MOLECULAR MECHANISMS OF CROP DOMESTICATION REVEALED BY COMPARATIVE ANALYSIS OF THE TRANSCRIPTOMES BETWEEN CULTIVATED AND WILD SOYBEANS

A Thesis

by

MURAT ACI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Hongbin Zhang
Committee Members,	Steve Hague
	Joshua Yuan
Head of Department,	David D. Baltensperger

August 2018

Major Subject: Plant Breeding

ABSTRACT

Soybean is one of the key crops necessary to meet the food requirement of the increasing global population. However, in order to meet this need, the quality and quantity of soybean yield must be greatly enhanced. Soybean yield advancement depends on the presence of favorable genes in the genome pool that have significantly changed during domestication. To make use of those domesticated genes, this study involved seven cultivated, *G. max*, and four wild-type, *G. soja,* soybeans. Their genomes were studied from developing pods to decipher the molecular mechanisms underlying crop domestication. Specifically, their transcriptomes were analyzed comparatively to previous related studies, with the intention of contributing further to the literature. For these goals, several bioinformatics applications were utilized, including *De novo* transcriptome assembly, transcriptome abundance quantification, and discovery of differentially expressed genes (DEGs) and their functional annotations and network visualizations. The results revealed 1,247 DEGs, 916 of which were upregulated in the cultivated soybean in comparison to wild type. Findings were mostly corresponded to literature review results, especially regarding genes affecting two focused, domesticated-related pod-shattering resistance and seed size traits. These traits were shown to be upregulated in cultivated soybeans and down-regulated in wild type. However, the opposite trend was shown in disease-related genes, which were down-regulated or not even present in the cultivated soybean genome. Further, 47 biochemical functions of the identified DEGs at the cellular level were revealed, providing some knowledge about the molecular mechanisms of genes related to the two aforementioned subjected traits. While our findings provide valuable insight about the molecular mechanisms of soybean domestication

attributed to annotation of differentially expressed genes and transcripts, these results must be dissected further and/or reprocessed with a higher number of samples in order to advance the field.

ACKNOWLEDGEMENTS

I would like to thank my committee chair and my advisor, Dr. Hongbin Zhang, my committee members, Dr. Steve Hague and Dr. Joshua Yuan, the head of department, Dr. David D. Baltensperger, for their guidance and support during my master's program in the major of Plant Breeding in department of Soil and Crop Science at Texas A&M University.

Thanks to my colleagues and seniors at Dr, Hongbin`s group, Meiping Zhang, Delin Xu, Yunhua Liu, Mustafa Cilkiz, Mehmet Dogan, who helped me to accomplish the program of Master of Science. Also, to my fellows, and friends, Nese Coskun from Department of Entomology, and Grace Samtani from Institute for Neuroscience for helping me to improve my knowledge and enhance my motivation during my major.

Thanks to my government for giving me a very important chance to get educated and to develop my knowledge as a plant breeder and to improve my personality by living in other country in the U.S.A.

Lastly, I deeply thank to my mother, father, sisters and brothers and my grandmother for supporting me emotionally and morally as anchors during my education.

CONTRIBUTORS AND FUNDING SOURCES

This work was edited and supervised by a thesis committee including Professor Hongbin Zhang and Associate Professor Steve Hague from department of Soil and Crop Science and Professor Joshua Yuan from department of Plant Pathology and Microbiology at Texas A&M University.

Additionally, all work written and done in this thesis was completed by the student under the advisement of Dr. Hongbin Zhang and great support of Delin Xu, Meiping Zhang and Yunhua Liu.

Graduate study was completed with a governmental scholarship supported by the Ministry of Food Agriculture and Livestock of Republic of Turkey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF GRAPHS

1. INTRODUCTION

The world's population is expected to be around 9.5 billion by the 2050s, which is nearly 2.3 billion more than the world's population today (Alexandratos and Bruinsma, 2012). Correspondingly, global food demand is projected to be roughly doubled and global food production must increase by 70 percent in the 2050s, as estimated in 2009 (Fao, Global agriculture towards 2050, 2009). Maize, rice, wheat and soybean are producing nearly 60% of global agricultural calories. Their production is currently getting higher at a rate of 1.0% - 1.5%, which is much lower than the expected 50% - 60% necessary to meet the projected food demand in 2050 (Deepak *et al*., 2013). Therefore, crop productivity needs to be increased by developing genetically super cultivars with favorable genes and continuously improving agronomy. For crop genetic improvement, the genomic diversity of breeding material or gene sources is very important to obtain or improve the targeted traits with gaining more favorable genes. However, most of the cultivated crops, such as soybean, have lost a significant number of favorable genes due to the intensive natural and artificial selection for elite cultivars during their domestication and cultivation. Therefore, the quantity of favorable alleles for food production has diminished in crops. Due to intensive artificial selection and cultivation, i.e., domestication, it has resulted in a serious genetic bottleneck effect on the crop species, and their genetic variation has decreased sharply (Hongye and Marilyn, 2004; Guo *et al*., 2010). Hence, efforts are needed to discover novel gene sources to transfer their favorable alleles and genes to elite cultivars. These favorable/desirable alleles and genes can be enriched by artificial selection and

introduced into elite cultivars by genetic recombination through plant breeding assisted with modern molecular technologies such as marker-assisted selection and genetic engineering (Cobb *et al.,* 2013; Park *et al*., 2015).

Soybean, *Glycine max* (L.) Merr., is one of the major crops suffering from lack of genetic diversity or favorable alleles in its genome. Domestication of soybean and intensive plant breeding applications and dramatic soybean genome modifications with the aim of meeting fastidious human demand have resulted in a severe genetic bottleneck (Guo *et al*., 2010; Li *et al.*, 2010; David *et al*., 2006). As a result of that case, numerous unique and important sequence variants, genetic diversity and desirable genes in the soybean genome have lost or decreased sharply (David *et al*., 2006; Teresa and Gunter, 2013; Hymowitz *et al*., 1980; Guo *et al.,* 2010; Tang *et al*., 2010). Studies have showed that wild soybean, Glycine soja Sieb & Zucc, still maintains many of those unique favorable genes and thus, this makes wild soybean species a valuable and a desirable source of necessary alleles for continued soybean genetic improvement (Li and Olsen, 2016; Lam et al., 2010). In the light of this information, studies have been pursued on both cultivated soybean (Schmutz *et al*., 2010), its wild relative (Kim *et al*., 2010), and their comparative analysis (Kim *et al*., 2012; Stupar, 2010; Xinpeng *et al*., 2014) to reveal the genomic changes and differences which have occurred over time between the two types of soybeans and to understand phenotypic and morphological effects of genetic loss on domesticated soybean, *G. max*. These genomic discrepancies have caused both types of soybean to have dramatically diverged in morphology and physiology, especially in regard to developing pods, seed size and seed

shattering. Therefore, these agronomic traits have been the main interest of this thesis. To understand these diversifications and their effects on soybean, it is necessary to decipher the molecular mechanisms underlying soybean domestication. The results of this research will provide valuable knowledge about candidate genes for genetic loss or differentiation and their agronomical importance on plants during crop domestication (Teresa and Gunter, 2013).

This study aimed to provide a deeper insight into mechanisms underlying crop domestication by comparatively analyzing the transcriptomes of cultivated soybean, *G. max*, and its wild progenitor, *G. soja*, using modern DNA sequencing methods and large-scale data analysis bioinformatics tools. To achieve these goals, this study has accomplished the following steps: (1) sequencing the transcriptomes of developing pods of seven and four germplasm lines representing the transcriptomes of developing pods of cultivated and wild-type soybean, respectively; (2) comparatively analyzing the transcriptomes of both types to discover differentially expressed genes in the developing pods and visualize their co-regulation and networks; (3) identifying and comparatively analyzing the candidate genes controlling agronomic traits crucial for soybean domestication mechanism; and (4) formulating the molecular mechanisms of soybean domestication by integrating the results obtained from above analyses.

1.1 Evolutionary and Domestication History of *G. max* and *G. soja*

The soybean genus, Glycine Willd., is a member of the Fabaceae (Leguminosae) family, the second largest family of flowering plants consisting of 650 genera and 18,000–20,000 species (Doyle and Lucknow, 2003). The species of the Glycine genus have 2n =40 chromosomes except *G. hirticaulis*, *G. tabacina* and *G. tomentella* (Brown *et al.*, 1987; Singh

et al., 1987b; Nguyen et al., 2017). This genus is further classified into two subgenera, *Glycine* and *Soja* (Moench) F.J. Her. The *Glycine* subgenus has at least 25 species, 16 of which are perennial and the rest annual (Raza *et al.,* 2016). The *Soja* subgenus contains only two species, the wild soybean, *G. soja* Sieb & Zucc and the cultivated soybean, *G. max* (L.) Merrill. Before these two species existed and originated from a common annual subgenus soja, the *G. soja/ G. max* complex has occurred as a result of two genomic duplications that occurred around 59 and 13 million years ago, respectively (Schmutz *et al.,* 2010). Further, genetic divergence began around 0.267 ± 0.03 million years ago between *G. max* and *G. soja*, resulting in genetic loss or novel genetic gain, and a wide range of chromosome rearrangements (Lam *et al*., 2010; Eric *et al.,* 2016). As a consequence of genomic differentiation, *G. max* has diverged from *G. soja*, and thus it is accepted that *G. soja* is the ancestor of the cultivated soybean (Kim *et al.,* 2010; Lam *et al.,* 2010). The *Soja* subgenus is the most diverse type among soja plants. It originated in northeastern China (Fukuda, 1933) and is the progenitor of *G. max* and the cultivated soybean. *G. max* has been known to be domesticated 6000 - 9000 years ago in the central China (Zhao and Gai 2004; Carter *et al.,* 2004; Kim *et al.,* 2010) and cultivated approximately 5,000 years ago in and around China (Hymowitz and Shurtleff, 2005).

1.2. Importance of Soybean

Soybean is a major world-wide grain legume crop. It is known as a high-plant protein and oil crop, with an attracting ability of nitrogen-fixation from nature. Therefore, the soybean harvested area in the world has increased from 100 million hectare (ha) to over 120 million ha and production has increased from 225 million tonnes to 335 million tonnes between 2006 and 2016. The most important soybean producers in 2016 are the United

States, Brazil, Argentina, China, with 103.4, 103, 57, 12.2 million tonnes, respectively (http://statistics.amis-outlook.org/data/index.html.). Because global demand for crop production is estimated to be doubled by 2050, crop yield, including soybean, needs to be increased at a rate of at least 2.4% per year. Although soybean is consumed by humans and animals at an increasing rate of around 56% globally as oil-seed production, the current soybean yield increase rate is only around 1.3% per year, which is clearly not enough to meet future estimated global yield demand (Zhou *et al*., 2014).

By 2050, crop production demand globally for all uses is predicted to rise by approximately 84% and the demand only for biofuel use is estimated to increase by 86%. Soybean yield acreage is needed to increase by 69%, while corn is needed to increase by 23% for both global food and feed demand. In the 2050s, the required world production of soybean is estimated to be about 150% higher than today's. Because the world-wide agricultural harvested area cannot be increased as fast as food demand is, the yield growth demand per year, without increasing production area, is necessary to be met for crops, including soybean (Lavlu, 2012; Dixon *et al*., 2001).

Soybean also has numerous important recognized effects on human health. It contains approximately 5 mg of isoflavones per gram of dry weight grain (Linus Pauling Institute, 2016), which is a potentially beneficial factor for cardiovascular diseases (Sacks *et al*., 2006). The isoflavones abundant in soybean seeds have been shown to decrease the risks of high blood cholesterol levels (Sacks *et al*., 2006, Qin *et al.*, 2013), respiratory infection or prostate cancer (Nutrition and Allergies, EFSA, 2011), and breast cancer (Linus Pauling Institute, 2016). The phytic acid contained in soybean acts as an antioxidant that is effective in preventing human diseases. Furthermore, the fatty acids of soybean reduce

cancers (Vucenik *et al.,* 2003), minimize diabetes (Yoon *et al.,* 1983), and prevent inflammation (Sudheer *et al.,* 2004).

Soybean, *G. max,* has a number of different kinds of uses because of its high protein (~40%) and high quality and amount of oil content (~20%). Soybean provides more than half of the global oil seeds, thus comprising 30% of the oil and ~70% of the proteins in a human diet (Lam *et al*., 2010). Approximately 85% of the world`s soybean is used (Soyatech, 2017) as livestock feed because of its high protein and oil content, and for human consumption in the form of soybean meal (Schmutz *et al.,* 2010). Soybean is also a very important oil source for biodiesel production. Approximately 80% of the domestic biodiesel production of the United States has been met by soybean (Natural Biodiesel Board, 2008).

Finally, soybean plays a significant role in the establishment of a sustainable agriculture system because it fixes nitrogen from the atmosphere to soil with its symbiotic microorganisms called rhizobia (Schmutz *et al*., 2010; Chung *et al*., 2014).

1.3. Genetic Diversity Between *G. max* and *G. soja*

Since the initial domestication and cultivation of soybean, *G. max*, has diverged gradually and dramatically from its wild ancestor, *G. soja*. In this process, numerous genomic changes have occurred, resulting in divergence in a number of morphological and physiological characteristics, otherwise known as domestication syndrome (Silvas., 2015; Liu *et al*., 2007; Wang *et al*., 2016). Genes differentiated after domestication of *G. max* are called ‘domestication genes’ and traits affected by those genes are called ‘domestication-related traits’ (Doebley *et al*., 2006). Domestication-related genes are those significantly affecting domestication-related traits, such as seed size, weight and color, plant height,

growth determinacy, flowering and maturity time, lodging (Zhou *et al.,* 2015), seed hardiness and pod dehiscence that causes pod shattering (Chachalis and Smith, 2000). In dicotyledonous plants like soybean, there is a dehiscence zone called ventral suture on seeds. The cells in this zone have a major role in seed shattering mechanisms because the dehiscence of the zone is dependent on adhesion force between those cells. The more gravitation between cells occurs, the less dehiscence will a plant pod have. Pod shattering results in shattering of ventral sutures on the pod. The result of pod shattering is dispersion of seeds, which is an effective method for seed propagation, especially in many wild-type species (Dong *et al.,* 2017). While some crop domestication traits are controlled by a small number of genes, qualitative traits, (Li *et al*., 2013), most of them, especially those related to seed quality, are quantitatively inherited and controlled by a relatively larger number of genes (Schmutz *et al.,* 2010).

During its cultivation, selection pressure has been carried out on soybean and the genetic diversity remaining in the genomic pool of cultivated type has been gradually reduced. This artificial selection pressure is called selective sweep, the hitchhiking effect or genetic draft, which is the intensive selection of plants according to desirable traits. Since soybean has been exposed to a genetic bottleneck, genetic differences exist between the cultivated and wild soybeans (Guo *et al*., 2010; Davit *et al*., 2006).

1.3.1 Morphological Differences

As a result of genomic variation and differentiation, the expression profiles of genes have changed, and that differentiation has given rise to numerous morphological differences and ultimately, specification (Carrol, 2008; Romeo *et al*., 2012). Generally, *G. max* has large shiny yellow seeds, while its wild relatives have small black seeds (Zhou *et al.,* 2015).

One of the most important morphological chances, often considered to be a domestication trait, is the seed size/weight of soybean. The seed weight of domesticated soybean has increased by approximately 15 times, relative to that of wild-type seeds. The seed size of *G. max*, which is one of the most important agronomical concerns, is much bigger than that of *G. soja* (Chen and Nelson, 2004). The number of seeds produced by the cultivated soybean has diminished by 8.2 times and the seed weight per plant of the cultivated soybean is roughly >32% higher (>9 g per 100 seeds) than wild-type soybean seed weight (<3 gram per 100 seeds). Moreover, despite the fact that the shape and color of pods of both types are almost the same and the number of seeds per pod has not changed, the cultivated soybean has 4.7 times bigger pods than the wild soybean. In plant structure, while *G. soja* has never-erect stem and low small stem ratio, *G. max* has a relatively thicker and erect stem (Chen and Nelson, 2004). The leaf surface area of the cultivated soybean is 2.6 times greater than that of the wild soybean. On the other hand, the wild soybean is much taller and has a much higher number of branches than the cultivated soybean (Steven *et al*., 1980; Silvas *et al.*, 2015). The root length and volume of the cultivated soybean has increased by at least 25%, relative to those of the wild-type (Silvas *et al.*, 2015). Even though both cultivated and wild soybeans have the same shape and color of flower, purple, *G. max* may also have a white colored flower (Steven *et al*., 1980). Moreover, another prominent domestication trait is seed shattering. *G. soja* has a severe shattering before plant maturity to promote long-distance seed dispersal. This highly undesirable trait gives rise to a considerable amount of yield loss but this dispersal capability is mostly removed from the cultivated types of soybean and other crops (Teresa and Gunter, 2013).

1.3.2. Physiological (Seed Ingredient or Chemical) Differences

In addition to their phenotypic differences, the cultivated and wild soybeans have seed ingredient differences. *G. max* seeds consist 41.0% of proteins, 20.0% of fat including 3% of saturated fat, 5.3% of ash, 2.7% – 3.9% of crude fibers and 25.0% of carbohydrates (Nutrient Data Laboratory, USDA, 2016). These nutritional values were much lower for the *G. soja* seeds (Raboy *et al*., 1984), as their seed oil content is only 8% (Trupti *et al*., 2013). As a daily value, 100 grams of the cultivated soybean seeds comprise proteins (36%), dietary fibers (37%), iron (121%), manganese (120%), phosphorus (101%) and several B vitamins, including folate (94%). The chemical components of the cultivated soybean seeds include vitamin K, magnesium, zinc and potassium. Among the staple foods consumed by humans, such as wheat, corn, rice, potato, cassava, sweet potato, yam, sorghum, plantain, etc., *G. max* seeds have the highest content of proteins, fat, calcium, vitamin C, thiamin B1, folate total (B9), saturated fatty acids, monounsaturated fatty acids, and polyunsaturated fatty acids. Soybean seed oil is also a good source of omega-6 and omega-3 with a ratio of 7:1 (Nutrient data laboratory, USDA, 2016). In comparison, *G. soja* seeds are much lower in oil (<130 mg) and oleic acid (<140 mg) contents but have a higher linoleic acid (<140 mg) concentration. *G. max* has higher oil (>185 mg), higher oleic acid (>190 mg) and lower linoleic acid concentration (<95 mg) (Chen and Nelson, 2004).

1.3.3. Genomic Differences

Although the cultivated soybean has gained numerous favorable and desirable alleles, genes or traits during domestication (Chen and Nelson, 2004; Teresa and Gunter, 2013), it has lost many rare sequence variants, single nucleotide polymorphisms (SNPs), and nucleotide diversity (Li *et al.,* 2013), and undergone changes of frequency of numerous

alleles (Hymowitz *et al.*, 1980). Lack of allelic diversity has been established in the cultivated soybean genome throughout its domestication history (Zhou *et al*., 2015; Guo *et al.*, 2010; Tang *et al*., 2010). Furthermore, the wild soybean has retained a high level of crucial allelic pool. The unique alleles retained in the wild soybean make it extremely valuable for genetic improvement of the cultivated soybean because these alleles are vital for gaining agronomical traits important to the modern soybean and restraining genetic loss in the cultivated type. From agronomical perspectives, both types of soybeans are naturally self-pollinated, but they can occasionally become cross-pollinated at a ratio of 0.04% - 4.52% (Ray *et al.*, 2003). Because the wild and cultivated soybeans have no sexual barrier to hybridize with each other and their chromosomes exhibit normal meiotic chromosome mating, they are capable of generating fertile hybrids (Kim *et al*., 2010). For these reasons, *G. soja* is a fundamental gene resource to enhance the allelic pool of *G. max* through introgression of beneficial alleles from the wild-type to the cultivated soybean (Li *et al.*, 2013). Thereby, the wild-type soybean can maintain and enhancing the cultivated soybean genome by providing desired alleles. This reveals the need for the wild soybean to be protected and paid attention to (Lam *et al*., 2010).

