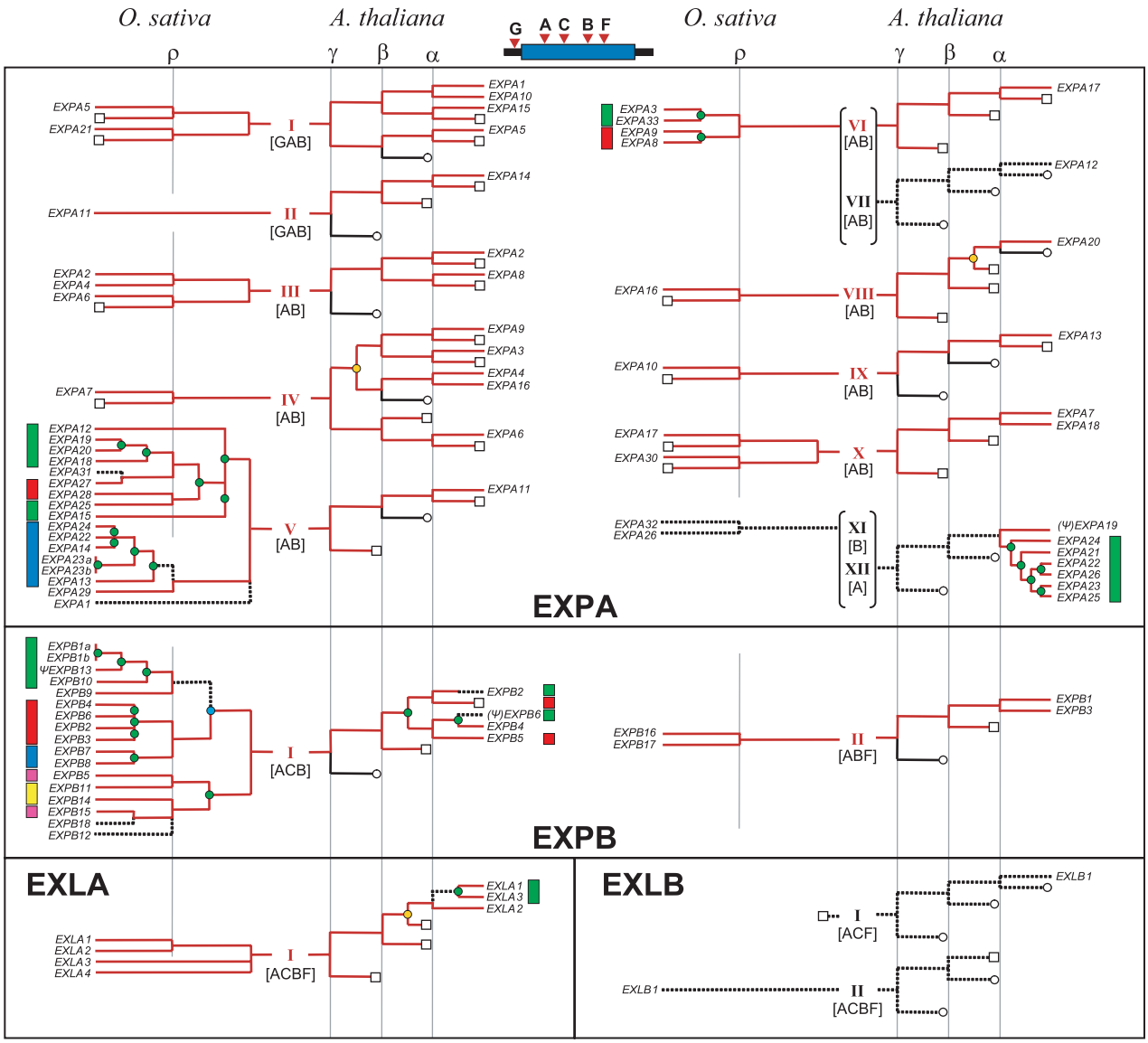


Figure 2. Integrated cladograms for the 17 clades of the expansin superfamily.
Clades, identified by roman numerals, represent putative independent lineages in the last ancestor of rice (left) and Arabidopsis (right). Brackets connect clades whose independence is uncertain. Intron pattern for the ancestral gene is shown below each clade. Intron positions are indicated at the top. Vertical gray lines indicate proposed polyploidy events. Rice event ρ is discontinuous due to its incomplete characterization. Continuous red line connects genes found in the same or homologous genomic locations; black dashed lines are used for genes in non-homologous locations. Green circles are for tandem duplications, blue for short range translocation, yellow for segmental duplication in tandem. For each clade, genes in tandem are indicated with boxes of the same color next to their names. White boxes are observed gene deaths, white circles on black branches are assumed deaths (see text). Collapsed branches have bootstrap values below 75 in clade trees. From the presence of the EXLB-II descendent in the Fabaceae (see Supplementary Text), we infer that this clade disappeared from the Arabidopsis lineage between events α and β (Bowers *et al.*, 2003).

