Genetics and Molecular Research

Identification of genes associated with the increased number of four-seed pods in soybean (*Glycine max* L.) using transcriptome analysis

Z.Z. Liu[1], D. Yao[1], J. Zhang[1], Z.L. Li[2], J. Ma[1], S.Y. Liu[1], J. Qu[1], S.Y. Guan[1], D.D. Wang[1], L.D. Pan[1], D. Wang[1] and P.W. Wang[1]

[1]Center for Plant Biotechnology, Jilin Agricultural University, Changchun, China
[2]Center for Applied Genetic Technologies, University of Georgia, Athens, GA, USA

*These authors contributed equally to this study.
Corresponding author: P.W. Wang
E-mail: davidmedsci@sina.com

Genet. Mol. Res. 14 (4): 18895-18912 (2015)
Received August 16, 2015
Accepted October 9, 2015
Published December 28, 2015
DOI http://dx.doi.org/10.4238/2015.December.28.39

ABSTRACT. Seed number per pod is an important component of yield traits in soybean (*Glycine max* L.). In 2010, we identified a natural mutant with an increased number of four-seed pods from a soybean variety named 'Jinong 18' (JN18). Subsequent observations indicated that the trait was stably inherited. To identify and understand the function of genes associated with this mutant trait, we analyzed the genetic differences between the mutant (JN18MT01) and source variety (JN18) by transcriptome sequencing. Three types of tissues, axillary buds, unfertilized ovaries, and young pods at three different growth stages, V6, R1, and R3, were analyzed, respectively. The sequencing results yielded 55,582 expressed genes and 4183 differentially expressed genes (DEGs). Among these, the log2 ratio value of 162 DEGs was >10, and 13 DEGs hadoverlapping expression at three different growth stages. Comparisons of DEGs among three different growth stages yielded similar results in terms of the percentage of genes classified into each gene ontology (GO) category. DEGs were classified into 25 different functional groups in clusters of orthologous groups analysis. Proportions of the main functional genes differed significantly over developmental stages.

A comparison of enriched pathways among the three developmental stages revealed that 646 unigenes were involved in 103 metabolic pathways. These results show that the development of four-seed pods is associated with a complex network involving multiple physiological and metabolic pathways. This study lays the foundation for further research on cloning and on the molecular regulation of genes related to the four-seed pod mutation.

Key words: Soybean; Four-seed pod mutant; Transcriptome analysis; High-throughput Illumina sequencing; Developmental stages

INTRODUCTION

Soybean (*Glycine max*) is a valuable food and forage crop and an important source of high-quality plant protein, oil, and various functional medical and health products (Zhai, 1988).This crop plays an important role in food and economic security, especially in developing countries. Despite the importance of soybean resources, the soybean crop improvement program lags behind those of cereals (Graham and Vance, 2003). One of the main reasons for this lag is a lack of sufficient genomic resources. However, with the recent development of effective and high-throughput sequencing technologies, research on soybean genomics has attracted widespread attention (Cannon et al., 2009; Severin et al., 2010; Chan et al., 2012). Full genome sequencing has been completed for five legumes; soybean, *lotus*, *Medicago*, pigeon pea, and chickpea (Sato et al., 2008; Schmutz et al., 2010; Varshney et al., 2011; Young et al., 2011; Jain et al., 2013; Varshney et al., 2013). However, researchers are now faced with the more difficult problem of trying to understand the relationships between genome sequences and molecular function and regulation in different cells. These relationships will directly affect the process of crop improvement.

Under conditions where the external environment and climate do not affect yield, the main factors affecting soybean yield are the number of pods per plant, number of seeds per plant, number of seeds per pod, and 100-seed weight. Previous research has shown that the number of four-seed pods is positively correlated with the number of seeds per pod in soybean, and varieties with a higher proportion of four-seed pods show higher yields and productivity (Peng et al., 1994). A previous study showed that the number of three-seed pods was positively correlated with the yield per plant, and that lines with more seeds per pod produced higher yields than lines with fewer seeds per pod (Zhou et al., 2005). A genetic linkage map was constructed using recombinant inbred soybean lines, and the gene controlling the number of seeds per pod was located on a linkage group. Although there were discrepancies in the heritability of different yield traits, the number of seeds per pod showed high heritability (80.07%). That study also showed that the gene controlling the number of seeds per pod was located at the main effective quantitative trait locus that strongly contributed to yield (Wang et al., 2007). Therefore, the number of seeds per pod has some practical value for increasing yield, and increasing the number of seeds per pod may be an effective way to improve yield.

Transcriptome sequencing is an effective strategy used to explore the relationship between gene sequences and molecular mechanisms (Wang et al., 2009; Ozsolak and Milos, 2011; Jain, 2012). Global transcriptome analyses can provide insight into the location of genes, their functions, transcriptional regulation, and the molecular basis of various cellular processes. Transcriptome analyses are an essential step for basic and applied research on any organism. Several studies on soybean have focused on the overall and specific transcriptional activity of genes across various tissues, organs, and developmental stages (Cheng et al., 2009; Wong et al., 2009; Fan et al., 2013;

Wong et al., 2013). Several genes putatively involved in the control of important agronomic traits such as nodule, flower, and seed development have been identified (Wong et al., 2009; Severin et al., 2010; Jung et al., 2012). However, the molecular mechanisms underlying the development of the four-seed pod and the related increase in yield remain poorly understood.

Branching, flowering, and pod formation are three important developmental periods in soybean, and represent the transition from vegetative to reproductive growth. Factors that contribute to the formation of more flowers will result in more pods and more seeds. In this study, we chose the axillary bud (Vegetative6 Stage, V6-Stage), unfertilized ovary (Reproductive Stage, R1-Stage), and young pod (Reproductive Stage, R3-Stage) as the experimental materials. High-throughput Illumina sequencing was performed for the three different tissues, representing three developmental stages, of the four-seed pod mutant and the control. Based on extensive data analyses, we identified different genes and pathways related to the soybean four-seed mutation. This dataset will serve as the foundation to understand the mechanisms regulating seed and pod development, and will also lay the foundation for the identification of key genes in soybean.

MATERIAL AND METHODS

Plant materials

The soybean four-seed pod mutant used in this study was selected by chance through drought-stress experiments at the Plant Biotechnology Center of Jilin Agricultural University. This line was mutated from the soybean variety Jinong 18. Compared with the control, the mutant shows a 20% higher ratio of four-grain pods. This soybean mutant has been continuously selected since its identification in 2008, and was included in a regional field trial in 2014.

In July 2013, fresh axillary buds (V6-Stage, Figure 1A), unfertilized ovaries (R1-Stage, Figure 1B), and young pods (R3-Stage, Figure 1C) (http://www.soybeanmanagement.info) were collected from the four-grain pod soybean mutant and control growing in the greenhouse of the Plant Biotechnology Center of Jilin Agricultural University. All fresh materials were wrapped in aluminum foil, frozen in liquid nitrogen, and stored at -80°C until analysis.

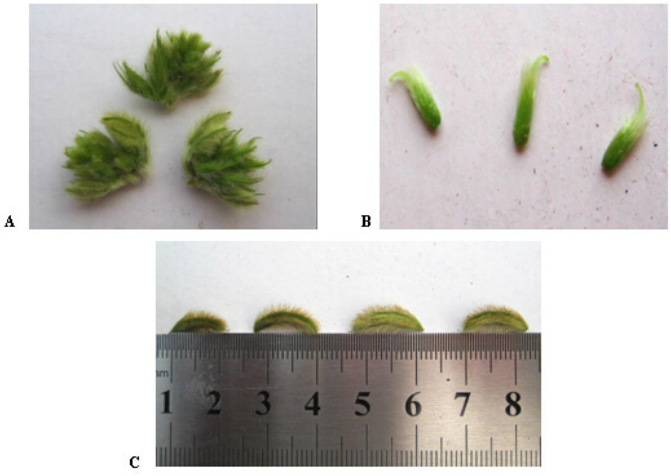

Figure 1. Plant materials in three different periods. **A.** Fresh axillary buds. **B.** Unfertilized ovaries. **C.** Young pods.

