Comparative transcriptome analysis during seeds development between two soybean cultivars

Li Peng [Equal first author, 1], **Linlin Qian** [Equal first author, 1], **Meinan Wang** [1], **Wei Liu** [1], **Xiangting Song** [1], **Hao Cheng** [1], **Fengjie Yuan** [2], **Man Zhao** [Corresp. 1]

[1] college of bioengineering and biotechnology, zhejiang university of technology, hang zhou, China

[2] Institute of Crop Science, Zhejiang Academy of Agricultural Sciences, hang zhou, China

Corresponding Author: Man Zhao
Email address: mzhao@zjut.edu.cn

Soybean is one of the important economic crops, which supplies a great deal of vegetable oil and proteins for human. The content of nutrients in different soybean seeds is different, which is related to the expression of multiple genes, but the mechanisms are complicated and still largely uncertain . In this study, to reveal the possible causes of the nutrients difference in soybeans A7 (containing low oil and high protein) and A35 (containing high oil and low protein), RNA-seq technology was performed to compare and identify the potential differential expressed genes (DEGs) at different seed developmental stages. The results showed that DEGs mainly presented at the early stages of seeds development and more DEGs were up-regulated at the early stage than the late stages. Gene Ontology and Kyoto Encyclopedia of Genes and Genomes analysis showed that the DEGs have diverged in A7 and A35. In A7, the DEGs were mainly involved in cell cycle and stresses, while in A35 were the fatty acids and sugar metabolism. Specifically, when the DEGs contributing to oil and protein metabolic pathways were analyzed, the differences between A7 and A35 mainly presented in fatty acids metabolism and seeds storage proteins (SSPs) synthesis. Furthermore, the enzymes, fatty acid dehydrogenase 2 (FAD2), 3-ketoacyl-CoA synthase (KCS) and 9S-lipoxygenase (LOX), in the synthesis and elongation pathways of fatty acids were revealed probably to be involved in the oil content difference between A7 and A35, while the SSPs content might be due to the transcription factors: *Leafy Cotyledon 2* (*LEC2)*, and *Abscisic acid-intensitive 3* (*ABI3*). Finally, six DEGs were selected to analyze their expression using qRT-PCR, and the results were consistent with the RNA-seq results. Generally, the study provided a comprehensive and dynamic expression trends for the seed development processes, and uncovered the potential DEGs for the differences of oil in A7 and A35.

Research article

Comparative Transcriptome Analysis during Seeds Development between Two Soybean Cultivars

Li Peng[1], Linlin Qian[1], Meinan Wang[1], Wei Liu[1], Xiangting Song[1], Hao Cheng[1], Fengjie Yuan[2], Man Zhao[1]*

1 College of Bioengineering and Biotechnology, Zhejiang University of Technology, 310014 Hangzhou, China;

2 Institute of Crop Science, Zhejiang Academy of Agricultural Sciences, 310014 Hangzhou, China

* Correspondence: mzhao@zjut.edu.cn; Tel./Fax: +86-0571-88320781

Author email list

L.P.: 15182613887@163.com

Q.L.: 1677325557@qq.com

M.W.: 2317532770@qq.com

W.L.: 1197028959@qq.com

X.S.: 292365976@qq.com

H.C.: 1124733507@qq.com

F. Y.: fjyuanhz@126.com

M.Z.: mzhao@zjut.edu.cn

Running title: Transcriptomes research of soybeans

ABSTRACT: Soybean is one of the important economic crops, which supplies a great deal of vegetable oil and proteins for human. The content of nutrients in different soybean seeds is different, which is related to the expression of multiple genes, but the mechanisms are complicated and still largely uncertain. In this study, to reveal the possible causes of the nutrients difference in soybeans A7 (containing low oil and high protein) and A35 (containing high oil and low protein), RNA-seq technology was performed to compare and identify the potential

differential expressed genes (DEGs) at different seed developmental stages. The results showed that DEGs mainly presented at the early stages of seeds development and more DEGs were up-regulated at the early stage than the late stages. Gene Ontology and Kyoto Encyclopedia of Genes and Genomes analysis showed that the DEGs have diverged in A7 and A35. In A7, the DEGs were mainly involved in cell cycle and stresses, while in A35 were the fatty acids and sugar metabolism. Specifically, when the DEGs contributing to oil and protein metabolic pathways were analyzed, the differences between A7 and A35 mainly presented in fatty acids metabolism and seeds storage proteins (SSPs) synthesis. Furthermore, the enzymes, fatty acid dehydrogenase 2 (FAD2), 3-ketoacyl-CoA synthase (KCS) and 9S-lipoxygenase (LOX), in the synthesis and elongation pathways of fatty acids were revealed probably to be involved in the oil content difference between A7 and A35, while the SSPs content might be due to the transcription factors: *Leafy Cotyledon 2* (*LEC2)*, and *Abscisic acid-intensitive 3* (*ABI3*). Finally, six DEGs were selected to analyze their expression using qRT-PCR, and the results were consistent with the RNA-seq results. Generally, the study provided a comprehensive and dynamic expression trends for the seed development processes, and uncovered the potential DEGs for the differences of oil in A7 and A35.

Subjects Plant biology, Comparative transcriptome

Keywords soybean; comparative transcriptome; differential expressed genes; oil; proteins

INTRODUCTION

Soybean (*Glycine max* L. Merr) is a staple economic crop, which provides about one-third of protein and vegetable oil for human diet (Graham and Vance, 2003). The content of oil and protein varies among soybean varieties, in which the seed oil and protein ranges from 15-25% and 35-50%, respectively (Hurburgh, 1994). Soybean oil is mainly composed of unsaturated oleic, linoleic and linolenic acid, and their proportion determines the quality of the oil. SSPs in soybean, the other important nutrient, are composed of β-conglycinin (7S) and glycinin (11S) (Harada et al., 1989; Nielsen et al., 1989; Meinke et al., 1981). In soybean breeding, one of the important objects is to improve the content of oil and protein. However, owing to their significantly negative correlation between oil and protein content, it is difficult to develop

58 soybean lines with high content of both (Hyten et al., 2004a, b; Li et al., 2018a,b; Wilcox and
59 Shibles, 2001).
60 In plants, the process of seed development is dynamic, which is regulated by both genetic and
61 environmental factors. The accumulation of nutrients, oil, carbohydrate and protein, are
62 governed by the programmed expression of a metabolic network during seed development
63 (Gupta et al., 2017; Le et al., 2007). Lots of studies have been performed to reveal the genetic
64 basis of seed development (Akond et al., 2014; Lee et al., 2007; Hwang et al., 2014; Hills, 2004).
65 For example, the studies of quantitative trait loci (QTL) and genome-wide association studies
66 (GWAS) have identified more than 100 QTLs related to the oil and protein content, which
67 widely distributed in the 20 soybean chromosomes (Hyten et al., 2004a,b; Akond et al., 2014;
68 Lee et al., 2007; Korte and Farlow, 2013; Panthee et al., 2006; Phansak et al., 2016; Wang et al.,
69 2014). In addition, along with the development in omics, the genes and metabolites that are
70 required by the seed development have also been systematically provided (Li et al., 2015). The
71 final chemical composition is a consequence of gene expression during seeds development.
72 Analyses of gene expression during seed development are providing clues to determine final
73 seed composition in different species, for example, *Arabidopsis thaliana* (Ruuska et al., 2002;
74 Palovaara et al., 2013; Fait et al., 2006), *Medicago truncatula* (Fedorova et al., 2002; Gallardo et
75 al., 2007), *Brassica napus* (Li et al., 2005), rice (Lan et al., 2005; Furutani et al., 2006), barley
76 (Watson et al., 2005), and soybean (Collakova et al., 2013). In soybeans, transcriptome of
77 different developmental stage seeds revealed that the most abundantly expressed genes
78 contributing to the metabolism and accumulation of oil and SSPs mainly presented at the middle
79 and late stages (Li et al., 2015; Libault et al., 2010; Severin et al., 2010). For example, during the
80 synthesis of SSPs, lots of amino acids-encoding genes and proline-rich proteins encoding genes
81 were highly expressed at the early seed development stages, and 7S and 11S were highly
82 expressed at the late stage, which were regulated by many transcription factors (Severin et al.,
83 2010; Verdier et al., 2008). As for lipids, there was a programmed expression of lipid
84 biosynthesis-related genes, in which FAD2-2B and FAD2-2C were highly expressed at early
85 stage and the FAD2-1A and FAD2-1B were highly expressed at later stages (O'Rourke et al.,
86 2014). Nevertheless, the genetic mechanism for the differences in seed composition has not been
87 revealed.
88 In this study, we selected two different soybean cultivars with distinct proteins and oil content

and RNA-Seq technology was applied to study three stages of seed development. Through comparative transcriptome analysis among developmental stages, we tried to reveal their dynamic expression trends, and identify the important DEGs and metabolism pathways involving in the accumulation of nutrients. Finally, this study reveals the possible divergence and accumulated mechanisms of oil and proteins in soybeans.

