Transcriptome analysis provides insights into the non-methylated lignin synthesis in Paphiopedilum armeniacum seed

lin fang (linfang@scbg.ac.cn)
South China Botanical Garden https://orcid.org/0000-0002-9971-4569

Xin Xu
South China Botanical Garden

Ji Li
South China Botanical Garden

Feng Zheng
South China Botanical Garden

Mingzhi Li
South China Botanical Garden

Jingwei Yan
Nanjing Agricultural University

Yuan Li
South China Botanical Garden

Xinhua Zhang
South China Botanical Garden

Lin Li
South China Botanical Garden

Guohua Ma
South China Botanical Garden

Aying Zhang
Nanjing Agricultural University

Fubin Lv
Guangdong Academy of Agricultural Sciences

Kunlin Wu
South China Botanical Garden

Songjun Zeng
South China Botanical Garden

Research article

Keywords: Paphiopedilum armeniacum, lignin, transcriptome analysis, germination

Posted Date: August 20th, 2019

DOI: https://doi.org/10.21203/rs.2.13152/v1

Version of Record: A version of this preprint was published on July 29th, 2020. See the published version at https://doi.org/10.1186/s12864-020-06931-1.

Abstract

Abstract Backgrounds Paphiopedilum is an important genus of orchid family (Orchidaceae) with high horticultural value. The wild populations are under the threat of extinction because of over collection and habitat destruction. Mature seeds of most Paphiopedilum species are difficult to germinate, which severely restricts the germplasm resources protection and commercial production. The germination inhibition factors are largely unknown. Results In this study, we found large amounts of non-methylated lignin were accumulated during seed maturation of Paphiopedilum armeniacum (P. armeniacum), which negatively correlates with the germination rate. We then further compared the transcriptome profiles of P. armeniacum seed at different development stages to explore molecular clues for the non-methylated lignin synthesis. KEGG enrichment analysis showed that a large number of genes associated with phenylpropanoid biosynthesis and phenylalanine metabolism as the seed maturation were differentially expressed. Several key genes in the lignin biosynthetic pathways displayed different expression patterns during the lignification process. PAL, 4CL, HCT and CSE were up-regulated to accelerate the C and H lignin accumulation. The expression of CCoAOMT, F5H and COMT were maintained at a low level or down-regulated to inhibit the conversion to the typical G and S lignin. Quantitative real-time RT-PCR analysis confirmed the altered expression levels of these genes among seeds and vegetative tissues. Conclusions This work demonstrated the plasticity of natural lignin polymer assembly in seed, and provided a better understanding of the molecular mechanism of seed-specific lignification process.

Background

Paphiopedilum Pftzer (Orchidacease) are commonly known as Lady's slipper orchids because of their slipper-shaped pouch. Members of this genus have high ornamental value because the unique flowers are available in a wide variety of colors, sizes and shapes. Besides the horticultural value, wild populations of *Paphiopedilum* are under the threat of extinction due to over-collection and habitat destruction, and all species are listed in the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) Appendix I and their trade is prohibited[1].

Many terrestrial orchid seeds including *Paphiopedilum* are tiny and contain no endosperm, in contrast to well-developed embryos in most angiosperm seeds. In the natural habitat, *Paphiopedilum* seeds are generally unable to germinate on their own. They need to engage in a mycorrhizal relationship with fungus that help to feed the emerging seedling[2,3]. Even offering enough nutrient supply in asymbiotic germination, mature seeds of most *Paphiopedilum* species are still difficult to germinate, which severely restricts their conservation and large-scale production. *Paphiopedilum* seeds developed heavily lignified secondary cell walls to reinforce mechanic support and to make it impermeable to water and nutrients[4,5]. In the natural habitat, this feature will enhance the survival of orchid seeds in harsh conditions. However, several studies have implied that the accumulation of lignin contributed to the germination inhibition during aseptic germination [6,7].

Lignin is a phenolic polymer formed by oxidative polymerization of the primary hydroxycinamyl alcohols (monolignols), ρ-coumaryl, caffeyl, coniferyl, sinapyl and 5-hydroxylcinnamyl alcohols, to give rise to ρ-hydroxylphenyl (H), catechyl (C), guaiacyl (G), syringyl (S) and 5-hydroxyguaiacyl (5H/5-OH-G) lignin. The five monolignols differ by their degree and position of methoxylation. The H and C units are non-methoxylated, whereas the G and 5H units are singly methylated on the 3-hydroxyl group and the S subunit is methylated on both the 3- and 5-hydroxyl moieties[8]. Lignin composition varies among plant species and tissue types. Lignins found in the vascular tissues are mainly composed of G and S units. While seed lignins have a significant variation in monolignol compositions. In addition to the typical H, G and S units, C unit is found in the seed of certain species belonging to *Orchidaceae*, *Cactaceae*, *Cleomaceae*, and *Euphorbiaceae*[9-11]. The significance of this variation remains unknown. The non-methylated monolignols are incorporated into lignin polymers via benzodioxane bonds forming a linear structure without side chains[12,13]. The linear lignin has less cross-linking with other cell wall components, and is capable of enhancing the hydrophobicity and stability of the plant tissue[14]. Lignin biosynthesis pathway is well established, involving functionalization of intermediated by O-methyltransferase (CCoAOMT and COMT) to generate monolignols differing in their degree of methylation[8]. Disruption or downregulation of those genes can lead to the lignin composition changed, and therefore change the cell wall integrity and mechanical properties of the tissue[15,16].

Here, we describe the presence of non-methoxylated H and C lignin in the seed of *Paphiopedilum armeniacum* (*P. armeniacum*). The deposition of the non-typical lignin started at the early stage of the seed development of *P. armeniacum*, and eventually made up approximately 39% of the total seed dry mass. Despite the unique structure and potentially important function, the biosynthesis of lignins in plant seeds has received little attention, relative to that in vascular tissues. In this study, a detailed gene expression profiling was performed at five key development stages of *P. armeniacum* seeds to help us understand the seed-specific regulation of lignin biosynthesis and potential regulator factors.

Results

Lignin accumulation during seed development

Several studies suggested the heavily lignified seed coat contributed to the germination inhibition in several species of orchids[3,6,7]. The germination was scored along with the lignin content during the seed development (Figure 1B). At the early stage of the seed development (45-95 DAP), the germination rate gradually increased as the embryo formation. At 95 DAP when the globular embryo has been fully developed, the germination rate reached 96.2%. During this stage, lignin was maintained at a relatively low level (5-9% of the dry mass). After that, even though the starch and lipid globules continue to accumulate as described in our previous study[3], the germination rate dropped sharply from 96% to 5%. Meanwhile, the amounts of lignin started to increase rapidly, as the seed coat turning from white to brown (Figure 1A). Lignin accumulation reached a plateau (~39% of the dry mass) at 150 DAP, by which time the seed coats became dark and hard.

Identification of non-methylated lignin in seeds

To analyze the monolignol composition from various tissue types, the samples were subjected to pyrolysis-gas chromatography (Py-GC/MS). The identities and relative molar abundances of the released compounds are listed in Table 1. The results indicated the lignin in the seeds contain high levels of C units and H units without any trace of G and S units. In contrast, lignins present in other tissues (stems, leaves and pods residues after seed isolation) were composed mainly of G and S units with essentially no C units.

Previous work has established the assignments for C lignin and G/S lignin in FTIR and NMR spectra in orchid seeds[9-11] . Next, the composition of monolignol structure in seeds was further characterized by those two techniques. FTIR spectra from the seeds at five developmental stages indicated C lignin deposited in the early stage with distinct C lignin bands at 1154 cm^{-1} and 823 cm^{-1} (Figure 2A). The rapid development of those two bands was observed at 108 DAP. In agreement with Py-GC analysis, no typical G/S lignin was detected in *P. armeniacum* seeds of all five developmental stages. Analysis of the aromatic region of the 2D NMR spectra of mature seeds confirmed the presence of H and C lignin (Figure 2B). The results indicated that the monolignol composition varied among tissue types in *P. armeniacum*, and the non-methylated C lignin and H lignin specifically deposited in the seed.

RNAseq analysis, *de novo* assembly and functional annotation

To characterize the transcriptome dynamics of *P. armeniacum* seed development and to identify genes involved in the seed-specific lignification, we performed an RNA-seq analysis of *P. armeniacum* seeds at five developmental stages (66, 87, 108, 122 and 150 DAP). Three biological replicate sequencing libraries were prepared from each stage. A total of 104.12 Gb clean data were originated from each library after filtering out low-quality data. The transcriptome details for each sample are given in Table S1. Q30 of the raw data ranged from 92.10 to 94.11% indicating a high-quality reads worthy of further analysis. Since no reference genome is available for *Paphiopedilum*, we *de novo* assembled the total 347,075,856 reads into 433,854 transcripts with an N50 length of 1180 bp and 183,737 unigenes with an average length of 860 bp, respectively. The transcripts and unigenes length distribution are shown in Table S2

For annotation, 183,737 unigenes were subjected to BLASTX searches against the sequences in the NCBI non-redundant protein sequences (NR), Swissprot, Gene Ontology (GO), the Clusters of Orthologous Groups (COG) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. As a result, a total of 92,235 unigenes (50.20% of all unigenes) could be assigned at least one putative function from one of these databases (Table S3). A total of 89,285 unigenes were annotated to the NR database. In a further analysis of the matched sequences, *P. armeniacum* transcript was highly similar to those *Dendrobium catenatum* (29.86%), *Apostasia shenzhenica* (8.38%) and *Phalaenopsis equestris* (5.45%), as shown in Figure 3A. For GO annotation, a total of 29,301 unigenes from *P. armeniacum* seeds were annotated into three GO pathways, as shown in Figure 3B. The functions of unigenes in biological process classifications contained cellular process, metabolic process and biological regulation. Cell part, cell and

organelle were the most abundant functions in terms of cellular component classifications. The most abundant biological process functions are metabolic process and cellular process. In the molecular function classification, binding and catalytic activity were more abundant. The main GO entries revealed that the cells divided frequently during seed development, and some catalytic, metabolic and binding activities were relatively high.