RNA isolation and construction of Illumina sequencing library

Total RNA was extracted from soybean at the three developmental phases described above using an Eastep™ Universal RNA Extraction Kit (Promega, Shanghai, China). Oligo (dT) beads were used to isolate poly(A) mRNA from total RNA, according to the Illumina manufacturer instructions. Fragmentation buffer was added to cut the mRNA into short fragments, and the fragments were then used to synthesize first-strand complementary DNA (cDNA) using random hexamer adaptors and reverse transcriptase (Invitrogen, Carlsbad, CA, USA). Second-strand cDNA was synthesized with RNase H (Invitrogen) and DNA polymerase I (NEB). A paired-end library was constructed from the cDNAs synthesized with a Genomic Sample Prep Kit (Illumina). The resulting short fragments with desired lengths were purified with a QIA quick polymerase chain reaction (PCR) (Qiagen, Valencia, CA, USA) Extraction Kit, and then end-repaired and linked to sequencing adapters (Margulies et al. 2005). After the unsuitable fragments were removed with AMPureXP beads, the sequencing library was constructed via PCR amplification. The cDNA library was quantified with Qubit 2.0 fluorescer (Invitrogen) and a cluster of the DNA fragments on the surface of a flow cell chip was amplified using bridge PCR. When the single molecular DNA cluster had been amplified many times, the products were sequenced using the Illumina GAII sequencing platform.

Illumina sequencing and functional annotation

Quality reads were successively assembled into contigs, scaffolds, and unigenes with Velvet and Oases software packages. We used the method of reads per kilobase of exon model per million mapped reads (RPKM) (Mortazavi et al., 2008) values to normalize the transcript levels. Genes were considered to be differentially expressed when there was a 2-fold difference in their transcript levels between two different tissues. All unigene sequences were used as queries to search the Clusters of Orthologous Groups (COG) database (E-value $<10^{-5}$), and were functionally annotated by Gene Ontology (GO) analysis with Blast2GO software (E-value $< 10^{-5}$) (http://www.blast2go.com/). Significantly enriched metabolic pathways were predicted by Kyoto Encyclopedia of Genes and Genomes (KEGG) mapping.

Quantitative real-time fluorescent PCR (qRT-PCR) analysis

To verify the results of transcriptome sequencing obtained from the soybean four-seed pod mutant at different developmental periods, the transcript levels of 10 genes were determined by quantitative real-time fluorescent transcription (qRT)-PCR. According to the sequenced transcriptome data, 10 differentially expressed genes (DEGs) between the control and mutants (log2 ratio = -19.43-14.36) were selected for analysis. All of these genes had GO functional annotations. For this analysis, the materials used for RNA extraction were the same as those used to construct the digital gene expression (DGE) library. Ten gene-specific primer pairs were designed based on sequences of the target genes using Primer Expression software (Table 1). Each reaction mixture (25 µL) contained 2 µL cDNA (25 ng/µL), 12.5 µL 2X SYBR premix Ex taq™ (Takara, Dalian, China), 2 µL forward and reverse primers (10 µM), 1.5 µL dNTP (10mM), and 5.0 µL ddH_2O. After gentle mixing, the PCR mixtures were placed in an Agilent Technologies Stratagene Mx3000P PCR instrument for the thermocycling reaction. The thermal cycling conditions were as follows: initial denaturation at 95°C for 3 min, denaturation at 94°C for 15 s, annealing at 60°C, and extension

for 40 s. The soybean tubulin gene (*TUB4*) (Fan et al., 2013) was used as the internal control. The relative expression levels of each transcript were calculated using the $2^{-\Delta\Delta Ct}$ method (Kenneth and Thomas, 2001): $(\Delta\Delta C_T = [C_{T, Target} - C_{T, Tub3}]_{Time\ x} - [C_{T, Target} - C_{T, Tub3})_{Time\ 0}$; where time x is any time point and time 0 represents the 1X expression of the target gene normalized to that of *TUB4*. The mean Ct values for both the target and internal control genes were determined at time zero.

Table 1. Primer sequences used for fluorescence quantification of candidate genes.

No.	ID#	Primer sequence (5'→3')	Annotation
1	Glyma08g22630.1	P1: GGAGCCAAATAGAACCGCACATAAC P2: AGTTCCTGTGTCCGAGTAAGTAGAG	Protein folding
2	Glyma16g01440.1	P1: GGATGTTCCTGGTCTCTCCAAGTCT P2: GAATCCAACCTCTCCTCTCCAACCT	Unfolded protein binding
3	Glyma15g41880.1	P1: CCCACATACCCATCTACCCAAAGC P2: TCTCTGAACCTCTCCGTCCTGAC	Protein folding
4	Glyma09g04500.1	P1: AGGTCAGTGAGGAGGTTGCCAAGTA P2: AGAAGGTCCACGAGTTGTAAGGATGG	Unknown
5	Glyma20g28360.1	P1: ACACAACTCTTTAAGGATGGCAACCC P2: AAGATACAAGACACAAGCCGAGTACG	Copper ion binding
6	Glyma02g03280.1	P1: GACTGCTTGCTTCAATGC P2: CGATGCTGCTATGTCTCA	Peptidase activity
7	Glyma08g11400.1	P1: ATCTTGAGTGGTATCTGGAA P2: TCTGTTCTTCAATTCTACCG	Protein binding
8	Glyma08g44950.1	P1: GGACTGGTGGAGGAACAATA P2: AAGTTGACACCGTAGCCT	DNA binding
9	Glyma08g45840.2	P1: CCCTGTGATGACCAAGAA P2: AACCTCAACAGAATCCATCC	Unknown
10	Glyma07g00260.1	P1: AGACTGCTTGCTTCAATG P2: CTTTCAGTGCTTTCTCCATT	Response to red or far red light
TUB4	EV263740	P1: GGCGTCCACATTCATTGGA P2: CCGGTGTACCAATGCAAGAA	Beta-tubulin

#ID is the ID of each differential gene. Annotation refers to the functional annotation of each gene in the GO database.

RESULTS

Illumina sequence data and assembly

The cDNAs obtained from the four-grain pod soybean mutant (JN18MT01) and the control (JN18) at three growth phases (axillary bud, young pod, and unfertilized ovary) were sequenced using Illumina sequencing technology. Large transcriptome datasets were generated for each sample. The smallest dataset was obtained from the young-pod mutant sample (4,017,760,362 bp) and the largest was from the unfertilized ovary control sample (4,686,416,658 bp). In each sample, 100% of the clean reads had quality scores at the cycleQ20 level, and the GC content was approximately 46-48% (Table 2). When sequence data were mapped against the soybean reference genome (allowing a ≤2-bp mismatch for each read), there was an average of 67.13% mapped reads, 44.13% perfect reads, and 96.89% InDels (Table 2).

Screening to detect differentially expressed genes

The strategies used to detect DEGs, the pairs of samples compared, and the number of DEGs between each pair of samples are shown in Table 3. In total, 4183 DEGs were detected in the mutant and the source variety. The highest number of DEGs was found between the axillary

bud mutant and the source variety (2138 up-regulated and 41 down-regulated genes). The fewest number of DEGs was found between the young pod mutant and the control (69 up-regulated and 22 down-regulated genes) (Table 3). Among them, the log2 ratio value of 13 DEGs was greater than 10. The DEGs in the three different growth phases of soybean and the overlaps among them are illustrated in a Venn diagram (Figure 2). This diagram shows that 13 DEGs were expressed in the axillary bud, young pod, and unfertilized ovary; 11 of these were up-regulated and two were down- regulated (Table 4). These genes were further evaluated to investigate their biological functions. The scatter plot showed that the expression of up- and down-regulated genes in the mutants and the control fitted a normal distribution at the V6-Stage (Figure 3A), R1- Stage (Figure 3C), and R3-Stage (Figure 3B).