MATERIALS AND METHODS

Plant materials

The two cultivars A7 (Yudou 12) and A35 (Fendou 53) were planted in a farm of Fuyang (Hangzhou, China) in summer. The 2-, 4- and 6-week after flowering pods were harvested at same time point. The harvested tissues were immediately stored in liquid N2 and then stored at -80 °C for total RNA extraction. The peeled seeds were used to extract total RNA using TRIzol reagent (Invitrogen). The RNA samples were used to RNA-seq and qRT-PCR experiments.

Determination of protein and oil content of seeds

Determination of the oil contents in beans: the mature seeds were ground to powder, and the powder was transferred into 10 mL glass tubes for oil extraction. Oil was extracted by ligarine and total lipids (TL) were determined (Dong et al., 2001). The content of proteins is determined by Kjeldahl's method. Grind the samples into fit powder, and weigh 1g for digestion, cooling and distillation-titration. The percentage of bean protein was the total nitrogen percentage multiplying by 6.25. Each experiment was duplicated three times.

cDNA library cConstruction and transcriptome sequencing

The total RNA (2 μg) was sent to GENEWIZ (SuZhou, China) to sequence and analyze. The total RNA was quantified and qualified by Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), NanoDrop (Thermo Fisher Scientific Inc.). 1 μg total RNA with RIN value above 7 was used for following cDNA library preparation. First, double-strand cDNA was synthesized, and then was treated to repair both ends and add a dA-tailing, followed by a T-A ligation to add adaptors to both ends. Size selection of Adaptor-ligated DNA was then performed using AxyPrep Mag PCR Clean-up (Axygen), and fragments of ~360 bp (with the approximate insert size of 300 bp) were recovered. Each sample was then amplified by PCR for 11 cycles

using P5 and P7 primers, with both primers carrying sequences which can anneal with flow cell to perform bridge PCR and P7 primer carrying a six-base index allowing for multiplexing. The PCR products were cleaned up using AxyPrep Mag PCR Clean-up (Axygen), validated using an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), and quantified by Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA).

Then libraries with different indices were multiplexed and loaded on an Illumina HiSeq instrument according to manufacturer's instructions (Illumina, San Diego, CA, USA). Sequencing was carried out using a 2x150bp paired-end (PE) configuration. The raw data has been deposited in NCBI with accession numbers: SRR12283471 (A35-2), SRR12283470 (A35-4), SRR12283469 (A35-6), SRR12283468 (A7-2), SRR12283467 (A7-4), SRR12283466 (A7-2).

RNA-seq data analysis

The software of Trimmomatic v0.30 (http://usadellab.org/cms/index.php?page=trimmomatic) was used to remove the technical sequences for the high quality clean data (Bolger et al., 2014). As for mapping, the reference genome sequences were downloaded from Phytozome (ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/early_release/Gmax_275_Wm82.a2.v1/, version Glyma 2.0, 975 Mb), and then the clean data were aligned to reference genome using Hisat2 (v2.0.1) (http://ccb.jhu.edu/software/hisat2/manual.shtml). Differential expression analysis used the DESeq Bioconductor package, a model based on the negative binomial distribution (http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html) (Anders and Huber, 2010, 2012). After adjusted by Benjamini and Hochberg's approach for controlling the false discovery rate, *P*-value of genes were set <0.05 to detect differential expressed ones. GO-TermFinder was used identifying Gene Ontology (GO) terms that annotate a list of enriched genes with a significant p-value less than 0.05 (http://search.cpan.org/dist/GO-TermFinder/lib/GO/TermFinder.pm). Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances (http://en.wikipedia.org/wiki/KEGG). We used scripts in house to enrich significant differential expression gene in KEGG pathways.

Principal component analysis

This is another way to visualize sample-to-sample distances. In this ordination method, the data

147 points are projected onto the 2D plane such that they spread out in the two directions that explain
148 most of the differences (Figure S4). The x-axis is the direction that separates the data points the
149 most. The values of the samples in this direction are written PC1. The y-axis is a direction (it
150 must be orthogonal to the first direction) that separates the data the second most. The values of
151 the samples in this direction are written PC2. The percent of the total variance that is contained
152 in the direction is printed in the axis label. Note that these percentages do not add to 100%,
153 because there are more dimensions that contain the remaining variance (although each of these
154 remaining dimensions will explain less than the two that we see). This analysis has been
155 performed using R language.

156 **Quantitative Real Time-PCR (qRT-PCR) analysis**

157 Two micrograms of total RNA were used to synthesize the first strand cDNA using a ReverTra
158 Ace qPCR RT Kit cDNA Synthesis Kit (TOYOBO). Quantitative RT-PCR (qRT-PCR) was
159 conducted using ChamQ TM SYBR qPCR Master Mix (Vazyme) in a CFX Connect Real-Time
160 system (BIO-RAD). ACTIN (Glyma.18G290800) was used to as an internal control. The
161 specific primers were shown in Table S1. Each experiment was performed using three
162 independent biological samples. PCR was performed in a 25.0 μL reaction mixture containing 5
163 μL ChamQ TM SYBR qPCR Master Mix (Vazyme), 10 ng cDNA template, 0.4 μL of each
164 primer (10.0 μM) and 3.2 μL of double distilled H_2O (dd H_2O). The optimized operational
165 procedure was performed as follows: 2 min at 95 °C (1 cycle), 15 s at 95 °C and 60 s at 60 °C
166 (40 cycles), and then 5 s at 65 °C and 5 s at 95 °C (1 cycle for melting curve analysis). Relative
167 gene expression was evaluated as previously described (Livak and Schmittgen, 2001).

168 **Statistical tests**

169 The statistical analyses were performed in the SPSS 18.0 statistical package. Standard errors
170 were calculated and Student's two-tailed t-test was used for experimental comparisons. Pearson,
171 spearman and kendall-tau correlation coefficient were used to indicate the correlationship
172 between groups. A single asterisk (*) indicates that the difference is statistically significant at P
173 < 0.05 while double asterisks (**) indicate that the difference is statistically significant at $P <$
174 0.01.

175

RESULTS

Phenotype identification and transcriptome sequencing

In this study, 2 week after flower (WAF), 4 WAF and 6 WAF seeds in both A7 and A35 (A7-2 WAF, A7-4 WAF, A7-6 WAF, A35-2 WAF, A35-4 WAF, A35-6 WAF) were collected, respectively (Figure 1A). The total content of proteins and oil was measured in mature seeds, respectively, in which the average protein content of A7 (50.8%) is significantly higher than in A35 (38.1%), while the oil content of A7 (17.6%) is much lower than A35 (22.11%) (Figure 1B).

To value the global changes in gene expression during seeds development, the transcriptome of above collected samples were sequenced. All the sequencing data was showed in Table 1. A total of 139 and 153 million clean reads were acquired in A7 and A35, respectively. The reads were highly matched to the soybean reference genome, and the statistical results showed that 54665387, 44537617, 40763561, 39120215, 44375847 and 41099599 reads were mapped for A7-2, A7-4, A7-6, A35-2, A35-4 and A35-6, respectively, with the average matching rate 90.30%. The saturation and evenness of the transcriptomes were further analyzed to estimate the sequencing depth and evenness. According to the gene expression levels, the saturation curves were analyzed in four classifications: below 25%, 25%-50%, 50%-75% and 75%-100% (Figure S1). The expression evenness was analyzed through the whole transcripts from 5' to 3' (Figure S2). The results showed that the overall quality of sequencing in this study was high and covered the vast majority of expressed genes. In addition, the correlation analysis of different samples showed that the average correlation efficient of all the comparison pairs was more than 90% (Figure S3).

