

Global analysis insights into coupling relationship between lignan and lignin metabolism of flax

Dongwei Xie

Nantong University

Zhigang Dai

Chinese Academy of Agricultural Sciences Institute of Bast Fiber Crops

Zemao Yang

Chinese Academy of Agricultural Sciences Institute of Bast Fiber Crops

Qing Tang

Chinese Academy of Agricultural Sciences Institute of Bast Fiber Crops

Jian Sun (✉ sunjian8416@163.com)

Nantong University

Jianguang Su (✉ su_changsha@163.com)

Chinese Academy of Agricultural Sciences Institute of Bast Fiber Crops <https://orcid.org/0000-0002-9109-4019>

Research article

Keywords: flax, coupling, lignin, lignin, PacBio Iso-Seq, UGTs

Posted Date: March 31st, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-18412/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Flax (*Linum usitatissimum* L.) is one of the most important economic crops in the world. The lignin content of flax stems directly determines the quality of flax fibers. Flax seeds contain lignans of highly beneficial for human health.

Results: To elucidate the metabolic relationship between these compounds and the regulatory nodes of their metabolic processes, third generation (PacBio Iso-Seq) and second generation (Illumina) sequencing technologies were used to sequence the transcriptomes of a pair of flax cultivars with significant differences in lignan content. It was discovered that the differential expressed genes (DEGs) are significantly enriched in the lignin and lignan biosynthesis pathways. Furthermore, there are seven genes with significant differences in expression that were annotated as UDP-glucosyl transferases (UGTs). We found that lignan and lignin content is significantly negatively correlated with each other. SEM observations on flax bast fibers provided further evidence of this relationship.

Conclusions: This is the first full-length transcriptome analysis on flax plants using third-generation sequencing technologies, and it is also the first study to observe a negative correlation between lignin and lignan content of flax plants. Furthermore, it was found that UGTs are likely to be regulatory node genes for lignan and lignin metabolism.

Background

Flax (*Linum usitatissimum* L.) is an important economic crop, as textiles made from flax fibers (linen) are in great demand. However, the linen industry is facing a severe shortage of high-quality flax fibers, and lignin content is a major determinant of flax fiber quality. Firstly, excessively high lignin content causes the fibers to become hard and brittle, which directly affects the quality and textile properties of the fibers. Secondly, lignin content is a limiting factor for degumming processes and textile quality, and the wastewaters produced by the lignin removal process are extremely polluting [1]. Consequently, there is a significant level of interest among researchers worldwide in the use of genetic engineering to reduce the lignin content of flax crops or to alter their composition, in order to address the aforementioned issues from their root cause [2]. In recent years, research into lignin synthesis have mainly focused on the metabolic pathway of phenylpropanoid compounds and lignin-specific biosynthesis. These studies have shown that the activities of enzymes like phenylalanine ammonia-lyase (PAL), cinnamate 4-hydroxylase (C4H), 4-coumarate-CoA ligase (4CL), cinnamyl alcohol dehydrogenase (CAD), and cinnamoyl-CoA reductase (CCR) are closely related to the total amount of lignin synthesis in flax fibers [3]. The reduction of PAL activity in transgenic tobacco decreases its lignin content, but this mutation also affects the normative growth of the plant [4]. The inhibition of C4H activity significantly decreases the lignin content of transgenic tobacco, and the regulation of C4H expression also decreases the syringyl/guaiacyl (S/G) ratio in lignin, thereby altering the lignin composition; however, no signs of abnormal plant growth were observed [5]. In studies on multiple transgenic plants, it was discovered that the function of 4CL is not always the same between different plants, and the results obtained after altering 4CL expression differs

from one plant to another [6–8]. Based on the histochemical staining of lignin, it was found that CAD overexpression increased vein and stem lignification in transgenic tobacco, whereas CAD overexpression in antisense-transformed plants reduced xylem lignification [9]. Antisense inhibition of the CCR gene in *Arabidopsis thaliana* significantly reduced the lignin content in the xylem, and it also dramatically loosened the secondary cell wall of interfascicular fibers and vessels, which resulted in abnormal growth [10].

Flax seeds are extremely valuable products, as research has shown that flaxseed is rich in functional nutrients like α -linolenic acid, flaxseed gum, flaxseed protein and lignans [11]. Flax is an extremely rich source of lignans, as their lignan content is tens or hundreds of times that of other plants [12, 13]. For this reason, flax is considered by some as the “king of lignans.” Flax lignans have been shown to be effective in aiding the prevention and treatment of several severe diseases, and they may also play important roles in cancer prevention, strengthening the immune system and improving human health [14–18]. Therefore, the development and utilization of flax lignans is an indispensable part of flax seed refinement and utilization, as lignans greatly increase the added value and competitiveness of flax products.

Furthermore, the development of flax lignans is also of utmost importance for ensuring that the flax industry will continue to develop sustainably. There are currently several reports about the genes related to lignan biosynthesis in flax plants. Renouard et al. (2012) conducted a quantitative PCR analysis of LuPLR1 expression during the growth of flax seeds, and found that the expression of LuPLR1 is positively correlated with the accumulation of lignan [19]. Ghose et al. (2014) discovered that the UGT74S1 uridine diphosphate glycosyl transferase (UGT) gene plays a role in the glycosylation of secoisolariciresinol (SECO) into secoisolariciresinol monoglucoside (SMG) and diglucoside (SDG) [20]. Therefore, UGT genes may also be important for lignan biosynthesis in flax plants. Fofana et al. (2017a) performed ethyl methane sulfonate (EMS) mutagenesis on CDC Bethune flax and obtained 1996 M₂ families. Between these mutants, 93 strains exhibited mutations in the exonic region of UGT74S1, which inhibited SDG synthesis in their M₃ and M₄ descendants [21]. In another study, Fofana et al. (2017b) conducted a preliminary gene expression and function study and found that UGT74S3 and UGT74S4 genes do not have the same function as their UGT74S1 paralog, as they do not play a role in SDG synthesis [22].

Phenylpropanoids refer to a family of compounds that contain a benzene ring attached to a three-carbon side chain, which includes flavonoids, anthocyanins, coumarins, and lignin precursors; these compounds and their derivatives are important components in secondary plant metabolism [23, 24]. The lignin polymers in plants mainly consist of syringyl units (S), guaiacyl units (G), and p-hydroxyphenyl units (H), which are derived from sinapyl alcohol, coniferyl alcohol, and p-coumaryl alcohol, respectively; coniferyl alcohol-derived G-lignin is especially prevalent in angiosperms [25]. In addition to G-lignins, research has shown that coniferyl alcohol also forms another product, lignin, after a series of enzyme reactions [26]. However, there are no reports in the literature about the relationship between lignin and lignan, or the application of this relationship to obtain high-quality flax cultivars with low lignin content and high lignan content.

In this study, the full-length transcriptome of flax was sequenced. Long-read third-generation sequencing and short-read second-generation sequencing was used in a complementary manner to construct a full-length transcriptome database of flax plants, which was used to identify and investigate alternate splicing events in the flax transcriptome. Based on the second-generation sequencing data of flax seeds in various stages of growth, we investigated the enrichment of DEGs in the metabolic pathways of lignin and lignan, to perform a preliminary study on the couplings that may exist between lignin and lignan in metabolic pathways. To provide further support of the relationship between lignin and lignan, a correlation analysis was performed between flaxseed lignan content and stem lignin content in various stages of growth, in flax cultivars that exhibit significant differences in lignan content. In addition, scanning electron microscopy was performed on the stem cross-sections of these flax cultivars in various stages of growth. The use of genetic engineering to cultivate high-quality flax cultivars with low lignin content and high lignan content is an ideal two-pronged solution to the issues currently being faced by the flax industry. This study will elucidate the biosynthetic relationship between flax lignan and lignin and provide new ideas for the regulation of lignin and lignan biosynthesis. In addition, the findings of this study will be informative for the cultivation of high-quality, environmentally friendly flax cultivars with high lignan content and low lignin content.

Results

The full-length transcriptome of flax

To construct a complete set of transcripts, we collected high-quality RNA samples from the root, stem, leaf, petals, anthers, and stigma of flax plants during their flowering period, and then mixed the high-quality RNAs of these tissues. Then the mixed RNAs were sequenced: RNA fragments of varying size were selected (1-2 kb, 2-3 kb, and 3-6 kb) to construct libraries, and sequencing was performed using a PacBio RS II sequencer with eight cells per RNA strand. A total of 1202336 raw sequence reads were obtained, and each cell generated 150292 reads on average. After quality control was performed, 550240 high-quality ROIs were obtained, with 213559, 211111, and 125570 ROIs being obtained from the 1-2 kb, 2-3 kb, and 3-6 kb libraries, respectively. As we had hoped, the average length of the data generated by each library was the same as the estimated value (Fig. S1). Between these reads, 239600 were non-full-length transcripts while 258948 were full-length transcripts. In this study, full-length transcripts were identified by the presence of 5' end primers, 3' end primers, and 3' end poly(A) tails in the read sequence. In the subsequent analyses, we excluded reads shorter than 300 bp and chimeric reads.