Table 2. Summary of raw reads and reference statistics for six samples.

Sample	Data (bp)	GC (%)	Cycle Q20 (%)	Mapped reads	Perfect reads	InDel
Axillary bud mutant	4,180,012,452	46.73	100	14,050,443/66.55%	6,426,153/45.73%	13,618,761/96.92%
Axillary bud control	4,408,624,836	48.21	100	15,045,492/67.57%	5,314,605/35.32%	14,559,026/96.76%
Young pod mutant	4,017,760,362	46.26	100	13,601,515/67.02%	6,168,703/45.35%	13,176,640/96.87%
Young pod control	4,121,206,254	46.72	100	13,598,653/65.33%	6,211,103/45.67%	13,181,916/96.93%
Ovary mutant	4,400,207,856	46.48	100	15,186,073/68.33%	7,018,883/46.21%	14,712,213/96.87%
Ovary control	4,686,416,658	46.34	100	16,094,490/67.99%	7,487,011/46.51%	15,605,648/96.96%

Data are total nucleotides. GC (%) is the proportion of guanidine and cytosine nucleotides among the total nucleotides. CycleQ20 (%) is the proportion of nucleotides with a quality value larger than 20. Mapped Reads represents the number of reads and comparison with reference genome alignment. Perfect Reads represents no mismatch, no insertion deletion, the only reads the comparison on the genome. InDel represents insertion deletion reads.

Table 3. Statistical analysis of differentially expressed genes between the mutant and the source variety.

Type	ABC1_*vs*_ABM2	OVC1_*vs*_OVM2	YPC1_*vs*_ YPM2
Total number of DEGs	2179	2060	91
Number of upregulated genes	2138	1381	69
Number of downregulated genes	41	679	22
Number of |log2 ration|≥ 10	74	75	13

ABC1 = axillary bud control; ABM2 = axillary bud mutant; YPC1 = young pod control; YPM2 = young pod mutant; OVC1 = unfertilized ovary control; OVM2 = unfertilized ovary mutant.

Figure 2. Venn diagram showing overlaps among differentially expressed genes of the soybean mutant and control in three different periods. Numbers outside the circles denote the total number of differentially expressed genes in three different periods. Numbers in one circle denote differentially expressed genes, and numbers in two or more intersecting circles denote overlapping genes.

Table 4. Different genes with overlapping expression in the axillary bud, young pod, and unfertilized ovary.

No.	ID#	log_2 ratio	Regulated	Annotation	Function
1	Glyma02g03230.1	12.3350	Up	Zinc ion binding	Glycine max cDNA, clone: GMFL01-27 - G05
2	Glyma02g03250.1	13.9117	Up	Peptidase activity; zinc ion binding	Glycine max metalloendoproteinase 1-like (LOC100814809), mRNA
3	Glyma02g03280.1	-5.0532	Down	Peptidase activity	Soybean clone JCVI-FLGm-9N16 unknown mRNA
4	Glyma02g03290.1	13.2829	Up	Peptidase activity; zinc ion binding	Glycine max metalloendoproteinase 1-like (LOC100820140), mRNA
5	Glyma02g03310.1	12.5956	Up	Peptidase activity; zinc ion binding	Glycine max metalloendoproteinase 1-like (LOC100782377), mRNA
6	Glyma06g47780.1	2.4507	Up	Protein binding	Glycine max disease resistance
7	Glyma12g31280.1	-3.0000	Down	-	Glycine max uncharacterized LOC100816278 (LOC100816278), ncRNA
8	Glyma13g37760.1	2.2224	Up	-	Glycine max uncharacterized LOC100795714 (LOC100795714)
9	Glyma15g03620.1	4.2095	Up	Beta-galactosidase activity	Glycine max cyanogenic beta-glucosidase-like (LOC100812431), transcript variant X5, misc_RNA
10	Glyma15g19910.1	11.4617	Up	DNA binding	Glycine max ethylene-responsive transcription factor ERF017-like (LOC100813688), mRNA
11	Glyma18g50470.1	11.7776	Up	Methyltransferase activity	Glycine max isoflavone 7-O-methyltransferase-like (LOC100776336), mRNA
12	Glyma19g05220.1	4.3923	Up	TRansferase activity	Glycine max coumaroyl-CoA: anthocyanidin 3-O-glucoside-6''-O-coumaroyltransferase 1-like (LOC100797080), mRNA
13	Glyma19g05290.1	3.3923	Up	Transferase activity	Glycine max coumaroyl-CoA: anthocyanidin 3-O-glucoside-6''-O-coumaroyltransferase

#ID is the ID of the differential gene. Regulated refers to the direction of the gene regulation pattern (up or down). Annotation is the functional annotation of the gene in the GO database.