DEGs in paired comparisons of A7 and A35

In order to detect significant DEGs, the Benjamini and Hochberg's approach was adjusted, and the log2foldchange > 2 and *P*-value of genes < 0.05 were set for controlling the false discovery ratio. In total, the DEGs between different development stages and different soybeans have been systematically compared and valued. In A7, 12638 DEGs (7237 up-regulated and 5401 down-regulated) were found in A7-2-VS-A7-4; 7595 DEGs were found in A7-4-VS-A7-6 (4196 up-regulated and 3399 down-regulated); and 17035 DEGs were found in A7-2-VS-A7-6 (9736 up-regulated and 7299 down-regulated) (Figure 2A). Similarly, in A35, the number of DEGs in

A35-2-VS-A35-4 was 9364 (6134 up-regulated and 3230 down-regulated); in A35-4-VS-A35-6, was 7375 DEGs (4034 up-regulated and 3341 down-regulated); in A35-2-VS-A35-6, was 14231 DEGs (8828 up-regulated and 5403 down-regulated). Generally, the DEGs in A7 and A35 mainly occurred in the 2- and 4-WAF pair and the 2- and 6-WAF pair, and the up-regulated genes was much more than the down-regulated, especially in the early developmental stages (Figure 2). In seeds, the storage nutrients gradually increase from early stage of the seed development (2 WAF) to the late stage of seed development (6 WAF). Therefore, as evidenced by a great deal of up-regulated DEGs in 2 WAF seeds compared with the 4-WAF or 6-WAF seeds in our study, many genes have been activated for meeting the needs of energy and materials during the accumulation of storage nutrients in seeds.

Furthermore, the comparisons between A7 and A35 in the same stages were also performed. Overall, 1312 DEGs (725 ups and 587 downs) were found in both A7 and A35 (Figure 2B). However, at different developmental stages, the DEGs have diverged. Firstly, the DEGs mainly presented in 2-WAF seeds between A7 and A35, in which the number of DEGs in 2-WAF seeds (4818) was far more than in the 4- (1371) and 6- (1515) WAF seeds. In addition, the principal component analysis (PCA) results also showed that the difference of A7 and A35 mainly occurred at early stage of seed development (2 WAF seeds) (Figure S4). Secondly, in the 2-WAF seeds, the upregulated DEGs (3317) were significantly more than the downregulated (1501) in A35 compared to A7, which indicated more genes were activated in A35 than A7 during the early developmental stages of seeds.

Analysis of DEG clustering and pathway enrichment

The hierarchical clustering was analyzed and presented in a boxplot dendrogram to better understand the change trends of DEGs during the seed development (Figure S5). The dendrogram provided a clear overview for the clade structure, in which the DEGs clustered to the up-regulated or down-regulated clades, according to their expression trends. In all, 25139 DEGs were showed in the heat map, but only half of them (13262) were annotated in GO involving in biological process (5631), cellular component (1200) and molecular function (6431). Furthermore, the DEGs concentrated in the functional groups such as catalytic activity, binding, metabolic process, cellular process and biological regulation, which were consistent with the

process of seeds development accompanying multiples of the energy metabolism, nutrients accumulation and cell proliferation (Figure 3A, Figure S6).

To further find out the involved metabolic pathways of DEGs during the seeds development in A7 and A35, the KEGG pathway database was detected (Table S5). In all, in A7, 85 pathways with Q value < 0.05 were identified. 58, 46 and 33 pathways were significantly rich in A7-2-VS-A7-4, A7-2-VS-A7-6 and A7-4-VS-A7-6 paired comparisons, respectively. In A35, 78 pathways were totally identified: 46 pathways were in A35-2-VS-A35-4, 37 in A35-2-VS-A35-6 and 28 in A35-4-VS-A35-6, respectively. Notably, eight and four significantly rich pathways were common to the three paired comparisons in A7 and A35, respectively (Figure 3B). In A7, the common pathways included ko05203 (Viral carcinogenesis), ko01524 (Platinum drug resistance), ko00073 (Cutin, suberine and wax biosynthesis), ko05418 (Fluid shear stress and atherosclerosis), ko00480 (Glutathione metabolism), ko05034 (Alcoholism), ko04110 (Cell cycle), and ko04914 (Progesterone-mediated oocyte maturation), which were mainly involved in cell cycle and stresses responses. However, in A35, the common pathways were ko00052 (Galactose metabolism), ko00073 (Cutin, suberine and wax biosynthesis), ko00591 (Linoleic acid metabolism) and ko00500 (Starch and sucrose metabolism), mainly involving in sugar and fatty acids metabolism. Generally, only one common pathway of ko00073 occurred in both A7 and A35, which were related to the biosynthesis of cutin, suberine and wax in cell wall (Figure 3). Meanwhile, when specific development stage of A7 was compared with A35, lipid related pathways (ko00592, ko00062, ko00071) and sugar related pathways (ko00196, ko00195, ko00710, ko00010) were found at the 2-WAF stage, while amino acids metabolic pathways (ko00360, ko00270) were found at 4- and 6-WAF stages, respectively (Table S5). The divergence of the pathways in different developmental stages of soybeans indicated their difference of nutrients content between A7 and A35.

Exploration of DEGs that may contribute to the oil and SSPs content in soybeans

Given the significant difference of oil and proteins content in A7 and A35, the DEGs involved in the synthetic and metabolic pathways of lipid, fatty acids, amino acids and proteins during the seed development were specifically investigated (Figure 4). First, the change trends of DEGs (up- and down-regulated DEGs) numbers involved in the lipid metabolism were V-shaped in A7 and A35 from 2-VS-4, 4-VS-6 to 2-VS-6. And correlation analysis showed that the trends were

consistent between A7 and A35 (Figure 4A). Similarly, the change trend of DEGs related to the amino acids metabolism was also V-shaped in A7 and A35 (Figure 4D). Nevertheless, the change trends of DEGs involving fatty acids and proteins metabolism were different in A7 and A35. As for the fatty acids, the numbers of up- and down-regulated DEGs in A7 were similar in 2-VS-4 and 4-VS-6 pairs, but sharply decreased in 2-VS-6. However, in A35 the number of down-regulated DEGs was increased linearly from 2-VS-4, 4-VS-6 to 2-VS-6, while the trend of the up-regulated DEGs was V-shaped (Figure 4B). Notably, the trends of DEGs involving proteins between A7 and A35 were oppositely. In A7, the up- and down-regulated DEGs were first increased and then decreased, while in A35 it was opposite (Figure 4C). Notably, the divergent trends of DEG numbers involving fatty acids and protein synthesis were basically consistent with their difference of oil and protein content in A7 and A35, which indicated that these DEGs probably contributed to their content differences in A7 and A35.

To further reveal DEGs involved in oil and proteins content difference in A7 and A35, the significant DEGs were matched to the QTL database in SoyBase (https://www.soybase.org/). Finally, the Glyma.06G214800 and Glyma.07G034900 were found involving in oil and linoleic acid content. Glyma.06G214800 and Glyma.07G034900 encoded 3-ketoacyl-CoA synthase (KCS) and linoleate 9S-lipoxygenase (LOX1-5), which catalyzing the elongation of C18 fatty acids, and forming a peroxide by adding oxygen in the double bond of linoleic acid, respectively. In our study, the expression levels of Glyma.06G214800 and Glyma.07G034900 were down-regulated along the development of soybean seeds in A35. However, the expression of Glyma.06G214800 was not detected in A7, and the expression levels of Glyma.07G034900 in A7 were significantly higher than in A35, which probably is related to the difference of oil content between A7 and A35 (Figure 5A).

Unfortunately, no specific DEGs involving protein content were matched to the QTL database. However, previous studies have revealed that the expression of seeds storage proteins (SSPs) is mainly controlled by transcription factors during the seed filling stages, the master regulators contain LEC1, LEC2, ABI3, and FUSCA3 (FUS3) (reviewed by Fedorova et al., 2002). In this study, the expression of *LEC1*s, *LEC2*s, *FUS3*s and *ABI3*s showed a little diverged, in which the expression trends of *LEC1*s (Glyma.07G268100 and Glyma.17G005600), *ABI3* (Glyma.02G099500, Glyma.01G087500, and Glyma.10G204400) were basically consistent in

A7 and A35, while *LEC2*s (Glyma.20G035800 and Glyma.20G035700) and *ABI3* (Glyma.20G186200) were diverged (Figure 5B). Besides, most of consistently expressed TFs were basically higher in A35 than A7 except for *ABI3* (Glyma.02G099500 and Glyma.01G087500) (Figure 5B). Considering the positive relationship of TFs expression and content of SSPs, the two *ABI3*s (Glyma.02G099500 and Glyma.01G087500) and *LEC2*s (Glyma.20G035800 and Glyma.20G035700) might be correlated with the difference of SSPs between A7 and A35.