Differential gene expression and KEGG enrichment analysis

With the restrictive conditions of False Discovery Rate (FDR) < 0.05 and log2 ratio > = 1.0, unigenes that were differentially expressed in the seeds at five stages were identified. In total 8,722 differentially expressed genes (DEGs) were identified among all the libraries. The number of DEGs increased with seed development, and peaked among 122 DAP and 66 DAP seeds (Figure 4A). A total of 576 common DEGs were identified in 87 DAP, 108 DAP, 122 DAP and 150 DAP among all developmental stages, implying that these DEGs might be responsible for seed development (Figure 4B).

To explore more insight into the DEGs regulating the lignin deposition at the later stage of seed development, we further performed KEGG enrichment analysis with the DEGs in 87 DAP vs 66 DAP, 108 DAP vs 66 DAP, 122 DAP vs 66 DAP and 150 vs 66 DAP (Figure 5). The top-enriched KEGG pathways of these DEGs were secondary metabolite biosynthesis pathways, including flavonoid, phenylpropanoid and flavone and flavonol biosynthesis pathways, etc. Among them, flavone and flavonol biosynthesis (ko00944) and flavonoid biosynthesis (ko00941) are highly enriched throughout the five key development stages, implying those pathways are generally important for seed development such as pigment accumulation, but not specifically related to lignin deposition. Our premise is that genes involved in non-methylated lignin accumulation would be induced primarily during the lignin rapid accumulation stage (from 101 DAP to 150 DAP). The analysis showed that phenylpropanoid biosynthesis (ko00940) and phenylalanine metabolism (ko00360) are indeed highly enriched in 108 DAP vs 66DAP, 122DAP vs 66DAP and 150 vs 66DAP. Other highly enriched pathways were DNA replication (ko03030) in 87DAP vs 66DAP, starch and sucrose metabolism (ko00500) in 87DAP vs 66DAP and 108DAP vs 66DAP, plant hormone signal transduction (ko04075) in 150DAP vs 66DAP, etc.

Identification of genes potentially involved in non-methylated lignin biosynthesis

The monolignol biosynthesis pathway is well established, and a series of enzymatic reactions catalyzed by specific enzymes have been identified, including phenylalanine ammonia lyase (PAL), cinnamic acid 4-hydroxylase (C4H), 4-coumarate: CoA ligase (4CL), cinnamoyl-CoA reductase (CCR), cinnamyl alcohol dehydrogenase (CAD), hydroxycinnamoyl CoA: shikimate hydroxycinnamoyl transferase (HCT), 4-coumarate 3-hydroxylase (C3H), caffeoyl shikimate esterase (CSE), ferulic acid 5-hydroxylase (F5H), caffeic acid O-methyltransferase (COMT), and caffeoyl-CoA 3-O-methyltransferase (CCoAOMT)[8,13] (Figure 6). We used the annotated lignin pathway gene in KEGG database, and identified 80 *Paphiopedilum* genes that are homologous to genes potentially involved in monolignol synthesis (Table S5).

Most lignin-related genes displayed a very similar transcript level between 66 DAP and 87 DAP, when the rapid accumulation of lignin has not started. Several monolignol synthesis related genes *PAL*, *4CL*, *HCT* and *CSE* demonstrated increased expression levels from 87 DAP to 150 DAP. Despite the high expression level of these genes, most *CCoAOMT*, *COMT* and *F5H* were absent or down-regulated from developing seeds, and a few of them displayed a significant decrease in the expression level. *CCoAOMT* expression level is relatively high at 66 DAP. But the low expression level of *F5H* made the conversion of caffeoyl moieties to feruloyl moieties inefficient, which resulted in no G and S lignin production at the early stage of the seed development. Later, the decrease in the expression level of *CCoAOMT* and upregulation of upstream genes in monolignol pathway caused the accumulation of H and C lignins instead of G and C lignins.

qRT-PCR validation of candidate genes involved in lignin biosynthesis

To verify the accuracy and reproducibility of the RNA-seq results, we selected contigs corresponding to *PAL*, *4CL*, *HCT*, *C4H*, *CCoAOMT*, *F5H* and *COMT* in five development stages of seed development for qRT-PCR validation. Methylated lignin rich tissue stems were also included to verify our predicted monolignol pathway. The sequences of the primers used are given in Table S6. In agreement with the transcriptomic data, *PAL*, *4CL* and *HCT* showed increased expression level as the seed approach maturity (Figure 7). *CCoAOMT* displayed relatively high transcript levels at the early stage of the seed development, and low transcript level during the later stage. *F5H* and *COMT* were also down-regulated during seed development. The results confirm the reliability of the transcriptomic data. Stems displayed relatively high transcript levels of *CCoAOMT*, in contrast to the mature seeds, which suggested that down-regulation of the first lignin methylation gene *CCoAOMT* was crucial for non-methylated lignin production.

Transcription factors (TFs) potentially involved in lignin synthesis

A total of 944 unigenes that encode TFs associated with 54 families were identified from our assembled transcripts. 40, 188, 274 and 247 differentially expressed TFs were found in 87 DAP vs 66 DAP, 108 DAP vs 66 DAP, 122 DAP vs 66 DAP and 150 vs 66 DAP (Table S6). The number of DEGs peaked at 122DAP vs 66DAP. At this stage, the bHLH family contained the most DEGs (28), followed by the C2H2 (22), ERF (21), NAC (18), WRKY (17), GRAS (17), MYB (16) and MYB-related (12). Lignin deposition is mainly regulated by TFs, such as MYB and NAC[8,17,18]. In our RNA-seq dataset, several differentially expressed MYB was screened out (Figure 8B). MYB2 and MYB308 displayed preferential expression in rapid lignin accumulation during seed development. These two TFs are positive regulators of most structure genes including *4CL*, *HCT* and *C3H* in Ecalyptus[19]. *MYB4* was down-regulated as the seed approach maturity. *MYB4* is a positive regulator of *CCoAOMT* and a negative regulator of *PAL*, *C4H*, *4CL* in Arabidopsis and Pinus tadeda[20,21]. *MYB21* is a negative regulator of *CCoAOMT*, the expression level of other structural genes remain unchanged in Poplar[18]. The expression of these TFs agreed with the expression of major structural genes involved in monolignol pathways, implying a potential role in lignin synthesis regulation.

Discussion

In contrast to well-developed embryo and endosperm present in most angiosperm seeds, the structure of orchid seeds is quite simple with a balloon like seed coat and a globular embryo. There is no endosperm. Orchids typically have tiny wind dispersed seeds, often called “dust seeds” [22]. Most species are difficult to germinate on their own under natural conditions. *P. armeniacum* is one of the most difficult species for propagation, and the difficulty in germination severely restricted its large-scale production. The factors that inhibit germination are still unknown. Seed germination of *Paphiopedilum* and other orchid species is significantly affected by the degree of seed maturity [7,23]. Immature orchid seeds with a low degree of lignification, exhibit higher germination percentages than mature seed [2,24]. Enhance germination of mature seeds was documented after seed lignin degradation[2,6,25]. Based on these findings, some researchers have proposed that the high amounts of lignin accumulation contributed to the germination inhibition. In this study, we first measured the lignin content during the seed development process to check whether the lignin content correlates with the germination rate. During the early stage of the development, the germination rate gradually increased as the embryo became fully developed with lignin maintained at a low level. The germination rate abruptly dropped when the seed lignin started rapid accumulating around 95 DAP. The results implied that the level of lignin negatively correlated with the germination rate. The lignin-rich seed coat provided mechanical restraint by preventing water uptake or interfere with diluting germination inhibitors such as abscisic acid[4].

Besides the content, the structure and composition of lignin have been shown to impact the cell wall mechanical properties[26]. Here, we discovered that lignins in the *Paphiopedilum* seeds are composed of H and C units, as clearly evident from Py-GC, two-dimensional NMR and FTIR analysis (Table 1and Figure 2). While the lignins in other parts of the *Paphiopedilum* plant are composed of typical G and S units. This discovery raised an important question: what’s the mechanism for regulating the seed specific deposition of non-methylated lignin. For non-model species lacking a well-studied genetic background, transcriptome analysis is a powerful approach to seeking answers for this question. Based on the dynamics of lignin accumulation and key anatomical features in embryo development, we selected five key stages for transcriptome analysis. A total of 104.12 Gb clean data were obtained from15 RNA-seq libraries of 45, 66, 87, 108, 122 and 150 DAP. The assembled data had an average N50 length of 1180 bp, which is similar to that in other orchid species, such as *Dendrobium catenatum* and *Apostasia shenzhenica* [27,28].

