Received Date : 19-Oct-2016

Accepted Date : 25-Jan-2017

Article type : Original Article

Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules

Colin Ruprecht[1], Sebastian Proost[1], Marcela Hernandez-Coronado[2], Carlos Ortiz-Ramirez[2], Daniel Lang[3], Stefan A. Rensing[4], Jörg D. Becker[2], Klaas Vandepoele[5], Marek Mutwil[1]*

[1]Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam, Germany

[2]Instituto Gulbekian De Ciencia, Rua da Quinta Grande 6, 2780-156 Oeiras, Portugal

[3]University of Freiburg, Schänzlestr. 1, D-79104 Freiburg, Germany

[4]University of Marburg, Karl-von-Frisch-Str. 8, D-35043 Marburg, Germany

[5]Department of Plant Systems Biology VIB, Department of Plant Biotechnology and Bioinformatics Ghent University , Technologiepark 927, B-9052 Gent, Belgium

***Corresponding author:**

Marek Mutwil

Max Planck Institute of Molecular Plant Physiology,

Am Muehlenberg 1,

14476 Potsdam,

Germany

Email: Mutwil@mpimp-golm.mpg.de

The author responsible for distribution of materials integral to the findings presented in this article is: Marek Mutwil (mutwil@mpimp-golm.mpg.de).

Running title: Evolutionary analysis of co-expression networks

Keywords: comparative co-expression, network evolution, phylostratigraphy, phylogenetics, gene function, *Arabidopsis thaliana*, *Oryza sativa*, *Physcomitrella patens*

Summary:

Molecular evolutionary studies correlate genomic and phylogenetic information with the emergence of new traits of organisms. These traits are, however, the consequence of dynamic gene networks composed of functional modules, which might not be captured by genomic analyses. Here, we established a method, which combines large-scale genomic and phylogenetic data with gene co-expression networks, to extensively study the evolutionary make-up of modules in the moss *Physcomitrella patens* and in the angiosperms *Arabidopsis thaliana* and rice. We first show that younger genes are less annotated than older genes. By mapping genomic data onto the co-expression networks, we found that genes from the same evolutionary period tend to be connected, while old and young genes tend to be disconnected. Consequently, the analysis revealed modules that emerged at a specific time

in plant evolution. To uncover the evolutionary relationships of the modules that are conserved across the plant kingdom, we added phylogenetic information which revealed duplication and speciation events on the module level. This combined analysis revealed an independent duplication of cell wall modules in bryophytes and angiosperms, suggesting a parallel evolution of cell wall pathways in land plants. We provide an online tool allowing plant researchers to perform these analyses at www.gene2function.de.

Introduction

Studying plant evolution can reveal principles that govern the establishment of multi-cellularity, hormone signalling, adaptation to terrestrial growth and sexual reproduction. Our initial understanding of plant evolution was based on the analysis of morphological and developmental features of different lineages, but with the advent of functional genomics, an improved toolkit to unravel molecular mechanisms driving plant evolution became available (Somerville and Somerville, 1999). Genomes of model species that represent major plant clades are now available for glaucophytes (*Cyanophora paradoxa*) (Price et al., 2012), red algae (*Cyanidioschyzon merolae*) (Matsuzaki et al., 2004), chlorophytes (*Chlamydomonas reinhardtii*) (Merchant et al., 2010), charophytes (*Klebsormidium flaccidum*) (Hori et al., 2014), early embryophytes (*Physcomitrella patens*) (Rensing et al., 2008), early vascular plants (*Selaginella moellendorfii*) (Banks et al., 2011), seed plants (*Picea abies*) (Nystedt et al., 2013), basal flowering plants (*Amborella*) (Amborella Genome Project, 2013), monocots (*Oryza sativa*) (International Rice Genome Sequencing Project, 2005) and dicots (*Arabidopsis thaliana*) (Arabidopsis Genome Initiative, 2000). These genomes, together with available

molecular evolutionary approaches, allow us to study plant evolution in unprecedented detail.

Molecular evolutionary approaches can be broadly divided into two categories: comparative genomics and molecular phylogenetics. Comparative genomics can identify similarities and differences of genomic features, such as gene families, of two or more organisms (Rubin, 2000; Hardison, 2003). To study the early evolution of plants, the moss *Physcomitrella patens* (*P. patens*) and the lycophyte *Selaginella moellendorffii* (*S. moellendorffii*) were chosen as model organisms for bryophytes and early diverging vascular plants, respectively (Cove et al., 1997; Rensing et al., 2008; Banks et al., 2011). Comparison of their gene family contents with *Arabidopsis thaliana* enabled the correlation of genomic data with functional innovations that occurred in early land plant evolution (Rensing et al., 2008; Banks et al., 2011). For example, the analysis of the *P. patens* genome revealed that many gene families involved in phytohormone signalling and transcriptional regulation were already present in the first land plants (Rensing et al., 2008). Similar analyses for *S. moellendorffii* identified novel gene families that emerged during the transition from vascular plants to seed plants, along with extensive yet independent duplication of secondary metabolism related genes in both plant lineages (Banks et al., 2011). These comparative genomic approaches, when combined with genomes of representative plant clades, can reveal in which plant clade a given gene or gene family appeared or was lost (Vandepoele and Van de Peer, 2005; Domazet-Loso et al., 2007; Guo, 2013). The evolutionary origin of genes and gene families can be identified by phylostratigraphy, which traces the earliest common ancestor of a gene or gene family, and consequently, estimates the age of the studied object (Domazet-Loso et al., 2007).

Molecular phylogenetics mainly study DNA or protein sequences to infer relationships between organisms and genes. The outcome of such analyses is a phylogenetic tree, which shows the inferred evolutionary relationships between the studied entities. For example, an analysis of 78 plastid genes from 360 diverse green plant taxa was used to infer a species tree, which revealed that cycads and Ginkgo are closest relatives (sisters) to extant gymnosperms, while horsetails are sisters to extant ferns (Ruhfel et al., 2014). On the other hand, a multi-species phylogenetic tree of a gene family can reveal the speciation and duplication history of genes found in the family (Proost et al., 2009; Vilella et al., 2009). By combining phylogenetic trees with species trees, the evolutionary timing of gene speciation or duplication can be uncovered (Proost et al., 2009; Vermeirssen et al., 2014).

These comparative genomic and phylogenomic analyses are based on the notion that biological functions emerge and evolve as gene families. To understand the evolution of biological pathways, however, such approaches have two important shortcomings that limit our understanding of evolution. First and foremost, genes and gene families rarely operate as single entities, but rather as functional gene modules (Hartwell et al., 1999). Their protein products can form higher order complexes and enzymatic pathways which often require multiple genes and gene families that operate together to perform a given task. For example, the photosystem II complex from cyanobacteria is composed of at least 20 protein subunits that interact with 77 cofactors (Loll et al., 2005). While genomic analyses could reveal family expansion and evolutionary rates of the individual subunits, it might not reveal evolutionary constraints of the whole complex without a priori knowledge of the underlying interactions. Second, plant gene families are typically large (Shiu and Bleecker, 2001; Shiu et

al., 2005) with homologs displaying divergent functions (Kliebenstein, 2001). Consequently, where duplications are abundant, sequence-based analyses are prone to placing a gene into an incorrect functional context in case of neo- or sub-functionalization of paralogs (Lynch and Katju, 2004; He and Zhang, 2005). Therefore, a more rewarding approach to explain the evolution of new traits and adaptations should by necessity integrate molecular evolutionary approaches with functional gene modules.

Genes that are involved in related biological processes tend to be co-expressed (i.e. transcriptionally co-regulated) across tissues and thus cluster together as gene modules (Usadel et al., 2009). In these networks, nodes correspond to genes and edges indicate co-expression relationships between genes (Lee et al., 2004). Gene co-expression networks have proven to be an invaluable method to predict functions of modules and have been applied to many model organisms (Stuart et al., 2003; Yu et al., 2003; Persson et al., 2005; Itkin et al., 2013). To facilitate access to these networks, several web-based tools emerged that allow researchers to exploit such networks to predict gene function and guide reverse-genetic approaches (e.g. (Mutwil et al., 2010; Obayashi et al., 2011; De Bodt et al., 2012; Lee et al., 2015)). Recently, the analyses were also extended to several plant crop species (Ficklin and Feltus, 2011; Movahedi et al., 2011; Mutwil et al., 2011; Tzfadia et al., 2016). Co-expression networks are conserved across species and even across distinct kingdoms of life, indicating that essential cellular processes, such as cell cycle, ribosome biogenesis, and the proteasome are implemented by the same orthologous modules in different species (Stuart et al., 2003; Gerstein et al., 2014; Zarrineh et al., 2014). These conserved modules can therefore be used to transfer knowledge obtained from better investigated model species to e.g. crop plants (Mutwil et al., 2011; Ruprecht et al., 2011; Heyndrickx and Vandepoele,

2012; Park et al., 2013; Tzfadia et al., 2016). In addition, several case studies suggested that some modules have been duplicated in some species in order to accommodate more complex plant organs and tissues (Kliebenstein, 2001; Persson et al., 2005; Matsuno et al., 2009; Busch et al., 2011). To systematically identify such duplicated modules, we recently introduced a platform that we named FamNet, which revealed that more than 30% of plant genes can be found in duplicated modules (Busch et al., 2011). Although these analyses gave insight into the conservation and duplication of modules, they lacked the required phylogenetic and genomic information to address the chronological emergence of different modules and the timing of module duplications.

Previous attempts to reconcile phylogenetics and genomics with modules have either used a limited amount of phylogenetic data or few co-expression networks. For example, integration of phylogenetic data of the *CONSTANS* gene family with the co-expression networks of *Chlamydomonas reinhardtii* (*C. reinhardtii*), *P. patens* and *Arabidopsis thaliana* (*A. thaliana*) could show the evolutionary relationships of the light-dependent regulatory modules from the three species (Romero-Campero et al., 2013). Fang and co-workers (Fang et al., 2013) used phylogenetic information from 236 bacterial species and found high conservation of genes for a majority of modules identified in a cell cycle-related co-expression network of the bacterium *Caulobacter crescentus*. Hansen and colleagues (Hansen et al., 2014) combined comparative co-expression analysis with genomic information to predict the composition of cell wall modules in the ancestor of angiosperms. Another study showed that younger gene families were highly expressed in pollen (Cui et al., 2015). Finally, by combining phylostratigraphy with developmental series expression data, a

convergent molecular and morphological pattern called the embryonic hourglass has been revealed in plants (Quint et al., 2012).

Despite these examples, little is known about how evolution contributes to the emergence, duplication and diversification of biological pathways in the different plant lineages. In this manuscript we first combined a phylostratigraphic approach (Domazet-Loso et al., 2007; Guo, 2013) with gene co-expression networks to show how genes derived from the different evolutionary periods are interconnected in gene co-expression networks of *A. thaliana*, *P. patens* and rice. Next, we used phylogenetics to reveal evolutionary relationships of gene modules in these three species. To facilitate access to these analyses, we generated an interactive web-based tool at www.gene2function.de.

Results

Phylostratigraphic analyses of plant gene families

In this study, we used 43391 PLAZA2.5 gene families based on genomes of five chlorophytes, one bryophyte (model land plant), one lycophyte (model vascular plant), five monocots and thirteen dicots (see Experimental procedure, (Van Bel et al., 2012)). The phylostratum of a gene family was determined by identifying the earliest plant lineage found within the gene family (Figure 1, Table S1, (Guo, 2013)). For example, a gene family present in *A. thaliana* (dicot) and *C. reinhardtii* (chlorophyte) was assigned to the Green Plants phylostratum (phylostrata are indicated by capital letters), as *C. reinhardtii* is the earliest plant lineage in this gene family (Figure 1A). Conversely, if a gene family contained genes found exclusively in *A. thaliana* and rice, the gene family was assigned to the

Angiosperm phylostratum, as this family was likely present in the ancestor of monocots and dicots. Since the order of divergence of the different plant lineages is known, the analysis can delineate the relative age of a gene family and genes found within the family (Figure 1A-B). For example, a gene family assigned to the Green Plants phylostratum is older than one assigned to the Angiosperms phylostratum.