Expression flux in the Fatty acids synthesis pathway

The metabolic pathway of fatty acids mainly involved in fatty acid biosynthesis (ko00061), fatty acid elongation (ko00062) and fatty acid degradation (ko00071) (Figure 6, Figure S7). All the DEGs and their expression trends in these pathways have been marked in Figure 6 to identify their expression flux. Generally, the expression levels of fatty acid dehydrogenase 2 (FAD2s) were the highest in the whole pathway, in which Glyma.10G278000 and Glyma.20G111000 were specifically expressed in seeds and the expression levels were increased along the development of seeds. In addition, the expression levels of *FAD2*s were higher in A35-2WAF seeds than in A7-2WAF, which suggested that *FAD2* genes might be essential to the content of oil in A7 and A35. Notably, the expression levels of key enzymes of fatty acid elongation pathway were drastically reduced with the development of seeds in A7 and A35 (Figure 6). Moreover, their expression levels in A35 were lower than in A7. The results might interpret why the fatty acids in soybean seeds were mainly present in the form of C18. As for fatty acid degradation, the expression of enoyl-CoA hydratase/3-hydroxyacyl-CoA dehydrogenase (MFP2/HAD) and acetyl-CoA C-acetyltransferase (atoB) were stable in the development of seeds in A7 and A35 (Figure 6).

The qRT-PCR analysis of DEGs in soybeans

To confirm the RNA-seq results in our research, we selected 6 significantly DEGs related to the fatty acids metabolism and the synthesis of amino acids for qRT-PCR analysis (Figure 7, Table S6). In all, the expression results of qRT-PCR were consistently with the RNA-seq results. Among them, the expression of Glyma.13G035200 was significantly increased along the seed development, while Glyma.16G147300 was gradually decreased. The expression trend of Glyma.06G211300 was decreased from 2 WAF to 4 WAF seeds, and then was increased in 6

WAF seeds. Oppositely, as for Glyma.17G047000, Glyma.17G027600 and Glyma.06G183900, their expression trends were bell-shaped, increasing from 2-WAF to 4-WAF seeds and then decreasing in 6-WAF seeds. The results verified our RNA-seq data was reliable.

DISCUSSION

In soybean, there is not only different nutrients content, but also a negative correlation of oil and protein content in seeds (Chuang et al., 2003). The two soybean cultivars A7 and A35, with similar genetically background, showed distinctive difference of storage proteins and oil content, which support natural materials for comparative studies of the regulation of seed nutrients accumulation. In our study, the transcriptome data of three stages in seed development, 2, 4, 6 WAF seeds, was compared to find their dynamic expression. Generally, similarly with the previous studies (Bao and Ohlrogge, 1999), there were more DEGs at the early stage of seeds development; and the up-regulated DEGs were more than the down-regulated ones along the seed development in both soybeans (Figure 2A). The accumulation of nutrients, especially oils and SSPs, were essential to seed development. The accumulation of nutrients was complicated and dynamic, which were influenced by multiple genetic and environmental factors (Gupta et al., 2017; Hills, 2006). Until now, studies including genetics, omics, QTL and GWAS have been performed to investigate the mechanisms of seed filling (Hyten et al., 2004a,b; Li et al., 2018a,b; Gupta et al., 2017; Agrawal et al., 2008). It has been revealed that the content of oil was gradually increased until 40 day after flowers (DAF), then began to keep steady in soybeans (Li et al., 2015, Collakova et al., 2013), which is consistent with our results. In our study, the comparison of transcriptome in different developmental stages of seeds and different soybean cultivars revealed that the change trends of DEGs involved in the synthesis process of fatty acids were basically consistent in A7 and A35, and the synthesis of fatty acids has begun to be down-regulated in 6 WAF seeds. It has been reported that the supply of fatty acids was the limiting factor to the accumulation of oil yield in embryos (Bao and Ohlrogge, 1999). Fatty acids, especially linoleic acid, are the main composition of soybean oil. FAD2 is essential to convert oleic acid into linoleic acid. Notably, the highest expressed *FAD2* genes in the fatty acid pathways were increased in all periods, and in A35 it was obviously higher than in A7, which indicated that the oil difference in A35 and A7 might be attributed to the expression difference of

354 *FAD2* genes. In our previous study, the bioactivity of FAD2 was also revealed to have
355 correlation with oil content in different plants (Zhao et al., 2019). Therefore, the oil content of
356 seeds might be controlled by both the expression and bioactivity of FAD2. Furthermore, the
357 DEGs in the fatty acids degradation pathway were stable between A7 and A35, and the result
358 showed that degradation pathway was not the main reason of divergence in the two cultivars.
359 Furthermore, we analyzed the fatty acid elongation pathway and found the expression trends
360 were drastically reduced with the seed development in A7 and A35 (Figure 5,6). Moreover, the
361 expression levels of KCS in A35 were lower than in A7. The results indicated that more fatty
362 acids were present in the form of C18 in the late development of seeds in A35, which was
363 consistent with the composition of soybean oil. Therefore, based on the above results, the
364 differences in both biosynthesis and elongation of fatty acids during the seed development
365 contributed to the difference of oil content in A35 and A7. Obviously, more experiments need to
366 be done for more details.
367 In addition, the content of proteins was also significantly different in A7 and A35. The change
368 trends of DEGs involved in the expression of proteins were different, especially in 4-VS-6 of A7
369 and A35 (Figure 4C), which suggested the content differences between A7 and A35 might be
370 due to the expression levels of the genes encoding proteins between 4 WAF seeds and 6 WAF
371 seeds. Unfortunately, in our study, we didn't find the candidate DEGs which matched into the
372 QTL database. The possible reason might be that proteins content is a quality trait, and which is
373 influenced by many different genes. Transcription factors (TF) are essential to regulate the gene
374 expression. Previous studies have revealed that the expression of SSPs is mainly controlled by
375 TFs, *LEC1*, *LEC2*, *ABI3*, and FUSCA3 (*FUS3*), during the seed filling stages (reviewed by
376 Verdier and Thompson, 2008). These TFs interacted in a network, in which LEC1 induces the
377 expression of *LEC2*, *ABI3* and *FUS3*. Moreover, *LEC1* and *LEC2* genes regulate each other, and
378 activate the expression of *FUS3* and *ABI3*, and then activate the *SSP* genes (Fedorova et al., 2002;
379 Kagaya et al., 2005; Santos Mendoza et al., 2005; Keith et al., 1994; Parcy et al., 1997; To et al.,
380 2006; Braybrook et al. 2006). In this study, the expression of *LEC1*s, *LEC2*s, *FUS3*s and *ABI3*s
381 showed a little diverged, in which the expression trends of *LEC1*s (Glyma.07G268100 and
382 Glyma.17G005600), *ABI3* (Glyma.02G099500, Glyma.01G087500, and Glyma.10G204400)
383 were basically consistent in A7 and A35, while *LEC2*s (Glyma.20G035800 and
384 Glyma.20G035700) and *ABI3* (Glyma.20G186200) were diverged (Figure 5). Besides, most of

consistently expressed TFs were basically higher in A35 than A7 except for *ABI3* (Glyma.02G099500 and Glyma.01G087500) (Figure 5). Considering the positive relationship of TFs expression and content of SSPs, we predicted that the two *ABI3*s (Glyma.02G099500 and Glyma.01G087500) and *LEC2*s (Glyma.20G035800 and Glyma.20G035700) might be correlated with the difference of SSPs between A7 and A35. However, more cultivars and samples should be further studied for the solid proofs.

It has been reported that the synthesis of amino acids, such as methionine and asparagine, is also related to the accumulation of proteins (Li et al., 2015; Galili et al., 2016; Molvig et al., 1997; Zhao et al., 2018). Furthermore, it has been reported that the asparagine was the major form of nitrogen importing from the vegetative organs to the seed development. And the asparaginase enzyme (Glyma.05G018300) was predicted to interconvert amino acids for protein synthesis during seed filling (Li et al., 2015). In our study, the expression of Glyma.05G018300 was also analyzed, but significantly differences were not found in any comparison pairs, which indicated the gene might be not involved in the difference in this study (Figure S8).