Phenylpropanoid biosynthesis and phenylalanine metabolism are highly enriched during lignin rapid accumulation stages. Further DEGs analysis of phenylpropanoid biosynthesis pathway showed that major structural genes were upregulated with the exception of *CCoAOMT*, *F5H* and *COMT.* Both CCoAOMT and COMT catalyze the O-methylation of the C3 and C5 hydroxyl groups of lignin precursors during G and S lignin formation. Earlier studies demonstrated that low O-methyltransferase (both CCoAOMT and COMT) enzyme activities in the C-lignin containing seed of *Cleome hasslerianan* (*C. hasslerianan*) [15]. CCoAOMT is the first OMT in monolignol pathway. Caffeyl alcohol would be formed if CCoAOMT became depleted or lost its function. Down regulation of *CCoAOMT* in *Pinus radiata* introduced caffeyl alcohol into lignification, resulting in low levels of C units present in the G lignin

dominant tissues[16]. In contrast to pine, suppression of *CCoAOMT* in several angiosperm species such as Arabidopsis, alfalfa, poplar and tobacco did not result in the incorporation of caffeyl alcohol [29-31]. This suggested other factors might also determine the lignin composition, such as transcriptional regulation or availability of the specific monolignol transporters. NAC and MYB families have been identified as regulators of lignin deopsotion[8,17,18]. Instead of the master switches, we are interested in the TFs that can specifically regulate *CCoAOMT*, since these TFs can control the non-methylated lignin accumulation. We identified a few MYBs that are regulators of *CCoAOMT*, such as MYB4 and MYB21. These TFs are good candidates for regulating different monolignol biosynthesis. Overall, the low expression level of *CCoAOMT*, *COMT* and *F5H* is likely the biochemical basis for the accumulation of non-methylated lignin in seeds. The detailed mechanism towards this fine coordination of monolignol biosynthetic genes needs further exploration.

Conclusions

In summary, the presence of H and C lignin in *Paphiopedilum* seeds exemplifies the flexibility in the lignin monomers assembly in nature. Transcriptome profiling revealed the lignin related candidate genes expression pattern, which clarified the seed specific monolignol pathways. However, the seed monolignol composition is not consistent even within the same families of *Cypripedium*, *Euphorbiaceae* and *Cleomaceae* plants, suggesting the feature is not conservative. The difference in the degree of methylation and molecular structure of C and H lignin strongly suggests a functional difference compared to typical G and S lignin. Future work should focus on linking these unique non-methylated lignin structure to the mechanistic role for seed development and germination.

Methods

Plant materials and Growth Conditions: *P. armeniacum* S. C. Chen et F. Y. Liu plants were maintained in a glass greenhouse in the South China Botanical Garden, Guangzhou, China. The plants were potted in a substrate of Zhijing stone for orchids under 800 μ mol m^{-2} s^{-1} natural light maintained by a sunshade net. The average temperature and relative humidity ranged from 10–32 °C and 70–98%, respectively. The flowers from three-year-old adult plants were labeled and artificially self-pollinated. Capsules of different developmental stages from 45 DAP to 180 DAP were collected at 7-day intervals from grown plants. Seeds were removed from capsules, and immediately frozen in liquid nitrogen.

Seed germination assay: Seed capsules were surface sterilized by the method described previously[32]. More than 100 seeds were sown on Hyponex N026 medium supplemented with 1.5 g/ l activated charcoal, 2 g/L peptone, 15g/L sucrose and 5% coconut water. Germination was scored after embryo swollen and testa rupture.

Lignin content and composition: Lignin content was estimated using the acetyl bromide method with slight modification[33]. Five (5) mg of finely ground seed was thoroughly exacted with ethanol/toluene (1:1) until the extracts no longer absorbed UV light at 280 nm. The dried samples were placed into loosely

capped glass tubes containing 1 mL of acetyl bromide/ acetic acid (1:3) and incubated at 70°C for 30 min. The sample was then cooled in an ice bath and then mixed with 0.9 mL of 2 M NaOH, 5 mL glacial acetic acid, and 0.1 mL 7.5 M hydroxylamine hydrochloride. The final volume was adjusted to 10 mL with glacial acetic acid and the absorbance measured at 280 nm with a microplate reader (Tecan Infinity). A standard curve was generated with alkali lignin (Sigma-Aldrich, 471003) for lignin content calculation.

The chemical composition of lignin was analyzed by pyrolysis-gas chromatography (GC)/ mass spectrometry (MS) using a previously described method with some modifications[34]. The pyrolysis was performed on Pyroprobe 5000 (CDS Analytical Inc.) with direct connection to Shimadzu GCMS-QP2010A equipped with HP-5MS column (30 m×0.25 mm×0.25 mm). The pyrolysis was carried out at 550 °C. The chromatograph was programmed from 50 °C to 250 °C at a rate of 15 °C/min, and the final temperature was held for 10 min. Helium was used as the carrier gas at a constant flow rate of 1 mL/min. The mass spectrometer was operated in scan mode and the ion source was maintained at 300 °C. The compounds were identified by comparing their mass spectra with those of the NIST library and those previously reported[35,36]. Peak molar areas were used to calculate the lignin degradation products, and the summed areas were normalized.

Fourier transform infrared spectroscopy (FTIR): The dried samples were embedded in KBr pellets in the concentration of about 1 mg/100 mg KBr. FTIR spectra in the range of 4000–400 cm^{-1} were recorded by using a Shimadzu IRAffinity-1S FTIR spectrophotometer. The spectra were recorded in the absorption mode at 64 scans per sample with a resolution of 4 cm^{-1}.

2D ^{13}C-^{1}H heteronuclear single-quantum coherence (HSQC) NMR spectroscopy: Seeds from 122 DAP was extracted and ball-milled as previously described[37]. The gels were formed using DMSO-d_6/pyridine-d_5 (4:1) and sonicated until homogenous. The homogeneous solutions were transferred to NMR tubes. HSQC spectra were acquired at 25 °C using a Bruker Avance-500 MHz instrument. The detailed running conditions and assignment of the spectra was described elsewhere[11].

RNA extraction, library construction and RNA sequencing: The five different development stages analyzed were 66, 87, 108, 122 and 150 days after pollination (DAP). The total RNA was extracted using Column Plant RNAout2.0 (Tiandz Inc., Beijing, China) according to the manufacturer's protocol, which was specifically designed for materials rich in polysaccharides and polyphenolics. Extracted RNA was treated with DNase (Tiandz Inc., Beijing, China) to remove genomic DNA. The RNA quality was validated using agarose gel electrophoresis, Nanodrop One (Nanodrop Technologies Inc., DE, USA), and Agilent 2100 (Agilent Technologies Inc., CA, USA) to confirm the purity, concentration, and integrity, respectively. Library construction and sequencing were performed using Illumina HiSeq4000 platform (Illumina Inc., CA, USA) by Genepioneer Technologies Corporation (Nanjing, China).

***De novo* assembly, functional annotation of unigenes and DEGs analysis:** The clean data was generated by removing adaptor sequences, ambiguous reads ('N' > 10%), and low-quality reads (that is, where more than 50% of bases in a read had a quality value $Q \leq 5$) using perl script. All sequence data was uploaded

into the BioProject databased hosted by the National Center for Biotechnology Information (NCBI) under the BioProject PRJNA550294. The transcripts were assembled using Trinity v2.4.0 program with default parameters[38], and gene expression was estimated by applying the fragments per kilobase per million mapped reads (FPKM). Functional annotations were performed using the public databases, including the NR [39], Swiss-Prot[40], COG [41] and GO[42]. Differentially expressed genes (DEGs) between libraries were identified as those with the fold change (FC) of the expression level (FC≥ 2 or FC≤0.5 under P-value ≤0.05), FDR ≤0.05). GO enrichment and KEGG enrichment were performed using the obtained DEGs. To identify genes encoding for TFs, all DEGs were compared with protein sequences downloaded from a plant TF database, PlantTFDB with the E-values threshold of 10^{-10}[43].

Verification of gene expression using qRT-PCR: Samples of RNA-seq were reversed transcribed into cDNA for real-time qPCR validation using GoScript Reverse Transcription System (Promega, CA, USA). The qRT-PCR reactions were performed on ABI 7500 Real-Time PCR System (Applied Biosystems, CA, USA) and the SYBR Premix ExTaq Kit (Takara, Dalian, China). Primer sequences are listed in Table S7. The expression level was calculated as $2^{-\Delta\Delta Ct}$, and normalized to the Ct values of *P. armeniacum Actin* (TRINITY_DN40678_c2_g1).

Abbreviations

P. armeniacum: *Paphiopedilum armeniacum*; Py-GC/MS: pyrolysis-gas chromatography; DAP: Days after pollination; FTIR: Fourier transform infrared spectroscopy; DEG: Differentially expressed genes; NR: non-redundant protein sequences; GO: Gene Ontology; COG: the Clusters of Orthologous Groups; KEGG: Kyoto Encyclopedia of Genes and Genomes; FDR: False Discovery Rate; PAL: Phenylalanine ammonia lyase; COMT: Caffeic acid O-methyltransferase; CCoAOMT: Caffeoyl-CoA 3-O-methyltransferase; TFs: Transcription factors.