Finally, in order to construct these cladograms it is necessary to determine the dating of the polyploidy events in relation to the divergence of the rice and Arabidopsis lineages. Phylogenetic trees of duplicated genes have been interpreted as indicating that event γ is shared between monocots and eudicots (Bowers *et al.*, 2003; Chapman *et al.*, 2004). We believe, on the other hand, that the evidence presented in these studies is not conclusive on this point and does not exclude the possibility that event γ happened in the Arabidopsis lineage after divergence from rice, where it has not yet been detected (see Discussion). Because this alternative hypothesis agrees better with the pattern of gene losses observed in the studied segments and also produces a more parsimonious tree for the expansins, we have adopted it for the segmental cladograms in Figure 3(a) and Figure S2. However, we have also

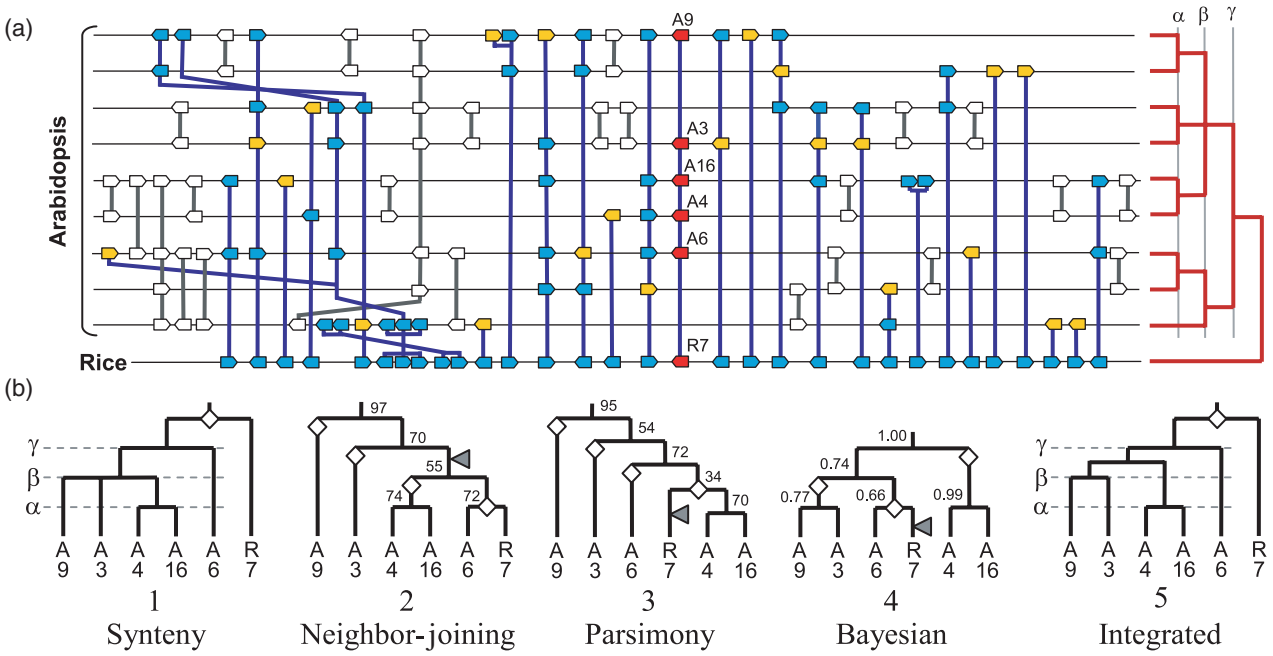


Figure 3. Microsynteny and sequence-based trees.
(a) Simplified synteny diagram for the orthologous group of genomic segments for clade EXPA-IV. Pentagons represent genes and their orientation. Distances are not to scale. Blue lines connect orthologous groups of genes. Yellow genes denote the 'best hit' (closest homolog) in the entire Arabidopsis genome for the connected rice gene. Selected Arabidopsis paralogs (in white connected by gray lines) are included for segment alignment. Expansins (red) are identified by species initial and EXPA gene number. A cladogram to the right explains the duplication history of the segments, with polyploidy events as gray lines. The triple node for event β is due to a small-scale segmental duplication in tandem, close in time to the polyploidy event (the segments that include expansins A3 and A4 are contiguous).
(b) Phylogenetic trees based on synteny (1), different phylogenetic methods (2, 3, 4) or a combination of both (5). Bootstrap or posterior probability values are shown above the nodes. Polyploidy events are indicated by dashed lines. Independent lineages in the last ancestor of monocots and eudicots are identified by diamonds. A gray triangle indicates the position of the closest *Pinus* protein when added to the tree.

considered the implications of a shared event γ. We explore below the consequences of this hypothesis for expansin phylogeny as well as for the estimated number of gene births and deaths.

Segmental cladograms are expected to parallel the cladograms of the individual genes included in them (see Discussion). With this assumption, we constructed position-based cladograms for the 12 groups of expansin genes in orthologous segments (Figure 2). A similar exercise was done assuming that event γ happened in the common lineage of monocots and eudicots (Figure S3).

A practical case

Figure 3(a) shows a simplified synteny diagram for an orthologous group of genomic segments, one from rice and nine from Arabidopsis. This group includes one rice expansin (*OsEXPA7*, abbreviated as *R7*) and five from Arabidopsis (*A3*, *A4*, *A6*, *A9* and *A6*). A total of 28 genes (or tandems of related genes) from the rice segment have putative orthologs in one or several of the Arabidopsis segments (23 best hits are shown in yellow). The cladogram for these segments is shown to the right and an expansin cladogram can be deduced directly from it (tree 1 in Figure 3b).