The soybean genome has been extensively studied to quantify the genetic variation between genotypes of interest using molecular markers (Chen and Nelson, 2004) and by QTL mapping (Liu *et al.*, 2007). After the first sequence of *G. max* (Schmutz *et al*., 2010), as well as that of *G. soja*, (Kim *et al.*, 2010) became available in 2010, comprehensive comparative genomics studies have been conducted on both the cultivated and wild-types of soybean (Trupti *et al.*, 2013). Comparative genome analysis between the two species has swiftly accelerated, especially with the recent development of high-throughput sequencing

(HTS), or next generation sequencing (NGS) technologies, to discover genomic differences between the cultivated and wild-type soybeans. It has been reported that 81% of the rare alleles and 50% of genetic diversity has been lost in the cultivated soybean and further, 60% of the genes have had substantial allelic frequency changes throughout soybean domestication (David *et al.,* 2006). Although elite soybean cultivars have maintained 72% of the sequence diversity, they have lost almost 80% of rare alleles with a frequency of ≤ 0.10, relative to the Asian landraces (David *et al.,* 2006). In another comparative genomic study, it has been seen that the sequence difference between the two species was 3.76%, including 0.267% substituted bases (SNPs), 0.043% of inserted/deleted bases (InDels) and 3.45% of large deleted sequences in the *G. soja* genome. Paired-end sequence alignment of the *G. soja* genome against that of its cultivated relative revealed 5,794 deletions, 194 inversions and 8,554 insertions in the *G. soja* genome. Moreover, 712 presence-absence variation (PAV) genes were found between these two genomes.

Additionally, it was observed that the *G. max* genome was devoid of 28 disease resistance or metabolism-related genes that were found in the *G. soja* genome (Kim *et al.*, 2010; Kim *et al*., 2012). Some differences of genes have led the cultivated soybean to gain important traits, such as increased size of seeds, higher oil and protein contents, higher resistance to pod shattering, etc (Li and Olsen, 2016). However, due to the divergence between the two species at the genomic and allelic levels, some genes conferring tolerance to dehydration stress (Chena *et al.,* 2006), disease and pest resistance (Kim *et al.,* 2012), salt tolerance (Lam *et al.,* 2010; Xinpeng *et al*., 2014), chilling and dehydration stress (Chen *et al*., 2006) and a high lutein content (Kanamaru *et al.,* 2006) have been observed in the wild soybean, but are absent or much fewer in its cultivated relative. These absent genes

can be readily transferred into *G. max* through traditional or molecular breeding, because the wild and cultivated soybeans can readily hybridize with each other and exhibit normal meiotic chromosome pairing. For that reason, the wild soybean is a great resource of alleles and novel genes for cultivated soybean improvement (Stupar, 2010).

1.4. Motivation of the Study

To help feed the growing global population, more endeavors are needed to improve the quality and quantity of agronomical plant species such as crops (Deepak *et al*., 2013). Soybean is one of the major crops feeding the world (Zhou *et al*., 2014). During domestication, the soybean genome has duplicated, having resulted in loss and mutation of numerous genes and alleles. Because of the genetic bottleneck caused by intensively artificial selective sweeps, the cultivated soybean has differentiated genomically and morphologically from its wild ancestor, *G. soja*. As a consequence, the allelic structure and frequency of the cultivated soybean has changed extremely (David *et al*., 2006). Although those allelic chances have allowed the cultivated soybean to gain some agronomically important traits, like better seed quality, they have also given rise to loss of some alleles contributing to some high-quality characters, such as disease resistance (Kim *et al*., 2012). To regain and improve those high-quality traits in the cultivated soybean, especially pod shattering, which is one of the main reasons to lose yield, and increased seed size, which is a very favorable trait, we need to enhance or transfer those crucial and favorable alleles and genes from the wild-type to the cultivated soybean. This process could be enhanced and accelerated by deciphering the molecular mechanisms of soybean domestication, especially those differentiating the domestication traits. Although some efforts have been

made to understand those molecular mechanisms, additional studies are still clearly needed (Lam *et al.,* 2010).

In this study, we aimed to provide a deeper insight into the molecular mechanisms of soybean domestication by comparatively analyzing the developing pod transcriptomes of a selection of germplasm lines representing the wild and cultivated soybeans. Transcriptome comparative analysis has been used for identification of genes underlying domestication-related traits (Daniel *et al.,* 2013; Olsen and Wendel, 2013; Yoo and Wendel, 2014). Because the seed size and pod shattering traits are two of the most important domestication traits (Song *et al.,* 2007; Shomura *et al*., 2008), we have focused in our study on those traits affected by domestication-related differentially expressed genes between wild and cultivated soybean. Furthermore, we have examined the molecular divergences between the wild and cultivated soybeans based on the developing pod transcriptomes, which have provided new knowledge for improving genetic structure of soybean and enhancing favorable allelic frequency and diversity in the cultivated soybean. Gene expression differentiation is known to play a prominent role in morphological divergence and ultimately, specification of plants (Carrol, 2008; Romeo *et al*., 2012).

2. LITERATURE REVIEW

There is no doubt that one of the most revolutionary explorations has been the discovery and sequencing the hereditary material of living creatures. Researchers aspired to find out the structure and sequence of DNA and RNA, and have been successful (Watson and Crick, 1953; Holley *et al*., 1965). Following such successes, some different sequencing techniques have been developed to make sequencing faster and cheaper (Sanger *et al.*, 1965; Maxam and Gilbert, 1977). Sequencing was done manually until the 1990s, beginning in the 1950s when first generation automated DNA sequencing machines were developed (Ansorge *et al.*, 1987; Prober *et al.*, 1987; Swerdlow and Gesteland, 1990). These machines enabled the sequence of genomic materials to be much longer than that of sequences done manually (Staden, 1979). This technology allowed the speed and quantity of sequencing to accelerate sharply and the workload and cost of sequencing to decrease dramatically, especially after second generation sequencing technologies were launched (Shendure and Ji, 2008). Shortly later in 2013, third generation machines were produced (Eisenstein, 2012), and it is predicted that development of sequencing technologies will continue to advance.

After using RNA-seq methods, high-throughput sequencing technologies have been especially successful, thriving on mapping and quantifying transcriptomes (Marioni *et al.*, 2008; Wang *et al.*, 2008; Xia *et al*., 2009). These machines used RNA extracted from organisms of interest as input to construct RNA-seq libraries. These libraries have been used in computational applications to dissect the genome in order to get deeper knowledge about molecular mechanisms of organisms (Garber *et al*., 2011).

To sequence RNA-seq libraries, the first approach involves genome-guided alignment, which is based on a reference genome taken from a model organism. This approach has quickly become a standard method for transcriptome studies (Haas *et al*., 2013). Nonetheless, the need for studying sequencing on non-model genomes caused the de novo method to be developed in 2008. To identify allelic diversity in wild soybean, the method of genome alignment against a reference genome has been used (Lam *et al.,* 2010). On the other hand, assembling has been done using the de novo sequencing method (Wang *et al*., 2008; Xia *et al.,* 2009). Comparative analysis studies have shown that results of both reference-based alignment and de novo assembly method were almost the same (Lam *et al.,* 2010). The number of SNPs discovered using both sequencing methods in wild soybean genomes was around 95-99%. In both methods, the amount of SNP was found to be 35% in the wild-type genome and 5% in cultivated soybean genome, and those SNPs have been found to be highly effective in the genes they are found in. It can be deduced from these results that wild soybean genome has much more allelic diversification than cultivated soybean genome. Additionally, in Lam`s research, 856 annotated genes were isolated from two different species of soybean using the de novo method, and it was seen that 40% of those genes were involved in metabolic and catalytic process in the cell. Further, it was inferred that half of annotated resistance-related genome sequences and 28 cellular metabolism-related genes were absent in all cultivated accessions (Lam *et al.,* 2010).

Plant domestication mechanisms have taken much attention amongst plant breeders and plant molecular breeders (Teresa and Gunter, 2013). Deciphering molecular mechanisms of crop domestication benefitting from high-throughput sequencing technologies and comparative bioinformatic analysis of transcriptomes taken from plants

of interest increasingly helps those improvements. During domestication, plants have had different characteristics, known as domestication syndrome, including seed shattering/size/color and plant height, flowering and maturing time, determinacy, lodging, number of leaf and branch etc. To enhance our knowledge about soybean domestication mechanisms, especially about two domestication traits including pod dehiscence and seed size/weight, we have examined previous surveys applied on comparative soybean mechanisms.

Seed shattering or the pod dehiscence trait is one of the most important traits which plants selected against during domestication. This trait has been interesting for researchers, so there have been some studies to decipher pod dehiscence mechanism. One of those studies has found quantitative trait loci (QTL) in the soybean genome, which has a high-level control on the *qPDH1* trait (Suzuki *et al.,* 2009). *qPDH1* was understood to be a sort of protein coding locus, having a function of lignin biosynthesis and providing lignin deposition on pod walls, which results in the forcing of pod walls to dehiscence. This protein in cultivated soybean is transformed into a non-functional gene by a stop codon in *pDH1* loci in order to have indehiscent pods (Funatsuki *et al.,* 2014). This phenomenon is mentioned as one of the mechanisms of pod dehiscence trait (Dong and Wang, 2015). In rice, the *Shattering4 (Sh4*) gene has been found to be associated with the shattering trait. This gene is effective to enhance pod dehiscence in wild rice, and thus, the expression of shattering-related allele is decreased in cultivated rice to decrease shattering of seeds (Li *et al.,* 2006). Annihilation of only a single nucleotide polymorphism (SNP) in the *qSH1* region results in reduction of shattering gene expression in cultivated rice. *SH5* is another gene which is homologous to *qSH1* found in Oryza sativa genome and some rice species

(Li *et al.,* 2013). That gene also causes reduction of the expression of shattering by promoting abscission layer occurrence. *SH5* also favorably regulates *Sh4* to promote development of the abscission layer on the pod of soybean (Yoon *et al.,* 2014). The abscission layer is the main physical cause of grain separation from the pedicel (Li *et al.,* 2006). It is affected by the *Sh4* gene, which positively regulates *SHAT1* expression in the abscission layer (Zhou *et al.,* 2012). In other QTL study, the *Sh4* allele of wild species has been observed to cause shattering dominantly in rice (Li *et al.,* 2006). In sorghum, the *Sh1* gene manipulates the expression of the abscission layer, contributing to seed shattering. This gene is also found to affect shattering-related genes in other cereals, such as in maize and rice (Lin *et al.,* 2012). The other gene controlling cell wall biosynthesis in the abscission zone is *SpWRKY* gene, which has a negative effect on genes controlling cell wall biosynthesis in Sorghum propinquum. However, the *SpWRKY* gene was not artificially selected against pod dehiscence trait during cultivation (Tang *et al.,* 2013).

Pectin, found primarily in the cell wall, is a high pod shattering-related carbohydrate (Braidwood *et al.,* 2013). Because pectin and cellulose are highly important components of the cell wall (Jung and Park, 2013; Muriira *et al.,* 2015) and the hydrolases enzyme has negative affect on pectin formation, they play an important role in cellular metabolism controlled by genes related to the pod dehiscence trait. Analysis of GO and KEGG metabolic enrichment pathways of some unigenes (DEGs) encoding hydrolase enzymes and cell wall modifications in a study (Dong *et al.,* 2017), showed that these genes break down glycosidic bonds of pectin and cellulose in cell walls and result in dehiscence of the cell wall in the ventral suture of the pod, resulting in the high tendency for pod- shattering. Thus, they were upregulated in shattering-susceptible Vicia sativa species and highly

involved in hydrolase activity and carbohydrate metabolic process. Dong`s findings showed that many shatter-related hydrolases were enriched in the "extracellular region," "membrane," and "cytoplasm" in the cellular component terms. On the other hand, expansin non-enzymatic protein families, the same family the pectin protein is in, has been dissected by Muhammad et al. (2017), Xinxin et al. (2015) and Cosgrove (2015), in soybean root. Because expansin causes cell wall stress relaxion (McQueen and Cosgrove, 1995), it helps plants to be resistant against cell wall degradation.

Seed size/weight is the other primary domesticated trait and a main factor controlling yield potential in soybeans (Song *et al.,* 2007; Shomura *et al.,* 2008). Some researchers have taken successful steps to investigate the molecular mechanisms of seed size/weight of plants, especially after NGS techniques were used. In a study with paired-end reads using Illumina platforms, the *Glyma03g35520.1* gene has found most likely to contribute to seed size/weight in soybean (Li *et al.,* 2013). This gene is known to be orthologous to the *Filling1 (GIF1)* gene found in rice (Ross-Ibarra, 2005). This domestication gene plays an important role in carbohydrate metabolism, resulting in grain development and higher seed weight in rice (Wang *et al.,* 2008). Furthermore, the Terminal *Flower 1 (TFL1)* gene found in sunflower was also orthologous to the *GIF1* gene (Blackman *et al.,* 2011; Li *et al.,* 2013). Further, previous research has revealed that the *WRKY* gene family is involved in some biological functions in Arabidopsis (Vanderauwera *et al.,* 2012) and in Tobacco (Yu *et al.,* 2012). Due to *SoyWRKY15* gene regulation of seed size, pod and seed development, and its higher expression level during pod development, we can deduce that it is clearly a domestication-related trait (Yongzhe *et al*., 2016). These genes have been supported in other research; it was found that *GsWRKY15a* in the *G. soja*

genome and *GmWRKY15a* in the *G. max* genome, which were members of *WRKY*-like gene family (*SoyWRKY*), have been observed to be differentially expressed in both species during pod development. The plant *WRKY* protein family has control over some biological processes including immune response, developmental processes and abiotic stress (Rushton *et al.,* 2010), seed development (Sun *et al.,* 2003; Luo *et al.,* 2005) and embryogenesis (Alexandrova and Conger, 2002; Lagacé and Matton, 2004). The *GsWRKY15a* gene is related to seed size variation in wild-type soybean. The expression level of the *GmWRKY15a* gene in *G. max* was higher than *GsWRKY15a* gene in *G. soja*. These research findings suggest that the *SoyWRKY15a* (*GmWRKY15a* and *GsWRKY15b*) orthologous genes play an important role in the improvement of seed/pod size and seed development.

Soybean also includes the *P450/CYP78A* gene family, which greatly contributes to seed and pod size development as in Arabidopsis (Adamski *et al.,* 2009; Fang *et al*., 2012). During soybean cultivation, the *GmCYP78A10* gene underwent selective sweeps. While the G*mCYP78A10a* allele is found in *G. soja* significantly more than in *G. max*, the

GmCYP78A10b allele is present in *G. max* in a much greater amount compared to that of in *G. soja*. Furthermore, seed weight variety with *GmCYP78A10b* is slightly higher than that of with *GmCYP78A10a*. ANOVA analysis of seed weight, width and thickness index of soybean showed that *GmCYP78A10b* is significantly more effective than *GmCYP78A10a*. Compared to the effect of the *GmCYP78A10a* gene, the genetic effect of *GmCYP78A10b* is measured to be 7.2% higher in terms of seed weight but 5.8% lower in terms of in pod number. Since the *GmCYP78A10b* allele has been the main selection criteria during soybean cultivation, the seed weight of the cultivated soybean has been higher, while pod number got slightly lower. Therefore, it has been deduced that there is a negative correlation between seed weight and pod number (Xiaobo *et al.*, 2015).

A *BIG BROTHER (BB)* allele was found to encode a novel E3 ubiquitin ligase, which negatively regulates organ size of Arabidopsis, primarily the size of leaves and petals (Disch *et al.,* 2006). The *QTL BB* allele is located in the locus tagged with *AT3G63530* in the genome. Another gene having an effect on ubiquitin activity in Arabidopsis is the *DA1* gene, which is orthologous to the *BB* gene (Li *et al.,* 2008). The *DA1* gene has more of an effect on seed size, rather than petal or another organ size. This gene encodes the ubiquitin receptor, which restricts the period of cell proliferation that causes smaller organ size, like the *BB* gene does. To prove this claim, Li and colleagues applied a mutation on the *DA1* gene to make it the *da1-1 (EOD1)* gene so that it could be observed whether the period of cell proliferation was limited to decreased seed and some other organ size. The result of down-regulation of *DA1* gene expression has resulted in the phenotypic enhancement of both seed and organ size. (Li *et al.,* 2008). Additionally,

another research study (Liang *et al.*, 2014) reported that *Ubiquitin-specific Protease15 (UBP15)* has a positive effect on seed and plant size in Arabidopsis. *UBP15 i*s encoded by *Suppresor2 of DA1 (SOD2)* gene; overexpression of this gene increases seed size. This gene`s QTL is tagged with *AT1G17110.*

Zhao has revealed that up-regulation of *GmCYP78A72* gene expression in soybean and Arabidopsis contributes to bigger pea and bean size. However, research has shown that the *CYP78A* gene family has interaction with each other. It has been observed that *GmCYP78A57* and *GmCYP78A72* have similar expression because of genome duplication millions of years ago. For that reason, silencing of only the *GmCYP78A72* gene was not shown to have a negative effect on seed size since other *CYP78A* family members, particularly *GmCYP78A57,* equally affected seed size (Zhao *et al.,* 2016). From this, we can conclude that in order to enhance seed size, we must consider enhancing of at least two genes of that gene family.

It has been seen in Liangfa`s research that the *BIG SEED1 (BS1)* gene is one of the major genes negatively controlling the size of lateral organs such as leaves, seed pods, and seeds. Two orthologs of *BS1* were identified in the soybean genome named *GmBS1 (Glyma10g38970)* and *GmBS2 (Glyma20g28840).* After down-regulating expression of those *BS1* genes using an artificial microRNA, leaf, pod and seed size was increased around 50% in soybean and likely also in other legumes.

3.3. METHODOLOGY

3.1 Materials

3.1.1 Plant Materials

Eleven soybean germplasm lines (Table 1) requested from the USDA Soybean Germplasm Collection, Urbana, Illinois, were used for this research, with seven representing the cultivated type of soybean and four representing the wild-type of soybean (Table 1). Of these 11 germplasm lines, William 82 is included as the reference because it has been sequenced as the genetic model of soybean genetic and genomic research. All plant materials were planted in a greenhouse for pod tissue sampling for transcriptome analysis and domestication-related trait phenotyping.

Table 1 Information of germplasms used for data analysis is shown in this table.

Sample ID	Accession	Name	Seed Source	Species
Cultivated Soybean	548655	Forrest	11S-8005	*G. max*
	291312	N/A	10U-7197	*G. max*
	437356	Ussurijscaja 19	06U-3485	*G. max*
	518671	**William 82**	**11U-31681**	***G. max***
	408342	N/A	03S-5163	*G. max*
	567452	Huang Mei Dou	05U-1649	*G. max*
	567361	Lu Fang Huang	05U-1640	*G. max*
Wild Soybean	407027	N/A	97Sg-68	*G. soja*
	423995	N/A	99Uc-1078	*G. soja*
	407275	N/A	98Uc-173	*G. soja*
	483463A	N/A	98Uc-316	*G. soja*

3.1.2 Plant Growing and Sampling

The germplasm lines were planted in two-gallon pots, with each line planted in five pods and each pot having two plants. The plants were grown at 24 - 28˚C with a 14 h light and 10 h dark cycle. Fertilization has been applied, as necessary, during plant growth and development (Jack`s Fertilizer, 20:20:20 general purpose). Because pod shattering and seed size variations are crucial to soybean domestication (Dong *et al.*, 2014; Epimaki *et al.*, 1996), developing pods that are significant for these traits were sampled 14 days after flowering, frozen in liquid nitrogen immediately and then stored at -80^{o}C for transcriptome sequencing. The phenotypes of leaf length and width, pod length and width, number of pods per plant, number of seeds per pod, and seed size/weight were measured.

Seed weights of each line have been obtained. Ninety-nine leaves have been collected as measurement samples, nine for each germplasm, and their height and width were recorded. Three of nine leaves from the top, three from the middle and three from the bottom side were collected from each germplasm.

3.2 Methods

3.2.1. RNA Extraction, RNA Qualification and RNA-seq Library Preparation

We took approximately 100 mg of the frozen developing pods of each sample to extract total RNA. The pod tissues were first ground into a fine powder in liquid nitrogen using a mortar and pestle. Then, total RNA was isolated using the Spectrum™ Plant Total

RNA Kit. The DNA contaminants, if any, in the RNA were removed by digestion with the On-column DNase Digest Set as described by the manufacturer (Sigma Aldrich, St. Louis). The integrity and quantity of the RNA were checked with an Experion™ Automated Electrophoresis System using Experion mRNA StdSens chips (Bio-Rad Laboratories, Inc., Hercules). The RNAs with an RNA quality index (RQI) of 9.0 or higher- with an RQI of 10.0 for perfect integrity- were used for shotgun RNA sequencing.

Massager RNA was purified and used to synthesize cDNA and shotgun RNA-seq libraries were constructed. These were accomplished using the TruSeq RNA sample Prep Kit v2 (Figure 1). The high-throughput sequencing technology Illumina platform, HiSeq 2000, was utilized to sequence the RNA-seq libraries. We used the Illumina platform because it could consistently generate quality sequencing reads, with the lowest cost per base and the highest base sequencing output (https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-rna-v2.html). The RNA-seq reads were used for several purposes, including assembling the transcripts of each gene using the *De novo* transcriptome assembly, quantifying gene and transcript expressions in the developing pods, looking for transcripts and genes differentially expressed between the cultivated and wild-type soybeans, discovering novel gene structure and SNP/InDel mutations, identifying alternatively spliced isoforms, discovering biological functions for each differentially expressed gene and determining co-expression networks of the domestication-related genes within and between the cultivated and wild-type soybeans.