Declarations

Acknowledgements

The authors would like to thank Yunfei Yuan for assistance with NMR experiments, Yuting Yin, Chuangxue Mao and Sha Deng for assistance with plant growth and pollination.

Consent for publication

Not Applicable.

Ethics approval and consent to participate

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and the supplementary information files. The sequence data was deposited at NCBI database

under the BioProject PRJNA550294.

Funding

This work was funded by National Key Research and Development Program of China (2016YFC0503104), Natural Science Foundation of Guangdong Province (2018A030313686), Guangdong Key Technology Research and Development Program (2018B020202001, 2017A020208036, 2017B090901050), and Guangdong Modern Agricultural Industry Technology System Program (2018LM2176, 2018LM1142), Biological Resources Program, Chinese Academy of Sciences (KFJ-BRP-007), Guangzhou Science and Technology Program (201904010305),Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Sciences (AB2018002, AB2018026). The Project was also supported by Program from Core Botanical Gardens, Chinese Academy of Sciences.

Authors' contributions

LF, KW and SZ conceived the research program and designed research. LF, XX, JL, FZ, JY and YL performed research, KW, ML, XZ, GM, FL and SZ analyzed the data. LF, XX, KW and SZ wrote the manuscript, AZ revised this manuscript.

References

1 CITES. http://www.cites.org/eng/app/appendices.php. Appendices I, II and III. (2015).

2 Zeng, S. *et al.* In vitro propagation of Paphiopedilum orchids. *Critical Reviews in Biotechnology* **36**, 521-534 (2016).

3 Zhang, Y. *et al.* Embryo development in association with asymbiotic seed germination in vitro of Paphiopedilum armeniacum SC Chen et FY Liu. *Scientific Reports* **5**, 16356 (2015).

4 Steinbrecher, T. & Leubner-Metzger, G. The biomechanics of seed germination. *Journal of Experimental Botany* **68**, 765-783, doi:10.1093/jxb/erw428 (2017).

5 Hossain, Z. et al. Transcriptome profiling of Brassica napus stem sections in relation to differences in lignin content. BMC Genomics 19, 255, doi:10.1186/s12864-018-4645-6 (2018).

6 Pierce, S., Spada, A., Caporali, E., Ceriani, R. M. & Buffa, G. Enzymatic scarification of Anacamptis morio (Orchidaceae) seed facilitates lignin degradation, water uptake and germination. *Plant Biology* (2018).

7 Yeung, E. C. A perspective on orchid seed and protocorm development. *Botanical Studies* **58**, 33 (2017).

8 Vanholme, R., Demedts, B., Morreel, K., Ralph, J. & Boerjan, W. Lignin biosynthesis and structure. *Plant physiology* **153**, 895-905 (2010).

9 Chen, F. *et al.* Novel seed coat lignins in the Cactaceae: structure, distribution and implications for the evolution of lignin diversity. *The Plant Journal* **73**, 201-211 (2013).

10 Barsberg, S. T., Lee, Y.-I. & Rasmussen, H. N. Development of C-lignin with G/S-lignin and lipids in orchid seed coats–an unexpected diversity exposed by ATR-FT-IR spectroscopy. *Seed Science Research* **28**, 41-51 (2018).

11 Chen, F., Tobimatsu, Y., Havkin-Frenkel, D., Dixon, R. A. & Ralph, J. A polymer of caffeyl alcohol in plant seeds. *Proceedings of the National Academy of Sciences* **109**, 1772-1777, doi:10.1073/pnas.1120992109 (2012).

12 Xie, M. *et al.* Regulation of lignin biosynthesis and its role in growth-defense tradeoffs. *Frontiers in plant science* **9**, 1427 (2018).

13 Hao, Z. & Mohnen, D. A review of xylan and lignin biosynthesis: foundation for studying Arabidopsis irregular xylem mutants with pleiotropic phenotypes. *Critical Reviews in Biochemistry Molecular Biology* **49**, 212-241 (2014).

14 Stone, M. L. *et al.* Reductive Catalytic Fractionation of C-Lignin. *ACS Sustainable Chemistry* **6**, 11211-11218 (2018).

15 Tobimatsu, Y. *et al.* Coexistence but independent biosynthesis of catechyl and guaiacyl/syringyl lignin polymers in seed coats. *The Plant Cell* **25**, 2587-2600 (2013).

16 Wagner, A. *et al.* CCoAOMT suppression modifies lignin composition in Pinus radiata. *The Plant Journal* **67**, 119-129 (2011).

17 Zhao, Q. & Dixon, R. A. Transcriptional networks for lignin biosynthesis: more complex than we thought? *Trends in plant science* **16**, 227-233 (2011).

18 Karpinska, B. *et al.* MYB transcription factors are differentially expressed and regulated during secondary vascular tissue development in hybrid aspen. **56**, 255-270 (2004).

19 Goicoechea, M. *et al.* EgMYB2, a new transcriptional activator from Eucalyptus xylem, regulates secondary cell wall formation and lignin biosynthesis. *The Plant Journal* **43**, 553-567 (2005).

20 Jin, H. *et al.* Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in Arabidopsis. **19**, 6150-6161 (2000).

21 Patzlaff, A. *et al.* Characterisation of a pine MYB that regulates lignification. *The Plant Journal* **36**, 743-754 (2003).

22 Barthlott, W., Große-Veldmann, B., Korotkova, N. J. A. s. e. m. s. T. N. y. R. M., editores. Berlin: Botanic Garden & Berlin-Englera, B. M. Orchid seed diversity. (2014).

23 Songjun, Z. *et al.* Seed biology and in vitro seed germination of Cypripedium. *Critical Reviews in Biotechnology* **34**, 358-371 (2014).

24 Pierce, S. & Cerabolini, B. Asymbiotic germination of the White Mountain Orchid (Pseudorchis albida) from immature seed on media enriched with complex organics or phytohormones. *Seed Science Technology* **39**, 199-203 (2011).

25 Barsberg, S., Rasmussen, H. N. & Kodahl, N. Composition of Cypripedium calceolus (Orchidaceae) seeds analyzed by attenuated total reflectance IR spectroscopy: in search of understanding longevity in the ground. *American journal of botany* **100**, 2066-2073 (2013).

26 Özparpucu, M. *et al.* Unravelling the impact of lignin on cell wall mechanics: a comprehensive study on young poplar trees downregulated for CINNAMYL ALCOHOL DEHYDROGENASE (CAD). *The Plant Journal* **91**, 480-490 (2017).

27 Zhang, L. *et al.* Origin and mechanism of crassulacean acid metabolism in orchids as implied by comparative transcriptomics and genomics of the carbon fixation pathway. *The Plant Journal* **86**, 175-185 (2016).

28 Zhang, G.-Q. *et al.* The Dendrobium catenatum Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. *Scientific reports* **6**, 19029 (2016).

29 Zhong, R., Morrison, W. H., Himmelsbach, D. S., Poole, F. L. & Ye, Z.-H. Essential role of caffeoyl coenzyme A O-methyltransferase in lignin biosynthesis in woody poplar plants. *Plant Physiology* **124**, 563-578 (2000).

30 Marita, J. M. *et al.* Structural and compositional modifications in lignin of transgenic alfalfa down-regulated in caffeic acid 3-O-methyltransferase and caffeoyl coenzyme A 3-O-methyltransferase. *Phytochemistry* **62**, 53-65 (2003).

31 Do, C.-T. *et al.* Both caffeoyl Coenzyme A 3-O-methyltransferase 1 and caffeic acid O-methyltransferase 1 are involved in redundant functions for lignin, flavonoids and sinapoyl malate

biosynthesis in Arabidopsis. *Planta* **226**, 1117-1129 (2007).

32 Zeng, S. *et al.* Asymbiotic seed germination, seedling development and reintroduction of Paphiopedilum wardii Sumerh., an endangered terrestrial orchid. *Scientia Horticulturae* **138**, 198-209 (2012).

33 Fang, L. *et al.* Loss of inositol phosphorylceramide sphingolipid mannosylation induces plant immune responses and reduces cellulose content in Arabidopsis. *The Plant Cell* **28**, 2991-3004 (2016).

34 Del Río, J. C. *et al.* Structural characterization of wheat straw lignin as revealed by analytical pyrolysis, 2D-NMR, and reductive cleavage methods. *Journal of agricultural food chemistry* **60**, 5922-5935 (2012).

35 Del Río, J. C. & Gutiérrez, A. Chemical composition of abaca (Musa textilis) leaf fibers used for manufacturing of high quality paper pulps. *Journal of agricultural food chemistry* **54**, 4600-4610 (2006).

36 Ralph, J. & Hatfield, R. D. Pyrolysis-GC-MS characterization of forage materials. *Journal of Agricultural Food Chemistry* **39**, 1426-1437 (1991).

37 Kim, H. & Ralph, J. Solution-state 2D NMR of ball-milled plant cell wall gels in DMSO-d6/pyridine-d5. *Organic biomolecular chemistry* **8**, 576-591 (2010).

38 Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**, 1494 (2013).

39 Deng, Y. *et al.* Integrated nr database in protein annotation system and its localization. *Comput Eng* **32**, 71-74 (2006).

40 Apweiler, R. *et al.* UniProt: the universal protein knowledgebase. *Nucleic acids research* **32**, D115-D119 (2004).

41 Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research* **28**, 33-36 (2000).