The full-length transcripts were clustered to obtain high-quality non-redundant transcripts. During the preliminary steps of the clustering analysis, it was found that 90% of the transcripts shorter than 6 kb could be uniquely mapped to the genome (Fig. 1A). In addition, the full-length transcripts produced by PacBio sequencing (average 2206 bp) were found to be longer than the reference genome (average 1253 bp) (Fig. 1B). Furthermore, our analysis discovered that 4258 genes only contain one exon, which is lower than that annotated in the reference genome (9227 genes). In the PacBio data, the number of genes that contain at least 10 exons is higher than that of the reference genome (13603 versus 6244). The PacBio

Iso-Seq technique determined that each gene contains 7.4 exons on average, which is 1.7 times that of the reference genome (Fig. 1C).

The full-length transcripts were mapped to the flax genome; it was determined that 47020 transcripts could be mapped to the reference gene model, and 43472 genes were discovered in the reference genome. As the full-length transcripts that were identified by the Iso-Seq technique were directly sequenced without requiring further assembly, these data are an excellent resource for improving the annotations of the reference genome. The new transcripts that were obtained from the analysis of alternative splicing in a variety of samples were merged. A total 1690 new gene models were generated by the merging of these transcripts with known transcripts. Additionally, we identified 1043 fusion transcripts that can be mapped to the flax genome; these fusion transcripts mainly originated from fusion events between different chromosomes (Fig. S2).

The identification of new transcripts and alternative polyadenylation (APA)

A total of 36353 new transcripts that could only be mapped to the intergenic region were identified in this study. Among these transcripts, 4057 may be lncRNAs. The average length of these putative lncRNAs is 1.8 kb, and their lengths vary between 0.3 kb and 4.4 kb. The potential of these new transcripts to code for proteins was determined via coding potential analyses, which resulted in the discovery of 294 lncRNAs (Fig. S3). By analyzing the origins of the lncRNA sequences, it was determined that 193 of these sequences are lncRNAs, 34 are antisense-lncRNAs, 14 are intronic-lncRNAs, and 53 are sense-lncRNAs (Fig. S4). Target gene prediction was performed on the predicted lncRNA sequences. Based on the acting mechanisms of lncRNA on its target genes, two approaches were used to predict the target genes; 7481 target genes were predicted based on positional relationships (Table S1), while 7661 target genes were predicted by the LncTar [27] target gene prediction tool, which predicts target genes based on the complementarity of base pairings.

Research has shown that APA plays an important role in the regulation of gene expression. APA in the last exon could lead to the generation of mRNA isoforms with different 3' UTR lengths, which may affect the stability, translational efficiency, transfer, and subcellular localization of RNA (mRNA) in cells [28]. Through an analysis of the 3' end of full-length transcripts, the APA sites of flax were accurately identified for the first time. In this study, we identified a total of 19048 genes with poly(A) sites; 13502 of these genes contained 26637 transcripts with multiple poly(A) sites, and 3707 genes contained a minimum of five poly(A) sites (Fig. 2). It was determined that each gene contained an average 2.98 poly(A) sites.

Splicing regions and alternative splicing analysis

One of the most important uses of PacBio Iso-Seq is its ability to compare different full-length transcripts of the same gene to identify alternative splicing events. However, there are no previous studies reporting alternative splicing events in the flax genome. Based on the high-confidence full-length transcripts that were identified in our study, we conducted a systematic analysis of the alternative splicing events that occurred in the flax genome. Firstly, we identified 16391 alternative splicing events in 9236 genes. Five

types of alternative splicing events were identified using the Astalavista [29] program, including exon skipping (ES), alternative 3' splice sites (A3'S), mutually exclusive exons (ME), alternative 5' splice sites (A5'S), and intron retention (IR) (Fig. 3A). By counting the number of genes corresponding to each type of splicing event, we discovered that IR occurred in 65.53% of the genes. Therefore, IR events account for most of the alternative splicing events in the flax genome (Table 2). The second and third most common alternative splicing events are A3'S and A5'S events, respectively. A similar number of A5'S and ES events were observed in this study. At the transcript level, 36353 new transcripts were identified in 24564 genes, and 13758 of these genes contain only one transcript. In the reference genome, most of the genes were annotated as single-transcript genes (43458 genes); only 13 genes have two transcripts, and no genes have three transcripts. From the PacBio data, there were 5190 genes with a minimum of three transcripts. The PacBio Iso-Seq technique determined that each gene contained 1.23 transcripts on average (Fig. 3B), which is 1.2 times that of the reference genome's annotations. These results indicate that alternative splicing events have increased the complexity of the flax transcriptome.

RT-PCR was performed on four randomly selected genes to validate the accuracy of alternative splicing event identification. Firstly, the transcripts of each gene were compared to locate a consensus sequence. Primers were then designed for the sequence, thereby allowing all of the transcripts to be amplified in a single experiment (Table S3). The results of this experiment indicate that the sizes of the amplified fragments were consistent with the expected sizes. In addition, some of the transcripts exhibited tissue specificity in their expression. For example, the PB.2775 gene codes for a MATE-like protein, and is able to transcribe seven transcript isoforms through A5'S events. Between these transcripts, at least two are only expressed in flax seeds during their development, and they are not expressed in other parts of the plant (Fig. 4, Fig. S5). Therefore, it is important to identify the transcripts that exhibit tissue-specific expression.

Differential gene enrichment analysis and functional annotation

DESeq was used to screen for mRNA DEGs in Shuangya 4 and NEW flax cultivars, which were subsequently subjected to functional annotation and screening (selection criteria: fold change > 2, FDR < 0.01). The DEGs are shown in Table S4 A. The mRNA expression levels of Shuangya 4 and NEW cultivars varied significantly during their development in the post-flowering stage. A total of 1643 DEGs were identified, of which 678 were upregulated while 965 were downregulated. We observed 965 DEGs 10 days after flowering (the highest number of DEGs, by far), with 432 being upregulated and 533 being downregulated; 404 DEGs were observed 30 days after flowering, and a mere 274 DEGs were observed 20 days after flowering. Functional annotation was performed on the 41 DEGs with |fold change| > 3 that were detected in two or more developmental stages. Seven, one, and one DEGs were annotated as *UDP-glycosyltransferase*, *laccase*, and *cytochrome P450*, respectively. In addition, five DEGs were annotated in all three developmental periods (Table S5). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was then performed on the mRNA DEGs. The functions of the '10 days after flowering' (henceforth denoted the '10-day' stage) DEGs in both cultivars were mainly related to plant-pathogen interactions, peroxisome, protein processing in the endoplasmic reticulum, spliceosomes, and ribosomes (Fig. S6 A). The functions of the '20 days after flowering' (20-day stage) DEGs were mainly related to

phenylpropanoid metabolism, spliceosomes, plant hormone signal transduction, ribosomes, and starch and sucrose metabolism (Fig. S6 B). The functions of the '30 days after flowering' (30-day stage) DEGs pertained to phenylpropanoid metabolism, starch and sucrose metabolism, photosynthesis-antenna proteins, DNA replication, and photosynthesis (Fig. S6 C). In particular, the raw materials required in all stages of plant growth and development are supplied by starch and sucrose metabolism [30], whereas spliceosomes and ribosomes are large and complex molecular machines that are indispensable for eukaryotic gene expression. We also identified DEGs that participate in photosynthetic antenna proteins and photosynthesis-related pathways, which hints that these DEGs may play a role in the formation and accumulation of flaxseed nutrients. More importantly, it was shown that many DEGs participate in phenylpropanoid metabolism, and these DEGs generally result in significant phenylpropanoid enrichment at the 20-day and 30-day stages. Hence, the phenylpropanoid metabolism pathway could produce large amounts of secondary metabolites [31,32]. Gene ontology (GO) annotation and clustering were also performed on the DEGs to reveal their functions. topGO directed acyclic graphs were plotted. Upon further analysis, it was discovered that the DEG functions of the Shuangya 4 and NEW cultivars exhibit significant enrichment in the (+)-pinoresinol catabolic process, (-)-lariciresinol biosynthetic process, (+)-lariciresinol biosynthetic process, protein-chromophore linkage, peptidyl-histidine phosphorylation, amine transport, and cellular responses to cold (Fig. S6 D-F). In addition to the pathways and processes necessary for plant growth, it was revealed by KEGG pathway analysis and GO annotation that many of the DEGs participate in the phenylpropanoid metabolism pathway in the KEGG pathway, as well as the (+)-pinoresinol catabolic process, (-)-lariciresinol biosynthetic process, and (+)-lariciresinol biosynthetic process in the GO annotations (Fig. S7). Phenylpropanoid metabolism is a part of the lignin metabolic pathway in plants [33], whereas the (+)-pinoresinol catabolic process, (-)-lariciresinol biosynthetic process, and (+)-lariciresinol biosynthetic process are necessary steps for lignan biosynthesis in flax plants [34]. Therefore, we concluded that these DEGs play an important role in the lignin and lignan metabolism of flax plants.

