CHARACTERIZATION OF PROGENY DERIVED FROM DISOMIC ALIEN ADDITION LINES FROM INTERSUBGENERIC CROSS BETWEEN *GLYCINE MAX* AND *GLYCINE TOMENTELLA*

BY

SUFEI WANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Crop Sciences
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Master's Committee:

Professor Randall L. Nelson
Assistant Professor Steven J. Clough
Professor A. Lane Rayburn

Abstract

Disomic alien addition lines (DAALs, 2n=42) were obtained from an intersubgeneric cross between *Glycine max* [L.] Merr. cv. Dwight (2n=40, G1G1) and *Glycine tomentella* Hayata (PI 441001, 2n=78, D3D3CC). They are morphologically uniform but distinct from either of the parents. These DAALs were all derived from the same monosomic alien addition line (MAAL, 2n=41), and theoretically they should breed true because they had a pair of homologous chromosomes from *G. tomentella* and 40 soybean chromosomes. However, in some selfed progenies of DAALs the extra *G. tomentella* chromosomes were eliminated resulting in plants with 2n=40 chromosomes. These progeny lines (2n=40) have a wide variation in phenotypes. The objective of this research was to document the phenotypic and chromosomal variation among the progeny of these DAALs, and to understand the genetics behind this phenomenon. In the replicated field study, variation was observed among the disomic progenies for the qualitative traits such as flower, seed coat, hilum, pod, and pubescence color, and stem termination; as well as the quantitative traits protein and oil concentrations, plant height, lodging, and time of maturity. Three disomic lines had protein concentrations significantly high than either the DAAL or Dwight. Studying the plant transcriptome via RNA-sequencing documented that many genes that are critical to fundamental plant growth processes and related to stress and defense responses were differentially expressed between the DAAL (LG13-7552) and one of the disomic progeny (LG12-7063). RNA-sequencing data indicated that the gray pubescence of LG12-7063 was not due to sequence change from *T*- to *t t* genotype, but the result of altered gene expression. The expression of *G. tomentella* sequences and higher expression of transposable elements (TEs) in the DAAL were also documented.

For my family,

for always believing in me.

Acknowledgements

I can never thank my advisor Dr. Randall Nelson enough for all his help in writing, research, fieldwork, and especially in teaching me how to be a qualified scientist. I am grateful to my committee members Dr. Steven Clough and Dr. A. Lane Rayburn for insightful feedback on my research and guidance throughout my Master's program.

For making this thesis possible, I give my special thanks to Dr. Ram Singh who created the progenitor plants from which the lines in my research were derived, and to Mohammad Belaffif who has served as the TA in my Bioinformatics class and private tutor during my RNA-sequencing data analysis.

I owe the success of this project to everyone who has offered me help over the years. I would like to thank Abraham Akpertey, Stephen Swarm and the farm crew in the Nelson lab for their help in the farm and in the field, David Neece in the Clough lab for helping my bench work, and Dr. Matt Hudson and Dr. Gopal Battu in the Hudson lab for sequencing data interpretation.

Table of Contents

Literature Review

Glycine taxonomy

The taxonomic hierarchy of cultivated soybean is as follow: (Retrieved [Feb 6th, 2015], from the Integrated Taxonomic Information System on-line database, http://www.itis.gov.)

Kingdom: *Plantae*
Division: *Tracheophyta*
Subdivision: *Spermatophytina*
Class: *Magnoliopsida*
Superorder: *Rosanae*
Order: *Fabales*
Family: *Fabaceae*
Genus: *Glycine Willd.*
Species: *Glycine max* (L.) Merr.

The genus *Glycine Willd.* has two subgenera; the subgenus *Soja (moench) F.J. Herm.* (annuals) and the subgenus *Glycine* (perennials). The subgenus *Soja* contains two annual species, soybean (*G. max* [L.] Merr.), and wild soybean (*Glycine soja* Seib. & Zucc.), and each contains 2n=40 chromosomes. Although described as two different species they are cross compatible and produce viable, vigorous and fertile F1 hybrid except for some lines with paracentric inversions (Ahmad et al., 1977, 1979) or reciprocal translocations (Karasawa, 1936; Palmer et al., 1987; Singh and Hymowitz, 1988). Pachytene chromosomes were all normally paired in crosses between soybean and wild soybean except for small structural differences for chromosome 6 and 11 (linkage group C2 and B1) (Singh and Hymowitz, 1988; Cregan et al., 1999). Based on the similarity and differences, *G. soja* and *G. max* were designated genome GG and G1G1, respectively (Singh et al., 2007). Molecular data confirmed that *G. soja* and *G. max* have the same genome and the ITS sequence divergence of rDNA is only 0.2% (Kollipara et al, 1997).

The subgenus *Glycine* contains 26 perennial *Glycine* species. They are indigenous to Australia and are found on some South Pacific Islands. These twining or trailing herbs are highly diverse morphologically and genetically and have adapted to a wide range of climatic conditions. Some of them may involve stolon or rhizome development. All of them have purple flowers; however, there is tremendous variation in color intensity among species (Hymowitz, 2004). Of these 26 perennial species, *G. pescadrensis* Hayata and *G. dolichocarpa* Tateishi and Ohashi have 80 chromosomes. *G. dolichocarpa* was collected from Taiwan by Tateishi and Ohashi, but it is considered endemic to Australia (Hymowitz, 1990). *G. hirticaulis* Tindale and Craven and *G. tabacina* (Labill.) Benth. have cytotypes of 40 and 80 chromosomes. And *G. tomentella* Hayata has accessions of chromosome number 2n = 38, 40, 78, and 80. All the rest have 40 chromosomes. A list of all the perennial species is in Table 1 (Chung and Singh, 2008).

The concept of three gene pools was developed to clarify the taxonomic and evolutionary relationships among a cultigen, a domesticated species, and its wild relatives (Harlan and de Wet, 1971). It groups species into primary (GP-1), secondary (GP-2) and tertiary (GP-3) gene pools based on the difficulty of making successful crosses between the species. GP-1 consists of species that can be easily crossed to produce vigorous hybrids. The hybrid shows normal meiotic pairing, produces fertile seeds, and genetic recombination is normal. According to this definition, all soybean and wild soybean germplasm are considered to be GP-1. GP-2 includes species between which gene transfer is relatively difficult because their hybrid progenies are weak and mostly sterile, and their meiotic chromosomes pairings are not normal. It takes considerable effort to recover the desired traits in advanced generations. Based on this concept, soybean has no known GP-2 relative. At the GP-3 level, hybridization is feasible; however, hybrids are anomalous, lethal or completely sterile. Genetic exchange is extremely difficult and requires radical measures, such as chromosome doubling, embryo rescue and bridging species to restore partial fertility. GP-3 is the extreme outer limit of accessible genetic resources. All 26 perennial *Glycine* species are considered the GP-3 of soybean. These species represent a very useful source of improving genetic diversity of soybean, but they have not yet been exploited (Singh et al, 1998).

Genomic relationships within subgenus *Glycine*

The understanding of genomic relationships among species provides information about the genome composition of ancestral species and helps us comprehend the evolution of the genus. Originally, identification and nomenclature of *Glycine* species were accomplished based on classical taxonomy. Due to the diversity among species, more complete phylogenetic relationships can only be defined by combining classical taxonomy with cytogenetics, molecular approaches and proteomics (Singh, 2003). The genomes of the *Glycine* species were designated with capital letters (A, B, C, D, E, H, G, and F) (Singh and Hymowitz, 1985). The designations were based on the genome affinity, and similar symbols are assigned to species with hybrids that exhibit normal chromosome pairing during meiosis, which indicates greater chromosome homoeology between the two. Molecular approaches assisted in designation when obtaining cytogenetic information is impossible (Singh et al., 1988, 1992; Kollipara et al., 1995, 1997; Brown et al., 2002; Doyle et al., 2002). A summary of the symbols for each species is provided in Table 1.

In the early stages of research, the genomic relationships were determined by interspecific crossability and chromosome pairing in interspecific hybrids, which provided direct information on phylogenetic relationship between the parental species. In general, hybrids from species with homoeologous genomes have complete or nearly complete pairing (bivalent) and normal chromosome migration to anaphase I poles. A small degree of chromosomal interchanges or paracentric inversions can occur. F1 hybrids from genomically distant species are commonly recovered using *in vitro* methods. Seed inviability, seedling lethality and sterility are common problems associated with intergenomic crosses (Newell and Hymowitz, 1983; Grant et al., 1984; Singh and Hymowitz, 1985, Singh et al., 1987; Kollipara et al., 1993). Meiotic chromosome pairing in the intergenomic hybrid varies greatly. Nineteen intraspecific and 30 interspecific F1 hybrids were produced among *G. canescens* (A2)*, G. clandestina* (A1)*, G. tomentella* (2n=78, 80; D3E, AE, EH2, DA6, DD2, H2), *G. falcata* (F), and *G. tabacina* (2n=40, 80; B, BB1, BB2, B1B2) by Putievsky and Broue (1979) in the first research on genomic relationships among perennial *Glycine* species. In later studies, all hybrids produced within A and B genomes exhibited 20 bivalents at metaphase I in most sporocytes (Putievsky and Broue, 1979; Newell and Hymowitz, 1983; Grant et al., 1984; Singh and Hymowitz, 1985; Singh et al., 1988, 1992, 2007). Crosses between genome A and B species exhibit an average chromosome pairing of 10.2 II + 19.7 I for *G. argynaria* (A2) x *G. latifolia* (B1) and 9.5 II + 20.9 I for *G. canescens* (A2) x *G. latifolia* (B1) (Singh et al., 1988). This pairing pattern shows genome A and B are distant from each other, yet the 50% bivalent pairing between the two genomes indicates that half of the genome might have come from the same progenitor. *G. falcata* has a unique genome (F) because it showed minimal chromosome association with *G. latifolia* (B1, 1.1 II + 37.8 I), *G. canescens* (A, 0.6 II+38.7 I), and *G. clandestina* (A1, 1.85 II + 36.1 I + 0.05 III) (Putievsky and Broue, 1979; Singh et al., 1998). No viable progeny was obtained from *G. falcata* (F) x *G. tabacina* (2n=40; B, BB1, BB2, B1B2). Nonviable hybrids were also reported with *G. canescens* (A2) and with *G. tomentella* (2n=40; D1A, H2, D2) (Newell and Hymowitz, 1983). These studies supported the uniqueness of *G. falcata* (F). It is likely that *G. falcata* (F) originated from a different progenitor and the subgenus of *Glycine* was formed through multiple independent events. Interspecific hybridization between species with similar genomes tends to form normal pod and produce fertile seeds. The F1 hybrids are usually vigorous and fertile. On the contrary, in crosses between distant species, pod or seed abortion is common, or the F1 hybrid is sterile. It was expected that crosses between *Glycine* species with morphological resemblance, or with genomes of similar letters, set normal pods and produce fertile F1 plants; however, this is not true among these species. In a study, out of 748 flowers that were pollinated between *G. cyrtoloba* (C1) and *G. curvata* (C), all gynoecia died after 2-3 days and no successful cross was recorded (Singh et al., 1992), although *G. cyrtoloba* (C1) and *G. curvata* (C) have nearly identical morphology and were both assigned genome C.

In recent decades, molecular techniques have been powerful tools for determining the genome resemblance among species to supplement the conventional techniques when conducting interspecific hybridization or obtaining F1 hybrid is impractical or extremely different. The earliest study employing isozyme groups confirmed the genomic relationships among *G. canescens, G. clandestina*, and *G. tomentella* (Kollipara et al., 1997). Research on the variation of nucleotide sequences in the internal transcript spacer (ITS) region of nuclear ribosomal DNA (nrDNA) revealed the phylogenic relationships among 16 species in the subgenus and provided evidence to assign genome symbols to *G. arenaria* (H)*, G. hirticaulis* (H1)*, G. pindanica* (H2)*, G. albicans* (I) *and G. lactovirens* (I1) (Kollipara et al., 1997). Little information was available for these five species, because only a few accessions were available, and they are very difficult to grow in the greenhouse. Based on histone 3-D gene sequences, *G. aphyonota* was assigned genome I3, *G. peratosa* was A5, *G. pullenii* was H3, and *G. stenophita* was B3 (Brown et al., 2002; Doyle et al., 2002).

G. tomentella Hayata

Among the few perennial species that have been successfully hybridized with *G. max* (Grant et al., 1986; Newell et al., 1987), *Glycine tomentella* Hayata is a unique species, because it is a polyploid complex that has four different cytotypes including aneudiploid (2n=38), diploid (2n=40), aneutetraploid (2n=78) and tetraploid (2n=80) (Singh et al., 1985). *G. tomentella* is an extremely variable species and widely distributed in Australia, and found in China, Papua New Guinea, Philippines, and Islet of Kinmen (Quemoy). *G. tomentella* was first included as a species complex in subgenus *Glycine* in 1981 (Hymowitz, 2004) and it has been characterized as the most compatible species to cross with soybean compare to other perennial species (Ladizinsky et al., 1979). Since then, various studies have been conducted regarding the genomic relationships within the species (Doyle and Brown, 1985; Doyle et al., 1986; Singh et al., 1988; Kollipara et al., 1993, 1997; Singh et al., 1998) and as well as its agronomic value (Loux et al., 1987; Hartman et al., 1992; Riggs et al., 1998; Hartman et al., 2000).

Aneudiploid (2n=38) and diploid *G. tomentella* are distinct from each other cytogenetically and genomically, although classical taxonomy cannot separate them into different species. There are eight isozyme banding groups among aneudiploid and diploid *G. tomentella* (Doyle and Brown, 1985). The aneudiploid (2n=38) is restricted in Queensland, Australia, but the diploid (2n=40) has wider geographic distribution in Papua New Guinea and Australia, including Queensland, Northern Territory and Western Australia. The isozymes divided aneudiploids into two groups, D1 and D2 (Doyle and Brown, 1985). These two D groups carry a similar genome but are disparate from the other 6 isozyme groups of diploid *G. tomentella* and were assigned to the E genome because they have good chromosome association at metaphase I when crossed with each other (Singh et al., 1988). The diploids (2n=40) were considered a species complex by cytogenetics and molecular studies (Kollipara et al., 1993, 1997). They were grouped into 6 isozyme groups, which are D3 (A, B, C), D4, D5 and D6 (Doyle and Brown, 1985). Biochemical analysis has shown that D1, D2 and D3 are genomically similar (Kollipara et al., 1993); however in a cross between the aneudiploid (D1, D2) and diploid (D3) *G. tomentella*, pod abortion and limited meiotic chromosome pairing was documented (Singh et al., 1988, 1998). The D4 group *G. tomentella* has close affinity to A-genome species, but it has less identity to A-genome species in morphology except for the long and narrow leaf and longer pod (Singh et al., 1998). Therefore, the D4 group had been considered close to A-genome species and later it was classified as *Glycine syndetika* (A6) (Pfeil et al., 2006; Singh et al., 2007). The D5 group contains highly heterogeneous accessions from Western Australia (Singh et al., 1998), but no viable seed was produced from crossing it with the other diploid groups. Tindale (1986) separated the D6 group collected in the Eastern Kimberley District of Western Australia and assigned it as an independent species *G. arenaria* Tindale.

Polyploid *G. tomentella* including aneutetraploid (2n=78) and tetraploid (2n=80) were proposed to originate from allopolyploidization. Based on classical taxonomy, they are indistinguishable from the diploids. Also, their chromosomes pair normally in bivalents at metaphase I the same as diploids (Singh and Hymowitz, 1985). Compare to diploids (2n=38 and 2n=40), tetraploids (2n=78 and 2n=80) are more morphologically and geologically diverse (Chung and Singh, 2008). Hybrids within aneutetraploid and tetraploid groups have normal chromosome pairing. Isozyme banding patterns suggested three isozyme groups in both aneutetraploid (T1, T5 and T6) and tetraploid (T2, T3 and T4) *G. tomentella* (Doyle and Brown, 1985; Doyle et al., 1986). Later, tetraploid accessions from Timor Island of Indonesia were assigned the T7 group; however, they were not examined for isozyme banding patterns (Kollipara et al., 1994). Among the aneutetraploids, the rate of producing mature pods from inter-isozyme group crosses (T1 x T5, T1 x T6 and T5 x T6) ranges from 3.8-10%, and Kollipara et al. (1994) determined that the aneutetraploid T1, T5 and T6 groups have a common genome EE (2n=38 *G. tomentella*) through chromosome association study. T1 accessions are mainly from Queensland. T5 was only collected from New South Wales and T6 in Western Australia (Chung and Singh, 2008). Among tetraploids the intra-isozyme group hybrids are fertile, but hybrids between T2 and T3, T3 and T4, and T2 and T7 were completely sterile. T3 and T7 genomes were not completely paired; however, a maximum of 30 II + 20 I chromosome associations at metaphase I was observed, and yielded several mature seeds. This partial pairing indicates a common genome may be present between T3 and T7 groups, yet geographical barrier had an irreplaceable effect on the divergence (Kollipara et al., 1994). Since a 2n=40 cytotype was never identified on Timor Island, an independent origination of T7 group from other tetraploids cannot be supported (Chung and Singh, 2008). To further confirm the ancestors of all aneutetraploids and tetraploids, amphidiploids were produced from possible parental species (Singh et al, 1987, 1989). By doubling the somatic chromosomes of 2n=39 and 2n=40 F1 hybrids, aneuallotetraploids (DDEE, AAEE; 2n=78) and allotetraploid (AADD; 2n=80) were synthesized and they were fertile and functioned like diploids (bivalent chromosome pairing) as well. These synthesized tetraploids were crossed with T1 (2n=78), T5 (2n=78) and T2 (2n=80) group accessions (Singh et al., 1989). Meiosis was normal and the progenies were fertile. Thus, genome letters were designated to T1 (D3D3EE), T5 (AAEE), and T2 (AAD3D3), which was later supported by molecular data (Kollipara et al., 1994). Considering the diversity of the tetraploids, as well as of the diploid genome donors, the polyploidy complex of *G. tomentella* may have been generated from multiple chromosome doubling events (Kollipara et al., 1994). Besides, these isozyme groups (T1, T2, T3, T4, T5, T6 and T7) may have multiple origins, which is supported by molecular studies (Rauscher et al., 2004).

Due to its complexity and variability in genetic composition, growth habits, and geographical distribution, *G. tomentella* harbors numerous desirable traits, including disease resistance (Burdon and Marshall, 1981; Zheng et al, 2005), pest resistance (Zhuang et al., 1996; Bauer et al., 2007) and stress resistance (White et al., 1990; Kao et al., 2005), which makes it a promising source for soybean improvement. For example, *G. tomentella* Hayata PI441001 (2n=78) was identified to be resistant to 3 Australian isolates of soybean rust (Schoen et al., 1992). And it has been discovered that PI441001 resists soybean rust by producing a growth inhibitor to fungus spores (Bilgin et al., 2008), which was isolated and applied to rust susceptible soybean lines as a natural fungicide. Also, some other *G. tomentella* accessions have been reported to have partial resistant to pathogens that cause Sclerotinia stem rot (*Sclerotinia sclerotiorum*), sudden death syndrome (*Fusarium solani*), (Hartman et al., 2000), and soybean rust (*Phakopsora pachyrhizi*) (Hartman et al., 1992). Resistance has also been shown to soybean cyst nematode (Riggs et al., 1998), and tolerance to glyphosate (Loux et al., 1987), 2, 4-D (Hart et al., 1991) and high chloride levels (Pantalone et al., 1997).

Hybridization between *G. max* and its perennial relatives

The wide hybridization needed to successfully introgress beneficial genes from perennial *Glycine*, which are species within the tertiary gene pool, into soybean is impossible unless radical techniques such as *in vitro* embryo rescue, chromosome doubling or bridging species are employed in obtaining fertile hybrids (Harlan and de Wet., 1971).

Even though the earliest wide hybridization in soybean trace back to 1979 (Broue et al., 1979; Ladizinsky et al., 1979), few fertile progenies and successful introgression have been reported until recently (Singh, 2010; Ratnaparkhe et al., 2011). Ladizinsky et al. (1979) attempted to cross soybean with five perennial species (*G. canescens, G. clandestine, G. falcata, G. tabacina, and G. tomentella*), but failed to produce viable F1 hybrids. Later, hybrids from intersubgeneric crosses with *G. canescens* (Broue et al., 1982; Grant et al., 1986; Newell et al., 1987)*, G. tomentella* (Newell and Hymowitz, 1982; Singh and Hymowitz, 1985; Sakai and Kaizuma, 1985; Newell et al., 1987; Chung and Kim, 1990; Bodanene-Zanettini et al., 1996; Hymowitz et al., 1998), *G. clandestine* (Singh et al., 1987), and *G. latifolia* (Chung and Kim, 1991) were obtained via *in vitro* technique; but they were all sterile. Singh et al (1990) successfully produced the first backcross-derived progenies from a synthetic amphidiploid (2n=118, GGDDEE) of *G. max* (2n=40, Altona) and *G. tomentella* (2n=78, PI483218). Monosomic alien addition lines (2n=41, MAALs) and modified diploids were screened and studied for introgressed genes responsible for favorable traits such as resistance to soybean rust, resistance to soybean cyst nematode and tolerance to salt and drought (Singh et al., 1998; Singh, 2007).

An improved protocol was developed to produce fertile progeny from crosses between soybean and perennial *Glycine* by Singh (2010, Figure 1), which facilitated the introduction of desirable genes from the tertiary gene pool to soybean. The F1 hybrid (2n=59, G1D3E) of *G. max* cv. Dwight (2n=40, G1G1) and *G. tomentella*, PI441001 (2n=78, D3D3EE) was rescued through *in vitro* embryo culture. It had 20 soybean chromosomes and 39 *G. tomentella* chromosomes, is vigorous but sterile. The hybrid was then treated with colchicine to double its chromosome number and produce the amphidiploid (2n=118, G1G1D3D3EE). Partial fertility was restored since the amphidiploid has two copies of all chromosomes that made up the original parental genomes. The amphidiploid was backcrossed to the *G. max* parent and to produce BC1 plants with 79 chromosomes (G1G1D3E), 40 from soybean and 39 from *G. tomentella.* Meiotic pairing showed that the soybean chromosomes were well paired and centered in the cells, while the chromosomes from the wild parent were floating free towards the pole. The BC1 generation was further backcrossed with *G. max* to produce the BC2 generation. The chromosome number of BC2 generations ranged from 50 to 60 with 40 soybean chromosomes and 10 to 20 *G. tomentella* chromosomes. BC2 plants were backcrossed to Dwight to produce BC3 plants, which contain 40 soybean chromosomes and one or more extra chromosomes from *G. tomentella*. To determine the actual chromosome number of each individual plant requires chromosome counting. These progenies include disomic lines (2n=40), monosomic alien addition lines (MAALs, 2n=41), disomic alien addition lines (DAALs, 2n=42) and some progenies with higher chromosome numbers, which generally require additional backcrossing to produce self-lines lines. Beneficial genes from *G. tomentella* possibly have been introgressed into the soybean chromosomes or carried by the additional *G. tomentella* chromosomes (Singh, 2010).

Application from other species

Crop wild relatives (CWR) are resources for novel variations or desirable traits especially species in secondary or tertiary gene pools. There has been a steady increase in the release of cultivars with CWR in pedigree as noted below; however, they remain underutilized, given the improved techniques for wide hybridization, increased number of accessible CWR in the gene banks and substantial literature on these CWR (Hajjar and Hodgkin, 2007).

Efforts in incorporating genetic materials from CWR have been made in other legume species to overcome the incompability of interspecific hybridization [Errico et al., 1991, 1996 (common bean); Ahmad et al., 1996 (chickpea); Campbell, 1997 (grass pea); Muños et al., 2004 (common bean); Gupta and Sharma, 2007 (lentil); Foncéka et al, 2009 (peanut); Smykal and Kosterin, 2010 (common bean)]. The use of *in vitro* embryo rescue overcame the post-fertilization interspecific barrier of crossing lentil (*Lens culinaria*) with *L. ervoides* and *L. nigricans* (Abbo and Ladizinsky, 1991) that was causing pod abortion (Fratini and Ruiz, 2006; 2011) and helped obtain hybrids between chickpea (*Cicer arietinum*) and *C. bijugum, C. judaicum and C. pinnatifidum* (Verma et al., 1995; Ahmad and Slinkard, 2004). Embryo rescue was also used to create hybrids from crosses between common bean (*Phaselus vulgaria*) and tepary bean (*P. acutifolius*), *P. coccineus* and *P. dumosus* and the introgression of increased seed size, more variable color and drought and heat tolerance (Muños et al., 2004; 2006). Advanced backcross QTL and marker-assisted selection methods were applied to introduce genes that increased seed mineral concentration and produce arcelin and APA cotyledonary proteins to enhance insect resistance from wild Andean accessions into cultivated common beans (Blair et al., 2006, 2010; Blair and Izquierdo, 2012). Molecular markers were also used to avoid the reduced fertility due to reciprocal translocations and facilitate the gene transfer when crossing *P. vulgaris* directly with *P. fulvum* (Errico et al., 1996; Smykal and Kosterin, 2010).

Three main pathways were established to tackle the differences in ploidy levels among *Arachis* genus species. One was backcrossing hexaploid lines generated from doubling the chromosome number of diploid and tetraploid hybrid (Garcia et al., 1996; Gowda et al., 2002; Burrow et al., 2013) which was successfully applied into the introgression of resistance to root-knot nematode and other foliar diseases (rust and late leaf spot) from *A. cardenasii*. A second was synthesizing autotetraploid by colchicine treatment which was used to introduce root-knot nematode resistance (Singh, 1985; Simpson., 1991; Mallikarjuna et al., 2011). The third was creating allotetraploid by doubling the chromosome number of the diploid hybrid which was used to introgress resistance for root-knot nematode and developing genetic mapping populations (Simpson et al., 1993; Fávero et al., 2006; Mallikarjuna et al., 2011). Although multiple methods of utilizing CWR have been incorporated in various legume crops, alien addition lines (AALs) have yet to be reported in any legume species except for soybean (Singh et al., 1998).

Constructing AALs as intermediate material is a common approach for introgressing genes through wide hybridization in families other than legume (Chang and de Jong, 2005). Through AALs, disease resistance to rust from *Aegilos ovata* and *Psathyrostachys huashanica* (Dhaliwal et al., 2002; Du et al., 2013), resistance to barley yellow dwarf virus from *Thinopyrum intermedium* (Larkin et al., 1995), resistance to *Fusarium* head blight from *Leymus racemosus* (Qi et al, 2008), resistance to wheat streak mosaic virus from *Aegilops speltoides* (Friebe et al, 1990), resistance to powdery mildew from *Elytrigia intermedium* (Luo et al., 2009), resistance to greenbug and curl mite and resistance to salinity from *Leymus multicaulis* (Zhang et al, 2006) were transferred to wheat. In rice, resistance to bacterial blight, brown planthopper, and whitebacked planthopper from *O. latifolia* were transferred. Derived progenies were observed to retain alien morphologies such as long awns, early maturity, black hulls, purple stigma and apiculus. Similar goals were also achieved by crossing rice with *O. officinalis* (Jena and Khush, 1990) and *O. australiensis* (Multani et al., 1994). The potential of AALs for breeding projects mainly depends on the genetic distance of the parents, and hence, the possibility of recombination between the alien chromosome and its homoeologous counterpart (Chang and de Jong, 2005). Additional limitations to crossing over are the heterochromatic pericentromeric chromosome regions (Chetelat et al., 2000) and chromosomal rearrangements such as duplication, inversion and translocations, (Tanksley et al., 2002; Ji and Chetelat, 2003).