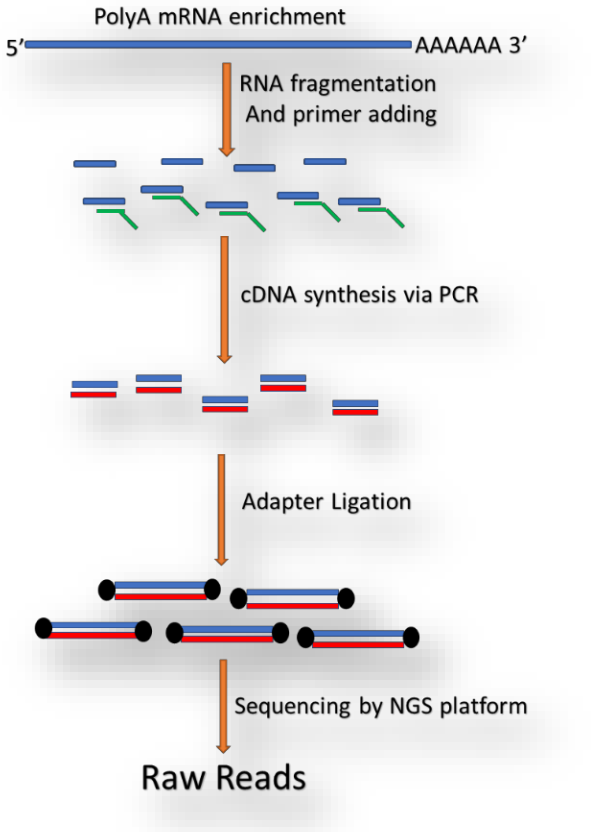


Figure 1 RNA-seq library preparation and sequencing. After obtaining mRNA, it is fragmented while cDNA is synthesized. Then, adapters were specifically designed for Illumina sequencing, and ligated to synthesized and fragmented cDNAs. Finally, the RNA-seq cDNA libraries were amplified by PCR and are sequenced.

As the sequencing reads of the Illumina platform would increase in errors with the increase of the read length, a 100-base paired-end reads (100PE) module was used for the RNA-sequencing to balance the read length and sequencing quality. The shot-gun RNA-seq data for each germplasm line consisted of 10 – 15 million clean reads in FASTQ format. These depths have been previously approved to be sufficient for our research purposes

(Magoc and Salzberg, 2011). Additionally, we used the PE module for the sequencing, due to the fact that it provides more read information, improves within-gene isoform frequencies and gives improved accuracy of transcriptome abundance estimation over SE (Katz *et al.*, 2010; Nicolae *et al.*, 2010).

3.2.2. Clean Read Preparation

NGS resulted in raw short sequence reads in FASTA or FASTQ format (Dundar *et al.*, 2015). The raw data were transformed into clean reads so that they could be used for downstream transcriptome analysis, such as *de novo* transcript assembly (Zhao *et al.*, 2011), transcript and gene expression quantification (Li and Dewey, 2011), transcript and gene differential expression analysis (Michael *et al* 2014), and gene annotation (Götz *et al.*, 2011). Therefore, the raw short-sequence data produced by the Illumina HiSeq 2000 platform were subjected to several treatments, including adaptor trimming, quality control and read normalization. We accomplished those treatments with a programming language syntax called the Trinity Bioconductor package. First, the adapters of the reads were trimmed, based on a threshold of Phred Q20 and a trimming cutoff of 50 base pairs (bp). This step was performed using the 'trimmomatic' software bundled with the Trinity package (Haas *et al.*, 2013) Then, quality control was done by removing low-quality reads by the FastQC option in the Trinity software. The last step of the clean read preparation process was the normalization of the read depth, achieved by running the RSEM software bundled with the Trinity package. RNA-seq by Expectation Maximization (RSEM) is a software package for estimating gene and isoform expression levels from RNA-Seq data (Li and Dewey, 2011).

3.2.3. De novo Transcript Sequence Assembling from RNA-seq Data

Two methods have been utilized to assemble full-length transcripts from the NGS short reads: alignment of the clean reads against a present reference genome sequence and direct assembly of transcripts, without any genome sequence guidance, called *de novo* assembly. The first method takes advantage of availability of a well-assembled genome of the organism under study. Nevertheless, even though there is a well-assembled genome available, the results could differ across different genome assembly versions and because the genome is much more complex than transcripts, it is far more difficult to accurately assemble. Therefore, the *de novo* assembly method has been one of the methods of choice for transcript assembly (Haas *et al.,* 2013). Due to its non-requirement for a well-assembled reference genome, the *de novo* assembly method allows researchers to comprehensively investigate the transcriptome of any organism (Martin *et al.,* 2011). In this study, although the genomes of cultivated (Schmutz *et al.,* 2010) and wild-type (Kim *et al.,* 2010) soybeans have been sequenced, the *de novo* method was utilized to assemble the transcripts, because some studies showed that the *de novo* transcript assembly method ensures superior results, including a higher proportion of reads mapping to a transcript assembly and therefore more accurate quantification of its expression, a higher recovery capability of widely expressed genes, statistics of N50 length, and a larger total number of assembled transcripts, etc. (Loren *et al.,* 2016; Duan *et al.,* 2012; Zhao *et al.,* 2011).

Trinity software was used to conduct *de novo* assembly of the transcripts of each germplasm line, since it is a very efficient program to collaborate with the downstream third-party analysis tools used in this study, such as RSEM (Li and Dewey, 2011) for

transcript and gene expression quantification, and EdgeR (Robinson *et al.*, 2010) and DESeq2 (Michael *et al*., 2014) for gene differential analysis. The RNA-seq clean reads were used for assembling the sequences of each gene and transcript expressed in the developing pods using the Trinity package. The Trinity software uses K-mer (K = 25) (Figure 2a) for the transcript sequence assembly by default. The gap free full-length alignment starts with RNA-seq clean reads being partitioned into many independent *de Bruijn* graphs (preferably one graph per expressed gene) (Figure 2b). These graphs are computed in parallel with increased algorithm to construct full-length transcripts and alternatively spliced isoforms (Garber *et al.*, 2011). After Trinity divides the sequence reads into many individuals *de Bruijn* graphs, it processes each graph independently to construct full-length splicing isoforms. To split the transcripts stemmed from paralogous genes, three different modules are applied including Inchworm, Chrysalis and Butterfly (Figure 3).

Trinity starts with Inchworm module which constructs contigs. Contigs are constructed according to overlapping of K-mers. Then, Chrysalis clusters all related Inchworm contigs that are compatible with the fragments of alternatively spliced transcripts. This process is also done by catching the most matching contigs with overlapping system. Then, Chrysalis builds a *de Bruijn* graph for each cluster of contigs related to each other, illustrating the complexity of overlaps among variants. At the end, de Bruijn graphs are analyzed on a computing grid in parallel and all possible transcript sequences are reported by Butterfly. Additionally, alternatively spliced isoforms and

transcripts stemming from paralogous genes are resolved at the end of the Butterfly process (Haas *et al.*, 2013).

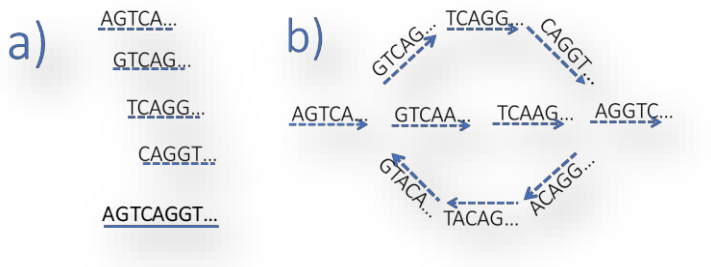


Figure 2 Figure a represents K-mers (assume that the length of k-mers are 25 bp). Figure b illustrates *de Bruijn* graph construction.

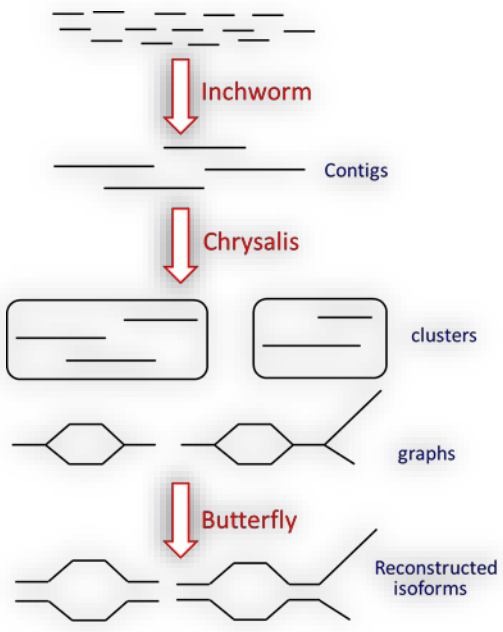


Figure 3 The process of transcript assembly using the Trinity software

In the Inchworm module, linear transcript contigs are constructed through six different steps. 1) Constituting a k-mer dictionary from entire sequence reads; 2) Removing all possible k-mer errors from the k-mer dictionary; 3) Constructing contigs by selecting the most frequent k-mers in the dictionary and removing the k-mers appearing only once and low-complexity k-mers: 4) Extending the contig sequence by adding the k-mer found in the dictionary with the highest abundance, with k-1 overlap, and removing the used k-mer from the dictionary; 5) Extending the sequence assembly till no sequence is left to be extended and reporting the linear contig; and 6) Repeating Steps 3 to 5 until all k-mers are constructed, and each repeat starts with the use of the most frequent k-mer.

The Chrysalis first iteratively classifies the Inchworm contigs into related-components. Those contigs are grouped only when there is an exact overlap of k-1 bases. Then, *de Bruijn* graphs are constituted for each component using k-1 size to reflect nodes, where k defines the edges connecting the nodes. Finally, each read is added to the component with the highest number of k-mers. Each region within reads is determined.

The Butterfly system consists of two parts: 1) Sequential nodes in linear paths are compounded in *de Bruijn* graph to constitute nodes representing longer sequences, and 2) Minor deviation reads, relatively short, resulting from sequencing errors are cut and removed. Butterfly benefits from a dynamic programming procedure to specify paths supported by reads in the graph.

High performance computing environment (called super computer) supported by Texas A&M University was utilized with the help of a network file transfer application named Putty (https://www.chiark.greenend.org.uk/~sgtatham/putty/) to run Trinity software. We uploaded forward and reverse read ‘fastq’ files for each line separately and an assembling

command to the super computer outlined in a red rectangle in Figure 4. All commands for running Trinity base commands were typed using Notpad++ programming language code editor software (https://notepad-plus-plus.org/). That template of commands can be found on vignettes or online bioinformatics software command sharing and development platforms such as GitHub (https://github.com/trinityrnaseq/trinityrnaseq/wiki/Running-Trinity). 'bsub < file_name' was typed as a starting command for assembling process. Assembling process was done by Bowtie software (http://bowtie-bio.sourceforge.net/index.shtml).

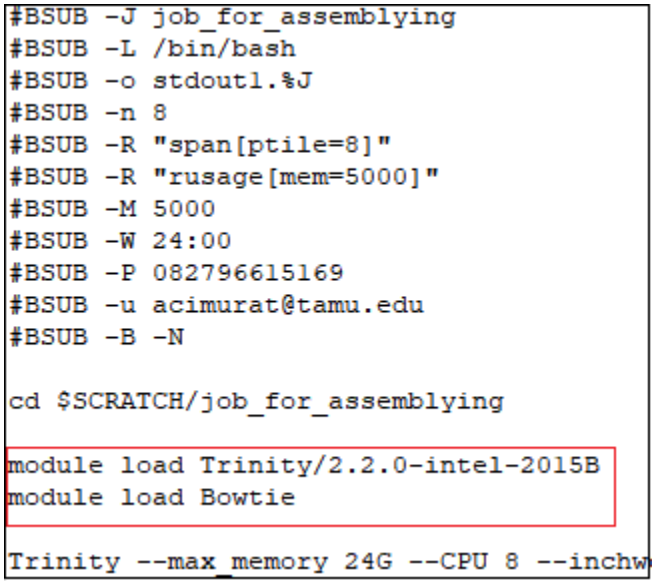


Figure 4 The command for read assembling.

3.2.4. Expression Quantification of Individual Transcripts

Transcript expression quantification is the estimation of abundance of genes or transcript isoforms (Li and Dewey, 2011). This process is necessary for discovering the genes expressed differentially between two divergent biological samples. Several softwares have been developed to quantify the expressions of genes and their transcripts from RNA-seq short sequence reads, including Cufflinks (v1.0.1) (Trapnell *et al.,* 2010), IsoEM (v1.0.5) (Nicolae *et al.,* 2010), RSEM (Li and Dewey, 2011), etc. The RSEM software was used in our study since it is compatible with the *de novo* transcriptome assembly method, with Trinity Software and downstream gene differential expression analysis, and is superior and comparable to the other methods, such as Cufflinks, in regard to quantification accuracy (Li and Dewey, 2011). RSEM estimates expression levels of both individual transcripts and genes (Haas *et al.,* 2013). Additionally, RSEM is capable of high-quality data normalization for reliable detection of transcriptional differences. These expression estimations result in fragments per kilo-base per million of mapped reads (FPKM) which is widely used for transcript and gene expression research (Peipei *et al.,* 2015) using paired-end sequence reads.

All germplasms were bootstrapped to have more biological replications and to enhance reliability of expression level quantification results of transcripts or genes. Super computer has been used for running the RSEM software package uploading bootstrapping comparison files (Figure 5), an assembled reference fasta file and command file for

expression level quantification (Figure 6). The template command for RSEM run shown in Figure 6 also can be found in online vignettes.

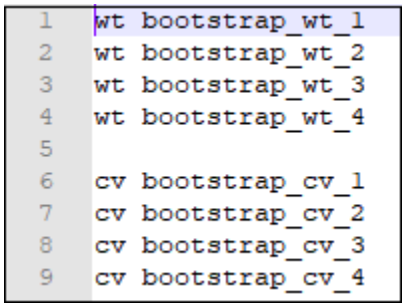
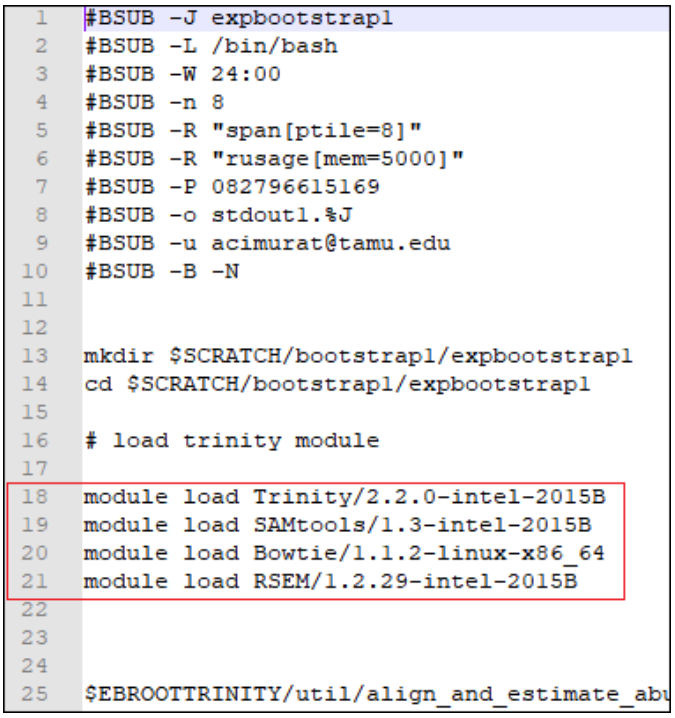


Figure 5 This Figure shows the order of comparison bootstrapping of both types of soybeans' assembled transcripts or genes. Each bootstrapping has different order of each line within each soybeans species.



```
#BSUB -J expbootstrap1
#BSUB -L /bin/bash
#BSUB -W 24:00
#BSUB -n 8
#BSUB -R "span[ptile=8]"
#BSUB -R "rusage[mem=5000]"
#BSUB -P 082796615169
#BSUB -o stdout1.%J
#BSUB -u acimurat@tamu.edu
#BSUB -B -N


mkdir $SCRATCH/bootstrap1/expbootstrap1
cd $SCRATCH/bootstrap1/expbootstrap1

# load trinity module

module load Trinity/2.2.0-intel-2015B
module load SAMtools/1.3-intel-2015B
module load Bowtie/1.1.2-linux-x86_64
module load RSEM/1.2.29-intel-2015B



$EBROOTTRINITY/util/align_and_estimate_abu
```

Figure 6 Figure b shows the template command typed in Notpad++ application. RSEM gets help from Bowtie to have gap-free assembled reads depicted in the red rectangle.

3.2.5. Identification of Transcripts Differentially Expressed in Developing Pods Between Cultivated and Wild-type Soybeans

The DeSeq2 software, a part of the R/Bioconductor package (DeSeq2 website; Gentleman *et al*., 2004), was used to identify the differentially expressed genes (DEGs) in developing pods between the cultivated and wild-type soybeans. DeSeq2 uses a statistical method to analyze quantitative differences between assembled and counted data in the presence of small replicate numbers, outliers, large dynamic range and discreteness. The

DeSeq2 method is capable of covering those requisites and its statistical power is more sensitive and more precise than other existing software (Michael *et al.,* 2014). It tests for differential expression of genes using negative binomial generalized linear models, including the statistical estimation of dispersion and logarithmic fold changes. With this method, it gains the ability to have shrinkage of log2 fold changes in order to reduce potential expression contrasts with drastically different numbers of differential genes. We adjusted the threshold value of log2-fold changes 2-fold for up-regulation or 2-fold for down-regulation to identify DEGs. The DeSeq2 package improves stability and interpretability of estimates using the shrinkage estimation method for logarithmic fold changes and dispersions. The significance value (*p*-value) and False Discovery Rate (FDR) were identified as < 0.05 to decrease false positives for the null hypothesis. The strength of shrinkage tends to decrease as the sample size increases. We took advantage of the Trinity Bioconductor to run DeSeq2 because it lists differentially expressed transcripts with fold-change and statistical significance values. The other important feature of DeSeq2 is that it can visualize the differences between gene expression levels and estimated fold changes based on stable estimation of effect sizes (logarithmic fold changes). Visualization results are in the format of Minus over Average (MA)-plots and Volcano plots showing log ratios and mean values, as well as clustered heat maps and correlation plots in PDF format. MA and Volcano plots facilitate visualization of the expression differences between two samples. For the MA plot, analyzed data is transformed into M (log ration) and A (mean average) scales, and then these values are plotted on a graph (Michael *et al.,* 2014). Volcano plots give information about whether fold chance of expression differences is statistically important or not (Chi and Churchill, 2003).

The DeSeq2 package run started with uploading expression result files of each bootstrapped replicate, which are results of the RSEM run for each soybean species with 'isoform.result' extension, to the super computer. Additionally, some commands for running DeSeq2 in Trinity were needed. Those commands were provided by Michael et al. (2014) and typed step by step to the high-performance computer. Lastly, bootstrap comparison command (Figure 5) typed on Notepad++ software was also needed as input.

3.2.6. Functional Annotation of DEGs

Blast2GO was used to annotate, functionally categorize and map the DEGs to pathways. Blast2GO is one of the few tools for these research purposes. It is used for not only annotation, but also data mining of new sequence data via the use of gene ontology (GO) vocabulary that is a consistent description of gene products in many different genomic databases (www.bioontology.org), KEGG pathways (Kyoto Encyclopedia of Genes and Genomes) (Ogata *et al.,* 1999), enzyme codes (EC) (Schomburg *et al.,* 2004), InterPro IDs (Conesa and Gotz, 2007), etc. DEGs are compared against the KEGG pathway enrichment database to explore which annotated genes are likely involved in which pathways (Liu *et al.,* 2016). Blast2GO identifies GO terms to define the genes by three primary categories, Cellular Component (C), Biological Process (P) and Molecular Function (F) (Dong *et al.,* 2017). This software is a user-friendly approach for function annotation of the genes. It also has some easy instructions about how to use on Blast2GO website (https://www.blast2go.com/).