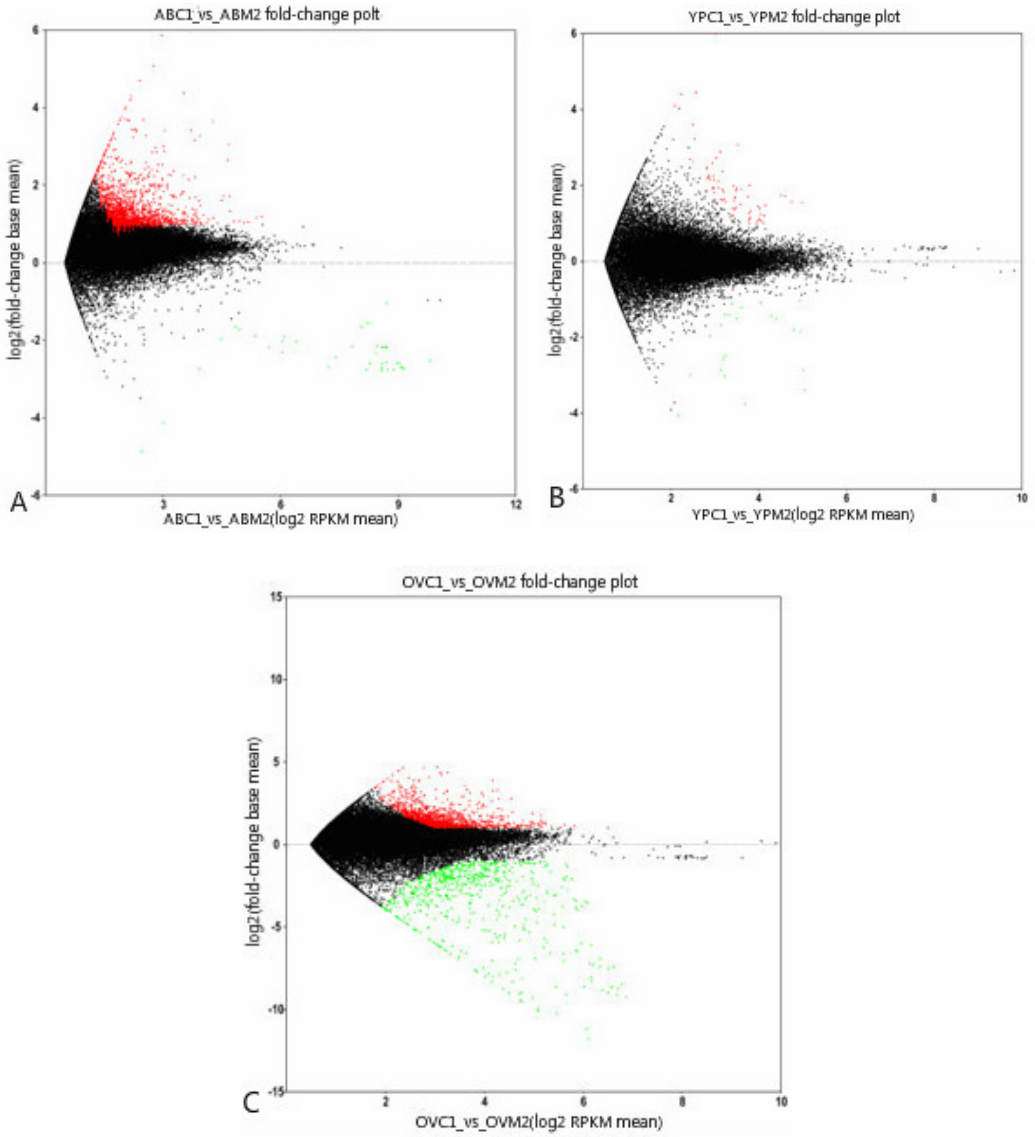

Figure 3. Scatter plot of soybean control and mutant differentially expressed genes in three different periods. **A.** Axillary bud control and mutant. **B.** Young pod control and mutant. **C.** Ovary control and mutant. Red represents upregulation genes, green represents downregulated genes. Abscissa represents the relative log2 RPKM mean, ordinate represented fold change base mean.

Functional annotation of differentially expressed genes

The assembled unigenes were further analyzed to determine their pathway annotations (KEGG), COG functional annotations, and GO functional annotations.

Gene Ontology functional analysis

GO is an internationally standardized gene function classification system. It provides dynamically updated and controlled vocabulary to fully describe the attributes of genes and gene products in an organism. The GO database is suitable to analyze genes from various species, and it can limit and describe genes and proteins. In GO analyses, genes are classified into three main categories; biological process, cellular component, and molecular function. The DEGs in the axillary bud (Figure 4A), young pod (Figure 4B), and unfertilized ovary (Figure 4C) were further classified into 18 subcategories in the cellular component category, 18 subcategories in the molecular function category, and 25 subcategories in the biological process category.

The percentage of DEGs in each category was similar across the three different growth phases. Of the 583,516 unigenes with known functions, 208,516 were assigned to the molecular function category, 74,563 were assigned to the cellular component category, and 300,437 were assigned to the biological process category. However, when the DEGs were compared among the three different developmental stages, there were clear differences in the distribution of subcategories. In the cellular component category, the cell part subcategory had the largest number of DEGs (65 or 28.26% of DEGs in the axillary bud, 1737 or 23.22% of DEGs in the young pod, and 1759 or 21.5% of DEGs in the unfertilized ovary). In the biological process category, DEGs in the cellular process and metabolic process subcategories were the most abundant (1563 or 13.61% of DEGs in the axillary bud, 54 or 13.37% of DEGs in the young pod, and 1557 or 12.42% of DEGs in the unfertilized ovary). In the molecular function category, DEGs in the binding protein subcategory were the most abundant (1254 or 42.95% of DEGs in the axillary bud, 42 or 43.3% of DEGs in the young pod, and 1303 or 44.62% of DEGs in the unfertilized ovary).

COG functional annotation

The orthologous classification of genetic products was conducted using tools at the COG database (COG, http://www.ncbi.nlm.nih.gov/COG/). It is assumed that each COG protein evolved from a specific ancestor (Tatusov et al., 2003). The COG database contains protein sequences encoded in the genomes of algae, bacteria, and eukaryotes, and provides information about their evolutionary relationships (Lulin et al., 2012). DEGs between the soybean mutant and the control at the three different growth periods were further annotated based on COG categories. The DEGs were assigned COG functional annotations, which could be grouped into 25 different functional categories. In the soybean axillary bud, the largest category of DEGs was general function prediction only (231, 19.66%), followed by transcription (93, 7.91%), carbohydrate transport and metabolism (91, 7.74%), posttranslational modification, protein turnover, chaperones (81, 6.89%), signal transduction mechanisms (79, 6.72%), amino acid transport and metabolism (73, 6.21%), energy production and conversion (57, 4.85%), secondary metabolites biosynthesis, transport, and catabolism (57, 4.85%), replication, recombination, and repair (54, 4.60%), function unknown (57, 4.85%), and others (Figure 5A). In the young pod, the largest category of DEGs was general function prediction only (11, 26.83%), followed by carbohydrate transport and metabolism (5, 12.20%), secondary metabolites biosynthesis, transport, and catabolism (5, 12.20%), amino acid transport and metabolism (3, 7.31%), posttranslational modification, protein turnover, chaperones (2, 4.88%), signal transduction mechanisms (2, 4.88%), function unknown (2, 4.88%), and others (Figure 5B). In the unfertilized ovary, the largest category of DEGs was general function prediction only (209, 16.95%), followed by posttranslational modification, protein turnover, chaperones

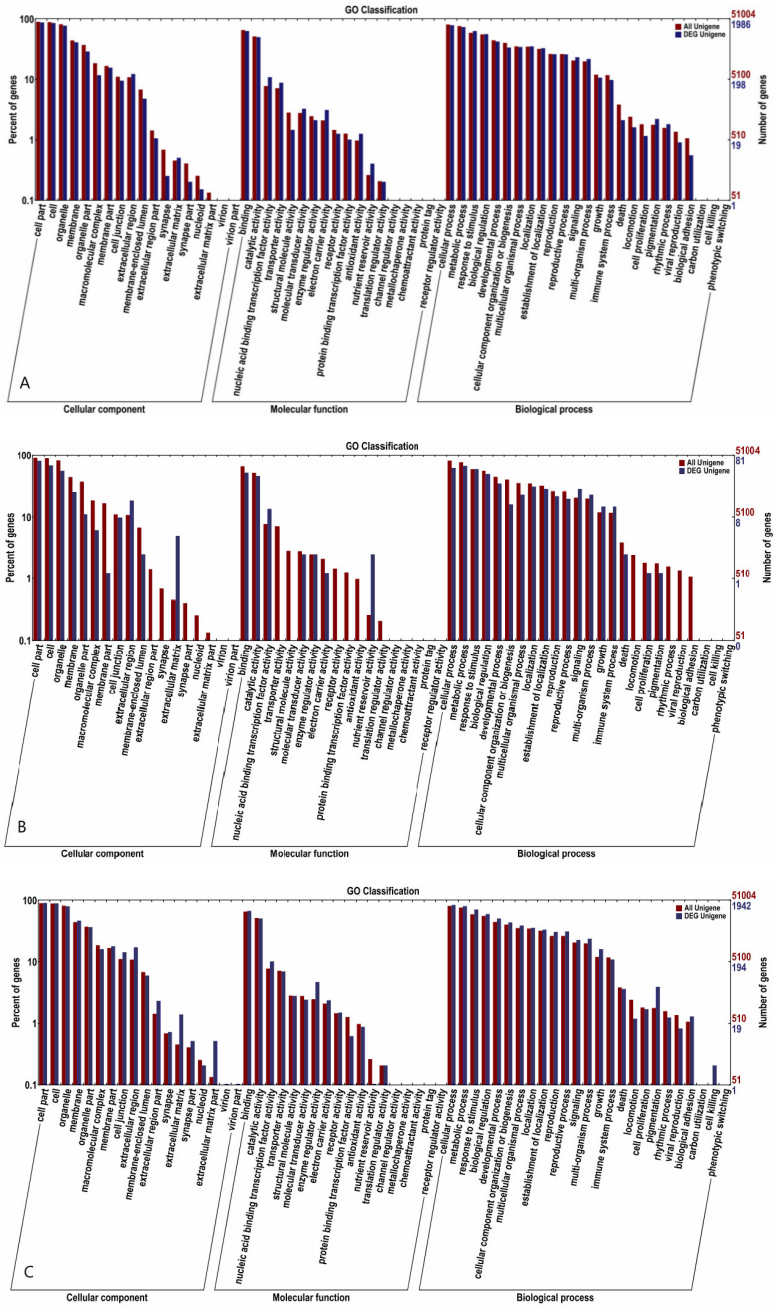

Figure 4. GO functional categorization of significantly differentially expressed genes in three different periods. Soybean axillary bud control and mutant (**A**), soybean young pod control and mutant (**B**), soybean ovary control and mutant (**C**). Abscissa represented gene function classification by GO analysis, from left to right in turn = Cellular component, Molecular function, Biological process. Left ordinate represents the percentage of genes, right ordinate represents the number of genes. In addition, the red column represents all unigenes and the blue column represents DEG unigenes.