The use of AALs is not limited to hybridization for crop improvement, but also for chromosome characterization, such as gene/marker localization (Jacobsen et al., 1995; Suen et al., 1997; Fox et al., 2001; Zhang et al., 2002), construction of chromosome specific libraries (Ananiev et al., 1997), and heterologous gene expression (Muehlbauer et al., 2000). Full sets of MAALs were used to characterize each alien chromosome in the genetic background of a distant relative. Full MAAL sets were developed in beet (*Beta vulgaris*) from *B. webbiana*, *B. patellaris* and *B. procumbens* (Reamon Ramos and Wricke, 1992; Mesbah et al., 1997; van Geyt et al., 1988), tomato (*Lycopersicon esculentum*) from *Solanum lycopersicoides* (Chetelat et al., 1998), potato (*Solanum tuberosum*) from tomato (Ali et al., 2001), rice (*Oryza sativa*) from *Oryza officinalisi* (Jena and Khush, 1989) and *Oryza latifolia* (Multani, et al., 2003), oat (*Avena sativa*) from maize (*Zea mays*) (Kynast et al., 2001), *Allium fistulosum* from onion (*Allium cepa*) (Shigyo et al., 1996) and different sets in wheat (*Triticum aestivum*) from species including rye (*Secale cereal*) and barley (*Hordeum vulgare*) (Friebe et al., 2000; Shepherd et al., 1988).

A full set of nine rapeseed-radish DAALs were developed to verify the resistance against root-knot nematodes from radish (*Raphanus sativus*) alien chromosomes in which two AALs have significantly reduced susceptibility (Zhang et al., 2014). Wheat-barley addition lines were used to assign the physical location of 1,257 barley genes using transcriptome analysis (Bilgic et al., 2007). Wheat alien addition lines were also used to obtain chromosome arm-specific BAC libraries of rye (Martis et al., 2013) and barley (Mayer et al., 2011) with highly reduced sequences (Kubaláková et al, 2003; Suchánková et al., 2006; Doležel et al., 2012). Oat DAALs with maize chromosome 9 were used to construct a cosmid library of maize chromosome-specific sequences, in which 29 out of 5000 clones were shown to contain maize DNA and eight of them produced a chromosome specific pattern that could be used for chromosome identification (Ananiev et al., 1997). Oat MAALs with maize chromosome 3 were used to study the ectopic expression of the maize *liguleless 3* homoeobox gene which resulted in a few characteristic morphological abnormalities of leaf and panicle, and the outgrowth of axillary buds in the MAALs (Muehlbauer et al., 2000).

In the identification of AALs, alien chromosomes are sometimes distinguishable by morphological traits [Jena and Khush, 1989 (rice); Morgan, 1991 (ryegrass); Reamon Ramos and Wricke, 1992 (beet); Shigyo et al., 1996 (Onion); Mesbah et al., 1997 (beet)], but sometimes require other approaches, such as isozyme markers [Quiros et al., 1987 (*Brassica oleracea*); Peffley and Currah, 1988 (Onion); Reamon Ramos and Wricke, 1992 (beet)], molecular markers including RFLPs [Garriga-Caldere et al., 1997 (potato); Friebe et al., 2000 (wheat); Jia et al., 2002 (wheat)], RAPDs [Jorgensen et al., 1996 (kale); Kaneko et al., 2000 (radish)], AFLPs [van Heusden et al., 2000 (onion)], microsatellites [Hernandez et al., 2002(wheat)] and repetitive sequence DNA fingerprints [Riera-Lizarazu et al., 1996 (oat); Mesbah et al., 1997 (beet)], and cytogenetics approaches including Giemsa C- or N- banding [Darvey and Gustafson, 1975 (wheat)], genomic *in situ* hybridization (GISH) [Schwarzacher et al., 1989 (wheat); Jacobsen et al., 1995 (potato); Gao et al., 2001 (beet)] and fluorescence *in situ* hybridization (FISH) [Dong et al., 2001 (potato)].

The problem associated with AALs is the difficulty in maintenance due to sterility, inferior viability and low chromosome transmission through the germline (Chang and de Jong, 2005). Seed fertility in AALs is reduced in crosses with recipient parents with lower ploidy level because of more imbalance of chromosome pairings (Islam et al., 1981), less genomic affinity between the parents (Blanco et al., 1987) and the genetic constitution of the added chromosomes (Islam et al., 1981; Miller et al., 1984; Singh, 2002). In MAALs, the alien chromosome usually fails to synapse and/or recombine in meiosis, lags behind in the equatorial plane and is lost at later stages of meiosis (Sybenga, 1992). Transmission rates are usually higher through the female than the male and vary among the addition sets (Multani et al., 1994; Shigyo et al., 2003; Ji and Chetelat, 2003) and among different alien chromosomes, in which larger chromosomes usually have higher transmission rates (Garriga-Caldere et al., 1998). Therefore, MAALs are usually kept *in vitro* or reproduce only vegetatively, if possible, for example in potato, onion and beets (Chang and de Jong, 2005). In the case of *O. sativa* x *O. latifolia*, the appearance of dominant mutants such as black-hulled plants and segregation distortion exhibited in various traits involving awns and hull color, plant pigmentation, grain shattering and height in the derived disomic progenies from the MAAL instigated the search for the origin of the newly emerged traits. The activation of transposable elements (TEs) is the primary suspect behind this phenomenon; however, there was no evidence or other convincing explanation available (Multani, et al., 2003).

In DAALs, which can only be obtained either when the extra chromosome in MAAL is transmitted through both female and male gametes or meiotic non-disjunction of the alien chromosome occurs in the MAAL parent, the extra two homologous chromosomes can have normal chromosome pairing and segregate at anaphase I (Chang and de Jong, 2005). DAALs are expected to breed true; however, in some cases they have been reported to fail to pair or form chiasmata, resulting in univalent and hence imbalanced gametes (O'Mara, 1940; Khush, 1973; Miller, 1984). A decrease of wheat and rye homoeologous chromosomes pairing has been observed in wheat-rye DAALs. It was proposed to be influenced by factors such as genes controlling meiotic pairing, heterochromatin, and cryptic wheat-rye interactions (Orellana et al., 1984) indicating that DAALs are not always stable. The chromosome pairing and transmission rate vary among DAALs and are determined by the alien chromosomes with the highest rate in DAALs with E chromosomes (99%)

and lowest in A (77%) in wheat-barley DAALs (Islam et al., 1981) and a range between 86% to 97% in wheat-*Agropyron elongatum* DAALs (Dvorak and Knott, 1974).

Glycine max - *G. tomentella* DAALs

Seeds from the first three confirmed DAALs (2n=42) from the intersubgeneric hybridization between *G. max* Dwight and *G. tomentella* PI441001 generated by Dr. Ram Singh at the University of Illinois were grown in the field in 2008. The three original DAALs (BC_3F_3) were all selfed progeny from the same BC_3 plant (2n=41). Three rows were grown from seeds from different 2n=42 plants. These rows were similar to each other but morphologically distinct from Dwight. Since they came from the same 2n=41 plant, each plant should carry the same pair of *G. tomentella* chromosomes and these 2n=42 plants should breed true, because they carry two homologous chromosomes from *G. tomentella* and 20 pairs of soybean chromosomes.

Since 2008 single plants have been harvested each year from progenies of these original rows. A burst of phenotypic variation from single plant progenies from the DAAL was first seen in 2010. Some of the selfed progenies of the DAALs produce progenies that are very different from the DAAL progenitor plant and from the recurrent parent, Dwight. The range of morphological variation that was observed in the selfed progenies of 2n=42 plants was much greater than has been observed in other lines derived from the same BC_2 plant. All of the off-type plants were determined to have lost one or both of the extra pair of *G. tomentella* chromosomes (unpublished data from Dr. Ram Singh). This phenomenon has never been observed in soybean before. The objective of this presented research was to document the phenotypic and chromosomal variation among the lines derived from the DAALs (2n=42) and to compare the transcriptome profile between a DAAL (2n=42, LG13-7552) and a disomic progeny (2n=40, LG12-7063) via RNA-sequencing in order to help understand the genetics behind the variation.

Materials and Methods

Field Study

Plant material and experimental design

In 2012, approximately 1400 single plant progenies (F6 Dwight (4) x PI 441001) that trace back to the original DAALs were planted. Some rows were still segregating phenotypically but the uniform rows were bulk harvested. Lines that were morphologically different from a typical DAAL were selected to represent the diversity of the lines. Field trials were conducted in 2013 and 2014 at the University of Illinois Crop Sciences Research and Education Center in Urbana, IL. The experimental design for both years was a randomized complete block design with two replications in 2013 and four replications in 2014. However, some of the lines continued to show segregation, therefore, only 73 lines with complete data of six replications in both years were included in the results. The rows were 1.2m long with 36 viable seeds planted per row. The checks included two rows of DAALs (LG12-11684 and LG12-7663) and one row of Dwight in each replication.

Field notes and soybean sampling

Flower color, pubescence color, pod color, stem termination, plant height (cm), lodging score (from 1 to 10, with 1 being erect) and maturity date (days after May 31st) were recorded for each plot. All the rows were bulk harvested, and seed coat and hilum color were recorded. Protein and oil concentrations were determined using Perten Diode Array 7200 Near-infrared spectroscopy (NIR, Perten Instrument, Sweden).

Protein and oil concentrations were not obtained for entries without a yellow seed coat because of the pigment interference with NIR measurements, therefore, LG12-7512 were excluded from quantitative trait analysis. *G. tomentella* PI441001 was not included as a control in the field because it is a perennial and will not produce seeds during regular field growing season and its seed composition was not measured.

Statistical analysis

The statistical model used for the analysis of the quantitative traits was:

$$Y_{i(j)k} = \mu + a_i + b_{(i)j} + G_k + ag_{ik} + \varepsilon_{ijkl}$$

$Y_{i(j)k}$ is the observation of kth GENOTYPE in the jth BLOCK of the ith YEAR.

μ is the grand mean.

a_i is the random effect of ith YEAR.

$b_{(i)j}$ is the random effect of jth BLOCK in ith YEAR.

G_k is the fixed effect of kth GENOTYPE.

ag_{ik} is the interaction between random effect of ith YEAR and kth GENOTYPE.

ε_{ijkl} is the random error of the kth GENOTYPE in the jth BLOCK of the ith YEAR, and assumed to be normally and independently distributed $(0, \sigma_e^2)$.

All the statistical analyses for the field study were done using R (R Core Team, 2014). The mixed model above was built using package "lme4". Package "lsmeans" was used to calculate Least Square Means for each entry and package "predictmeans" was used to obtain Least Significant Difference for each quantitative traits (Appendix A).

Quantitative traits of check rows and disomic progeny rows (2n=40), excluding black seed coat line LG12-7512, were analyzed using ANOVA. LSMEANS were acquired for each quantitative trait for each entry. And least significant differences at $\alpha=0.05$ ($LSD_{0.05}$) were calculated for each trait for pair-wise comparison between entries.

Root Tip Chromosomal Count

The methodology for counting soybean mitotic chromosomes was as described by Chung and Singh (2008). Root tips from selected lines were collected from a sand bench in the greenhouse when the unifoliolate leaves were fully expanded. An 8-hydroxyquinoline solution (0.5g/L) was used to pretreat the root tip to cause chromosome contraction and improve the spreading of the chromosomes on the glass slides. After washing out the pretreatment solution, fixative containing 3 parts of ethanol and 1 part of glacial acetic acid was added to the root tips. After storing at room temperature for at least 24 hours, the fixative was removed through pipetting; root tips were washed with distilled H_2O, hydrolyzed in 1N HCl and then washed again with distilled H_2O. Treated root tips were stained with Feulgen stain to create purple colored root tips and then stored in cold water. Before counting, the purple colored root tips were further stained with Carbol Fuchsin on a clear glass slide then covered with a cover slide and squashed to spread the chromosomes for observation under a compound microscope.

RNA Sequencing

Sample preparation

LG13-7552 (2n=42) and LG12-7063 (2n=40) derived from the same DAAL were used for RNA sequencing (RNA-Seq). LG12-7063 was chosen because it is phenotypically very different from either of the parents or the progenitor DAAL. LG12-7063 has gray pubescence, white flowers, and a buff hilum. It is taller and lodged more than Dwight. Seeds from these two lines were grown in sand in the growth chamber at 24 C until the unifoliolate leaf was fully expanded. Plants were extracted from the sand and root tips excised, and then stored at 4 C until chromosome numbers were counted. Plants were transplanted into individual pots. When the second trifoliate was fully unfolded, all tissue above the first trifoliolate leaf of each plant was collected for RNA-Seq. Each line assayed for RNA-Seq was represented by three biological replicates. Tissue was immediately frozen in liquid nitrogen after harvesting and stored at -80 C until RNA isolation.

RNA isolation

RNA was isolated using TRIzol Reagent and Phase Lock Gel-Heavy (Zou et al., 2005) and further cleaned using DNase I (Ambion) and RNeasy (QIAGEN). Purified RNA samples were analyzed for quality using an Agilent®2100 Bioanalyzer™ (Agilent®, Santa Clara, CA, USA) and samples with RNA content of more than 1ug/sample were sent to the Roy J. Carver Biotechnology Center at University of Illinois for high-throughput RNA-Seq (Illumina Hi-Seq 2500).

Alignment

In the first step of alignment, USeq (http://useq.sourceforge.net/) pooled predicted gene coding regions of the *Glycine max* Wm82.a2.v1 (Schmutz et al, 2010) reference genome by collecting sequences assigned as transcripts (including all possible combinations of RNA splicing) together with genome regions predicted to not encode for any genes (masked genome) to create an index reference genome. This newly created reference index was used by the program Novoalign (Novocraft Technologies, Selangor, Malaysia) to align all the RNA-Seq reads. Novoalign is a highly accurate aligner for mapping sequences to a reference database (http://www.novocraft.com/products/novoalign/). Novoalign does not allow the user to define the number of mismatched alignments, but it does allow the user to define a threshold of alignment scores and the program reports the best alignment with the lowest score and any other alignment with similar scores (Yu et al, 2012). After mapping, Novoalign sorted mapped reads by the start location in the genome of each sequence read, and then merged all files into one BAM file (a binary format for storing sequence information) (Appendix B).

Differential gene expression

Using the BAM file generated by Novoalign, the program HTSeq (htseq-count) (Anders et al, 2014) determines how many aligned reads have overlapping exons for each gene according to the reference gene models. HTSeq is a Python script specifically coded for differential expression analysis as it only counts reads unambiguously aligned to a reference gene model and discards reads that overlapped with multiple genes or are aligned to more than one location. HTSeq yields a count table for each gene with a summary of discarded reads at the end (Appendix B).

The three biological replicates allows for testing of significance in differential expression between genes in LG13-7552 (2n=42) versus LG12-7063 (2n=40). The R/Bioconductor package DESeq was used to detect genes that have significantly different expression level using negative binomial distribution and a shrinkage estimator for the variance of distribution. DESeq addresses the over-dispersion problem of Poisson distributions and extends the single proportionality constant model of *edgeR* to provide a better fitted model and determines the mean and variance according to the linear regression of sample factors, expression strength and the smooth function raw variance. It estimates a dispersion value for each gene, and then fits a curve through the estimates. Based on the per-gene estimate and the fitted value, it assigns a dispersion value to each gene that result in a more balanced selection of differentially expressed genes (Figure 2). In the conservative approach of DESeq, if a per-gene estimate lies below the red line, it will be shifted towards the regression line. However, the above per-gene estimate will be kept as is (Anders and Huber, 2010) (Appendix C).

With the estimated dispersion per-gene, "2n=40" and "2n=42" were set as the conditions to conduct binomial test on expression level between LG12-7063 and LG13-7552. DESeq uses the Benjamini-Hochberg multiple comparison correction method (Anders and Huber, 2010). Differentially expressed genes (DEGs) were detected at an adjusted p-value and false discovery rate (FDR) of 0.05 (Figure 3). Principal component analysis (PCA) of the 6 plants using all sequence reads was done within DESeq (Appendix C).

Enriched Gene Ontology analysis

Differentially expressed genes acquired from DESeq were extracted for Gene Ontology (GO) singular enrichment analysis (SEA) using AgriGO (http://bioinfo.cau.edu.cn/agriGO/index.php), an ontology analyzer for the agriculture community that supports 45 species and 292 datatypes. SEA will sum GO terms into one set, and then performs Fisher's t test to statistically determine if a particular GO term occurs more frequently between samples as compared to the frequency of a reference set.

$$\mathrm{F} = \frac{\frac{\text{Number of enriched GO term in input set}}{\text{Number of total genes in input set}}}{\frac{\text{Number of enriched GO term in reference set}}{\text{Number of total genes in reference set}}}$$

For soybean, Williams 82 genome is the reference. Its GO information is utilized as the reference with which to compare other gene sets. In AgriGO, there are in total of 29,641 annotated genes in the soybean reference database. The GlymaID identifier of the DEGs was input for SEA to acquire GO information. Annotated genes were clustered into functional groups and they were compared to the reference. Graphic outputs of the analysis were automatically generated by AgriGO.

Trinity *de novo* assembly

The Trinity RNA-seq assembler (Grabherr et al, 2011) was used as the platform for *de novo* assembly because it allows for the study of the transcriptome in the absence of a reference genome. Although a soybean reference genome is available (Williams 82), the difference between the parent Dwight (as well as *G. tomentella*) and the reference is unknown, and the SNPs between them may result in mismatches in the alignment process. Also, Trinity handles the strand-specific Illumina paired-end libraries well (Appendix B).

Trinity consists of three processes (Grabherr et al, 2011). The first step is to break the long reads (computationally challenging) to smaller overlapping fragments called k-mers (with k-mer set to 25 bases) that are used by the program Inchworm to generate sequence contigs. The second program, Chrysalis, clusters the Inchworm contigs based on their relatedness and constructs a *de Bruijn* graph for each cluster, partitioning the RNA-Seq reads into cluster graphs that allow subsequent computation. In the third step, the program Butterfly processes each graph by tracing the reads through the graph to determine the connectivity to other contig graphs. The final output is a report of potential full-length transcripts for differently spliced isoforms that reflects the original cDNA molecules.

In Trinity, sequence reads of all 6 samples were pooled together for contig assembly. Thus, an integrated contig library would be available for subsequent comparisons. Minimum contig length was set at 100 bp.

Differential gene expression (Trinity)

RSEM, a program within Trinity, was used to estimate transcript abundance. RSEM executes the program Bowtie to directly align RNA-Seq reads to the contigs and then normalizes the expression level base on fragments per kilobase of transcript per million mapped reads (FPKM) (Li and Dewey, 2011). The counts for each gene in all 6 samples are merged into one table. Bowtie is a very stringent aligner, which allows only 3 mismatches and does not tolerate gaps. All the contigs generated by Trinity were exported into R/Bioconductor package DESeq to test significant difference in gene expression levels between 2n=40 and 2n=42 lines. The expression was analyzed on a synthetic gene level. DEGs were detected at an adjusted p-value and FDR of 0.05 (Figure 4) (Appendix C).

Contig annotation

Assembled contigs from Trinity were annotated using a series of sequential BLAST searches (Altschul et al, 1998). To separate the sequences from the 40 *G. max* chromosomes and the two *G. tomentella* chromosomes, all assembled contigs were first aligned by BLAST against the soybean primary transcript (*Glycine max* Wm82.a2.v1). The GlymaID associated with the best match, or top hit, for each contig was extracted and assigned to the contig as its gene identifier. Contigs that had no hits in the primary transcript were aligned by BLAST to the soybean genome to identify by location contigs that hit the soybean genome, but not in a region that had been described or predicted previously to be a coding section. The remaining contigs that did not align to any part of the soybean genome were considered foreign sequences to the soybean genome and were aligned by tblastx to the NCBI nt sequence database (http://www.ncbi.nlm.nih.gov) to obtain a possible annotation. The potential functional annotations of the assembled contigs that matched *G. tomentella* sequences in tblastx were acquired from NCBI nr database using blastx (E-value < 1e-06) (Appendix B).

In parallel to the work above, all contigs were blasted against the eudicots repetitive element database in Repbase (http://www.girinst.org) using RepeatMasker (http://www.repeatmasker.org) to detect transposable elements (TEs) and repetitive sequences. It reports the contigs that contain repetitive sequences and classification of the repetitive sequences. Contigs within the categories of "simple repeat" and "low complexity" are removed because their functions are not clear. The remaining contigs are considered to be possible transposable elements (Appendix B).

Gene information

Functional information for each gene is required to relate the differential gene expression with the observed phenotype. Soybean Gene Expression Database (SGED, http://sged.cropsci.illinois.edu/index.cgi) is a powerful tool for finding gene descriptions. It contains BLAST results for all the predicted gene models in soybean reference genome against the NCBI nr protein database. By uploading the GlymaIDs to SGED, an Excel table containing the gene information of DEGs was produced providing sequence matches with their alignment scores. The top hit of some of the genes may be ambiguous; therefore the top 10 hits for each GlymaID were requested from the database and the best hit with a descriptive annotation was assigned to each DEG.

T allele sequence information

To confirm if LG12-7063 carries the *t* allele, the Trinity assembled contig (c23278_g1) that maps to the *T* locus was aligned to the *T* locus coding gene Glyma06g21920 sequence from *G. max* Wm82.a2.v1 (Phytozome, http://www.phytozome.net/) and *G. max* sf3'h1 mRNA for flavonoid 3'-hydroxylase (GenBank: AB061212.1) sequence deposited in NCBI (http://www.ncbi.nlm.nih.gov) by Toda et al (2002). Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/) was used for the multiple DNA sequence alignment.

Results and Discussion

Associations with chromosome loss and soybean phenotypic variation

Chromosome counts were made on all lines selected for the replicated field test and for RNA-sequencing analysis, LG13-7552 (2n= 42) (Figure 5) and LG12-7063 (2n=40) (Figure 6). All lines in the replicate field test, except the 2n=42 checks, were confirmed to have 40 chromosomes (Table 2). And genomic *in situ* hybridization result has shown that the DAAL (LG13-7552) had 40 soybean chromosomes and 2 *G. tomentella* chromosomes, whereas, the disomic progeny (LG12-7063) had 40 soybean chromosomes (unpublished data from Dr. Gopal Battu at the University of Illinois).

All DAALs (2n=42) derived from the original three lines had distinctive, but nearly identical, morphology with stunted plant growth, short internodes, reduced pod set, wrinkled leaves with delayed leaf abscission, and green stem at maturity. The plants had tawny pubescence, yellow seed coats with a black hilum, brown pods, purple flowers and indeterminate stem termination (Table 2). These qualitative traits were consistent with Dwight, the soybean parent. The original three DAALs in our research originated from one common MAAL, so the extra chromosomes in each DAAL were homologous and it was assumed the DAALs would be genetically stable. However, this assumption proved to be wrong and it is still not known why the chromosome loss occurred in some progenies, and not in others.

The characteristics of the disomic progenies (2n=40) were highly variable, including various traits that were not present in either of the parents, Dwight and PI441001, or in the progenitor DAAL. Some lines resemble Dwight in certain traits, but some others look completely different from either of the parents and their progenitor DAAL. Unexpected qualitative variation that was confirmed in our replicated test included gray pubescence, black seed coat; buff, imperfect black, gray, brown, and yellow hilum color; white flower color; tan pod color; and determinate stem termination (Table 2).

The two DAALs included in the replicated test were very similar to each other but were much shorter than Dwight (58 vs 85 cm), much later in maturity (130 vs 114 days) and had different seed composition (416 g/kg vs 389 g/kg for protein and 181 g/kg vs 215 g/kg for oil) (the value of the DAALs is the mean of the two DAAL lines LG12-11684 and LG12-7663). The protein concentration among the 2n=40 progenies ranged from 371 g/kg to 486 g/kg on a dry weight basis (Table 2 and Figure 7). The low extreme was not significantly different than Dwight, but 19 lines had higher protein concentrations than Dwight, and three lines, LG12-7063, LG12-7086 and LG12-7072, had protein concentrations higher than the highest 2n=42 check (Table 2 and Figure 7). This was surprising since the protein concentrations of all perennial accessions that have been measured have all been significantly lower than cultivated soybean (Kollipara and Hymowitz, 1992). Oil concentration varied from 168 g/kg to 223 g/kg among the disomic lines (Table 2 and Figure 7). There was a significant difference within the population, but none of the lines were significantly higher or lower than Dwight or the 2n=42 checks. DAALs (58 cm) were shorter than most of the disomic progeny lines, but 2 lines were less than 72 cm and not statistically different from the DAALs. The tallest line was 120 cm, and 38 other lines were significantly taller than Dwight (Table 2 and Figure 7). The lodging scoring of this population ranged from 1.7 to 7.2. There were no lines that lodged significantly less than Dwight and the DAALs, but 25 lines lodged significantly more than Dwight (Table 2 and Figure 7). Some of the lines with extreme lodging also had long branches with tips touching the ground. These quantitative traits involve multiple genes controlling different biological processes; therefore, it is difficult to speculate the cause of the variation among the disomic population.

The DAALs matured 16 days later than Dwight (Table 2 and Figure 7). There was a 35 day difference in maturity among the disomic progeny lines. LG12-7326, a disomic progeny, was significantly earlier in maturity than Dwight; however, 22 lines matured significantly later than Dwight. The latest one, LG12-7063, matured the same as the mean of the DAALs. The DAALs were poorly podded, which was likely to contribute to the delay in maturity. Secondly, AALs, trisomics or polyploids grew slower than normal diploids, and this could be attributed to the observation that individual cells with more DNA content take a longer time for each cell division than its lower ploidy cytotype (Cavalier-Smith, 1978). Therefore, it is possible that DAALs would take a longer time to mature. It is also possible that specifically expressed genes were somehow involved in delaying the mature process of the DAALs as *G. tomentella* is a perennial species without a seasonal life cycle. To have progeny that were 16 days later (LG12-7063) in maturity than Dwight and the same as DAALs without the two extra chromosomes seems highly likely to involve gene introgression from *G. tomentella* and/or gene regulation.

There are at least three possible explanations for the observed morphological changes. One of the hypotheses is that the unique phenotype of the DAAL was due to the suppression induced by the two alien chromosomes on the gene expression in the 40 soybean chromosomes. If the two alien chromosomes were then lost during meiosis, the suppressed genes on the 40 soybean chromosomes could then be expressed. The variation observed would be from the effect of the *G. tomentella* genes already introgressed into the *G. max* chromosomes. Another hypothesis is that the extra pair of *G. tomentella* chromosomes was lost because of pairing with a homoeologous soybean chromosome. The *G. tomentella* genes were expressed differently once they are integrated into the soybean genome and this novel expression caused the increased variation. The third hypothesis is that transposable elements or epigenetic factors were activated in disomic progenies that lost the extra chromosomes and this apparently has random effect on gene expression that caused the increased phenotypic variation. These hypotheses are not mutually exclusive, yet, none of the hypotheses seem adequate to explain why disomic progenies derived from 2n=42 plants in this cross had a greater range of variation than any other 2n=40 plants from the same BC2 plant.