In contrast, sequence-based phylogenetic analyses of this same group of expansin genes yielded three different results (trees 2, 3 and 4 in Figure 3b), none of which agrees with the synteny tree (whatever the timing of event γ). To make these sequence-based phylogenies compatible with the synteny tree would require the existence of a very ancient gene tandem and an unlikely sequence of gene losses. Neighbor-joining and parsimony trees are close to each other, but they are incompatible with the Bayesian tree. All of the sequence-based phylogenies require more independent lineages in the ancestor of monocots and eudicots (diamonds in Figure 3b), lineages that are not supported by extant descendents in rice.

In this case, the main difference between position and sequence-based analyses concerns the correct rooting of the group. Trees 2 and 4 (and tree 3, but for a poorly supported branch) would be identical to the synteny tree if the root were moved to the rice branch. It is noteworthy that protein distances between R7 and the group (A3, A4, A16 and A6) are the smallest in the entire superfamily for interspecific pairs (R7/A9 is the eighth smallest). Surprisingly, when a

gymnosperm gene (The Institute for Genomic Research, TIGR gene index no. TC46521 from *Pinus*) is included in the sequence trees; it groups with rice in trees 3 and 4 (triangle in Figure 3b). This topology would thus require some of the closest Arabidopsis/rice homologs to be paralogs that predate the angiosperm–gymnosperm split, which is twice as old as that of monocots and eudicots, an unlikely proposition. A simpler hypothesis is found in the synteny tree. The rooting problem in trees 2–4 could be due to unequal rates of evolution and long-branch attraction (Felsenstein, 1978). Once we decided to accept the synteny tree, we studied the three alternative topologies for the apparent triple node linked to event β. We adopted the solution shown in tree 5, with the tandem segmental duplication happening before the polyploidy event, as most compatible with sequence-based trees (see Supplementary Text).

The other position-based cladograms were merged with sequence-based phylogenetic trees in a similar way (see Supplementary Text for detailed explanations), with the end result shown in Figure 2. To resolve branches not linked to segmental duplications, DNA trees were created for each orthologous group of expansins (Figure S4). In cases of conflict, preference was given to positional information and the most parsimonious solution was always chosen. A few branches well supported by sequence-based trees were ignored in the integrated solution (asterisks in Figure 1; see also Figure S1). All these cases showed suspected rooting problems similar to the one just described (see Supplementary Text). Position-based cladograms were also useful in providing independent confirmation for topologies that were suggested by sequence-based trees but had low support, or where different trees contradicted each other. In a couple of cases, they also resolved orthologous groups missed by all the sequence-based trees (see below). An alternative set of cladograms under the assumption that event γ happened in the common lineage of Arabidopsis and rice is provided as Figure S3.

The integrated tree can be viewed as the most parsimonious hypothesis with respect to the number of orthologous groups, gene births and deaths. It incorporates new information about topologies and branch lengths that is independent of expansin sequences, makes testable predictions and alerts us to problematic nodes that require further study. In the end, deciding in particular cases between sequence-based and position-based topologies is a matter of judgment and should be seen as a provisional solution. An increased taxon sampling in the problematic areas could eventually help to reconcile the contradicting topologies.

A fresh view of the superfamily

Using our integrated cladograms, the four families can be divided into 17 orthologous clades, which we have designated with roman numerals (Figure 2). Each clade contains all Arabidopsis and rice expansins that descend from the same gene in their last common ancestor. A previous attempt at dividing the EXPA family (Link and Cosgrove, 1998), while limited to the few sequences known at that time, is nonetheless in general agreement with our results. Subgroups A, B, C and D in this classification correspond to clades IV, III, I and V, respectively.

According to our analysis, the last common ancestor of monocots and eudicots had 15 to 17 expansin genes (10–12 EXPA, 2 EXPB, 1 EXLA and 2 EXLB). The uncertainty is due to two cases where phylogenetic trees suggest (although with poor support) that a pair of clades without synteny might actually be a single orthologous group (indicated by brackets in Figure 2; see Supplementary Text for details). If this were true, gene movement or severe loss of flanking genes could account for the lack of synteny. Analyses of additional genomes might resolve this uncertainty. It is also possible that additional ancestral genes existed but were lost in both lineages. However, the small number of unilateral clade losses argues against this view. The rice lineage seems to have lost at most three clades unilaterally, while Arabidopsis only one or two. Assuming that clade losses are random, the likelihood of many double losses is very low. New genomic sequences from early branching eudicots and monocots would allow this assumption to be tested more thoroughly. Finally, two more EXPA genes would be required in the last common ancestor of rice and Arabidopsis if event γ had already happened by then.

Another conclusion we can draw from this analysis is that most expansin genes have not moved from their genomic neighborhood since the separation of the Arabidopsis and rice lineages. Only eight translocation events suffice to explain all the cases where synteny was not detected. Moreover, microsynteny allows us to determine which gene locations are ancestral. We can say, for example, that a tandem of two Arabidopsis genes (*AtEXPB6*/*AtEXPB2*) has moved recently from the neighborhood of *AtEXPB4* to a new location in chromosome I (see Supplementary Text). It is notable that translocations only occurred in clades with tandem duplications.