The annotation using the Blast2Go software is composed of three primary steps, blasting or InterPro mining, mapping and annotation (Figure 7). In the first step, the sequences homologous and similar to the query sequences are found using the Basic Local

Alignment Search Tool (BLAST), which compares biological information of the query sequence with a public library or database sequence, such as NCBI. Nucleotide sequences are uploaded to the software as input in FASTA-formatted files. GO terms related to the hits (the most similar sequences found in databases to the query sequences) are provided by NCBI. Then, the functional terms obtained from the GO vocabulary pool are assigned to the query sequence in the annotation step. At the last step of running the software, an excel file is given by the software, which generates information about each annotated DEG, including putative biological functions, sequence length, number of hits, GO IDs and names, enzyme code and names, InterPro IDs, InterPro GO IDs and their names (Figure 8). Moreover, some graphs showing the enzyme pathway in which the query sequence is involved are also presented as visual outputs. Furthermore, the Blast2GO software has a user-friendly interface with user support websites. There are some ready-to-use instructions on Blast2GO home page.

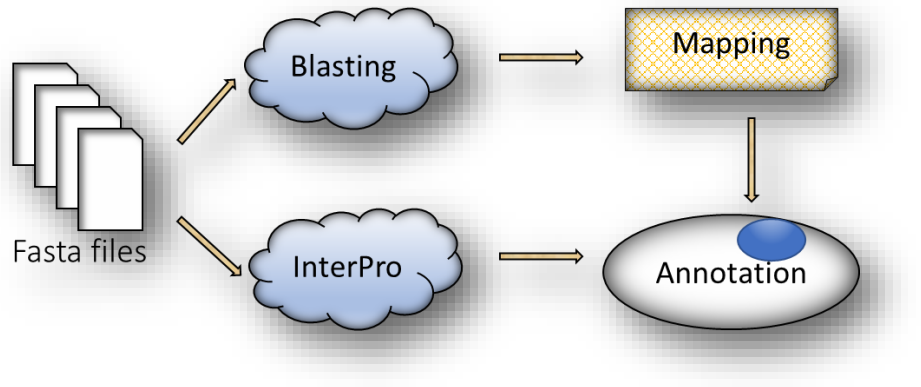


Figure 7 Annotation, functional categorization and pathway mapping of genes using the Blast2GO software. During the run of Blast2Go software, DEGs are blasted against public databases (NCBI blast service) and InterPro discovers if there are any protein families convenient to DEGs in the public database. After mapping those blasted sequences or proteins matching the DEGs, they are annotated to describe which biological and function metabolism in which they are involved.

Blast2GO also has the function of visualization by differentially coloring annotated genes or transcripts to describe the process in which those DEGs are. There are some different colored rows and tags representing active sequences analysis status. Some colors are shown on the first column in Figure 8. Dark red colored rows are categorized as 'no positive result of blasting', light red means there is a positive blast result, mapped sequences are colored with green, annotated sequences are in the blue color, etc. (Conesa and Gotz, 2008).

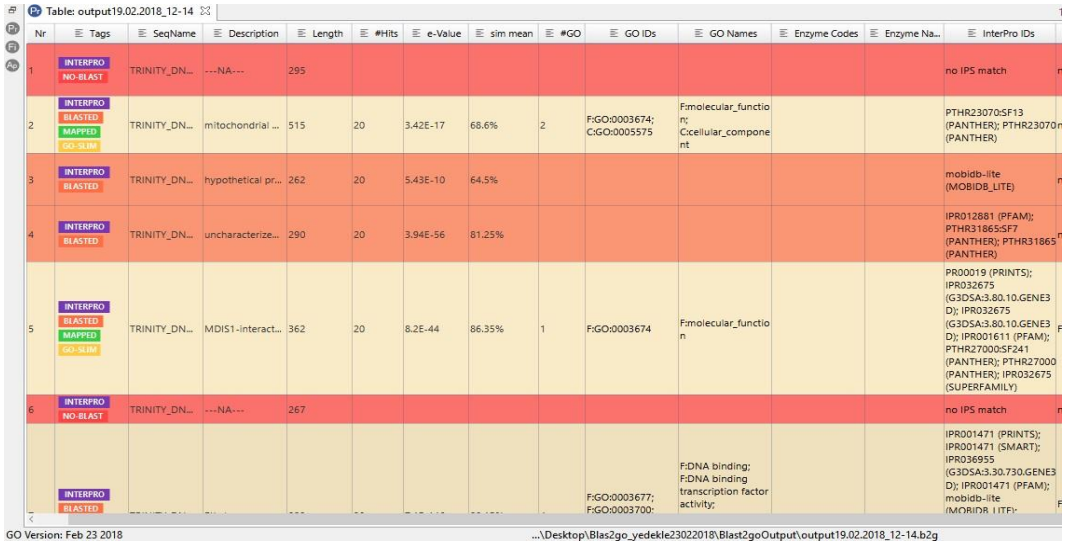


Figure 8 Blast2GO informs users about the status of an input sequence with either colors or descriptions during the run.

3.2.7. Functional Network Visualization of DEGs

Network analysis of DEGs is a technique visualizing graphical relationships between genes. Statistical or computational bioinformatic data can be transformed from electronic files into two or three-dimensional network graphs. After annotation of DEGs, it is desirable to figure out the relationships between the genes to have a deeper insight into how they are involved in a biological process. Because we were looking for the biological relationships of the DEGs, the Biolayout Express^{3D} easy-to-use software (Freeman *et al.*, 2007; Athanasios *et al.*, 2009) was used to construct their co-expression network visualization. In this software, gene or transcript nodes are connected by interaction edges, represented by the expression or functional correlation of the genes, to construct their network. Nodes represents DEGs or differentially expressed transcripts and their functional correlation or interaction are presented by edges (Bader and Enright, 2005). This software has an easy-to-use interface and online vignettes for help in running it. We used the

significance values of $p \leq 0.05$ and 0.01 for expression correlations as the cutoff for the network construction of the cultivated soybean and the wildtype soybean. After running it, visual and colored networks were achieved, which depict groups of nodes representing specific classes of genes with different functional properties. Each color represents a different cluster of nodes interacting closely with each other, according to relatively co-expression levels. While some of the nodes have reciprocal interactions, others may have only single-sided interactions.

In summary, we compared the transcriptomes of the cultivated and wild-type soybean, each represented by a set of selected germplasm lines, to provide an insight into the molecular mechanisms underlying soybean domestication. We first generated the transcriptomes of the cultivated and wild-type soybean germplasm lines (Figure 9) including total RNA extraction mRNA purification and cDNA synthesis, RNA-seq library construction and RNA-seq library sequencing, followed by preparation of clean reads. Then, we analyzed the transcript sequence clean reads (Figure 10) using several software and bioinformatics tools, such the Trinity software for *de novo* transcriptome assembly, the RSEM software for quantification of gene and transcript expression, DeSeq2 for identification of the DEGs in developing pods between the cultivated and wild-type soybeans, Blast2GO for functional annotation, categorization and pathway mapping of DEGs, and the Biolayout Express^{3D} software for functional network analysis.

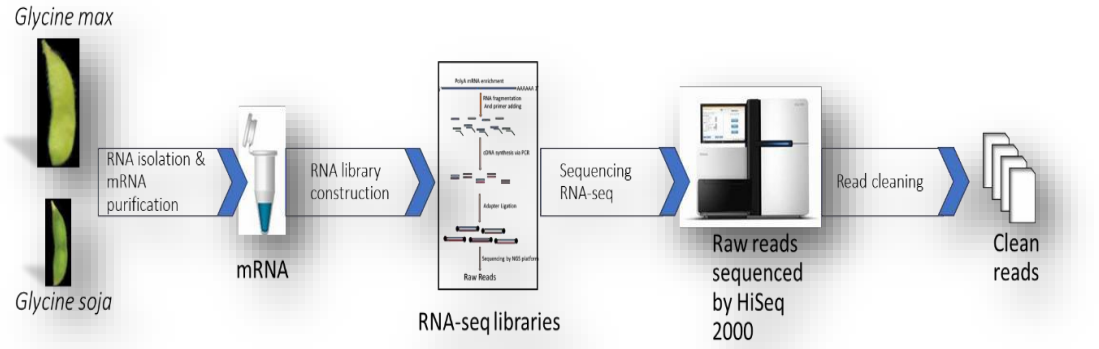


Figure 9 Flowchart of the generation of RNA-seq clean reads. Total RNA was isolated from developing pods of different germplasm lines representing the cultivated and wild-type soybeans. After mRNA was purified and cDNA was synthesized and fragmented, the RNA-seq.

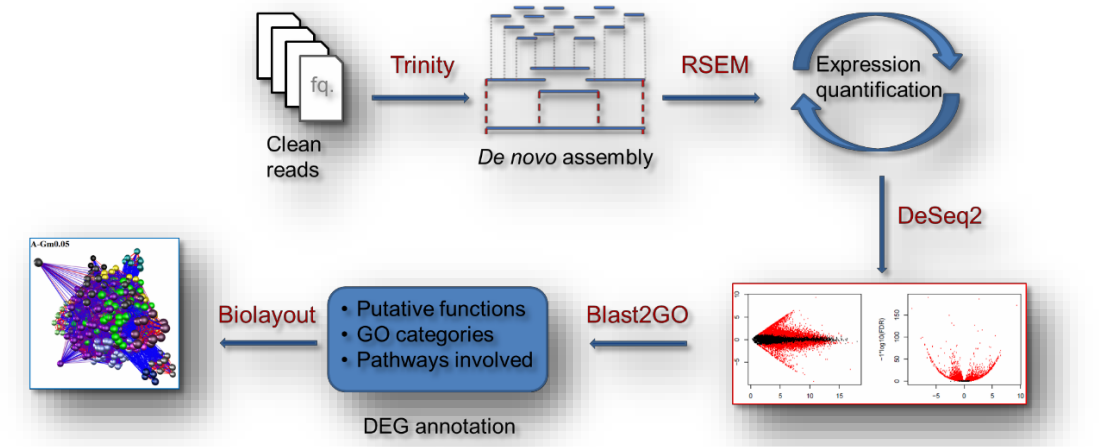


Figure 10 Flowchart of RNA-seq clean read data analysis. Clean reads resulted from each germplasm line were assembled into full-length transcripts with the de novo method using the Trinity software. The expression levels of the assembled transcripts were determined using the RSEM software, and the DEGs were identified between the cultivated and wild-type soybeans using the DeSeq2 bioconductor. The DEGs were annotated, categorized and pathway mapped using Blast2GO. The data analysis resulted in visualization of interactions between the DEGs.

4. RESULTS

4.1. Morphological Differences Between *G. max* and *G. soja*

The plants of each germplasm line representing the cultivated and wild-type soybeans were subjected to phenotypic measurements in several morphological traits important to domestication, such as seed size/weight, leaf length and width, and pod size and number. The 100-seed weight mean and deviation of seeds for two different soybean types have been weighed and results are shown in Table 2. The row with light gray in the table belongs to William 82 line, which is the genetic model utilized in this project. Leaf size shown in Table 3 has been calculated in the centimeter-denomination. In soybean, leaves are grown in the triple leaf shape, which has three leaves connected to one footstalk. It consists of the connection of a top leaf and two leaves located oppositely to each other. The top leaf is the tallest leaf in triple leaves and the side leaves are generally almost equal to each other. The average leaf size of the domesticated soybean germplasm was more than three-fold higher than that of the wild-type soybean, 11.6 and 7.3 cm respectively.

Table 2 Seed weight (g) per 100 seeds for each line, species mean and within-species variation.

Accession PI	Species	g/100 seeds	Mean	Standard deviation	C.V. (%)
548655	max	10.43	12.8	3.32	25.9
291312	max	17.64			
518671	max	14.71			
408342	max	14.06			
567452	max	13.48			
567361	max	11.18			
407027	soja	3.02	2	0.58	29
423995	Soja	2.69			
407275	Soja	1.96			
483463A	Soja	1.89			
437356	Max	15.81			

Table 3 This table shows the leave sizes of each germplasm. The first seven columns were colored green, representing the ID of cultivated lines, while wild-type lines were colored by yellow. Numbers in first column represent the number of plants measured for each line. On the other hand, numbers on the right column depict individual leaves in triple shape leaves. Numbers with blue colors illustrate the average number of leaves of each line.

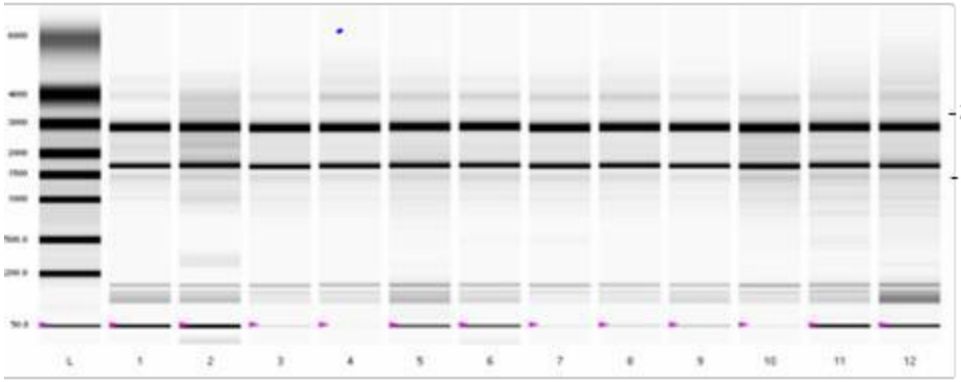
	548655	518671	437356	567452	291312	567361	408342	407027	407275	483464A	423995	
#1	11.5	13.5	12	8.5	7	16	12	6	7	9	9	top leaf
#1	9	7	8	7.5	6	12	11	5	6	4.5	6	side leaf 1
#1	10	7.5	8	5	6	12	11	5	5	5	7	side leaf 2
#2	11	14	11	9		16	13	6	7.5	12	10	top leaf
#2	10	8	7.5	7		12.5	11	4.5	6	7	6	side leaf 1
#2	10	8	7.5	7		12.5	10	4.5	5	6	7.5	side leaf2
#3	12	13.5	10.5	8.5		12.5	12	6.5	5	5	6	top leaf
#3	10	7	7	6.5		10	10	5	4	4.5	5	side leaf 1
#3	9	6.5	4	6.5		10	10	5	4	4.5	4.5	side leaf3
			top leaf	average	11.689				top leaf	average	7.3234	
			side leaf	average	9				side leaf	average	5.2917	
			side leaf	average	9.1667				side leaf	average	5.25	

4.2. Total RNA Isolation and RNA-seq Library Construction

4.2.1. Total RNA Isolation and Qualification

Total RNA was isolated (Figure 11a) from developing pods for all 11 germplasm lines representing the cultivated and wild-type soybeans for shotgun RNA sequencing. The quality and quantity of the total RNA were measured using an Experion™ Automated Electrophoresis System with Experion mRNA StdSens chips (Bio-Rad Laboratories, Inc., Hercules). The RNA isolated from all germplasm lines were shown to have an RNA quality index (RQI) of 9.0 or higher (Figure 11b), with an RQI of 10.0 for perfect integrity. The RNAs had a concentration varying from 18.17 to 815.58 ng/µl. These results suggest that we obtained high-quality RNA and a proper amount of RNA that was well suited for RNA-seq for every germplasm line.

a



b

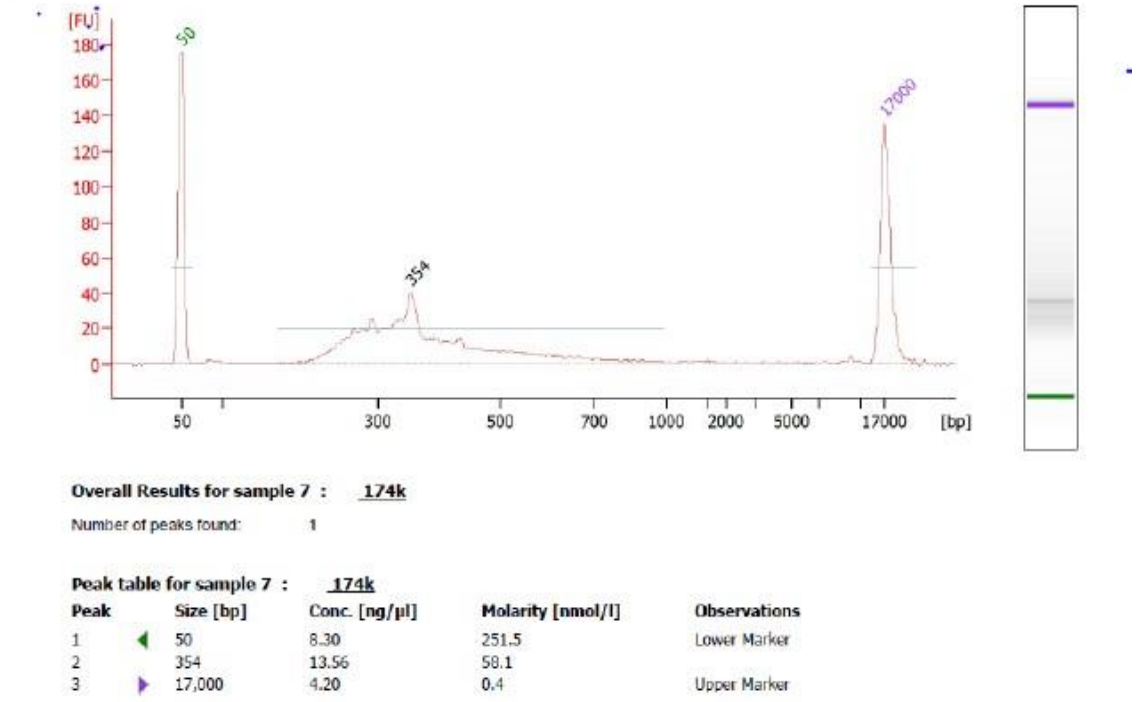
well ID	Sample Name	RNA Area	RNA Concentration (ng/µl)	Ratio [28S:18S]	RQI	RQI Classification
L	Ladder	268.50	160.00			
1	wx13	69.15	41.21	1.91	10.0	Green
2	wx14	30.49	18.17	1.94	9.4	Green
3	wx15	536.90	319.94	1.76	9.8	Green
4	wx16	1,368.66	815.58	1.74	9.9	Green
5	wx17	101.61	60.55	1.96	9.6	Green
6	wx18	66.04	39.35	2.15	9.8	Green
7	wx19	721.83	430.14	1.80	9.8	Green
8	wx20	438.70	261.42	1.96	9.8	Green
9	wx21	268.58	160.05	1.91	9.9	Green
10	wx44	866.54	516.37	1.72	9.6	Green
11	wx63	63.11	37.61	1.95	9.2	Green
12	wx68	94.15	56.10	1.88	9.0	Green

Figure 11 Qualification and quantification of total RNA isolated from the soybean germplasm lines. Figure a shows the RNAs fractionated using the Experion™ Automated Electrophoresis System with Experion mRNA StdSens chips and the ratio of 28S:18S rRNA. Figure b indicates the RNA quality index (RQI) and concentrations of the RNAs.

4.2.2. RNA-seq Library Construction andSequencing

After obtaining high-quality RNA, we constructed the shotgun RNA-seq libraries from the RNAs. The RNA-seq libraries were qualified and quantified as shown in Figure 12. The results showed that we constructed high-quality shotgun RNA-seq libraries with proper concentrations from the RNA. As shown in the example of these RNA-seq libraries in Figure 12, the library had an average size of 352 bp and a concentration of 13.54 ng/µl

suggesting that it was well qualified for RNA seq. Therefore, the libraries were sequenced using the Illumina HiSeq 2000 platform. We evaluated the quality of 100 paired-end (PE) short reads. We trimmed the cloning adapters and filtered the reads of low quality using a quality cutoff of Q20, an equivalent to the probability of an incorrect base call 1 in 100 times. After filtration, a slight increase in percentage of reads having a quality of Q20 or better was observed, leading to an average of 98% of the reads, varying from 97.6% to 98.08%. The clean reads had a GC content ranging from 44.7% to 49.4%, with an average of 45.4% (Table 4), We obtained from 1.20 to 1.66 billion bases of clean reads or 12.0 – 16.6 million 100-base clean reads, with an average of 14.0 million clean reads for each germplasm line. In comparison, the clean reads sequenced for the wild-type soybean lines were 13.5 million while those for the cultivated soybean germplasm lines were 14.3 million, the former being 0.8 million fewer than the latter. According to our other studies, this sequencing depth is sufficient for our research purposes.