3.5 The differentially expressed genes of lignin and lignin metabolism were verified by qRT-PCR

The seven differential genes screened in this study were all annotated as UDP-glycosyltransferase one. So far, 98 families (CAZy database) of glycosyltransferase have been divided according to amino acid sequence similarity, among which secondary metabolites capable of glycosylation generally belong to UDP-glycosyltransferase family one [35]. Therefore, it is highly likely that these seven differential genes have glycosylation effects on the secondary metabolite of lignan or lignin. The expression quantity of these seven *UGTs* at different stages of flaxseed and stem development in different flax varieties were tested by qRT-PCR (Fig. 5, Table S6). We found that except for *lus10025514* and *lus10016461*, the expression levels of the other five *UGTs* in Shuangya 4 were significantly higher than that in NEW at different stages of flax seed and stem development. Since the functional annotation of these differentially expressed genes was related to the metabolism of lignan and lignin, and the expression level of these genes in the two varieties had certain rules. So, it is speculated that these 5 *UGTs* were most likely closely related to lignan and lignin metabolism. In addition, it was found that the expression

patterns of 7 *UGTs* in each stage of flax seed development were basically consistent with transcriptome sequencing results (Table S5), which proved the reliability of this study.

Correlation between lignin and lignan content at each stage of development

Although it has been demonstrated that lignin and lignan are both produced by the same metabolic pathway [26], a relationship between lignin and lignan has yet to be reported in the literature. Therefore, revelations about the relationship between lignin and lignan will be of significant importance for the cultivation of high-quality flax cultivars with high lignan and low lignin content. A correlation analysis was performed between the seed lignan content and tissue lignin content of flax plants in various stages in growth (Table 3). At 0 days after flowering, the lignan content of flax seeds was extremely significantly negatively correlated with lignin content in the roots, lower stem, and upper stem, and significantly negatively correlated with lignin content in the leaves. At 5-30 days after flowering, the lignan content of flax seeds was negatively correlated with the lignin content of the roots, but this was not significant. Significant and extremely significant negative correlations were observed between seed lignan content and the lignin content of the lower stem and upper stem, respectively. A positive correlation was observed between seed lignan content and leaf lignin content, albeit at a non-significant level (Table 3). Therefore, the accumulation of seed lignan and stem lignin in the post-flowering period may have competed for the same substrate, which resulted in a negative correlation between seed lignan content and stem lignin content in flax plants.

SEM observations of the stem in various stages of flax plant growth

SEM observations were performed on the stem cross-sections of two flax cultivars with significant differences in lignan content, Shuangya 4 (Fig. 6 A-C) and NEW (Fig. 6 D-E), in various stages of growth. This was performed to provide further evidence supporting the relationship between lignan and lignin. 10 days after flowering, the bast fiber bundles of the high-lignan cultivar (Shuangya 4, Fig. 6A) and low-lignan cultivar (NEW, Fig. 6D) did not differ significantly in appearance, as they were still in their initial stages of formation. At 20 days after flowering, the bast fiber bundles of the high-lignan Shuangya 4 cultivar (Fig. 6B) and low-lignan NEW cultivar (Fig. 6E) began to thicken and form hollow fibers. However, we observed that the lignification of the NEW cultivar's fiber bundle was significantly greater than that of the Shuangya 4 cultivar, indicating that the NEW cultivar has accumulated more lignin in the stem than the Shuangya 4 cultivar. At 30 days after flowering, the lignification of the NEW cultivar had increased further (Fig. 6F), whereas the fiber bundle of the Shuangya 4 cultivar displayed a lower degree of lignification. The SEM observations are consistent our correlation analyses (Table 3), thus providing further evidence that a negative correlation exists between the lignan and lignin content of a flax plant.

Discussion

Flax is an important economic crop, and the findings that have been published about flax genome sequencing have helped to promote basic research about flax plants [36]. However, the reported flax genome still lacks high-quality transcript annotations, which hinders in-depth functional genomics

studies on flax plants. In this present study, Iso-Seq sequencing technology was used to produce 20.34 Gb of clean data from eight cells, and a total of 113471 consensus transcript sequences were obtained. After corrections were performed using non-full-length sequences, we obtained 95403 high-quality full-length transcripts, as well as 18068 low-quality full-length transcripts. Error correction was then performed on the 18,068 low-quality sequences using data derived from a second-generation sequencing technique. The gene annotations, isoform transcripts, and UTRs of the flax genome were updated using these data. In addition, lncRNAs were also identified in our study. The merging of these data with second-generation sequencing datasets will help to promote functional genomics research on flax plants.

The analysis of alternative splicing events performed in our study will accelerate the development of functional genomics research with respect to flax plants in the future. Owing to the lack of an accurately annotated flax transcriptome data, functional differences between transcripts of the same gene have not been noted in previous flax plant studies. In eukaryotic lifeforms, alternative splicing often leads to the generation of different transcripts. Based on Iso-Seq data, we have identified 16391 alternative splicing events in 9236 genes, with IR events accounting for the vast majority of these events. This result is consistent with the findings of previous studies in other plants [37–40]. By analyzing the transcripts of various tissues, we have found that alternative splicing may have resulted in the same gene with different functions across different tissues. These results generally indicate that alternative splicing events have greatly increased the complexity of the flax transcriptome, and they also remind us that transcript-level approaches are necessary in gene studies.

The processing and utilization of plant materials are mainly limited by the presence of lignin. Excessively high levels of lignin content in plant fibers will result in hard and brittle fibers, which are difficult to weave into textiles. In addition, due to the immense difficulty of lignin removal, plant materials with high lignin content will result in high degumming costs and industrial pollutant discharges [41]. Therefore, the lignin content of a flax plant will directly affect the quality of its fibers and thus, its economic value. Flax seeds also contain lignan, which is a functional component [11]; the development and utilization of this functional component will greatly increase the added value of flax products. However, studies on flax plants have generally focused either on lignan or lignin, and there have been no reports about the relationship between these compounds. Huis et al. (2012) utilized a multi-omics approach that combined metabolomics and transcriptomics to study lignification in the woody part of flax stems and extraxylary bast fibers [26]. In their study, they briefly mentioned that coniferyl alcohol, a metabolic intermediate, could ultimately be converted into lignin and lignan. However, there was no mention of the derivative relations between lignin and lignan. By measuring the seed lignan content and stem lignin content of flax plants in various stages of the post-flowering period, we discovered that lignan and lignin content is negatively correlated with each other. Based on SEM observations on the bast fibers of a pair of flax cultivars with significant differences in lignan content, in different periods of growth, we discovered that lignin accumulation was significantly lower in the high-lignan cultivar (which has thin bast fiber walls) than in the low-lignan cultivar (which has thick bast fiber walls), during the middle and later periods of flax growth. This finding supports the hypothesis that the lignan and lignin content of a flax plant is

negatively correlated with each other. In the transcriptome data of flaxseeds that were collected from these flax cultivars at different stages of growth, significant DEG enrichment was observed in the KEGG phenylpropanoid metabolism pathway. The derivatives of phenylpropanoid are p-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol, which serve as substrates for the biosynthesis of G-lignin, H-lignin, and S-lignin production, respectively [42]. In particular, coniferyl alcohol also serves as the substrate for lignan production. As the DEGs are significantly enriched in the phenylpropanoid metabolism pathway, it is implied that these DEGs are closely related to lignin and lignan metabolism. In the GO annotations, the DEGs are mainly enriched in the (+)-pinoresinol catabolic process, (-)-lariciresinol biosynthetic process, and (+)-lariciresinol biosynthetic process, which are necessary processes for lignan biosynthesis [19]. These results imply that the genes related to lignin and lignan metabolism are strongly coupled to each other. Functional annotation analysis was performed on the DEGs that were significantly enriched in the phenylpropanoid metabolism pathway ($|\text{fold change}| > 3$) and detected in a minimum of two stages of growth; a total of 41 DEGs were selected, and seven of these were annotated as UDP-glycosyltransferase 1. At present, glycosyltransferases have been divided into 98 families (CAZy database) based on similarity of their amino acid sequences. In particular, the glycosyltransferases that are able to glycosylate secondary metabolites belong to family 1 (i.e., UDP-glycosyltransferase 1) [35]. It is noteworthy that all seven of the DEGs that were selected in our study have been annotated as a part of the UDP-glycosyltransferase 1 family, implying that these DEGs play a role in the glycosylation of secondary metabolites like lignan or lignin. Lin et al. 2016 created a *ugt72b1* mutant in *Arabidopsis thaliana*, which resulted in severe ectopic lignification and the enhancement of cell wall lignification; lignin content analyses on this mutant also revealed significantly higher levels of lignin than the wildtype. In addition, Lin et al. 2016 also measured the p-coumaryl alcohol, coniferyl alcohol and sinapyl alcohol (i.e., lignin precursor) content of the mutant [43]. It was determined that the coniferyl alcohol content of the mutant greatly exceeds that of the wildtype and even surpasses that of the overexpression line; this observation greatly puzzled the authors. Our research has resolved this puzzle, as we have found that lignan and lignin compete for the same substrate-coniferyl alcohol. Furthermore, UGTs are regulatory hub genes in lignan and lignin metabolism. Therefore, we speculate that the function of the UGTs was disrupted in the mutant strain, which blocked or weakened the pathway that converts coniferyl alcohol into lignan, thus resulting in large amounts of coniferyl alcohol being accumulated within the mutant. Our hypothesis also finds support from the results of previous studies, albeit from a different viewpoint. A preliminary study by Ghose et al. (2014) discovered that the *UGT74S1* gene plays a role in the glycosylation of SECO into lignan, thus indicating that UGTs could play a significant role in flax lignan biosynthesis [20]. Fofana et al. (2017a) performed EMS mutagenesis on CDC Bethune flax and obtained 1996 M_2 mutants; between these mutants, 93 strains exhibited mutations in the exonic region of *UGT74S1*, which inhibited lignan synthesis in their M_3 and M_4 descendants [21].