The analysis assigned each family and genes found in the family to one of the phylostrata shown on Figure 1 (Table S1). We exemplify the outcome of the analysis with a representative of early land plants (bryophyte *Physcomitrella patens*), a eudicot (*Arabidopsis thaliana*) and a monocot (rice, *Oryza sativa*). *P. patens*, *A. thaliana*, and rice contain similar numbers of genes and gene families for the common phylostrata (Green Plants and Land Plants, Figure 1A). While *A. thaliana* and rice contain a similar amount of gene families in the Angiosperm phylostratum (519 vs. 619 for *A. thaliana* and rice, respectively, Figure 1A), rice has roughly double the amount of Angiosperm phylostratum genes, indicating either extensive duplications in rice, or conversely, an extensive loss of these genes in *A. thaliana*. Furthermore, rice contains 1348 Monocot gene families, while *A. thaliana* contains 241 Dicot gene families (Figure 1B), indicating that monocots contain more gene families than dicots, which is in agreement with previous studies (Vandepoele and Van de Peer, 2005; Cui et al., 2015). Finally, *P. patens*, *A. thaliana* and rice contain 7680, 1362 and 4963 lineage specific (also called orphan) families, respectively, representing recently established genes. Using this phylostratigraphic information for each gene in the three species as a basis, we first analyzed the functional annotation of genes from the different phylostrata, and then mapped the phylostratigraphic information onto the co-expression networks of *P. patens*, *A. thaliana* and rice to investigate properties of the phylostrata in the networks.

Inverse relationship between age and functional information of gene families

To gain insight into available functional information for each phylostrata, we have retrieved GO terms derived from experimentally characterized genes in *A. thaliana*, and estimated their distribution in the phylostrata (Table S2A). The analysis revealed that percentages of experimentally characterized genes decreased as the phylostrata become younger (Figure 2A), indicating that older gene families are more functionally characterized than younger families. Interestingly, the percentage of experimentally characterized genes per phylostratum almost perfectly corresponded to the order of plant clades (Figure 1), with the exception of the Brassicales phylostratum (Figure 2A). However, we hypothesized that this exception is caused by a very low number of Brassicales phylostratum genes (22 genes in *A. thaliana*, Figure 1, Table S2A), of which four genes (18.1%) have been characterized, resulting in seemingly high percentage.

Next, we investigated which types of experiments were used to characterize these genes (Table S2B). The experimental evidence codes are inferred from direct assay (IDA, comprising enzyme assay, physical interaction/binding, and immunofluorescence), protein interaction (IPI, yeast-two-hybrid and pull-down assays), mutant phenotype (IMP, characterization of mutant phenotypes), genetic interaction (IGI, phenotype suppressors and synthetic lethals) and expression (IEP, transcript or protein levels). Similarly to the distribution of functional information per phylostrata, the amount of experimental evidences derived from direct assays, mutant phenotypes and genetic interactions decreased for younger phylostrata (Figure 2B). Exceptions included protein interaction (IPI, e.g. genes from the Vascular Plants phylostratum are more characterized by this assay than

from the Land Plants phylostratum) and expression profile (IEP, e.g. genes from the Eudicots phylostratum are more characterised than from the Angiosperms phylostratum, Figure 2B).

Functionally related genes are enriched for distinct phylostrata

To gain a genome-wide view on the functions of the phylostrata, we investigated which biological functions are assigned to the phylostrata. To this end, we asked whether certain phylostrata are enriched for functional categories represented by MapMan gene ontology terms (Thimm et al., 2004). We applied a hypergeometric distribution analysis to all MapMan bins in *A. thaliana*, rice and *P. patens*, and report significant associations at significance level of $P<0.05$ (Table S3).

We found that most MapMan bins representing fundamental biological features, such as photosynthesis, primary metabolism, RNA (transcription and processing), signalling and DNA metabolism were significantly enriched ($P<0.05$) for the Green Plants phylostratum, indicating that genes and gene families involved in these processes are ancient (Figure 3, Table S3A). Conversely, biological features that appeared in land plants, such as secondary metabolism (isoprenoids, phenylpropanoids and flavonoids), hormone metabolism (abscisic acid, auxin, ethylene, gibberelin, jasmonate and salicylic acid), development (embroygenesis), RNA (regulation of transcription) were significantly enriched in the Land Plants or younger phylostrata (Figure 3, Table S3). Interestingly, the MapMan bin related to ubiquitin-mediated protein degradation is enriched in Eudicots and Rosids phylostrata ("protein", Figure 3, Table S3), suggesting the appearance and expansion of complex protein degradation pathways in the dicot lineage. Furthermore, the "not assigned" bin, which

contains genes of unknown function is significantly enriched for phylostrata younger than the Green Plants phylostratum (Figure 3), which corresponded to our result that genes from younger phylostrata tend to be less functionally characterized than genes from older phylostrata (Figure 2).

With few exceptions, such as the functional categories stress, development (seed storage proteins) and metal handling, which appeared in separate evolutionary periods in *A. thaliana* (Figure 3, Table S3), the analysis showed enrichments for only one phylostratum or adjacent phylostrata (e.g. Land Plant and Vascular Plant phylostrata for hormone metabolism in *A. thaliana*) per each MapMan bin. Thus, this analysis suggested that new biological features emerged at distinct evolutionary periods, without significant addition of new genetic material during later stages of evolution (Figure 3).

Genes from the same phylostratum tend to be connected in the co-expression networks

Based on our finding that biological processes can be assigned to certain phylostrata, and since functionally related genes are often co-expressed (Usadel et al., 2009), we hypothesized that genes assigned to the same phylostrata might be connected (i.e. co-expressed) in the co-expression networks. To test this hypothesis in well-characterized plant species with abundant expression data, we used the existing co-expression networks of *A. thaliana* and rice (Mutwil et al., 2011). To include a representative species of early land plants, we also generated a new co-expression network for *P. patens* using available microarray data (see Methods).

To investigate whether genes assigned to the same phylostrata are preferentially co-expressed, we first counted how often these genes are connected in the networks. For example, the toy network in Figure 4A consists of three Green Plant genes and one Vascular Plant gene, resulting in three Green Plant-Green Plant (GP-GP) edges and one Green Plant-Vascular Plant (GP-VP) edge (observed edges). Then, we employed a randomization method where we shuffled the gene-phylostratum assignments and estimated the empirical p-value, by comparing the number of the observed GP-GP edges with the number of GP-GP edges in the shuffled networks (Figure S1A).

In our *A. thaliana* co-expression network, we found 35017 GP-GP edges (observed) and on average 26457 GP-GP edges in the networks where gene-phylostratum assignments were shuffled (shuffled, Figure 4B), indicating that the Green Plant phylostratum genes are preferentially connected. We also observed similar pattern for GP-GP edges in the rice and *P. patens* co-expression networks (Figure 4B, $P<0.05$, Table S4). The analysis was performed for all phylostratum combinations, and revealed also a significant association of the Land Plant phylostratum (LP-LP) in all three species, as well as *A. thaliana*-specific and Monocot-specific phylostrata in co-expression networks of *A. thaliana* and rice, respectively (*Arabidopsis-Arabidopsis*, AT-AT, Monocot-Monocot, Oryza-Oryza edges). As the only exception of this trend, we observed a significant dissociation between *P. patens*-specific edges (PP-PP, Figure 4F). We hypothesized that this is due to lack of other bryophyte genomes, which collapsed 500 million years of genetic material into one phylostratum. We also observed a significant association of GP-LP edges in all three species (Figure 4D-F), indicating that genes belonging to the two ancient phylostrata have a higher chance of being co-expressed. Conversely, we found a clear depletion of Green Plant-*A. thaliana* (GP-AT)

edges in *A. thaliana*, GP-*Oryza sativa* (GP-OS) edges in rice and GP-*P. patens* (GP-PP) edges in *P. patens*, indicating that the ancient genes are preferentially disconnected with new genes (Figure 4C, P<0.05, Table S4). We observed similar dissociation of LP-AT, LP-OS and LP-PP edges in the three plants (Figure 4D-F, P<0.05, Table S4).

We concluded that genes which emerged during the same evolutionary period tend to be connected more frequently in co-expression networks than expected by chance, and that genes forged during distinct evolutionary periods tend to be disconnected. To corroborate this, we investigated the sizes of connected components formed by the phylostrata. A connected component is defined here as a group of genes from the same phylostrata in the co-expression network for which a path exists between all genes. If genes belonging to the same phylostratum are preferentially connected, the sizes of phylostratum-specific connected components should be larger than expected by chance. This would indicate that the co-expression network modules are composed of specific phylostrata. To test this hypothesis, we employed a similar gene-phylostratum shuffling analysis as used in the previous analysis, but in this case we compared the sizes of connected components obtained from observed and shuffled gene-phylostratum networks (Figure S1B).

We show connected components for rice (Figure 5A), *A. thaliana* (Figure S2) and *P. patens* (Figure S3). As expected, the sizes of connected components are proportional to the number of genes assigned to the phylostrata (Figure 1), with the Green Plant-specific connected component being the largest (Figure 5A, Table S5). By comparing the size of the largest components in the actual network with 1000 shuffled networks (Figure S1B), the majority of connected components were larger than expected by chance for nearly all considered phylostrata in all species (Figure 5B, P<0.05, Table S5). Exceptions included the Angiosperms

phylostratum in *A. thaliana* and rice, the Vascular Plants phylostratum in *A. thaliana*, and the PP phylostratum in *P. patens* (observed: 4942 genes, average of permutations: 7182 genes for *P. patens*, Table S5). While it was unclear why these phylostrata are disconnected, the exceptions corresponded to the lack of significant association of these phylostrata in the gene co-expression network (Figure 4). To conclude, genes from the same phylostrata tend to form gene modules in the co-expression networks, which implied that certain regions of the co-expression network emerged at a specific time in evolution.

Examples of modules enriched for phylostrata

In our genome-wide analyses, we found that certain modules have been created during distinct evolutionary periods. To identify such modules, we investigated if first neighborhoods in the co-expression networks are enriched for certain phylostrata. A first neighborhood consists of a query gene and all genes co-expressed with it. To determine if a phylostratum is enrichment or depleted in a neighborhood, we employed the same approach (hypergeometric test, Materials and methods) that was used to determine if MapMan bins are enriched for a specific phylostratum. The analysis revealed which neighborhoods are significantly ($P<0.05$) enriched or depleted for the phylostrata in *A. thaliana*, rice and *P. patens* (Table S6A-C). For example, the analysis revealed that 2366 and 938 neighborhoods were significantly enriched and depleted for the Green Plant phylostratum in *A. thaliana*, respectively (Table S6D). To visualize these results, we have updated the PlaNet database (Mutwil et al., 2011) with phylostratigraphic information (see online tutorial describing how to use this feature).

We exemplify four neighborhoods that are significantly enriched for Green Plants, Land Plants, Monocots and *A. thaliana* (AT) phylostrata (Figure 6). The neighborhoods are shown as networks, where nodes correspond to genes, edges connect co-expressed genes, and node borders are used to indicate the phylostratum of a gene.

The first neighborhood from *P. patens* is based on the query gene *Pp1s38_194v6.1*, encoding a ribosomal S4 protein (Figure 6A, large central node). GO enrichment analysis provided by PlaNet indicated that the other genes in this neighborhood are also involved in protein synthesis (http://aranet.mpimp-golm.mpg.de/responder.py?name=gene!ppa!13964). All genes that are assigned to a phylostratum in this neighborhood are belonging to Green Plants phylostratum (22 out of 22 genes, Green Plants enrichment $P<0.05$, Table S6), which is expected for an ancient biological process, such as protein synthesis. The second example is based on *At5g10130,* an uncharacterized extensin from *A. thaliana* (Figure 6B). The neighborhood is significantly enriched for the Land Plants phylostratum (22 out of 31 genes, $P<0.05$) and is putatively involved in cell wall modification and ion uptake processes, as revealed by GO enrichment analysis (http://aranet.mpimp-golm.mpg.de/responder.py?name=gene!ath!16968). Expression profile analyses indicate that this putative cell wall modification and ion uptake module is active in the root cap and the endodermis (links to expression profiles are provided in Methods). The third neighborhood is based on *Loc_os01g12580*, a LEA14 (Late Embryogenesis Abundant) gene from rice (Figure 6C). The neighborhood is significantly enriched for the Monocots phylostratum (8 out of 15 genes, $P<0.05$) and showed expression in mature seeds and embryo. Our GO enrichment analysis indicated that the module is involved in embryo development (http://aranet.mpimp-

golm.mpg.de/responder.py?name=gene!osa!4852), which suggested that the module is implicated in a monocot-specific aspect of seed development. The fourth neighborhood contains mostly genes derived from the *A. thaliana* phylostratum (Figure 6D, 17 out of 19 genes are AT, P<0.05), and is predicted to be involved in carpel biogenesis based on GO enrichment analysis (http://aranet.mpimp-golm.mpg.de/responder.py?name=gene!ath!14921). In line with our finding that young genes are not likely to be functionally characterized (Figure 2 and 3), many genes in this neighborhood do not have any GO term annotation (14 out of 22 genes).

Combining phylostratigraphic analysis with phylogenetic data - evolution of photosynthesis

In addition to phylostratigraphic information that can reveal the evolutionary period, in which a given gene module appeared, we have also included phylogenetic gene trees, to uncover the speciation and duplication events between genes found in the modules ((Proost et al., 2009), Table S7). PlaNet database uses FamNet pipeline to detect duplicated and conserved modules (Ruprecht et al., 2016), and is now updated with phylostratigraphic and phylogenetic analyses.