Overall Results for sample 7 : 174k

Number of peaks found: 1

Peak table for sample 7 : 174k

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	50	8.30	251.5	Lower Marker
2	354	13.56	58.1	
3	17,000	4.20	0.4	Upper Marker

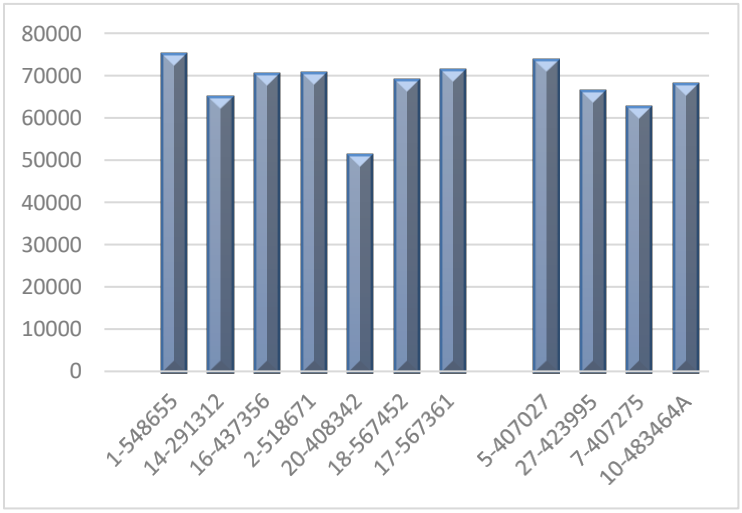
Figure 12 Example of capillary electropherogram of RNA-seq library quality and quantity constructed for RNA-seq of soybean germplasm lines. The RNA-seq library was analyzed using the Agilent 2100 Bioanalyzer (Monica et al., 2014).

Table 4 Preparation of clean reads and their qualification and quantification.

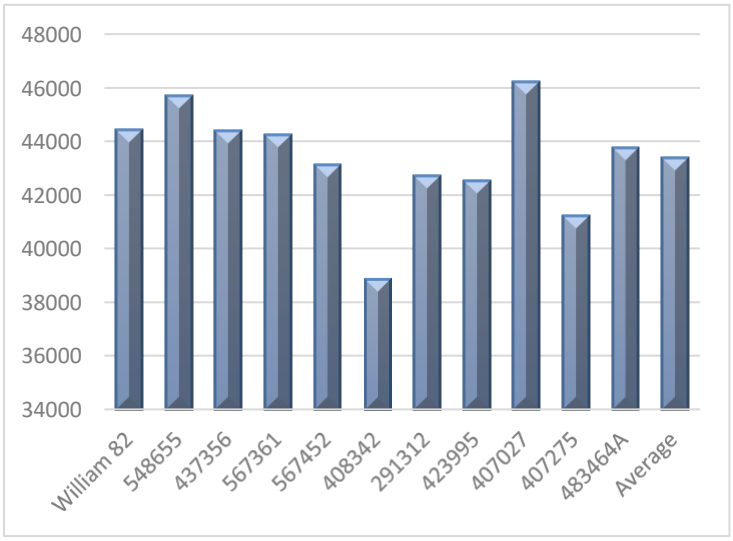
Sample ID	%Q20 Before Filtering	%Q20 After Filtering	%GC Before Filtering	%GC After Filtering	Filter Adapter	Filter Low Quality	Total Reads nt	Clean Reads
548655	97.05%	97.97%	44.76%	44.69%	1.02%	1.76%	1.71G	1.66G
291312	97.24%	98.02%	44.94%	44.88%	0.68%	1.60%	1.35G	1.32G
437356	97.28%	98.07%	44.86%	44.80%	1.10%	1.61%	1.42G	1.38G
518671	97.17%	98.03%	45.16%	45.08%	1.20%	1.71%	1.53G	1.48G
408342	96.31%	97.60%	49.52%	49.41%	1.54%	2.61%	1.44G	1.37G
567452	97.09%	98.06%	45.50%	45.42%	1.14%	1.73%	1.36G	1.31G
567361	97.17%	98.00%	45.30%	45.23%	0.91%	1.67%	1.53G	1.49G
407027	97.19%	98.08%	44.91%	44.83%	2.75%	1.61%	1.64G	1.56G
423995	97.19%	98.07%	44.96%	44.89%	1.89%	1.58%	1.34G	1.29G
407275	97.22%	98.00%	45.05%	44.98%	0.97%	1.56%	1.24G	1.20G
483464	97.13%	98.08%	45.21%	45.14%	1.42%	1.61%	1.40G	1.36G

4.3. Transcript Assembly

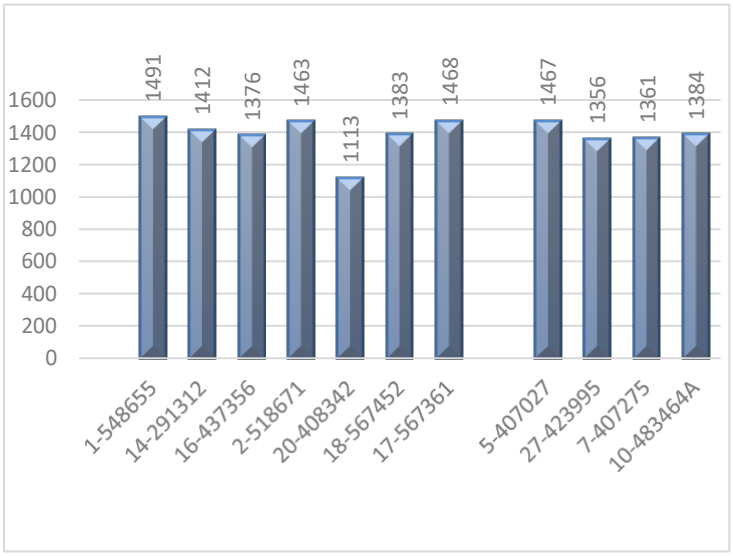
We assembled transcripts expressed in the developing pods for each germplasm line using the Trinity R/bioconductor software (Loren *et al.*, 2016; Duan *et al*., 2012; Zhao *et al.*, 2011). Graph 1, 2 and 3 describe statistics of transcripts and gene assemblies. The seven bars on the right-hand side are the cultivated soybean germplasm lines, and the remaining four bars represent the wild-type soybean germplasm lines. From 51,029 to 74,872 transcripts (Graph 1), an average of 67,407 transcripts, were assembled for each germplasm line. These transcripts were shown to be derived from 38,853 to 46,178 genes (graph 2) with an average of 43,349 genes. The N50 length of the transcript assemblies (Graph 3) was from 1,113 bases to 1,491 bases, with an average of 1,389 bases and the N50 length of the wild-type soybean lines was longer than that of the cultivated soybean lines by 5 bases. In comparison, 176 more transcripts (67,519 transcripts for the wild-type soybean vs. 67,343 transcripts for the cultivated soybean) or 74 more genes (43,386 genes for the wild-type soybean vs. 43,322 genes for the cultivated soybean) were assembled for the wild-type soybean than for the cultivated soybean.



Graph 1 The number of transcripts assembled for each germplasm line.



Graph 2 The number of genes assembled for each germplasm line.



Graph 3 The N50 length of the transcript assembly for each germplasm line.

4.4 Transcript Expression Quantification

Quantification of expression levels of transcripts or genes is necessary to dissect DEGs. Therefore, we quantified the expression of every transcript in the developing pods of every cultivated and wild-type germplasm line using the RSEM software (Li and Dewey, 2011). The transcript sequences of William 82 were used as the reference and the clean reads of each germplasm line were used to quantify their expression levels. FPKM was used to normalize the sequencing depth. We quantified the expressions of all 70,464 transcripts in the developing pods of the 11 cultivated and wild-type soybean germplasm lines. Table 4 shows examples of the expressions of the transcripts. The expression levels of a number of transcripts varied dramatically, not only between the cultivated and wild-type soybeans, but also within each of them. For example, the expression level of

TRINITY_DN11431_c0_g1_i2 varied from 0.63 FPKM to 15.29 FPKM among the cultivated and wild-type soybean lines, from 0.63 FPKM to 12.29 FPKM among the cultivated soybean lines, and 4.70 FPKM to 15.29 FPKM among the wild-type soybean lines. The variance of coefficient (CV) was 43.8% within the cultivated soybean and 58.8% within the wild-type soybean. It was also noted that substantial variation of expression existed among the transcripts alternatively spliced from a single gene, such as among those of TRINITY-DN11434_c0_g1_i1, TRINITY-DN11434_c0_g1_i2 and TRINITY DN11434_c0_g1_i3.

Table 5 This table demonstrates the example RSEM result with the number of FPKMs of each line, green columns for domesticated and yellow columns for wild- type soybeans. Every row belongs to other genes or transcripts and transcript ids were specified. These numbers give some idea about expression levels of transcripts or genes, which means the high expression level has effects on the trait the gene has control of more than low expression level genes do on genes affecting certain traits.

Transcript_id	Cultivated soybean							Wild soybean			
	William 82	548655	437356	567361	567452	2408342	291312	423995	407027	407275	483464A
TRINITY_DN11431_c0_g1_i1	5.73	5.32	7.31	6.71	8.57	0.38	6.15	3.71	5.86	8.68	3.46
TRINITY_DN11431_c0_g1_i2	12.29	7.74	10.48	10.66	11.21	0.63	10.29	4.70	9.34	15.29	4.75
TRINITY_DN11432_c0_g1_i1	7.87	5.70	2.59	2.07	1.71	3.74	7.09	5.41	3.97	8.03	4.57
TRINITY_DN11432_c0_g2_i1	10.89	15.65	15.97	13.15	12.79	3.58	14.10	12.31	13.39	9.49	10.46
TRINITY_DN11433_c0_g1_i1	24.20	19.69	30.41	7.30	3.34	11.45	25.34	17.92	21.64	24.64	8.05
TRINITY_DN11433_c0_g1_i2	166.66	168.99	168.81	210.52	199.77	65.36	160.88	213.83	120.56	143.25	153.30
TRINITY_DN11434_c0_g1_i1	7.49	5.97	6.59	6.13	7.72	1.89	5.59	8.87	3.96	8.93	7.88
TRINITY_DN11434_c0_g1_i2	4.29	2.25	1.77	2.00	4.21	1.13	2.34	6.20	5.09	6.19	1.82
TRINITY_DN11434_c0_g1_i3	2.36	5.89	6.94	5.94	3.90	3.70	1.92	3.61	7.00	7.21	5.47
TRINITY_DN11435_c0_g1_i1	13.45	8.56	14.35	12.91	15.61	4.83	11.56	8.66	14.31	10.65	17.46
TRINITY_DN11435_c0_g1_i2	5.90	8.60	3.22	2.94	3.50	0.97	1.89	2.30	2.00	3.26	4.76
TRINITY_DN11436_c0_g1_i1	21.39	15.67	15.63	14.54	18.88	8.31	15.98	14.49	17.24	9.84	18.28
TRINITY_DN11437_c0_g1_i1	10.51	10.76	10.00	10.95	7.03	3.39	8.90	8.06	8.82	11.16	10.69
TRINITY_DN11438_c0_g1_i1	42.05	49.54	57.40	48.96	59.16	27.76	44.33	51.12	46.94	41.35	57.90
TRINITY_DN11438_c0_g1_i2	37.83	37.57	42.47	40.97	33.62	16.44	32.42	51.97	50.36	65.18	48.01
TRINITY_DN11438_c0_g1_i3	73.64	75.16	75.27	80.82	84.32	38.18	61.72	69.43	94.19	75.36	80.76
TRINITY_DN11439_c0_g1_i1	38.08	50.30	32.24	34.77	24.73	16.80	37.80	31.35	16.12	17.44	37.56
TRINITY_DN11439_c0_g1_i2	51.72	62.17	47.53	46.48	31.78	16.79	51.42	45.07	31.33	19.20	31.23

4.5. Genes Differentially Expressed in the Developing Pods between Cultivated and Wild-type soybeans

Divergence has occurred between the cultivated soybean and its wild-type relative during and after soybean domestication, not only in gene sequence mutation, but also in expression and network interaction. Due to this knowledge, we conducted gene differential expression analysis using the DeSeq2 software to identify the DEGs between the two species. Since we were interested in the DEGs between the cultivated soybean and its wild relative and there was a different number of germplasm lines representing each, we randomly grouped the seven cultivated soybean germplasm lines into two subgroups, with each subgroup consisting of William 82, the model genotype for soybean genetic and genomics studies, and three other germplasm lines. Therefore, each subgroup of the cultivated soybean consisted of four germplasm lines, which was the same number as the four germplasm lines of the wild-type soybean. Moreover, to facilitate the analysis using the DeSeq2 software, a bootstrap sampling method was used to create "biological replicates" from each four-line subgroup. Four "biological replicates" were created for each subgroup. Each subgroup of four germplasm lines of the cultivated soybean was then subjected to differential expression analysis independently against the four germplasm lines of the wild-type soybean. Finally, the DEGs identified between the two pairs of the cultivated soybean versus the wild-type soybeans were compared. Interestingly, the same set of DEGs was achieved between the two pairs of the cultivated and wild soybean germplasm lines, suggesting that the strategy of DE analysis was well acceptable for our

research purposes. The results are summarized in Figure 13. We identified a total of 1,403 DEG transcripts between the two species, when a cutoff of $P \leq 1.0E-03$ was applied for DEG transcripts. Examination of the 1,403 DEG transcripts revealed that they were alternatively spliced from 1,247 genes. Of the 1,403 DEG transcripts or 1,247 DEGs, 1.035 DEG transcripts or 916 DEGs were up-regulated in the cultivated soybean and down- regulated in the wild soybean, while 368 DEG transcripts or 331 DEGs were up-regulated in the wild soybean and down-regulated in the cultivated soybean.

Additionally, DeSeq2 supported visualization of results with MA-plot and Volcano plot graphics as in Graph 4, in pdf format. In the MA plot graphic, each red-colored dot represents one expressed gene. As the expression level of genes is more upregulated or downregulated, the spread of the plot increases in either direction. Furthermore, the Volcano plot graphic also works well to show DEGs. As we can see on the graph, the expression-levels of genes advances as the plots biases from the logFC limits (2 or more for up-regulation, -2 or less for down-regulation).

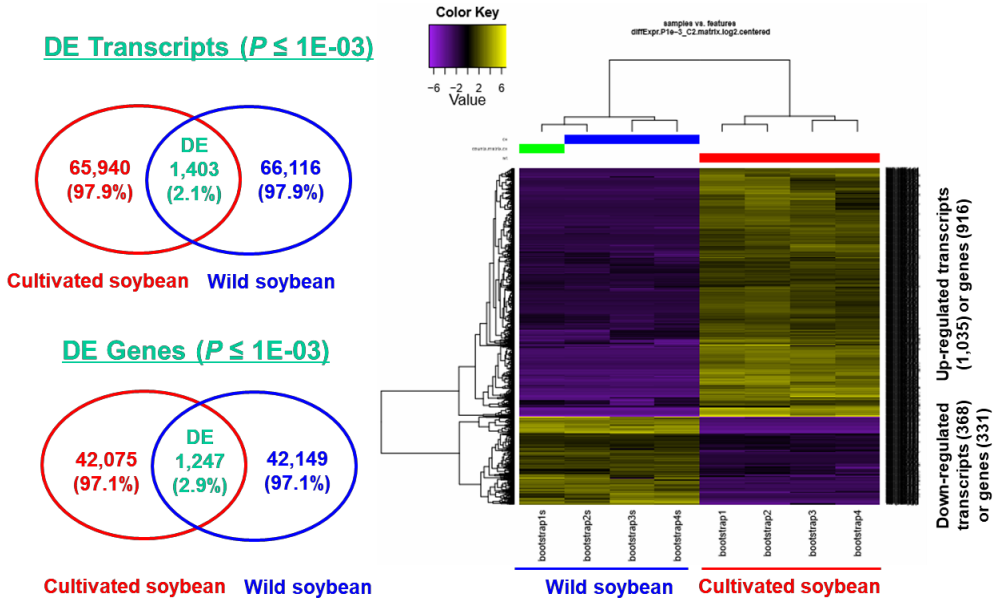
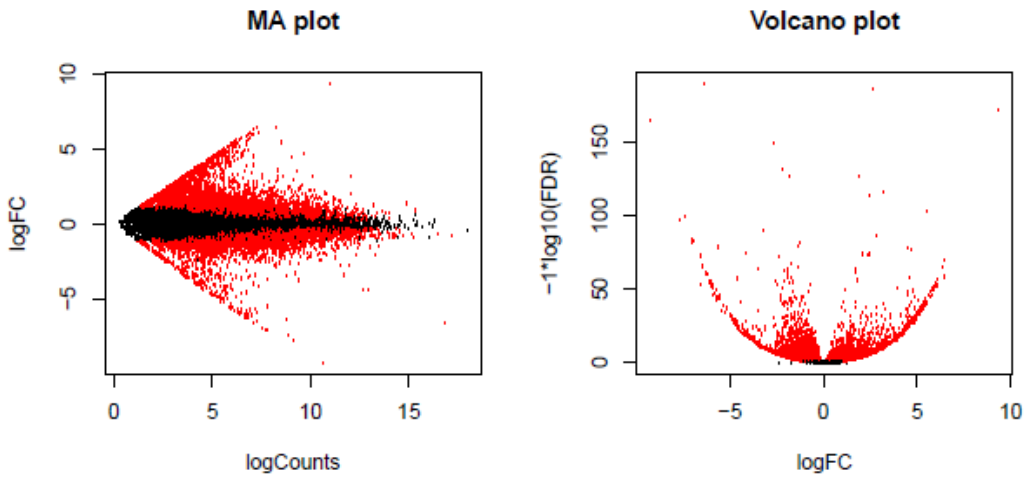


Figure 13 Circles on the top-left side of figure shows the number and percentage of transcripts expressed and differentially expressed. In the circle on the bottom-left side, numbers or percentage belong to expressed and differentially expressed genes between soybean species. The picture on the right side of figure (heat map) illustrates visually up and down-regulated transcripts and genes. The purple and yellow colors represent down and up-regulation, respectively. The level of regulation increases as the color key value gets further away from the center, as in the little schematic in the top middle of the figure.



Graph 4 These graphs represent visualized results of the DeSeq2 package, with the MA plot on the left and Volcano plot on the right. All the dots in both graphs represent genes. While black dots represent statistically not significant fold change values, red dots indicate that there are statistical differences in fold change between gene expression levels. In the MA plot, as long counts (x axis) increase, expression changes (logFC on y axis) increase positively (upwards) or negatively (downwards) (Michael et al., 2014). For the volcano plot graph, the p-value is shown on the y axis, where 50 indicates that the p value is 0.05 where fold change equals to 2. The p value goes down when the significance value for fold change enhances. On the contrary, expression differences are less than 2-fold changed when the p value is higher than 0.05, which means expression differences are not significant (Chi and Churchill, 2003).

4.6. Annotation, Categorization and Pathway Mapping of DEGs

To find what the DEGs are, we have annotated and categorized 1.403 DEG transcripts in gene ontology (GO) and mapped their pathway using the Blast2GO program. We could annotate 1,255 (89.5%) of the 1,403 DEG transcripts while 148 were reported as 'NA', meaning that the sequences could not be annotated. The other unknown and unannotated sequences or proteins are illustrated in pie chart with their ratio among all DE

transcripts (Figure 14). Table 6 shows examples of annotation results of the DEG transcripts.

The 1,255 annotated DEG transcripts were categorized into 47 secondary GO functional categories of all three primary categories, 16 of which belong to the Cellular Component category, 9 to the Molecular Function category and 22 to the Biological Process category (Graph 5). The annotated DEG transcripts were involved in a total of 56 KEGG pathways. The top 10 included Starch and Sucrose Metabolism, Citrate Cycle (TCA cycle), Fatty Acid Biosynthesis, etc. (Graph 6). Figure 15 shows the Citrate Cycle pathway in which the DEGs are involved. The annotation results showed that 108 of the DEGs coded different enzymes. After the isoforms of genes that coded the same enzymes were removed, 52 enzymes were coded by the DEGs. Integrative analysis of the results from GO categorization, KEGG enrichment pathway analysis, and literature review found that 12 of the 52 DEG-coding enzymes were putatively involved in pod dehiscence and seed size metabolisms in soybean (Table 7). Eight of these 12 enzymes were members of the hydrolase class that causes pectin bonds to be broken. They are involved in cellular metabolism, acting as peptide bond and ester bond, and expansin protein deformation relating to shattering trait. One of the hydrolase class enzymes was pectin hydrolase, which was down-regulated (-4.96) with a higher expression level in the cultivated soybean, causing shattering resistance. Another enzyme of the hydrolase class has a negative effect on carbon-nitrogen bonds, rather than peptide bonds. The remaining four of the 12 enzymes were seed size-related domestication genes. One of the four enzymes was ubiquitinyl hydrolase, which is known to affects seed size negatively. Another one of the four enzymes play role in ubiquitin ligase activity and the expression level of the transcript encoding the enzyme was up-

regulated at a log2fold change of 4.6 in the cultivated soybean. One of those enzymes was found to be involved in histidine amino acid synthesis, which inhibits development of seeds (Ma and Wang, 2016). On the other hand, one enzyme was found to enhance plant branching metabolism with an up-regulated expression level of 4.6-fold in the wild-type soybean.