Alignment

A total of 392 million reads were generated from the high-throughput RNA sequencing. The reads were 100nt in length, thus, the sequence coverage of this RNA-Seq is at least 35X. Novoalign aligned 331,328,696 pairs of reads to the reference genome, which is 84% of the total reads. Among the paired reads, 72 million had a mapping score (MapQ) below 10, and they were considered too low in quality and were therefore discarded. Among the low MapQ reads, 94% had a score of 3, meaning that the read maps to multiple locations, presumably the result of highly repetitive nature of the soybean genome.

In total, 73,320 gene models are predicted in the soybean genome and excluding splicing isoforms of genes, 54,175 of them are considered as primary transcripts (Schmutz et al., 2010). In this study, the total number of captured expressed genes was 41,602 regardless of the line, which is 77% of the total predicted. The statistical analysis identified 2,499 of these transcripts as being differentially expressed genes (DEGs) between the DAAL (LG13-7552) and the disomic progeny line (LG12-7063). There were 1,192 DEGs in higher abundance in the 2n=40 progeny line and 1,179 DEGs in higher abundance in the DAAL line. Among the detected DEGs, 128 DEGs were uniquely transcribed in the disomic line, while 79 DEGs were unique in DAAL line. These uniquely expressed genes were excluded from the discussion of DEGs in alignment, because these DEGs resulted from mapping to the sequence of soybean reference genome Williams 82, which is different from the sequence of recurrent parent Dwight, and the zero reads were more likely due to insufficient sequencing depth and/or high mapping stringency. Among the 2,292 DEGs that were expressed in both lines, 641 DEGs had expression levels of at least 4 fold differences.

Principal component analysis of RNA-sequencing reads from aligning to the soybean reference genome

Principal component analysis provided general information on the between- and within-group variances of RNA-sequencing reads from the six plants. In Figure 8, three biological replicates of each genotype, green or blue dots, do not always group closely together, which suggests within-group variance. Principal component 2 (PC2) is the vector that represents the within group variance, which may consist of the effects of several influential genes or a group of differentially expressed genes affected by extrinsic microenvironment factors. This variance could be made up of various components, including variance from the RNA-Seq sampling, technical issues and biological differences in responses to microenvironment, and the variation in expressing the unknown genetic trigger leading to these unstable phenotypes. Illumina high-throughput sequencing yields millions of reads in every run, but only part of RNA sequences in the cells was sampled and presented in the sequencing data, so there is always the possibility of insufficient sequencing depth. Only the difference in the partial RNA that was sampled contributes to the sampling variance. During the procedures in sample isolation and library preparation, some variance will be introduced as the handling of each sample cannot be exactly the same. Although there is considerable variance within each chromosomal group, PC1 can clearly separate DAALs and disomic progenies on one dimension. Therefore, PC1 represents the differential expressions of genes that are responsible for the main differences between DAAL and 2n=40 progeny (Figure 8). These genes may be associated with the phenotypic variations that were documented previously.

Gene Ontology enrichment analysis of DEGs from aligning to soybean reference genome

Gene Ontology (GO) provides a controlled language to describe the attributes of genes and gene products associated with biological processes, cellular components and molecular functions. The principal rationale of enrichment analysis is that if a biological process is different between or among group of comparison, the co-functioning genes should have a high (enriched) possibility of being selected as a relevant GO group (Huang et al., 2008). Among the differentially expressed gene list from the alignment analysis, 1,529 genes have annotations. Their GO terms enrichments were compared with reference Williams 82 (Figure 9).

GO terms in the biological process category associated with apoptosis, oxidation reduction, and defense and immune response were significant (Figure 10), which suggests that one of the lines in this study were functioning under stress compare to the other. As previously stated, the growth of DAALs appears stunted and distorted in vegetative and reproductive development. Therefore, the enrichment of stress related response could be expected. When the plants are growing under stress, stress responsive genes are expressed. If a plant spends a lot of energy to offset the stress, less resource would be allocated for normal plant growth. The stress, as well as the energy and nutrient deficiency caused by stress response, could lead to less biomass accumulation and slower plant growth.

Stress can be defined as any deviation from optimal growth condition. The stress existed in the DAALs even at the very early growth stage, as the plant tissue used for this RNA-seq was from young seedlings including the meristem. All 6 plants were growing in the growth chamber under favorable conditions and were provided with sufficient light, water and nutrient and favorable temperature. Therefore, the stress is likely the effect from internal genetic rather than external environmental factors. The enrichment analysis in the molecular function category of significant receptor activity, nucleoside binding, transporter activity and catalytic activity indicates difficulties in normal cell activity, which also supports the stress theory that the extra *G. tomentella* chromosomes suppressed the normal transcription and function of the genes on 40 Dwight chromosomes.

The GO enrichment analysis provided a relevant gene group-oriented view instead of an individual gene-based view to help understand the biological themes behind the large gene list; however, the approach contains some important limitations. First, only approximately 55% predicted soybean genes in the annotation database of the analysis were annotated and it is possible that certain enriched GO terms were overestimated. Then, certain annotations may be imprecise or incorrect, since most databases were established by curators manually reviewing literature, which may limit the accuracy of the result. Third, some ambiguous genes are categorized in multiple biological processes but were not adjusted for weight in different processes by the context of the experiment and it may result in false positives in certain groups. The biggest limitation specific to this study is that this analysis may not help understand the biological function of the *G. tomentella* chromosomes because the analysis was performed on known genes and biological processes only and the *G. tomentella* is not well represented in these categories. A more precise analysis of these results will depend on a better understanding of the *G. tomentella* genome and its gene functions.

Trinity *de novo* assembly

To analyze the RNA-Seq data in a manner that would allow for the study of all the sequences independent from a reference genome, we used the software Trinity, which generates contigs from the RNA-Seq data and uses the assembled *de novo* contigs as the reference genome. Trinity identified 133,872 contigs and 4,206 contigs (considered as genes) were differentially expressed between LG13-7552 (2n=42) and LG12-7063 (2n=40). Among the total DEGs, 2,894 had expression in both lines, while the expression level of 1,457 contigs was higher in LG12-7063 and 1,437 contigs higher in LG13-7552. There were 989 contigs expressed only in LG13-7552 which is much higher than the number of gene expressed only in LG12-7063 (323).

In the BLAST results (Table 3), 61,569 contigs had hits against the primary transcript of the soybean reference genome, which made up 46% of the total contigs. Of these, 1,563 contigs with annotations from GlymaID were differentially expressed between LG13-7552 (2n=42) and LG12-7063 (2n=40) and 397 of them had expression levels differences of at least 4 fold between the two lines. Among those without a hit against a gene model, 69,525 (52%) contigs found matches in the soybean genome. Therefore, 98% of the assembled contigs matched sequences from the 40 soybean chromosomes of the reference Williams 82 genome.

Comparison of DEGs from direct alignment to Williams 82 and DEGs from *de novo* assembly

DEGs identified from aligning to the reference genome and from *de novo* assembly located to every chromosome (Figure 11). The majority of the two DEG sets are similarly distributed across all chromosomes but the most are on chromosome 20, which suggests that differential expression is a global occurrence in the genome rather than chromosome/region specific. The Venn diagram (Figure 12) demonstrates that among the 2,292 DEGs from alignment and 1,563 DEGs from *de novo* assembly, only 752 DEGs matched the exact same GlymaID and the distributions of the common DEGs also resembles distributions of the previous two DEG sets (Figure 11).

Although the statistical analysis of differential expression in both methods (Trinity and direct alignment to a reference genome) was conducted in DESeq, abundance estimation tools used to determine an actual transcript differed. In the method without *de novo* assembly, transcripts were identified by direct alignment of RNA-Seq reads to the reference genome using Novoalign. Positive matches would correspond to a GlymaID, and Novoalign would calculate the frequency of each GlymaID that was represented by an RNA-Seq read. Novoalign uses a high sensitivity that encourages the mapping of more reads than other aligners. It also allows gaps in aligning, but lowers the MapQ of the sequence. The threshold of alignment scores was set to the default, which normally corresponds to alignments with 85% identity or better. But in *de novo* assembly, the build-in abundance estimator RSEM was used, which incorporates Bowtie to align sequences to the contigs that were derived from the RNA-Seq reads, and therefore this alignment could be much more stringent, and the mismatch parameter was set to 3 and gaps to zero. The use of different alignment methods is one of the main reason why the number of DEGs from the Trinity *de novo* assembly was much less than that from direct alignment of RNA-Seq reads to the reference genome using Novoalign, where a less stringent alignment was used to ensure alignments between Dwight or *G. tomentella* genomes and Williams 82 reference genome. Only about half of the DEGs from the Trinity analysis matched a GlymaID from the direct alignment analysis, which is also likely the result of mismatching nearly identical paralogs, or due to *de novo* analysis identification of sequences not present in the Williams 82 reference.

To narrow down the analysis, the 752 DEGs that were common to the two analysis methods were further processed. The top 20 most abundant transcripts and their descriptions are summarized in Table 4 and Table 5 for each line. Most genes have definitive functional descriptions, but some, such as Glyma02g03280, Glyma11g27920, Glyma17g06540, Glyma07g07145, Glyma10g26990, Glyma10g39800 and Glyma05g36750 have no known functional annotation.

Several stress and pathogen related plant defense genes were among the most differentially expressed genes between the DAAL (LG13-7552) and the disomic line (LG12-7063) (Table 4 and Table 5). Genes Glyma16g31341, Glyma17g31730, Glyma08g06730, Glyma03g16620, and

Glyma18g29400 had higher expression in LG13-7552, whereas, genes Glyma01g05710, Glyma02g42315, Glyma02g10320, Glyma03g22060, Glyma03g03170, and Glyma11g00510 were higher in LG12-7063. Of these top defense DEGs, three are predicted to be leucine-rich repeat (LRR) receptor-like serine/threonine-protein kinase, which is reported to play a central role in signaling during pathogen recognition, the subsequent activation of plant defense mechanisms, and developmental control (Afzal et al., 2008). This result is consistent with the previous GO analysis that shows large portion of defense genes among the most differentially expressed genes; however, whether these stress related genes are causing the phenotypic differences between LG13-7552 and LG12-7063 and the specific trigger to the activation of stress/pathogenic responsive genes and the defense mechanism is to be determined.

Of the top 20 most abundant genes in the DAAL (Table 4), six are predicted to be genes involved in controlling a number of fundamental aspects of plant growth and development, such as meristem differentiation, shoot structure, flowering time, branching, and leaf initiation rate. These genes include Glyma06g36140, Glyma02g13401, Glyma17g08890, Glyma20g23220, Glyma02g13420, and Glyma20g30740 and two of these genes are predicted to be a MADS-box protein, which can be critical to gametophyte development, embryo and seed development, and root, flower and fruit development as well (Gramzow and Theissen, 2010). The high ratio of differentially expressed critical genes that contribute to the fundamental plant growth captured in young DAAL seedlings suggests abnormal gene regulation of early plant development, which could explain curled and wrinkled trifoliolates of the DAALs and occasionally in defect shape or opposite growth position observed when harvesting tissue for sequencing. All six developmental genes identified are at least 16 times higher in expression in LG13-7552 than in LG12-7063, therefore, apart from the previous speculation of the DAAL in stress, overexpression of the critical genes involved in plant development could possibly be involved in the abnormal and stunted development of the DAAL as well. However, among the top 20 most abandance genes in the disomic progeny (Table 5), except for the genes with ambiguous functional annotations and the stress related genes, the rest are involved in multiple pathways, such as ABC transporter, N-hydroxycinnamoyl/benzoyltransferase and purple acid phosphatase, which are difficult to associate with any of the phenotypic variation or to interpret at this point of time.

Introgression of *G. tomentella* sequences

The functional annotations available from Phytozome (www.phytozome.net) for the Williams 82 reference are limited, and therefore 2,752 (2% of total assembled) of the Trinity assembled contigs had no annotation found in the soybean reference genome. When these unmatching contigs were tblastx to NCBI database, 1,454 of them had high similarity to known sequences, and 213 of them matched *G. tomentella* sequences with percent identities ranging from 37% to 100% (Table 3). Of the 213 contigs, 20 had homologous protein matches (E-value < 1e-06) from NCBI database (blastx) and are listed in Appendix D.

These 213 assembled contigs that matched *G. tomentella* sequences were grouped into 175 genes that matched 22 publicly deposited *G. tomentella* sequences, which were made up of 10 BAC-clones and 12 retrotransposon sequences. Of the matched 175 genes, 174 were expressed only in the DAAL whereas the other one had no transcript, and 51 of the genes had statistically significant differential expression between the two lines ($p_{adj} <0.05$). This shows that part of the *G. tomentella* genome was present in the DAAL. These *G. tomentella* sequences may be in the extra pair of *G. tomentella* chromosome or were introgressed into the 40 soybean chromosomes. Unfortunately, all 22 matched *G. tomentella* sequences were annotated as either BAC clones or transposons and no additional details regarding their functions is available.

One of the 213 contigs that matched *G. tomentella* sequences had reads mapped to it from both lines and two were expressed only in the disomic progeny (2n=40); however, all of these three have very low read counts (<10) and none of them had significantly different expression levels, which means they are likely to be false positive read counts. Therefore we found no evidence of expressed, introgressed *G. tomentella* sequences in the disomic progeny (2n=40).

There are not many *G. tomentella* sequences available in the database, so those that matched *G. tomentella* sequences are not likely to be all of the introgressed *G. tomentella* genes. Many of the contigs that did not match the soybean reference matched sequences from other legume species, such as *Medicago truncluata*. It is highly probable that these sequences could be genes from *G. tomentella* that are highly conserved across species in *Facaceae*.

Expression of transposable elements

RepeatMasker (RM) found that 4% of the contigs were retroelements (Appendix E). In the 26,126 contigs containing repetitive sequences, 1,077 were significantly different in expression level between LG13-7552 (2n=42) and LG12-7063 (2n=40) (Table 3). Of these, 399 contigs were expressed in both lines and up-regulated in the DAAL, compared to 242 up-regulated in the disomic progeny. Gene expressions of 342 contigs were unique to the DAAL and 94 were only expressed in the disomic progeny (2n=40). Nearly four times the number of transposable elements (TEs) was uniquely expressed in DAAL as in disomic progeny line, indicating that transposable elements were more highly activated in the DAAL line.

Previously we proposed a hypothesis that the phenotypic variation in the disomic progenies may be due to the activation of transposable elements; the fact that more retrotransposons are expressed in the DAAL line seems to contradict this hypothesis. However, there are other possibilities. The SoyTEdb shows that 42% of the soybean genome consists of class I TEs (retrotransposon), whereas class II TEs (DNA transposons) account for 16% of the soybean genome (Du et al., 2010). Class I TEs involve an RNA intermediate when inserted at a new position, but class II does not. Since only RNA was sequenced in this study, the effect of class II TEs on the phenotype cannot be determined. Although DNA transposons constitute only a small portion of the whole genome, it is still possible that they are responsible for at least some of the genetic variation. Secondly, a decrease in TE expression can lead to variation as well, as long as the expression of TEs is influencing phenotype. Therefore, removal of those regulatory TEs could reveal the new variation, and this scenario could fit the hypothesis of suppression of the phenotypic variation by the *G. tomentella* genome. Thirdly, if the phenotypic variation was caused by epigenetic effects, the expression of many genes could be repressed, including the reduction in the number of activated retrotransposon in the disomic progeny lines.

The TEs detected in this study likely originated from the *G. tomentella* genome. In the first use of RM, total assembled contigs were blasted against the *Glycine* repetitive element database, but few TEs were found. Since the 4% retro-elements were found when blasting against eudicotyledons database instead of the *Glycine* database, they are unlikely to be soybean sequences. The 12 matched *G. tomentella* retrotransposon sequences from NCBI database support this idea, as well as RM confirming that 56 of the 175 genes that matched *G. tomentella* sequences deposited in NCBI were retrotransposons. These uniquely expressed TEs in DAAL could locate on the two *G. tomentella* chromosomes, but could also be introgressed into the soybean genome.

Regulation of pigmentations

The gene that had largest expression difference was Glyma20g33810 (Table 4), which was predicted to be an anthocyanidin 3-O-glucosyltransferase with 100% identity. Its expression was more than 512 times higher in DAAL lines than disomic progeny line. Anthocyanidin 3-O-glucosyltransferase catalyzes the reversible reaction of converting uridine diphosphate (UDP)-D-glucose and an anthocyanidin to UDP and an anthosyanidin-3-O-beta-glucoside (Kamsteeg et al, 1978). It was postulated to be one of the two key enzymes in the conversion from colorless leucoanthocyanidin to purple anthocyanidin 3-glucoside (Nakajima et al, 2001). This result is consistent with the phenotypic observation of DAALs possess purple hypocotyl and darker leaf tissue, whereas, LG12-7063 has green hypocotyl and lighter leaf.

Glyma06g21920 (*Glycine max* v1.1: 6.902 kbp from Gm06:18,534,606 to 18,541,507) is the coding gene located at the *T* locus, which in classical genetics is the dominant *T* allele producing tawny pubescence, whereas the recessive *t t* genotype is associated with grey pubescence. Buzzell et al. (1987) reported that the soybean *T* allele is involved in the formation of cyanidin-3-glucoside by dihydroxylate of the flavonol B-ring, which is an overlapping pathway with anthocyanidin 3-O-glucosyltransferase to produce the bronze pigmentation in the trichomes. Therefore, because the DAALs had tawny pubescence, it can be assumed that the *T* allele was present and expressed, as recurrent parent Dwight carries the same. Toda et al (2002) concluded that the *t* allele is only a single base deletion difference from the *T* allele, which results in the early termination of gene translation and leads to gray pubescence. However, there was only one isoform assembled (c23278_g1_i1) under this gene cluster. The sequence of this isoform matched 100% with the Glyma06g21920 sequence of Williams 82 (tawny) reference genome, and matched 100% with the protein sequence of flavonoid 3'-hydroxylase (*Glycine max* sf3'h1 mRNA for flavonoid 3'-hydroxylase, complete cds, GenBank: AB061212.1) deposited by Toda et al. (2002) and only one base (C/A) different from its nucleotide sequence which is a synonymous mutation. No trace of single base deletion at the reported position was detected in the assembled contigs (Appendix F). Although the dominant allele seems to be present at the *T* locus in both lines, it was found that this gene, Glyma06g21920, was transcribed four times higher in the DAALs than in LG12-7063. Therefore, the conclusion is that the gray pubescence of LG12-7063 was not a result of genotypic sequence change from *T* – to *t t*, but instead was most likely due to regulatory factors leading to reduced expression of Glyma06g21920.

If the down-regulation of Glyma06g21920 in LG12-7063 was responsible for the change in phenotype, it is possible that down-regulation of genes might be the reason for other phenotypic changes observed that are normally considered to be conditioned by recessive alleles. For example, this could be the cause of the changes from black hilum in LG13-7552 to buff hilum in LG12-7063 and from purple flower in LG13-7552 to white flower in LG12-7063, which is controlled by the W1 locus (Woodworth, 1923); however, the expression of the locus is tissue specific (hypocotyl, flower, and pod), the speculation can not be confirmed in this RNA-Seq analysis. This could also be the case in phenotypic changes in other disomic progeny lines from yellow to black seed coat (Reese and Boerma, 1989); brown to tan pod color (Bernard, 1967); indeterminate to determinate stem termination (Bernard, 1972). Conventionally going from black to gray hilum would be changing the *I* locus from i^i i^i to *I I*, where *I* is dominant to i^i; however, we know that the dominate allele *I* functions by gene silencing (Tuteja et al., 2004). Epigenetic methylation and transposable elements (TEs) activations are two other mechanisms, because they are both capable of inducing changes throughout the genome. For example, DNA methylation can happen to any gene in the genome and prohibit or decrease their expression, while duplicates of TEs can randomly bind to any site of the genome and can perform different functions based on the family of the TEs. Without additional experimental evidence, the roles of TEs and methylation are strictly speculation at this point.

Conclusions

This research documented that derived disomic progenies (2n=40) from the DAALs (2n=42) exhibit a wide phenotypic variation in numerous traits after the extra pair of *G. tomentella* chromosomes in the DAALs (2n=42) was lost. Some phenotypes of the derived disomic progenies do not exist in either of the parents or in the progenitor DAALs. Variation was observed for quantitative traits such seed composition, plant height, lodging and time of maturity as well as qualitative traits such as flower, seed coat, hilum, pod, and pubescence color. Of potential practical interest are three high protein lines (> 450 g/kg). Whether these disomic progeny lines would be valuable in any future breeding projects requires further evaluations and comparisons with existing high protein soybean lines. Why the extra pair of *G. tomentella* chromosomes is occasionally eliminated from the DAALs is still not known.

Aligning RNA-sequencing reads to the soybean reference genome shows 2,292 differentially expressed genes between the DAAL (LG13-7552) and one of the disomic progeny (LG12-7063) randomly spread across the genome. Genes critical to fundamental growth are among the most differentially expressed and high number of DEGs related to stress and pathogen response explained the abnormal and stunted development the DAAL and partially its differences from the disomic progeny.

Failure to detect the single base pair deletion at the *T* allele that has been shown to produce gray pubescence and the down regulation of Glyma06g21920, the gene at the *T* locus, indicates that the gray pubescence of LG12-7063 was not due to *t t* genotype, but the result of gene regulation. It is possible that such regulation may be responsible for changes in other qualitative traits where the observed phenotype is generally conditioned by a recessive allele or a loss of function of a gene.

RNA-sequencing data also strongly support the expression of *G. tomentella* sequences and higher expression levels of transposable elements (TEs). However, *G. tomentella* sequence expression was only observed in the DAAL (LG13-7552) and the number of TEs sequences was lower in the disomic progeny than in DAAL. In the future, to further clarify the mechanisms that caused the wide variation, more studies focusing on transposable elements and epigenetic effects are needed, such as confirming TE duplication on the 40 soybean chromosomes using Southern plots and examining epigenetic effects including DNA methylation and siRNA.

Figures and Tables

Table 1 Taxonomy of *Glycine* Species

	Species	Isozyme Group	2n	Nuclear Genome Symbol	Chloroplast Genome Symbol	Geographical Distribution[1]
	Subgenus *Soja* (Moench) F. J. Hernann					
1	*G. soja* Sieb. & Zucc.		40	G	G	China, Japan, Russia, Korea, Taiwan
2	*G. max* (L.) Merr.		40	G1	G	World wide
	Subgenus *Glycine*					
1	*G. albicans* Tindale and Craven		40	I	A	WA
2	*G. aphyonota* B. Pfeil		40	I3	A	WA
3	*G. arenaria* Tindale		40	H	A	WA
4	*G. argyrea* Tindale		40	A2	A	Q
5	*G. canescens* F. J. Hermann		40	A2	A	Q, NSW, V, SA, NT, WA
6	*G. clandestina* Wendl.		40	A1	A	Q, NSW, V, SA, T
7	*G. curvata* Tindale		40	C1	C	Q
8	*G. cyrotoloba* Tindale		40	C1	C	Q, NSW
9	*G. falcata* Benth.		40	F	A	Q, NT, WA
10	*G. gracei* B. E. Pfeil and Craven		40			NT
11	*G. hirticaulis* Tindale and Craven		40	H1	A	NT
			80			NT
12	*G. lactovirens* Tindale and Craven		40	I1	A	WA
13	*G. latifolia* (Benth.) Newell and Hymowitz		40	B1	B	Q, NSW
14	*G. latrobeana* (Meissn.) Benth.		40	A3	A	V, SA, T
15	*G. microphylla* (Benth.) Tindale		40	B1	B	Q, NSW, V, SA, T
16	*G. montis-douglas* B. E. Pfeil and Craven		40			NT
17	*G. peratoda* B. E. Pfeil and Tindale		40	A5	A	WA
18	*G. pescadrensis* Hayata		40	AB1	A	Q, NSW; Taiwan, Japan
19	*G. pindanica* Tindale and Craven		40	H2	A	WA
20	*G. pullenii* B. Pfeil, Tindale and Craven		40	H3, A4	A	WA
21	*G. rubiginosa* Tindale and B.E. Pfeil		40	B3	A	NSW, SA, WA
22	*G. stenophita* B. Pfeil and Tindale		40		B	Q, NSW
23	*G. syndetika* B. E. Pfeil and Craven	D4	40	A6		Q
24	*G. dolichocarpa* Tateishi and Ohashi		80	D1A		Taiwan
25	*G. tabacina* (Labill.) Benth.		40	B2	B	Q, NSW, V, SA; West Central & South Pasific Islands
			80	BB1, BB2, B1B2	B	
26	*G. tomentella* Hayata	D1, D2	38	E	A	Q
		D3	40	D1A	A	Q, WA; PNG
		D5B	40	H2	A	WA
		D5A	40	D2	A	WA, NT
		T1	78	D3E	A	Q, NSW; PNG

Table 1 (cont.)

	T5	78	AE	A	NSW
	T6	78	EH2	A	WA
	T2	80	DA6	A	Q; Taiwan
	T3	80	DD2	A	Q, NT, WA; Philippines, Taiwan
	T4	80	H2	A	

[1] WA: Western Australia; Q: Queensland; NT: Northern Territory; SA: South Australia; T: Tasmania; V: Victoria; NSW: New South Wales; PNG: Papua New Guinea.

This table is developed from Chung and Singh, 2008.