Conclusion

In our study, full-length transcriptome analysis of flax was performed for the first time in this study, which has helped to fill gaps in the flax reference genome. We discovered that alternative splicing events have

increased the complexity of the flax transcriptome. A concerted analysis based on second-generation sequencing revealed that DEGs are mainly enriched in the phenylpropanoid metabolism pathway, which is the pathway that produces lignan and lignin. Based on the measurement of stem lignin content and flaxseed lignan content in different stages of plant growth, as well as SEM observations on flax cultivars with significant differences in lignan content, we found that the lignan content of a flax plant is negatively correlated with its lignin content. Many of the DEGs were functionally annotated as UGTs; based on the results of previous studies, we speculate that UGTs are regulatory node genes for lignan and lignin metabolism.

Methods

Plant materials

Flax plants of Shuangya 4 and NEW were sown at the Harbin Experimental Base of Heilongjiang Academy of Agricultural Sciences (45°65'N, 126°68'E), Harbin, China, in April of 2018. The average annual rainfall from seeding to harvest was 350.2 mm and the average annual temperature was 4.26°C. Each cultivar was planted in 10-m lines, with a 20-cm inter-line gap, and each test had ten rows and the area was 20 m², with three replicates. The field management was the same as the local field management.

PacBio sequencing

Root, stem, leaf, flower petal, anther, and stigma samples were acquired from the Shuangya 4 flax cultivar when it began to flower. RNA from each sample was extracted and equally mixed, and the SMARTerTM PCR cDNA Synthesis Kit was used to synthesize full-length cDNA from the mRNA. The BluePippin size-selection system was used to select full-length cDNA fragments, which were used to construct cDNA libraries of varying size. The selected full-length cDNAs were PCR amplified, and end repair was performed on the full-length cDNAs, followed by the ligation of PacBio Single Molecule, Real-Time (SMRT) bell adaptors and exonuclease digestion. Secondary screening was performed using the BluePippin system to construct the sequencing libraries, which were quantified using a Qubit 2.0 fluorimeter. An Agilent 2100 Bioanalyzer was used to inspect the size of the sequencing libraries. Sequencing was performed only after the sizes of the sequencing libraries matched the expected sizes. Sequencing was performed on a Pacific Bioscience (PacBio) RS II sequencer, with a movie time of 240 minutes.

Raw SMRT sequences were deposited in the National Center for Biotechnology Information (NCBI) and can be accessed in the database (<https://www.ncbi.nlm.nih.gov/>) under accession PRJNA478803.

Illumina RNA-seq sequencing

The flax seeds of Shuangya 4 and NEW flax cultivars were obtained 10, 20, and 30 days after flowering, with three biological replicates for each sample. RNA extraction was conducted on a total of 18 samples.

Second-generation transcriptome sequencing was performed using an Illumina HiSeq sequencing system. Detailed process of transcriptome analysis as described in our team previous research [44]. Raw Illumina sequences were deposited in the National Center for Biotechnology Information (NCBI) and can be accessed in the database (<https://www.ncbi.nlm.nih.gov/>) under accession PRJNA476039.

Identification of long non-coding RNA (lncRNA)

The full-length transcriptome was compared to the genetic model of flax, so as to enable the identification of new transcripts. The transcripts that could not be matched to the genetic model were deemed as new transcripts. lncRNAs were then identified in these new transcripts. As they do not code for proteins, lncRNAs can be identified by screening the transcripts for coding potential. lncRNA prediction was performed on the newly discovered transcripts using the most widely used methods for coding potential analysis, including the Coding Potential Calculator (CPC) [45], Coding-Non-Coding-Index (CNCI) [46], Coding Potential Assessment Tool (CPAT) [47], and the Pfam protein family's database. The transcripts that were identified as lncRNA by all four methods were treated as high-confidence lncRNA transcripts.

Alternative polyadenylation site analysis

The TAPIS pipeline [48] was used to identify the alternative polyadenylation sites of the full-length transcripts. In nucleotide sequences that contain eight adenylates and a maximum of two non-adenosines in the 3' end, each starting base was identified as an alternative polyadenylation site. The alternative polyadenylation sites of the genome were then extracted by comparing the transcripts to the reference genome.

Quantitative real-time PCR

Fresh seed and stem (5 cm long, 5 cm-10 cm above the cotyledon nodes) samples at three different developmental stages of 10, 20 and 30 day after flowering were collected. The samples were put into liquid nitrogen immediately and stored at -80 °C until RNA isolation. The method of quantitative real-time PCR (qRT-PCR) was described by our previous report [49]. All the qRT-PCR experiments were performed for three biological replicates, and each with three technical repeats under the same conditions.

Determination of lignan content

The seeds of Shuangya 4 and NEW flax cultivars were obtained 0, 5, 10, 20 and 30 days after flowering, with three biological replicates prepared for each sample. Fixation was performed at 105°C for 30 minutes, and the samples were then baked at 80°C until they reached a constant weight. The dried and fixed samples were then grounded into a powder and passed through a #100 mesh. The resulting powder was stored in a desiccator until further use.

In this study, the High-Performance Liquid Chromatography (HPLC) procedure of Charlet et al. (2002) (with minor modifications) was used to quantify the lignan content of the flax seeds [50]. HPLC was

performed using an Agilent 1260 HPLC and a diode array detector (DAD). The mobile phases were methanol (A) and ultrapure water (B). The chromatography column was an Agilent C18 (4.6 mm × 250 mm × 5 μm), and the temperature of the column was 35°C. The detection wavelength was 290 nm, and 20 μL of sample was injected into the column.

Determination of lignin content

The lignin content of the experimental materials was determined using the NYT 2337-2013 Agricultural Industry Standard of the People's Republic of China (*Testing Retted Jute/Retted Kenaf Fibre Lignin with Sulfuric Acid Method*), with minor modifications. The root, lower stem (5 cm long, 5 cm-10 cm above the cotyledon nodes), upper stem (5 cm long, 5 cm-10 cm below the first branch), and leaves of Shuangya 4 and NEW flax cultivars were obtained 0, 5, 15, 20, and 30 days after flowering. Three biological replicates were prepared for each sample.

Preparation for scanning electron microscopy

The lower stem (5 cm segment, 5 cm - 10 cm above the cotyledon node) of Shuangya 4 and NEW flax plants were collected 0, 5, 10, 20, and 30 days after flowering, as samples for SEM. The flax stem samples were fixed in 2.5% glutaraldehyde at 4°C, washed with PBS buffer three times (10 minutes for each wash), and then fixed in 1% osmium tetroxide for 2 hours, followed by another round of PBS buffer washes (3 washes, 10 minutes for each wash). The samples were dehydrated using a series of ethanol concentrations in distilled water (30%, 50%, 70%, 90%, and 100% ethanol, for 10 minutes each); then the samples were dehydrated twice in 100% ethanol. Critical point drying was then performed on the samples, which were then sputtered in gold. An Eiko (Japan) IB-3 ion coater was used to coat the samples in gold, and SEM was performed using a JEOL (Japan) JSM-6390LV scanning electron microscope.

Abbreviations

DEGs: Differential expressed genes; UGTs: UDP-glucosyl transferases; SEM: Scanning electron microscope; PAL: Phenylalanine ammonia-lyase; C4H: Cinnamate 4-hydroxylase; 4CL: 4-coumarate-CoA ligase; CAD: Cinnamyl alcohol dehydrogenase; CCR: cinnamoyl-CoA reductase; SECO: Secoisolariciresinol; SMG: Secoisolariciresinol monoglucoside; EMS: Ethyl methane sulfonate; SMRT: PacBio Single Molecule, Real-Time; NCBI: National Center for Biotechnology Information; lncRNA: Long non-coding RNA; CPC: Coding Potential Calculator; CNCI: Coding-Non-Coding-Index; CPAT: Coding Potential Assessment Tool; qRT-PCR: Quantitative real-time PCR; HPLC: High-Performance Liquid Chromatography; DAD: Diode array detector; APA: Alternative polyadenylation; ES: exon skipping; A3'S: Alternative 3' splice sites; ME: Mutually exclusive exons; A5'S: Alternative 5' splice sites; IR: Intron retention ; GO: Gene ontology

Declarations

Ethics approval and consent to participate

The plant materials were obtained from the Chinese Crop Germplasm Resources System (CGRIS; <http://www.cgris.net/>). Sampling of plant materials were performed in compliance with institutional, national, and international guidelines. The materials were publicly available for non-commercial purposes.