Table 6 As an example, some of sequence names of DEGs are given on the list in this table including some information about their biological function for each differentially expressed gene in the soybeans` genomes.

SeqName	Description	#GO	GO IDs	GO Names	InterPro IDs
TRINITY_DN2	mitochondrial chaperone bcs1	2	C:GO:0005575; F:GO	C:cellular_componen	PTHR23070:SF13 (PANTHER); PTHR2
TRINITY_DN5	hypothetical protein CQW23_34413				mobidb-lite (MOBIDB_LITE)
TRINITY_DN6	uncharacterized protein LOC100811994				IPR012881 (PFAM); PTHR31865:SF7 (
TRINITY_DN9	MDIS1-interacting receptor lik	1	F:GO:0003674	F:molecular_functio	PR00019 (PRINTS); IPR032675 (G3DS
TRINITY_DN1	---NA---				no IPS match
TRINITY_DN1	Ethylene-responsive transcrip	4	F:GO:0003677; F:GO	F:DNA binding; F:tra	IPR001471 (PRINTS); IPR001471 (SMA
TRINITY_DN1	probable 2-oxoglutarate-depe	1	F:GO:0016491	F:oxidoreductase act	PR00682 (PRINTS); IPR026992 (PFAM
TRINITY_DN3	1-aminocyclopropane-1-carbo	1	F:GO:0016491	F:oxidoreductase act	IPR027443 (G3DSA:2.60.120.GENE3D
TRINITY_DN3	BTB POZ domain-containing A	2	C:GO:0016020; C:GO	C:membrane; C:inte	IPR027356 (PFAM); PTHR32370 (PAN
TRINITY_DN3	DUF581 family	5	F:GO:0016702; F:GO	F:oxidoreductase act	IPR007650 (PFAM); PTHR33059 (PAN
TRINITY_DN4	transmembrane	1	C:GO:0016021	C:integral componen	IPR007493 (PFAM); IPR036758 (G3DS
TRINITY_DN4	transmembrane	1	C:GO:0016021	C:integral componen	IPR007493 (PFAM); IPR036758 (G3DS
TRINITY_DN4	myosin heavy chain	2	C:GO:0016020; C:GO	C:membrane; C:inte	PTHR34360 (PANTHER); PTHR34360:
TRINITY_DN5	toll interleukin-like receptor-	1	P:GO:0007165	P:signal transduction	IPR035897 (G3DSA:3.40.50.GENE3D);
TRINITY_DN5	IQ-DOMAIN 31-like	1	F:GO:0003674	F:molecular_functio	IPR000048 (SMART); IPR000048 (PFA

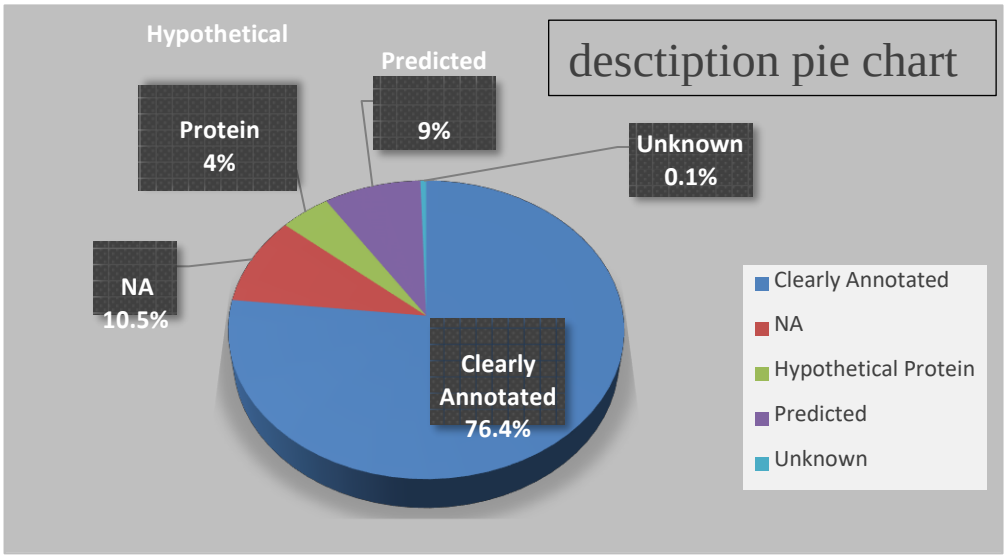
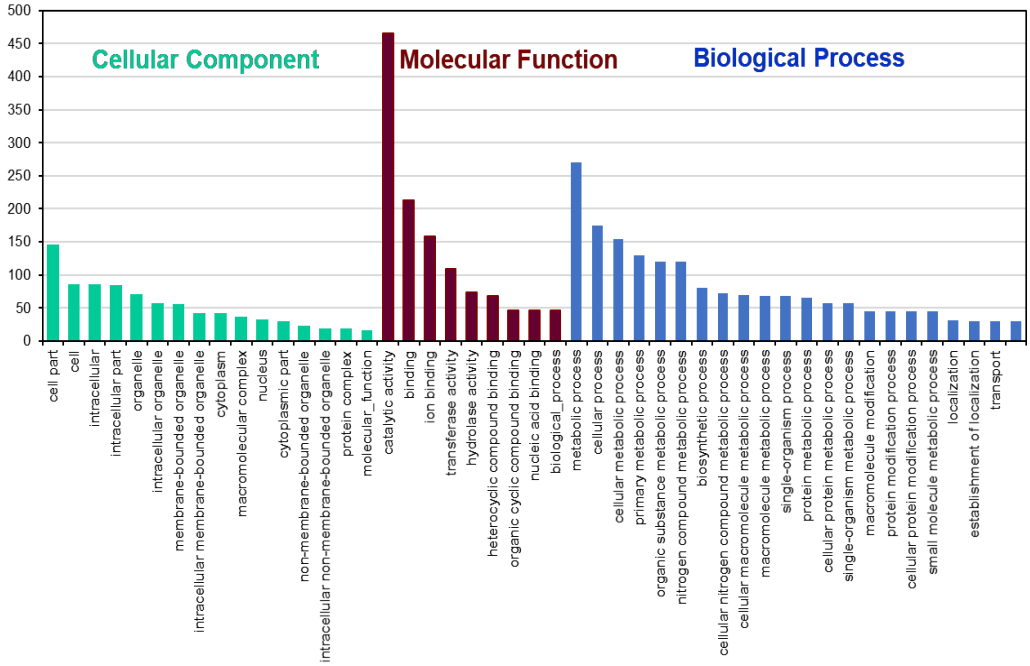
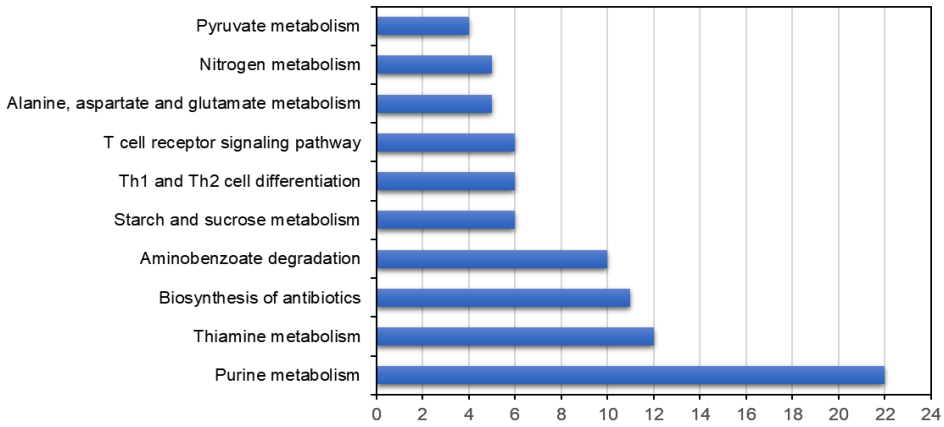


Figure 14 Percentage of annotated and non-annotated proteins.



Graph 5 The primary and secondary functional categories into which the annotated DEG transcripts were categorized. The x-axis represents different functional categorizes into which the DEG transcripts were categorized and y-axis indicates the number of DEG transcripts were categorized and y-axis indicates the number of DEG transcripts that were categorized into each functional category.



Graph 6 The top 10 pathways in which the annotated DE transcripts were involved. While x-axis represents the number of transcripts, y-axis shows the metabolic pathways in which the transcripts were involved.

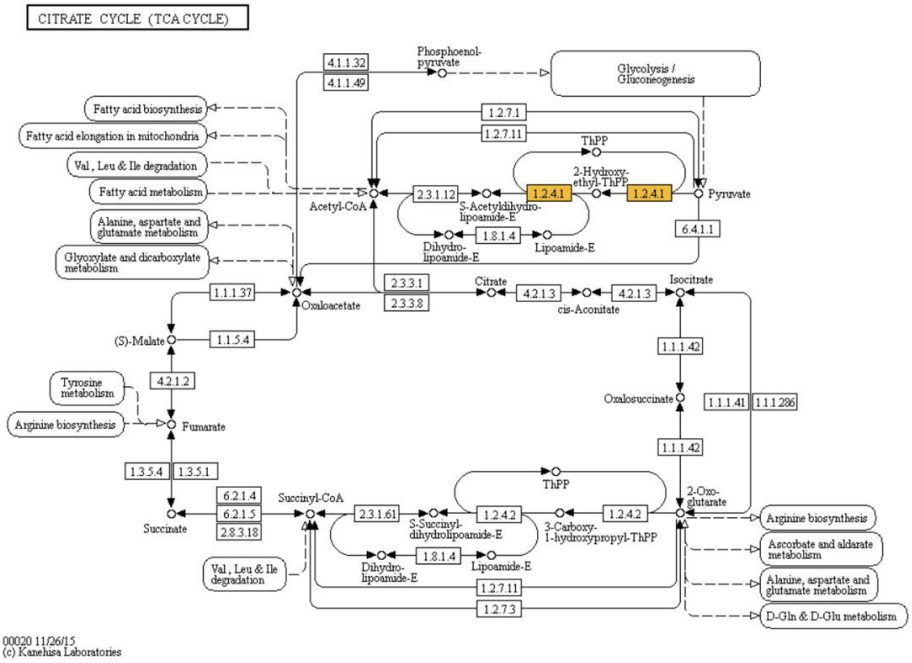


Figure 15 The Citrate Cycle pathway; one of the top 10 pathways in which the annotated DEG transcript enzymes are involved. The boxes highlighted in orange indicate the enzymes coded by the DEG transcripts.

Table 7 12 different pod shattering and seed size-related enzymes encoded by some of DEGs.

	A	B	C	D
1	Enzyme Codes	Enzyme Names	Description	
2	EC:3.6.1.15	Nucleoside-triphosphate phosphatase	kinesin KIN-12B	
3	EC:3.4.24	Acting on peptide bonds (peptidases)	membrane-bound transcription factor site-2 protea	
4	EC:2.7.1.1	Hexokinase	Hexokinase-3	
5	EC:3.1.30/3.1.26; EC:	Acting on ester bonds; Acting on ester bonds; Ribonuclease H	ribonuclease H	
6	EC:4.2.1.19	Imidazoleglycerol-phosphate dehydratase	Imidazoleglycerol-phosphate dehydratase	
7	EC:6.1.1.10	1,4-alpha-glucan branching enzyme	starch branching enzyme I	
8	EC:2.7.1.107	Acting on peptide bonds (peptidases); Acting on peptide bonds	serine carboxypeptidase-like 18	
9	EC:3.1.3.16/3.1.3.41	4-nitrophenylphosphatase; Acid phosphatase	purple acid phosphatase	
10	EC:3.6.1.15	Amidase	Amidase 1	
11	EC:3.1.30/3.1.26; EC:	Carbamoyl-phosphate synthase (glutamine-hydrolyzing)	carbamoyl-phosphate synthase small chloroplastic	
12	EC:3.6.1.15	Polygalacturonase	probable polygalacturonase isoform X2	
13	EC:3.4.19.12	Ubiquitinyl hydrolase 1	josephin	

4.7. Co-expression Network Analysis of the DEGs

To see genetic interaction, pathways and relationships among the DEGs, we conducted co-expression network analysis of the genes using the annotated results of DE transcripts and genes as input files for Biolayout Express 3D software. Since gene expression is an indication of functionality, the co-expression of the genes will provide an indication of the functional relationship among the DEGs. After adjusting p-value to 0.05 and 0.01 and the correlation value we calculated, we ran the software. DEG interaction was in three dimensional visual format with the extension of '.layout' as output of the software. As a result of running the analysis, we were given some colored pictures (Figure 16) showing interactions the networks of the DEGs between differentially expressed genes of the cultivated soybean, the wild-type soybean, and the combined cultivated and wild-type soybeans, respectively (*G. max, G. soja* and mixture of both species). When a cutoff of $P \leq 0.01$ was applied for the network construction within domesticated type DEGs for the cultivated soybean, we were able to observe total nodes. The total nodes were a network consisting of 140 DEG nodes and 2,966 gene-gene interactions, while the total number of edges was 2.966.

The average connectivity of each of the 140 DEGs was just over 54, for each DEGs with a maximum connectivity of 107. The network constructed at the same cutoff for the wild-type soybean, however, with the same significance value, total nodes consisted of 83 DEG nodes with 3,403 gene interaction of total edges, and an average and maximum connectivity of 82. However, for the network of the mixture DEGs of the combined cultivated and wild-type soybeans, the numbers of total both nodes and edges were much higher than that of interactions within DEGs either or both of the cultivated or wild soybeans, with 856 and 78.279 values, respectively. Additionally, the average and maximum connectivity were also much higher, with 1954.815 and 480 values, respectively. When a cutoff of $P \leq 0.05$ was applied for the network construction, the numbers of nodes, edges and connectivity were observed to be much higher than those obtained with the cutoff of $P \leq 0.01$. For the *G. max*. DEG network in the cultivated soybeans, while total interactions of the gene nodes were increased to 273, total network between those genes was 12,.048, and the average connectivity between genes was increased to 122,.066 with a maximum connectivity of 202. Within the DEG network in the wild soybeans of *G. soja*, the total gene nodes remained 83, and total network the gene interaction edges remained 3,.403, and the average connectivity of each DEG was also kept the same (83). Interaction average within DEGs was 82 with the same number of average connectivity. Mixture of the DEG networks of the combined two species had 1,403 total gene nodes interacting with each other and 217,.920 total gene interaction number edges. The average connectivity of the network was 3110.649 and the maximum connectivity was 795. Furthermore, each network consists of multiple clusters representing some the DEGs groups that were more closely co-expressed. Each cluster has been colored differentially by Biolayout software. In either case of $P \leq 0.01$ and $P \leq 0.05$, DEGs of domesticated soybean the DEG networks for the cultivated soybean had many fewer

clusters than those for the wild soybean. Both types of results were observed to have structure of interwoven clusters, meaning that each cluster has interactions with each other. Undomesticated soybean DEGs have 28 different clusters with no interaction with each other. DEGs in each cluster tend to be involved in the common or close trait mechanisms.

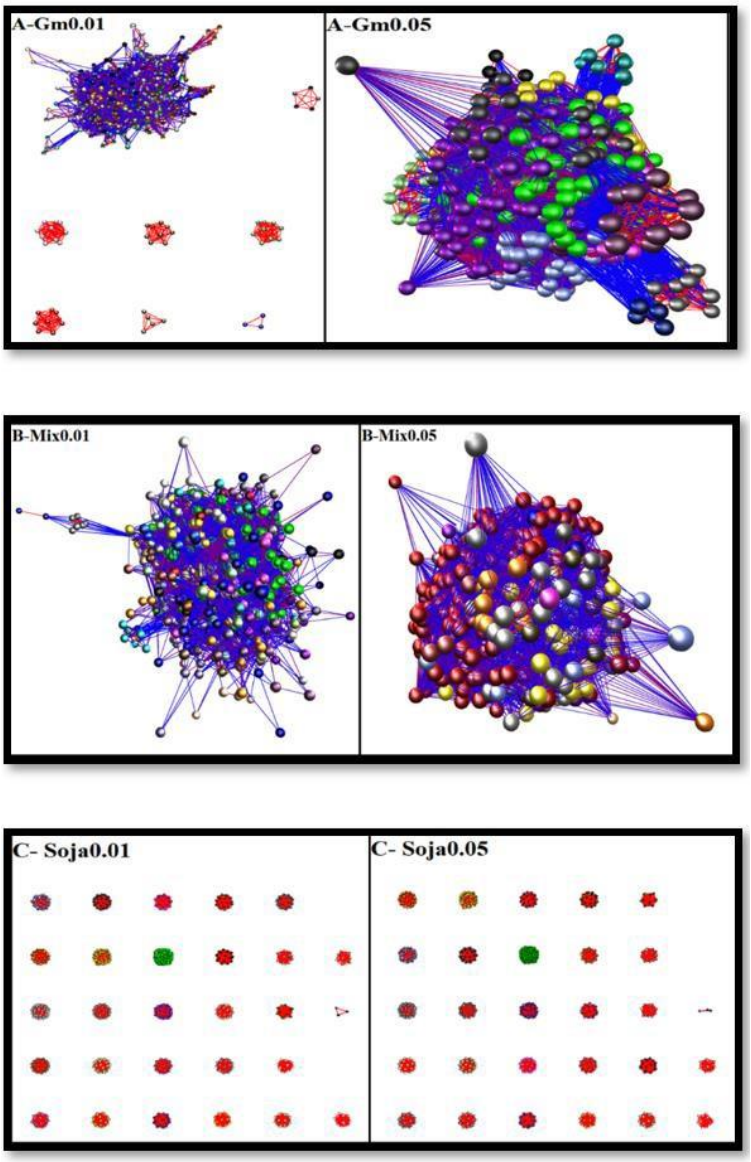


Figure 16 Images A, B and C are results of Biolayout Express 3D visualization software showing DEGs interactions with each other. Image A shows network structure within DEGs of *G. max* with 0.01 and 0.05 of p-value, in the left and right side of the picture respectively. The middle image (B) shows mixture of both types DEGs with the 0.01 and 0.05 significance value. The bottom image (C) represents DEGs network of *G. soja* with 0.01 in the left and 0.05 of significance value in the right.

5. DISCUSSION

This study has generated transcriptomes taken from developing pods of seven cultivated and four wild soybean germplasm lines. Interestingly, assembling the transcripts of each germplasm line has resulted in 176 more transcripts or 74 more genes for the wild soybean than for the cultivated soybean even though the wild soybean lines have fewer clean reads and its transcript assemblies have longer N50 lengths. This result provides the first hint of the divergence of the expressed genes between the cultivated and wild soybeans. The divergence has been further confirmed by the identification of 1,403 transcripts or 1,247 genes differently expressed in the developing pods. Although only 1,255 of the 1,403 DEG transcripts have been successfully annotated with the public databases and 148 have not been annotated yet, the identification of the 1,247 DEGs have already provided an overall insight into the molecular mechanisms of soybean domestication.

Although *qPDHI* and *qSHI* pod shattering-related domestication genes (Suzuki *et al*., 2009; Kamisha *et al.,* 2006) have protein sequence available in NCBI databases, they are not annotated yet. Nonetheless, their homologous, shattering 1-5 genes, named *Glyma16g02200*, have 3 different GO terms found in our results, e.g. GO:0003677, GO:0006355, GO:0005634. They have been found to be up-regulated in the domesticated soybean genome, which results in having shattering-resistance pods. This result was robustly supported by Dong et al. (2017).

Moreover, 20 DEGs have been found in the results of this study, encoding for cell wall modification and hydrolase activity, which reduces the capability of deposition pectin carbohydrate in cell walls, resulting in having shattering-susceptible pods. The expression of those DE genes was up-regulated in the undomesticated soybean genome. Gene Ontology and

KEGG metabolic enrichment pathway analysis of those genes showed that each gene had more than one GO ID such as F:GO:0016757, P:GO:0045489, C:GO:0005794, F:GO:0016740, C:GO:0048046, etc. They were commonly involved in cellular components such as the membrane, extracellular region and cell wall, and biological processes such as cell wall organization and pectin biosynthetic process. Further, they were involved in molecular functions, including transferase activity, hydrolase activity, and peptidase activity. Our results regarding the amount of discovered DEGs, up and down-regulated Unigenes and pectin deposition are almost entirely corroborated by Dong`s findings et al. (2017).