Introns

An improved tree topology is also useful for understanding intron evolution. We reconstructed eight different intron patterns in the 17 ancestral genes (Figure 2). To explain present-day intron patterns, it is necessary to assume at least five independent events of intron loss (some involving several introns) in the Arabidopsis lineage and a minimum of 18 losses for the rice lineage. We have detected only two intron gains, both in *OsEXPB12* (Figure S5). Thus, intron losses have greatly outnumbered intron gains in this family since the rice and Arabidopsis lineages diverged.

Gene births

The last common ancestor of monocots and eudicots could have lived as recently as 140–170 Ma (Leebens-Mack *et al.*, 2005; Sanderson *et al.*, 2004). Since then, the size of the expansin superfamily appears to have doubled in the Arabidopsis lineage and more than tripled in rice. In Arabidopsis, at least 12 out of 21 new and surviving genes appeared through segmental duplications, all but one probably in the course of a genome doubling. In rice, segmental duplications explain 17 of the 44 new and surviving genes and event ρ seems responsible for 10 of those. At least six segmental duplications appear to predate this event, pointing to the possibility of an older genome duplication in the rice lineage.

At least 20 surviving expansin genes arose through tandem duplications in rice, compared with eight in Arabidopsis. This is in line with the relative deficit of recent tandems in Arabidopsis when compared with rice or other plants (Blanc and Wolfe, 2004b). Furthermore, tandem duplications are concentrated in just five clades. In contrast, segmental duplications have increased gene numbers in at least 10 of the 17 clades. The massive and asymmetrical growth of clades EXPA-V and EXPB-I accounts for most of the extant expansins in rice and may be related to the evolution of a distinctive cell wall composition in grasses (Carpita, 1996). It seems clear from our analysis that the growth of these clades involved both tandem and segmental duplications and that it was already well under way before event ρ: that is, before the divergence of the cereal grasses (Paterson *et al.*, 2004).

Gene deaths

In our analysis, we distinguish two classes of inferred gene deaths: 'observed' deaths (inferred from segments in the orthologous groups lacking expansin genes) and 'assumed' deaths (inferred from the assumption of full genome duplications).

We count 'observed' deaths from genomic segments that descended from an expansin-containing segment, but that no longer contain an expansin gene. When phylogenetic trees exclude the possibility of translocation, we can safely conclude that an expansin gene once existed there and later disappeared. Observed deaths are indicated by empty boxes in Figure 2 (see also Figure S2). In some cases, two or more empty paralogous segments can be explained by a single gene death in an ancestral segment.

In the Arabidopsis branch of clade EXPB-I, a tandem duplication predates event α (Figure 2). The absence of the expected duplicate of *AtEXPB2* next to *AtEXPB5* is taken as evidence of another gene death. Two similar cases could be argued in rice for clade EXPA-V, but due to uncertain phylogenies we have excluded them from our estimates. We conclude that a minimum of 28 expansin genes were lost in this way in the Arabidopsis lineage and nine in the rice lineage. In most cases, the expansin genes have disappeared without leaving a trace. In a single case, a small gene fragment can still be identified in one of the empty segments. It dates to event α and is 78% identical, over 212 bp, to the end of *AtEXPA17* (Figure S2 and Supplementary Text).

In addition to these 'observed' deaths, the assumption of whole genome duplications lets us infer additional deaths, even if we cannot identify the empty paralogous segments, which may have disappeared in large deletions. Event α requires at least three such 'assumed' deaths (marked as circles in Figure 2). With event β, the number increases to 11. If event γ was also a genome duplication (the evidence is weaker), nine more deaths are required for a total of 20 assumed deaths. We did not make similar estimates for rice because its genomic evolution is less well understood.

In summary, assuming three polyploidy events since divergence from the rice lineage, we estimate that growth of the expansin superfamily in the Arabidopsis lineage involved a minimum of 48 gene deaths, which were more than offset by a minimum of 67 gene births (branch points in Figure 2), 56 of which are due to events α–γ, three due to independent segmental duplications, and eight due to tandem duplications. The relevant numbers if event γ happened in the common lineage of Arabidopsis and rice are 34 gene deaths and 53 gene births since the last common ancestor (Figure S3). It is noteworthy that gene deaths equaled gene births in three of the clades preserved in both species (EXPA-II, EXPA-VIII and EXPA-IX), so that one-to-one orthologous relationships have been preserved despite multiple rounds of polyploidy. Perhaps the genes in these clades are highly specialized, thus impeding functional divergence.

Because polyploidy events allow us to tentatively date most of the gene duplications in the Arabidopsis lineage, we can also estimate the minimal size of the superfamily at different times. We estimate at least 16 genes before event γ, 19 before event β, and 22 before the most recent polyploidy event (Figure 4). Gene loss is shown as a straight line in Figure 4, but it is not expected to be uniform in the long intervals between polyploidy events. Maize, which has been a tetraploid for just 12 Myr, already seems to have lost 50% of its duplicated genes (Lai *et al.*, 2004). This kind of analysis could also allow predictions to be made as to the structure of the expansin superfamily in particular eudicot species, once their divergence from the Arabidopsis lineage had been dated with respect to the polyploidy events. We would know, for example, if we should be looking for orthologs of individual Arabidopsis genes or for orthologs of several duplicated genes. This sort of information could help in exporting knowledge from model species to economically important ones.