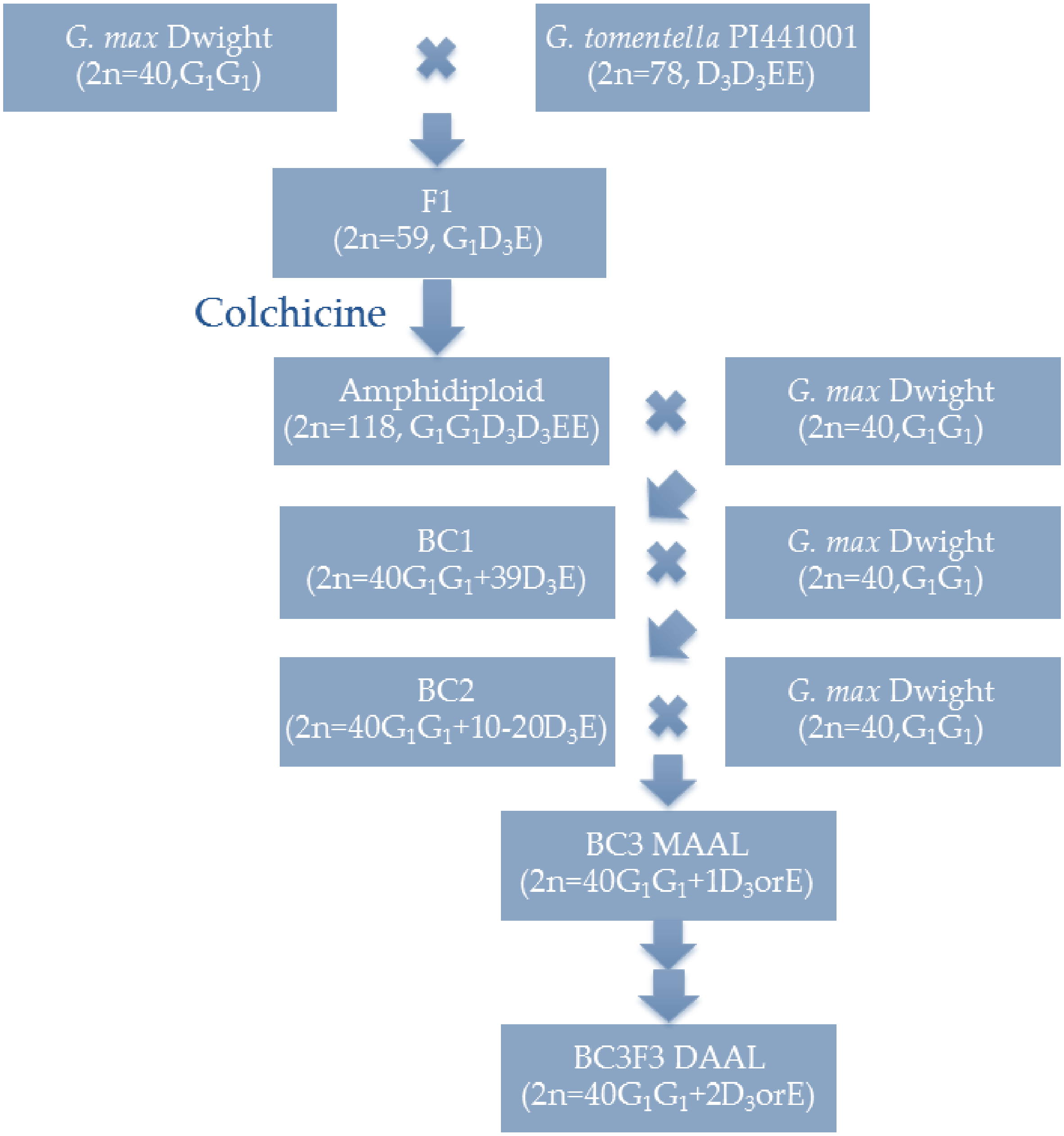

Figure 1 A schematic diagram of the production of fertile intersubgeneric progenies between *G. max* and *G. tomentella*

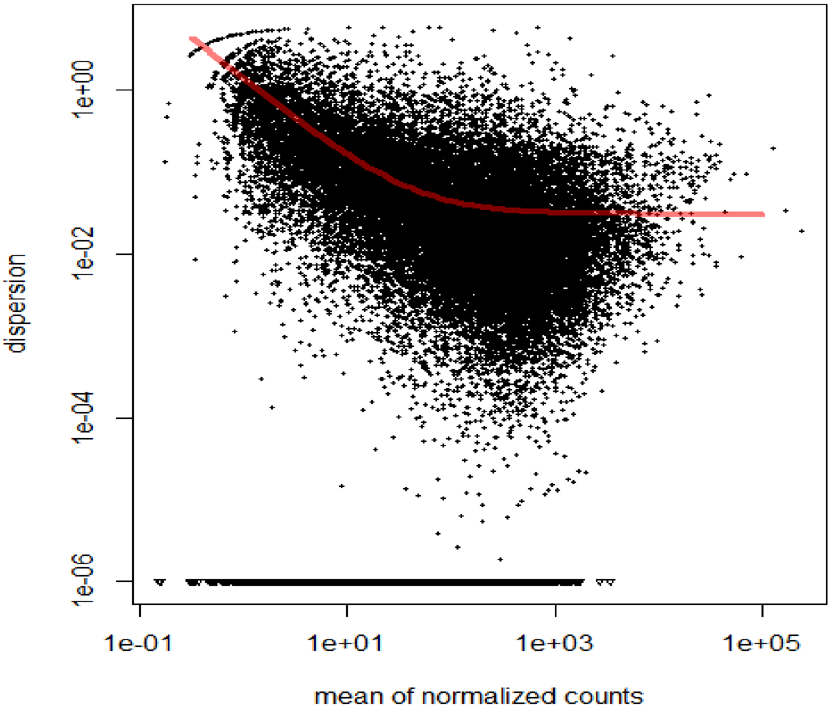

Figure 2 Empirical and fitted dispersion values plotted against the mean of the normalized RNA sequencing data counts of each gene of the DAAL (LG13-7552) and the disomic progeny (LG12-7063) by HTSeq. The red line is the fitted curve of dispersion values estimated for each gene (per-gene estimate, black dots). HTSeq chose a dispersion value for each gene between the per-gene estimation and the fitted value for subsequent inference (Anders and Huber, 2013).

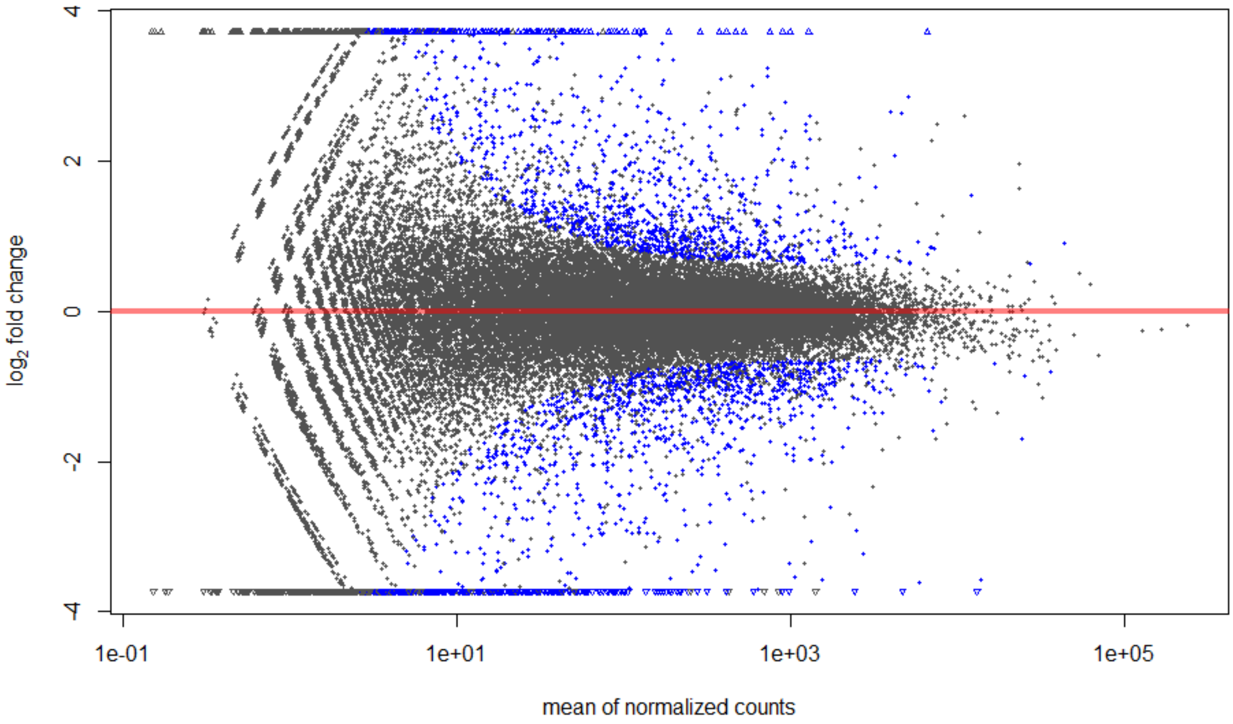

Figure 3 Plot of normalized means of aligning RNA sequencing reads against the soybean reference genome Williams 82 versus $\log_2$ fold change for the contrast of disomic progeny (LG13-7552) versus DAAL (LG12-7063). Blue dots indicate all the differentially expressed genes (DEGs, $p_{adj} < 0.05$).

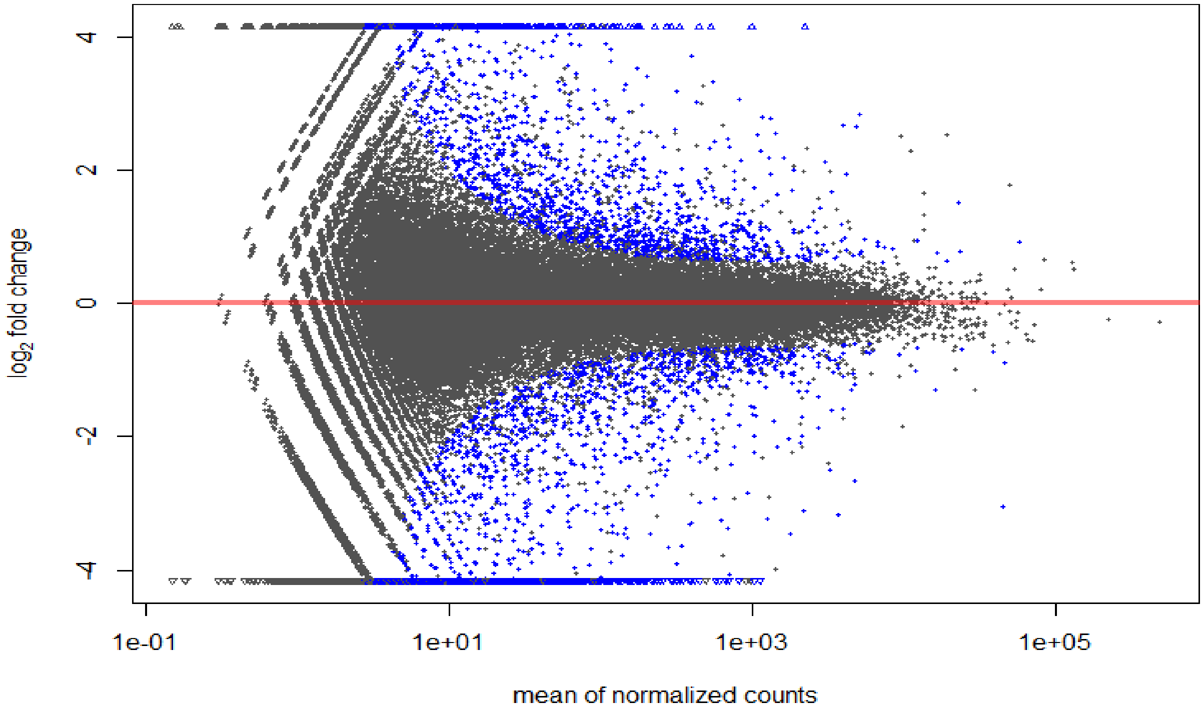

Figure 4 Plot of normalized means of mapping RNA sequencing reads to assembled contigs from Trinity *de novo* assembly versus $\log_2$ fold change for the contrast of disomic progeny (LG13-7552) versus DAAL (LG12-7063). Blue dots indicate all the DEGs ($p_{adj} < 0.05$).

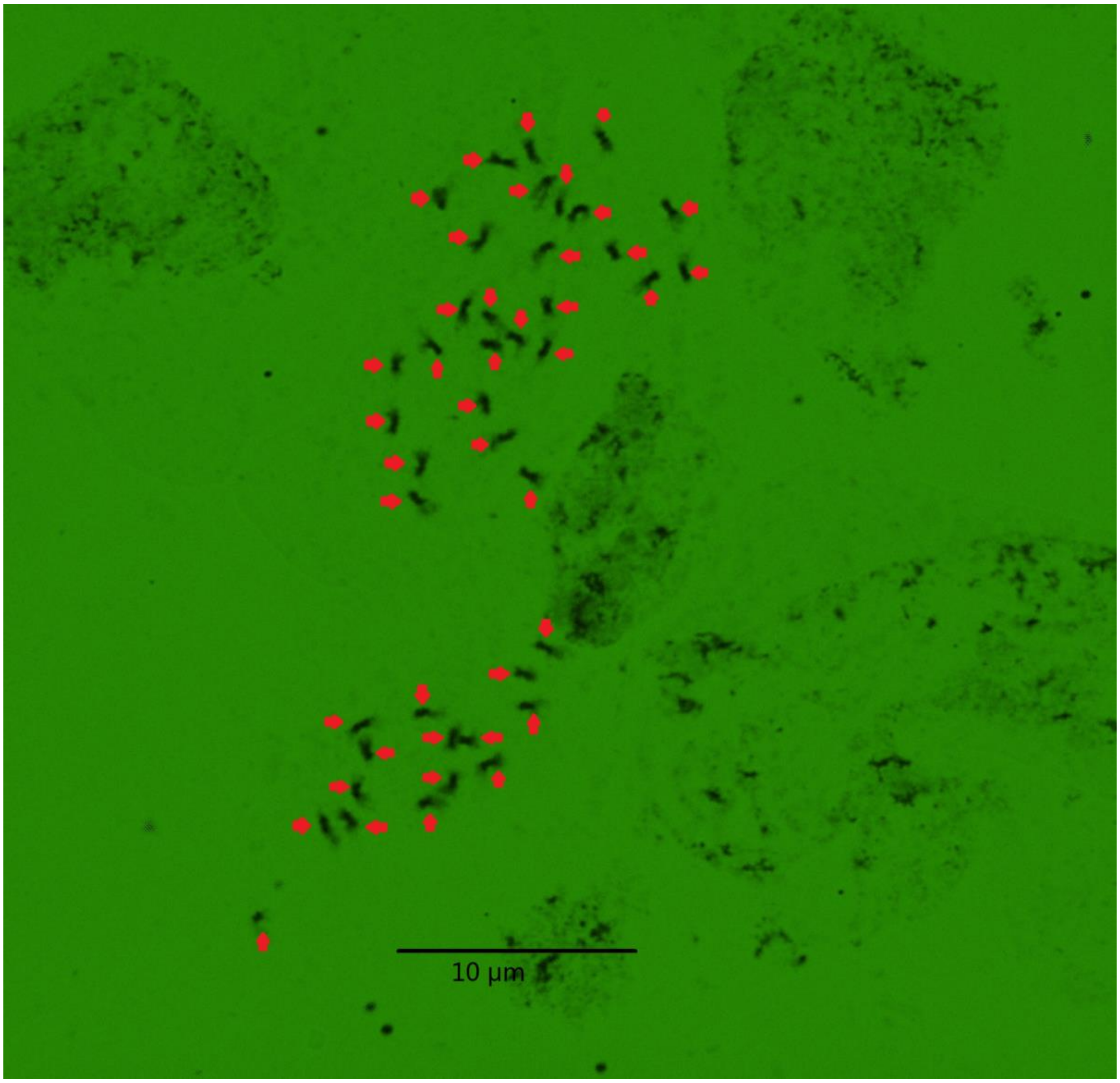

Figure 5 Mitotic chromosomes in the root tip cell of the disomic alien addition line (LG13-7552, 2n=42).

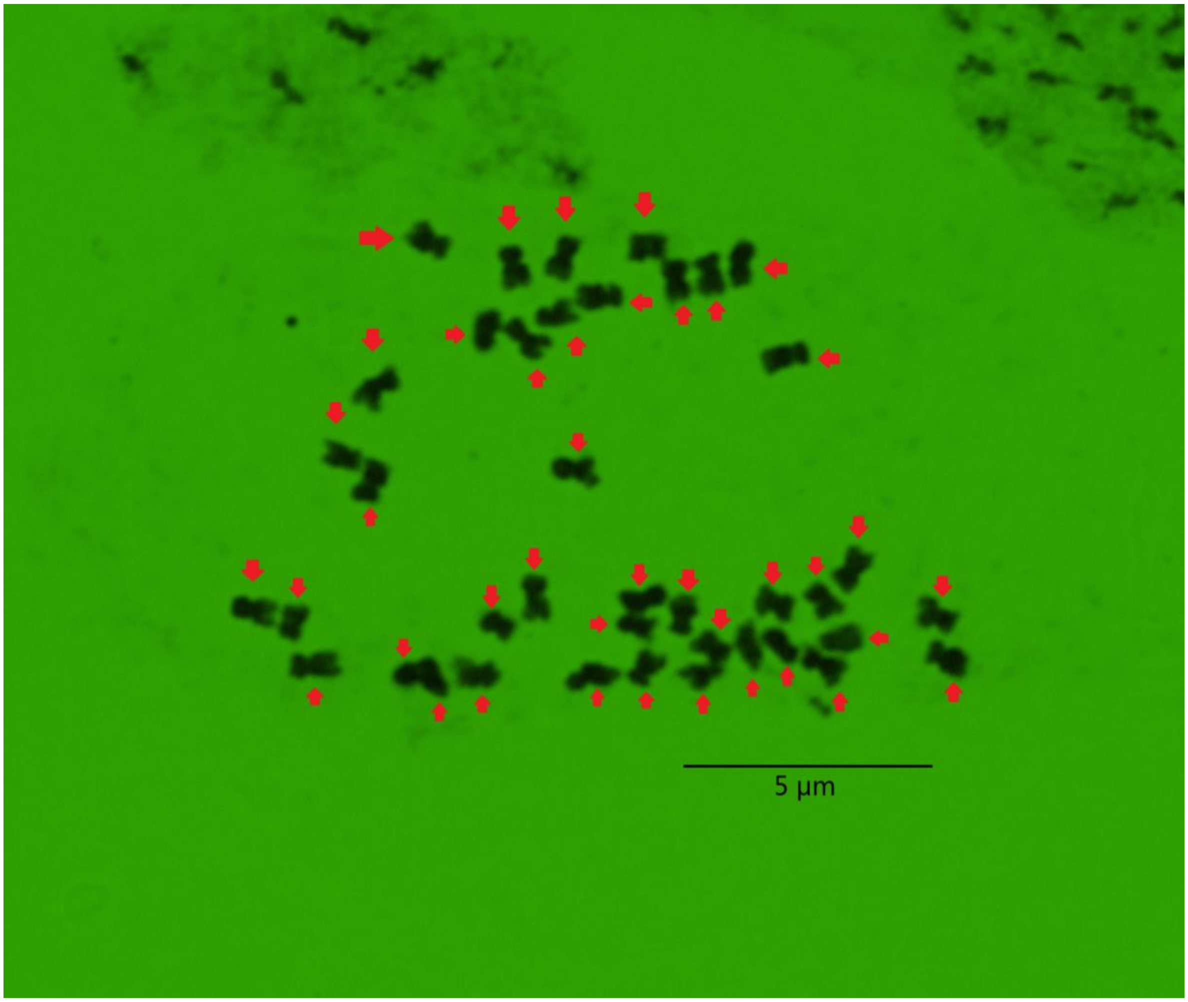

Figure 6 Mitotic chromosomes in the root tip cells of the disomic progeny (LG12-7063, 2n=40).

Table 2 Chromosome number, qualitative descriptors, and trait means for entries grown in 2013 (2 replications) and 2014 (4 replications) at Urbana, Illinois. Data from PI 441001 were not collected in this research but are provided for comparison purposes.

Entry	Chr No[1]	Pub[2]	SC[3]	Hlm[4]	FC[5]	PC[6]	Stem Term[7]	Pro[8] (g/kg)	Oil[8] (g/kg)	Hgt[9] (cm)	Ldg[10]	Mat[11] (day)
PI441001	78	T	Bl	Bl	P	Br	-[12]	-	-	-	-	-
LG12-11684	42	T	Y	Bl	P	Br	Ind	409	201	60	1.9	131
LG12-7663	42	T	Y	Bl	P	Br	Ind	423	193	56	1.9	129
Dwight	40	T	Y	Bl	P	Br	Ind	389	216	85	2.8	114
LG12-11612	40	T	Y	Bl	P	Tn	Ind	391	214	80	2.9	110
LG12-11672	40	G	Y	IB	P	Br	Ind	373	215	76	2.5	113
LG12-11711	40	T	Y	Bl	P	Tn	Ind	377	215	71	3.0	111
LG12-11829	40	T	Y	Bl	W	Tn	Ind	397	213	84	2.6	118
LG12-11832	40	T	Y	Br	W	Tn	Ind	402	211	87	2.4	121
LG12-11954	40	T	Y	Br	P	Tn	Det	422	185	86	3.6	125
LG12-12024	40	T	Y	Br	P	Tn	Ind	394	213	83	2.9	110
LG12-12110	40	G	Y	IB	P	Br/ Tn	Ind	409	203	87	2.7	114
LG12-12125	40	G	Y	IB	P	Tn	Det	408	199	65	1.7	116
LG12-7063	40	G	Y	Bf	W	Br	Ind	486	168	99	6.3	130
LG12-7072	40	G	Y	IB	P	Tn	Ind	445	184	93	5.1	126
LG12-7074	40	T	Y	Bl	P	Tn	Ind	407	192	86	5.6	122
LG12-7076	40	T	Y	Bl	P	Br	Ind	420	195	89	4.8	117
LG12-7080	40	T	Y	Bl	P	Tn	Ind	418	192	105	4.5	125
LG12-7081	40	T	Y	Bl	P	Tn	Ind	409	195	102	4.5	115
LG12-7086	40	G	Y	Bl	P	Br	Ind	469	177	103	6.4	129
LG12-7090	40	G	Y	Bl	P	Tn	Ind	382	217	89	4.3	113
LG12-7103	40	G	Y	Bl	P	Tn	Ind	421	189	99	5.5	123
LG12-7105	40	G	Y	IB	P	Br	Ind	428	188	94	5.7	119
LG12-7106	40	T	Y	Bl/ Br/ G	P	Tn	Ind	421	196	97	6.2	110
LG12-7108	40	G	Y	Br	P	Tn	Ind	417	198	90	5.5	114
LG12-7112	40	G	Y	Br	P	Tn	Ind	415	191	81	4.0	115
LG12-7113	40	G	Y	Bl	P	Br	Ind	404	194	90	6.4	116
LG12-7118	40	G	Y	IB	P	Tn	Ind	421	185	90	6.2	124
LG12-7120	40	T	Y	Bl	P	Tn	Ind	409	189	88	6.7	124
LG12-7127	40	T	Y	Bl	P	Br	Ind	412	191	103	6.6	124
LG12-7133	40	G	Y	IB	P	Tn	Ind	409	191	89	4.9	123
LG12-7135	40	T	Y	Bl	P	Tn	Ind	401	205	100	5.9	118
LG12-7137	40	T	Y	Bl	P	Br	Ind	387	213	90	2.6	124
LG12-7140	40	G	Y	IB	P	Br	Ind	431	186	83	5.5	116

Table 2 (cont.)

LG12-7148	40	T	Y	Bl	P	Tn	Ind	407	201	94	5.7	121
LG12-7151	40	T	Y	Bl	P	Tn	Ind	407	192	95	5.6	122
LG12-7155	40	T	Y	Bl	P	Tn	Ind	431	183	93	7.2	118
LG12-7157	40	G	Y	IB	P	Tn	Ind	426	184	100	6.7	125
LG12-7159	40	T	Y	Bl	P	Tn	Ind	420	190	99	6.9	118
LG12-7171	40	G	Y	Bl	P	Tn	Ind	401	201	100	5.5	124
LG12-7174	40	G	Y	IB	P	Br	Ind	415	188	99	5.6	127
LG12-7177	40	G	Y	Bf	P	Br	Ind	402	215	110	3.1	126
LG12-7180	40	G	Y	Bf	P	Br	Ind	397	215	108	3.5	123
LG12-7181	40	G	Y	Bf	P	Br	Ind	396	218	107	3.2	125
LG12-7182	40	G	Y	Bf	P	Br	Ind	389	218	100	2.9	119
LG12-7187	40	G	Y	G/IB	P	Br	Ind	398	215	105	3.2	112
LG12-7191	40	G	Y	G	P	Tn	Ind	393	212	118	3.2	116
LG12-7196	40	G	Y	Bf	P	Br	Ind	404	208	120	3.2	125
LG12-7201	40	G	Y	IB	P	Br	Ind	390	219	85	3.2	116
LG12-7204	40	G	Y	G/IB	P	Br	Ind	403	217	91	3.1	110
LG12-7214	40	T	Y	G	P	Br	Ind	399	213	115	3.8	115
LG12-7220	40	T	Y	Bl/G	P	Br/Tn	Ind	390	214	108	3.9	111
LG12-7224	40	G	Y	G/Y	P	Br	Ind	395	210	110	4.2	111
LG12-7227	40	G	Y	IB	P	Br	Ind	408	200	117	3.0	128
LG12-7235	40	G	Y	Bf	P	Br	Ind	395	212	79	3.8	108
LG12-7236	40	G	Y	Br	P	Br	Ind	400	211	99	3.5	116
LG12-7238	40	G	Y	G/IB	P	Br	Ind	404	215	78	2.4	107
LG12-7244	40	T	Y	Bl/G	P	Tn	Ind	397	215	119	3.3	114
LG12-7248	40	G	Y	Y	P	Tn	Ind	371	223	104	3.7	114
LG12-7251	40	G	Y	G/Y	P	Br	Ind	399	208	89	2.9	113
LG12-7253	40	T	Y	G	P	Br	Ind	402	207	97	3.2	108
LG12-7260	40	G	Y	Bf/IB	P	Br	Ind	408	212	111	3.7	123
LG12-7261	40	G	Y	IB	P	Br	Ind	389	214	103	3.3	121
LG12-7266	40	G	Y	IB	P	Br	Ind	389	213	94	3.3	111
LG12-7274	40	G	Y	G/IB	P	Br	Ind	396	216	108	4.2	112
LG12-7276	40	G	Y	IB	P	Br	Ind	384	218	98	3.8	112
LG12-7277	40	G	Y	Y	P	Br	Ind	388	218	104	3.6	111
LG12-7278	40	G	Y	G/Y	P	Br	Ind	389	219	103	3.3	119
LG12-7285	40	G	Y	G/Y	P	Br	Ind	388	216	109	3.2	121

Table 2 (cont.)

LG12-7287	40	G	Y	Bl	P	Br	Ind	401	216	88	2.5	111
LG12-7296	40	T	Y	G	P	Tn	Ind	407	220	112	3.7	112
LG12-7298	40	T	Y	G	P	Br	Ind	396	214	102	4.1	109
LG12-7322	40	T	Y	Bl/IB	P	Tn	Ind	401	206	113	4.2	120
LG12-7326	40	T	Y	Bl/Br	P	Br	Ind	392	214	100	3.8	106
LG12-7330	40	T	Y	Bl/G	P	Br	Ind	406	206	107	3.4	121
LG12-7335	40	G	Y	IB	P	Br	Ind	381	221	104	3.4	112
LG12-7512	40	T	Bl	Bl	P	Br	Ind	-	-	-	-	-
$LSD_{0.05}$								20	14	13	1.6	7

[1] Chromosome number;
[2] Pubescence color; T: Tawny; G: Gray; / means a mixture of phenotypes.
[3] Seed coat color; Y: Yellow; Bl: Black;
[4] Hilum color; Y: Yellow; Bl: Black; IB: Imperfect Black; Br: Brown; Bf: Buff;
[5] Flower color; P: Purple; W: White;
[6] Pod Color; Br: Brown; Tn: Tan;
[7] Stem termination; Det: Determinate; Ind: Indeterminate.
[8] Protein concentration and oil concentration reported on a dry weight basis;
[9] Height;
[10] Lodging score, 1 being upright and 10 being prostrate;
[11] Maturity, days after May 31st.
[12] Dash (-) means missing data.