To exemplify the analysis, we compared the conserved photosynthetic PsbW modules from *P. patens*, *A. thaliana* and rice using our PlaNet database (see web-link in the Methods for an online tutorial on how to perform this analysis). The output of PlaNet showed gene contents, together with phylostratigraphic and phylogenetic information of the selected modules (Figure 7A). Phylogenetic information is indicated by colored edges between

modules, which denote the type of event (speciation or duplication represented by dashed or solid edges, respectively) and evolutionary period of the event (e.g. common ancestor of green plants - green edge, or land plants - red edge) found between genes in two modules (Figure 7A-B). In addition to displaying the phylostratigraphic and phylogenetic relationships of the modules, PlaNet also provides the number and P-values indicating significant enrichment of depletion of these relationships.

As anticipated, *A. thaliana* and rice modules showed significant speciation events corresponding to the split of monocots and dicots (4 purple dashed edges, $P < 0.05$, angiosperm split), while the *P. patens* module showed land plant speciation events (12 and 6 red dashed edges for *A. thaliana*-*P. patens* and rice-*P. patens*, respectively, corresponding to the bryophytes-vascular plants split) relative to *A. thaliana* and rice (Figure 7A, Table S8A). As expected for such an ancient process as photosynthesis, all three modules were enriched for genes from the Green Plants phylostratum (Figure 7B), with 14 out of 14 genes in *A. thaliana*, 10 out of 12 genes in rice and 20 out of 22 genes in *P. patens* assigned to the Green Plants phylostratum ($P < 0.05$ for all three species, Table S8). It is important to note that the modules are not identical, as only 63%, 60% (Mutwil et al., 2011) and 85% (32,569/38,357) of all protein coding sequences are present on the microarrays for *A. thaliana*, rice and *P. patens*. Consequently, some genes are not found in the co-expression networks and are therefore missing in the modules. This is exemplified by unequal presence of some of the gene families found in the photosynthetic modules (Figure 7C).

Taken together, the combination of phylostratigraphic and phylogenetic data showed that the photosynthesis module can be found in green plants (or earlier, as our analysis does not include cyanobacterial genomes) and revealed speciation events concordant with the

species tree of plants (Figure 7C). These results reflect the expected evolution of photosynthesis and thus validate the power of combining genomic and phylogenetic data with co-expression networks, to study evolution of gene modules. Since the phylostratigraphic and phylogenetic relationships of the modules are inferred by multiple gene families and phylogenetic trees, our approach should be more robust to artefacts that arise due to large-scale construction of gene families and phylogenetic trees.

An independent duplication of cellulose biosynthesis pathways in angiosperms and bryophytes

To explore if the combination of phylostratigraphic and phylogenetic inferences can provide new insights, we analyzed gene modules involved in cellulose biosynthesis. Our previous studies have shown that gene modules related to cellulose biosynthesis are present in multiple copies in angiosperms, with primary cell wall (PCW), secondary cell wall (SCW), pollen and root modules being the most prominent in *A. thaliana* (Ruprecht et al., 2016). While these examples represent specialized modules of the cell wall biosynthesis pathway found in higher plants, it is currently not known when in angiosperm evolution the PCW and SCW modules emerged. To elucidate the evolution of cellulose biosynthesis related modules, we queried the PlaNet database with a *P. patens* gene *Pp1s144_8V6*, a multicopper oxidase family gene shown to be associated with cell wall biosynthesis (Ruprecht et al., 2011; Ruprecht et al., 2016). Our FamNet pipeline detected gene modules similar to the *Pp1s144_8V6* module in all angiosperms, indicating that the cellulose biosynthesis module is highly conserved in land plants. Interestingly, two other *P. patens*

gene modules strongly similar to the *Pp1s144_8V6* module were detected, indicating that similarly to angiosperms, *P. patens* also contains multiple cellulose biosynthesis modules.

Expression profiles of the representative genes for each module (*Pp1s60_109V6*, *Pp1s175_122V6* and *Pp1s144_8V6*) revealed that the three selected *P. patens* genes have distinct expression profiles, with *Pp1s60_109V6* highly expressed in archegonia and gametophore, *Pp1s175_122V6* highly expressed in caulonema, and *Pp1s144_8V6* predominantly expressed in protonema (Figure 8A, Figure S4). The expression patterns of *Pp1s175_122V6* and *Pp1s144_8V6* are partially overlapping (protoplasts, protonema and caulonema) with high expression levels in tip-growing tissues, whereas *Pp1s60_109V6* shows a distinct expression in tissues corresponding to isotropic/anisotropic growth (Figure 8A). Based on our previous finding that specialized modules exist for tip-growth in pollen tubes and root hairs in *A. thaliana* (Ruprecht et al., 2016), it appears possible that *P. patens* also evolved distinct cell wall modules that are adapted for two fundamentally different types of growth.

To gain insight into the evolution of cell walls, we have selected the three cell wall biosynthesis modules of *P. patens* along with the primary and secondary cell wall modules from *A. thaliana* for phylogenetic and phylostratigraphic analysis (Figure 8B, Table S9, Figure S5). As expected for modules representing a trait that emerged in land plants, all five modules showed significant ($P < 0.05$) enrichment for genes from the Land Plants phylostratum, indicating that the ancestor of these modules appeared in an ancestor of land plants (Table S9).

Surprisingly, phylogenetic analysis of the three *P. patens* cell wall modules revealed exclusive *P. patens*-specific gene duplications of these modules (black edges), which indicated that the *P. patens* modules were duplicated in the bryophyte lineage, after the split of mosses and vascular plants (Table S9). This result thus represents an independent duplication of the same pathway in two separate plant lineages. This furthermore suggested that the ancestor of land plants contained one cellulosic module, since none of the *P. patens* modules showed land plant-specific gene duplications.

Phylogenetic analysis revealed that the *A. thaliana* PCW and SCW modules are connected by edges representing gene duplications in land plants and angiosperms, indicating that parts of these modules were duplicated in two evolutionary periods (red and purple edges, Figure S5). However, if PCW and SCW modules were partially duplicated in the ancestor of land plants, we would expect to observe modules connected by land plant-specific edges in *P. patens*. Since this was not the case, we investigated the PCW and SCW phylogenetic edges in more detail and observed that the land plant duplication edges were based on the phylogenetic tree of the *CELLULOSE SYNTHASE A* (*CESA*) family (Figure S5). Previous studies indicated that *CESA*s can be divided into two classes, i.e. *CESA1*, *CESA3*, and *CESA6*-related (PCW), and *CESA4*, *CESA7* and *CESA8* (SCW) (Taylor et al., 2003; Persson et al., 2007). The *CESA* phylogenetic tree shows that the *P. patens* and *S. moellendorfii* (representing vascular plants) *CESAs* are found on the same clade as *A. thaliana* PCW *CESAs*, whereas surprisingly, the SCW clade does not contain genes found in *P. patens* or *S. moellendorfii* (Figure S6A-B). This result would indicate that either both *P. patens* and *S. moellendorfii* independently lost the SCW clade *CESAs*, or that the tree topology of *CESA* phylogenetic tree is incorrect. A study of *CESA* evolution indicated that PCW and SCW *CESAs* duplicated after the divergence

of land plants, along the lineage leading to *A. thaliana* and rice (Roberts and Roberts, 2004). We have therefore amended the phylogenetic edges stemming from the *CESA* tree to angiosperm-specific duplication edges, which resulted in PCW and SCW modules showing exclusively angiosperm-specfic edges (Figure 8C).

Taken together, by combining phylostratigraphic and comparative genomic approaches, these results indicated that PCW and SCW modules duplicated in the ancestor of angiosperms, while *P. patens* independently duplicated additional cell wall modules.

Discussion

Functional analysis of phylostrata

In this study we used a comparative genomics method, phylostratigraphy, to define the evolutionary origin of gene families (Figure 1). Surprisingly, our analysis clearly demonstrated an inverse relationship between the age and the percentage of characterized genes found in a given phylostratum (Figure 2), and this trend almost perfectly reflected the order of plant clades (Figure 1). While functional knowledge from *E. coli*, yeast, *Drosophila melanogaster* and human systems is used to annotate functions of many plant gene families by sequence similarity-based approaches (Quevillon et al., 2005), these gene families have to be present in the last common ancestors of plants and the species that are used to obtain functional knowledge. In our analysis, this corresponds to the oldest phylostratum, i.e. Green Plants. However, this trend was still visible for phylostrata that are plant-specific and which therefore could not be annotated by functional knowledge obtained from other

kingdoms of life, (e.g. Land Plants, Vascular Plants and others, Figure 2). We hypothesized that older gene families are more extensively studied, because (i) they tend to show stronger phenotypes in genetic screens (e.g. they represent an essential process, Figure 3), (ii) they are present in more plant species (and are consequently more likely to be investigated), and (iii) old genes tend to be more broadly expressed and are more likely to be identified by early gene detection methods, such as expressed sequence tags (Rutter et al., 2012; Guo, 2013). Finally, while microarray platforms are invaluable source to study gene function, they often lack genes from younger phylostrata, as only 31%, 60% and 56% of genes assigned to youngest phylostrata are found on the microarrays of *A. thaliana*, rice and *P. patens*, respectively (Figure S7). Since information gained from microarrays is often used to identify genes relevant for a given process, this underrepresentation could cause a bias towards selecting older genes for functional analysis.

While younger genes are under-investigated, they might represent an untapped source to study the yet uncharacterized pathways and identify a molecular basis for differences of species. We therefore propose to use co-expression networks (Table S6 and www.gene2function.de) to systematically investigate recently evolved modules. Such an approach would turn around the usual direction of biological research, where genes are characterized due to an association to a known biological pathway. Instead, novel pathways important for e.g. environmental adaptations could possibly be found with this approach.

Preferential co-expression of genes that appeared at the same evolutionary period

Phylostratigraphic analyses of biological processes showed that functionally related genes tend to be associated to a specific phylostratum (Figure 3), indicating that general biological processes, such as photosynthesis, glycolysis, DNA synthesis and others were already present in ancestors of green plants. Conversely, developmental programs, secondary metabolism, cell wall biosynthesis, stress responses, and regulation of gene expression appeared and expanded in land plants or later (Figure 3, Table S3). This is in line with a genomic analysis of *Selaginella moellendorfii*, which revealed a punctuated expansion of secondary metabolism in vascular plant lineage (Banks et al., 2011).

Since functionally related genes tend to be co-expressed, we hypothesized that genes assigned to the same phylostratum should be preferentially connected in the co-expression networks. Indeed, we found that genes from the same phylostratum tend to be connected (Figure 4) and consequently, form larger connected components than expected by chance (Figure 5). Conversely, we observed preferential dissociation between genes separated by large phylostratigraphic distances (e.g. Green Plants-*Arabidopsis*, Green Plants-*Oryza*, Figure 4), showing that younger genes tend not to be co-expressed with old genes. This shows that certain modules in the co-expression networks were created at a specific time in plant evolution, and we identified such modules by phylostratigraphic enrichment analysis (Figure 6, Table S6). It is important to note that our results do not mean that new biological pathways are completely separated from existing pathways in the co-expression network. For example, the evolution of cellulosic cell walls in land plants is integrated into existing factors that are important for cellular growth, such as actin and microtubules (Ruprecht et al., 2011), and other studies have experimentally shown the

importance of actin and microtubules for cellulose synthesis (Gutierrez et al., 2009; Sampathkumar et al., 2013). Consistent with this, we found several significant associations of adjacent phylostrata, such as GP-LP and LP-VP in rice (Figure 4E).

New insights gained from combining phylostratigraphy, phylogenetics and co-expression networks

To elucidate the emergence time and evolutionary relationships between modules, we have updated PlaNet (www.gene2function.de) with phylostratigraphic and phylogenetic data, respectively. As a proof of concept, we used photosynthesis to show how our combined approach can reveal evolution of modules (Figure 7).

To demonstrate how our approach can be used to gain new insights, we analyzed cell wall biosynthesis modules of *P. patens* and *A. thaliana*. Our phylostratigraphic results indicated that the cell wall module arose in an ancestor of land plants, which is in line with the absence of cellulosic cell wall in chlorophytes (Domozych et al., 2012). However, since our analysis does not include charophytes (Figure 1), it is unclear whether the cell wall modules appeared in the ancestor of charophytes or bryophytes. Based on the presence of cellulosic cell walls in charophytes (Popper and Fry, 2003; Sarkar et al., 2009; Harholt et al., 2016), it is likely that the cell wall module arose in this lineage. Addition of a charophyte genome, such as *Klebsormidium flaccidum* (Hori et al., 2014), will provide a better phylostratigraphic resolution for future studies.