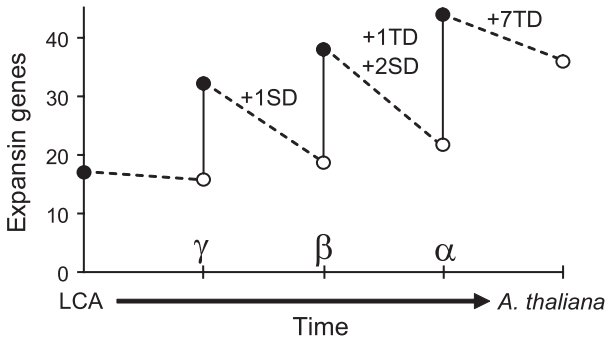


Figure 4. Changes in the minimal estimated number of expansin genes from the last common ancestor (LCA) of eudicots and monocots to present-day Arabidopsis.
Polyploidy events (vertical lines) are identified as in the text. Other gene gains are shown in the intervals. TD = tandem duplications, SD = segmental duplications.

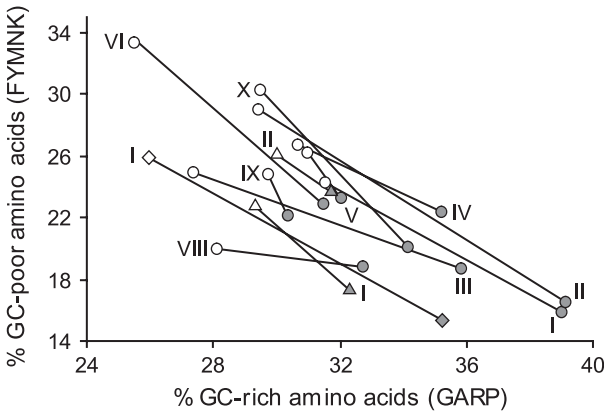


Figure 5. Changes in amino acid composition related to %GC.
For each expansin clade, the average proportions of GC-rich and GC-poor amino acids are plotted for rice (gray) and Arabidopsis (white), connected by a line. Circles are for EXPA clades, triangles for EXPB and diamonds for EXLA. Clade numbers as in Figure 2.

Pseudogenes

The Columbia ecotype of Arabidopsis contains two expansin pseudogenes: Ψ*AtEXPB6* is missing the end of the promoter and most of exon 1, whereas Ψ*AtEXPA19* has a 2-bp insertion that creates a premature stop codon. However, when we sequenced these genes in the Landsberg ecotype, we found them to be normal (undisrupted) genes (GenBank accession nos AY619565 and AY843212). In rice, the *Japonica* genome contains one pseudogene, Ψ*OsEXPB13*, with a premature stop codon. In the *Indica* genome, this stop codon does not exist, but a 1-bp deletion causes a similar result (GenBank accession no. AAAA02007691). It appears that this gene was independently inactivated in both cultivars.

Nucleotide and amino acid composition

Due to its potential to distort phylogenetic trees, we investigated amino acid bias in the expansin superfamily. In Arabidopsis, the distribution of guanine + cytosine (GC) content for coding regions is unimodal and centered around 44%, while in rice it is much broader, with two main peaks, around 45–50 and 65–70% (Wang *et al.*, 2004). Rice expansins, with an average 66% GC content, are mostly in the GC-rich category. The differences in GC content between Arabidopsis and rice are mainly due to third codon positions, but recently it was shown that rice GC-rich genes, when compared to their closest Arabidopsis homologs, are also enriched in certain amino acids (GARP; glycine, atanine, arginine and proline) encoded by GC-rich codons, while simultaneously depleted of amino acids (FYMNK; phenylalanine, tyrosine, methionine, isoleucine, asparagine and lysine) encoded by GC-poor codons (Wang *et al.*, 2004).

We compared the average GC-rich and GC-poor amino acid content in the rice and Arabidopsis branches of each of the 12 clades represented in both species (Figure 5). All of them show a consistent shift, with rice clades showing, on average, an 18% increase in GC-rich amino acids when compared with their Arabidopsis orthologs (from 0.29 to 0.34) and a 25% decrease in GC-poor amino acids (from 0.26 to 0.20). This shift in amino acids probably contributes to the difficulties in phylogenetic analysis noted earlier and may be a common problem in sequence-based phylogenetic analyses that include rice and Arabidopsis genes.

Discussion

Through microsynteny analysis in Arabidopsis and rice, we have uncovered the patterns and mechanisms of gene births and deaths in the expansin gene family since the last common ancestor of monocots and eudicots. Sequence-based phylogenetic trees lacked the resolving power for this job, but when integrated with information about the duplication history of the genomic segments where expansins are located, a much clearer picture emerged. We note that position-based phylogenetic analyses similar to ours could become an important tool for dissecting the evolution of many other gene families, thus adding to the already major contributions that the genomics revolution has made to molecular evolution research (Wolfe and Li, 2003).

Segmental and tandem duplications are the main sources of growth in the expansin superfamily and both preserve positional information. Translocations are rare, and so most genes are still found in the same genomic neighborhood as their predecessors in the last ancestor of monocots and eudicots (Figure 2). In a recent genome-wide analysis that used an automated approach somewhat similar to ours, 14% of the Arabidopsis genome showed microsynteny with 30% of the rice genome (Simillion *et al.*, 2004). In contrast, our

detailed analysis found cross-species synteny in all 12 shared clades. If expansin-containing segments are typical, our results suggest that the extent of microsynteny between rice and Arabidopsis has been seriously underestimated in previous analyses and that it may actually encompass the greater part of both genomes.