Consent for publication

Not applicable.

Competing Interests

The authors declare no conflict of interest. All the experiments undertaken in this study were complied with the current laws of China.

Funding

This work was supported by the National Natural Science Foundation of China (31801409); The Project of Heilongjiang Natural Science Foundation of China (C2018061); China Agriculture Research System (CARS-16-E01); The Scientific and Technological Innovation Project of Chinese Academy of Agricultural Sciences (CAAS-ASTIP-2017-IBFC01); National Infrastructure for Crop Germplasm Resources(NICGR-13). The founders did not play any roles in the design, analysis, and interpretation of this study or relevant data.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files. Raw Illumina sequences were deposited in the National Center for Biotechnology Information (NCBI) and can be accessed in the database (<http://www.ncbi.nlm.nih.gov/biosample/7551536>) under accession PRJNA478803 and PRJNA476039.

Author contributions

D.X., Q.T. and Z.Y. carried out most of the experimental work; Collections of flax germplasm resources were performed by D.Z.; D.X. designed the research and wrote the manuscript; and J.S. and J.S. revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thanks the National Bast Fiber Crops Germplasm Improvement Center of Flax Branch Center for kindly supplying the experimental platform.

References

- [1] Magdalena WK, Starzycki M, Zebrowski J, Oszmiański J, Szopa J. Lignin deficiency in transgenic flax resulted in plants with improved mechanical properties. *Journal of Biotechnology*. 2007;128:919-934.
- [2] Summerscales J, Dissanayake NPJ, Virk AS, Hall W. A review of bast fibres and their composites. Part 1-Fibres as reinforcements. *Composites Part A: Applied Science and Manufacturing*. 2010; 41:0-1335.
- [3] Chen HY, Babst BA, Nyamdari B, Hu H, Sykes R, Davis MF, Harding SA, Tsai CJ. Ectopic Expression of a Loblolly Pine Class II 4-Coumarate: CoA Ligase Alters Soluble Phenylpropanoid Metabolism but not Lignin Biosynthesis in *Populus*. *Plant & Cell Physiology*. 2014; 55: 1669.
- [4] Elkind Y, Edwards R, Mavandad M, Hedrick SA, Ribak O, Dixon RA, Lamb CJ. Abnormal plant development and down-regulation of phenylpropanoid biosynthesis in transgenic tobacco containing a heterologous phenylalanine ammonia-lyase gene. *Proceedings of the National Academy of Sciences*. 1990; 87: 9057-9061.
- [5] Sewalt V, Ni W, Blount JW, Jung HG, Masoud SA, Howles PA, Lamb C, Dixon RA. Reduced Lignin Content and Altered Lignin Composition in Transgenic Tobacco Down-Regulated in Expression of L-Phenylalanine Ammonia-Lyase or Cinnamate 4-Hydroxylase. *Plant Physiology*. 1997;115:41-50.
- [6] Lu H, Zeng Q, Zhao YL, Wang SS, Jiang XN. Xylem-specific expression of a GRP1.8 promoter::4CL gene construct in transgenic tobacco. *Plant Growth Regulation*. 2003;41: 279-286.
- [7] Voelker SL, Lachenbruch B, Meinzer FC, Jourdes M, Ki C, Patten AM, Davin LB, Lewis NG, Tuskan GA, Gunter L, et al. Antisense Down-Regulation of 4CL Expression Alters Lignification, Tree Growth, and Saccharification Potential of Field-Grown Poplar. *Plant Physiology*. 2010;154:874-886.
- [8] Voelker SL, Lachenbruch B, Meinzer FC, Strauss SH. Reduced wood stiffness and strength, and altered stem form, in young antisense 4CL transgenic poplars with reduced lignin contents. *New Phytologist*. 2011;189: 1096-1109.
- [9] Wang Y, Zhang X, Yang S, Wang C, Lu G, Wang R, Yang Y, Li D. Heterogenous expression of *Pyrus pyrifolia* PpCAD2 and PpEXP2 in tobacco impacts lignin accumulation in transgenic plants. *Gene*. 2017;637:181-189.
- [10] Goujon TFV, Ferret V, Mila I, Pollet B, Ruel K, Burlat V, Joseleau JP, Barriere Y, Lapierre C, Jouanin L. Down-regulation of the *AtCCR1* gene in *Arabidopsis thaliana*: effects on phenotype, lignins and cell wall degradability. *Planta*. 2003; 217:218-228.
- [11] Meagher LP, Beecher GR. Assessment of Data on the Lignan Content of Foods. *Journal of Food Composition and Analysis*. 2000;13: 935-947.
- [12] Kiely M, Faughnan M, Wähälä K, Brants H, Mulligan A. Phyto-oestrogen levels in foods: the design and construction of the VENUS database. *British Journal of Nutrition*. 2003;89:S19-S23.

- [13] Blitz CL, Murphy SP, Au DLM. Adding lignan values to a food composition database. *Journal of Food Composition and Analysis*. 2007;20:99-105.
- [14] Axelson M, Setchell KDR. The excretion of lignans in rats: evidence for an intestinal bacterial source for this new group of compounds. *FEBS Letters*. 1981;123:337-342.
- [15] Alhassane T, Xu XM. Flaxseed Lignans: Source, Biosynthesis, Metabolism, Antioxidant Activity, Bio-Active Components, and Health Benefits. *Comprehensive Reviews in Food Science and Food Safety*. 2010;9:261-269.
- [16] Zanwar AA, Hegde MV, Bodhankar SL. Chapter 71-Flax Lignan in the Prevention of Atherosclerotic Cardiovascular Diseases. *Polyphenols in Human Health & Disease*. 2014; 2:915-921.
- [17] Kreydin EI, Kim MM, Barrisford GW, Rodriguez D, Sanchez A, Santiago-Lastra, Y, Ko DS. Urinary Lignans Are Associated With Decreased Incontinence in Postmenopausal Women. *Urology*, 2015; 86: 716-720.
- [18] Chen FP, Chang CJ, Chao AS, Huang HY, Huang JP, Wu MH, Tsai CC, Kung FT, Chang CW, Tsai YC. Efficacy of Femelle for the treatment of climacteric syndrome in postmenopausal women: An open label trial. *Taiwanese Journal of Obstetrics and Gynecology*. 2016; 55: 336-340.
- [19] Renouard S, Corbin C, Lopez T, Montguillon J, Gutierrez L, Lamblin F, Lainé E, Hano C. Abscisic acid regulates pinoresinol–lariciresinol reductase gene expression and secoisolariciresinol accumulation in developing flax (*Linum usitatissimum* L.) seeds. *Planta*. 2012; 235: 85-98.
- [20] Ghose K, Selvaraj K, McCallum J, Kirby CW, Sweeney-Nixon M, Cloutier SJ, Deyholos M, Datla R, Fofana B. Identification and functional characterization of a flax UDP-glycosyltransferase glucosylating secoisolariciresinol (SECO) into secoisolariciresinol monoglucoside (SMG) and diglucoside (SDG). *BMC Plant Biology*. 2014; 14:1-17.
- [21] Fofana B, Ghose K, Somalraju A, McCallum J, Main D, Deyholos MK, Rowland GG, Cloutier S. Induced Mutagenesis in UGT74S1 Gene Leads to Stable New Flax Lines with Altered Secoisolariciresinol Diglucoside (SDG) Profiles. *Frontiers in plant science*. 2017a; 8: 1638.
- [22] Fofana B, Ghose K, McCallum J, You FM, Cloutier S. UGT74S1 is the key player in controlling secoisolariciresinol diglucoside (SDG) formation in flax. *BMC plant biology*. 2017b; 17: 35.
- [23] Vogt T. Phenylpropanoid Biosynthesis. *Molecular Plant*. 2010;3: 2-20.
- [24] Gray J, Caparrós-Ruiz D, Grotewold E. Grass phenylpropanoids: regulate before using. *Plant Science*. 2012; 184: 112-120.
- [25] Voxeur A, Wang Y, Sibout R. Lignification: different mechanisms for a versatile polymer. *Current Opinion in Plant Biology*. 2015; 23: 83-90.