CONCLUSIONS

In this study, the comparative transcriptome analysis was performed to research the dynamic expression of different developmental stages of soybean seeds in A7 and A35. The results showed that more DEGs occurred in early stage of seed development, especially at 2 WAF stage, and there were more up-regulated DEGs at the early stages compared with the late stages. As for the DEGs involving in nutrients metabolism such as oil and SSPs, their change trends were different. The DEGs of the synthesis and elongation of fatty acids , *FAD2*, *KCS* and *LOX*, have been revealed might contribute to the oil content difference in A7 and A35, while SSPs might be due to the transcription regulators such as *ABI3* and *LEC2*.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by Zhejiang Provincial Major Agriculture Science and Technology Special Sub-project (Grant No. 2016C02050-10-3).

Competing Interests

The authors declare no competing financial interest.

Authors' contributions

M.Z. and L.P. conceived and designed the analyses. L. Q., M. W. and H.C. performed most of the experiments and carried out most of the analysis. X. S. and W. L. performed RT-PCR expression analyses. M. Z. and L.P. analyzed and interpreted the data. F. Y. and Z.W. coordinated the work. M.Z. drafted the manuscript. All authors have read and approved the final manuscript.

SUPPLEMENTARY MATERIALS

Figure S1. The saturation curves of all the samples in A7 and A35.

Figure S2. The evenness curves of samples in A7 and A35.

Figure S3. The correlation analysis of samples in A7 and A35.

Figure S4. The principal component analysis (PCA) analysis of samples in A7 and A35.

Figure S5. Clustering analysis of the DEGs. The tree was constructed with log10 (RPKM+1). Blue, white and red indicate high, intermediate and expression, respectively. The clustered genes were shown in the different color such as green, light blue and purple. 2 WAF, 4 WAF and 6 WAF seeds were labeled with A7-2,-4,-6 and A35-2, -4, -6. Each sample has three repetitions labeled a, b and c.

Figure S6. GO analysis of the DEGs. A1-A4: the enriched Go term in A35-2-VS-A7-2, A35-4-VS-A7-4, A35-6-VS-A7-6 and A35-VS-A7. B: the DEGs numbers in different comparative pairs. The functions related to molecular function, cellular component and biological process were shown in red, green and blue, respectively.

Figure S7. The fatty acids and amino acids metabolism pathway in KEGG database.

Figure S8. The expression trends of asparaginase enzyme family in different comparison pairs. The Neighbor joining tree was constructed with amino acids sequences. The red, green and gray boxes represented their expression were down-regulated, up-regulated and not changed, respectively.

Table S1. The primers of DEGs used in qRT-PCR.

Table S2. The overview of DEGs in A7-2-VS-A7-4, A7-2-VS-A7-6 and A7-4-VS-A7-6.

Table S3. The overview of DEGs in A35-2-VS-A35-4, A35-2-VS-A35-6 and A35-4-VS-A35-6

444 **Table S4.** The overview of DEGs in A35-2-VS-A7-2, A35-4-VS-A7-4, A35-6-VS-A7-6 and
445 A35-VS-A7.
446 **Table S5.** The significant KEGG pathways in different compared pairs.
447 **Table S6.** The expression levels of qRT-PCR, and the protein and oil content in A7 and A35.
448 **REFERENCES**
449 **Graham PH, Vance CP**. 2003. Legumes: importance and constraints to greater use. *Plant*
450 *Physiology* **131:**872-877.
451 **Hurburgh CR. 1994.** Long-term soybean composition patterns and their effect on processing.
452 *Journal of the American Oil Chemists Society* **71:**1425-1427.
453 **Harada JJ, Barker SJ, Goldberg RB. 1989.** Soybean beta-conglycinin genes are clustered in
454 several DNA regions and are regulated by transcriptional and posttranscriptional processes. *the*
455 *Plant Cell* **1:**415-425.
456 **Nielsen NC, Dickinson CD, Cho TJ, et al. 1989.** Characterization of the glycinin gene family in
457 soybean. *the Plant Cell* **1:**313-328.
458 **Meinke D, Chen J, Beachy R. 1981.** Expression of storage-protein genes during soybean seed
459 development. *Planta* **153:**130-139.
460 **Hyten DL, Pantalone VR, Sams CE, Saxton AM, Landau-Ellis D, Stefaniak TR, Schmidt**
461 **ME. 2004a.** Seed quality QTL in a prominent soybean population. *Theoretical and Applied*
462 *Genetics* **109:**552-561.
463 **Hyten DL, Pantalone VR, Saxton AM, Schmidt ME, Sams CE. 2004b.** Molecular mapping
464 and identification of soybean fatty acid modifier quantitative trait loci. *Journal of the American*
465 *Oil Chemists Society* **12:**1115-1118.
466 **Li D, Zhao X, Han Y, Li W, Xie F. 2018a.** Genome-wide association mapping for seed protein
467 and oil contents using a large panel of soybean accessions. *Genomics* **111(1):**90-95.
468 **Li Y, Reif JC, Hong H, Li H, Liu Z, Ma Y, Li J, et al. 2018b**. Genome-wide association
469 mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean
470 accessions. *Plant Science* **266:**95-101.
471 **Wilcox JR, Shibles RM. 2001.** Interrelationships among seed quality attributes in soybean.
472 *Crop Science* **41:**11-14.
473 **Gupta M, Bhaskar PB, Sriram S, Wang PH. 2017.** Integration of omics approaches to
474 understand oil/protein content during seed development in oilseed crops. *Plant Cell Reports*

36:637-652.

Le BH, Wagmaister JA, Kawashima T, Bui AQ, Harada JJ, Goldberg RB. 2007. Using genomics to study legume seed development. *Plant Physiology* **144:**562-574.

Akond M, Liu S, Boney M, Kantartzi SK, Meksem K, Bellaloui N, Lightfoot DA, Kassem MA. 2014. Identification of quantitative trait loci (QTL) underlying protein, oil, and five major fatty acids' contents in soybean. *American Journal of Plant Sciences* **5:**158-167.

Lee JD, Bilycu KD, Shannon JG. 2007. Genetics and breeding for modified fatty acid profile in soybean seed oil. *Journal of Crop Science and Biotechnology* **10:**201-210.

Hwang EY, Song Q, Jia G, Specht JE, Hyten DL, Costa J, Cregan PB. 2014. A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* **15:**1.

Hills MJ. 2004. Control of storage-product synthesis in seeds. *Current Opinion in Plant Biology* **7:**302-308.

Korte A, Farlow A. 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9:**29.

Panthee DR, Pantalone VR, Saxton AM. 2006. Modifier QTL for fatty acid composition in soybean oil. *Euphytica* **152:**67-73.

Phansak P, Soonsuwon W, Hyten DL, Song Q, Cregan PB, Graef GL, Specht JE. 2016. Multi-population selective genotyping to identify soybean (*Glycine max* (L.) Merr.) seed protein and oil QTLs. *G3-Genes Genomes Genetics* **6(6):**1635-1648.

Wang X, Jiang GL, Green M, Scott RA, Hyten DL. 2014. Quantitative trait locus analysis of unsaturated fatty acids in a recombinant inbred population of soybean. *Molecular Breeding* **33:**281-296.

Li L, Hur M, Lee JY, et al. 2015. A systems biology approach toward understanding seed composition in soybean. *BMC Genomics* **16(3):**S9.

Ruuska SA, Girke T, Benning C, Ohlrogge JB. 2002. Contrapuntal networks of gene expression during *Arabidopsis* seed filling. *the Plant Cell* **14(6):**1191-1206.

Palovaara J, Saiga S, Weijers D. 2013. Transcriptomics approaches in the early *Arabidopsis* embryo. *Trends in Plant Science* **18(9):**514-521.

Fait A, Angelovici R, Less H, Ohad I, Urbanczyk-Wochniak E, Fernie AR, Galili G. 2006. *Arabidopsis* seed development and germination is associated with temporally distinct metabolic switches. *Plant physiology* **142(3):**839-854.

Fedorova M, van de Mortel J, Matsumoto PA, Cho J, Town CD, VandenBosch KA, Gantt JS, Vance CP. 2002. Genome-wide identification of nodule-specific transcripts in the model legume *Medicago truncatula*. *Plant Physiology* **130(2):**519-537.

Gallardo K, Firnhaber C, Zuber H, Hericher D, Belghazi M, Henry C, Kuster H, Thompson R. 2007. A combined proteome and transcriptome analysis of developing *Medicago truncatula* seeds: evidence for metabolic specialization of maternal and filial tissues. *Molecular & cellular proteomics* **6(12):**2165-2179.

Li F, Wu X, Tsang E, Cutler AJ. 2005. Transcriptional profiling of imbibed *Brassica napus* seed. *Genomics* **86(6):**718-730.