Problems with sequence-based phylogenies

The low phylogenetic resolution of the Arabidopsis and rice dataset could be due in part to the differences in amino acid usage in these two species. When combined with strong functional constraints, the magnitude of the bias could be enough to erase the weak phylogenetic signal of short internal branches. It is probably not a coincidence that some of the largest changes in amino acid composition (Figure 5) have taken place precisely in the more problematic clades. In sequence-based trees, the Arabidopsis member of clade EXPA-II (AtEXP14) tends to branch independently or with Arabidopsis clade EXPA-I, while the rice representative of the same clade (OsEXPA11) tends to branch with rice clade EXPA-III. This is likely due to the fact that the rice sequence shows a 33% increase in GC-rich amino acids and a 43% decrease in GC-poor amino acids when compared with its Arabidopsis ortholog. A similar problem could be affecting EXPA-VI, the only other clade not recognized as a monophyletic group by any sequence-based tree.

It has recently been estimated that 21% of duplicated genes in event α show evidence of significantly different rates of evolution at the protein level (Blanc and Wolfe, 2004a). This is likely to apply also to older duplications as well. Unequal evolution rates have the potential to cause weak resolution in sequence-based phylogenies and even misleading topologies due to long-branch attraction (Felsenstein, 1978). As we saw before, this seems to be the case for expansin clade EXPA-IV (Figure 3), and a similar problem could be affecting clades EXPA-V and EXPB-I (Figure 1; see Supplementary Text for details).

Position-based phylogeny

Segmental duplications are the only source of gene growth in most expansin clades and so the phylogeny of expansin-containing segments can clarify the evolution of the expansin superfamily. Before segment phylogeny is taken as a proxy for gene phylogeny, some caveats have to be considered. A tandem of two genes, for example, could be part of a segmental duplication event. If alternative copies of the two genes are then lost in the two segments, the date of duplication for the segments would underestimate the gene duplication date. Gene conversion between genes in paralogous segments would have the opposite effect, but this phenomenon appears to be common only when genes are in close physical proximity (Baumgarten *et al.*, 2003). Despite these potential problems, it can be argued that a segment-averaged date of duplication should, in general, be a better estimate of the real age of the individual gene duplications in the segments than those obtained by directly comparing any pair of duplicated genes.

Luckily, most segmental duplications in the Arabidopsis and rice lineages seem to be linked to well-spaced polyploidy events, making their dating relatively easy. Nonetheless, an improved census of segmental duplications in Arabidopsis and rice that also identifies and dates those not linked to polyploidy events would be a valuable resource for the general application of position-based phylogeny in these two species.

Orthologs or paralogs

An essential question, when trying to identify true orthologs by microsynteny, is the relation between polyploidy events and speciation. Some studies have suggested that event γ, detected in the Arabidopsis genome, predates the divergence of the rice lineage (Bowers *et al.*, 2003; Chapman *et al.*, 2004). This conclusion is based on rooted trees obtained using pairs of Arabidopsis paralogs from event γ and the closest rice homolog from a nearly completed genome. In 53% of the cases, the rice sequence branched with one of the Arabidopsis paralogs (Chapman *et al.*, 2004). However, assuming a pre-rice event γ, the remaining 47% of trees imply double clade losses. Considering the long distances involved and the likelihood of unequal evolution in paralogs (Blanc and Wolfe, 2004a), this result could also be seen as indicating that event γ happened too close to the last common ancestor of monocots and eudicots for a strong phylogenetic signal to have been preserved.

After event α, only 15% of duplicated gene pairs were preserved in two copies (Blanc and Wolfe, 2004a) and even lower values are expected for event γ (Bowers *et al.*, 2003; Maere *et al.*, 2005). If this event had occurred before the divergence of the rice and Arabidopsis lineages, we would expect individual rice segments to show much better synteny with the Arabidopsis segments on one of the branches of each γ node and we would also expect a clearly skewed distribution of best matches. However, none of this is apparent in the segments we have studied (Figure 3a and Figure S2). An in-depth analysis of the rice genome would probably resolve this issue. As far as the expansins are concerned, the most parsimonious hypothesis is that event γ is specific to the Arabidopsis lineage, as shown in Figure 2. If, on the other hand, this event had happened in the common lineage of Arabidopsis and rice, two more genes would be required in their last common ancestor (i.e. clades EXPA-I and EXPA-IV would have to be divided as shown in Figure S3). Although the consequences for expansin phylogeny are relatively minor, the dating of event γ and its possible presence in the rice genome need to be clarified to

allow a more widespread application of position-based phylogeny to angiosperm evolution.

Potential applications

We have shown how position-based phylogeny can detect orthologous relations obscured by amino acid bias, alert us to the presence of possible artifacts in sequence-based trees, and provide additional evidence for poorly supported branches. Because expansins seem to be fairly typical in their process of growth when compared with other large gene families (Cannon *et al.*, 2004) and because three-quarters of the genes in Arabidopsis and more than half in rice belong to multigene families (Horan *et al.*, 2005), our method could prove to be useful in understanding the evolution of many other gene families since the separation of the monocots and eudicot lineages. In particular, this kind of analysis will provide an excellent framework for studying the stability or divergence of gene function and the fate of gene duplications. Furthermore, a family subdivision like the one we propose, based on the independent lineages existing in the last ancestor of eudicots and monocots and supported by microsynteny evidence, could be useful in other large families and should be readily applicable to at least 90% of all angiosperms. As genomic sequences from basal angiosperms and early branching eudicots and monocots become available, the resolving power of position-based phylogeny will be greatly increased.