Higher plants contain two cell wall types (PCW and SCW) which are synthesized by two duplicated gene modules present in angiosperms (Brown et al., 2005; Ruprecht et al., 2011). Our finding of three cell wall modules in *P. patens* could have supported that these duplicated modules were already present in the ancestor of land plants. However, phylogenetic analysis of the *P. patens* modules indicated that they were duplicated independently in the lineage leading to *P. patens* (Figure 8C), suggesting that the ancestor of land plants contained one cell wall module and that angiosperms and bryophytes independently duplicated these modules. While the function of the three *P. patens* cell wall modules is currently unknown, the expression patterns of representative genes in *P. patens* suggested that these modules could have acquired distinct functions, similarly to the PCW and SCW modules found in flowering plants (Ruprecht et al., 2016). To verify this hypothesis, genes in the *P. patens* modules could be functionally analysed in relation to cell wall synthesis and expansion. Furthermore, including additional bryophyte genomes, will help to establish whether the three cell wall modules found in *P. patens* were duplicated in the bryophyte lineage, or if they are *P. patens*-specific.

In our example of the duplicated cell wall biosynthesis modules, the analysis revealed incorrect phylogenetic relationships based on one phylogenetic tree. This clearly highlighted that in some cases individual phylogenetic trees derived from high-throughput generated data might be erroneous and misleading. However, a major advantage of our approach is that the age and phylogenetic relationships of the modules are inferred by multiple gene families and phylogenetic trees, which should make the evolutionary inferences more robust to artefacts that arise due to large-scale construction of gene families and phylogenetic trees (Figure S6).

Finally, the limitation of purely genomic approaches is that the functional context of the younger genes is unknown. As biological functions often require an intricate and coordinated co-operation of multiple gene products, a more complete understanding of evolution can only be obtained by taking these relations into account.

Materials and Methods

Estimating phylostrata for genes and gene families

Gene family data for 25 plant species was downloaded from the PLAZA2.5 database (Van Bel et al., 2012) as file genefamily_data.hom.csv.gz from http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v2_5/download/index. The species used in this analysis include 5 chlorophytes (*Chlamydomonas reinhardtii*, *Ostreococcus lucimarinus*, *Ostreococcus tauri*, *Micromonas sp. RCC299*, *Volvox carteri*), 1 bryophyte (*Physcomitrella patens*), 1 lycophyte (*Selaginella moellendorfii*), 5 monocots (*Oryza sativa ssp. Indica, Oryza sativa ssp. Japonica, Brachypodium distachyon, Sorghum bicolor, Zea mays*) and 13 dicots (*Lotus japonicus, Medicago truncatula, Glycine max, Malus domestica, Fragaria vesca, Manihot esculenta, Ricinus communis, Populus trichocarpa, Arabidopsis thaliana, Arabidopsis lyrata, Carica papaya, Theobroma cacao, Vitis vinifera*). To determine the phylostratum of each family, we investigated which species are present and absent in each family (Guo, 2013). For example, if a family contained any genes from green algae, the family was deemed to be created during “green plant” phylostratum. Conversely, if the family contained genes found in monocots and dicots only, the family was assigned to the “Angiosperm” phylostratum. Table S1 contains gene IDs, family and phylostratum assignments for each gene. Since PLAZA2.5 is based on *P. patens* v1.2 genome and

Nimblegen arrays are based on v1.6 genome, we used Table S10 (obtained from http://www.cosmoss.org/) to annotate the Nimblegen arrays with PLAZA gene family information. Overall, 9340 out of 32569 (28.6%) V1.6 Nimblegen loci present on the Nimblegen microarray could not be unambiguously mapped to V1.2 PLAZA2.5 loci. TAIR10 and MSU RGAP 6.1 genome versions were used for *A. thaliana* and rice, respectively.

Percentage of functionally characterized genes in *Arabidopsis thaliana*

A table containing experimentally characterized genes was obtained from TAIR, (https://www.arabidopsis.org).

Phylostrata enrichment within MapMan bins

Phylostrata enrichment within Mapman bins was calculated using hypergeometric distribution. This method estimates the probability of obtaining k successes (i.e. number of a specific phylostrata in a Mapman bin) in n draws (i.e. number of genes in a Mapman bin), from a finite population of size N (i.e. number of genes in the genome) that contains exactly K successes (i.e. number of times the phylostrata of interest is found in the genome). To correct for multiple hypothesis testing, we used Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), at significance level of 0.05.

Microarray normalization, functional annotation and construction of co-expression networks for *P. patens*

We have obtained the Nimblegen microarray data for *P. patens* from ArrayExpress (Table S11). Nimblegen array data (32569 probesets designed with *P. patens* genome version v1.6) was RMA-normalized by DEVA software package (Roche, Nimblegen). Highest reciprocal rank (HRR) co-expression networks for the two datasets were calculated, as described in Mutwil et al. (2010). HRR co-expression networks for *A. thaliana* and rice were obtained from the PlaNet database (Mutwil et al., 2011). Gene Ontology annotations for *P. patens* were downloaded from (http://www.cosmoss.org/). The interactive networks are implemented using Cytoscape Web (Lopes et al., 2010). Additional quality control of microarrays revealed no large outliers when clustering the microarray profiles (Figure S8). The normalized expression matrix and HRR-based co-expression networks are available for download at http://www.gene2function.de/download.html. Co-expression network neighborhoods of *P. patens* were compared to the co-expression networks present in PlaNet database, by using the FamNet platform (described in Ruprecht et al., 2016). The resulting conserved and multiplied modules can be downloaded from www.gene2function.de/download.html.

Estimating performance of *P. patens* co-expression networks

To estimate the performance of the *P. patens* co-expression networks and to determine the optimal HRR value cut-off, we estimated how well the co-expression network predicts gene function (Figure S9). Such an approach usually necessitates gold standard data in form of

experimentally characterized genes (Rhee and Mutwil, 2014). While the majority of *P. patens* gene annotations are *in silico* predictions (Gene Ontology evidence code IEA) based on sequence similarity annotations (e.g. blast2go, (Conesa and Götz, 2008)), co-expression analyses have not been used to predict gene function in this species. We therefore assume that (i) majority of the *in silico* annotations are correct and (ii) the annotations are not derived from co-expression analyses in *P. patens*. We use a simple Neighborhood Counting method to estimate gene function (Schwikowski et al., 2000). For each node in the network, functions present in the first neighborhood are counted. For example, given a gene with function A that is connected to three genes with function A and two genes with function B, the number of counted functions is 3A and 2B. A correct prediction is made when the most frequent function in the first neighborhood is equal to the function of the query node, which is the case in this example. To obtain a numeric estimate of performance, we use F-measure, which is defined as F = 2 * ((*precision***recall*)/(*precision* + *recall*)), where *precision* = (#*correct predictions*)/(#*made predictions*) and recall = (#*correct predictions*)/(#*feasible predictions*). #*Made predictions* is equal to the number of nodes with edges at a given HRR cut-off. #*Feasible predictions* is equal to the number of genes with assigned function. To remove general GO terms (e.g. “biological process”, “plastid”), terms that are present in >10% of genes were removed from the analysis (Table S12).

Estimating association between phylostrata

The analysis was performed by counting the phylostratum edges between co-expressed gene pairs (Figure S1A). For example, two co-expressed genes belonging to the GP phylostratum, would produce 1 GP-GP association. Since plants employ large gene families,

co-expressed genes that belong to the same family would lead to an overestimation of the number of associations between genes from the same phylostratum. To account for this bias, our analysis only considers gene pairs that belong to different families. To estimate whether the observed number of association edges is significantly larger or smaller than expected by chance, we performed permutation analysis, where the gene-phylostratum assignments were shuffled 1000 times (Figure S1A). For each permutation, the number of phylostratum associations was counted. The empirical P-value for enrichment is given by the number of scores from the permuted networks which are larger than the score from the original network. For example, if the observed number of GP-GP edges was 100 and for 563 of the 1000 randomized networks the value was larger than 100, the empirical P-value is 0.563. To correct for multiple hypothesis testing, we used Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), at significance level of 0.05.

Estimating sizes of connected components

Starting from each gene in a given phylostratum, our algorithm iteratively collects genes from the same phylostratum that are connected to it (Figure S1B). The collection stops until no more genes from the same phylostratum can be collected. The outcome of the analysis is a collection of groups of genes that belong to the same phylostratum and are connected in the networks. From this collection, the largest group of genes from a given phylostratum is used to represent the connected component. To estimate whether the observed connected components are larger or smaller than expected by chance, we have kept the network topology constant, but shuffled the gene-phylostratum assignments 1000 times and estimated the size of the largest connected component per phylostratum for each

permutation (Figure S1B). The empirical P-value was calculated by estimating how many times the observed connected components were larger or smaller than the connected components in the permuted network. To correct for multiple hypothesis testing, we used Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), at significance level of 0.05.

Estimating neighborhoods enriched for different phylostrata

For each first neighborhood in the analyzed co-expression networks, the number of found phylostrata was determined by first isolating the set of families present in a neighborhood. By using gene families for the estimation, rather than genes, we account for the large gene families found in plants. For example, consider a neighborhood of 10 genes, with 8 of the genes assigned to family A (created in Green Plant lineage), 1 of the genes assigned to family B (created in Land Plant lineage) and 1 of the genes assigned to family C (created in Angiosperm lineage). By counting genes rather than gene families, we would obtain AAAAAAAABC, giving us 8 x Green Plant, 1 x Land Plant, 1 x Angiosperm counts. This result, caused by the large size of gene family A, would likely indicate that the neighborhood arose in the Green Plant lineage. Conversely, by isolating the set of gene families, we would obtain 1 x Green Plant, 1 x Land Plant, 1 x Angiosperm counts, which are not likely to be indicated as enriched for Green Plant lineage. Similarly to the previous statistical analyses, the empirical P-value describing significance of enrichment or depletion of a phylostratum was performed by permuting the family-phylostrata assignments 1000 times. To correct for multiple hypothesis testing, we used Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), at significance level of 0.05.

Gene Ontology and expression profile analysis of modules

GO enrichment analysis of a module is performed by using hypergeometric distribution, which estimates the probability of obtaining k successes (i.e. number of genes in a module with GO term of interest) in n draws (i.e. number of genes in a module), from a finite population of size N (i.e. number of genes in the genome) that contains exactly K successes (i.e. number of times the GO term of interest is found in the genome). Expression profiles, co-expression neighborhoods, GO enrichment analyses and FamNet analyses of discussed modules can be accessed online by querying PlaNet with the respective gene identifier. For example, analysis of *Loc_os01g09510* can be found at http://www.gene2function.de/responder.py?name=gene!osa!24234.

Obtaining phylogenetic speciation/duplication data

Phylogenetic trees representing PLAZA HOM and ORTHO families were used to (i) estimate speciation or duplication relationship between genes and to (ii) elucidate the phylostratum of the speciation/duplication event (Table S3). For each gene pair in the tree, the first common ancestor node with bootstrap (BS) >50 was isolated and used to estimate the phylostratum. The duplication Consistency score (C, (Van Bel et al., 2012)) was used to differentiate between speciation and duplication events, where duplication was defined by C>0.05 (Van Bel et al., 2012). Table S7 contains gene pairs, BS and C values, event types, phylostratum, trees and identity of first common ancestor nodes.

Phylogenomic analysis of photosynthesis and cell wall modules

We provide a step-by-step tutorial on how to replicate the evolutionary analysis of photosynthesis modules at http://www.gene2function.de/publications.html. The cell wall module analysis is done by the same procedure.

Acknowledgments

M.M. conceived the project, M.M. performed the analyses, M.M., C.R., S.P., K.V., D.L. and S.A.R wrote the article with help from all authors. M.H.C, C.O.R and J.D.B. provided early access to the expression data of *Physcomitrella patens*. This work was supported by the Max-Planck-Gesellschaft (M.M., S.P., C.R), and ERA-CAPS grant EVOREPRO (S.P.). We would like to thank Bjoern Oest Hansen and Manuel Hiss for their initial analysis of *P. patens* arrays. We would also like to thank Daniela Geisler, Staffan Persson and Mark Stitt for their useful comments. Finally, we would like to thank the ERA-CAPS EVOREPRO consortium for funding. The authors declare no conflicts of interest.

Short Supporting Information Legends

Figure S1. Permutation analyses.

Figure S2. *A. thaliana* connected components.

Figure S3. *P. patens* connected components.

Figure S4. Expression profiles of genes found in the *P. patens* cell wall modules *Pp1s144_8V6.1* (red lines), *Pp1s60_109v6.1* (blue lines) and *Pp1s175_122v6.1* (black lines) shown in Figure 8.

Figure S5. PlaNet output of the 5 cell wall modules found in *A. thaliana* and *P. patens*.

Figure S6. PLAZA Phylogenetic tree of Cellulose synthase (CESA) and Cellulose Synthase-like gene family (CSL).

Figure S7. Representation of phylostrata on used microarrays.

Figure S8. Hierarchical clustering of *P. patens* Nimblegen microarrays.

Figure S9. HRR cut-off versus F-measure and density of the *Physcomitrella patens* co-expression network.