- [26] Huis R, Morreel K, Fliniaux O, Lucau-Danila A, Fenart S, Grec S, Neutelings G, Chabbert B, Mesnard F, Boerjan W, et al. Natural Hypolignification Is Associated with Extensive Oligolignol Accumulation in Flax Stems. *Plant Physiology*. 2012; 158: 1893-1915.
- [27] Li J, Ma W, Zeng P, Wang J, Geng B, Yang J, Cui Q. LncTar: a tool for predicting the RNA targets of long noncoding RNAs. *Briefings in Bioinformatics*. 2015;16:806.
- [28] Ma W, Chen C, Liu Y, Li J, Xia R, Zeng M, Meyers BC. Coupling of microRNA-directed phased small interfering RNA generation from long noncoding genes with alternative splicing and alternative polyadenylation in small RNA-mediated gene silencing. *New Phytologist*. 2018;217:1535-1550.
- [29] Foissac S, Sammeth M. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Research*. 2007;35: 297-9.
- [30] Vriet C, Smith AM, Wang TL. Root Starch Reserves Are Necessary for Vigorous Re-Growth following Cutting Back in *Lotus japonicus*. *Plos One*. 2014; 9:e87333-.
- [31] Shelton D, Leach D, Baverstock P, Henry R. Isolation of genes involved in secondary metabolism from *Melaleuca alternifolia* (Cheel) using expressed sequence tags (ESTs). *Plant Science*. 2002; 162:9-15.
- [32] Tian HL, Xu XS, Zhang FS, Wang YQ, Guo SH, Qin XM, Du GH. Analysis of *Polygala tenuifolia* Transcriptome and Description of Secondary Metabolite Biosynthetic Pathways by Illumina Sequencing. *International Journal of Genomics*. 2015; 2015:1-11.
- [33] Barber MS, Mitchell HJ. Regulation of Phenylpropanoid Metabolism in Relation to Lignin Biosynthesis in Plants. *International Review of Cytology*. 1997;172: 243-293.
- [34] Ayella AK, Trick HN, Wang W. Enhancing lignan biosynthesis by over-expressing pinoreosinol lariciresinol reductase in transgenic wheat. *Molecular Nutrition & Food Research*. 2010; 51:1518-1526.
- [35] Mackenzie PI, Owens IS, Burchell B, Bock KW, Bairoch A, Belanger A, Fournel-Gigleux S, Green M, Hum DW, Iyanagi T, et al. The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence. *Pharmacogenetics & Genomics*. 1997; 7:255.
- [36] Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L, Hawkins S, Neutelings G, Datla R, et al. The genome of flax (*Linum usitatissimum* L.) assembled de novo from short shotgun sequence reads. *Plant Journal*. 2012; 72:461.
- [37] Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, Zhuang R, Lu Z, He Z, Fang X, et al. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Research*. 2010; 20:646–654.
- [38] Shen Y, Zhou Z, Wang Z, Li W, Fang C, Wu M, Ma Y, Liu T, Kong LA, Peng DL, et al. Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell*. 2014; 26: 996–1008.

- [39] Thatcher SR, Zhou WG, Leonard A, Wang BB, Beatty M, Zastrow-Hayes G, Zhao XY, Baumgarten A, Li BL. Genome-wide analysis of alternative splicing in *Zea mays*: landscape and genetic regulation. *Plant Cell*. 2014; 26:3472–3487.
- [40] Wang M, Wang P, Liang F, Ye Z, Li J, Shen C, Pei L, Wang F, Hu J, Tu L, et al. A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. *New Phytologist*. 2017; 217: 163.
- [41] Vanholme R, Storme V, Vanholme B, Sundin L, Christensen JH, Goeminne G, Halpin C, Rohde A, Morreel K, Boerjan W. A Systems Biology View of Responses to Lignin Biosynthesis Perturbations in *Arabidopsis*. *Plant Cell*. 2012; 24:3506-3529.
- [42] Le Roy J, Blervacq AS, Créach A, Huss B, Hawkins S, Neutelings G. Spatial regulation of monolignol biosynthesis and laccase genes control developmental and stress-related lignin in flax. *BMC Plant Biology*. 2017; 17:124.
- [43] Lin JS, Huang XX, Li Q, Cao YP, Bao Y, Meng XF, Li YJ, Fu CX, Hou BK. UDP-glycosyltransferase 72B1 catalyzes the glucose conjugation of monolignols and is essential for the normal cell wall lignification in *Arabidopsis thaliana*. *Plant Journal*. 2016; 88:26-42.
- [44] Xie D, Dai Z, Yang Z, Tang Q, Sun J, Yang X, Song X, Lu Y, Zhao D, Zhang LG, et al. Genomic variations and association study of agronomic traits in flax. *BMC Genomics*. 2018; 19:512.
- [45] Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei LP, Gao G. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*. 2007; 35:345-9.
- [46] Sun L, Luo H, Bu D, Zhao GG, Yu KT, Zhang CH, Liu YN, Chen RS, Zhao Y. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research*. 2013; 41:e166-e166.
- [47] Wang LG, Park HJ, Dasari S, Wang SQ, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research*. 2013; 41: e74-e74.
- [48] Abdelghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, Ben-Hur A, Reddy ASN. A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications*. 2016; 7: 11706.
- [49] Sun J, Xie DW, Zhao HW, Zou DT. Genome-wide identification of the class III aminotransferase gene family in rice and expression analysis under abiotic stress: *ATIII* gene family in japonica and indica rice, *Genes Genom*. 2013; 35:597-608.
- [50] Charlet S, Bensaddek L, Raynaud S, Gillet F, Mesnard F, Fliniaux MA. An HPLC procedure for the quantification of anhydrosecoisolariciresinol. Application to the evaluation of flax lignan content. *Plant Physiology & Biochemistry*. 2002; 40: 225-229.

Tables

Table 1 Classification of high-quality reads of insert

Category	1-2 kb		2-3 kb		3-6 kb		Total
	Value	Ratio	Value	Ratio	Value	Ratio	
No. of reads of insert	213559	100%	211111	100%	125570	100%	550240
No. of five primer transcripts	127897	59.89%	114263	54.12%	70127	55.84%	312287
No. of three primer transcripts	141481	66.25%	126590	59.96%	75398	60.04%	343469
No. of poly(A) transcripts	135010	63.22%	122846	58.19%	74496	59.33%	332353
No. of filtered short transcripts	26552	12.43%	19697	9.33%	5443	4.33%	51692
No. of nonfull-length transcripts	79816	37.37%	97921	46.38%	61863	49.26%	239600
No. of full-length transcripts	107191	50.19%	93493	44.28%	58264	46.40%	258948
No. of chimeric transcripts	106609	49.92%	93286	44.19%	58187	46.34%	258082
Average length of full-length transcripts	1301	—	2285	—	3359	—	

Table 2 Alternative splicing (AS) event statistics

AS event	Number	Percent	Gene	Percent
Mutually exclusive exon	111	0.68%	96	0.39%
Intron retention	10741	65.53%	5010	20.40%
Exon skipping	1035	6.31%	823	3.35%
Alternative 5' splice site	1397	8.52%	1082	4.40%
Alternative 3' splice site	3107	18.96%	2225	9.06%

Table 3. Correlation analysis between seed lignan content and tissue lignin content of flax plants in various stages of growth

Trait	Sample	Time	Lignan				
			0 d	5 d	10 d	20 d	30 d
Lignin	Root		-0.9932**	-0.2726	-0.2151	-0.1819	-0.3014
	Lower stem		-0.7224**	-0.6749*	-0.5888*	-0.6274*	-0.5946*
	Upper stem		-0.9909**	-0.6013*	-0.6395*	-0.7961**	-0.5864*
	Leaf		-0.6279*	0.3567	0.4753	0.5435	0.4381

Note: n = 10, *P < 0.05 indicates a significant correlation, **P < 0.01 indicates an extremely significant correlation.