The major role of polyploidy events in the growth of many angiosperm gene families is what makes this approach particularly attractive. In the future, it may be possible to relate gene duplications in different families involved in the same process, by first assigning them to particular polyploidy events. We must consider, for example, the possibility that duplicated expansin genes may form part of duplications of entire gene networks, including transcription factors and other wall proteins (Blanc and Wolfe, 2004a). Our hypothesis will likely need refinement when the evolution of the rice and Arabidopsis genomes is better understood. A number of polyploidy events in other angiosperm lineages have also been proposed (Blanc and Wolfe, 2004b) and it should be possible to incorporate them into this framework as genomic sequences from the relevant species become available.

Experimental procedures

Phylogenetic trees

Expansin protein sequences were aligned with CLUSTALW from DNASTAR 5.01 using default parameters (Gonnet matrices, gap penalty of 15, gap length penalty of 6.66). Neighbor-joining trees were constructed using MEGA 2.1 (Kumar *et al.*, 2001) with Poisson-corrected distances and complete gap deletion. Bootstrap values are based on 1000 replicates. The EXPA tree was manually rooted at clade EXPA-X and the EXPB family was used to root the second tree. Methods for other trees are detailed in Figure S1.

Microsynteny diagrams

We used published data (Bowers *et al.*, 2003; Simillion *et al.*, 2002) to identify and align paralogous groups of Arabidopsis genomic segments containing at least one expansin gene. Rice protein sequences from expansin-containing genomic segments were downloaded from the TIGR Rice Genome Annotation Resource (Yuan *et al.*, 2003) and correspond to release 3 of TIGR Rice Pseudomolecules. We used these rice protein sequences as queries in BLASTP searches (Madden *et al.*, 1996) against the Arabidopsis proteome build 5.0, deposited at Entrez Genome. Expect threshold was 0.01. If one of the first four hits for a rice protein was located in one of the previously identified Arabidopsis groups of expansin-containing paralogous segments, all hits for that protein in the same group of segments were considered potential orthologs, up to the 12th best hit. A minimum of 300 kbp surrounding each rice expansin or group of expansins was analyzed this way. Diagrams for orthologous groups were constructed to show conservation of relative position and orientation.

We considered a group of segments as orthologous when at least three rice proteins in a segment had their best hit in the same group of Arabidopsis segments (we counted tandems as a single gene). With a randomly shuffled genome, starting with a rice segment of 50 genes and a target region in Arabidopsis of 250 genes, the probability of finding three best hits by chance is below 1%, and this figure does not take into account the conservation of order and orientation, which reduces the probability even further.

PCR

We obtained Landsberg *erecta* (Ler-0) seeds from the Arabidopsis Biological Resource Center. Primers used to amplify *AtEXPA19* and *AtEXPB6* are listed in Table S1.

Acknowledgements

This work was supported by the National Science Foundation (NSF) grant IBN-9874432 to DJC.

Supplementary Material

The following supplementary material is available for this article online:

Figure S1. Protein phylogenetic trees for expansin families.

Figure S2. Synteny diagrams and segmental cladograms for identified orthologous groups of segments.

Figure S3. Integrated cladograms for the expansin superfamily under the hypothesis that event γ happened in the common lineage of monocots and eudicots.

Figure S4. DNA parsimony trees for clades with more than two rice or Arabidopsis genes.

Figure S5. Intron pattern evolution.

Table S1 Sequence of primers used in this work

Supplementary Text. Clade discussions.

This material is available as part of the online article from http://www.blackwell-synergy.com

References

Baumgarten, A., Cannon, S., Spangler, R. and May, G. (2003) Genome-level evolution of resistance genes in *Arabidopsis thaliana*. *Genetics*, **165**, 309–319.

Blanc, G. and Wolfe, K.H. (2004a) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell*, **16**, 1679–1691.

Blanc, G. and Wolfe, K.H. (2004b) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, **16**, 1667–1678.

Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**, 433–438.

Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D. and May, G. (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* **4**, 10.

Carpita, N.C. (1996) Structure and biogenesis of the cell walls of grasses. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 445–476.

Chapman, B.A., Bowers, J.E., Schulze, S.R. and Paterson, A.H. (2004) A comparative phylogenetic approach for dating whole genome duplication events. *Bioinformatics*, **20**, 180–185.

Cho, H.T. and Cosgrove, D.J. (2000) Altered expression of expansin modulates leaf growth and pedicel abscission in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **97**, 9783–9788.

Cho, H.T. and Cosgrove, D.J. (2002) Regulation of root hair initiation and expansin gene expression in Arabidopsis. *Plant Cell*, **14**, 3237–3253.

Cho, H.T. and Kende, H. (1997) Expansins in deepwater rice internodes. *Plant Physiol.* **113**, 1137–1143.

Cho, H.T. and Kende, H. (1998) Tissue localization of expansins in deepwater rice. *Plant J.* **15**, 805–812.

Cosgrove, D.J. (2000) Loosening of plant cell walls by expansins. *Nature*, **407**, 321–326.