42 Consortium, G. O. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* **32**, D258-D261 (2004).

43 Jin, J. *et al.* PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic acids research*, gkw982 (2016).

Table 1

Table 1. Relative molar abundances (%) of the compounds released after pyro-GC/MS of seeds, stems, leaves and pods of *P. armeniacum*. Values in brackets are the standard

deviations from three biological replicates.

Compound name	Origin	Formula	Molecular Mass	Main Mass fragments	Seeds (%)	Stems (%)	Leaves (%)	Pods (%)
Phenol	H	C_6H_6O	94	65,66,94	16.2 (0.1)	-	23.0 (6.9)	5.7 (1.2)
2-Methylphenol	H	C_7H_8O	108	77,107,108	18.7 (0.4)	-	-	-
3-Methylphenol	H	C_7H_8O	108	77,107,108	18.3 (0.4)	-	-	-
2,5-Dimethylphenol	H	$C_8H_{10}O$	122	77,107,122	7.2 (1.3)	-	-	-
Catechol	C	$C_6H_6O_2$	110	53,64,110	19.6 (2.7)	-	-	-
4-Methylcatechol	C	$C_7H_8O_2$	124	51,97, 125	13.4 (0.7)	-	-	-
4-Ethylcatechol	C	$C_8H_{10}O_2$	138	51,78,124	6.5 (0.1)	-	-	-
2-Methoxylphenol	G	$C_7H_8O_2$	124	81,109,124	-	14.9 (0.8)	16.1 (2.1)	19.0 (2.1)
2-Methoxy-5-methylphenol	G	$C_8H_{10}O_2$	138	95,123,138	-	5.6 (0.2)	-	13.4 (2.7)
4-Ethenyl-2-methoxyphenol	G	$C_9H_{10}O_2$	150	107,135,150	-	31.6 (4.6)	23.1 (3.2)	31.7 (3.2)
4-Hydroxy-3-methoxybenzaldehyde	G	$C_8H_8O_3$	152	109,151,152	-	4.5 (0.1)	-	5.5 (0.3)
2-Methoxy-4-propenylphenol	G	$C_{10}H_{12}O_2$	164	131,149,164	-	7.9 (0.1)	-	11.2 (0.6)
4-Hydroxy-3-methoxyphenyl acetone	G	$C_{10}H_{12}O_3$	180	122,137,180	-	4.1 (0.3)	18.1 (0.7)	-
2,6-Dimethoxyphenol	S	$C_8H_{10}O_3$	154	111,139,154	-	-	19.8 (4.7)	13.6 (2.6)
4-Hydroxy-3,5-dimethoxystyrene	S	$C_{10}H_{12}O_3$	180	137,165,180	-	31.6 (2.7)	-	-
% H units					60.5 (2.0)	-	23.0 (4.3)	5.7 (0.9)
% C units					39.5 (2.0)	-	-	-
% G units					-	68.4 (6.9)	57.2 (3.1)	80.7 (2.5)
% S units					-	31.6 (2.1)	19.8 (2.9)	13.6 (1.9)

Supplemental Information

Supplementary Table S1. Overview of transcriptome sequencing and *de novo* assembly results.

Supplementary Table S2. Overview of transcriptome assembly showing length distribution of transcript and unigene.

Supplementary Table S3. Summary of functional annotation of contigs from BLAST searches against public databases.

Supplementary Table S4. KEGG enrichment analysis list.

Supplementary Table S5. Lignin related genes with RPKM values.

Supplementary Table S6. Differentially expressed transcription factors (TFs) list.

Supplementary Table S7. Primers used for qRT-PCR.

Figures

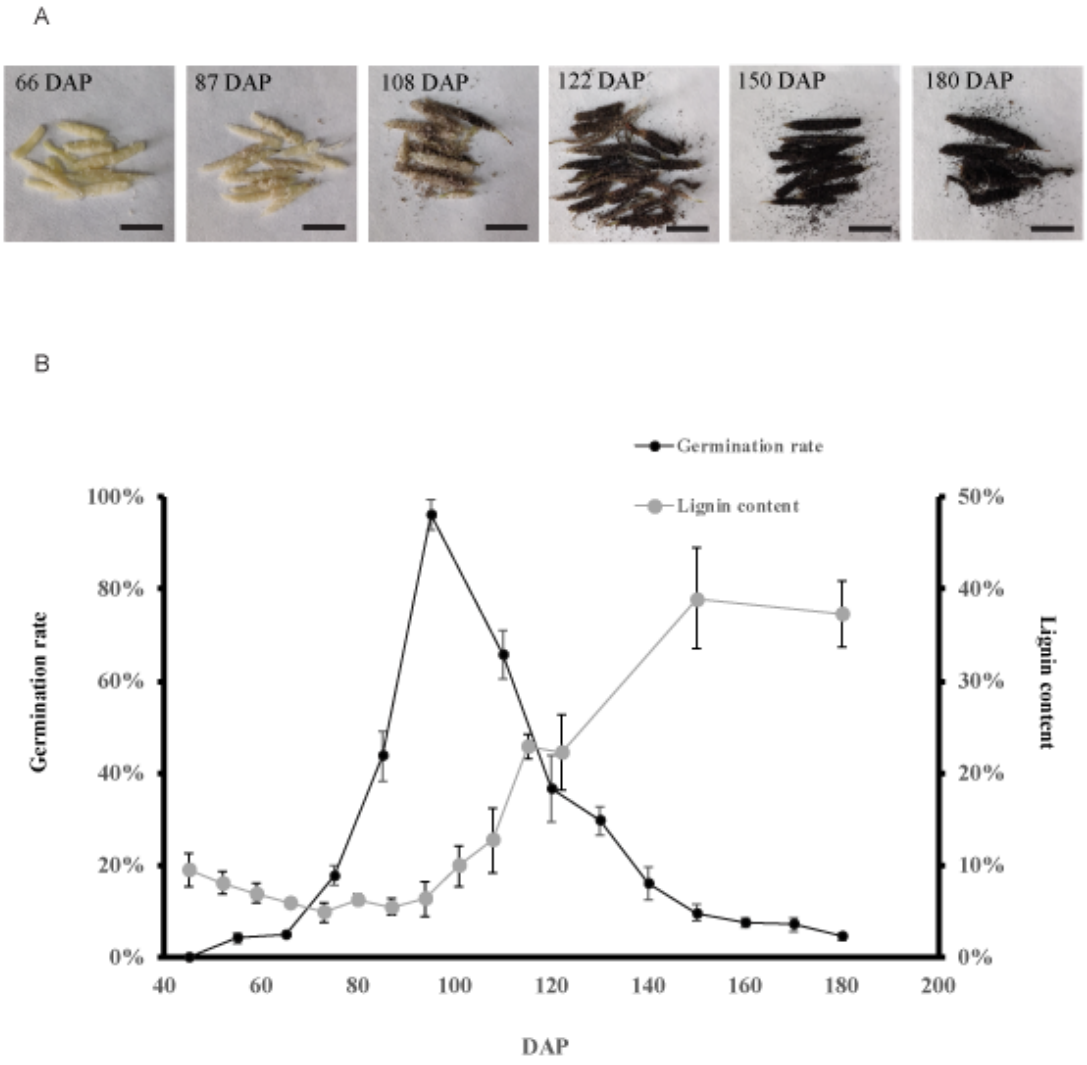

Figure 1. Dynamic changes in seed morphology, lignin accumulation and germination rate during *P. armeniacum* seed development. (A) Phenotypes of *P. armeniacum* seeds. Scale bar = 1 cm. (B) Lignin content and germination rate during *P. armeniacum* seed development.

Figure 1

Dynamic changes in seed morphology, lignin accumulation and germination rate during P. armeniacum seed development. (A) Phenotypes of P. armeniacum seeds. DAP, days after pollination. Scale bar = 1 cm. (B) Lignin content and germination rate during P. armeniacum seed development. The lignin content was estimated using the acetyl bromide method, n = 3.