Figures

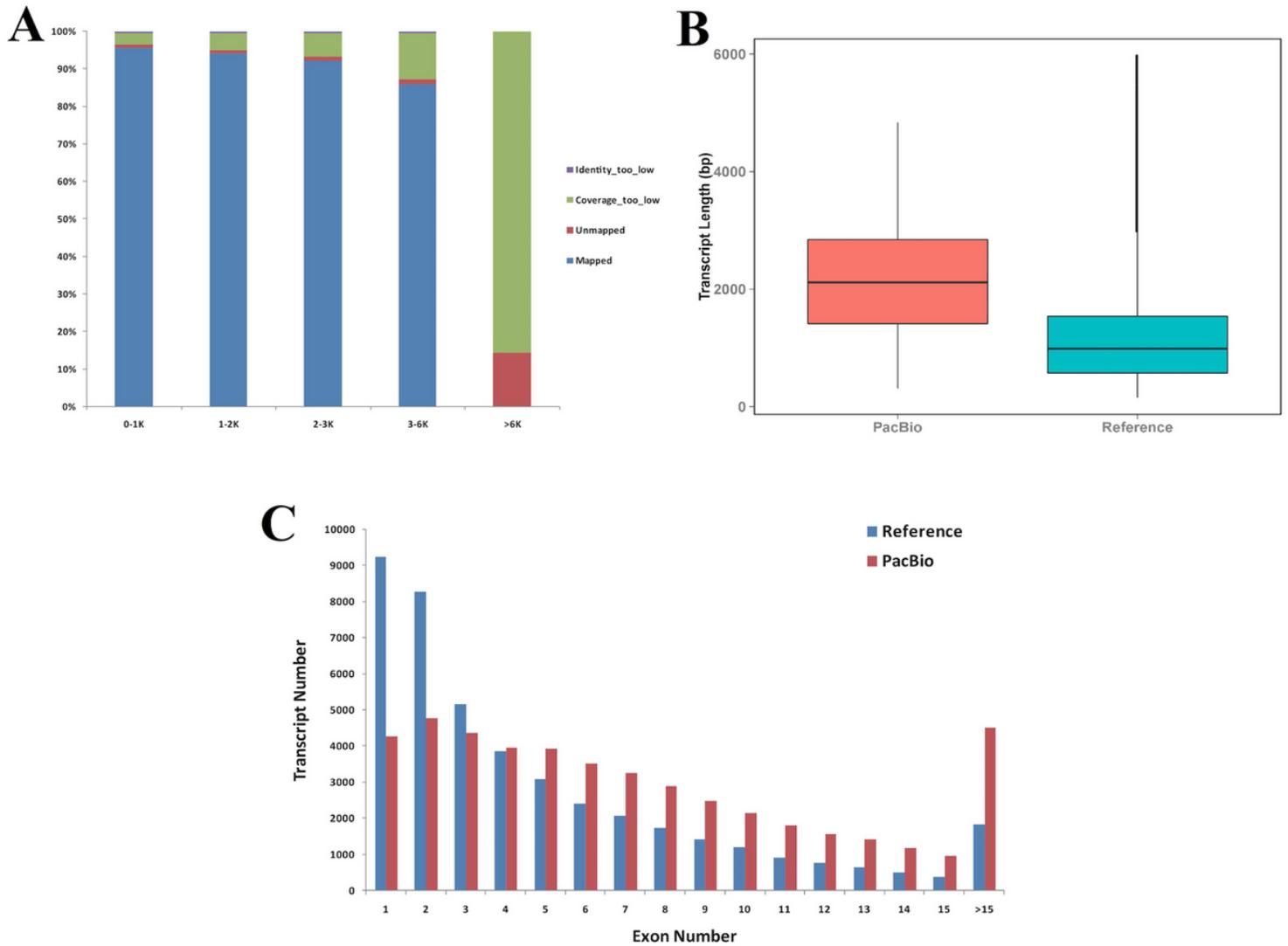


Figure 1

Characterization of flax (*Linum usitatissimum* L.) transcriptome using Iso-Seq. A: Classification of high-quality reads. B: Summary of full-length isoform mapping in 1-2 kb, 2-3 kb, and 3-6 kb libraries. This includes multiple mapping, unique mapping, and unmapped isoforms. C: Box-plots show distribution patterns of transcript length in reference annotation and PacBio Iso-Seq data ($p < 0.001$).

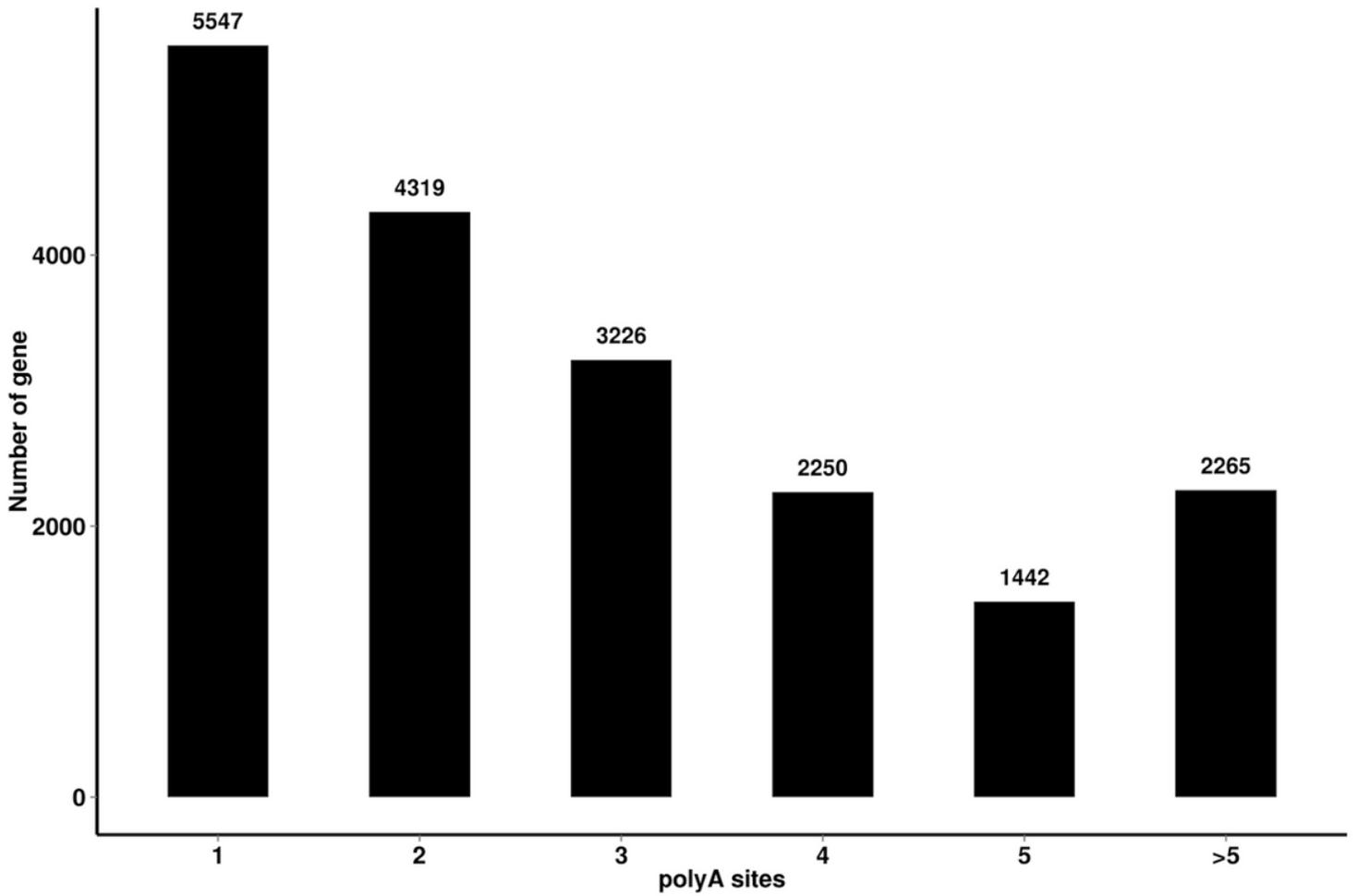


Figure 2

Distribution of the number of poly(A) sites per gene.

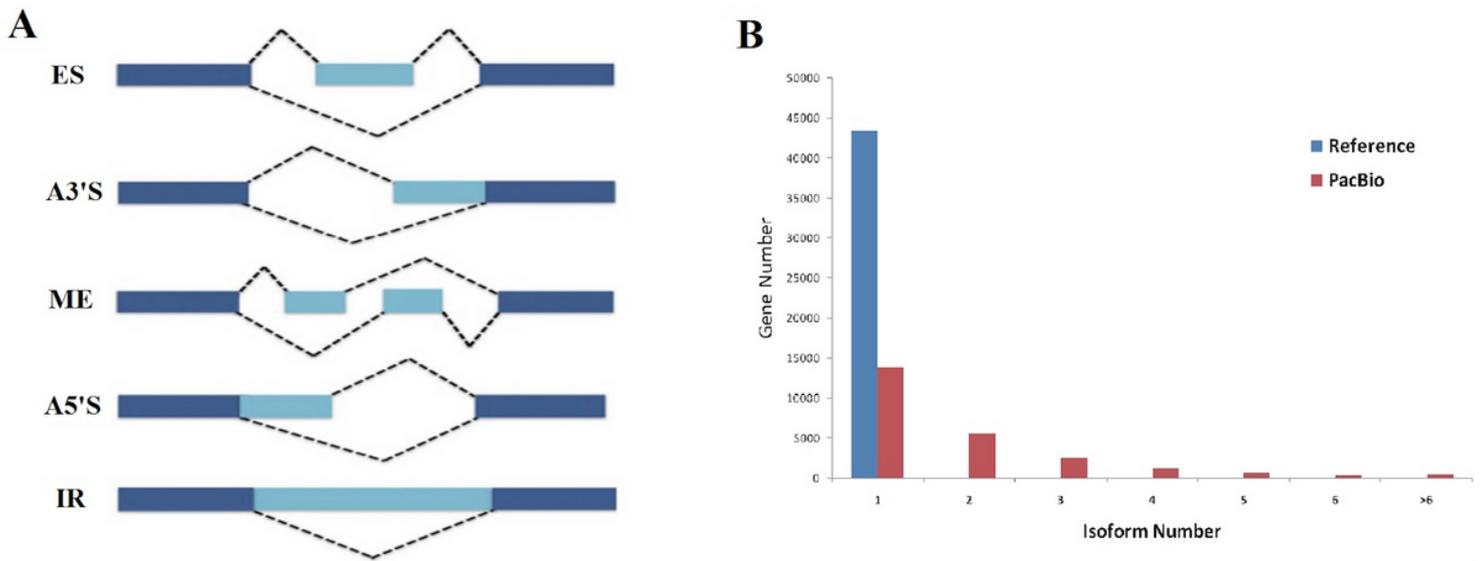


Figure 3

Classification of AS events and distribution of the number of isoforms per gene. A: Classification of AS events. images display the AS events: exon skipping (ES), alternative 3' splice site (A3'S), mutually exclusive exon (ME), alternative 5' splice site (A5'S), and intron retention (IR). B: Distribution of the number of isoforms per gene. The average number of isoforms for all genes in reference annotation and PacBio data are shown in blue and red.