(183, 14.84%), transcription (105, 8.52%), carbohydrate transport and metabolism (85, 6.89%), signal transduction mechanisms (83, 6.73%), replication, recombination, and repair (78, 6.33%), translation, ribosomal structure, and biogenesis (77, 6.25%), amino acid transport and metabolism (69, 5.60%), secondary metabolites biosynthesis, transport, and catabolism (61, 4.95%), function unknown (40, 3.24%), and others (Figure 5C).

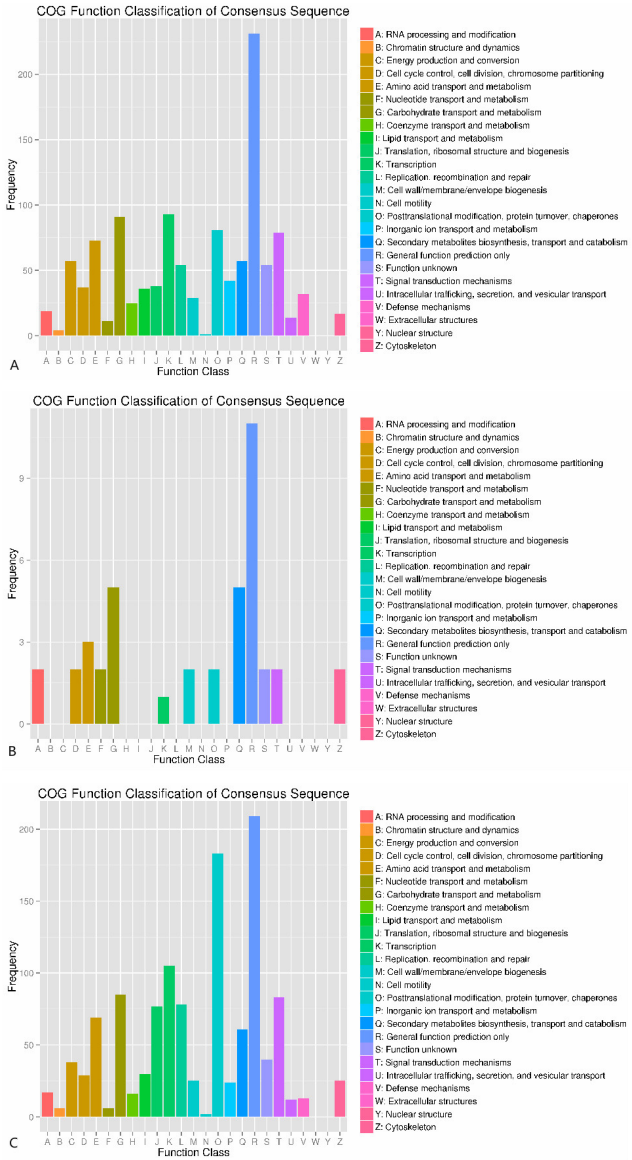

Figure 5. COG function classification of consensus sequences in three different periods. Soybean axillary bud contrast and mutant (**A**), soybean young pod contrast and mutant (**B**), soybean ovary contrast and mutant (**C**). Abscissa represents gene function classification by COG analysis, including 25 different functions; different capital letters represent different functions; ordinates represents gene frequency. Different colored columns represent different gene functions.

Significant enrichment analysis of KEGG pathways

Within an organism, different genes coordinate to perform biological functions. Analyses based on pathways can provide information about the biological functions of genes. Gene functions were analyzed using the hypergeometric model in KEGG to identify pathways significantly enriched with DEGs. These analyses revealed the main signal transduction pathways and biochemical metabolic pathways in which the DEGs were involved.

In this study, pathways with a Q_value of ≤ 0.05 were defined as those that were significantly enriched with DEGs. The sequencing results showed that a total of 646 unigenes of DEGs were involved in 103 metabolic pathways. At the axillary bud growth phase, a total of 305 unigenes of DEGs were involved in 93 metabolic pathways. At the young pod growth stage, three unigenes of DEGs were involved in four pathways. At the unfertilized ovary growth stage, 338 unigenes of DEGs were involved in 81 metabolic pathways. Among these metabolic pathways, eight had Q_values of ≤ 0.05 (Table 5). These eight pathways consisted of those associated with amino acid metabolism (beta-alanine metabolism, arginine and proline metabolism, and amino sugar and nucleotide sugar metabolism), those related to carbon metabolism (starch and sucrose metabolism, and glyoxylate and dicarboxylate metabolism), and those related to plant hormone signal transduction and protein processing in the endoplasmic reticulum. Metabolic maps were constructed after aligning DEGs to the KEGG database.

Table 5. Pathways of differentially expressed genes that were significantly enriched with a Q_value ≤0.05.

No.	#KEGG_Pathway	ko_ID	Cluster_frequency	Genome_frequency	P value	Q_value
1	Plant hormone signal transduction	ko04075	57 of 305/18.69%	742 of 9217/8.05%	1.04E-09	9.66E-08
2	Glyoxylate and dicarboxylate metabolism	ko00630	15 of 305/4.92%	96 of 9217/1.04%	5.15E-07	4.79E-05
3	Beta-Alanine metabolism	ko00410	11 of 305/3.61%	101 of 9217/1.10%	0.000487	0.045289
4	Arginine and proline metabolism	ko00330	17 of 305/5.57%	208 of 9217/2.26%	0.000528	0.049122
5	Amino sugar and nucleotide sugar metabolism	ko00520	2 of 3/66.67%	249 of 9217/2.70%	0.002142	0.008568
6	Starch and sucrose metabolism	ko00500	2 of 3/66.67%	296 of 9217/3.21%	0.003018	0.012073
7	Valine, leucine, and isoleucine biosynthesis	ko00290	1 of 3/33.33%	90 of 9217/0.98%	0.029012	0.116047
8	Protein processing in the endoplasmic reticulum	ko04141	52 of 338/15.38%	429 of 9217/4.66%	2.01E-12	1.63E-10

Cluster analysis of differentially expressed genes based on expression patterns

We conducted a cluster analysis of DEGs based on their expression patterns; that is, those genes with the same or similar expression patterns were grouped together. Such analyses can provide important clues to the roles of functionally unknown genes in biological processes, because genes sharing the same expression pattern have related functions. Cluster analysis of DEGs between the soybean mutant and the control was conducted through average linkage distance, i.e., genes with similar functions were clustered together in the taxonomic tree. In this way, genes with a known function can shed light on the functions of unknown genes. The results of the cluster analyses are summarized in Table 6 and Figure 6 with the young pod mutant and control as an example.

Validation of transcriptome data by qRT-PCR

We conducted quantitative real-time polymerase chain reaction (qRT-PCR) analyses to verify the Solexa sequencing results. Specific primers used in the qRT-PCRs were designed for

10 candidate genes using Beacon Designer 7.5 SYBR® Green Design, and *TUB4* served as the internal control. Reverse transcription was conducted using the same RNA samples as were used in the Solexa sequencing, and internal control and candidate genes were amplified at the same time. The qRT-PCR results of the expression patterns of candidate genes were generally consistent with the Solexa sequencing results (Figure 7), confirming the reliability of the transcriptome data. The significant differences in transcript abundance may result from the different algorithms used in the two technologies.