Table S1. Gene (first column) to family (second column) to phylostrata (third column) assignments for *A. thaliana*, rice and *P. patens*.

Table S2. Percentage of functionally characterized genes per phylostratum.

Table S3. Enrichment or depletion of phylostrata within Mapman terms for *A. thaliana* (Table S3A), rice (Table S3B) and *P. patens* (Table S3C).

Table S4. Statistical analysis of connections between phylostrata for *A. thaliana* (Table S74A), rice (Table S4B) and *P. patens* (Table S4C). The "Observed" column and "Mean of 1000 permutations" column show the number of edges in observed and permuted networks, respectively.

Table S5. Statistical analysis of largest connected components for *A. thaliana* (Table S5A), rice (Table S5B) and *P. patens* (Table S5C). The "Observed" column and "Mean of 1000 permutations" column show the number of edges in observed and permuted networks, respectively.

Table S6. Enrichment of phylostrata in first neighborhoods of *A. thaliana* (Table S6A, left), rice (Table S6B, middle) and *P. patens* (Table S6C, right). The positive and negative P-values indicate enrichment and depletion of phylostrata, respectively.

Table S7. Speciation/duplication events found between genes from *A. thaliana*, rice and *P. patens*.

Table S8. Speciation/duplication events found between photosynthetic modules (Table S8A) and enrichment of phylostrata in the photosynthetic modules (Table S8B). Positive and negative P-values indicate enrichment and depletion, respectively.

Table S9. Speciation/duplication events found between cell wall modules (Table S9A) and enrichment of phylostrata in the cell wall modules (Table S9B).

Table S10. Conversion table between *P. patens* genome V1.2 and V1.6.

Table S11. PlaNet microarray identified (first column) and identifier found on ArrayExpress (second column).

Table S12. Frequency of Gene Ontology terms in the *P. patens* genome V1.6. terms that are present in more than 10% of genes are marked red.

References

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

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**: 960–3

Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol **158**: 590–600

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B **57**: 289–300

De Bodt S, Hollunder J, Nelissen H, Meulemeester N, Inzé D (2012) CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations. New Phytol **195**: 707–20

Brown DM, Zeef LAH, Ellis J, Goodacre R, Turner SR (2005) Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. Plant Cell **17**: 2281–95

Busch BL, Schmitz G, Rossmann S, Piron F, Ding J, Bendahmane A, Theres K (2011) Shoot branching and leaf dissection in tomato are regulated by homologous gene modules. Plant Cell **23**: 3595–609

Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics **2008**: 619832

Cove DJ, Knight CD, Lamparter T (1997) Mosses as model systems. Trends Plant Sci **2**: 99–105

Cui X, Lv Y, Chen M, Nikoloski Z, Twell D, Zhang D (2015) Young Genes out of the Male: An Insight from Evolutionary Age Analysis of the Pollen Transcriptome. Mol Plant **8**: 935–45

Domazet-Loso T, Brajković J, Tautz D (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet **23**: 533–9

Domozych DS, Ciancia M, Fangel JU, Mikkelsen MD, Ulvskov P, Willats WGT (2012) The Cell Walls of Green Algae: A Journey through Evolution and Diversity. Front Plant Sci **3**: 82

Fang G, Passalacqua KD, Hocking J, Llopis PM, Gerstein M, Bergman NH, Jacobs-Wagner C (2013) Transcriptomic and phylogenetic analysis of a bacterial cell cycle reveals strong associations between gene co-expression and evolution. BMC Genomics **14**: 450

Ficklin SP, Feltus FA (2011) Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. Plant Physiol **156**: 1244–1256

Gerstein MB, Rozowsky J, Yan K-K, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, et al (2014) Comparative analysis of the transcriptome across distant species. Nature **512**: 445–8

Guo Y-L (2013) Gene family evolution in green plants with emphasis on the origination and evolution of Arabidopsis thaliana genes. Plant J **73**: 941–51

Gutierrez R, Lindeboom JJ, Paredez AR, Emons AMC, Ehrhardt DW (2009) Arabidopsis cortical microtubules position cellulose synthase delivery to the plasma membrane and interact with cellulose synthase trafficking compartments. Nat Cell Biol **11**: 797–806

Hansen BO, Vaid N, Musialak-Lange M, Janowski M, Mutwil M (2014) Elucidating gene function and function evolution through comparison of co-expression networks of plants. Front Plant Sci **5**: 1–9

Hardison RC (2003) Comparative genomics. PLoS Biol. doi: 10.1371/journal.pbio.0000058

Harholt J, Moestrup Ø, Ulvskov P (2016) Why Plants Were Terrestrial from the Beginning. Trends Plant Sci **21**: 96–101

Hartwell LH, Hopfield JJ, Leibler S, Murray AW From molecular to modular cell biology.

He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics **169**: 1157–1164

Heyndrickx KS, Vandepoele K (2012) Systematic Identification of Functional Plant Modules through the Integration of Complementary Data Sources. PLANT Physiol **159**: 884–901

Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, Seo M, Sato S, Yamada T, Mori H, Tajima N, et al (2014) Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. Nat Commun **5**: 3978

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature **436**: 793–800

Itkin M, Heinig U, Tzfadia O, Bhide a J, Shinde B, Cardenas PD, Bocobza SE, Unger T, Malitsky S, Finkers R, et al (2013) Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. Science (80-) **341**: 175–9

Kliebenstein DJ (2001) Gene Duplication in the Diversification of Secondary Metabolism: Tandem 2-Oxoglutarate-Dependent Dioxygenases Control Glucosinolate Biosynthesis in Arabidopsis. PLANT CELL ONLINE **13**: 681–693

Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res **14**: 1085–94

Lee T, Yang S, Kim E, Ko Y, Hwang S, Shin J, Shim JE, Shim H, Kim H, Kim C, et al (2015) AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. Nucleic Acids Res **43**: D996–1002

Loll B, Kern J, Saenger W, Zouni A, Biesiadka J (2005) Towards complete cofactor arrangement in the 3.0 A resolution structure of photosystem II. Nature **438**: 1040–4

Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD (2010) Cytoscape Web: an interactive web-based network browser. Bioinformatics **26**: 2347–8

Lynch M, Katju V (2004) The altered evolutionary trajectories of gene duplicates. Trends Genet **20**: 544–549

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

Matsuzaki M, Misumi O, Shin-I T, Maruyama S, Takahara M, Miyagishima S-Y, Mori T, Nishida K, Yagisawa F, Nishida K, et al (2004) Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. Nature **428**: 653–657

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz J, Witman GB, Terry A, Salamov A, Fritz-laylin LK, Maréchal-drouard L, et al (2010) The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions. Science (80-) **318**: 245–250

Movahedi S, Van de Peer Y, Vandepoele K (2011) Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice. Plant Physiol **156**: 1316–1330

Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S (2011) PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. Plant Cell **23**: 895–910

Mutwil M, Usadel B, Schütte M, Loraine A, Ebenhöh O, Persson S (2010) Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. Plant Physiol **152**: 29–43

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature **497**: 579–84

Obayashi T, Nishida K, Kasahara K, Kinoshita K (2011) ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. Plant Cell Physiol **52**: 213–219

Park CY, Wong AK, Greene CS, Rowland J, Guan Y, Bongo LA, Burdine RD, Troyanskaya OG (2013) Functional knowledge transfer for high-accuracy prediction of under-studied biological processes. PLoS Comput Biol **9**: e1002957

Persson S, Paredez A, Carroll A, Palsdottir H, Doblin M, Poindexter P, Khitrov N, Auer M, Somerville CR (2007) Genetic evidence for three unique components in primary cell-wall cellulose synthase complexes in Arabidopsis. Proc Natl Acad Sci U S A **104**: 15566–15571

Persson S, Wei H, Milne J, Page GP, Somerville CR (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci U S A **102**: 8633–8

Popper ZA, Fry SC (2003) Primary cell wall composition of bryophytes and charophytes. Ann Bot **91**: 1–12

Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber AP, Schwacke R, Gross J, Blouin NA, Lane C, et al (2012) Cyanophora paradoxa genome elucidates origin of photosynthesis in algae and plants. Science (80-) **335**: 843–847

Project AG (2013) The Amborella genome and the evolution of flowering plants. Science (80-) **342**: 1241089

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. Plant Cell **21**: 3718–3731

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan:

Protein domains identifier. Nucleic Acids Res. doi: 10.1093/nar/gki442

Quint M, Drost H-G, Gabel A, Ullrich KK, Bönn M, Grosse I (2012) A transcriptomic hourglass in plant embryogenesis. Nature **490**: 98–101

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud P-F, Lindquist EA, Kamisugi Y, et al (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science **319**: 64–9

Rhee SY, Mutwil M (2014) Towards revealing the functions of all genes in plants. Trends Plant Sci **19**: 212–21

Roberts AW, Roberts E (2004) Cellulose synthase (CesA) genes in algae and seedless plants. Cellulose **11**: 419–435

Romero-Campero FJ, Lucas-Reina E, Said FE, Romero JM, Valverde F (2013) A contribution to the study of plant development evolution based on gene co-expression networks. Front Plant Sci **4**: 291

Rubin GM (2000) Comparative Genomics of the Eukaryotes. Science (80-) **287**: 2204–2215

Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG (2014) From algae to angiosperms - inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol Biol **14**: 23

Ruprecht C, Mendrinna A, Tohge T, Sampathkumar A, Klie S, Fernie AR, Nikoloski Z, Persson S, Mutwil M (2016) FamNet: A Framework to Identify Multiplied Modules Driving Pathway Expansion in Plants. Plant Physiol **170**: 1878–94

Ruprecht C, Mutwil M, Saxe F, Eder M, Nikoloski Z, Persson S (2011) Large-Scale Co-Expression Approach to Dissect Secondary Cell Wall Formation Across Plant Species. Front Plant Sci **2**: 1–13

Rutter MT, Cross K V, Van Woert PA (2012) Birth, death and subfunctionalization in the Arabidopsis genome. Trends Plant Sci **17**: 204–12

Sampathkumar A, Gutierrez R, McFarlane HE, Bringmann M, Lindeboom J, Emons A-M, Samuels L, Ketelaar T, Ehrhardt DW, Persson S (2013) Patterning and lifetime of plasma membrane-localized cellulose synthase is dependent on actin organization in Arabidopsis interphase cells. Plant Physiol **162**: 675–88

Sarkar P, Bosneaga E, Auer M (2009) Plant cell walls throughout evolution: Towards a molecular understanding of their design principles. J Exp Bot **60**: 3615–3635

Schwikowski B, Uetz P, Fields S (2000) A network of protein-protein interactions in yeast. Nat Biotechnol **18**: 1257–61

Shiu SH, Bleecker AB (2001) Plant receptor-like kinase gene family: diversity, function, and signaling. Sci STKE **2001**: re22

Shiu S-H, Shih M-C, Li W-H (2005) Transcription factor families have much higher expansion rates in plants than in animals. Plant Physiol **139**: 18–26

Somerville C, Somerville S (1999) Plant functional genomics. Science **285**: 380–3

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science **302**: 249–55

Taylor NG, Howells RM, Huttly AK, Vickers K, Turner SR (2003) Interactions among three distinct CesA proteins essential for cellulose synthesis. Proc Natl Acad Sci U S A **100**: 1450–5

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M (2004) MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J **37**: 914–939

Tzfadia O, Diels T, De Meyer S, Vandepoele K, Aharoni A, Van de Peer Y (2016) CoExpNetViz: Comparative Co-Expression Networks Construction and Visualization Tool. Front Plant Sci **6**: 1194

Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ (2009) Co-expression tools for plant biology: Opportunities for hypothesis generation and caveats. Plant, Cell Environ **32**: 1633–1651

Vandepoele K, Van de Peer Y (2005) Exploring the plant transcriptome through phylogenetic profiling. Plant Physiol **137**: 31–42

Vermeirssen V, De Clercq I, Van Parys T, Van Breusegem F, Van de Peer Y (2014) Arabidopsis Ensemble Reverse-Engineered Gene Regulatory Network Discloses Interconnected Transcription Factors in Oxidative Stress. Plant Cell **26**: 4656–4679

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res **19**: 327–335

Yu H, Luscombe NM, Qian J, Gerstein M (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet **19**: 422–427

Zarrineh P, Sánchez-Rodríguez A, Hosseinkhan N, Narimani Z, Marchal K, Masoudi-Nejad A (2014) Genome-scale co-expression network comparison across Escherichia coli and Salmonella enterica serovar Typhimurium reveals significant conservation at the regulon level of local regulators despite their dissimilar lifestyles. PLoS One **9**: e102871

Figure legends

Figure 1. Species tree of plants. A) Species tree showing major plant lineages. The numbers below the tree nodes indicate the number of gene families and genes assigned to a phylostratum, for *A. thaliana* (Ath), rice (Osa) and *P. patens* (Ppa). For example, Ath: 3948, 15715 indicates that *A. thaliana* has 3948 gene families and 15714 genes assigned to the Green Plant phylostratum. Grayed-out clades are not present in the PLAZA database and are not included in the phylostratigraphic analysis. B) Species tree of angiosperms. The BEP clade (also called BOP clade) includes sub-families Bamboos, Oryzoideae and Pooideae,

while the PACMAD clade includes Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae, and Danthonioideae sub-families.