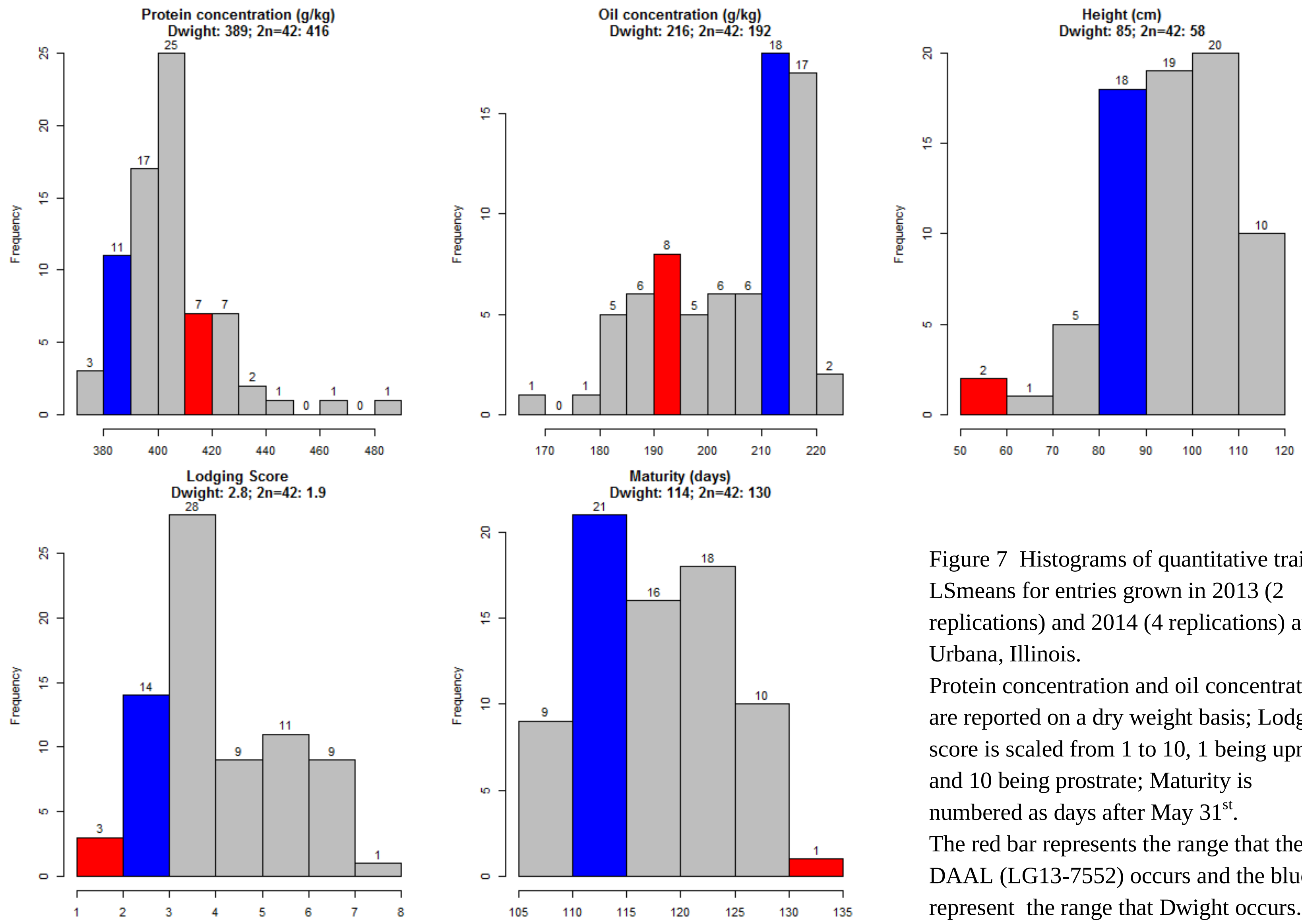

Figure 7 Histograms of quantitative trait LSmeans for entries grown in 2013 (2 replications) and 2014 (4 replications) at Urbana, Illinois.

Protein concentration and oil concentration are reported on a dry weight basis; Lodging score is scaled from 1 to 10, 1 being upright and 10 being prostrate; Maturity is numbered as days after May 31st.

The red bar represents the range that the DAAL (LG13-7552) occurs and the blue bar represent the range that Dwight occurs.

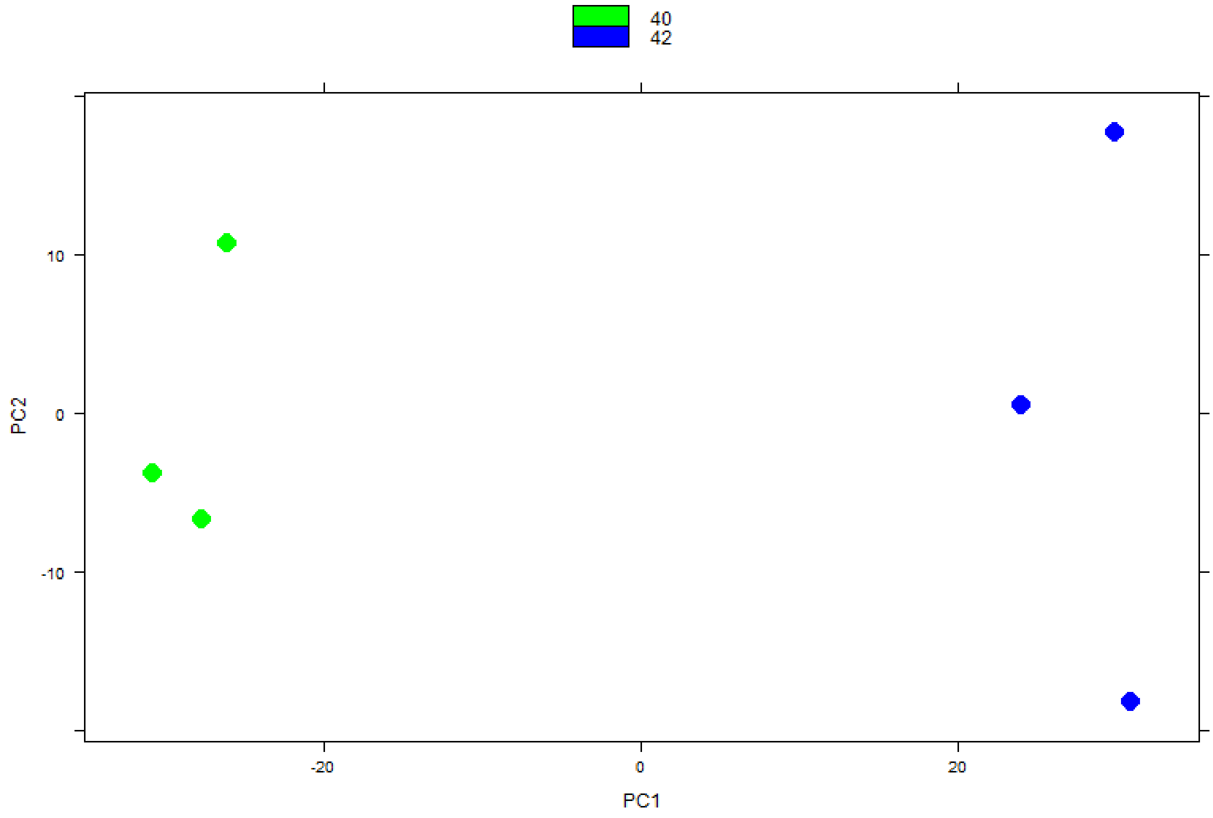

Figure 8 Principal component analysis of all sequence reads from alignment to the soybean reference genome Williams 82 from all six DAAL and disomic plants. Green dots represent the three disomic progeny plants (LG12-7063) and blue dots represent the three DAAL plants (LG13-7552).

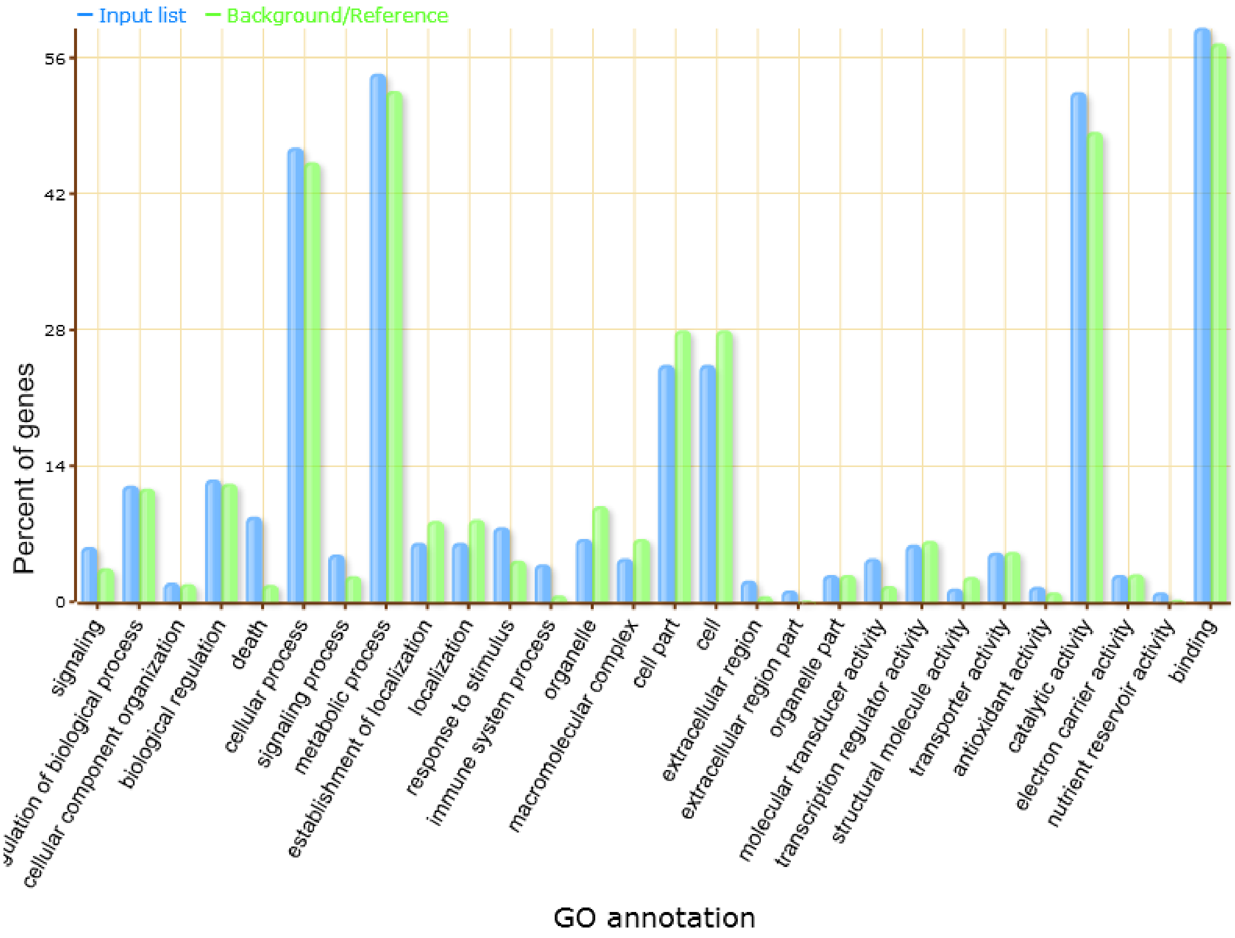

Figure 9 Gene Ontology (GO) classifications of differentially expressed genes (DEGs) between the DAAL (LG13-7552) and the disomic progeny (LG12-7063) generated from AgriGO using DEGs from aligning RNA sequencing reads to the Williams 82 soybean reference genome. These results combined the three main categories, biological process, cellular component and molecular function. The blue bar represents the DEGs list obtained from aligning to the soybean reference genome Williams 82. The green bar represents the Williams 82 genome. The x-axis annotates the GO terms and y-axis indicates the percentage of genes associated with the GO terms.

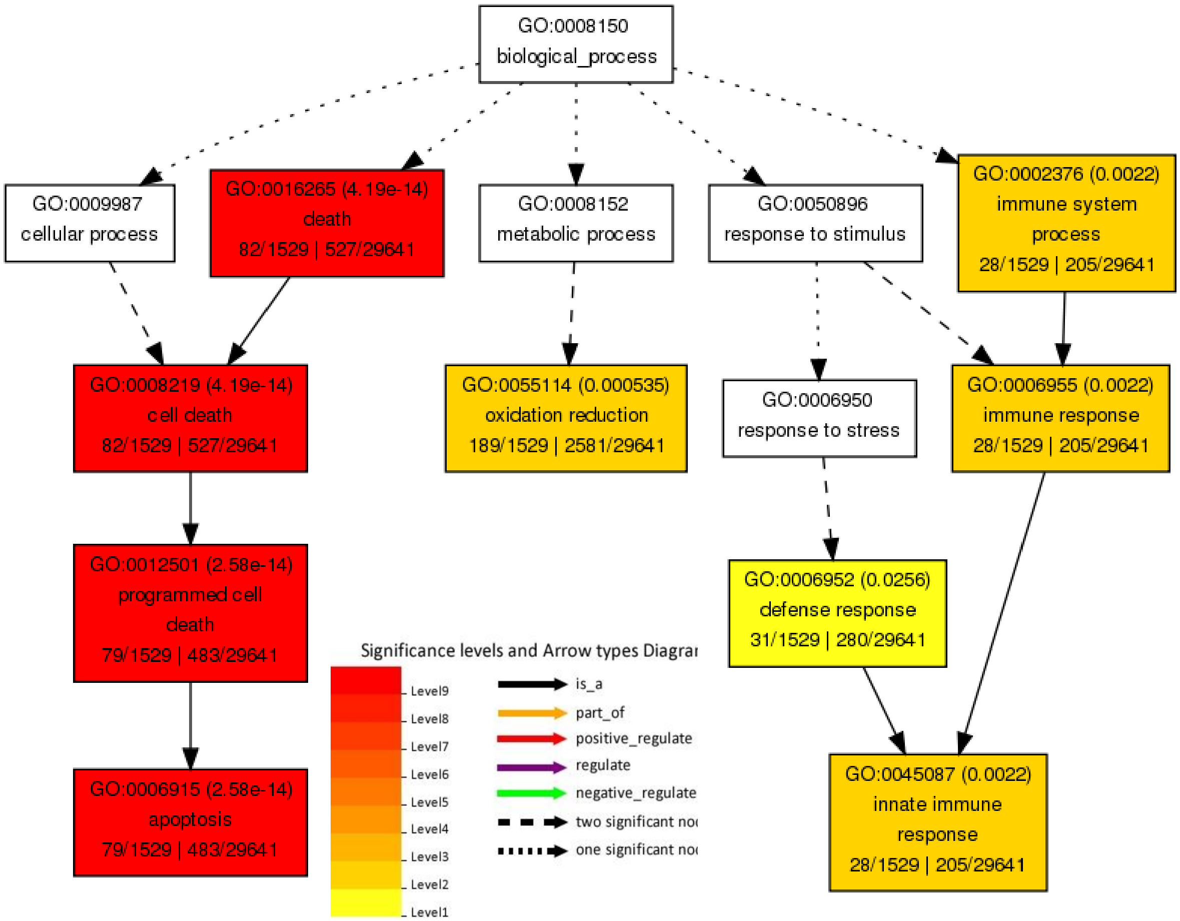

Figure 10 Enriched Gene Ontology analyses of differentially expressed genes (DEGs) between the DAAL (LG13-7552) and the disomic progeny (LG12-7063) in the biological process category generated from AgriGO using DEGs from aligning RNA sequencing reads to the Williams 82 soybean reference genome. Each box contains the GO term number and GO term, if significant, the p-value in parenthesis, the number of genes in the input list that are associated with the GO term, total number of annotated genes in the input list, the total number of genes in the reference that are associated with the GO term and the total number of genes in the reference are also indicated. The box colors reflect the level of significance of the analysis.

Table 3 Summary of functional annotations of contigs from Trinity *de novo* assembly using merged RNA sequencing data of the DAAL (LG13-7552) and the disomic progeny (LG12-7063)

	Total Contigs[1]	Primary Transcript[2]	Soybean Genome[3]	*G. tomentella*[4]	Repetitive Element Database[5]
Total	133,872	61,569	69,525	213	26,136
DEGs[6]	4,206	1,563	-[7]	51	1077
Up-regulated in LG13-7552 (DAAL)	1,437	701	-	0	399
Up-regulated in LG12-7063 (2n=40)	1,457	862	-	0	242
Only in LG13-7552 (DAAL)	989	-	-	51	342
Only in LG12-7063 (2n=40)	323	-	-	0	94

[1] Total number of *de novo* contigs from Trinity *de novo* assembly;

[2] Number of *de novo* contigs that hit to the primary transcript of the soybean reference genome Williams 82 from BLAST;

[3] Number of *de novo* contigs that hit to the soybean reference genome Williams 82 other than the primary transcript from BLAST;

[4] Number of *de novo* contigs that were not identified in the soybean reference genome Williams 82 but hit *G. tomentella* sequences publicly deposited in National Center for Biotechnology Information (NCBI) database from BLAST;

[5] Number of *de novo* contigs that were identified to be repetitive elements by RepeatMasker (RM).

[6] Differentially expressed genes

[7] Dash (-) indicates contig number does not apply to the category.

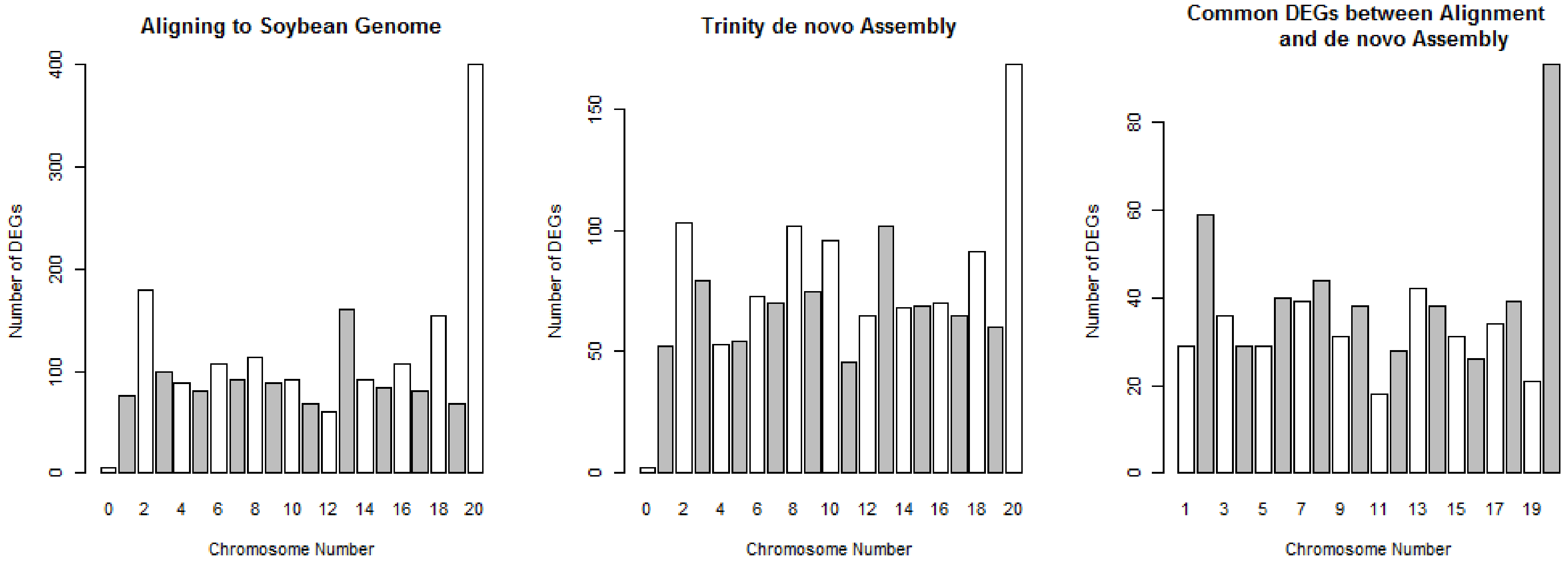

Figure 11 Distributions of differentially expressed genes (DEGs) between the DAAL (LG13-7552) and the disomic progeny (LG12-7063) on each chromosome from aligning RNA sequencing reads to the Williams 82 soybean reference genome, Trinity *de novo* assembly and common DEGs between the two methods. The x-axis indicated the chromosome number. The y-axis indicated the number of DEGs located on each chromosome.

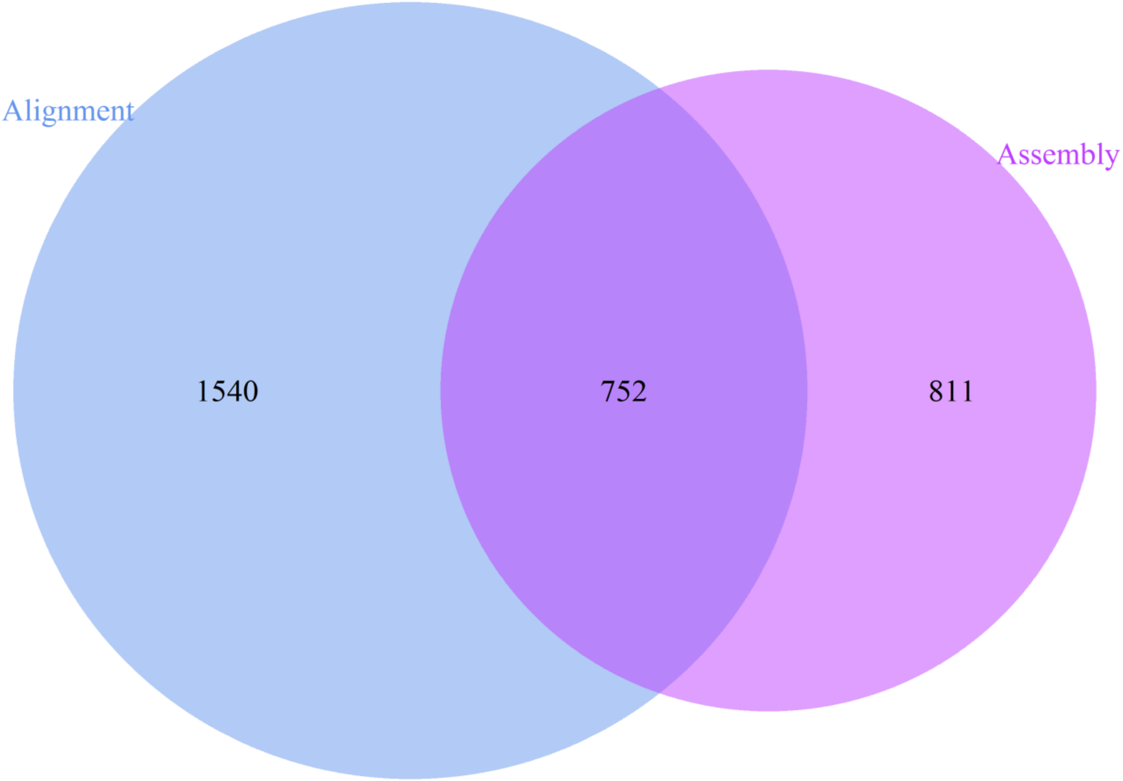

Figure 12 Venn diagram of lists of differentially expressed genes (DEGs) with transcripts in both the DAAL (LG13-7552) and the disomic progeny (LG12-7063) from aligning RNA sequencing reads to Williams 82 reference genome and Trinity *de novo* assembly. The blue area indicates the number of DEGs only present in the alignment DEGs list, the pink area indicates the number of DEGs only present in the Trinity *de novo* assembly DEGs list and the magenta area indicates the number of common DEGs in the two lists.

Table 4 Functional annotations of top 20 most abundant genes in the DAAL (LG13-7552, 2n=42) compared to the disomic progeny (LG13-7063, 2n=40) from the common differentially expressed genes between aligning RNA sequencing reads to the soybean reference genome Williams 82 and Trinity *de novo* assembly. All the annotations were acquired from Soybean Gene Expression Database (SGED).

GlymaID[1]	Annotations from NCBI BLAST[2]	Description[3]	Score[4]	E-value[5]	%ID[6]
Glyma20g33810	gi|356576401|ref| XP_003556320.1|	PREDICTED: anthocyanidin 3-O-glucosyltransferase-like [Glycine max]	941	0	100
Glyma16g31341	gi|356561629|ref| XP_003549083.1|	PREDICTED: LRR receptor-like serine/threonine-protein kinase GSO1-like [Glycine max]	2108	0	100
Glyma06g36140	gi|356514675|ref| XP_003526029.1|	PREDICTED: squamosa promoter-binding protein 1-like [Glycine max]	292	2E-95	100
Glyma02g13401	gi|356499925|ref| XP_003518786.1|	PREDICTED: MADS-box protein CMB1-like [Glycine max]	504	3E-175	100
Glyma17g31730	gi|356553315|ref| XP_003545002.1|	PREDICTED: 12-oxophytodienoate reductase 1-like [Glycine max]	585	0	77.1
Glyma17g08890	gi|356562644|ref| XP_003549579.1|	PREDICTED: agamous-like MADS-box protein AGL8-like [Glycine max]	490	2E-171	100
Glyma02g46311	gi|351722745|ref| NP_001235463.1|	uncharacterized protein LOC100527650 [Glycine max]	147	4E-41	99
Glyma14g37210	gi|356551857|ref| XP_003544289.1|	PREDICTED: dehydrodolichyl diphosphate synthase 2-like [Glycine max]	546	0	100
Glyma10g04230	gi|356536985|ref| XP_003537012.1|	PREDICTED: inorganic phosphate transporter 1-4-like [Glycine max]	1003	0	100
Glyma18g47090	gi|356566913|ref| XP_003551669.1|	PREDICTED: nucleolar protein 14-like [Glycine max]	1268	0	92.2
Glyma08g06730	gi|357476059|ref| XP_003608315.1|	Pathogenesis-related protein [Medicago truncatula]	450	8E-149	60.2
Glyma20g23220	gi|356575375|ref| XP_003555817.1|	PREDICTED: WUSCHEL-related homeobox 13-like [Glycine max]	577	0	100
Glyma03g16620	gi|357505745|ref| XP_003623161.1|	Protease inhibitor/seed storage/LTP family protein [Medicago truncatula]	179	1E-52	68
Glyma02g13420	gi|356499927|ref| XP_003518787.1|	PREDICTED: floral homeotic protein APETALA 1-like [Glycine max]	429	6E-147	100
Glyma20g30740	gi|225460644|ref| XP_002266350.1|	PREDICTED: thioredoxin Y1; chloroplastic [Vitis vinifera]	207	7E-62	79.3
Glyma18g29400	gi|356566177|ref| XP_003551311.1|	PREDICTED: AP2-like ethylene-responsive transcription factor PLT2-like [Glycine max]	908	0	100
Glyma17g06540	gi|356562052|ref| XP_003549289.1|	PREDICTED: uncharacterized protein LOC100793322 [Glycine max]	535	0	100
Glyma07g07145	gi|356520357|ref| XP_003528829.1|	PREDICTED: uncharacterized protein LOC100783381 [Glycine max]	698	0	87.4
Glyma18g53780	gi|356567298|ref| XP_003551858.1|	PREDICTED: probable methyltransferase PMT19-like [Glycine max]	1203	0	100
Glyma10g26990	gi|356533451|ref| XP_003535277.1|	PREDICTED: uncharacterized protein LOC100807304 [Glycine max]	691	0	100

[1] GlymaID identifier of differentially expressed genes (DEGs);
[2] Functional annotations of the DEGs from National Center for Biotechnology Information (NCBI) (blastx) database (2010);
[3] Descriptions of the functional annotations of the DEGs;
[4] BLAST score to the alignment quality between the functional annotations and the DEGs;
[5] The probability of the functional annotations and the DEGs sequences were matched by chance;
[6] Percent identity between the functional annotations and the DEGs.