Lan L, Chen W, Lai Y, Suo J, Kong Z, Li C, Lu Y, Zhang Y, Zhao X, Zhang X. et al. 2004. Monitoring of gene expression profiles and isolation of candidate genes involved in pollination and fertilization in rice (*Oryza sativa* L.) with a 10K cDNA microarray. *Plant Molecular Biology* **54(4):**471-487.

Furutani I, Sukegawa S, Kyozuka J. 2006. Genome-wide analysis of spatial and temporal gene expression in rice panicle development. *Plant Journal* **46(3):**503-511.

Watson L, Henry RJ. 2005. Microarray analysis of gene expression in germinating barley embryos (*Hordeum vulgare* L.). *Functional & Integrative Genomics* **5(3):**155-162.

Collakova E, Aghamirzaie D, Fang Y, Klumas C, Tabataba F, Kakumanu A, Myers E, Heath LS, Grene R. 2013. Metabolic and transcriptional reprogramming in developing soybean (Glycine max) embryos. *Metabolites* **3(2):**347-372.

Libault M, Farmer A, Joshi T, et al. 2010. An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. *Plant Journal* **63:**86-99.

Severin AJ, Woody JL, Bolon YT, et al. 2010. RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. *BMC Plant Biology* **10:**160.

Verdier J, Thompson RD. 2008. Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. *Plant Cell Physiology* **49(9):**1263-1271.

O'Rourke JA, Bolon YT, Bucciarelli B, Vance CP. 2014. Legume genomics: understanding biology through DNA and RNA sequencing. *Annals of Botany* **113:**1107-1120.

Chung J, Babka HL, Graef GL, Staswick PE, Lee DJ, Cregan PB, Shoemaker RC, Specht JE. 2003. The seed protein, oil, and yield QTL on soybean linkage group I. *Crop Science* **43(3):**1053-1067.

Agrawal GK, Hajduch M, Graham K, Thelen JJ. 2008. In-depth investigation of the soybean seed-filling proteome and comparison with a parallel study of rapeseed. *Plant Physiology* **148:**504-518.

Bao X, Ohlrogge J. 1999. Supply of fatty acid is one limiting factor in the accumulation of triacylglycerol in developing embryos. *Plant Physiology* **120 (4):**1057-1062.

Kagaya Y, Toyoshima R, Okuda R, Usui H, Yamamoto A, Hattori T. 2005. LEAFY COTYLEDON1 controls seed storage protein genes through its regulation of FUSCA3 and ABSCISIC ACID INSENSITIVE3. *Plant and Cell Physiology* **46:**399-406.

Santos Mendoza M, Dubreucq B, Miquel M, Caboche M, Lepiniec L. 2005. LEAFY COTYLEDON 2 activation is sufficient to trigger the accumulation of oil and seed specific mRNAs in *Arabidopsis* leaves. *FEBS Letters* **579:**4666-4670.

Keith K, Kraml M, Dengler NG, McCourt P. 1994. fusca3: a heterochronic mutation affecting late embryo development in *Arabidopsis*. *the Plant Cell* **6:**589-600.

Parcy F, Valon C, Kohara A, Mise´ ra S, Giraudat J. 1997. The ABSCISIC ACID-INSENSITIVE3, FUSCA3, and LEAFY COTYLEDON1 loci act in concert to control multiple aspects of *Arabidopsis* seed development. *the Plant Cell*. **9:**1265-1277.

To A, Valon C, Savino G, Guilleminot J, Devic M, Giraudat J, Parcy F. 2006. A network of local and redundant gene regulation governs *Arabidopsis* seed maturation. *the Plant Cell.* **18:**1642-1651.

Braybrook SA, Stone SL, Park S, Bui AQ, Le BH, Fischer RL, Goldberg RB, Harada JJ. 2006. Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis. *Proceedings of the National Academy of Sciences USA* **103:**3468-3473.

Galili G, Amir R, Fernie AR. 2016. The regulation of essential amino acid synthesis and accumulation in plants. *Annual Review of Plant Biology* **67:**153-178.

Molvig L, Tabe LM, Eggum BO, Moore AE, Craig S, Spencer D, et al. 1997. Enhanced methionine levels and increased nutritive value of seeds of transgenic lupins (*Lupinus angustifolius* L.) expressing a sunflower seed albumin gene. *Proceedings of the National Academy of Sciences USA* **94:**8393-8398.

Zhao M, Chen P, Wang W, Yuan F, Zhu D, Wang Z, Ying X. 2018. Molecular evolution and expression divergence of *HMT* gene family in plants. *International Journal of Molecular*

Sciences **19:**1248.

Dong XL, Bai PL, Wang JM, Ruan CJ. 2011. Comparative study on determination of seed oil content of energy plants by using NMR and Soxhlet Extraction. *RENEW ENERG.* **29(3):**21-24.

Bolger AM, Lohse M, Usadel B. 2014. Trimmer for Illumina sequence data. *Bioinformatics* btu170.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biology* **11:**R106.

Anders S, Huber W. 2012. Differential expression of RNA-Seq data at the gene level-the DESeq package. *EMBL*.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCt method. *Methods*. **25:**402-408.

Zhao M, Wang W, Wei L, Chen P, Peng L, Qin Z, Yuan F, Wang Z, Ying X. 2019. The evolution and biocatalysis of FAD2 indicate its correlation to the content of seed oil in plants. *International Journal of Molecular Sciences* **20(4)**:849.

Figure 1

Figure 1. Seed phenotypes, and the protein and oil content in A7 and A35

A. Phenotypes of 2-WAF, 4-WAF and 6-WAF seeds. B. The total protein and oil content in the mature seeds of A7 and A35. Error bar: standard deviation. The significance was tested in comparison with the contents of oil and protein in A7 (blue columns). The * and ** represent the significance at a $P < 0.05$ and $P < 0.01$ levels, respectively.

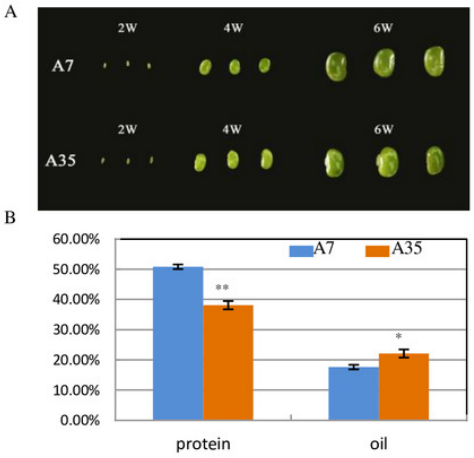

A
2W
4W
6W
A7
2W
4W
6W
A35
B
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
A7
A35
**
*
protein
oil

Figure 2

Figure 2. The gene numbers of significantly different expression among samples.

The x-axis represented the comparison pairs. The y-axis represented the number of DEGs. A. The comparison between 2-, 4-, 6-WAF in cultivars. B. The comparison between A7 and A35. The red and blue lines represent up-regulated and down-regulated DEGs, respectively.

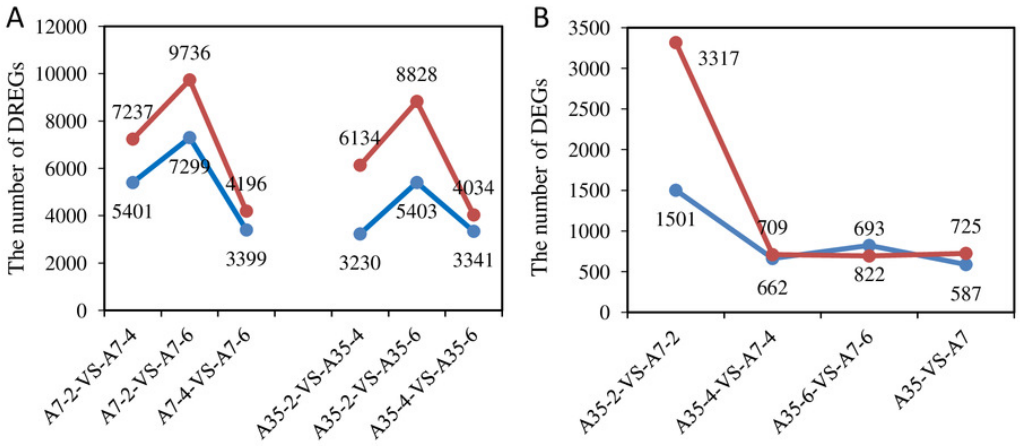

A
The number of DREGs
12000
10000
8000
6000
4000
2000
0
9736
7237
7299
4196
5401
3399
8828
6134
5403
4034
3230
3341
A7-2-VS-A7-4
A7-2-VS-A7-6
A7-4-VS-A7-6
A35-2-VS-A35-4
A35-2-VS-A35-6
A35-4-VS-A35-6
B
The number of DEGs
3500
3000
2500
2000
1500
1000
500
0
3317
1501
709
693
725
662
822
587
A35-2-VS-A7-2
A35-4-VS-A7-4
A35-6-VS-A7-6
A35-VS-A7