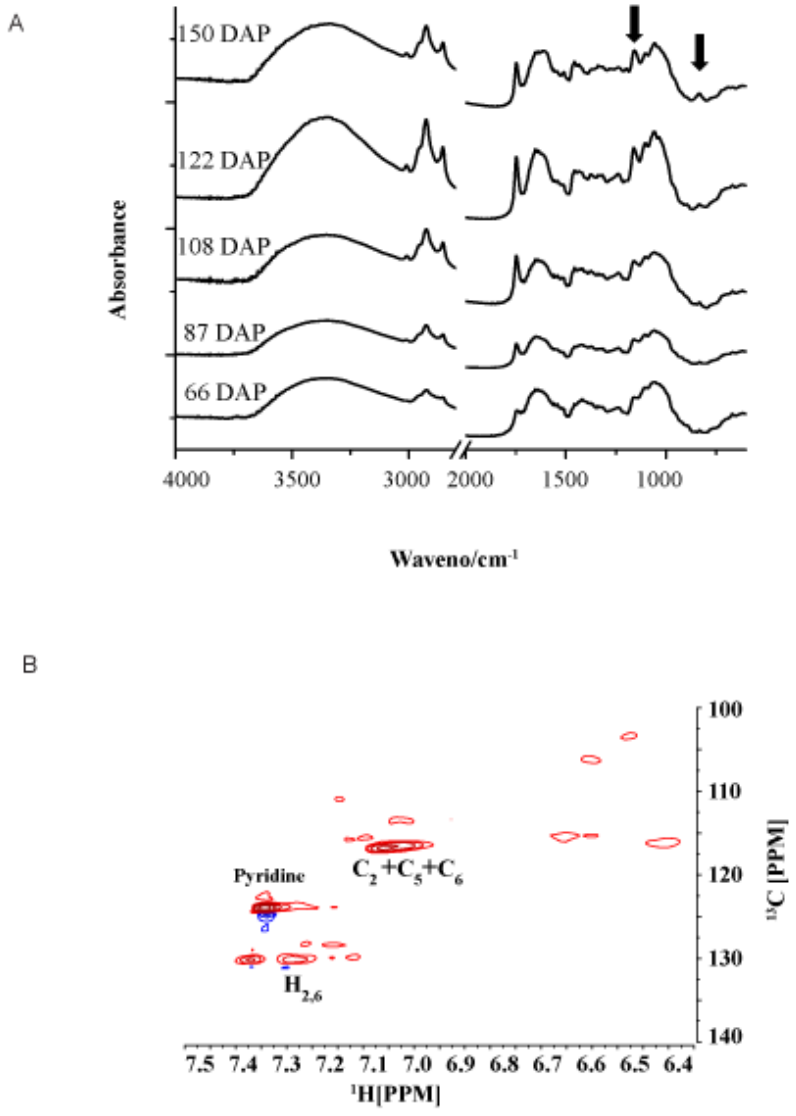

Figure 2. Lignin structure characterization. (A) FTIR spectra of *P. armeniacum* seeds at five developmental stages. Positions of C lignin bands are marked by arrows. (B) Partial short-range ^{13}C–^{1}H (HSQC) spectra (aromatic region) of 122 DAP mature seeds.

Figure 2

Lignin structure characterization. (A) FTIR spectra of P. armeniacum seeds at five developmental stages. Positions of C lignin bands are marked by arrows. (B) Partial short-range 13C–1H (HSQC) spectra (aromatic region) of 122 DAP mature seeds.

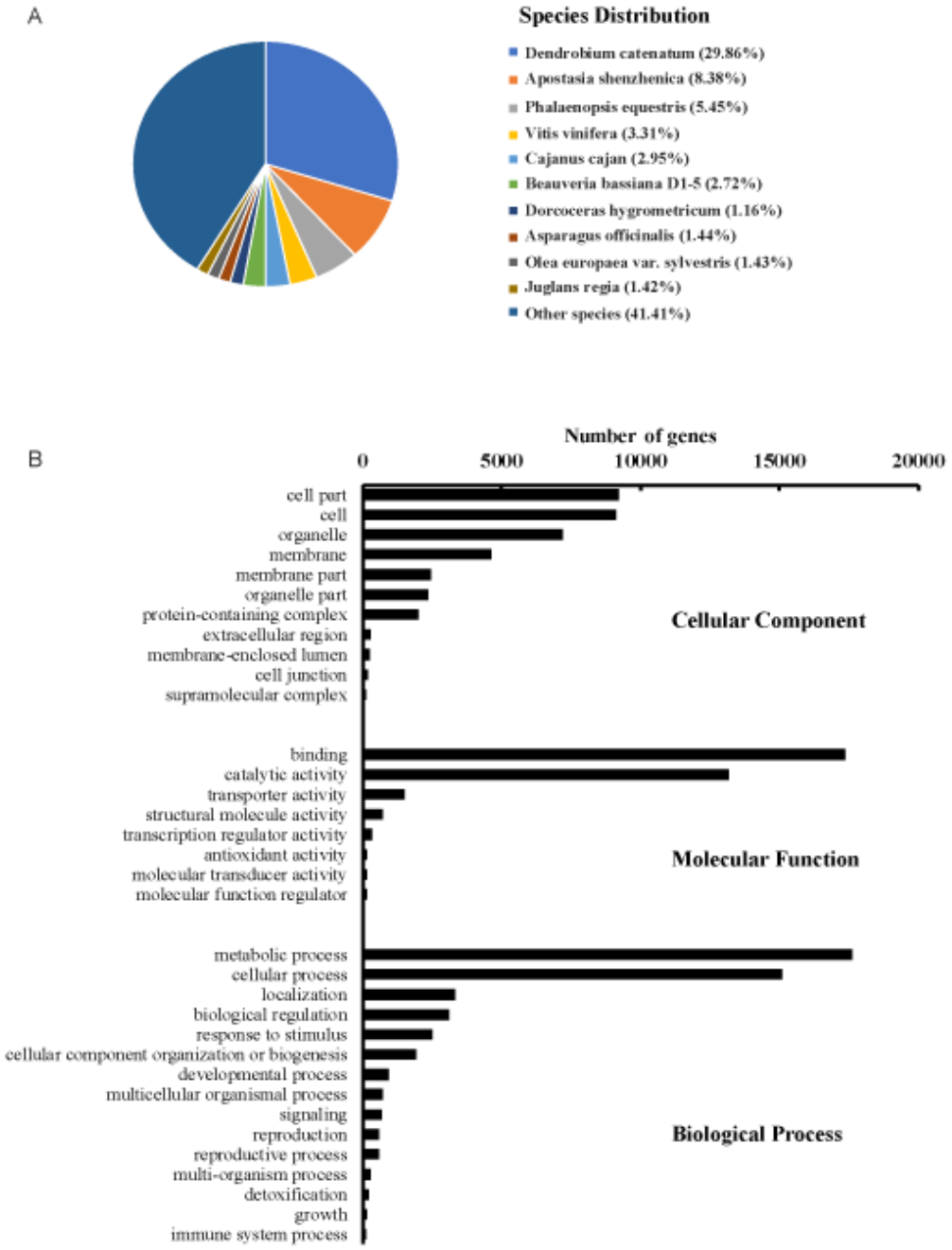

Figure 3. Functional annotations of the unigenes of the *P. armeniacum* eed transcriptome. (A) NR annotated species distribution map similar to the *Paphiopedilum armeniacum* transcriptome. (B) GO function annotation.

Figure 3

Functional annotations of the unigenes of P. armeniacum seed transcriptome. (A) NR annotated species distribution map similar to the Paphiopedilum armeniacum transcriptome. Dendrobium catenatum

shows the highest similarity. (B) GO function annotation. The most abundant functions are binding and catalytic activity in terms of molecular function, and metabolic process and cellular process in terms of biological process.

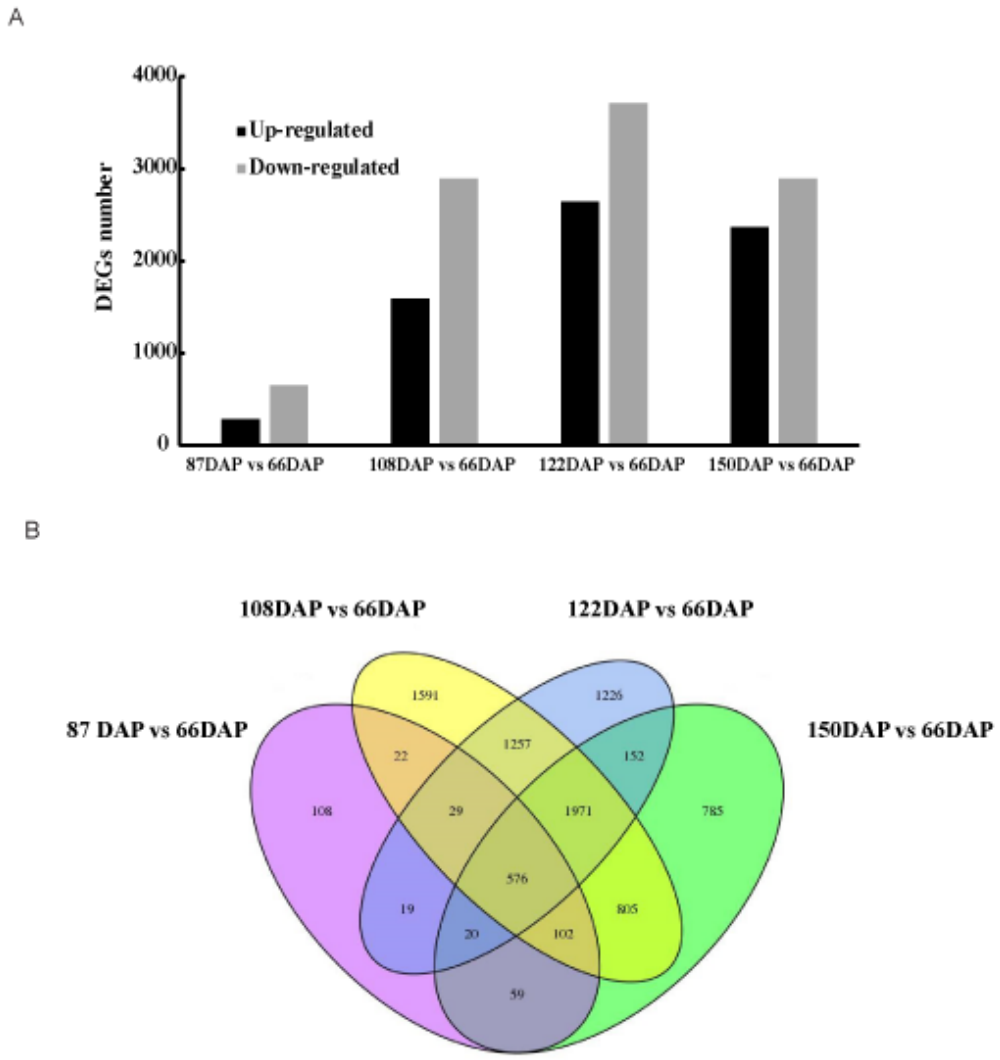

Figure 4. Statistical analysis of differentially expressed unigenes (DEGs) during *P. armeniacum* seed development. (A) up/down-regulated unigenes in all development stages. (B) Venn diagram of all DEGs.