Cosgrove, D.J., Bedinger, P. and Durachko, D.M. (1997) Group I allergens of grass pollen as cell wall-loosening agents. *Proc. Natl Acad. Sci. USA*, **94**, 6559–6564.

Devos, K.M., Beales, J., Nagamura, Y. and Sasaki, T. (1999) Arabidopsis-rice: will colinearity allow gene prediction across the eudicot-monocot divide? *Genome Res.* **9**, 825–829.

Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410.

Gray-Mitsumune, M., Mellerowicz, E.J., Abe, H., Schrader, J., Winzell, A., Sterky, F., Blomqvist, K., Queen-Mason, S., Teeri, T.T. and Sundberg, B. (2004) Expansins abundant in secondary xylem belong to subgroup A of the alpha-expansin gene family. *Plant Physiol.* **135**, 1552–1564.

Horan, K., Lauricha, J., Bailey-Serres, J., Raikhel, N. and Girke, T. (2005) Genome cluster database. A sequence family analysis platform for Arabidopsis and rice. *Plant Physiol.* **138**, 47–54.

Kende, H., Bradford, K., Brummell, D. *et al.* (2004) Nomenclature for members of the expansin superfamily of genes and proteins. *Plant Mol. Biol.* **55**, 311–314.

Kumar, S., Tamura, K., Jakobsen, I.B. and Nei, M. (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics*, **17**, 1244–1245.

Lai, J., Ma, J., Swigonova, Z. *et al.* (2004) Gene loss and movement in the maize genome. *Genome Res.* **14**, 1924–1931.

Lee, Y., Choi, D. and Kende, H. (2001) Expansins: ever-expanding numbers and functions. *Curr. Opin. Plant Biol.* **4**, 527–532.

Leebens-Mack, J., Raubeson, L.A., Cui, L., Kuehl, J.V., Fourcade, M.H., Chumley, T.W., Boore, J.L., Jansen, R.K. and de Pamphilis, C.W. (2005) Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. *Mol. Biol. Evol.* doi:10.1093/molbev/msi191.

Li, Z.-C., Durachko, D.M. and Cosgrove, D.J. (1993) An oat coleoptile wall protein that induces wall extension in vitro and that is antigenically related to a similar protein from cucumber hypocotyls. *Planta*, **191**, 349–356.

Li, Y., Darley, C.P., Ongaro, V., Fleming, A., Schipper, O., Baldauf, S.L. and McQueen-Mason, S.J. (2002) Plant expansins are a complex multigene family with an ancient evolutionary origin. *Plant Physiol.* **128**, 854–864.

Li, Y., Jones, L. and McQueen-Mason, S. (2003) Expansins and cell growth. *Curr. Opin. Plant Biol.* **6**, 603–610.

Link, B.M. and Cosgrove, D.J. (1998) Acid-growth response and alpha-expansins in suspension cultures of bright yellow 2 tobacco. *Plant Physiol.* **118**, 907–916.

Madden, T.L., Tatusov, R.L. and Zhang, J. (1996) Applications of network BLAST server. *Methods Enzymol.* **266**, 131–141.

Maere, S., De, B.S., Raes, J., Casneuf, T., Van, M.M., Kuiper, M. and Van de Peer, Y. (2005) Modeling gene and genome duplications in eukaryotes. *Proc. Natl Acad. Sci. USA*, **102**, 5454–5459.

McQueen-Mason, S., Durachko, D.M. and Cosgrove, D.J. (1992) Two endogenous proteins that induce cell wall expansion in plants. *Plant Cell*, **4**, 1425–1433.

Paterson, A.H., Bowers, J.E. and Chapman, B.A. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA*, **101**, 9903–9908.

Salse, J., Piegu, B., Cooke, R. and Delseny, M. (2002) Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. *Nucleic Acids Res.* **30**, 2316–2328.

Sanderson, M.J., Thorne, J.L., Wikström, N. and Bremer, K. (2004) Molecular evidence on plant divergence times. *Am. J. Bot.* **91**, 1656–1665.

Sasaki, T. and Burr, B. (2000) International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.* **3**, 138–141.

Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M. and Van de Peer, Y. (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **99**, 13627–13632.

Simillion, C., Vandepoele, K., Saeys, Y. and Van de Peer, Y. (2004) Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Res.* **14**, 1095–1106.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Vandepoele, K., Saeys, Y., Simillion, C., Raes, J. and Van de Peer, Y. (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome Res.* **12**, 1792–1801.

Wang, H.C., Singer, G.A. and Hickey, D.A. (2004) Mutational bias affects protein evolution in flowering plants. *Mol. Biol. Evol.* **21**, 90–96.

Wolfe, K.H. and Li, W.H. (2003) Molecular evolution meets the genomics revolution. *Nat. Genet.* **33**, 255–265.

Yuan, Q., Ouyang, S., Liu, J., Suh, B., Cheung, F., Sultana, R., Lee, D., Quackenbush, J. and Buell, C.R. (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res.* **31**, 229–233.

Zenoni, S., Reale, L., Tornielli, G.B. *et al.* (2004) Downregulation of the Petunia hybrida alpha-expansin gene PhEXP1 reduces the amount of crystalline cellulose in cell walls and leads to phenotypic changes in petal limbs. *Plant Cell*, **16**, 295–308.

Sequence data: GenBank accession nos AY619565 and AY843212.