Table 6. List and annotation of genes that were differentially expressed between soybean young pod mutant and contrast.

No.	ID#	YPM	YPC	Chi	Q_value	log2 ratio	Regulated	Annotation
1	Glyma02g03250.1	96	1003	0	0	-3.3851	Down	Metalloendopeptidase activity
2	Glyma02g03290.1	125	918	0	0	-2.8766	Down	Metalloendopeptidase activity
3	Glyma02g03230.1	175	612	0	0	-1.8062	Down	Metalloendopeptidase activity
4	Glyma02g03310.1	212	755	0	0	-1.8324	Down	Metalloendopeptidase activity
5	Glyma02g13290.1	125	918	0	0	-2.8766	Down	Metalloendopeptidase activity
6	Glyma06g08860.1	116	335	0	0	-1.5300	Down	Catalytic activity
7	Glyma04g08760.1	103	281	0	0	-1.4479	Down	Nucleotide transport and metabolism
8	Glyma14g06990.1	11	155	0	0	-3.8167	Down	Serine-type endopeptidase activity
9	Glyma06g45050.1	540	181	0	0	1.5770	Up	O-methyltransferase activity
10	Glyma12g12230.1	397	118	0	0	1.7504	Up	O-methyltransferase activity
11	Glyma12g31280.1	454	137	0	0	1.7285	Up	Unknown
12	Glyma14g02240.1	752	258	0	0	1.5434	Up	Lipid transport
13	Glyma02g44200.1	116	31	0	0	1.9038	Up	Plasma membrane
14	Glyma12g26780.1	142	56	0	0	1.3424	Up	Sequence-specific DNA binding transcription factor activity
15	Glyma17g07740.1	169	80	0	0	1.0790	Up	Nucleotide binding
16	Glyma02g37020.1	177	82	0	0	1.1101	Up	Nucleotide binding
17	Glyma10g36910.1	141	43	0	0	1.7133	Up	Response to stress
18	Glyma15g13750.1	188	90	0	0	1.0627	Up	Unknown
19	Glyma13g39890.1	117	14	0	0	3.0630	Up	Cytoplasmic membrane-bounded vesicle
20	Glyma16g28150.1	201	100	0	0	1.0072	Up	Response to stress
21	Glyma06g45380.1	212	91	0	0	1.2201	Up	Unknown
22	Glyma09g10340.1	185	67	0	0	1.4653	Up	Very long-chain fatty acid metabolic process
23	Glyma17g03350.1	191	73	0	0	1.3876	Up	Response to biotic stimulus
24	Glyma14g04580.1	154	38	0	0	2.0189	Up	Plasmodesma
25	Glyma15g13770.1	191	81	0	0	1.2376	Up	Unknown
26	Glyma15g13760.1	233	83	0	0	1.4891	Up	Unknown
27	Glyma11g14580.1	68	13	0	0	2.3870	Up	Ubiquitin-protein ligase activity
28	Glyma05g02950.1	89	29	0	0	1.6178	Up	Carboxylesterase activity
29	Glyma15g19910.1	67	9	0	0	2.8962	Up	DNA binding
30	Glyma12g35550.1	91	34	0.000001	0.000232	1.4203	Up	DNA binding
31	Glyma02g03280.1	63	0	0	0	15.9431	Up	Peptidase activity
32	Glyma12g09830.1	92	29	0	0	1.6656	Up	Nucleic acid binding
33	Glyma06g45370.1	125	62	0.00001	0.001872	1.0116	Up	Unknown
34	Glyma20g30700.1	92	28	0	0	1.7162	Up	Response to stress
35	Glyma05g35170.1	105	37	0	0	1.5048	Up	Unknown
36	Glyma19g29880.1	91	23	0	0	1.9842	Up	2-isopropylmalate synthase activity
37	Glyma15g10710.1	127	60	0.000003	0.000633	1.0818	Up	Unknown
38	Glyma12g07190.1	133	60	0	0	1.1484	Up	Monooxygenase activity
39	Glyma17g15460.1	100	26	0	0	1.9434	Up	DNA binding
40	Glyma11g18460.1	98	27	0	0	1.8598	Up	Nucleic acid binding
41	Glyma08g48040.1	131	60	0.000001	0.000232	1.1265	Up	Regulation of transcription, DNA-dependent
42	Glyma17g03360.1	134	55	0	0	1.2847	Up	Response to biotic stimulus
43	Glyma18g53440.1	159	79	0.000001	0.000232	1.0091	Up	Regulation of transcription, DNA-dependent
44	Glyma13g35950.1	64	17	0	0	1.9125	Up	Copper ion binding
45	Glyma13g37760.1	66	18	0	0	1.8745	Up	Unknown
46	Glyma14g04300.1	58	8	0	0	2.8580	Up	Cell part
47	Glyma06g45410.1	81	31	0.000005	0.000986	1.3857	Up	Unknown
48	Glyma07g37270.1	85	33	0.000004	0.000815	1.3650	Up	Metal ion binding
49	Glyma09g01610.1	50	10	0	0	2.3219	Up	Protein binding
50	Glyma06g35710.1	58	18	0.000008	0.001541	1.6881	Up	Sequence-specific DNA binding transcription factor activity

Continued on next page

Table 6. Continued.