Figure 4

Statistical analysis of differentially expressed unigenes (DEGs) during P. armeniacum seed development. (A) up/down-regulated unigenes in all development stages. The total DEG number peaked between 122

DAP and 66 DAP. (B) Venn diagram of all DEGs.

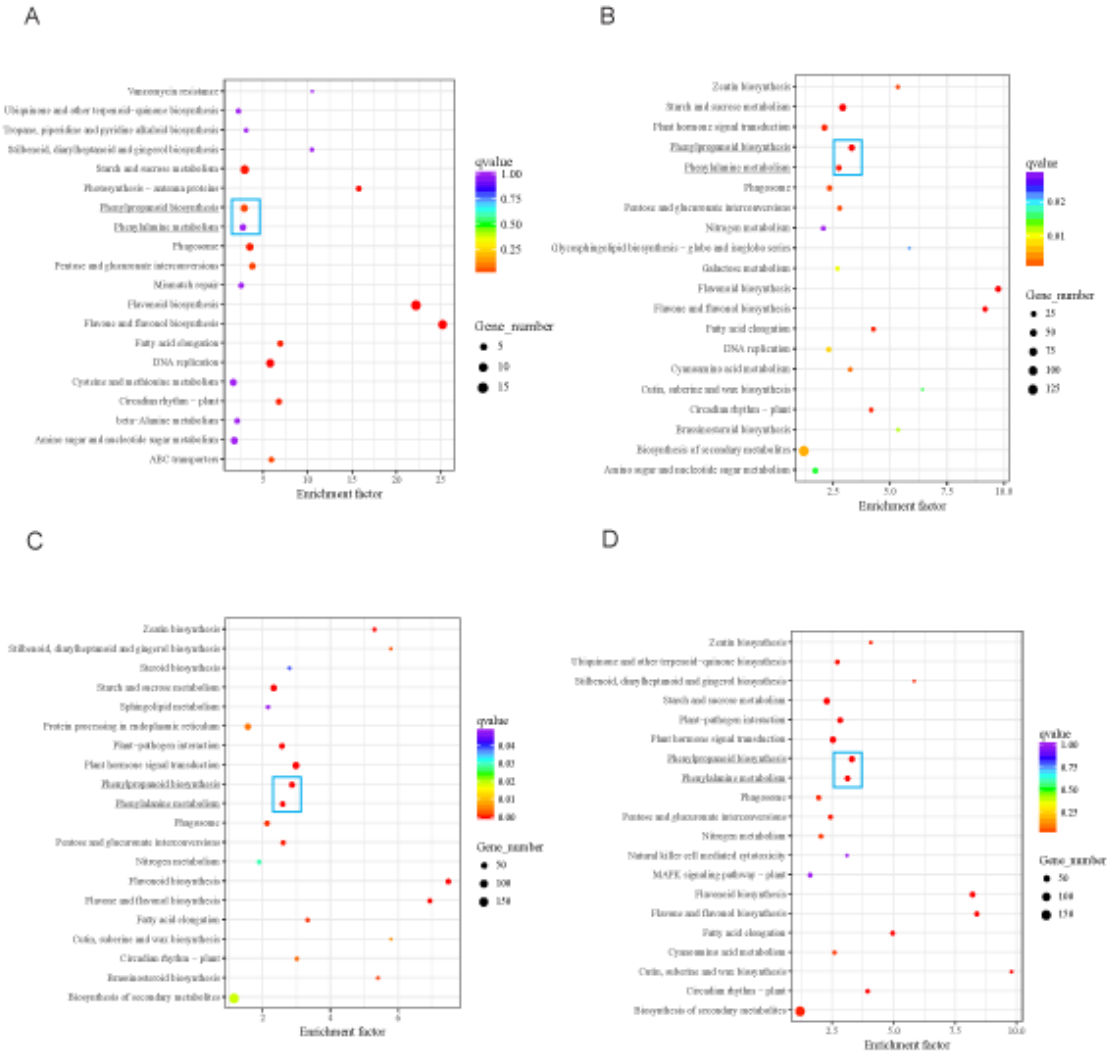

Figure 5. KEGG enrichment analysis with the DEGs. (A) 87 DAP vs 66DAP, (B) 108DAP vs 66DAP, (C) 122DAP vs 66DAP, (D) 150DAP vs 66 DAP. Flavone and flavonol biosynthesis (ko00944) and flavonoid biosynthesis (ko00941) are highly enriched throughout the five key development stages. Phenylpropanoid biosynthesis (ko00940) and phenylalanine metabolism (ko00360) were enriched during the seed lignification.

Figure 5

KEGG enrichment analysis with the DEGs. (A) 87 DAP vs 66 DAP, (B) 108 DAP vs 66 DAP, (C) 122 DAP vs 66 DAP, (D) 150 DAP vs 66 DAP. Flavone and flavonol biosynthesis (ko00944) and flavonoid biosynthesis (ko00941) are highly enriched throughout the five key development stages. Phenylpropanoid biosynthesis (ko00940) and phenylalanine metabolism (ko00360) were enriched during the seed lignification.

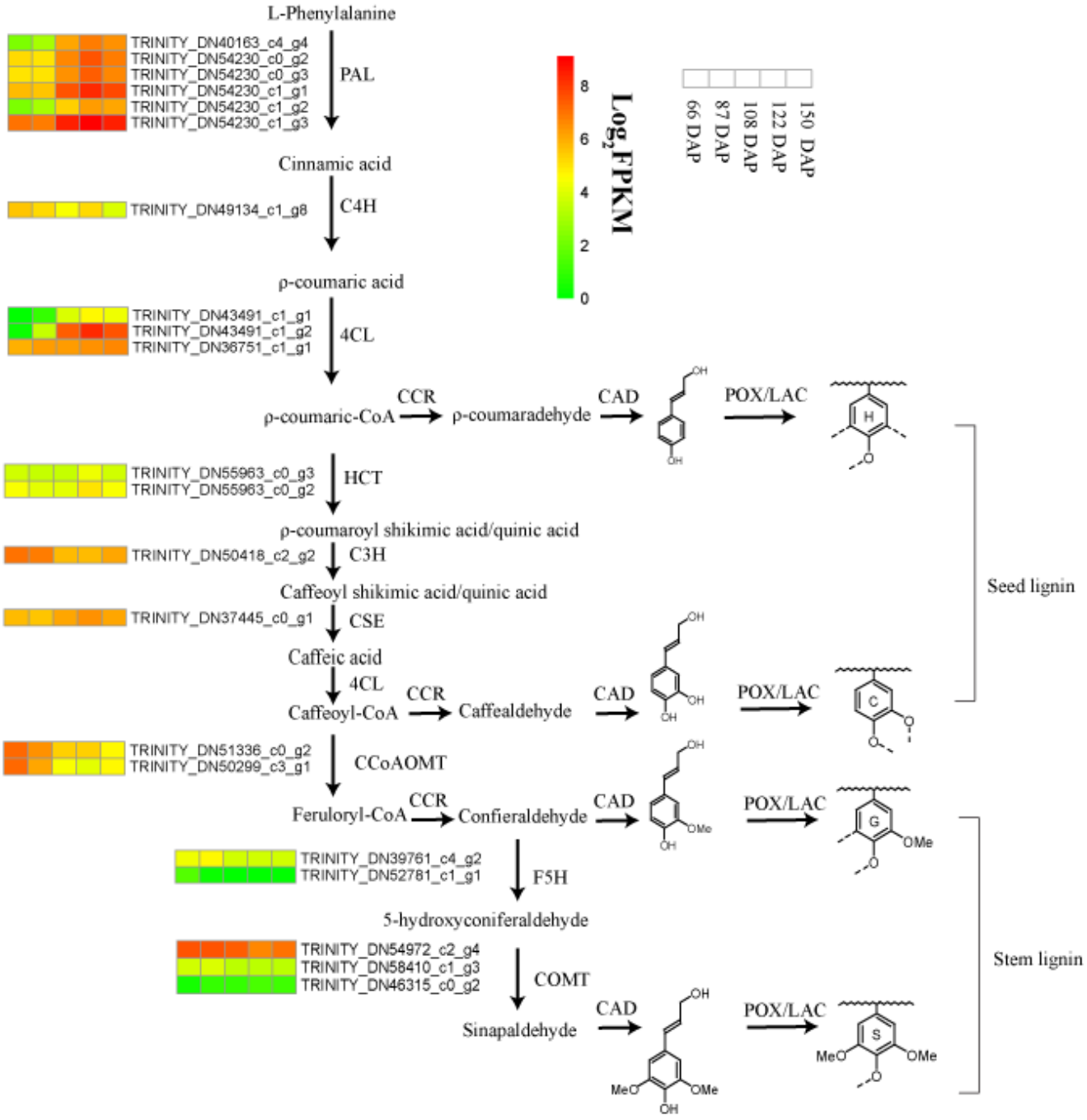

Figure 6. Lignin biosynthesis pathway and monolignol biosynthesis related gene expression in *P. armeniacum* .