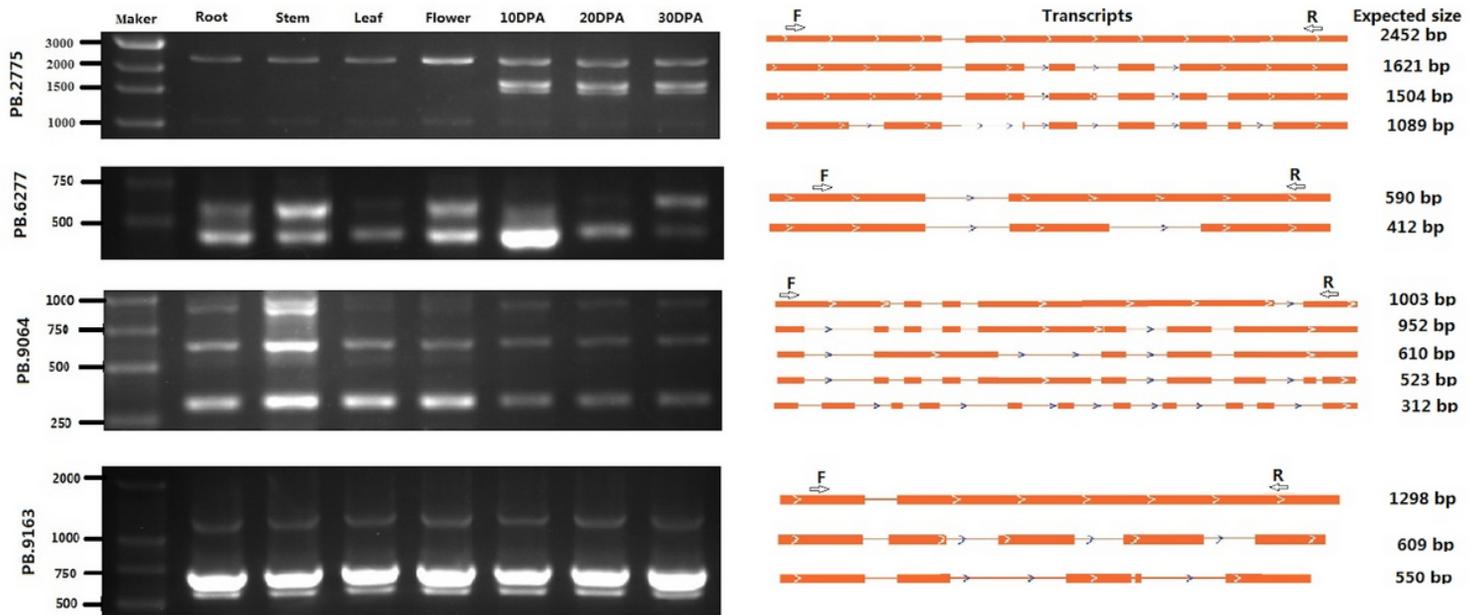


Figure 4

RT-PCR validation of AS events for four genes. Gel bands in each figure show DNA markers and PCR results in seven samples (DPA, days post-anthesis). Transcript structure of each isoform is shown in the right panel. Yellow boxes indicate exons and lines with arrows indicate introns. PCR primers (F, forward and R, reverse) are shown on the first isoform of each gene. The length of each full-length isoform is shown after the transcript structure. (Note: In order to facilitate understanding, we had cut the gel picture accordingly. Please refer to the original picture in the Fig. S5)

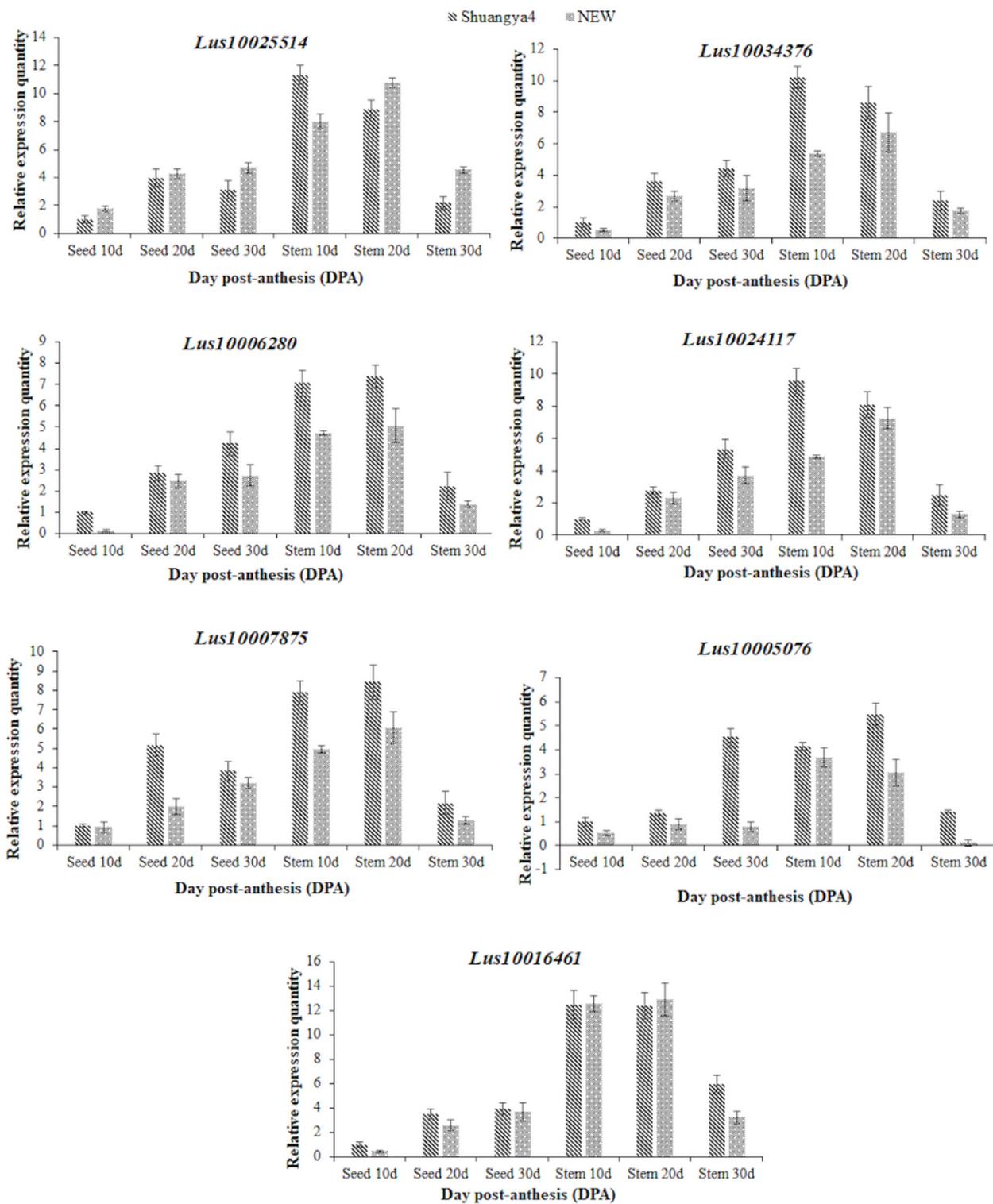


Figure 5

The expression patterns of UGTs between Shuangya 4 and NEW in each stage of flaxseed and stem development. The total RNA of the flax plants was obtained from 0-30 days after flowering, to facilitate an analysis of the expression patterns of each UGT gene in each stage of flaxseed and stem development. The quantitative PCR data are the mean of three biological replicates, whereas the bars represent the standard deviations.



Figure 6

SEM observations on the accumulation of lignin in the stem fibers of flax plants, in various stages of growth. (A-C) Stem cross-sections of the Shuangya 4 cultivar illustrate the gradual development of fiber bundles in various stages of growth, A = 10 days after flowering, B = 20 days after flowering, C = 30 days after flowering; (D-F) stem cross-sections of the NEW cultivar illustrate the gradual development of fiber bundles in various stages of growth, D = 10 days after flowering, E = 20 days after flowering, F = 30 days after flowering. The SEM images are 100× magnified with a scale of 50 μm .

Supplementary Files

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

- [Additionalfile5.docx](#)
- [Additionalfile10.xlsx](#)
- [Additionalfile9.xls](#)
- [Additionalfile11.docx](#)
- [Additionalfile12.docx](#)
- [Additionalfile13.docx](#)
- [Additionalfile6.docx](#)
- [Additionalfile7.docx](#)
- [Additionalfile8.xls](#)
- [Additionalfile2.docx](#)
- [Additionalfile3.docx](#)
- [Additionalfile4.docx](#)
- [Additionalfile1.docx](#)