No.	ID#	YPM	YPC	Chi	Q_value	log2 ratio	Regulated	Annotation
51	Glyma12g06460.1	50	11	0.000001	0.000232	2.1844	Up	Ubiquitin-protein ligase activity
52	Glyma13g35850.1	49	12	0.000004	0.000815	2.0297	Up	Copper ion binding
53	Glyma09g04750.1	58	13	0	0	2.1575	Up	Ubiquitin-protein ligase activity
54	Glyma20g38140.1	79	34	0.000045	0.007421	1.2163	Up	Protein binding
55	Glyma02g09000.1	67	22	0.000004	0.000815	1.6067	Up	Response to mechanical stimulus
56	Glyma07g31300.1	65	21	0.000004	0.000815	1.6301	Up	Copper ion binding
57	Glyma09g08330.1	50	8	0	0	2.6439	Up	DNA binding
58	Glyma07g05620.1	61	19	0.000005	0.000986	1.6828	Up	DNA binding
59	Glyma04g06720.1	57	14	0.000001	0.000232	2.0255	Up	Unknown
60	Glyma13g01930.1	75	32	0.000061	0.009565	1.2288	Up	DNA binding
61	Glyma10g32830.1	33	1	0	0	5.0444	Up	Serine-type endopeptidase inhibitor activity
62	Glyma11g03900.1	41	8	0.000004	0.000815	2.3576	Up	DNA binding
63	Glyma16g29670.1	48	15	0.000052	0.008332	1.6781	Up	Response to biotic stimulus
64	Glyma19g01910.1	42	7	0.000001	0.000232	2.5850	Up	Transferase activity, transferring hexosyl groups
65	Glyma06g10700.1	44	9	0.000003	0.000633	2.2895	Up	Plant-type cell wall
66	Glyma08g16810.1	48	13	0.000012	0.002218	1.8845	Up	Ethylene biosynthetic process
67	Glyma13g04780.1	41	7	0.000002	0.000438	2.5502	Up	Transferase activity, transferring hexosyl groups
68	Glyma08g45600.1	43	9	0.000004	0.000815	2.2563	Up	Endopeptidase inhibitor activity
69	Glyma13g35820.1	46	12	0.000013	0.002385	1.9386	Up	Unknown
70	Glyma17g12710.1	43	9	0.000004	0.000815	2.2563	Up	Unknown
71	Glyma13g17250.1	29	2	0.000002	0.000438	3.8580	Up	Sequence-specific DNA binding transcription factor activity
72	Glyma17g38010.1	30	3	0.000004	0.000815	3.3219	Up	Nucleic acid binding
73	Glyma12g34580.1	39	10	0.000053	0.008455	1.9635	Up	Copper ion binding
74	Glyma11g25660.1	21	0	0.000006	0.001168	14.3581	Up	Calcium ion binding
75	Glyma07g30520.1	25	3	0.000045	0.007421	3.0589	Up	Detection of biotic stimulus
76	Glyma04g10870.1	27	4	0.000051	0.008207	2.7549	Up	Plant-type cell wall
77	Glyma11g25670.1	17	0	0.000048	0.00781	14.0533	Up	Calcium ion binding
78	Glyma07g13790.1	88	186	0	0	-1.0797	Down	Hydrolase activity
79	Glyma08g13440.1	9	77	0	0	-3.0969	Down	Nutrient reservoir activity
80	Glyma03g28850.1	33	94	0	0	-1.5102	Down	Glucan endo-1,3-beta-D-glucosidase activity
81	Glyma19g05220.1	13	75	0	0	-2.5284	Down	Transferase activity
82	Glyma07g13780.1	44	103	0	0	-1.2271	Down	Hydrolase activity
83	Glyma14g06980.1	9	68	0	0	-2.9175	Down	Protein binding
84	Glyma19g05290.1	12	72	0	0	-2.5850	Down	O-acyltransferase activity
85	Glyma07g13800.1	44	103	0	0	-1.2271	Down	Hydrolase activity
86	Glyma17g07690.1	18	59	0.000002	0.000438	-1.7127	Down	Amino acid transport
87	Glyma18g50470.1	34	75	0.000046	0.007535	-1.1414	Down	8-hydroxyquercitin 8-O-methyltransferase activity
88	Glyma13g01570.2	24	63	0.000016	0.002871	-1.3923	Down	Response to abiotic stimulus
89	Glyma06g47780.1	11	44	0.000005	0.000986	-2.0000	Down	Protein kinase activity
90	Glyma05g30300.1	3	26	0.000014	0.002549	-3.1155	Down	Nutrient reservoir activity
91	Glyma15g03620.1	1	19	0.000043	0.007171	-4.2479	Down	Glucan exo-1,3-beta-glucosi -dase activity

#ID is the ID of each differential gene. YPM is the young pod mutant. YPC is the young pod contrast. Chi is the P value obtained by the Chi-square test. Q_value is the P value of the Chi-square test after FDR correction. log2(A/B) is the log2 value of two samples ratio. Regulated refers to the direction of the gene regulation pattern (up or down). Annotation is the functional annotation of the gene in the GO database.

DISCUSSION

Illumina sequencing technology has become an indispensable tool in genomics, and has been used in a diverse range of biological studies. Comparisons between the transcriptomes of the soybean four-seed pod mutant and the control plant at three developmental stages have laid a foundation for the understanding of changes in gene expression during different development phases. We obtained a large dataset (25.81 Gbp raw data) by Illumina sequencing, and 55,582 expressed genes were identified by comparison with the soybean reference genome. The sequencing depth and assembly efficiency in this study were greater than those reported previously (Fan et al., 2013).

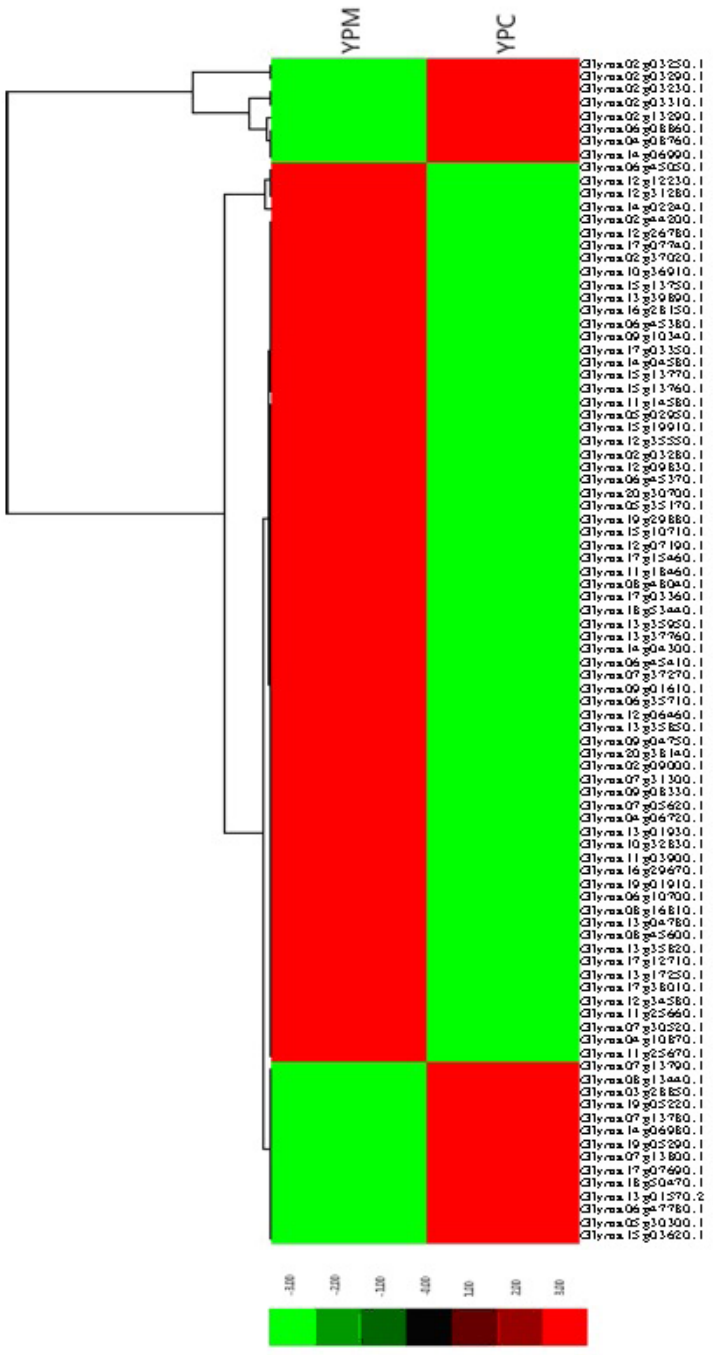

Figure 6. Differential gene expression pattern clustering analysis in soybean young pod mutant and control. Clustering figure columns represent two samples; lines represented different genes. Red represents high gene expression; green represents lower gene expression.

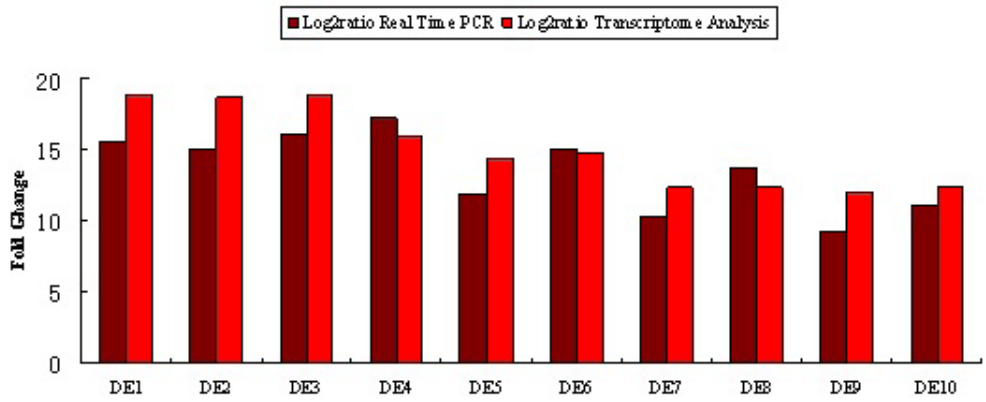

Figure 7. Relative expression levels obtained by real time PCR and transcriptome analyses of 10 candidate genes. Abscissa represents differentially expressed candidate genes; ordinates represent fold change (log_2 ratio).