Table 5 Functional annotations of top 20 most abundant genes in the disomic progeny (LG12-7063, 2n=40) compared to the DAAL (LG13-7552, 2n=42) from the common differentially expressed genes between aligning RNA sequencing reads to the soybean reference genome Williams 82 and Trinity *de novo* assembly. All the annotations were acquired from Soybean Gene Expression Database (SGED).

GlymaID[1]	Annotations from NCBI BLAST[2]	Description[3]	Score[4]	E-value[5]	%ID[6]
Glyma02g03280	gi|255640859|gb|ACU20712.1|	Unknown [Glycine max]	213	2E-63	100
Glyma05g25130	gi|356510754|ref|XP_003524099.1|	PREDICTED: reticuline oxidase-like protein-like [Glycine max]	825	0	78.1
Glyma01g05710	gi|357486941|ref|XP_003613758.1|	Disease resistance-like protein [Medicago truncatula]	1162	0	57.6
Glyma11g27920	No_hit				
Glyma02g42315	gi|356553790|ref|XP_003545235.1|	PREDICTED: probable LRR receptor-like serine/threonine-protein kinase At3g47570-like [Glycine max]	425	1E-137	65
Glyma02g10320	gi|356502432|ref|XP_003520023.1|	PREDICTED: heat shock cognate 70 kDa protein 2-like [Glycine max]	1201	0	100
Glyma06g16010	gi|356518775|ref|XP_003528053.1|	PREDICTED: ABC transporter G family member 5-like [Glycine max]	1264	0	100
Glyma03g22060	gi|356503056|ref|XP_003520328.1|	PREDICTED: TMV resistance protein N-like [Glycine max]	1343	0	97.3
Glyma04g38940	gi|356518777|ref|XP_003528054.1|	PREDICTED: pentatricopeptide repeat-containing protein At3g29230-like [Glycine max]	278	2E-85	91.9
Glyma06g43880	gi|356515120|ref|XP_003526249.1|	PREDICTED: UDP-glycosyltransferase 79B6-like [Glycine max]	929	0	100
Glyma04g04230	gi|83853828|gb|ABC47860.1|	N-hydroxycinnamoyl/benzoyltransferase 1 [Glycine max]	761	0	80.4
Glyma10g39800	gi|356577793|ref|XP_003557007.1|	PREDICTED: uncharacterized protein LOC100790784 [Glycine max]	95.5	1E-19	43.1
Glyma20g01470	gi|356577045|ref|XP_003556640.1|	PREDICTED: purple acid phosphatase 17-like [Glycine max]	653	0	100
Glyma03g03170	gi|356506370|ref|XP_003521957.1|	PREDICTED: probable LRR receptor-like serine/threonine-protein kinase At4g08850-like [Glycine max]	1447	0	100
Glyma19g01840	gi|356571919|ref|XP_003554118.1|	PREDICTED: cytochrome P450 82A4-like [Glycine max]	1087	0	100
Glyma05g36750	gi|351722649|ref|NP_001238275.1|	uncharacterized protein LOC100305633 [Glycine max]	277	4E-90	99.4
Glyma08g18033	gi|356528695|ref|XP_003532935.1|	PREDICTED: LOW QUALITY PROTEIN: 1-aminocyclopropane-1-carboxylate oxidase homolog 10-like [Glycine max]	634	0	99.7
Glyma07g08950	gi|356520493|ref|XP_003528896.1|	PREDICTED: gibberellin 20 oxidase 2-like [Glycine max]	830	0	100
Glyma11g00510	gi|356540317|ref|XP_003538636.1|	PREDICTED: cysteine-rich receptor-like protein kinase 10-like [Glycine max]	1151	0	100
Glyma02g40620	gi|356500976|ref|XP_003519306.1|	PREDICTED: medium-chain-fatty-acid--CoA ligase-like [Glycine max]	1111	0	100

[1] GlymaID identifier of differentially expressed genes (DEGs);
[2] Functional annotations of the DEGs from National Center for Biotechnology Information (NCBI)(blastx) database (2010);
[3] Descriptions of the functional annotations of the DEGs;
[4] BLAST score to the alignment quality between the functional annotations and the DEGs;
[5] The probability of the functional annotations and the DEGs sequences were matched by chance;
[6] Percent identity between the functional annotations and the DEGs.

References

Abbo, S., and G. Ladizinsky. 1991. Anatomical Aspects of Hybrid Embryo Abortion in the Genus *Lens* L. Botanical Gazette 152(3): 316–320.

Afzal, A.J., A.J. Wood, and D.A. Lightfoot. 2008. Plant receptor-like serine threonine kinases: roles in signaling and plant defense. Mol. Plant-Microbe Interact. MPMI 21(5): 507–517.

Ahmad, F., and A.E. Slinkard. 2004. The extent of embryo and endosperm growth following interspecific hybridization between *Cicer arietinum* L. and related annual wild species. Genetic Resources and Crop Evolution 51(7): 765–772.

Ahmad, M., D.L. McNeil, A.G. Fautrier, K.F. Armstrong, and A.M. Paterson. 1996. Genetic relationships in *Lens* species and parentage determination of their interspecific hybrids using RAPD markers. Theoret. Appl. Genetics 92(8): 1091–1098.

Ahmad, Q.N., E.J. Britten, and D.E. Byth. 1977. Inversion bridges and meiotic behavior in species hybrids of soybeans. J. Hered. 68(6): 360–364.

Ahmad, Q.N., E.J. Britten, and D.E. Byth. 1979. Inversion heterozygosity in the hybrid soybean × *Glycine soja* Evidence from a pachytene loop configuration and other meiotic irregularities. J Hered 70(6): 358–364.

Ali, S.N.H., M.S. Ramanna, E. Jacobsen, and R.G.F. Visser. 2001. Establishment of a complete series of a monosomic tomato chromosome addition lines in the cultivated potato using RFLP and GISH analyses. Theor Appl Genet 103(5): 687–695.

Altschul, S., T. Madden, A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, and D. Lipman. 1998. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Faseb J. 12(8): A1326–A1326.

Ananiev, E.V., O. Riera-Lizarazu, H.W. Rines, and R.L. Phillips. 1997. Oat–maize chromosome addition lines: A new system for mapping the maize genome. PNAS 94(8): 3524–3529.

Anders, S., and W. Huber. 2010. Differential expression analysis for sequence count data. Genome Biol. 11(10): R106.

Anders, S., P.T. Pyl, and W. Huber. 2014. HTSeq: A Python framework to work with high-throughput sequencing data. bioRxivAvailable at http://biorxiv.org/content/early/2014/02/20/002824 (verified 16 July 2014).

Bauer, S., T. Hymowitz, and G.R. Noel. 2007. Soybean cyst nematode resistance derived from *Glycine tomentella* in amphiploid (*G. max* × *G. tomentella*) hybrid lines. Nematropica 37(2): 277–285.

Bernard, R.L. 1967. The Inheritance of Pod Color in Soybeans. J. Hered. 58(4): 165–168.

Bernard, R.L. 1972. Two Genes Affecting Stem Termination in Soybeans. Crop Sci. 12(2): 235.

Bilgic, H., S. Cho, D.F. Garvin, and G.J. Muehlbauer. 2007. Mapping barley genes to chromosome arms by transcript profiling of wheat–barley ditelosomic chromosome addition lines. Genome 50(10): 898–906.

Blair, M.W., and P. Izquierdo. 2012. Use of the advanced backcross-QTL method to transfer seed mineral accumulation nutrition traits from wild to Andean cultivated common beans. Theor Appl Genet 125(5): 1015–1031.

Blair, M.W., G. Iriarte, and S. Beebe. 2006. QTL analysis of yield traits in an advanced backcross population derived from a cultivated Andean × wild common bean (*Phaseolus vulgaris* L.) cross. Theor Appl Genet 112(6): 1149–1163.

Blair, M.W., J.I. Medina, C. Astudillo, J. Rengifo, S.E. Beebe, G. Machado, and R. Graham. 2010. QTL for seed iron and zinc concentration and content in a Mesoamerican common bean (*Phaseolus vulgaris* L.) population. Theor Appl Genet 121(6): 1059–1070.

Blanco, A., R. Simeone, and P. Resta. 1987. The addition of *Dasypyrum villosum* (L.) Candargy chromosomes to durum wheat (*Triticum durum* Desf.). Theoret. Appl. Genetics 74(3): 328–333.

Bodanese-Zanettini, M.H., M.S. Lauxen, S.N.C. Richter, S. Cavalli-Molina, C.E. Lange, P.J. Wang, and C.Y. Hu. 1996. Wide hybridization between Brazilian soybean cultivars and wild perennial relatives. Theoret. Appl. Genetics 93(5-6): 703–709.

Boerma, H.R., J.E. Specht, and T. Hymowitz. 2004. Speciation and Cytogenetics. In Agronomy Monograph. American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

Broué, P., D.R. Marshall, and J.P. Grace. 1979. Hybridization among the Australian wild relatives of the soybean [*Glycine* spp.]. Journal of the Australian Institute of Agricultural Science (Australia)Available at http://agris.fao.org/agris-search/search.do?recordID=AU8001006 (verified 19 March 2015).

Broué, P., J. Douglass, J.P. Grace, and D.R. Marshall. 1982. Interspecific hybridisation of soybeans and perennial *Glycine* species indigenous to Australia via embryo culture. Euphytica 31(3): 715–724.

Brown, A.H.D., J.L. Doyle, J.P. Grace, and J.J. Doyle. 2002. Molecular phylogenetic relationships within and among diploid races of *Glycine tomentella* (Leguminosae). Aust. Syst. Bot. 15(1): 37–47.

Burdon, J.J., and D.R. Marshall. 1981. Evaluation of Australian native species of *Glycine* for resistance to soybean rust. Plant Dis. 65(1): 44–45.

Buzzell, R.I., B.R. Buttery, and D.C. MacTavish. 1987. Biochemical genetics of black pigmentation of soybean seed. J. Hered. 78(1): 53–54.

Campbell, C.G. 1997. Grass Pea, *Lathyrus Sativus* L. Bioversity International.

Cavalier-Smith, T. 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J. Cell Sci. 34(1): 247–278.

Chang, S.-B., and H. de Jong. 2005. Production of alien chromosome additions and their utility in plant genetics. Cytogenetic and Genome Research 109(1-3): 335–343.

Chetelat, R.T., C.M. Rick, P. Cisneros, K.B. Alpert, and J.W. DeVerna. 1998. Identification, transmission, and cytological behavior of *Solanum lycopersicoides* Dun. monosomic alien addition lines in tomato (*Lycopersicon esculentum* Mill.). Genome 41(1): 40–50.

Chung, G., and R.J. Singh. 2008. Broadening the Genetic Base of Soybean: A Multidisciplinary Approach. Crit. Rev. Plant Sci. 27(5): 295–341.

Chung, G.H., and J.H. Kim. 1990. Production of interspecific hybrids between *Glycine max* and *G. tomentella* through embryo culture. Euphytica 48(2): 97–101.

Chung, G.H., and K.S. Kim. 1991. Obtaining intersubgeneric hybridization between *Glycine max* and *G. latifolia* through embryo culture. Korean Journal of Plant Tissue Culture (Korea Republic)Available at http://agris.fao.org/agris-search/search.do?recordID=KR9300130 (verified 19 March 2015).

Cregan, P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, N. Kaya, T.T. VanToai, D.G. Lohnes, J. Chung, and J.E. Specht. 1999. An Integrated Genetic Linkage Map of the Soybean Genome. Crop Sci. 39(5): 1464.

Darvey, N.L., and J.P. Gustafson. 1975. Identification of Rye Chromosomes in Wheat-rye Addition Lines and Triticale by Heterochromatin Bands. Crop Science 15(2): 239.

Dhaliwal, H.S., Harjit-Singh, and M. William. 2002. Transfer of rust resistance from *Aegilops ovata* into bread wheat (*Triticum aestivum* L.) and molecular characterisation of resistant derivatives. Euphytica 126(2): 153–159.

Doležel, J., J. Vrána, J. Šafář, J. Bartoš, M. Kubaláková, and H. Šimková. 2012. Chromosomes in the flow to simplify genome analysis. Funct Integr Genomics 12(3): 397–416.

Dong, F., J.M. McGrath, J.P. Helgeson, and J. Jiang. 2001. The genetic identity of alien chromosomes in potato breeding lines revealed by sequential GISH and FISH analyses using chromosome-specific cytogenetic DNA markers. Genome 44(4): 729–734.

Doyle, J.J., J.L. Doyle, A.H.D. Brown, and R.G. Palmer. 2002. Genomes, multiple origins, and lineage recombination in the *Glycine tomentella* (Leguminosae) polyploid complex: histone H3-D gene sequences. Evolution 56(7): 1388–1402.

Doyle, M., J. Grant, and A. Brown. 1986. Reproductive Isolation Between Isozyme Groups of *Glycine tomentella* (Leguminosae), and Spontaneous Doubling in Their Hybrids. Aust. J. Bot. 34(5): 523–535.

Doyle, M.J., and A.H. Brown. 1985. Numerical analysis of isozyme variation in *Glycine tomentella*. Biochemical systematics and ecology 13(4): 413–419.

Du, J., D. Grant, Z. Tian, R.T. Nelson, L. Zhu, R.C. Shoemaker, and J. Ma. 2010. SoyTEdb: a comprehensive database of transposable elements in the soybean genome. BMC Genomics 11(1): 113.

Du, W., J. Wang, M. Lu, S. Sun, X. Chen, J. Zhao, Q. Yang, and J. Wu. 2013. Molecular cytogenetic identification of a wheat–*Psathyrostachys huashanica* Keng 5Ns disomic addition line with stripe rust resistance. Mol Breeding 31(4): 879–888.

Dvorak, J., and D.R. Knott. 1974. Disomic and Ditelosomic Additions of Diploid *Agropyron elongatum* Chromosomes to *Triticum aestivum*. Can. J. Genet. Cytol. 16(2): 399–417.

Errico, A., T. de Martino, R. Ercolano, L.M. Monti, and C. Conicella. 1996. Chromosome reconstruction in *Pisum sativum* through interspecific hybridization with *P. fulvum*. J. Genet. Breed. ItalyAvailable at http://agris.fao.org/agris-search/search.do?recordID=IT1998061308 (verified 17 February 2015).

Fávero, A.P., C.E. Simpson, J.F.M. Valls, and N.A. Vello. 2006. Study of the Evolution of Cultivated Peanut through Crossability Studies among *Arachis ipaënsis*, *A. duranensis*, and *A. hypogaea*. Crop Sci. 46(4): 1546.

Foncéka, D., T. Hodo-Abalo, R. Rivallan, I. Faye, M.N. Sall, O. Ndoye, A.P. Fávero, D.J. Bertioli, J.-C. Glaszmann, B. Courtois, and J.-F. Rami. 2009. Genetic mapping of wild introgressions into cultivated peanut: a way toward enlarging the genetic basis of a recent allotetraploid. BMC Plant Biology 9(1): 103.

Fox, S.L., E.N. Jellen, S.F. Kianian, H.W. Rines, and R.L. Phillips. 2001. Assignment of RFLP linkage groups to chromosomes using monosomic F1 analysis in hexaploid oat. Theor Appl Genet 102(2-3): 320–326.

Fratini, R., and M.L. Ruiz. 2006. Interspecific Hybridization in the Genus *Lens* Applying *in Vitro* Embryo Rescue. Euphytica 150(1-2): 271–280.

Friebe, B., L.L. Qi, S. Nasuda, P. Zhang, N.A. Tuleen, and B.S. Gill. 2000. Development of a complete set of *Triticum aestivum-Aegilops speltoides* chromosome addition lines. Theor Appl Genet 101(1-2): 51–58.

Friebe, B., Y. Mukai, H.S. Dhaliwal, T.J. Martin, and B.S. Gill. 1991. Identification of alien chromatin specifying resistance to wheat streak mosaic and greenbug in wheat germ plasm by C-banding and in situ hybridization. Theoret. Appl. Genetics 81(3): 381–389.

Garcia, G.M., H.T. Stalker, E. Shroeder, and G. Kochert. 1996. Identification of RAPD, SCAR, and RFLP markers tightly linked to nematode resistance genes introgressed from *Arachis cardenasii* into *Arachis hypogaea*. Genome 39(5): 836–845.

Geyt, J.P.C.V., M. Oléo, W. Lange, and T.S.M.D. Bock. 1988. Monosomic additions in beet (*Beta vulgaris*) carrying extra chromosomes of *Beta procumbens*. Theoret. Appl. Genetics 76(4): 577–586.

Goodstein, D.M., S. Shu, R. Howson, R. Neupane, R.D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, and D.S. Rokhsar. 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40(D1): D1178–D1186.

Gowda, M. V. C., Motagi, B. N., Sheshagiri, R., Naidu, G. K., & Rajendraprasad, M. N. 2002. Mutant 28-2: A Bold seeded Disease and Pest Resistant Groundnut Genotype for Karnataka, India. International Arachis Newsletter, 22, 32-33.

Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29(7): 644–652.

Gramzow, L., and G. Theissen. 2010. A hitchhiker's guide to the MADS world of plants. Genome Biol. 11(6): 214.

Grant, J., J. Grace, A. Brown, and E. Putievsky. 1984. Interspecific Hybridization in *Glycine* Willd. Subgenus *Glycine* (Leguminosae). Aust. J. Bot. 32(6): 655–663.

Grant, J.E., R. Pullen, A.H.D. Brown, J.P. Grace, and P.M. Gresshoff. 1986. Cytogenetic affinity between the new species *Glycine argyrea* and its congeners. J Hered 77(6): 423–426.

Gupta, D., and S.K. Sharma. 2007. Widening the gene pool of cultivated lentils through introgression of alien chromatin from wild *Lens* subspecies. Plant Breeding 126(1): 58–61.

Hajjar, R., and T. Hodgkin. 2007. The use of wild relatives in crop improvement: a survey of developments over the last 20 years. Euphytica 156(1-2): 1–13.

Harlan, J.R., and J.M.J. de Wet. 1971. Toward a Rational Classification of Cultivated Plants. Taxon 20(4): 509–517.

Hartman, G.L., M.E. Gardner, T. Hymowitz, and G.C. Naidoo. 2000. Evaluation of Perennial Species for Resistance to Soybean Fungal Pathogens That Cause Sclerotinia Stem Rot and Sudden Death Syndrome. Crop Science 40(2): 545.

Hartman, G.L., T.C. Wang, and T. Hymowitz. 1992. Sources of resistance to soybean rust in perennial *Glycine* species. Plant Dis. 76(4): 396–399.

Huang, W.-L., C.-W. Tung, S.-W. Ho, S.-F. Hwang, and S.-Y. Ho. 2008. ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. BMC Bioinformatics 9(1): 80.

Jacobsen, E., J.H. de Jong, S.A. Kamstra, P.M.M.M. van den Berg, and M.S. Ramanna. 1995. Genomic *in situ* hybridization (GISH) and RFLP analysis for the identification of alien chromosomes in the backcross progeny of potato (+) tomato fusion hybrids. Heredity 74(3): 250–257.

Jena, K.K., and G.S. Khush. 1989. Monosomic alien addition lines of rice: production, morphology, cytology, and breeding behavior. Genome 32(3): 449–455.

Jena, K.K., and G.S. Khush. 1990. Introgression of genes from *Oryza officinalis* Well ex Watt to cultivated rice, *O. sativa* L. Theoretical and Applied Genetics 80(6): 737–745.

Ji, Y., and R. Chetelat. 2003. Homoeologous pairing and recombination in *Solanum lycopersicoides* monosomic addition and substitution lines of tomato. Theor Appl Genet 106(6): 979–989.

Kamsteeg, J., J. van Brederode, and G. van Nigtevecht. 1978. Identification and properties of UDP-glucose: Cyanidin-3-O-glucosyltransferase isolated from petals of the red campion (*Silene dioica*). Biochem. Genet. 16(11-12): 1045–1058.

Kaneko, Y., S.W. Bang, and Y. Matsuzawa. 2000. Early-bolting trait and RAPD markers in the specific monosomic addition line of radish carrying the e-chromosome of *Brassica oleracea*. Plant Breed. 119(2): 137–140.

Kao, W.-Y., T.-T. Tsai, H.-C. Tsai, and C.-N. Shih. 2006. Response of three *Glycine* species to salt stress. Environmental and Experimental Botany 56(1): 120–125.

Karasawa, K. 1936. Crossing experiments with *Glycine soja* and *G. ussuriensis.* Jap. J. Bot, 8, 113-118.

Khush, G. S. 1973. Cytogenetics of Aneuploids (Academic, New York). KhushCytogenetics of Aneuploids1973.

Kollipara, K.P., and T. Hymowitz. 1992. Characterization of trypsin and chymotrypsin inhibitors in the wild perennial *Glycine* species. J. Agric. Food Chem. 40(12): 2356–2363.

Kollipara, K.P., R.J. Singh, and T. Hymowitz. 1993. Genomic diversity in aneudiploid (2n=38) and diploid (2n=40) *Glycine tomentella* revealed by cytogenetic and biochemical methods. Genome 36(3): 391–396.

Kollipara, K.P., R.J. Singh, and T. Hymowitz. 1994. Genomic diversity and multiple origins of tetraploid (2n = 78, 80) *Glycine tomentella*. Genome 37(3): 448–459.

Kollipara, K.P., R.J. Singh, and T. Hymowitz. 1995. Genomic Relationships in the Genus *Glycine* (Fabaceae: Phaseoleae): Use of a Monoclonal Antibody to the Soybean Bowman-Birk Inhibitor as a Genome Marker. Am. J. Bot. 82(9): 1104–1111.

Kollipara, K.P., R.J. Singh, and T. Hymowitz. 1997. Phylogenetic and genomic relationships in the genus *Glycine* Willd. based on sequences from the ITS region of nuclear rDNA. Genome 40(1): 57–68.

Kubaláková, M., M. Valárik, J. Bartoš, J. Vrána, J. Cíhalíková, M. Molnár-Láng, and J. Dolezel. 2003. Analysis and sorting of rye (*Secale cereale* L.) chromosomes using flow cytometry. Genome 46(5): 893–905.

Kynast, R.G., O. Riera-Lizarazu, M.I. Vales, R.J. Okagaki, S.B. Maquieira, G. Chen, E.V. Ananiev, W.E. Odland, C.D. Russell, A.O. Stec, S.M. Livingston, H.A. Zaia, H.W. Rines, and R.L. Phillips. 2001. A Complete Set of Maize Individual Chromosome Additions to the Oat Genome. Plant Physiol. 125(3): 1216–1227.

Ladizinsky, G., C.A. Newell, and T. Hymowitz. 1979. Wide crosses in soybeans: Prospects and limitations. Euphytica 28(2): 421–423.

Larkin, P.J., P.M. Banks, E.S. Lagudah, R. Appels, C. Xiao, X. Zhiyong, H.W. Ohm, and R.A. McIntosh. 1995. Disomic *Thinopyrum intermedium* addition lines in wheat with barley yellow dwarf virus resistance and with rust resistances. Genome 38(2): 385–394.

Loux, M.M., R.A. Liebl, and T. Hymowitz. 1987. Examination of wild perennial *Glycine* species for glyphosate tolerance. Soybean genetics newsletter - United States, Agricultural Research Service (USA)Available at http://agris.fao.org/agris-search/search.do?recordID=US874656088 (verified 17 March 2015).

Luo, P.G., H.Y. Luo, Z.J. Chang, H.Y. Zhang, M. Zhang, and Z.L. Ren. 2009. Characterization and chromosomal location of Pm40 in common wheat: a new gene for resistance to powdery mildew derived from *Elytrigia intermedium*. Theor Appl Genet 118(6): 1059–1064.

Mallikarjuna, N., S. Senthilvel, and D. Hoisington. 2010. Development of new sources of tetraploid Arachis to broaden the genetic base of cultivated groundnut (*Arachis hypogaea* L.). Genet Resour Crop Evol 58(6): 889–907.

Martis, M.M., R. Zhou, G. Haseneyer, T. Schmutzer, J. Vrána, M. Kubaláková, S. König, K.G. Kugler, U. Scholz, B. Hackauf, V. Korzun, C.-C. Schön, J. Doležel, E. Bauer, K.F.X. Mayer, and N. Stein. 2013. Reticulate Evolution of the Rye Genome. Plant Cell 25(10): 3685–3698.

Mayer, K.F.X., M. Martis, P.E. Hedley, H. Šimková, H. Liu, J.A. Morris, B. Steuernagel, S. Taudien, S. Roessner, H. Gundlach, M. Kubaláková, P. Suchánková, F. Murat, M. Felder, T. Nussbaumer, A. Graner, J. Salse, T. Endo, H. Sakai, T. Tanaka, T. Itoh, K. Sato, M. Platzer, T. Matsumoto, U. Scholz, J. Doležel, R. Waugh, and N. Stein. 2011. Unlocking the Barley Genome by Chromosomal and Comparative Genomics. Plant Cell 23(4): 1249–1263.

Mesbah, M., T.S.M. de Bock, J.M. Sandbrink, R.M. Klein-Lankhorst, and W. Lange. 1997. Molecular and morphological characterization of monosomic additions in *Beta vulgaris*, carrying extra chromosomes of *B. procumbens* or *B. patellaris*. Molecular Breeding 3(2): 147–157.

Methods for producing fertile crosses between wild and domestic soybean species. 2010.

Miller, T.E. 1984. The homoeologous relationship between the chromosomes of rye and wheat. Current status. Can. J. Genet. Cytol. 26(5): 578–589.

Morgan, W.G. 1991. The morphology and cytology of monosomic addition lines combining single *Festuca drymeja* chromosomes and *Lolium multiflorum*. Euphytica 55(1): 57–63.

Muehlbauer, G.J., O. Riera-Lizarazu, R.G. Kynast, D. Martin, R.L. Phillips, and H.W. Rines. 2000. A maize chromosome 3 addition line of oat exhibits expression of the maize homeobox gene *liguleless*3 and alteration of cell fates. Genome 43(6): 1055–1064.

Multani, D.S., G.S. Khush, B.G. delos Reyes, and D.S. Brar. 2003. Alien genes introgression and development of monosomic alien addition lines from *Oryza latifolia* Desv. to rice, *Oryza sativa* L. TAG Theoretical and Applied Genetics 107(3): 395–405.

Multani, D.S., K.K. Jena, D.S. Brar, B.G. de los Reyes, E.R. Angeles, and G.S. Khush. 1994. Development of monosomic alien addition lines and introgression of genes from *Oryza australiensis* Domin. to cultivated rice *O. sativa* L. Theoret. Appl. Genetics 88(1): 102–109.

Muñoz, L.C., M.C. Duque, D.G. Debouck, and M.W. Blair. 2006. Taxonomy of Tepary Bean and Wild Relatives as Determined by Amplified Fragment Length Polymorphism (AFLP) Markers. Crop Science 46(4): 1744.

Muñoz, L.C., M.W. Blair, M.C. Duque, J. Tohme, and W. Roca. 2004. Introgression in Common Bean × Tepary Bean Interspecific Congruity-Backcross Lines as Measured by AFLP Markers. Crop Science 44(2): 637.

Nakajima, J., Y. Tanaka, M. Yamazaki, and K. Saito. 2001. Reaction Mechanism from Leucoanthocyanidin to Anthocyanidin 3-Glucoside, a Key Reaction for Coloring in Anthocyanin Biosynthesis. J. Biol. Chem. 276(28): 25797–25803.