Table 1(on next page)

table 1 Summary of soybeans seeds transcriptome data sequenced by the Illumina platform

Table 1. Summary of soybeans seeds transcriptome data sequenced by the Illumina platform

Sample	A35-2WAF	A35-4WAF	A35-6WAF	A7-2WAF	A7-4WAF	A7-6WAF	Sum/Ave
Raw reads	45,623,515	48,593,831	46,469,858	60,380,729	49,145,343	44,862,351	295,075,627
Raw bases	6,843,527,200	7,289,074,700	6,970,478,700	9,057,109,400	7,371,801,500	6,729,352,700	44,261,344,200
Q20(%)	95.97	96.89	96.77	96.25	96.13	96.46	96.41
Q30(%)	90.78	92.57	92.32	91.06	90.98	91.70	91.57
Clean reads	45,229,794	48,357,101	46,210,171	59,926,314	48,762,965	44,562,063	293,048,408
Clean bases	6,703,514,302	7,186,695,710	6,868,762,709	8,891,051,948	7,243,899,179	6,621,751,911	43,515,675,759
Q20(%)	96.42	97.20	97.13	96.78	96.59	96.83	96.83
Q30(%)	91.40	92.99	92.81	91.80	91.62	92.20	92.14
Mapped reads	39120215	44375847	41099599	54665387	44537617	40763561	264,562,226
Proportion (%)	86.64	91.56	89.61	91.17	91.30	91.50	90.30

Figure 3

Figure 3. GO analysis and Pathways enrichment of DEGs in three paired comparisons in A7 and A35.

The functions related to molecular function, cellular component and biological process were shown in red, green and blue, respectively. The numbers of significantly rich pathways were marked on the map, respectively.

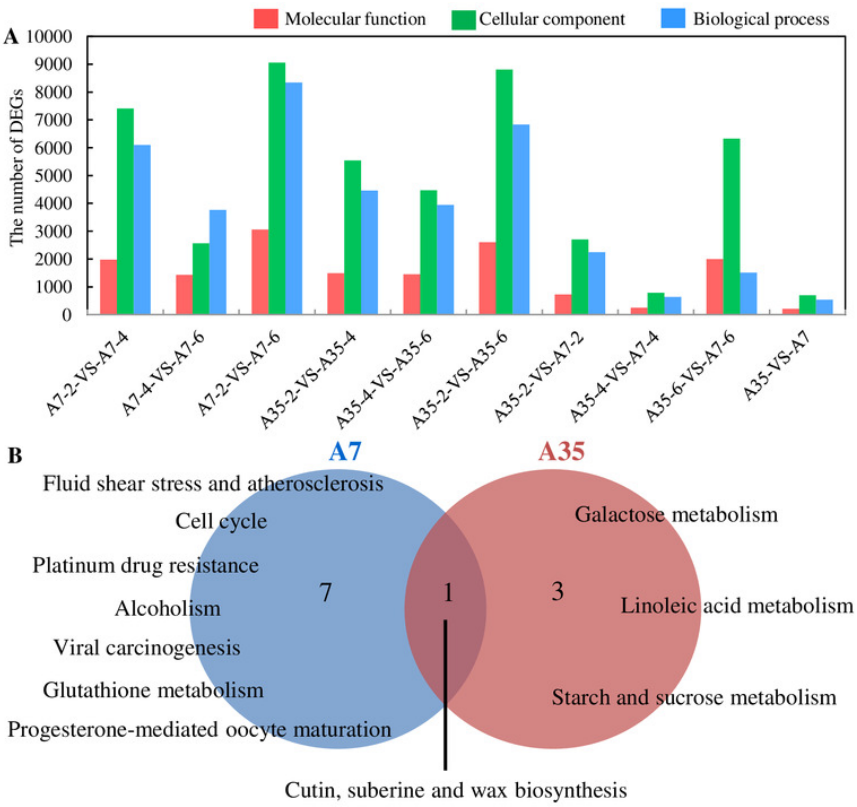

A
Molecular function
Cellular component
Biological process
The number of DEGs
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
A7-2-VS-A7-4
A7-4-VS-A7-6
A7-2-VS-A7-6
A35-2-VS-A35-4
A35-4-VS-A35-6
A35-2-VS-A35-6
A35-2-VS-A7-2
A35-4-VS-A7-4
A35-6-VS-A7-6
A35-VS-A7
B
A7
A35
Fluid shear stress and atherosclerosis
Cell cycle
Platinum drug resistance
Alcoholism
Viral carcinogenesis
Glutathione metabolism
Progesterone-mediated oocyte maturation
7
1
3
Galactose metabolism
Linoleic acid metabolism
Starch and sucrose metabolism
Cutin, suberine and wax biosynthesis

Figure 4

Figure 4. The analysis of the DEGs numbers involved in oil and protein metabolism.

A-D: the numbers of DEGs in lipid, fatty acids, proteins and amino acids metabolism in A7 and A35, respectively. The x-axis represented the comparison pairs. The y-axis represented the number of DEGs. The white and gray backgrounds represented the down- and up-regulated DEGs, respectively.

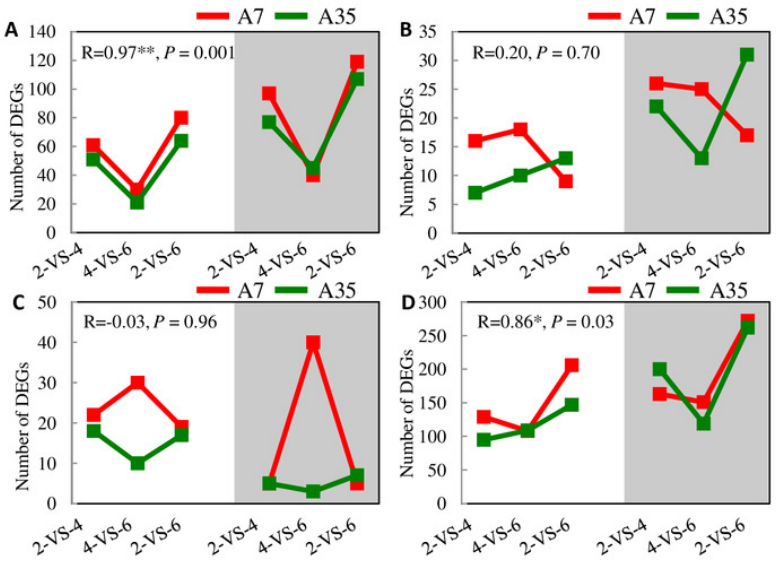

A
B
C
D
A7
A35
R=0.97**, P = 0.001
R=0.20, P = 0.70
R=-0.03, P = 0.96
R=0.86*, P = 0.03
Number of DEGs
2-VS-4
4-VS-6
2-VS-6

Figure 5

Figure 5. The expression trends of DEGs related to oil and SSPs.

X-axis represented the developmental stages; Y-axis represented FPKM in RNA-Seq.