Figure 2. Percentage of genes in phylostrata annotated with experimental evidence. A) Percentage (y-axis) of functionally characterized genes for *Arabidopsis thaliana* phylostrata (x-axis). B) Percentage of functionally characterized genes (y-axis) derived from IDA (inferred from direct assay), IPI (inferred from protein interaction), IMP (inferred from mutant phenotypes), IGI (inferred from genetic interaction) and IEP (inferred from expression pattern).

Figure 3. Functional analysis of phylostrata. The rows in the enrichment matrix represent MapMan terms, while columns indicate phylostrata for *A. thaliana* (Ath), rice (Osa) and *P. patens* (Ppa). Red cells indicate instances where a given phylostratum is significantly enriched ($P<0.05$) within a MapMan bin (e.g. Green Plant phylostratum in photosynthesis bin). Blue and gray cells indicate a significant depletion ($P<0.05$) or no significant enrichment or depletion within the MapMan bins, respectively.

Figure 4. Association between phylostrata in the co-expression networks. A) An example of a network consisting of three Green Plant (GP) and one vascular plant (VP) genes. The network contains 3 GP-GP and 1 GP-VP edge. B) Number of observed (black bars) Green Plant-Green Plant (GP-GP) edges compared to an average number GP-GP edges found in 1000 networks with shuffled gene-phylostratum assignments (white bars). Error bars denote

standard deviation of the shuffled observations. C) Number of observed (black bars) edges compared to shuffled (white bars) gene-phylostratum assignments for GP-AT (*A. thaliana*), GP-OS (*O. sativa*) and GP-PP (*P. patens*). The asterisks indicate cases where the number of edges in the network was significantly (P<0.05) larger or smaller than in the shuffled gene-phylostratum assignments, respectively. D-F) Lower triangular heatmap indicates significant (BH<0.05) association (red) and dissociation (blue) of all phylostratum combinations in *A. thaliana* (D) rice (E) and *P. patens* (F). The upper triangular matrix indicates the empirical P-values.

Figure 5. Sizes of connected components in rice. A) Depiction of a genome-wide network of rice. Nodes represent genes, while node color indicates the phylostratum assignment of a gene (see legend below). For brevity, only connected gene pairs that belong to the same phylostratum are shown. B) Observed size of largest connected component per phylostratum (indicated by filled bars), as compared to a randomized analysis with 1000 permutations, where gene-phylostratum assignments have been shuffled (white bars). Error bars denote standard deviation, while red asterisk indicate significant (P<0.05) difference between observed and permuted sizes.

Figure 6. Examples of neighborhoods enriched for the different phylostrata. A) Green Plants phylostratum enriched neighborhood of *P. patens* Pp1s38_194v6.1 (large central node). Nodes present genes, grey edges indicate co-expression edges, while genes with GO term “translation” are colored blue. The phylostratigraphic information is visualized as node border colors, where e.g. the Green Plants phylostratum is indicated by green borders, the

Land Plants phylostratum by red borders, and the Vascular Plants phylostratum by blue borders (see edge color legend below the figure). Note that the colors are contextual based on the species, where orange border indicates a monocot- and dicot-specific phylostrata for rice and *Arabidopsis thaliana*, respectively. Rosids, brassicales, malvids and BEP clade are represented by brown color only, since only few genes are assigned to these phylostrata. B) Land Plants phylostratum enriched neighborhood from *Arabidopsis thaliana*. Nodes colored blue and orange have assigned GO terms "ion transport" and "cell wall organization or biogenesis", respectively. C) Monocots phylostratum enriched neighborhood from rice. Nodes colored blue have assigned GO term "embryo development". D) AT phylostratum enriched neighborhood from *Arabidopsis thaliana*. Nodes colored blue have assigned GO terms "carpel morphogenesis.

Figure 7. Phylostratigrapic and phylogenetic analysis of photosynthesis modules. A) Photosynthetic gene modules for *A. thaliana*, rice and *P. patens*. Nodes represent genes, while colored edges represent timing of speciation events found between modules. Nodes with same shapes and colors indicate genes that belong to the same family and contain same Pfam domains. The description of the colored shapes is given to the right of the figure; for brevity, only Pfam families are shown. B) Description of the elements displayed in the gene module pages. Colored shapes indicate which genes belong to the same family and contain same Pfam domains, edge styles indicate speciation/duplication events and edge colors indicate the phylostratum of the events. C) An evolutionary model of the photosynthetic modules of *A. thaliana*, rice and *P. patens*. The colored shapes correspond to the gene families shown in the modules in Figure 7A.

Figure 8. Evolutionary analysis of cell wall modules. A) Average expression profiles of the genes present in the three *P. patens* cell wall modules provided by PlaNet. Error bars indicate standard error for 32, 21 and 48 genes found in the *Pp1s144_8v6.1*, *Pp1s60_109v6.1* and *Pp1s175_122v6.1* modules, respectively. DAF (days after fertilization), KO (knock-out). B) Comparison of five cell wall modules from *A. thaliana* and *P. patens*. All five modules are significantly enriched ($P<0.05$) for genes generated in the Land Plants phylostratum (denoted by red borders). C) Evolutionary model of the five cell wall modules from *A. thaliana* and *P. patens*. The colored shapes indicate which gene families are present in the modules.

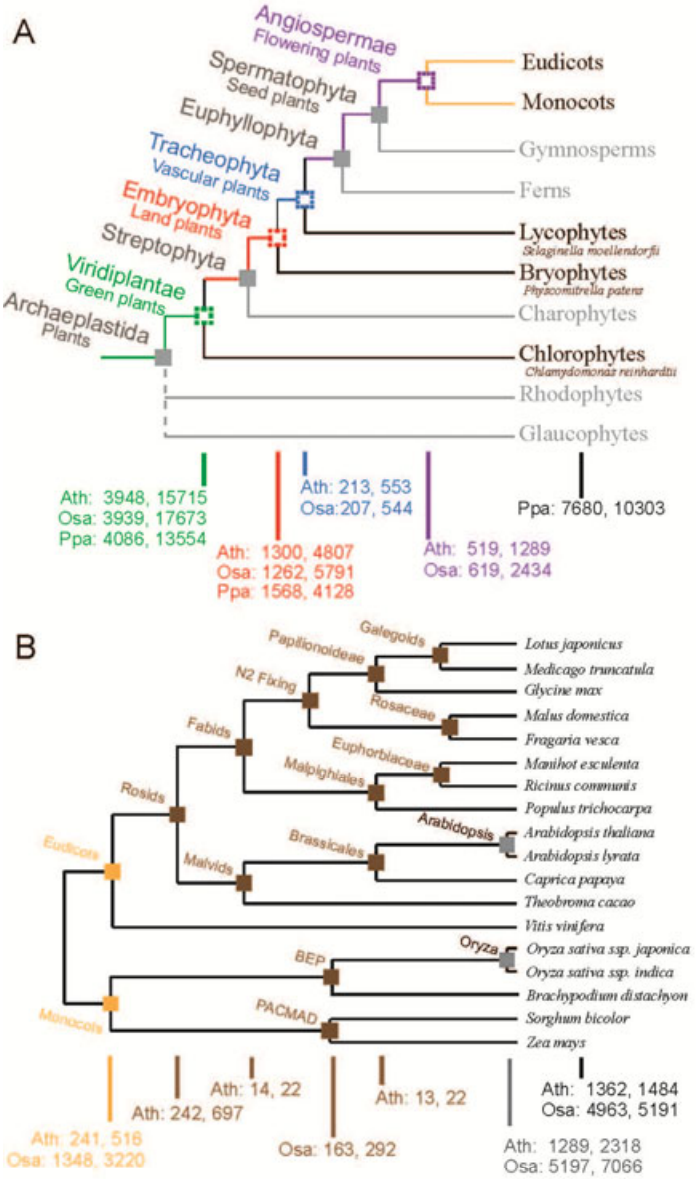

Figure 1. Species tree of plants. A) Species tree showing major plant lineages. The numbers below the tree nodes indicate the number of gene families and genes assigned to a phylostratum, for Arabidopsis (Ath), rice (Osa) and Physcomitrella (Ppa). For example, Ath: 3948, 15715 indicates that Arabidopsis has 3948 gene families and 15714 genes assigned to the Green Plant phylostratum. Grayed-out clades are not present in the PLAZA database and not included in the phylostratigraphic analysis. B) Species tree of angiosperms. BEP clade (also called BOP clade) includes sub-families Bamboos, Oryzoideae and Pooideae, while PACMAD clade includes Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae, and Danthonioideae sub-families.

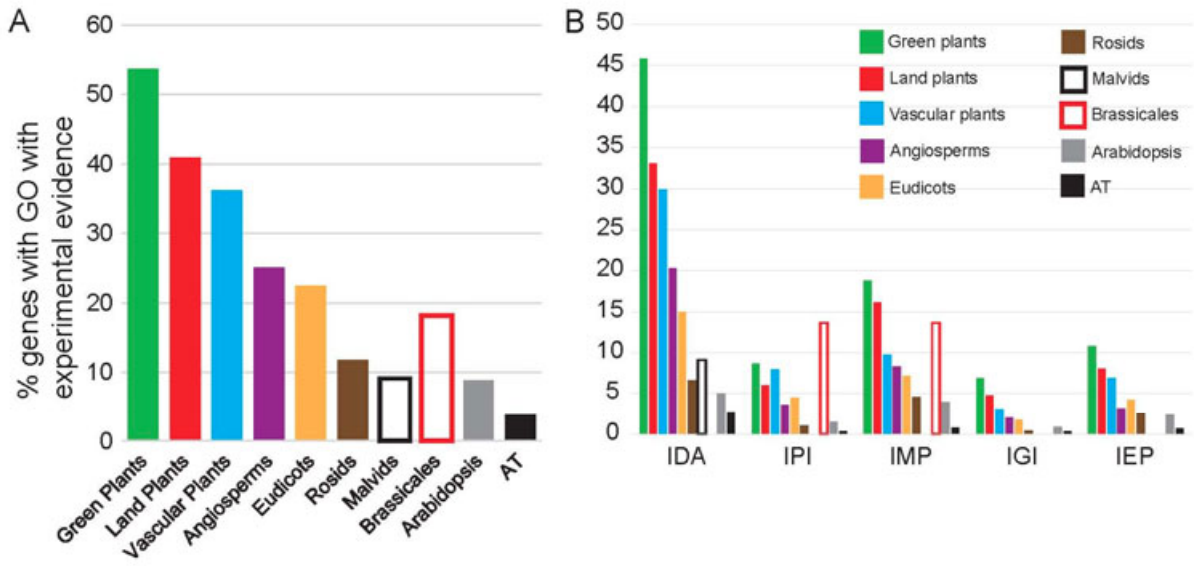

Figure 2. Percentage of genes in phylostrata annotated with experimental evidence. A) Percentage (y-axis) of functionally characterized genes for *Arabidopsis thaliana* phylostrata (x-axis). B) Percentage of functionally characterized genes (y-axis) derived from IDA (inferred from direct assay), IPI (inferred from protein interaction), IMP (inferred from mutant phenotypes), IGI (inferred from genetic interaction) and IEP (inferred from expression pattern).