Newell, C.A., and T. Hymowitz. 1982. Successful Wide Hybridization Between the Soybean and a Wild Perennial Relative, *G. tomentella* Hayata1. Crop Science 22(5): 1062.

Newell, C.A., X. Delannay, and M.E. Edge. 1987. Interspecific hybrids between the soybean and wild perennial relatives. J Hered 78(5): 301–306.

Nix, D.A., S.J. Courdy, and K.M. Boucher. 2008. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC Bioinformatics 9(1): 523.

O’mara, J.G. 1940. Cytogenetic Studies on Triticale. I. a Method for Determining the Effects of Individual *Secale* Chromosomes on *Triticum*. Genetics 25(4): 401–408.

Orellana, J., M.C. Cermeño, and J.R. Lacadena. 1984. Meiotic pairing in wheat–rye addition and substitution lines. Can. J. Genet. Cytol. 26(1): 25–33.

Palmer, R.G., K.E. Newhouse, R.A. Graybosch, and X. Delannay. 1987. Chromosome structure of the wild soybean Accessions from China and the Soviet Union of *Glycine soja* Sieb. & Zucc. J. Hered. 78(4): 243–247.

Pantalone, V.R., W.J. Kenworthy, L.H. Slaughter, and B.R. James. 1997. Chloride tolerance in soybean and perennial *Glycine* accessions. Euphytica 97(2): 235–239.

Putievsky, E., and P. Broue. 1979. Cytogenetics of Hybrids Among Perennial Species of *Glycine* Subgenus *Glycine*. Aust. J. Bot. 27(6): 713–723.

Qi, L.L., M.O. Pumphrey, B. Friebe, P.D. Chen, and B.S. Gill. 2008. Molecular cytogenetic characterization of alien introgressions with gene Fhb3 for resistance to Fusarium head blight disease of wheat. Theor Appl Genet 117(7): 1155–1166.

Rauscher, J.T., J.J. Doyle, and A.H.D. Brown. 2002. Internal transcribed spacer repeat-specific primers and the analysis of hybridization in the *Glycine tomentella* (Leguminosae) polyploid complex. Molecular Ecology 11(12): 2691–2702.

Reamon-Ramos, S.M., and G. Wricke. 1992. A full set of monosomic addition lines in *Beta vulgaris* from *Beta webbiana*: morphology and isozyme markers. Theoret. Appl. Genetics 84(3-4): 411–418.

Reese, P.F., and H.R. Boerma. 1989. Additional Genes for Green Seed Coat in Soybean. J. Hered. 80(1): 86–88.

Riggs, R.D., S. Wang, R.J. Singh, and T. Hymowitz. 1998. Possible Transfer of Resistance to *Heterodera glycines* from *Glycine tomentella* to *Glycine max*. J Nematol 30(4S): 547–552.

Robinson, M.D., and G.K. Smyth. 2008. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9(2): 321–332.

Sakai, T., and N. Kaizuma. 1985. Hybrid Embryo Formation in an Intersubgeneric Cross of Soybean (*Glycin max* MERILL) with a Wild Relative (*G.tomentella* HAYATA). Japanese Journal of Breeding 35(4): 363–374.

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D.L. Hyten, Q. Song, J.J. Thelen, J. Cheng, D. Xu, U. Hellsten, G.D. May, Y. Yu, T. Sakurai, T. Umezawa, M.K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X.-C. Zhang, K. Shinozaki, H.T. Nguyen, R.A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R.C. Shoemaker, and S.A. Jackson. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463(7278): 178–183.

Schoen, D.J., J.J. Burdon, and A.H.D. Brown. 1992. Resistance of *Glycine tomentella* to soybean leaf rust *Phakopsora pachyrhizi* in relation to ploidy level and geographic distribution. Theoret. Appl. Genetics 83(6-7): 827–832.

Schwarzacher, T., A.R. Leitch, M.D. Bennett, and J.S. Heslop-Harrison. 1989. *In Situ* Localization of Parental Genomes in a Wide Hybrid. Ann Bot 64(3): 315–324.

Shepherd, K.W., and A.K.M.R. Islam. 1988. Fourth compendium of wheat-alien chromosome lines. p. 1373–1398. In Institute of Plant Science Research.

Sievers, F., A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, and D.G. Higgins. 2011. Fast, scalable generation of high quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7(1): 539.

Simpson, C. E., J. L. Starr, S. C. Nelson, K. E. Woodard and O. D. Smith, 1993. Registration of TxAG-6 and TxAG-7 peanut germplasm. Crop Sci 33: 1418.

Simpson, C.E. 1991. Pathways for Introgression of Pest Resistance into *Arachis hypogaea* L. Peanut Science 18(1): 22–26.

Singh, A. K. (1985, October). Genetic introgression from compatible wild species into cultivated groundnut. In Proceedings of the International Workshop on Cytogenetics of Arachis. ICRISAT Press, Patancheru (p. 107).

Singh, R.J. 2006. Genetic Resources, Chromosome Engineering, and Crop Improvement: Oilseed Crops. CRC Press.

Singh, R.J., and T. Hymowitz. 1985. An intersubgeneric hybrid between *Glycine tomentella* Hayata and the soybean, *G. max* (L.) Merr. Euphytica 34(1): 187–192.

Singh, R.J., and T. Hymowitz. 1987. Intersubgeneric Crossability in the Genus *Glycine* Willd. Plant Breeding 98(2): 171–173.

Singh, R.J., and T. Hymowitz. 1988. The genomic relationship between *Glycine max* (L.) Merr. and *G. soja* Sieb. and Zucc. as revealed by pachytene chromosome analysis. Theor. Appl. Genet. 76(5): 705–711.

Singh, R.J., G.H. Chung, and R.L. Nelson. 2007. Landmark research in legumes. Genome 50(6): 525–537.

Singh, R.J., K.P. Kollipara, and T. Hymowitz. 1987. Intersubgeneric hybridization of soybeans with a wild perennial species, *Glycine clandestina* Wendl. Theor. Appl. Genet. 74(3): 391–396.

Singh, R.J., K.P. Kollipara, and T. Hymowitz. 1988. Further data on the genomic relationships among wild perennial species (2 n= 40) of the genus *Glycine* Willd. Genome 30(2): 166–176.

Singh, R.J., K.P. Kollipara, and T. Hymowitz. 1990. Backcross-Derived Progeny from Soybean and *Glycine tomentella* Hayata Intersubgeneric Hybrids. Crop Science 30(4): 871.

Singh, R.J., K.P. Kollipara, and T. Hymowitz. 1992a. Genomic relationships among diploid wild perennial species of the genus *Glycine* Willd. subgenus *Glycine* revealed by crossability, meiotic chromosome pairing and seed protein electrophoresis. Theor. Appl. Genet. 85(2-3): 276–282.

Singh, R.J., K.P. Kollipara, and T. Hymowitz. 1992b. Genomic relationships among diploid wild perennial species of the genus *Glycine* Willd. subgenus *Glycine* revealed by crossability, meiotic chromosome pairing and seed protein electrophoresis. Theor. Appl. Genet. 85(2-3): 276–282.

Singh, R.J., K.P. Kollipara, and T. Hymowitz. 1998. Monosomic Alien Addition Lines Derived from *Glycine max* (L.) Merr. and *G. tomentella* Hayata: Production, Characterization, and Breeding Behavior. Crop Science 38(6): 1483.

Smýkal, P., & Kosterin, O. 2010. Towards introgression library carrying wild pea (*Pisum fulvum*) segments in cultivated pea (*Pisum sativum*) genome background. In Book of Abstracts of Vth International Congress on Legume Genetics and Genomics (p. 123).

Suchánková, P., M. Kubaláková, P. Kovářová, J. Bartoš, J. Číhalíková, M. Molnár-Láng, T.R. Endo, and J. Doležel. 2006. Dissection of the nuclear genome of barley by chromosome flow sorting. Theor Appl Genet 113(4): 651–659.

Suen, D.F., C.K. Wang, R.F. Lin, Y.Y. Kao, F.M. Lee, and C.C. Chen. 1997. Assignment of DNA markers to *Nicotiana sylvestris* chromosomes using monosomic alien addition lines. Theor Appl Genet 94(3-4): 331–337.

Sybenga, J. 1992. Cytogenetics in plant breeding. : xvi + 469 pp.

Tanksley, S.D., M.W. Ganal, J.P. Prince, M.C. de Vicente, M.W. Bonierbale, P. Broun, T.M. Fulton, J.J. Giovannoni, S. Grandillo, and G.B. Martin. 1992. High density molecular linkage maps of the tomato and potato genomes. Genetics 132(4): 1141–1160.

Tindale, M. 1986. Taxonomic notes on three Australian and Norfolk Island species of *Glycine* Willd. (Fabaceae: Phaseolae) including the choice of a Neotype for *G.clandestina* Wendl. Brunonia 9(2): 179–191.

Toda, K., D. Yang, N. Yamanaka, S. Watanabe, K. Harada, and R. Takahashi. 2002. A single-base deletion in soybean flavonoid 3'-hydroxylase gene is associated with gray pubescence color. Plant Mol. Biol. 50(2): 187–196.

Tuteja, J.H., S.J. Clough, W.-C. Chan, and L.O. Vodkin. 2004. Tissue-Specific Gene Silencing Mediated by a Naturally Occurring Chalcone Synthase Gene Cluster in *Glycine max*. Plant Cell Online 16(4): 819–835.

Verma, M.M., Ravi, and J.S. Sandhu. 1995. Characterization of the interspecific cross *Cicer anetinum* L. ×*C. judaicum* (Boiss). Plant Breeding 114(6): 549–551.

White, R.H., R.A. Liebl, and T. Hymowitz. 1990. Examination of 2,4-D tolerance in perennial *Glycine* species. Pestic. Biochem. Physiol. 38(2): 153–161.

Wilson, J.N., M.R. Baring, M.D. Burow, W.L. Rooney, and C.E. Simpson. 2013. Generation Means Analysis of Oil Concentration in Peanut. Journal of Crop Improvement 27(1): 85–95.

Woodworth, C.M. 1923. Inheritance of growth habit, pod color, and flower color in soybeans. Agron. J.Available at http://agris.fao.org/agris-search/search.do?recordID=US201302674386 (verified 9 April 2015).

Yu, X., K. Guda, J. Willis, M. Veigl, Z. Wang, S. Markowitz, M.D. Adams, and S. Sun. 2012. How do alignment programs perform on sequencing data with varying qualities and from repetitive regions BioData Min. 5: 6.

Zhang, J., X. Li, R. R. ‐ C. Wang, A. Cortes, V. Rosas, and A. Mujeeb ‐ Kazi. 2002. Molecular Cytogenetic Characterization of Eb-Genome Chromosomes in *Thinopyrum bessarabicum* Disomic Addition Lines of Bread Wheat. International Journal of Plant Sciences 163(1): 167–174.

Zhang, S., E. Schliephake, and H. Budahn. 2014. Chromosomal assignment of oil radish resistance to *Meloidogyne incognita* and *M. javanica* using a set of disomic rapeseed-radish chromosome addition lines. Nematology 16: 1119–1127.

Zheng, C., P. Chen, T. Hymowitz, S. Wickizer, and R. Gergerich. 2005. Evaluation of *Glycine* species for resistance to Bean pod mottle virus. Crop Protection 24(1): 49–56.

Zhuang B., Sun Y., Lu Q., Wang Y., and Xu B.. 1996. A study on resistance to soybean mosaic virus and *Aphis glycinece* of perennial wild soybean. Soybean Genetics Newsletter 23: 66–69.

Zou, J., S. Rodriguez-Zas, M. Aldea, M. Li, J. Zhu, D.O. Gonzalez, L.O. Vodkin, E. DeLucia, and S.J. Clough. 2005. Expression Profiling Soybean Response to Pseudomonas syringae Reveals New Defense-Related Genes and Rapid HR-Specific Downregulation of Photosynthesis. Mol. Plant. Microbe Interact. 18(11): 1161–1174.

Zou, J.J., R.J. Singh, J. Lee, S.J. Xu, P.B. Cregan, and T. Hymowitz. 2003. Assignment of molecular linkage groups to soybean chromosomes by primary trisomics. Theor. Appl. Genet. 107(4): 745–750.

Appendix A

R code for statistical analysis and partial results

```
setwd ("C:/Users/swang130/Desktop/Research")
savehistory("Quantitative trait.Rhistory")
loadhistory("Quantitative trait.Rhistory")

# Read Table and load module
Quanti_table = read.csv("C:/Users/swang130/Desktop/Research/Quantitative traits1.csv", header=T,
na.strings = ".")
Quanti_42 = read.csv("C:/Users/swang130/Desktop/Research/Quantitative traits42.csv", header=T,
na.strings = ".")

library("pbkrtest")
library("lsmeans")
library("lme4")
library("lattice")
library("predictmeans")

#########################################
#         Quantitative traits           #
#########################################

# Combine 2n=40 and 2n=42 quantitative data
Quan = rbind(Quanti_42, Quanti_table)
attach(Quan)
as.factor(Block)

# Protein
  # Linear Model
  Protein_DB_lm = lmer(Protein_DB ~ 1+ Entry + (1|Year) + (1|Entry:Year) + (Block|Year),
data=Quan, REML=T)
  summary(Protein_DB_lm)
        #Random effects:
        #Groups     Name        Variance Std.Dev. Corr
        #Entry.Year (Intercept) 0.4903   0.7002
        #Year       (Intercept) 1.5381   1.2402
        #           Block       0.1469   0.3833   -1.00
        #Year.1     (Intercept) 0.0000   0.0000
        #Residual               1.3137   1.1462
        #Number of obs: 450, groups:  Entry:Year, 150; Year, 2
  # ANOVA
  anova(Protein_DB_lm)
        #Analysis of Variance Table of type 3  with  Satterthwaite
        #approximation for degrees of freedom
        #      Sum Sq Mean Sq NumDF  DenDF F.value   Pr(>F)
        #Entry 706.13  9.5423    74 66.489  7.2634 1.51e-14 ***
        #  ---
        #  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  # LSMean
  Protein_DB_rg = ref.grid(Protein_DB_lm)   # establish the reference grid for LSMEANS.
  summary(Protein_DB_rg)
  lsm_Protein = lsmeans(Protein_DB_rg, "Entry")
  sum_lsm_Protein = summary(lsm_Protein)       # convert lsmean into dataframe
  sum_lsm_Protein = sum_lsm_Protein[order(sum_lsm_Protein$lsmean, decreasing = T),,drop=F]   #
sort data from the largest to the smallest
    # Plot by Entry
  plot(sum_lsm_Protein, by = "Entry",main="Protein")
    # LSD
  LSD_Protein = predictmeans(Protein_DB_lm,"Entry")$LSD
        #[1] 1.95

###############################################################################

# Oil
  # Linear Model
  Oil_DB_lm = lmer(Oil_DB ~1 + Entry + (1|Year) + (1|Entry:Year) + (Block |Year),
data=Quan,REML=T)
  summary(Oil_DB_lm)
        #Random effects:
        #  Groups     Name        Variance     Std.Dev.  Corr
        #Entry.Year (Intercept) 0.3283695471 0.5730354
        #Year       (Intercept) 0.9529915298 0.9762129
        #           Block       0.0001679705 0.0129603 1.00
        #Year.1     (Intercept) 0.0000000101 0.0001005
        #Residual               0.3735099174 0.6111546
        #Number of obs: 450, groups:  Entry:Year, 150; Year, 2
  # ANOVA
  anova(Oil_DB_lm)
          # Analysis of Variance Table
          #       Df Sum Sq Mean Sq F value
          #Entry  74 194.92  2.6341  7.0522
  # LSMean
  Oil_DB_rg = ref.grid(Oil_DB_lm)   # establish the reference grid for LSMEANS.
  summary(Oil_DB_rg)
  lsm_Oil = lsmeans(Oil_DB_rg, "Entry")
  sum_lsm_Oil = summary(lsm_Oil)      # convert lsmean into dataframe
  sum_lsm_Oil = sum_lsm_Oil[order(sum_lsm_Oil$lsmean, decreasing = T),,drop=F]   # sort data from
the largest to the smallest
    # Plot by Entry
  plot(sum_lsm_Oil, by = "Entry", main="Oil")
    # LSD
  LSD_Oil = predictmeans(Oil_DB_lm,"Entry")$LSD
        #[1] 1.36

###############################################################################

# Height
  # Linear Model
    HGT_lm = lmer(HGT ~ Entry + (1|Year) + (1|Entry:Year) + (Block|Year), data=Quan,REML=T)
  summary(HGT_lm)
        #Random effects:
        #Groups     Name        Variance Std.Dev. Corr
        #Entry.Year (Intercept) 22.354   4.728
        #Year       (Intercept) 27.521   5.246
        #           Block        3.787   1.946    1.00
        #Year.1     (Intercept)  0.000   0.000
        #Residual               50.331   7.094
        #Number of obs: 450, groups:  Entry:Year, 150; Year, 2
  # ANOVA
  anova(HGT_lm)
        #Analysis of Variance Table
        #      Df Sum Sq Mean Sq F value
        #Entry 74  31956  431.84  8.5801
  # LSMean
  HGT_rg = ref.grid(HGT_lm)   # establish the reference grid for LSMEANS.
  summary(HGT_rg)
  lsm_HGT = lsmeans(HGT_rg, "Entry")
  sum_lsm_HGT = summary(lsm_HGT)      # convert lsmean into dataframe
  sum_lsm_HGT = sum_lsm_HGT[order(sum_lsm_HGT$lsmean, decreasing = T),,drop=F]   # sort data from
the largest to the smallest
    # Plot by Entry
  plot(sum_lsm_HGT, by = "Entry", main="Height")
    # LSD
  predictmeans(HGT_lm,"Entry")
  LSD_HGT =predictmeans(HGT_lm,"Entry")$LSD
      #[1] 12.7

###############################################################################

#Lodging
  # Linear Model
    LDG_lm = lmer(LDG ~ 1 + Entry + (1|Year) + (1|Entry:Year) + (Block |Year), data=Quan,REML=T)
  summary(LDG_lm)
        #Random effects:
        #Groups     Name        Variance Std.Dev. Corr
        #Entry.Year (Intercept) 0.52944  0.7276
        #Year       (Intercept) 0.08703  0.2950
        #           Block       0.04098  0.2024   -1.00
        #Year.1     (Intercept) 0.00000  0.0000
        #Residual               0.39204  0.6261
        #Number of obs: 450, groups:  Entry:Year, 150; Year, 2
  # ANOVA
  anova(LDG_lm)
          #Analysis of Variance Table
          #      Df Sum Sq Mean Sq F value
          #Entry 74 167.73  2.2667  5.7818
  # LSMean
  LDG_rg = ref.grid(LDG_lm)   # establish the reference grid for LSMEANS.
  summary(LDG_rg)
  lsm_LDG = lsmeans(LDG_rg, "Entry")
  sum_lsm_LDG = summary(lsm_LDG)      # convert lsmean into dataframe
  sum_lsm_LDG = sum_lsm_LDG[order(sum_lsm_LDG$lsmean, decreasing = T),,drop=F]   # sort data from
the largest to the smallest
    # Plot by Entry
  plot(sum_lsm_LDG, by = "Entry", main="Lodging score")
    # LSD
  LSD_LDG = predictmeans(LDG_lm,"Entry")$LSD
        #[1] 1.64

###############################################################################

# R8
  # Linear Model
    R8_lm = lmer(R8 ~ 1 + Entry + (1|Year) + (1|Entry:Year) + (Block |Year), data=Quan,REML=T)
  summary(R8_lm)
        #Random effects:
        #Groups     Name        Variance Std.Dev. Corr
        #Entry.Year (Intercept)   8.7547  2.9588
        #Year       (Intercept) 117.8450 10.8556
        #           Block         0.1321  0.3635  -1.00
        #Year.1     (Intercept)   0.0000  0.0000
        #Residual                11.3484  3.3687
        #Number of obs: 450, groups:  Entry:Year, 150; Year, 2
  # ANOVA
  anova(R8_lm)
          #Analysis of Variance Table
          #      Df Sum Sq Mean Sq F value
          #Entry 74 5294.1  71.542  6.3041
  # LSMean
  R8_rg = ref.grid(R8_lm)   # establish the reference grid for LSMEANS.
  summary(R8_rg)
  lsm_R8 = lsmeans(R8_rg, "Entry")
  sum_lsm_R8 = summary(lsm_R8)      # convert lsmean into dataframe
  sum_lsm_R8 = sum_lsm_R8[order(sum_lsm_R8$lsmean, decreasing = T),,drop=F]   # sort data from
the largest to the smallest
    # Plot by Entry
  plot(sum_lsm_Protein, by = "Entry", main="Maturity")
    # LSD
  LSD_R8 = predictmeans(R8_lm,"Entry")$LSD
        #[1] 7.15

################################################################################
# LSMEANS for all traits
LS_table = Reduce(function(x, y) merge(x, y, all=TRUE, by = "Entry"),
list(sum_lsm_Protein[,1:2],sum_lsm_Oil[,1:2],
                sum_lsm_HGT[,1:2],sum_lsm_LDG[,1:2],sum_lsm_R8[,1:2]))
colnames(LS_table)= c("Entry","Protein_LS","Oil_LS", "Height_LS","Lodging_LS","Maturity_LS")

###############################################################################
#Histogram
par(mfrow=c(2,3), bg="transparent",mai=c(0.45,0.7,0.45,0.45))
hist(sum_lsm_Protein$lsmean*10, main="Protein concentration (g/kg DB)
     Dwight: 389; 2n=42: 416", label=TRUE,
     col =
c("grey","blue","grey","grey","red","grey","grey","grey","grey","grey","grey","grey"),breaks=10)
hist(sum_lsm_Oil$lsmean*10, main="Oil concentration (g/kg DB)
     Dwight: 216; 2n=42: 192", label=TRUE,
     col =
c("grey","grey","grey","grey","grey","red","grey","grey","grey","Blue","grey"),breaks=8)
hist(sum_lsm_HGT$lsmean, main="Height (cm)
     Dwight: 85; 2n=42: 58", label=TRUE,
     col =
c("red","grey","grey","blue","grey","grey","grey","grey","grey","grey","grey","grey"),breaks=8)
hist(sum_lsm_LDG$lsmean, main="Lodging Score *
     Dwight: 2.8; 2n=42: 1.9", label=TRUE,
     col = c("red","blue","grey","grey","grey","grey","grey"),breaks=6)
hist(sum_lsm_R8$lsmean, main="Maturity (days after July 1st)
     Dwight: 114; 2n=42: 130", label=TRUE,
     col = c("grey","blue","grey","grey","grey","red","grey"),breaks=6)
plot.new()

###############################################################################

# Multivariate
  # MANOVA
Y2 = cbind(Protein_DB,Oil_DB,HGT,LDG,R8)

MANOVA2 = manova( Y2 ~ Entry,Quanti_table)
summary(MANOVA2)

        #           Df Pillai approx F num Df den Df    Pr(>F)
        #Entry      74 2.6019   6.3048    370   2150 < 2.2e-16 ***
        #Residuals 430
        #---
        #Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  # Correlation
plot(Quanti_table[,5:9],main="Figure 3.4: Correlation Matrix of Quantitative Traits",cex=1.4)
Cor_40 = cor(na.omit(Quanti_table[,5:9]), method="pearson")
###############################################################################
```