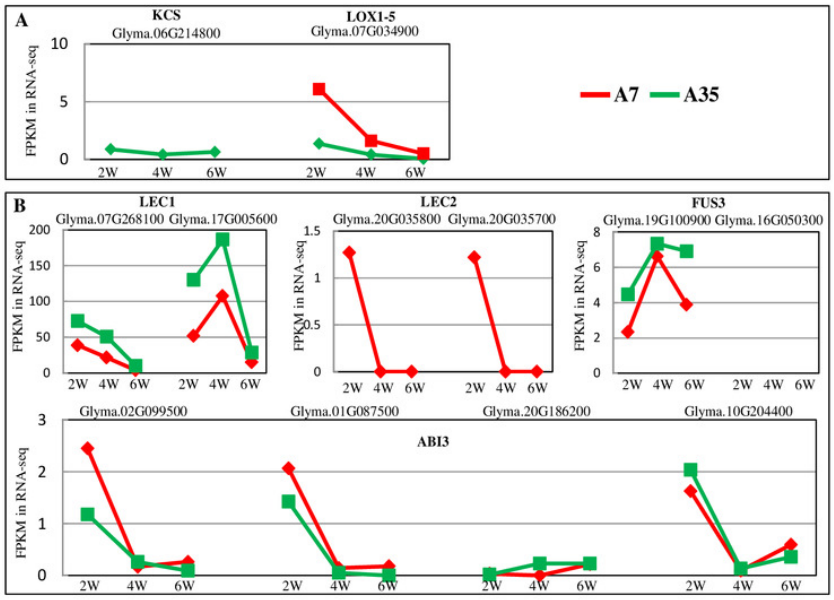

A
KCS
Glyma.06G214800
LOX1-5
Glyma.07G034900
FPKM in RNA-seq
A7
A35
B
LEC1
Glyma.07G268100 Glyma.17G005600
LEC2
Glyma.20G035800 Glyma.20G035700
FUS3
Glyma.19G100900 Glyma.16G050300
Glyma.02G099500
Glyma.01G087500
Glyma.20G186200
Glyma.10G204400
ABI3
2W 4W 6W

Figure 6

Figure 6. DEGs in soybeans associated with the fatty acid biosynthetic, elongation and degradation pathways.

Each square represents a comparative pair. The squares from left to right represent: A7-2, A7-4, A7-6, A35-2, A35-4 and A35-6. The red, blue squares indicate their expression levels. The enzymes are marked in blue in pathways. ACCase: acetyl-CoA carboxylase; MAT: malonyl-CoA; ACP transacylase; ACP: acyl carrier protein; FAS: fatty acid synthases; FATA/B: oleoyl-[acyl-carrier-protein] hydrolase; FAD2: fatty acid desaturase 2; ACOX: acyl-CoA oxidase; MFP2: enoyl-CoA hydratase; HAD: 3-hydroxyacyl-CoA dehydrogenase; atoB: acetyl-CoA C-acetyltransferase; KCS: 3-ketoacyl-CoA synthase; KAR: very-long-chain 3-oxoacyl-CoA reductase; PHS1: very-long-chain (3R)-3-hydroxyacyl-CoA dehydratase; TER: very-long-chain enoyl-CoA reductase; ACOT: acyl-CoA thioesterase.

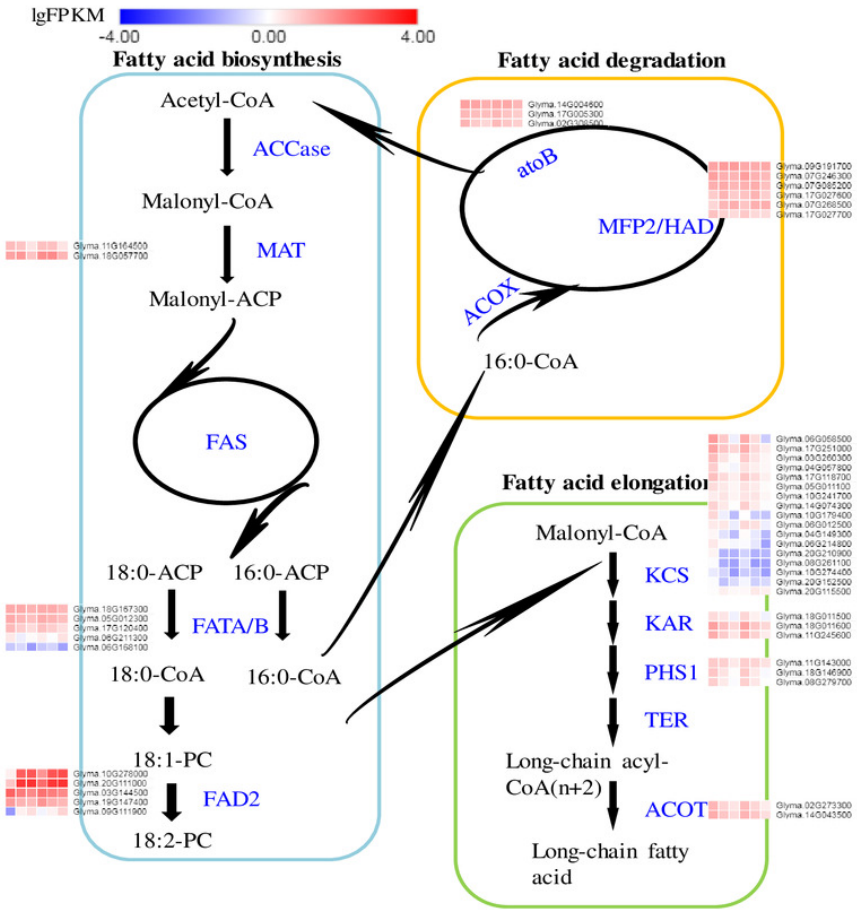
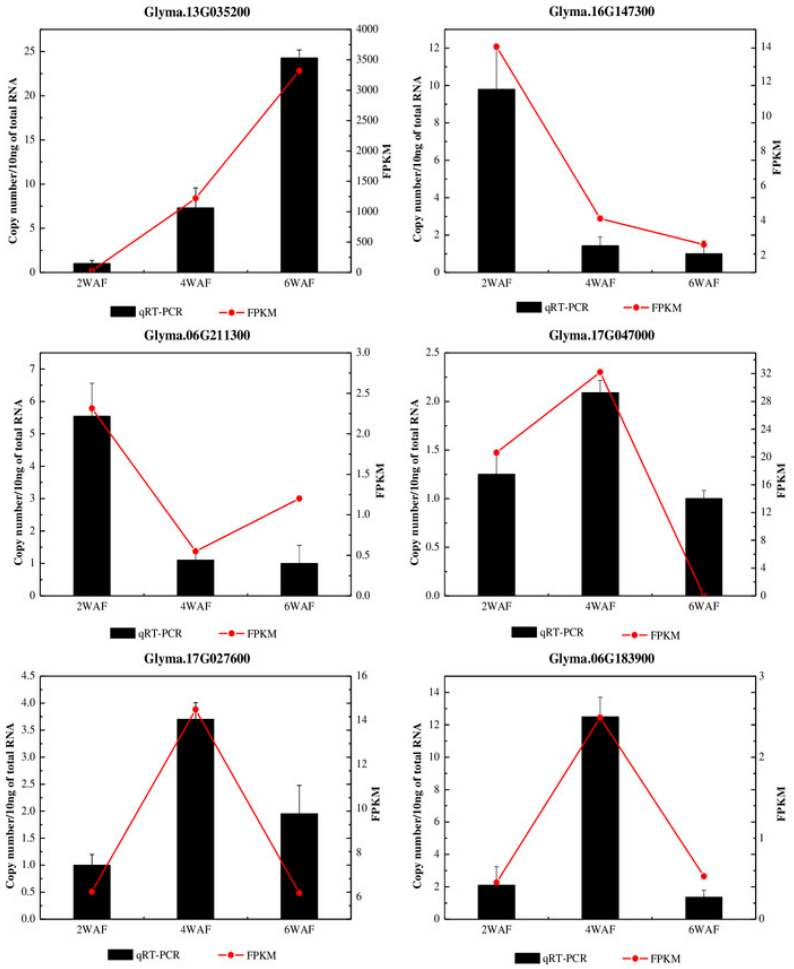

lgFPKM
-4.00
0.00
4.00
Fatty acid biosynthesis
Fatty acid degradation
Acetyl-CoA
ACCase
Malonyl-CoA
MAT
Malonyl-ACP
FAS
18:0-ACP
16:0-ACP
FATA/B
18:0-CoA
16:0-CoA
18:1-PC
FAD2
18:2-PC
atoB
MFP2/HAD
ACOX
16:0-CoA
Fatty acid elongation
Malonyl-CoA
KCS
KAR
PHS1
TER
Long-chain acyl-CoA(n+2)
ACOT
Long-chain fatty acid

Figure 7

Figure 7. qRT-PCR analysis of DEGs with different samples from that in RNAseq.

X-axis represented the developmental stages, the black columns represented qRT-PCR results, and the red dots represented RNA-Seq results; Y-axis represented the relative level of gene expression in qRT-PCR (left) and FPKM in RNA-Seq (right).


Glyma.13G035200
Glyma.16G147300
Glyma.06G211300
Glyma.17G047000
Glyma.17G027600
Glyma.06G183900
Copy number/10ng of total RNA
FPKM
2WAF
4WAF
6WAF
qRT-PCR
FPKM