Functional annotations were assigned to 4010 of the DEGs (95.86%). DEGs between the mutant and the control were involved in the processing of genetic information, metabolism, transcription, translation, and signal transduction related to major substances such as sugars, fats, amino acids, and plant hormones. The results indicated that gene expression was affected by many factors, and that the transfer of genetic information at the different growth stages of soybean is a very complicated process. In addition, we found there was about a 40% overlap in the expressed DEGs associated with peptidase activity and metal ion binding. In this study, the smallest number of DEGs was found between the young pod mutant and the young pod control. The expression level of most genes was relatively low in the mutant at the young pod growth stage, and the fold-changes in expression were lower than those detected at other growth stages. Further research should be conducted to explore the reasons for the low levels of gene expression observed in the young pod of the mutant.

The main objective of plant transcriptome sequencing is to evaluate variations in gene expression under different conditions or among different growth stages, and to identify new genes or those specifically expressed during a particular period via comparisons of sequences with those in databases. For example, transcriptome sequencing of soybean leaves and roots identified many genes that were expressed under different salinity and drought stress conditions (Fan et al., 2013). Similarly, transcriptome sequencing analyses of cucumber female flowers and bisexual flowers identified genes related to sex differentiation of cucumber flowers (Guo et al., 2010). In this study, we compared DEGs between the soybean four-seed pod mutant and the control at three different developmental stages using transcriptome analysis. There were 206 up-regulated genes with a log_2 ratio value greater than 3, including 103 genes with functional annotations. GO analysis showed that these up-regulated genes were involved in catalytic reactions, transcript level regulation, modulation of enzyme regulator activity, primary and secondary metabolism, tolerance responses, and other related biological metabolic pathways. Among these 103 genes, the most abundant were those encoding products related to protein binding/DNA binding/nucleotide binding (36%), followed by those involved in metal ion binding (15%). A previous study showed that the leaf photosynthetic rate and the accumulation of photosynthetic products increased in transgenic tobacco plants expressing the soybean ferritin gene (Wang, 2010). That study also showed that

the lack of ferritin genes inhibits vegetative growth and the reproductive ability of the plant. Based on the results of the present study, we tentatively suggest that the up-regulated DEGs related to protein binding, DNA binding, nucleotide binding and metal ion binding may be related to the formation of four-seed pods or yield enhancement in soybean.

In this study, eight significantly enriched metabolic pathways (Q_value ≤ 0.05) in three different growth periods were identified. These pathways included those related to amino acid synthesis and metabolism, carbon metabolism, plant hormone signal transduction, and protein processing in the endoplasmic reticulum. We conclude that these pathways may be related to the formation of four- or multiple-seed pods. These results lay the foundation for further research on the molecular mechanism of the formation of four- or multiple-seed pods at the transcriptome level.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The authors wish to thank Professor Zenglu Li (Center for Applied Genetic Technologies, University of Georgia, GA, USA) for his repair and advice. Research supported by the Education Department of Jilin province (#2015-191), the Agency of Science and Technology of Jilin Provincial Science & Technology Department (#20140204021NY and #20140101015JC), and the Campus Startup Funds of Jilin Agricultural University (#201242).

REFERENCES

Cannon SB, May GD and Jackson SA (2009). Three sequenced legume genomes and many crop species: rich opportunities for translational genomics. *Plant Physiol*. 151: 970-977.

Chan C, Qi XP, Li MW, Wong FL, et al. (2012). Recent developments of genomic research in soybean. *J. Genet. Genomics* 39: 317-324.

Cheng LB, Li SY and He GY (2009). Isolation and expression profile analysis of genes relevant to chilling stress during seed imbibition in soybean [*Glycine max* (L.) Meer.]. *Agric. Sci. China* 8: 521-528.

Fan XD, Wang JQ, Yang N, Dong YY, et al. (2013). Gene expression profiling of soybean leaves and roots under salt, saline-alkali and drought stress by high-throughput Illumina sequencing. *Gene* 512: 392-402.

Graham PH and Vance CP (2003). Legumes: importance and constraints to greater use. *Plant Physiol.* 131: 872-877.

Guo SG, Zheng Y, Joung JG, Liu SQ, et al. (2010). Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. *BMC Genomics* 11: 384.

Jain M (2012). Next-generation sequencing technologies for gene expression profiling in plants. *Brief. Funct. Genomics* 11: 63-70.

Jain M, Misra G, Patel RK, Priya P, et al. (2013). A draft genome sequence of the pulse crop chickpea (*Cicer arietinum* L.). *Plant J*. 74: 715-729.

Jung CH, Wong CE, Singh MB and Bhalla PL (2012). Comparative genomic analysis of soybean flowering genes. *PLoS One* 7: e38250.

Kenneth JL and Thomas DS (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta Ct}$ method. *Methods* 25: 402-408.

Lulin H, Xiao Y, Pei S, Wen T, et al. (2012). The first Illumina-based de novo transcriptome sequencing and analysis of safflower flowers. *PLoS One* 7: e38653.

Margulies M, Egholm M, Altman WE, Attiya S, et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.

Mortazavi A, Williams BA, McCue K, Schaeffer L, et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5: 621-628.

Ozsolak F and Milos PM (2011). RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet*. 12: 87-98.

Peng YH, Zhu JC, Yang GB and Yuan JZ (1994). Relation of soybean leaf shape distribution to 4-seeded pods. *Acta Agron. Sin*. 20: 501-503.

Sato S, Nakamura Y, Kaneko T, Asamizu E, et al. (2008). Genome structure of the legume, *Lotus japonicus*. *DNA Res*. 15: 227-239.

Schmutz J, Cannon SB, Schlueter J, Ma JX, et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178-183.

Severin AJ, Woody JL, Bolon YT, Joseph B, et al. (2010). RNA-Seq Atlas of *Glycine max:* A guide to the soybean transcriptome. *BMC Plant Biol*. 10: 160.

Soybean Growth and Development [http://www.soybeanmanagement.info].

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.

Young ND, Debellé F, Oldroyd GE, Geurts R, et al. (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* 480: 520-524.

Varshney RK, Chen W, Li Y, Bharti AK, et al. (2011). Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol*. 30: 83-89.

Varshney RK, Song C, Saxena RK, Azam S, et al. (2013). Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol*. 31: 240-246.

Wang XZ, Zhang XJ, Zhou R, Sha AH, et al. (2007). QTL analysis of seed and pod traits in soybean RIL population. *Acta Agron. Sin*. 33: 441-448.

Wang Z, Gerstein M and Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet*. 10: 57-63.

Wang SL (2010). Cloning of soybean Fettitin gene, construction of its expression vector and transformation with *Arabidopsis*. PhD thesis, Henan Agricultural University, Biotechnology Department.

Wong CE, Singh MB and Bhalla PL (2009). Molecular processes underlying the floral transition in the soybean shoot apical meristem. *Plant J*. 57: 832-845.

Wong CE, Singh MB and Bhalla PL (2013). The dynamics of soybean leaf and shoot apical meristem transcriptome undergoing floral initiation process. *PLoS One* 8: e65319.

Zhai FL (1988). Crop Quality Breeding. China Agricultural Press, Beijing.

Zhou XA, Wang XZ, Wu XJ, Cai SP, et al. (2005). Relation of three-seed and four-seed pods with yield of RIL in soybeans. *Chin. J. Oil Crop Sci*. 27: 22-25.