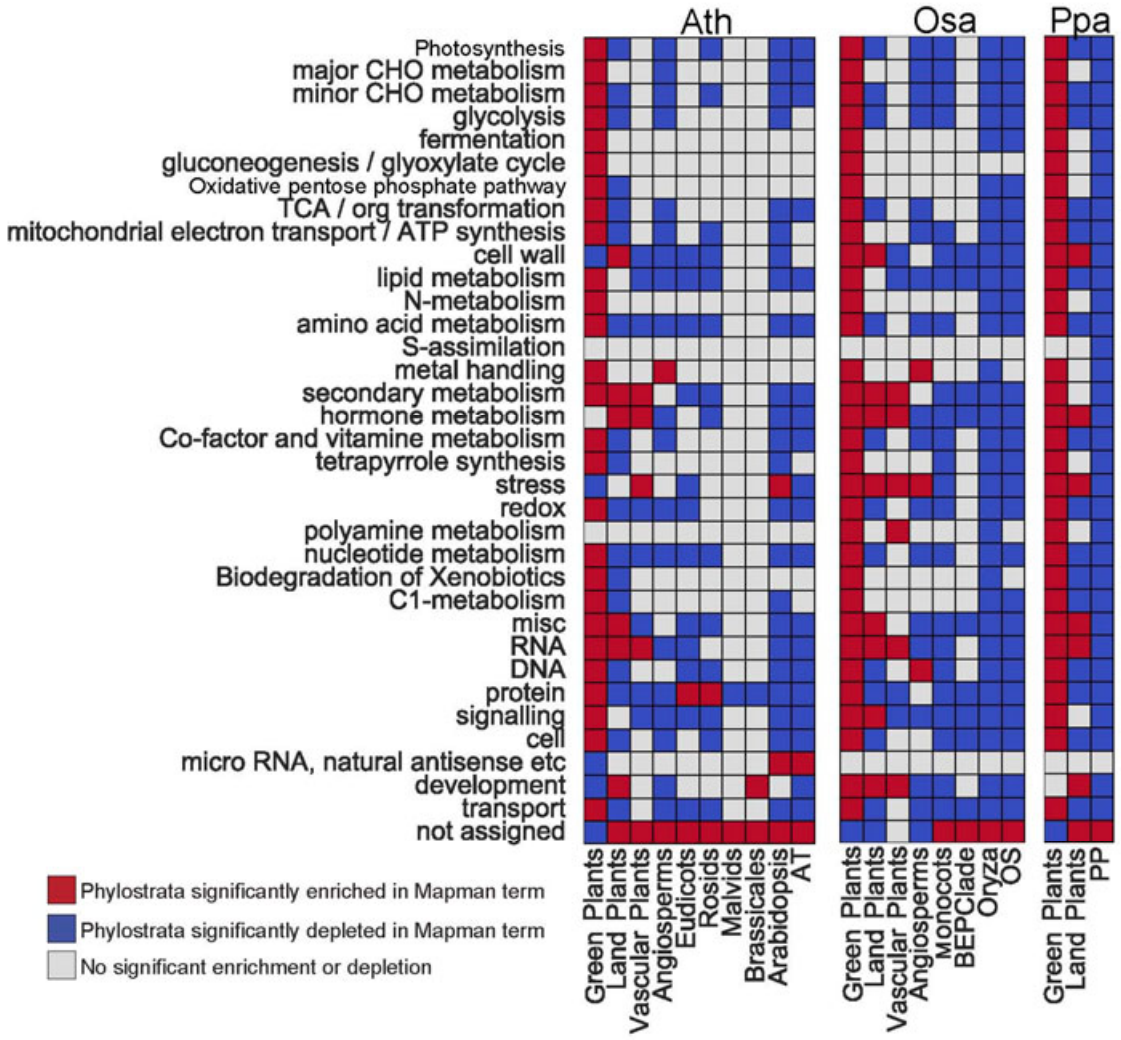

Figure 3. Functional analysis of phylostrata. The rows in the enrichment matrix represent MapMan terms, while columns indicate phylostrata for *A. thaliana* (Ath), rice (Osa) and *P. patens* (Ppa). Red cells indicate instances where a given phylostratum is significantly enriched ($P<0.05$) within a MapMan bin (e.g. Green Plant phylostratum in photosynthesis bin). Blue and gray cells indicate a significant depletion or no significant enrichment within the MapMan bins, respectively.

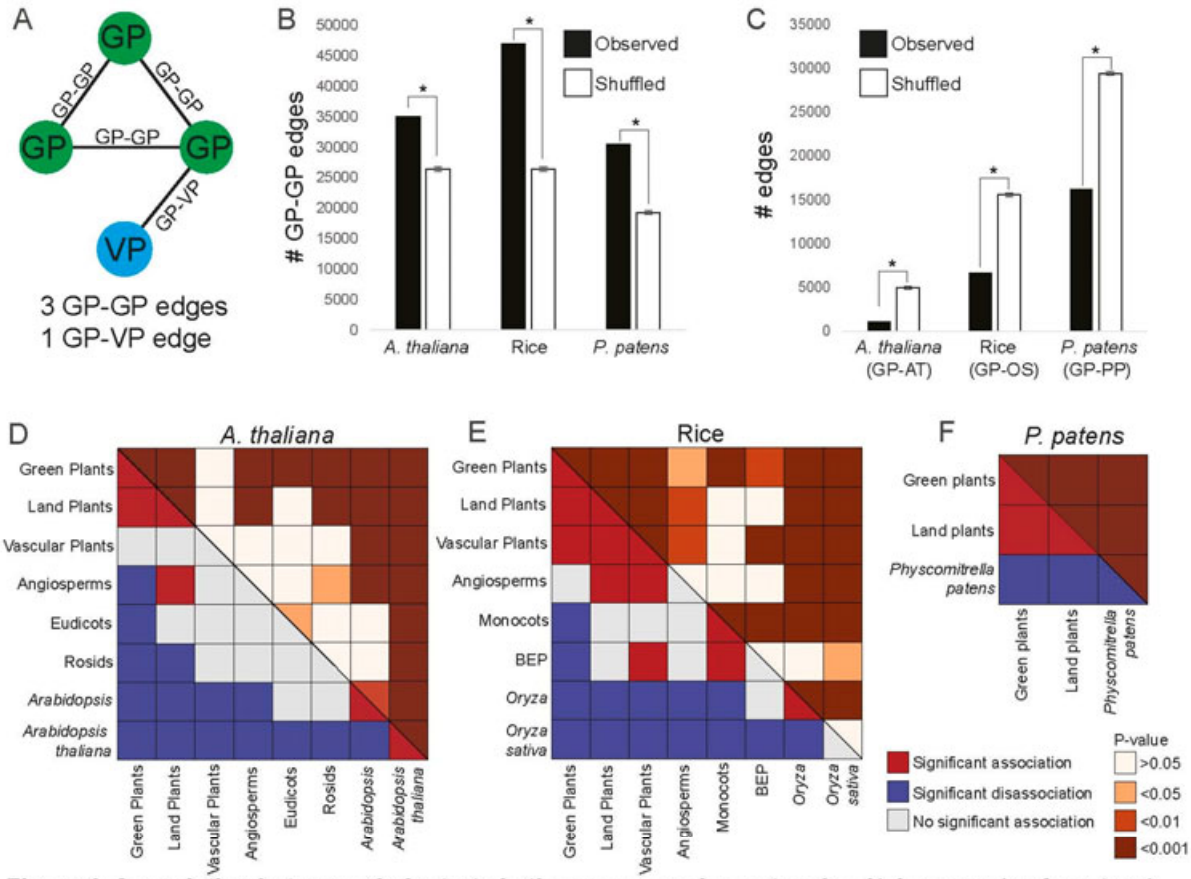

Figure 4. Association between phylostrata in the co-expression networks. A) An example of a network consisting of three Green Plant (GP) and one vascular plant (VP) genes. The network contains 3 GP-GP and 1 GP-VP edge. B) Number of observed (black bars) Green Plant-Green Plant (GP-GP) edges compared to an average number GP-GP edges found in 1000 networks with shuffled gene-phylostratum assignments (white bars). Error bars denote standard deviation of the shuffled observations. C) Number of observed (black bars) edges compared to shuffled (white bars) gene-phylostratum assignments for GP-AT (*A. thaliana*), GP-OS (*O. sativa*) and GP-PP (*P. patens*). The asterisk indicate cases where the number of edges in the network was significantly (P<0.05) larger or smaller than in the shuffled gene-phylostratum assignments, respectively. D-F) Lower triangular heatmap indicates significant (BH<0.05) association (red) and dissociation (blue) of all phylostratum combinations in *A. thaliana* (D) rice (E) and *P. patens* (F). The upper triangular matrix indicates the empirical P-values.

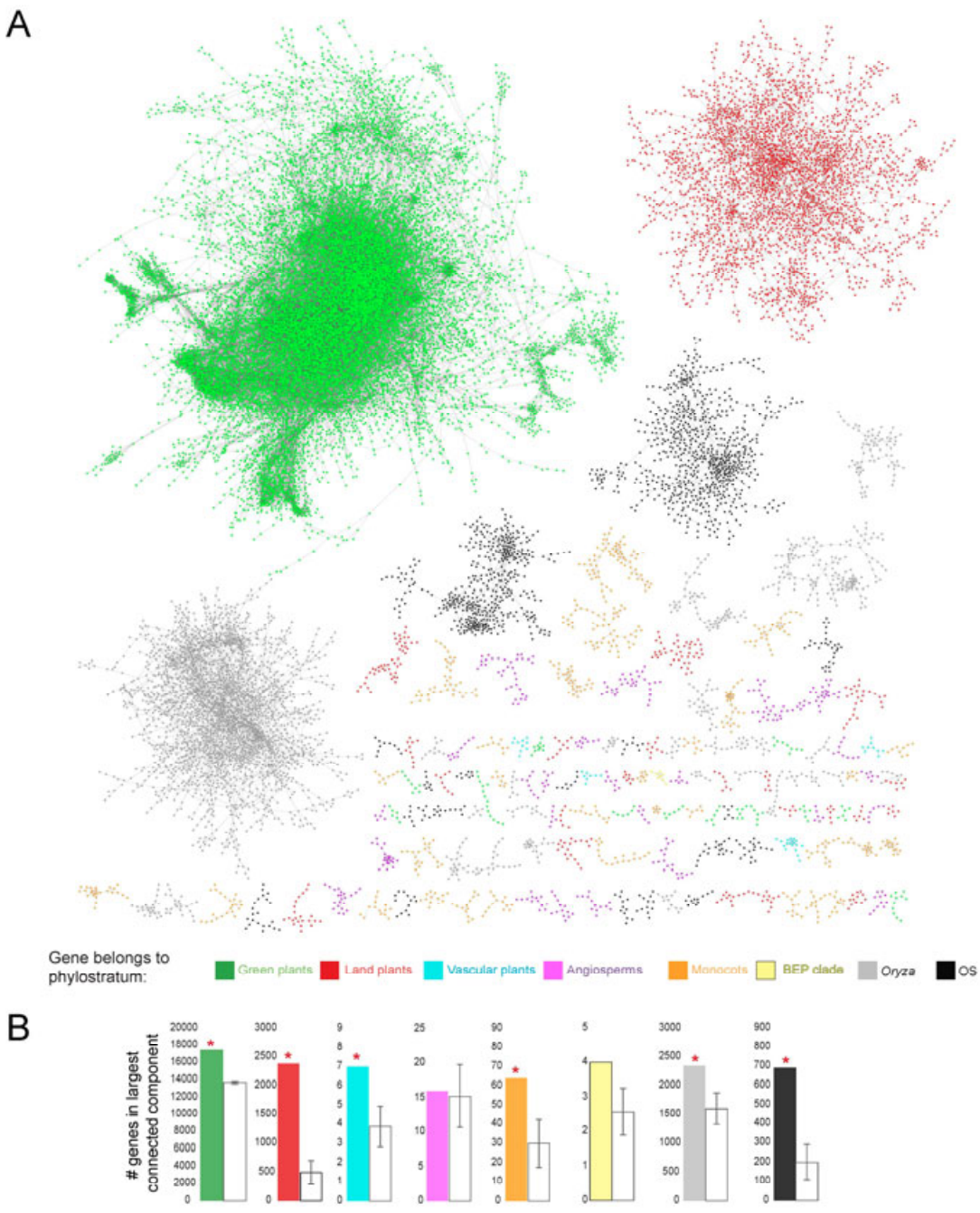

Figure 5. Sizes of connected components in rice. A) Genome-wide network of rice. Nodes represent genes, while node color indicates the phylostratum assignment of a gene (legend below). For brevity, only connected gene pairs that belong to the same phylostratum are shown. B) Observed size of largest connected component per phylostratum (indicated by filled bars), as compared to a randomized analysis with 1000 permutations, where gene-phylostratum assignments have been shuffled (white bars). Error bars denote standard deviation, while red asterisk indicate significant ($P<0.05$) difference between observed and permuted sizes.