In addition to pectin, the other most important gene family causing cell wall durability against shattering is the expansin protein family (McQueen and Cosgrove, 1995). As a result of this study, two genes encoding the expansin protein, which decreases pod dehiscence, were observed to be up-regulated in shattering-resistance accession, G. max. These genes were involved in cellular components in the extracellular region and biological processes with cell wall organization.

The QTL where *BIG BROTHER (BB)* gene is located in, has locus tag with *AT3G63530*. This gene is known to effect seed size negatively (Dish *et al.,* 2016). While this gene has a lot of synonymous GO enrichment terms, we have found in our results only one of those synonymous terms, with the GO ID of 0061630. Its expression level was high in wild-type soybean genome and that gene was absent in its cultivated relative. With this information, it can be concluded that undomesticated soybean organ size has been negatively affected by that DE gene, which influences ubiquitin ligase activity negatively and is one of the reasons that *G. soja* has smaller leaves and petals sizes. Our findings were verified with research from Dish et al. (2006).

Additionally, 21 unigenes encoding the ubiquitin protein were found to be highly up-regulated in *G. soja* genome and absent in *G. max* genome.

BB gene and its allele *DA1* gene have involved in molecular function and some biological process including metal (zinc) ion binding, peptidase activity, ubiquitin binding, ubiquitin ligase enzyme. These genes limit cell proliferation and of organ growth, seed development etc. Conversely, Liang et al. (2014) has observed *UBP15,* a new ubiquitin- related protein like *BB, DA1, EOD1*. Although it has the same molecular function and biological process as those three proteins, it affects seed size positively, promoting cell proliferation when it is highly expressed. After searching GO terms of this gene in our result, with *AT1G17110* of locus tag, we have faced with six different GO terms encoded by around 500 total DEG results highly expressed in undomesticated soybean, which means our results were supported by Liang`s finding.

Additionally, we pursued Liangfa`s findings about *BS1* gene orthologous in soybean to see if it has a negative effect on lateral organs in soybean as it was claimed (Ge *et al.*, 2016). The results revealed a GO term numbered with 0070375 representing *BIGSEED1 (BS1)* gene. Previously, there was no gene or GO term with that number annotated before.

Zhao et al. (2016) claimed that *GmCYP78A72* gene had an effect on seed size. We searched this gene on NCBI; the QTL of that gene on soybean genome was numbered with *LOC100790231*. In our results, *LOC100790514* was closest code of that locus having some different GO IDs such as C:GO:0016020; C:GO:0016021. Nonetheless, although Zhao claims that gene is over expressed in cultivated soybean, our findings say the exact opposite, meaning that the gene was mostly down-regulated in domesticated soybean or sometimes up-regulated in

undomesticated soybean. This comparison shows that this gene needs to be annotated and dissected at least one more time.

As mentioned before, wild-type soybean is also vital for incorporating some genes related to abiotic stress resistance, such as drought, toxicity etc., and biotic stress resistance, like restriction of pathogen and bacterium growth and some virus resistance. During domestication, the domesticated soybean has lost those resistance-related genes. We found over 40 genes related to disease resistance that were overexpressed in wild-type, absent in cultivated soybean and related to some molecular functions like ADP and ATP binding and hydrolase activity. Further, they were involved in biological processes such as defense response, response to stress and immune system function.

In addition to those comparisons, 12 different enzymes relating to pod shattering and seed size metabolism have been observed in the result of Gene Ontology and KEGG enrichment analysis (Table 6). Eight of them, active pod dehiscence-related domestication genes, were coded by some DEGs for hydrolase and the ligase activity enzyme. The vast majority of them were up-regulated in undomesticated soybean and one of them was down- regulated in this genome. These expression levels are considered to contribute to the shattering-susceptibility of wild soybean. The four other enzymes were involved in the seed size domestication trait. While one of those genes was up-regulated in the uncultivated type, three of them positively affected seed size and were up-regulated in domesticated soybean genome. Differentially, one of the genes contributes to enhancement of branching in the plant and is an undesirable characteristic reducing seed size and quality. Thus, we concluded, supported by Ali et al. (2006), that it has indirect negative effect on the seed size of soybean.

When biological network analysis results of this study are compared, with consideration only of the results of 0.05 significance level, it has been deduced that genes in *G. max* have more interaction than that of *G. soja*. As it can be seen from the cluster structure of DEGs in both accessions (Figure 15), the genes in the same clusters involving in common trait mechanisms in wild-type soybean are independent from each other. We have implicated that we could enhance or remove a trait from undomesticated soybean with minimal effect on other traits, as we discovered clusters influence specific traits only. However, the molecular mechanisms of DE genes in domesticated soybean are related closely, thus it may be hard to change some genes independently from the others.

Although we used only 11 germplasm lines to minimize the time spent and to accomplish more in a short time, our results were approved and supported by previous studies` results. We suggest that future researches use a higher quantity of accessions and samples to prove and improve our findings. Even if the number of accessions cannot be enhanced, our results can be enough to examine and interpret from different perspectives. We believe that there must be some more comments or findings in our results waiting for being discovered to contribute to previous information about molecular mechanisms of crop domestication. Additionally, we were successful to annotate 1.255 DE genes, but failed to annotate the rest 148 transcripts and genes like *qPDHI* and *qSHI*. We are predicting those non-annotated genes have some more important characters involving in molecular mechanisms of plant domestication, and thus they need to be dissected and annotated properly.

Findings from this research demonstrated some genes, related to hydrolase activities that reduce lignin biosynthesis and pectin carbohydrate deposition in cell walls, play an important role in the molecular mechanisms of pod-shattering-related domestication traits. Previous studies

and our findings confirmed that those shatter-related hydrolases are enriched in the "extracellular region," "membrane," and "cytoplasm" in cells. Furthermore, although the expansin protein family has been previously expected to be involved in some different molecular mechanisms, our literature review showed that it has been for the first time evaluated in this thesis as an important cell wall tension inhibiter in the pods of soybeans. The cell wall tension relaxation aspect, as well as the significant findings mentioned above, highlight the value of this thesis in its contributions to the field.

6. CONCLUSION

Plants are needed to be well-known, particularly at the gene level, so that the increasing demand for high quality and quantity of food is met. For the soybean plant, the need to regain some lost genes or enhance the amount and effectiveness of genes influencing favorable traits, such as seed size and pod dehiscence, and the morphological and genomic diversity of domesticated and wild-type soybean has to be known well.

This study has investigated differentially expressed genes between those two types of soybean with the aim of deciphering the molecular mechanisms of crop domestication by comparatively analyzing the transcriptome obtained from both types of soybean by means of some bioinformatical processes. Findings from this research have not only supported previous comparative gene level-studies` results about soybean domestication but also attained novel discoveries and aspects about those domestication-related traits. As a result of this study, between 1,255 successfully annotated DEGs, some important genes have been recognized to have a high impact on those traits. Their roles in molecular activity in the cell have been partly demonstrated according to their GO ontology terms and KEGG metabolic enrichment pathways analysis results.

However, it should be kept in mind that we still have 148 DEGs waiting to be annotated and our findings reveal the remarkable possibility of that some of them are involved in some molecular mechanisms of those domestication traits. Although this study has provided valuable information about soybean genome and domestication mechanism of plants, we highly encourage researchers to either repeat our research workflow with more biological

samples and replications or deeply evaluate and dissect our findings, because we strongly believe that the findings of this study have more meanings, findings and deductions than what we have concluded so far. Moreover, this study most likely will be widened and intensified for PhD dissertation and would be mostly utilized even for post- doctoral work.

REFERENCES

Adamski N. M., Anastasiou E., Eriksson S., O'Neill C. M., Lenhard M., 2009. Local maternal control of seed size by KLUH/CYP78A5- dependent growth signaling. Proc Natl Acad Sci. USA 106(47):20115–20120.

Alexandratos N. and Bruinsma J., 2012. World agriculture towards 2030/2050: the 2012 revision. ESA Working Paper 3.

Alexandrova K. S., Conger B. V., 2002. Isolation of two somatic embryogenesis-related genes from orchardgrass (Dactylis glomerata). Plant Science 162, 301–307.

Ansorge W., Sproat B., Stegemann J., Schwager C., Zenke M., 1987. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Res. 1987 Jun 11; 15(11): 4593–4602.

Athanasios Theocharidis, Stjin van Dongen, Anton J. Enright & Tom C. Freeman, 2009. Network visualization and analysis of gene expression data using BioLayout Express 3D. doi:10.1038/nprot.2009.177.

Bader G. D. and Enright A. J., 2005. In Bioinformatics: A Practical Analysis of Genes and Proteins (ed. Baxevanis, A.D.) 540 (John Wiley, New York).

Blackman B. K., Rasmussen D. A., Strasburg J. L., Raduski A. R., Burke J. M., Knapp S. J., Michaels S. D., Rieseberg L. H., 2011. Contributions of flowering time genes to sunflower domestication and improvement. Genetics 2011, 187(1):271–287.

Braidwood Luke, Breuer Christian, Sugimoto Keiko, 2013. My body is a cage: mechanisms and modulation of plant cell growth. New Phytologist. 201: 388-402. doi:10.1111/nph.12473.

Brown, A.H.D., Grace, J.P., Speer, S.S., 1987. Designation of a core collection of perennial Glycine. Soybean Genet. Newsl. 14, 59–70.

Carroll S. B., 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134, 25–36.

Carter T. E., Jr., Nelson R., Sneller C. H., Cui Z.. Genetic diversity in soybean. Soybeans: Improvement, Production and Uses. In: Boerma HR, Specht JE, editors. Am Soc of Agro. Madison, Wisconsin: 2004. pp. 303–416.

Chachalis Demos and Smith M. L., 2000. Imbibition behavior of soybean (Glycine max (L.) Merrill) accessions with different testa characteristics. Seed Science and Technology, January 2000. 28(2):321-331.

Chen Y. W., Nelson R. L., 2004. Genetic variation and relationships among cultivated, wild, and semiwild soybean. Crop Science 44, 316–325.

Chena Yanyun, Pengyin Chen and Benildo G. de los Reyes, 2006. Differential Responses of the Cultivated and Wild Species of Soybean to Dehydration Stress. V ol. 46 No. 5, p. 2041-2046. doi:10.2135/cropsci2005.12.0466.

Chung Won-Hyong, Namhee Jeong, Jiwoong Kim, Woo Kyu Lee, Yun-Gyeong Lee, Sang-Heon Lee, Woongchang Yoon, Jin-Hyun Kim, Ik-Young Choi, Hong-Kyu Choi, Jung-Kyung Moon, Namshin Kim, and Soon-Chun Jeong. Population Structure and Domestication Revealed by High-Depth Resequencing of Korean Cultivated and Wild Soybean Genomes. DNA Res. 2014;21(2): 153–167. doi: 10.1093/dnares/dst047.

Circle, Sidney Joseph; Smith, Allan H. (1972). Soybeans: Chemistry and Technology. Westport, CT: Avi Publishing. pp. 104, 163. ISBN 0-87055-111-6.

Cobb, J. N., G. DeClerck, A. Greenberg, R. Clark, and S. McCouch, 2013. Next generation phenotyping: requirements and strategies for enhancing our understanding of genotype - phenotype relationships and its relevance to crop improvement. Theoretical and Applied Genetics 126: 867-887.

Conesa A., Gotz S., 2007. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. Hindawi Publishing Corporation International Journal of Plant Genomics Volume 2008, Article ID 619832, 12 pages doi:10.1155/2008/619832.

Cosgrove D. J., 2015. Plant expansins: diversity and interactions with plant cell walls. Current opinion in plant biology. 2015;25:162-172. doi:10.1016/j.pbi.2015.05.014.

Cui X., Churchill G. A., 2003. Statistical tests for differential expression in cDNA microarray experiments. Genome Biology. 4 (4): 210. doi:10.1186/gb-2003-4-4-210. PMC 154570 Freely accessible. PMID 12702200.

Daniel Koenig, José M. Jiménez-Gómez, Seisuke Kimura, Daniel Fulop, Daniel H. Chitwood, Lauren R. Headland, Ravi Kumar, Michael F. Covington, Upendra Kumar Devisetty, An V. Tat, Takayuki Tohge, Anthony Bolger, Korbinian Schneeberger, Stephan Ossowski, Christa Lanz, Guangyan Xiong, Mallorie Taylor- Teeples, Siobhan M. Brady, Markus Pauly, Detlef Weigel, Björn Usadel, Alisdair R. Fernie, Jie Peng, Neelima R. Sinha, and Julin N. Maloof, 2013. Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato.
PNAS July 9, 2013. 110 (28) E2655-E2662; https://doi.org/10.1073/pnas.1309606110

David L. Hyten, Qijian Song, Youlin Zhu, Ik-Young Choi, Randall L. Nelson, Jose M.

Costa, James E. Specht, Randy C. Shoemaker and Perry B. Cregan, 2006. Impacts of genetic bottlenecks on soybean genome diversity. PNAS November 7, 2006. 103 (45) 16666-16671.

Deepak K. Ray, Nathaniel D. Mueller, Paul C. West, Jonathan A. Foley, 2013. Yield Trends Are Insufficient to Double Global Crop Production by 2050. PLoS ONE 8(6): e66428. doi:10.1371/journal.pone.0066428

Disch S., Anastasiou E., Sharma V. K., Laux T., Fletcher J. C., Lanhard M., 2006. The E3 ubiquitin ligase BIG BROTHER controls Arabidopsis organ size in a dosage- dependent manner. Curr. Biol. 16:272–279.

Dixon, J., A. Gulliver, and D. Gibbon, (2001). "Farming Systems and Poverty: Improving Farmers' Livelihoods in a Changing World". Rome and Washington, DC: FAO and World Bank.

Doebley J.F., Gaut B.S., Smith B. D., 2006. The molecular genetics of crop domestication. Cell 127, 1309–1321.

Dong R., Dong D., Luo D., Zhou Q., Chai X., Zhang J., Xie W., Liu W., Dong Y., Wang Y., Liu Z., 2017. Transcriptome Analyses Reveal Candidate Pod Shattering- Associated Genes Involved in the Pod Ventral Sutures of Common Vetch (Vicia sativa L.). Front. Plant Sci. 8:649. doi: 10.3389/fpls. 2017.00649.

Dong R., Dong D., Luo D., Zhou Q., Chai X., Zhang J., Xie W., Liu W., Dong Y., Wang Y., Liu Z., 2017. Transcriptome Analyses Reveal Candidate Pod Shattering- Associated Genes Involved in the Pod Ventral Sutures of Common Vetch (Vicia sativa L.). Front. Plant Sci. 8:649. doi: 10.3389/fpls. 2017.00649.

Dong Y. and Wang Y-Z., 2015. Seed shattering: from models to crops. Front. Plant Sci. 6:476. doi: 10.3389/fpls.2015.00476.

Dong Y., Yang X., Liu J., Wang B. H., Liu B. L., Wang Y. Z., 2014. Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. Nat Commun. 2014; 5:3352. doi: 10.1038/ncomms4352.

Doyle J.J. and Lucknow M.A., 2003. The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiology 131:900-910.

Duan J., Xia C., Zhao G., Jia J., Kong X., 2012. Optimizing de novo common wheat transcriptome assembly using short-read RNA-seq data. BMC Genomics 13, 392.

Eisenstein Michael., 2012. Oxford Nanopore announcement sets sequencing sector abuzz. Nature Biotechnology 30, 295–296 (2012). doi:10.1038/nbt0412-295.

Epimaki M. K., Koinange, Shree P., Singh, and Paul Gept, 1996. Genetic Control of the Domestication Syndrome in Common Bean. Published in Crop Sci. 36:1037-1045.

Eric J. Sedivy, Faqiang Wu and Yoshie Hanzawa, 2016. Soybean domestication: The origin, genetic architecture and molecular bases. New Phytologist. 214: 539 553 doi: 10.1111/nph.14418.

Fang W. J., Wang Z. B., Cui R. F., Li J., Li Y. H., 2012. Maternal control of seed size by EOD3/CYP78A6 in Arabidopsis thaliana. Plant J. 70(6):929–939.

Fao, Global agriculture towards 2050, 2009. http://www.fao.org/fileadmin/templates/wsfs/docs/Issues_papers/HLEF2050_Global_Agriculture.pdf.

Freeman T. C., *et al.,* 2007. Construction, visualisation, and clustering of transcription networks from microarray expression data. PLoS Comput. Biol. 3, 2032 – 2042.

Friederike Dündar, Luce Skrabanek, Paul Zumbo, 2015. Introduction to differential gene expression analysis using RNA-seq. September 2015, updated March 20, 2018. 2015-2018 Applied Bioinformatics Core - Weill Cornell Medical College.

Fukuda Y., 1933. Cyto-genetical studies on the wild and cultivated Manchurian soy beans (Glycine L.). Jpn J Bot 6:489–506.

Funatsuki Hideyuki, Masaya Suzuki, Aya Hirose, Hiroki Inaba, Tetsuya Yamada, Makita Hajika, Kunihiko Komatsu, Takeshi Katayama, Takashi Sayama, Masao Ishimoto, and Kaien Fujino, 2014. Molecular basis of a shattering resistance boosting global dissemination of soybean. PNAS December 16, 2014. 111 (50) 17797-17802. https://doi.org/10.1073/pnas.1417282111

Garber M., Manfred G. Grabherr, Mitchell Guttman & Cole Trapnell, 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. NATURE METHODS VOL.8 NO.6. DOI:10.1038/NMETH.1613.

Ge Liangfa, Jianbin Yu, Hongliang Wang, Diane Luth, Guihu Bai, Kan Wang, and Rujin Chen, 2016. Increasing seed size and quality by manipulating BIG SEEDS1 in legume species. PNAS November 1, 2016. 113 (44) 12414-12419; published ahead of print October 17, 2016. https://doi.org/10.1073/pnas.1611763113.

Gentleman R. C., Carey V. J., Bates D. M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F., Li C., Maechler M., Rossini A. J., Sawitzki G., Smith C., Smyth G., Tierney L., Yang J. Y., Zhang J., 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004, 5: R80.

Götz S., Arnold R. Sebastián-León P. Martín-Rodríguez S. Tischler P. Jehl M. A. Dopazo J. Rattei T. Conesa A., 2011. B2G-FAR, a species-centered GO annotation repository. Bioinformatics. 27 (7): 919–24. doi:10.1093/bioinformatics/btr059. PMC 3065692 Freely accessible. PMID 21335611.

Guo J., Yunsheng Wang, Chi Song, Jianfeng Zhou, Lijuan Qiu, Hongwen Huang and Ying Wang. 2010. A single origin and moderate bottleneck during domestication of soybean (Glycine max): implications from microsatellites and nucleotide sequences. Annals of Botany 106: 505–514, 2010 doi:10.1093/aob/mcq125.

Haas Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew B. Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D. MacManes, Michael Ott. Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N. Dewey, Robert Henschel, Richard D. LeDuc, Nir Friedman, Aviv Regev. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc. 2013; 8(8): 10.1038/nprot.2013.084.

Holley Robert W., Jean Apgar, George A. Everett, James T. Madison, Mark Marquisee, Susan H. Merrill, John Robert Penswick and Ada Zamir, 1965. Structure of Ribonucleic acid. New Series, Vol. 147, No. 3664 (Mar. 19, 1965), pp. 1462-1465.

Hongye Li and Marilyn J. Roossinck. 2004. Genetic Bottlenecks Reduce Population Variation in an Experimental RNA Virus Population. JOURNAL OF VIROLOGY, Oct. 2004, p. 10582–10587 Vol. 78, No. 19 0022-538X/04/$08.000 DOI: 10.1128/JVI.78.19.10582–10587.2004.

Hymowitz T, Newell C: Taxonomy, speciation, domestication, dissemination, germplasm resources, and variation in the genus Glycine. In Advances in Legume Science. Edited by Summerfield RJ, Bunting AH. Kew, Richmond, Surrey: Royal Botanical Gardens; 1980:251-264.

Hymowitz T., and Shurtleff W.R., 2005. Debunking Soybean Myths and Legends in the Historical and Popular Literature. Crop Sci. 45:473–476.

Hymowitz T., Newell C., 1980. Taxonomy, speciation, domestication, dissemination, germplasm resources, and variation in the genus Glycine. In Advances in Legume Science. Edited by Summerfield RJ, Bunting AH. Kew, Richmond, Surrey: Royal Botanical Gardens; 1980:251-264.

Jung J., and Park W., 2013. Comparative genomic and transcriptomic analyses reveal habitat differentiation and different transcriptional responses during pectin metabolism in Alishewanella species. Appl. Environ. Microb. 79, 6351–6361. doi: 10.1128/AEM.02350-13.