Figure 6

Lignin biosynthesis pathway and monolignol biosynthesis related gene expression in P. armeniacum. PAL, phenylalanine ammonia lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate:CoA ligase; CCR, cinnamoyl-CoA reductase; CAD, cinnamyl alcohol dehydrogenase; C3H, 4-coumarate 3-hydroxylase; HCT, hydroxycinnamoyl-CoA:shikimate/quinate hydroxycinnamoyl transferase; CCoAOMT, caffeoyl-CoA 3-O-methyltransferase; COMT, caffeic acid O-methyltransferase; F5H, ferulate-5-hydroxylase; COMT, caffeic

acid O-methyltransferase; CAD, cinnamyl alcohol dehydrogenase; POX/LAC, peroxidases/laccase. Gene expression was scaled using the Z-score of FPKM (mean value of three biological replicates) in the heatmap. For each heatmap, the key is located at right side with FPKM values increasing from green to red.

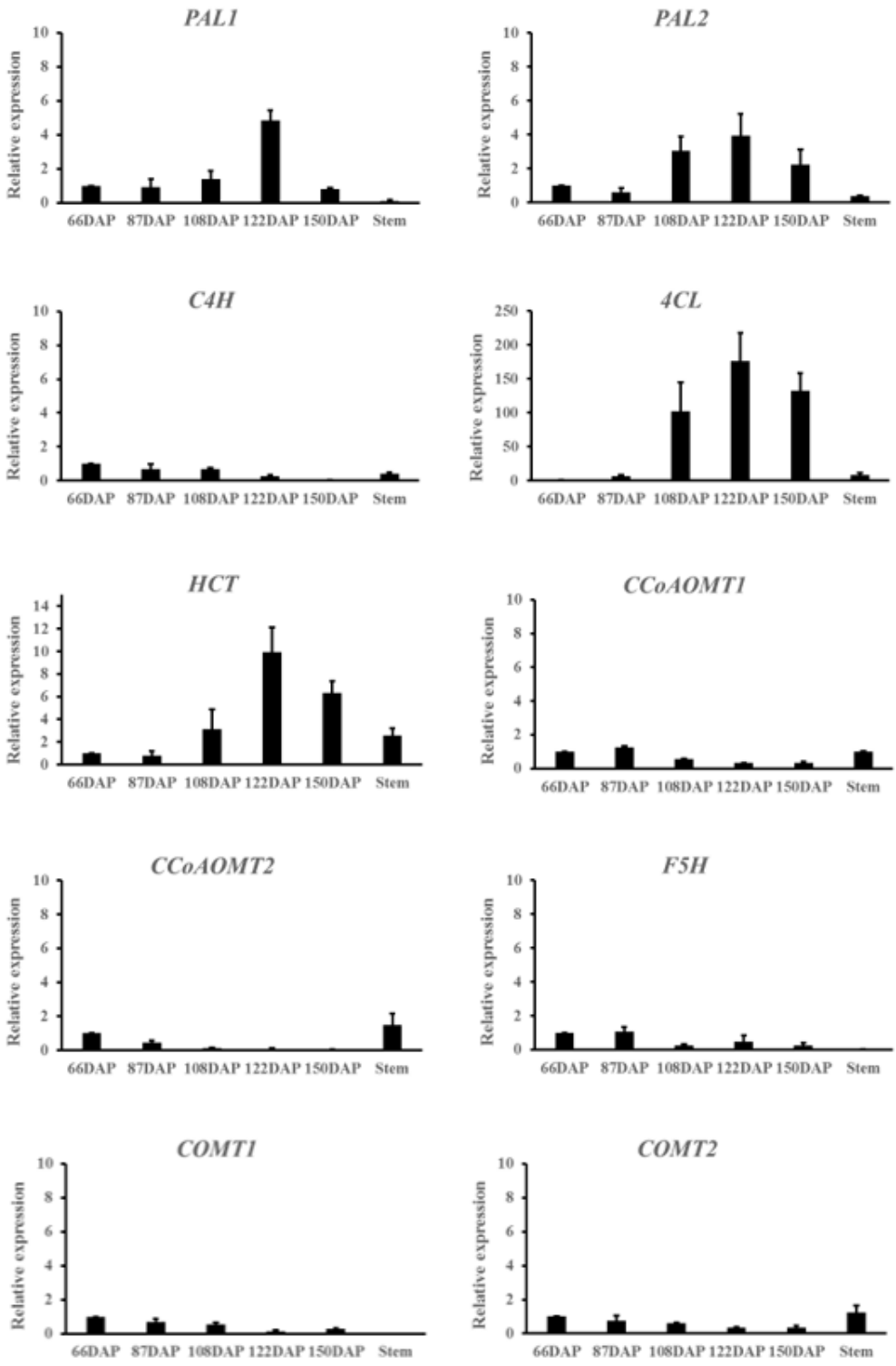

Figure 7

Validation of lignin-related gene expression in five stages of seed development, and stems. Expression levels were normalized to expression levels of actin.

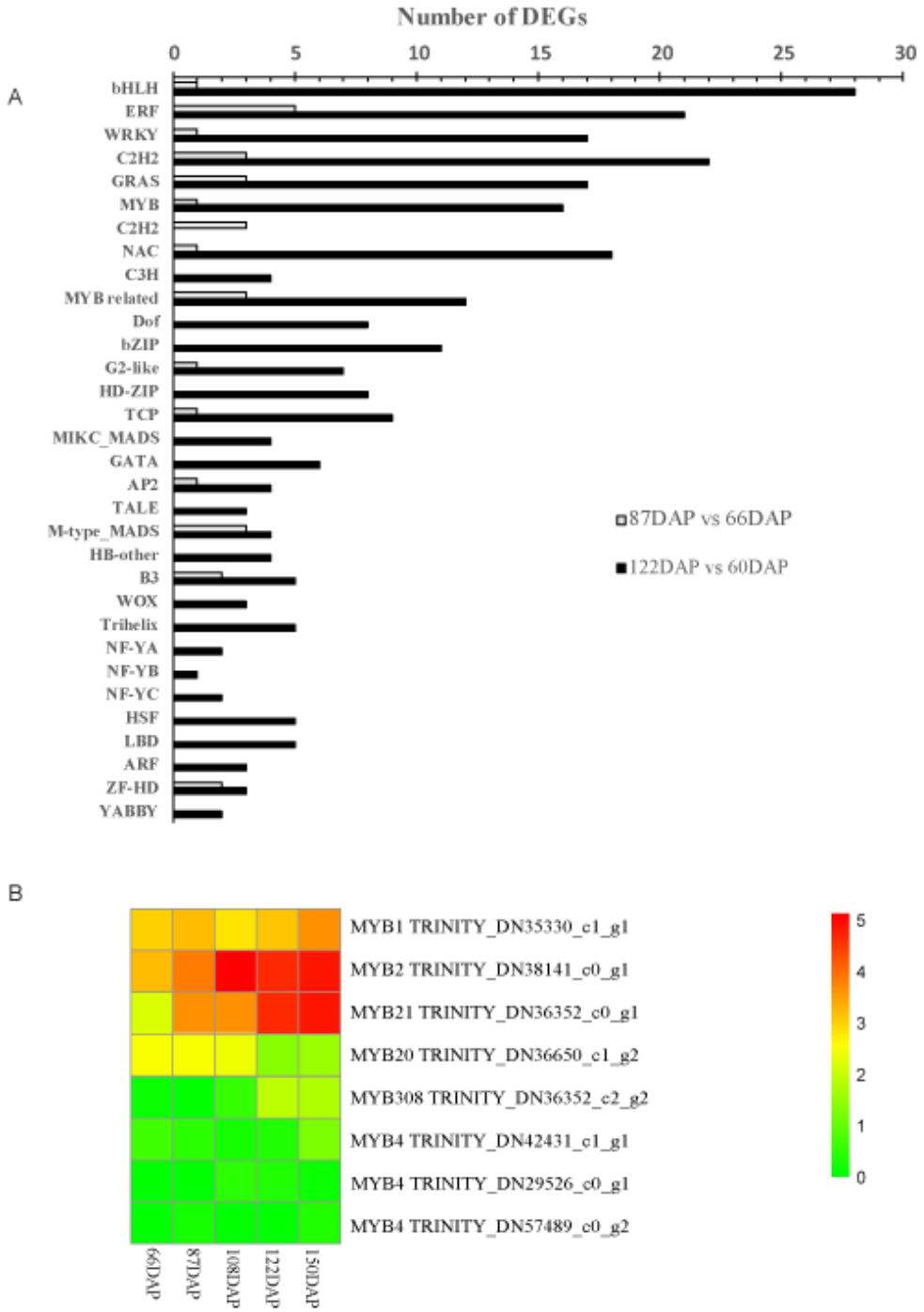

Figure 8. (A)Distribution of differentially expressed transcription factors (TFs); (B) Expression profiles of TFs potentially related to lignin synthesis.

Figure 8

Distribution of differentially expressed transcription factors (TFs); (B) Expression profiles of MYBs potentially related to lignin synthesis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.xlsx