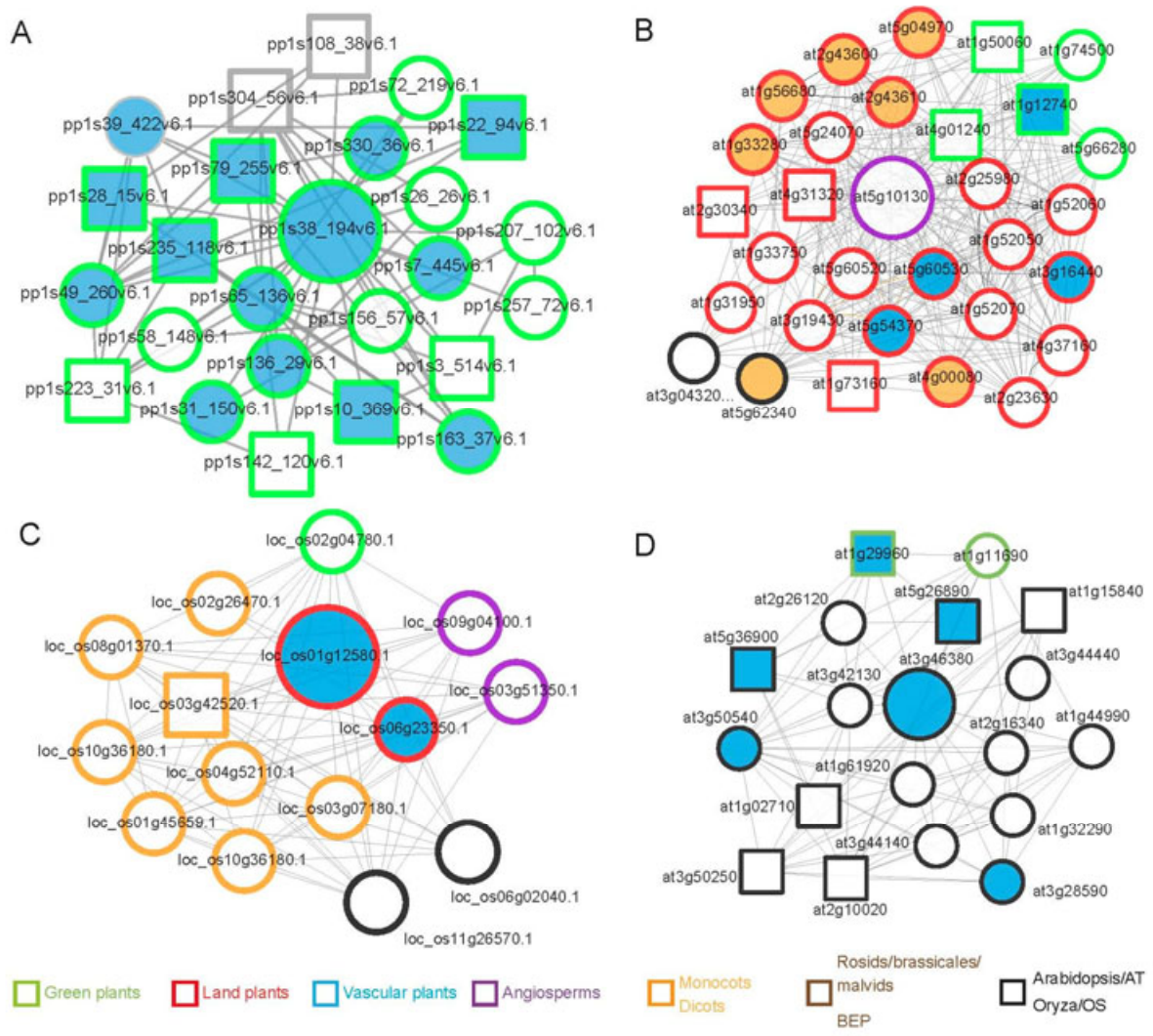

Figure 6. Examples of neighborhoods enriched for the different phylostrata. A) Green Plants phylostratum enriched neighborhood of *P. patens* Pp1s38_194v6.1 (large central node). Nodes present genes, grey edges indicate co-expression edges, while genes with GO term "translation" are colored blue. The phylostratigraphic information is visualized as node border colors, where e.g. the Green Plants phylostratum is indicated by green borders, the Land Plants phylostratum by red borders, and the Vascular Plants phylostratum by blue borders (see edge color legend below the figure). Note that the colors are contextual based on the species, where orange border indicates a monocot- and dicot-specific phylostrata for rice and *Arabidopsis thaliana*, respectively. Rosids, brassicales, malvids and BEP clade are represented by brown color only, since only few genes are assigned to these phylostrata. B) Land Plants phylostratum enriched neighborhood from *Arabidopsis thaliana*. Nodes colored blue and orange have assigned GO terms "ion transport" and "cell wall organization or biogenesis", respectively. C) Monocots phylostratum enriched neighborhood from rice. Nodes colored blue have assigned GO term "embryo development". D) AT phylostratum enriched neighborhood from *Arabidopsis thaliana*. Nodes colored blue have assigned GO terms "carpel morphogenesis.

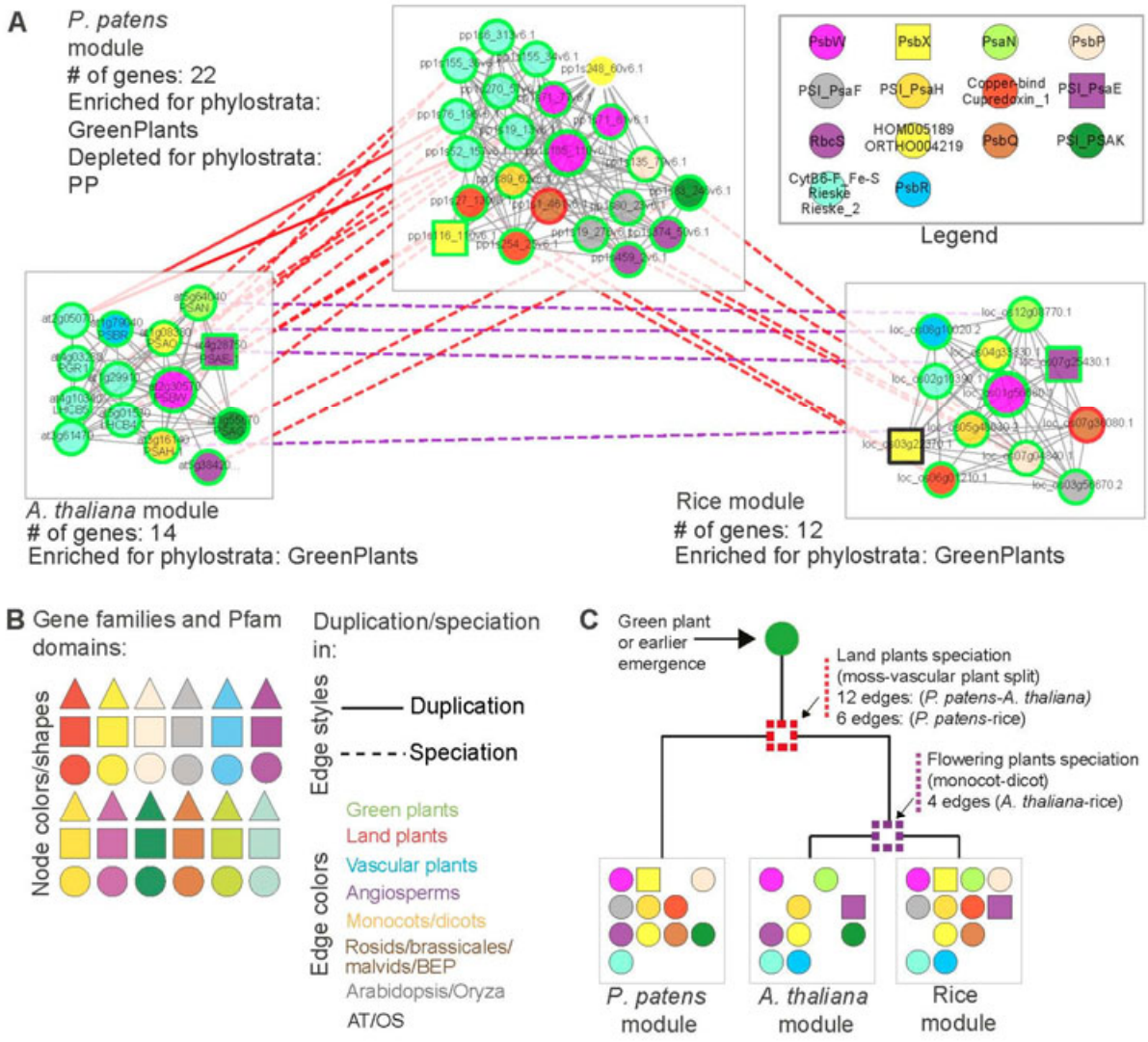

Figure 7. Phylostratigrapic and phylogenetic analysis of photosynthesis modules. A) Photosynthetic gene modules for *A. thaliana*, rice and *P. patens*. Nodes represent genes, colored shapes denote gene families, while colored edges represent timing of speciation events found between modules, as indicated in the legend. B) Description of the elements displayed in the gene module pages. Colored shapes indicate which genes belong to the same family and contain same Pfam domains, edge styles indicate speciation/duplication events and edge colors indicate the phylostratum of the events.C) An evolutionary model of the photosynthetic modules. The colored shapes correspond to the gene families shown in the modules in Figure 7A.

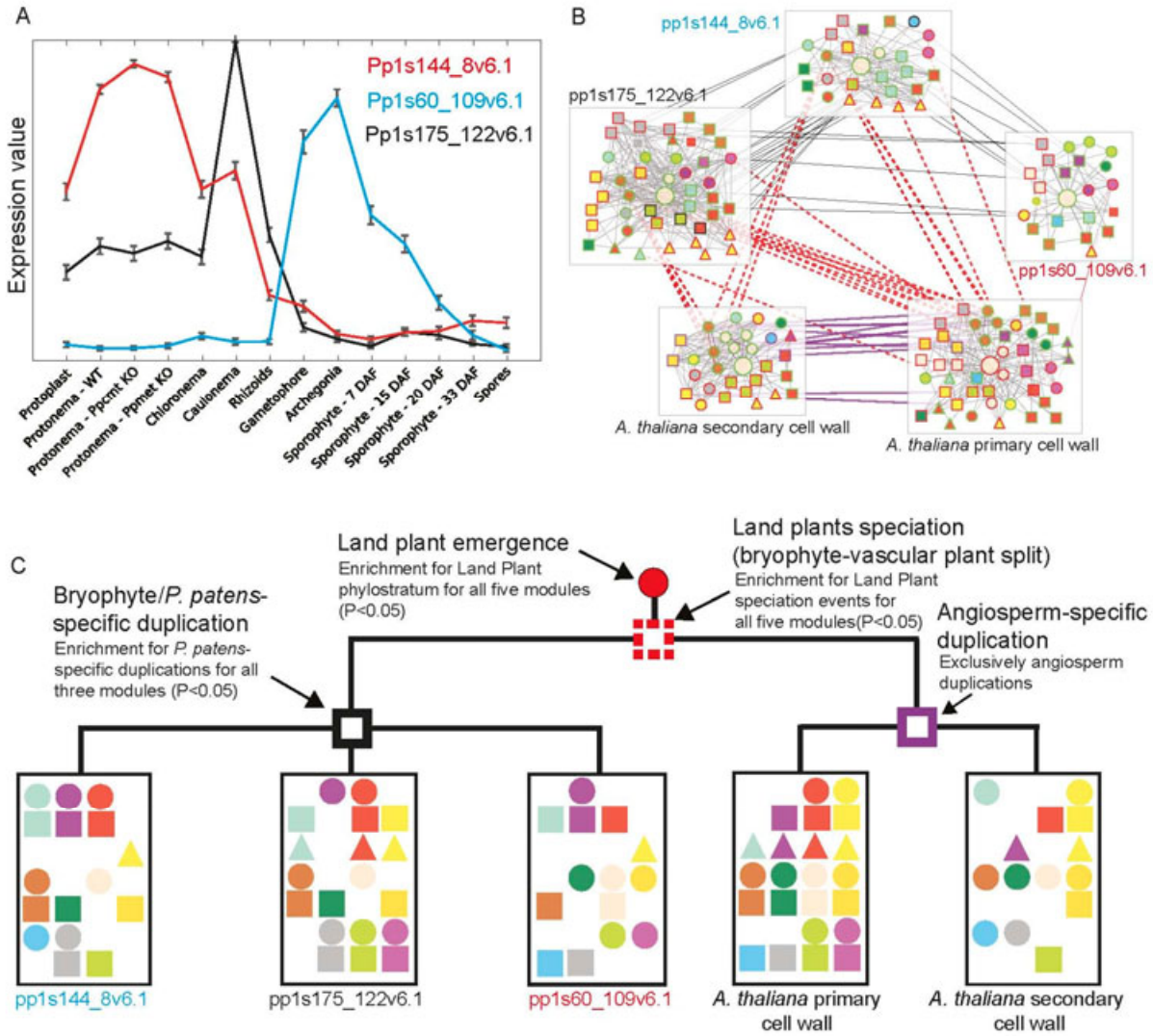

Figure 8. Evolutionary analysis of cell wall modules. A) Average expression profiles of the genes present in the three P. patens cell wall modules provided by PlaNet. Error bars indicate standard error for 32, 21 and 48 genes found in the *Pp1s144_8v6.1*, *Pp1s60_109v6.1* and *Pp1s175_122v6.1* modules, respectively. DAF (days after fertilization), KO (knock-out). B) Comparison of five cell wall modules from *A. thaliana* and *P. patens*. All five modules are significantly enriched (P<0.05) for genes generated in the Land Plants phylostratum (denoted by red borders). C) Evolutionary model of the five cell wall modules from *A. thaliana* and *P. patens*. The colored shapes indicate which gene families are present in the modules.