Kanamaru Kyohei, Shaodong Wang, Jun Abe, Tetsuya Yamada, Keisuke Kitamura. 2006. Identification and Characterization of Wild Soybean (Glycine soja Sieb. et Zecc.) Strains with High Lutein Content. Online ISSN : 1347-3735. Print ISSN : 1344-7610.

Katz Y., Wang E. T., Airoldi E. M., Burge C. B., 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods 2010, 7(12):1009-15.

Kim M. Y., Van K., Kang Y. J., Kim K. H., Lee S. H.. Tracing soybean domestication history: From nucleotide to genome. Breed Science. 2012 Jan;61(5):445-52. doi: 10.1270/jsbbs.61.445.

Kim M.Y., Lee S., Van K., Kim T-H., Jeong S-C., Choi I-Y., Kim D-S., Lee Y-S., Park D., Ma J., 2010. Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. Proc. Natl. Acad. Sci USA.; 107:22032–22037.

Lagacé M., Matton D. P., 2004. Characterization of a WRKY transcription factor expressed in late torpedo-stage embryos of Solanum chacoense. Planta 219, 185–189.

Lam Hon-Ming, Xun Xu, Xin Liu, Wenbin Chen, Guohua Yang, Fuk-Ling Wong, Man- Wah Li, Weiming He, Nan Qin, Bo Wang, Jun Li, Min Jian, Jian Wang, Guihua Shao, Jun Wang, Samuel Sai-Ming Sun & Gengyun Zhang, 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nature Genetics volume 42, pages 1053–1059 (2010) doi:10.1038/ng.715.

Lavlu Mozumdar, 2012. Agricultural productivity and food security in the developing world, Bangladesh J. Agric. Econs. XXXV, 1&2(2012) 53-69.

Li Changbao, Ailing Zhou and Tao Sang, 2006. Rice Domestication by Reducing Shattering. Science 31 Mar 2006: Vol. 311, Issue 5769, pp. 1936-1939 DOI: 10.1126/science.1123604.

Li L.-F., Olsen K. M., 2016. Chapter Three - To Have and to Hold: Selection for Seed and Fruit Retention During Crop Domestication. Current Topics in Developmental Biology, Volume 119, 2016, Pages 63-10.

Li YH, Li W, Zhang C, Yang L, Chang RZ, Gaut BS, Qiu LJ., 2010. Genetic diversity in domesticated soybean (Glycine max) and its wild progenitor (Glycine soja) for simple sequence repeat and single-nucleotide polymorphism loci. New Phytol. 2010, 188: 242-253. 10.1111/j.1469-8137.2010.03344.x.

Li Ying-hui, Shan-cen Zhao, Jian-xin Ma, Dong Li, Long Yan, Jun Li, Xiao-tian Qi, Xiao- sen Guo, Le Zhang, Wei-ming He, Ru-zhen Chang, Qin-si Liang, Yong Guo, Chen Ye, Xiao-bo Wang, Yong Tao, Rong-xia Guan, Jun-yi Wang, Yu-lin Liu, Long- guo Jin, Xiu-qing Zhang, Zhang-xiong Liu, Li-juan Zhang, Jie Chen, Ke-jing Wang, Rasmus Nielsen, Rui-qiang Li, Peng-yin Chen, Wen-bin Li, Jochen C Reif, Michael Purugganan, Jian Wang, Meng-chen Zhang, Jun Wang and Li-juan Qiu. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genomics 2013 (14):579.

Li Yunhai, Leiying Zheng, Fiona Corke, Caroline Smith, and Michael W. Bevan, 2008. Control of final seed and organ size by the DA1 gene family in Arabidopsis thaliana. Genes & Dev. 2008. 22: 1331-1336 doi: 10.1101/gad.463608.

Li, B. & Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011).

Liang Du, Na Li, Liangliang Chen, Yingxiu Xu, Yu Li, Yueying Zhang, Chuanyou Li, Yunhai Lia, 2014. The Ubiquitin Receptor DA1 Regulates Seed and Organ Size by Modulating the Stability of the Ubiquitin-Specific Protease UBP15/SOD2 in Arabidopsis. The Plant Cell, Vol. 26: 665–677.

Lin Z., Li X., Shannon L. M., Yeh C. T., Wang M. L. and Bai G., 2012. Parallel domestication of the Shattering1 genes in cereals. Nat. Genet. 44, 720–724.doi: 10.1038/ng.2281.

Linus Pauling Institute, Soy isoflavones. Micronutrient Information Center, Oregon State University, Corvallis. 2016. Retrieved 23 May 2016.

Liu Baohui, Toshiro Fujita, Ze-Hong Yan, Shinichi Sakamoto, Donghe Xu, and Jun Abe. QTL Mapping of Domestication-related Traits in Soybean (Glycine max). Ann Bot. 2007 Oct; 100(5): 1027–1038. Published online 2007 Aug 7. doi: 10.1093/ aob/ mcm149.

Loren A., Honaas Eric K., Wafula Norman J., Wickett Joshua P., Der Yeting Zhang, Patrick P., Edger Naomi S., Altman J., Chris Pires, James H., Leebens-Mack, Claude W., dePamphilis, 2016. Selecting Superior De Novo Transcriptome Assemblies: Lessons Learned by Leveraging the Best Plant Genome. https://doi.org/10.1371/journal.pone.0146062.

Luo M., Dennis E. S., Berger F., Peacock W. J., Chaudhury A., 2005. MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in Arabidopsis. Proceedings of the National Academy of Sciences, USA 102, 17531–17536.

Ma H. and Wang S., 2016. Histidine Regulates Seed Oil Deposition through Abscisic Acid Biosynthesis and β-Oxidation. Plant Physiology. 2016;172(2):848-857. doi:10.1104/pp.16.00950.

Marioni J., Mason C., Mane S., Stephens M., Gilad Y., 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 11 Jun 2008 (doi:10.1101/ gr.079558.108).

Martin Jeffrey A., Wang Zhong., Next-generation transcriptome assembly. Nature Reviews Genetics. 12 (10): 671–682. doi:10.1038/nrg3068.

Maxam A. M. and Gilbert W., 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA. 74 (2): 560–64. Bibcode:1977PNAS. 74. 560M. doi:10.1073/pnas.74.2.560. PMC 392330. PMID 265521.

McQueen-Mason S. J., Cosgrove D. J., 1995. Expansin mode of action on cell walls. Analysis of wall hydrolysis, stress relaxation, and binding. Plant Physiol. 107 (1): 87–100. doi:10.1104/pp.107.1.87. PMC 161171 Freely accessible. PMID 11536663.

Michael I. Love, Wolfgang Huber and Simon Anders, 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. DOI 10.1186/s13059-014-0550-8.

Muhammad Amjad Nawaz, Hafiz Mamoon Rehman, Muhammad Imtiaz, Faheem Shehzad Baloch, Jeong Dong Lee, Seung Hwan Yang, Soo In Lee & Gyuhwa Chung, 2017. Systems Identification and Characterization of Cell Wall Reassembly and Degradation Related Genes in Glycine max (L.) Merill, a Bioenergy Legume.

Muriira N. G., Xu W., Muchugi A., Xu J., and Liu A., 2015. De novo sequencing and assembly analysis of transcriptome in the Sodom apple (Calotropis gigantea). BMC Genomics 16:723. doi: 10.1186/s12864-015-1908-3.

Mustakas, G. C. (1964). "Production and nutritional evaluation of extrusion-cooked full- fat soybean flour". Journal of the American Oil Chemists' Society. 41: 607–614. National Biodiesel Board. April 2008. Retrieved February 18, 2012.

Nguyen, Henry T., Bhattacharyya, Madan Kumar, 2017. The Soybean Genome. ISBN; 978-3-319-64198-0.

Nicolae M., Mangul S., Măndoiu I. and Zelikovsky A., 2010. Estimation of alternative splicing isoform frequencies from RNA-Seq data. In Algorithms in Bioinformatics, Lecture Notes in Computer Science. Edited by: Moulton V, Singh M. Liverpool, UK: Springer Berlin/Heidelberg; 2010:202-214.

Nutrient data laboratory". United States Department of Agriculture. Retrieved August 10, 2016.

Ogata H., Goto S., Sato K., Fujibuchi W., Bono H. and Kanehisa M., 1999. “KEGG: Kyoto Encyclopedia of Genes and Genomes,” Nucleic Acids Research, vol. 27, no. 1, pp. 29–34.

Olsen M. K. and Wendel J. F., 2013. A bountiful harvest: genomic insights into crop domestication phenotypes. Annual Review of Plant Biology, April 2013. Vol. 64:47-70.

Park, S.-Y., F. C. Peterson, A. Mosquna, J. Yao, B. F. Volkman, et al., 2015. Agrochemical control of plant water use using engineered abscisic acid receptors. Nature 520: 545-548.

Peipei Li, Yongjun Piao, Ho Sun Shon and Keun Ho Ryu, 2015. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinformatics201516:347 https://doi.org/10.1186/s12859-015-0778-7.

Prober J. M., Trainor G. L., Dam R. J., Hobbs F. W., Robertson C. W., Zagursky R. J., Cocuzza A. J., Jensen M. A., Baumeister K., 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science. 1987 Oct 16;238(4825):336-41.

Raboy V., Dickinson D. B. and Below F. E. Variation in Seed Total Phosphorus, Phytic Acid, Zinc, Calcium, Magnesium, and Protein among Lines of Glycine max and G. soja, 1984. Vol. 24 No. 3, p. 431-434. doi:10.2135/cropsci1984.0011183X002400030001x.

Ray, Jeffery D; Kilen, Thomas C; Abel, Craig A; Paris, Robert L., 2003. Soybean cross-pollination rates under field conditions. Environ. Biosafety Res. 2 (2003) 133138 ISBR, EDP Sciences, 2003DOI: 10.1051/ebr:2003005.

Raza G., Ahmad N., Hussain M., Zafar Y., Rahman M., 2016. Role of Genetics and Genomics in Mitigating Abiotic Stresses in Soybeans. DOI; 10.1016/B978-0-12-801535-3.00009-7.

Robert M. Stupar, 2010. Into the wild: The soybean genome meets its undomesticated relative. 107 (51) 21947-21948.

Robinson M. D., McCarthy D. J., Smyth G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139-140.

Romero I. G., Ruvinsky I., Gilad Y., 2012. Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews. Genetics 13, 505–516.

Romero I. G., Ruvinsky I., Gilad Y., 2012. Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews. Genetics 13, 505–516.

Ross-Ibarra J., 2005. Quantitative trait loci and the study of plant domestication. Genetics of Adaptation 2005, 123:197–204.

Rushton P. J., Somssich I. E., Ringler P., Shen Q. J., 2010. WRKY transcription factors. Trends in Plant Science 15, 247–258.

Sanger F., Brownlee G., Barrell B., 1965. A two-dimensional fractionation procedure for radioactive nucleotides. J. Mol. Biol., 13 (1965), p. 373-IN4.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. Genome sequence of the paleopolyploid soybean. Nature. 2010;463:178–183.

Shendure J. and Ji H., 2008. Next-generation DNA sequencing. Nat. Biotechnol., 26 (2008), pp. 1135-1145.

Shomura A., Izawa T., Ebana K., Ebitani T., Kanegae H., Konishi S., Yano M., 2008. Deletion in a gene associated with grain size increased yields during rice domestication. Nature Genetics 40, 1023–1028.

Shomura A., Izawa T., Ebana K., Ebitani T., Kanegae H., Konishi S., Yano M., 2008. Deletion in a gene associated with grain size increased yields during rice domestication. Nature Genetics 40, 1023–1028.

Silvas J. Prince, Li Song, Dan Qiu, Joao V. Maldonado dos Santos, Chenglin Chai, Trupti Joshi, Gunvant Patil, Babu Valliyodan, Tri D. Vuong, Mackensie Murphy, Konstantinos Krampis, Dominic M. Tucker, Ruslan Biyashev, Anne E. Dorrance, MA. Saghai Maroof, Dong Xu, J. Grover Shannon and Henry T Nguyen., 2015.
Genetic variants in root architecture-related genes in a Glycine soja accession, a potential resource to improve cultivated soybean. BMC Genomics 16:132.

Silvas J. Prince, Li Song, Dan Qiu, Joao V. Maldonado dos Santos, Chenglin Chai, Trupti Joshi, Gunvant Patil, Babu Valliyodan, Tri D. Vuong, Mackensie Murphy, Konstantinos Krampis, Dominic M. Tucker, Ruslan Biyashev, Anne E. Dorrance, MA. Saghai Maroof, Dong Xu, J. Grover Shannon and Henry T Nguyen., 2015. Genetic variants in root architecture-related genes in a Glycine soja accession, a potential resource to improve cultivated soybean. BMC Genomics 16:132.

Singh R.J., Kollipara K.P., Hymowitz T. (1987b) Polyploid complexes of Glycine tabacina (Labill.) Benth. and G. tomentella Hayata revealed by cytogenetic analysis. Genome 29:490–497.

Song X. J., Huang W., Shi M., Zhu M. Z., Lin H. X., 2007. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. Nature Genetics 39, 623–630.

Song X. J., Huang W., Shi M., Zhu M. Z., Lin H. X., 2007. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. Nature Genetics 39, 623–630.

Soyatech. Archived from the original on January 12, 2017. Soy Facts. Retrieved January 24, 2017.

Staden, R., 1979. A strategy of DNA sequencing employing computer programs. Nucleic Acids Research. 6 (70): 2601–10. doi:10.1093/nar/6.7.2601. PMC 327874. PMID 461197.

Steven L. Broich and Reid G. Palmer. A cluster analysis of wild and domesticated soybean phenotypes. Euphytica, 29 (1980) 23332.

Sudheer, Kumar M.; Sridhar, Reddy B.; Kiran, Babu S.; Bhilegaonkar, P. M.; Shirwaikar, A.; Unnikrishnan, M. K., 2004. Antiinflammatory and Antiulcer Activities of Phytic Acid in Rats. Indian Journal of Experimental Biolotgy. National Institute of Science Communication and Information Resources. 42 (2): 179–185. PMID 15282951.

Sun C., Palmqvist S., Olsson H., Borén M., Ahlandsberg S., Jansson C., 2003. A novel WRKY transcription factor, SUSIBA2, participates in sugar signaling in barley by binding to the sugar-responsive elements of the iso1 promoter. The Plant Cell 15, 2076–2092.

Suzuki Masaya, Kaien Fujino, Yumi Nakamoto, Masao Ishimoto, Hideyuki Funatsuki, 2009. Fine mapping and development of DNA markers for the qPDH1 locus associated with pod dehiscence in soybean. Molecular Breeding 25(3):407-418 DOI10.1007/s11032-009-9340-5.

Swerdlow H. and Gesteland R., 1990. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. Nucleic Acids Res., 18 (1990), pp. 1415-1419.

Tang H, Sezen U, Paterson AH. Domestication and plant genome. Curr Opin Plant Biol. 2010;13:160–166.

Tang H., Cuevas H. E., Das S., Sezen, U. U., Zhou C. and Guo H., 2013. Seed Shattering in a wild sorghum is conferred by a locus unrelated to domestication. Proc. Natl. Acad. Sci. U.S.A. 110, 15824–15829.doi:10.1073/pnas. 1305213110.

Tanja Magoč and Steven L. Salzberg, 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics, Volume 27, Issue 21, 1 November 2011, Pages 2957–2963, https://doi.org/10.1093/bioinformatics/btr507.

Teresa Lenser, Günter Theißen, 2013. Molecular mechanisms involved in convergent crop domestication. Trends in Plant Science,Volume 18, Issue 12, December 2013, Pages 704-714.

Trupti Joshi, Babu Valliyodan, Jeng-Hung Wu, Suk-Ha Lee, Dong Xu, and Henry T Nguyen. Genomic differences between cultivated soybean, G. max and its wild relative G. soja. BMC Genomics. 2013; 14(Suppl 1): S5. Published online 2013 Jan 21. doi: 10.1186/ 1471-2164-14-S1-S5.

Vanderauwera S, Vandenbroucke K, Inzé A, et al., 2012. AtWRKY15 perturbation abolishes the mitochondrial stress response that steers osmotic stress tolerance in Arabidopsis. Proceedings of the National Academy of Sciences, USA 109, 20113–20118.

Vucenik, Ivana; Shamsuddin, AbulKalam M., 2003. Cancer Inhibition by Inositol Hexaphosphate (IP6) and Inositol: From Laboratory to Clinic. The Journal of Nutrition. American Society for Nutrition. 133 (11): 3778S–3784S. PMID 14608114.

Wang Y., Gu Y., Gao H., Qiu L., Chang R., Chen S., He C., 2016. Molecular and geographic evolutionary support for the essential role of GIGANTEAa in soybean domestication of flowering time. BMC Evolutionary Biology 16, 79.

Wang Z., Gerstein M. and Snyder M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 10 (1): 57–63. doi:10.1038/nrg2484. PMC 2949280 Freely accessible. PMID 19015660.

Watson J. D. and Crick F. H. C., 1953. A Structure for Deoxyribose Nucleic Acid. April 25, 1953 (2), Nature (3), 171, 737-738.

Xia Q. *et al*., 2009. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). Science 326, 433–436.

Xiaobo Wang, Yinhui Li, Haowei Zhang, Genlou Sun, Wenming Zhang, Lijuan Qiu, 2015. Evolution and association analysis of GmCYP78A10 gene with seed size/weight and pod number in soybean. Mol Biol Rep. 42:489–496 DOI 10.1007/s11033-014-3792-3.

Xinpeng *et al*., 2014 Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing Nature Communications volume5, Article number: 4340 (2014) doi:10.1038/ncomms5340.

Xinxin Li, Jing Zhao, Zhiyuan Tan, Rensen Zeng, and Hong Liao, 2015. GmEXPB2, a Cell Wall β-Expansin, Affects Soybean Nodulation through Modifying Root Architecture and Promoting Nodule Formation and Development. Plant Physiology, December 2015, Vol. 169, pp. 2640–2653. ORCID ID: 0000-0002-6435-3054 (R.Z.).

Yongzhe Gu, Wei Li, Hongwei Jiang, Yan Wang, Huihui Gao, Miao Liu, Qingshan Chen, Yongcai Lai and Chaoying He, 2016. Differential expression of a WRKY gene between wild and cultivated soybeans correlates to seed size. Journal of Experimental Botany, Vol. 68, No. 11 pp. 2717 2729, 2017 doi:10.1093/jxb/erx147.

Yoo Mi-Jeong and Wendel Jonathan F., 2014. Comparative Evolutionary and Developmental Dynamics of the Cotton (Gossypium hirsutum) Fiber Transcriptome. https://doi.org/10.1371/journal.pgen.1004073

Yoon, J., Cho, L. H., Kim, S. L., Choi, H., Koh, H. J., and An, G., 2014. The BEL1- Type homeobox gene SH5 induces seed shattering by enhancing abscission- Zone development and inhibiting lignin biosynthesis. Plant J. 79, 717–728.doi: 10.1111/tpj.12581.

Yoon, Jane H.; Thompson, Lilian U.; Jenkins, David J. A., 1983. The Effect of Phytic Acid on In Vitro Rate of Starch Digestibility and Blood Glucose Response. American Journal of Clinical Nutrition. American Society for Nutrition. 38 (6): 835–842. PMID 6650445.

Yu F, Huaxia Y, Lu W, Wu C, Cao X, Guo X., 2012. GhWRKY15, a member of the WRKY transcription factor family identified from cotton (Gossypium hirsutum L.), is involved in disease resistance and plant development. BMC Plant Biology 12, 144.

Zhao Q. Y., et al., 2011. Optimizing de novo transcriptome assembly from short-read RNA-seq data: a comparative study. BMC Bioinformatics 12 (suppl. 14), S2 (2011).

Zhao T., Gai J., Zhongguo N.K., 2004. The origin and evolution of cultivated soybean; 37(7):954-962.

Zhou Y., Lu D., Li C., Luo J., Zhu B. F. and Zhu J., 2012. Genetic control of seed shattering in rice by the APETALA2 transcription factor shattering abortion1. Plant Cell 24, 1034 1048. doi:10.1105/ tpc. 111. 094383.

Zhou Zhengkui, Yu Jiang, Zheng Wang, Zhiheng Gou, Jun Lyu, Weiyu Li, Yanjun Yu, Liping Shu, Yingjun Zhao, Yanming Ma, Chao Fang, Yanting Shen, Tengfei Liu, Congcong Li, Qing Li, Mian Wu, Min Wang, Yunshuai Wu, Yang Dong, Wenting Wan, Xiao Wang, Zhaoli Ding, Yuedong Gao, Hui Xiang, Baoge Zhu, Suk-Ha Lee, Wen Wang & Zhixi Tian, 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nature Biotechnology volume 33. doi: 10.1038/ nbt.3096.