Appendix B

Unix and Perl command used in RNA-sequencing data analysis

```
# Making gene reference using Useq
# create fasta file for each chromosome and scaffolds
# do this on seqs folder where you will execute USeq
split_fasta_by_seq.pl /home/swang130/scratch/phytozome_v9.1/Gmax_189/assembly/Gmax_189.fa

# run USeq
module load java
java -jar -Xmx4G ~/bin/MakeTranscriptome -f ./seqs -u Gmax_189_gene.refFlat  -r 96 -n 60000
# -r: read length - 4 bp
# -n: create max 60k combination

# install trinity
wget -O trinityrnaseq_r20140413p1.tar.gz
http://sourceforge.net/projects/trinityrnaseq/files/trinityrnaseq_r20140413p1.tar.gz/download
cd trinityrnaseq_r20140413p1
make

# install bowtie
wget -O bowtie-1.0.1-linux-x86_64.zip http://sourceforge.net/projects/bowtie-
bio/files/bowtie/1.0.1/bowtie-1.0.1-linux-x86_64.zip/download
unzip bowtie-1.0.1-linux-x86_64.zip

# install samtools
wget -O  samtools-0.1.19.tar.bz2
http://sourceforge.net/projects/samtools/files/samtools/0.1.19/samtools-0.1.19.tar.bz2/download
tar xvjf samtools-0.1.19.tar.bz2
cd samtools-0.1.19
make

# Create novoalign index file
 novoindex -k 14 -s 1 -t 12 Gmax_189_genome_transcript_splices.nix Gmax_189_genome_masked.fa
Gmax_189_geneRad96Num60kMin10Transcripts.fasta Gmax_189_geneRad96Num60kMin10Splices.fasta

# OUTPUT:
# novoindex (3.2) - Universal k-mer index constructor.
# # (C) 2008 - 2011 NovoCraft Technologies Sdn Bhd
# # novoindex -k 14 -s 1 -t 12 Gmax_189_genome_transcript_splices.nix
# Gmax_189_genome_masked.fa Gmax_189_geneRad96Num60kMin10Transcripts.fasta
# Gmax_189_geneRad96Num60kMin10Splices.fasta
# # Creating 12 indexing threads.
# Warning: Adjusting s to 2 due to large reference sequence.
# tcmalloc: large alloc 1073750016 bytes == 0x2b06000 @  0x503a0c 0x4038d0
# 0x4006aa 0x491630 0x400c99
# tcmalloc: large alloc 11284127744 bytes == 0x42c44000 @  0x504948 0x403aa5
# 0x4006aa 0x491630 0x400c99
# # novoindex construction dT = 170.5s
# # Index memory size  10.509Gbyte.
# # Done.

# Alignment
       # LG13-7552-1
/home/swang130/bin/novocraft/novoalign -d
~/scratch/data/reference/Gmax_189_genome_transcript_splices.nix -F ILM1.8 -c 12 -i PE 250,50 -o
SAM -r Random -f LG13-7552-1_ATCACG_L001_R1_001.fastq LG13-7552-1_ATCACG_L001_R2_001.fastq >
LG13-7552-1_ATCACG_L001_R12_001.fastq.sam
# Convert splice junctions coordinates back to genome coordinates
java -jar ~/bin/SamTranscriptomeParser -f LG13-7552-1_ATCACG_L001_R12_001.fastq.sam -a 50000 -n
100 -u -s LG13-7552-1_ATCACG_L001_R12_001.fastq.Converted.sam
# Sort and convert to bam
~/samtools view -uS LG13-7552-1_ATCACG_L001_R12_001.fastq.Converted.sam 2> LG13-7552-1.err |
~/novosort -o LG13-7552-1_ATCACG_L001_R12_001.fastq.Converted.Sorted.bam -i LG13-7552-
1_ATCACG_L001_R12_001.fastq.Converted.Sorted.bam.bai -c 12 -
```

```
        # LG13-7552-4
/home/swang130/bin/novocraft/novoalign -d
~/scratch/data/reference/Gmax_189_genome_transcript_splices.nix -F ILM1.8 -c 12 -i PE 250,50 -o
SAM -r Random -f LG13-7552-4_CGATGT_L001_R1_001.fastq LG13-7552-4_CGATGT_L001_R2_001.fastq >
LG13-7552-4_CGATGT_L001_R12_001.fastq.sam
java -jar ~/bin/SamTranscriptomeParser -f LG13-7552-4_CGATGT_L001_R12_001.fastq.sam -a 50000 -n
100 -u -s LG13-7552-4_CGATGT_L001_R12_001.fastq.Converted.sam
~/samtools view -uS LG13-7552-4_CGATGT_L001_R12_001.fastq.Converted.sam 2> LG13-7552-4.err |
~/novosort -o LG13-7552-4_CGATGT_L001_R12_001.fastq.Converted.Sorted.bam -i LG13-7552-
4_CGATGT_L001_R12_001.fastq.Converted.Sorted.bam.bai -c 12 -

        # LG13-7552-5
/home/swang130/bin/novocraft/novoalign -d
~/scratch/data/reference/Gmax_189_genome_transcript_splices.nix -F ILM1.8 -c 12 -i PE 250,50 -o
SAM -r Random -f LG13-7552-5_TTAGGC_L001_R1_001.fastq LG13-7552-5_TTAGGC_L001_R2_001.fastq >
LG13-7552-5_TTAGGC_L001_R12_001.fastq.sam
java -jar ~/bin/SamTranscriptomeParser -f LG13-7552-5_TTAGGC_L001_R12_001.fastq.sam -a 50000 -n
100 -u -s LG13-7552-5_TTAGGC_L001_R12_001.fastq.Converted.sam
~/samtools view -uS LG13-7552-5_TTAGGC_L001_R12_001.fastq.Converted.sam 2> LG13-7552-5.err |
~/novosort -o LG13-7552-5_TTAGGC_L001_R12_001.fastq.Converted.Sorted.bam -i LG13-7552-
5_TTAGGC_L001_R12_001.fastq.Converted.Sorted.bam.bai -c 12 -

        # LG13-20153-1
/home/swang130/bin/novocraft/novoalign -d
~/scratch/data/reference/Gmax_189_genome_transcript_splices.nix -F ILM1.8 -c 12 -i PE 250,50 -o
SAM -r Random -f LG13-20153-1_TGACCA_L001_R1_001.fastq LG13-20153-1_TGACCA_L001_R2_001.fastq >
LG13-20153-1_TGACCA_L001_R12_001.fastq.sam
# Convert splice junctions coordinates back to genome coordinates
java -jar ~/bin/SamTranscriptomeParser -f LG13-20153-1_TGACCA_L001_R12_001.fastq.sam -a 50000 -n
100 -u -s LG13-20153-1_TGACCA_L001_R12_001.fastq.Converted.sam
# Sort and convert to bam
~/samtools view -uS LG13-20153-1_TGACCA_L001_R12_001.fastq.Converted.sam 2> LG13-20153-1.err |
~/novosort -o LG13-20153-1_TGACCA_L001_R12_001.fastq.Converted.Sorted.bam -i LG13-20153-
1_TGACCA_L001_R12_001.fastq.Converted.Sorted.bam.bai -c 12 -

        # LG13-20153-2
/home/swang130/bin/novocraft/novoalign -d
~/scratch/data/reference/Gmax_189_genome_transcript_splices.nix -F ILM1.8 -c 12 -i PE 250,50 -o
SAM -r Random -f LG13-20153-2_ACAGTG_L001_R1_001.fastq LG13-20153-2_ACAGTG_L001_R2_001.fastq >
LG13-20153-2_ACAGTG_L001_R12_001.fastq.sam
java -jar ~/bin/SamTranscriptomeParser -f LG13-20153-2_ACAGTG_L001_R12_001.fastq.sam -a 50000 -n
100 -u -s LG13-20153-2_ACAGTG_L001_R12_001.fastq.Converted.sam
~/samtools view -uS LG13-20153-2_ACAGTG_L001_R12_001.fastq.Converted.sam 2> LG13-20153-2.err |
~/novosort -o LG13-20153-2_ACAGTG_L001_R12_001.fastq.Converted.Sorted.bam -i LG13-20153-
2_ACAGTG_L001_R12_001.fastq.Converted.Sorted.bam.bai -c 12 -

        # LG13-20153-3
/home/swang130/bin/novocraft/novoalign -d
~/scratch/data/reference/Gmax_189_genome_transcript_splices.nix -F ILM1.8 -c 12 -i PE 250,50 -o
SAM -r Random -f LG13-20153-3_GCCAAT_L001_R1_001.fastq LG13-20153-3_GCCAAT_L001_R2_001.fastq >
LG13-20153-3_GCCAAT_L001_R12_001.fastq.sam
java -jar ~/bin/SamTranscriptomeParser -f LG13-20153-3_GCCAAT_L001_R12_001.fastq.sam -a 50000 -n
100 -u -s LG13-20153-3_GCCAAT_L001_R12_001.fastq.Converted.sam
~/samtools view -uS LG13-20153-3_GCCAAT_L001_R12_001.fastq.Converted.sam 2> LG13-20153-3.err |
~/novosort -o LG13-20153-3_GCCAAT_L001_R12_001.fastq.Converted.Sorted.bam -i LG13-20153-
3_GCCAAT_L001_R12_001.fastq.Converted.Sorted.bam.bai -c 12 -

# install HTSeq for counting reads mapped to genes
module load python/2.7.3

# download package
wget https://pypi.python.org/packages/source/H/HTSeq/HTSeq-
0.6.1p1.tar.gz#md5=c44d7b256281a8a53b6fe5beaeddd31c
tar xvzf HTSeq-0.6.1p1.tar.gz
cd HTSeq-0.6.1p1

# install to ~/.local
python setup.py install --user
python ~/bin/HTSeq-0.6.1p1/scripts/htseq-count
```

```
# install pysam locally per requirement of htseq-count
pip install --install-option="--prefix=$HOME/.local" pysam

# update PYTHONPATH, make sure to do this before running htseq-count
export PYTHONPATH=$HOME/.local/lib/python2.7/site-packages/:$PYTHONPATH

# Abundance estimation
        # LG13-7552-1
python ~/bin/HTSeq-0.6.1p1/scripts/htseq-count -f bam -r pos --stranded=reverse -t gene -i ID
LG13-7552-1_ATCACG_L001_R12_001.fastq.Converted.Sorted.bam
~/scratch/phytozome_v9.1/Gmax_189/annotation/Gmax_189_gene_exons.gff3 > LG13-7552-1_gene.count

        # LG13-7552-4
python ~/bin/HTSeq-0.6.1p1/scripts/htseq-count -f bam -r pos --stranded=reverse -t gene -i ID
LG13-7552-4_CGATGT_L001_R12_001.fastq.Converted.Sorted.bam
~/scratch/phytozome_v9.1/Gmax_189/annotation/Gmax_189_gene_exons.gff3 > LG13-7552-4_gene.count

        # LG13-7552-5
python ~/bin/HTSeq-0.6.1p1/scripts/htseq-count -f bam -r pos --stranded=reverse -t gene -i ID
LG13-7552-5_TTAGGC_L001_R12_001.fastq.Converted.Sorted.bam
~/scratch/phytozome_v9.1/Gmax_189/annotation/Gmax_189_gene_exons.gff3 > LG13-7552-5_gene.count

        # LG13-20153-1
python ~/bin/HTSeq-0.6.1p1/scripts/htseq-count -f bam -r pos --stranded=reverse -t gene -i ID
LG13-20153-1_TGACCA_L001_R12_001.fastq.Converted.Sorted.bam
~/scratch/phytozome_v9.1/Gmax_189/annotation/Gmax_189_gene_exons.gff3 > LG13-20153-1_gene.count

        # LG13-20153-2
python ~/bin/HTSeq-0.6.1p1/scripts/htseq-count -f bam -r pos --stranded=reverse -t gene -i ID
LG13-20153-2_ACAGTG_L001_R12_001.fastq.Converted.Sorted.bam
~/scratch/phytozome_v9.1/Gmax_189/annotation/Gmax_189_gene_exons.gff3 > LG13-20153-2_gene.count

        # LG13-20153-3
python ~/bin/HTSeq-0.6.1p1/scripts/htseq-count -f bam -r pos --stranded=reverse -t gene -i ID
LG13-20153-3_GCCAAT_L001_R12_001.fastq.Converted.Sorted.bam
~/scratch/phytozome_v9.1/Gmax_189/annotation/Gmax_189_gene_exons.gff3 > LG13-20153-3_gene.count


###########################################################################
# Trinity assembly of LG13-20153 & LG13-7552
# Reads: R1 --> map to antisense, R2 --> map to sense strand
# Merged assembly of LG13-20153 and LG13-7552
# Trying to create trinity assembly by merging all the reads from LG13-20153
# and LG13-7552 into single assembly. Look it up in folder merged_RNAseq

# Load java for USeq
module load java
#
# Set stack size unlimited for Chrysalis
ulimit -s unlimited

# create named pipe for LG13-7552
# no need for LG13-20153 since we concatenate all reads
cat /home/swang130/scratch/data/LG13-7552-1_ATCACG_L001_R1_001.fastq
/home/swang130/scratch/data/LG13-7552-4_CGATGT_L001_R1_001.fastq
/home/swang130/scratch/data/LG13-7552-5_TTAGGC_L001_R1_001.fastq >
/home/swang130/scratch/trinity/LG13-7552/LG13-7552_R1_all.fastq

cat /home/swang130/scratch/data/LG13-7552-1_ATCACG_L001_R2_001.fastq
/home/swang130/scratch/data/LG13-7552-4_CGATGT_L001_R2_001.fastq
/home/swang130/scratch/data/LG13-7552-5_TTAGGC_L001_R2_001.fastq >
/home/swang130/scratch/trinity/LG13-7552/LG13-7552_R2_all.fastq

# run Trinity
/home/swang130/bin/trinityrnaseq_r20140413p1/Trinity --seqType fq --JM 80G --left
/home/swang130/scratch/trinity/LG13-7552/LG13-7552_R1_all.fastq
/home/swang130/scratch/trinity/LG13-20153/LG13-20153_R1_all.fastq --right
/home/swang130/scratch/trinity/LG13-7552/LG13-7552_R2_all.fastq
/home/swang130/scratch/trinity/LG13-20153/LG13-20153_R2_all.fastq --SS_lib_type RF --CPU 12 --
min_contig_length 300 --full_cleanup --bflyHeapSpaceMax 20G --bflyCalculateCPU
```

```
# OUTPUT[swang130@taubh1 merged_RNAseq]$ abyss-fac.pl -H trinity_out_dir.Trinity.fasta
n       n:100   n:N50  min    median mean   N50    max    sum
133872  133872  28424  301    750    1138   1744   13871  152.4e6 trinity_out_di
                                                                 r.Trinity.fasta

# percent converage
#hit_pct_cov_bin      count_in_bin   >bin_below
#100   21648  21648
#90    1707   23355
#80    1684   25039
#70    1830   26869
#60    1985   28854
#50    1805   30659
#40    1771   32430
#30    1946   34376
#20    1992   36368
#10    0      36368
#0     0      36368

# number of contigs: 133872
# BLASTN against Gmax_189 gene model
# output: Trinity_v_Gmax_189.blastn
#
# process using extract_blast_tophitv0.3.pl
# CMD: perl ~/bin/extract_blast_tophitv0.3.pl -i Trinity_v_Gmax_189.blastn -l 300 -t 1
# output: Trinity_v_Gmax_189.blastn.hits --> blastn table
#         Trinity_v_Gmax_189.blastn.list --> list of top hit
# number of contigs w/ hits on soybean gene model (alignment length >= 100):61569
# number of contigs w/o hits on soybean gene model: 72303
#
# create fasta file of contigs w/o hit or not passing filtering criteria
# CMD: create_nohit_fastav0.2.pl trinity_out_dir.Trinity.fasta Trinity_v_Gmax_189.blastn.list
# output: Trinity_v_Gmax_189.blastn.list.nohit.fasta
#
# Take contigs w/o hit against Gmax gene model & do tblastx
# against soybean genome ver. 1.1
# output: Trinity_nohit_v_Gmax_genome.tblastx
# blastn under megablast (default)
tblastx -query Trinity_v_Gmax_189.blastn.list.nohit.fasta -db
/home/swang130/scratch/trinity/blastdb/Gmax_189.fa -out Trinity_nohit_v_Gmax_genome.tblastx -
evalue 1e-10 -num_threads 12 -max_target_seqs 1 -outfmt 6
# process using extract_blast_tophitv0.3.pl
# length cut off 66 aa/198 bp
# CMD : perl ~/bin/extract_blast_tophitv0.3.pl -i Trinity_nohit_v_Gmax_genome.tblastx -l 66 -t 1
# output: Trinity_nohit_v_Gmax_genome.tblastx.hits
#         Trinity_nohit_v_Gmax_genome.tblastx.list
#
# create fasta file of contigs w/ hit against soybean genome using tblastx &
# min length of 66 aa
# CMD: create_fasta_from_list.pl Trinity_v_Gmax_189.blastn.list.nohit.fasta
Trinity_nohit_v_Gmax_genome.tblastx.list
# output : Trinity_nohit_v_Gmax_genome.tblastx.list.fasta
#
# create fasta file of contigs w/o hit or not passing filtering criteria
# CMD: create_nohit_fastav0.2.pl Trinity_v_Gmax_189.blastn.list.nohit.fasta
Trinity_nohit_v_Gmax_genome.tblastx.list
# output: Trinity_v_Gmax_189.blastn.list.nohit.fasta
#
# Count expression for each libraries
# merge all RSEM count table into single table for DESeq analysis
# per gene counts
perl ~/bin/trinityrnaseq_r20140413p1/util/abundance_estimates_to_matrix.pl --est_method RSEM --
name_sample_by_basedir --cross_sample_fpkm_norm none --out_prefix mergedRNASeq_genes LG13-20153-
1/RSEM.genes.results LG13-20153-2/RSEM.genes.results LG13-20153-3/RSEM.genes.results LG13-7552-
1/RSEM.genes.results LG13-7552-4/RSEM.genes.results LG13-7552-5/RSEM.genes.results

# annotate interval data from Gopal with assembled contigs
# CMD: perl genome_interval_annotation.pl -t dw_tom_overlapping_regions -i
test,contigs_genome_annotation

# add contig annotation (soybean gene model) to DESeq significance genes
# perl add_annotation -i <gene id table> -o <output file name> -g <DEG# list of significant genes>
perl add_annotation.pl -i contigs_genemodel_annotation -o resSig_Anno -g resSig

# using EdgeR for counting DEG
# run abundance_estimates_to_matrix.pl
# by genes
/home/swang130/bin/trinityrnaseq_r20140413p1/util/abundance_estimates_to_matrix.pl --est_method
RSEM --name_sample_by_basedir --cross_sample_fpkm_norm none
/home/swang130/scratch/trinity/merged_RNAseq/LG13-20153-1/RSEM.genes.results
/home/swang130/scratch/trinity/merged_RNAseq/LG13-20153-2/RSEM.genes.results
/home/swang130/scratch/trinity/merged_RNAseq/LG13-20153-3/RSEM.genes.results
/home/swang130/scratch/trinity/merged_RNAseq/LG13-7552-1/RSEM.genes.results
/home/swang130/scratch/trinity/merged_RNAseq/LG13-7552-4/RSEM.genes.results
/home/swang130/scratch/trinity/merged_RNAseq/LG13-7552-5/RSEM.genes.results

# Blast nohit to soybean genome to NCBI
# blastn under megablast (default)
tblastx -query Trinity_v_Gmax_189.blastn.list.nohit.fasta -db nr -remote -out
Trinity_Gmax_genome_nohit_v_ncbi.tblastx -evalue 1e-10 -max_target_seqs 1 -outfmt 6
```

Appendix C

R code used for differential gene expression

```
###############################################################################
#                                                                             #
#                                       Alignment                             #
#                                                                             #
###############################################################################

# install bioconductor + DESeq
source("http://bioconductor.org/biocLite.R")
biocLite()
biocLite("DESeq")

getwd()
setwd("C:/Users/swang130/Desktop/Research/RNAseq/Gene Count")

# import gene count table
lg13_20153.1 <- read.table("LG13-20153-1_gene.count", header=F, row.names=1)
lg13_20153.2 <- read.table("LG13-20153-2_gene.count", header=F, row.names=1)
lg13_20153.3 <- read.table("LG13-20153-3_gene.count", header=F, row.names=1)
lg13_7552.1 <- read.table("LG13-7552-1_gene.count", header=F, row.names=1)
lg13_7552.4 <- read.table("LG13-7552-4_gene.count", header=F, row.names=1)
lg13_7552.5 <- read.table("LG13-7552-5_gene.count", header=F, row.names=1)

# merge table
GeneCount= cbind(lg13_20153.1, lg13_20153.2, lg13_20153.3, lg13_7552.1, lg13_7552.4, lg13_7552.5)
GeneCount= GeneCount[1:(nrow(GeneCount)-5),]


# colnames
colnames(GeneCount) <- c("40_1", "40_2", "40_3", "42_1", "42_2", "42_3")

# set up metadata of gene counts
GeneCountDesign=data.frame (
  row.names = colnames(GeneCount),
  condition = c("40", "40", "40", "42", "42", "42"),
  libType   = c("paired-end", "paired-end","paired-end","paired-end","paired-end","paired-end")
  )

# load DESeq
library("DESeq")

###############################################################################
#                     Select Differentially expressed genes                   #
###############################################################################
# initiate DESeq data structure
cds = newCountDataSet(GeneCount, GeneCountDesign$condition)

# Normalisation
cds = estimateSizeFactors(cds)
sizeFactors(cds)
    # head( counts(cds, normalized=T))

# variance estimation
cds = estimateDispersions(cds)
str(fitInfo(cds))
plotDispEsts(cds)

# Calling differential expression
res=nbinomTest(cds, "42", "40")
res = na.omit(res)

  #head(res)
```

```
plotMA(res, col=ifelse(res$padj>=0.05, "gray32", "blue"))

# filter at FDR=0.05
resSig = res[res$padj < 0.05,]
plotMA(resSig, col=ifelse(abs(resSig$log2FoldChange)>=2, "gray32", "blue"))

# remove NA
resSig = na.omit(resSig)
resSig = subset(resSig,resSig$log2FoldChange!=Inf & resSig$log2FoldChange!=-Inf)
write.csv(resSig, file="DEGs_gene_model")

# filter |log2foldChange| > 2
resSig = resSig[abs(resSig$log2FoldChange)>=2,]
write.csv(resSig, file = "up in 40 genes.csv")

resSig = resSig[resSig$log2FoldChange>0,]

# save table as
write.table(res,file="DESeq table.csv")
write.csv(resSig, file = "Significant genes.csv")

# save image and history
loadhistory(file="DEG_42_RNASeq.Rhistory")
save.image(file="DEG_42_RNASeq.RData")

################################################################################
#                                Creat Heatmap                                 #
################################################################################

# creat dataset with multiple factors
cdsFull = newCountDataSet(GeneCount, GeneCountDesign)
#cdsFull = newCountDataSet(GeneCountMod, GeneCountDesignMod)

# estimate the size factors and dispersions
cdsFull = na.omit(cdsFull)
cdsFull = estimateSizeFactors( cdsFull )
cdsFull = estimateDispersions( cdsFull )


# heatmap of the count table
cdsFullBlind = estimateDispersions ( cdsFull, method = "blind")
vsdFull = varianceStabilizingTransformation(cdsFullBlind)

library("RColorBrewer")
library("gplots")

select = order(rowMeans(counts(cdsFull)), decreasing = TRUE) [1:200]   # TOP200 highest expressed
genes
hmcol=colorRampPalette(brewer.pal(9, "GnBu"))(100)
heatmap.2(exprs(vsdFull)[select,], col = hmcol, trace="none", margin=c(10,6))
heatmap.2(counts(cdsFull)[select,], col = hmcol, trace="none", margin=c(10,6))

################################################################################
#                                PCA                                           #
################################################################################

print (plotPCA(vsdFull), intgroup = c("condition", "libType"))

################################################################################
#                                                                              #
#                                     Trinity                                  #
#                                                                              #
################################################################################
# install bioconductor + DESeq
source("http://bioconductor.org/biocLite.R")
biocLite()
biocLite("DESeq")

getwd()
setwd("C:/Users/swang130/Desktop/Research/RNAseq/mergedRNAseq Count")
```

```
# save image and history
savehistory(file="DEG_42_RNASeq_merged.Rhistory")
loadhistory(file="DEG_42_RNASeq_merged.Rhistory")
load(file="DEG_42_RNASeq_merged.RData")

# import gene count table
count_matrix = read.table("C:/Users/swang130/Desktop/Research/RNAseq/mergedRNAseq
Count/mergedRNASeq_genes.counts.matrix", header=T, row.names=1)

count_matrix[,1] = as.integer(count_matrix[,1])
count_matrix[,2] = as.integer(count_matrix[,2])
count_matrix[,3] = as.integer(count_matrix[,3])
count_matrix[,4] = as.integer(count_matrix[,4])
count_matrix[,5] = as.integer(count_matrix[,5])
count_matrix[,6] = as.integer(count_matrix[,6])

# set up metadata of gene counts
GeneCountDesign=data.frame (
  row.names = colnames(count_matrix),
  condition = c("40", "40", "40", "42", "42", "42"),
  libType   = c("paired-end", "paired-end","paired-end","paired-end","paired-end","paired-end")
  )

# load DESeq
library("DESeq")

###############################################################################
#                     Select Differentially expressed genes                   #
###############################################################################
# initiate DESeq data structure
cds = newCountDataSet(count_matrix, GeneCountDesign$condition)

# Normalisation
cds = estimateSizeFactors(cds)
sizeFactors(cds)
    # head( counts(cds, normalized=T))

# variance estimation
cds = estimateDispersions(cds)
str(fitInfo(cds))
plotDispEsts(cds)

# Calling differential expression
res=nbinomTest(cds, "42", "40")
res = na.omit(res)
res[is.na(res$padj),]     #NA list: has 0 counts on all the columns

  #head(res)
plotMA(res, col=ifelse(res$padj>=0.05, "gray32", "blue"))

# filter at FDR=0.05
resSig = res[res$padj < 0.05,]

plotMA(resSig, col=ifelse(abs(resSig$log2FoldChange)>=2,"red",  "gray32"))

# Seperate NA and Infs
resSig = na.omit(resSig)
Infs = subset(res,res$log2FoldChange == Inf | res$log2FoldChange ==-Inf)
Sig_Infs = subset(resSig,resSig$log2FoldChange == Inf | resSig$log2FoldChange ==-Inf)
resSig = subset(resSig,resSig$log2FoldChange!=Inf & resSig$log2FoldChange!=-Inf)
upin40 = resSig[resSig$log2FoldChange > 0, ]
upin42 = resSig[resSig$log2FoldChange < 0, ]

    write.table(resSig, file = "resSig")
    write.table(Infs, file = "Infs")
    write.table(Sig_Infs, file = "Sig_Infs")


#########################################################################
# Contigs are blasted against gene model and added annotation to in taub#
#########################################################################
```

```
# add gene_id to contigs
resSig_Anno = read.table("C:/Users/swang130/Desktop/Research/RNAseq/mergedRNAseq
Count/resSig_Anno", header=T, row.names=NULL)
resSig_Anno = unique(resSig_Anno)

upin40_Anno = resSig_Anno[resSig_Anno$log2FoldChange > 0, ]
upin42_Anno = resSig_Anno[resSig_Anno$log2FoldChange < 0, ]

# filter |log2foldChange| > 2
#resSig = resSig[abs(resSig$log2FoldChange)>=2,]
#write.csv(resSig, file = "up in 40 genes.csv")

#resSig = resSig[resSig$log2FoldChange>0,]

# save table as
#write.table(res,file="DESeq table.csv")
write.csv(resSig_Anno, file = "DEGs_trinity")

#########################################################################
#                      Infinites and -Infinites                         #
#########################################################################

Tom = read.table("C:/Users/swang130/Desktop/Research/RNAseq/mergedRNAseq
Count/Trinity_Gmax_genome_nohit_v_ncbi.tblastx_annotated_as_tomentella.xls",sep="\t")
Tom1 = gsub("_i\\d+","",Tom[,1],perl=TRUE)
Tom = cbind(Tom1,Tom[,2:14])
unique(Tom[,1])

# number and list of contigs that map to tomentella sequence
Con_Tom = res[res$id %in% Tom$Tom1,]
nrow( res[res$id %in% Tom$Tom1,])
                      # 174
# number of Significant contigs
Sig_Con_Tom = Con_Tom[Con_Tom$padj < 0.05, ]

# count table of contigs that map to tomentella sequence
count_matrix[row.names(count_matrix) %in% Tom1,]
                      # 175

#########################################################################
#                       Repeat Master Output                            #
#########################################################################

Repe = read.table("C:/Users/swang130/Desktop/Research/RNAseq/mergedRNAseq
Count/trinity_out_dir.Trinity.fasta.out", strip.white=T, header = T)
          #dim(Repe)
          #[1] 79770    15
table(Repe$class.family)

          # ARTEFACT                 DNA       DNA/CMC-EnSpm             DNA/hAT
DNA/hAT-Ac        DNA/hAT-Tag1
          #        1                161                1918                   4
739                 729
          #DNA/hAT-Tip100       DNA/MULE-MuDR      DNA/MULE-MuDR?   DNA/PIF-Harbinger
DNA/TcMar-Pogo  DNA/TcMar-Stowaway
          #           111              2379                   3                 595
1                  49
          #3DNA/TcMar-Stowaway?                DNA?             LINE/L1       LINE/RTE-BovB
Low_complexity                 LTR
          #                   1                  13                3386               1535
6934                 152
          #LTR/Caulimovirus           LTR/Copia          LTR/Copia?           LTR/Gypsy
LTR?     Other/Composite
          #             159                7374                  22                5475
23                   9
          #RC/Helitron       RC?/Helitron?                rRNA           Satellite
Simple_repeat                SINE
          #        880                  36                  62                  58
46700                  12
```

```
          #SINE/tRNA       SINE/tRNA-RTE          SINE/tRNA?           SINE?/RTE
Unknown
          #     3191                   9                 18                    3
28

Repeseq = gsub("_i\\d+","",Repe$sequence, perl=TRUE)
Repe = cbind(Repe[,1:4],Repeseq,Repe[,6:15])

# trim simple repeats and low complexity
Repe = subset(Repe, !(Repe$class.family %in% "Simple_repeat"))
          #dim(Repe)
          # 33070    15
Repe = subset(Repe, !(Repe$class.family %in% "Low_complexity"))
          #dim(Repe)
          # 26136    15
# common contigs in repeatmaster and map to tomentella genes
intersect(Repe$Repeseq, Tom$Tom1)

# Significant repeat master contigs
Con_repe = res[res$id %in% Repe$Repeseq,]
Sig_Con_repe = Con_repe[Con_repe$padj < 0.05, ]
          # 1077    8
up40_Con_repe = Sig_Con_repe[Sig_Con_repe$log2FoldChange > 0, ]
          # 336 (242)

Inf40_Con_repe = up40_Con_repe[up40_Con_repe$log2FoldChange=="Inf",]
          #94

up42_Con_repe = Sig_Con_repe[Sig_Con_repe$log2FoldChange < 0, ]
          #741 (399)

Inf42_Con_repe = up42_Con_repe[up42_Con_repe$log2FoldChange=="-Inf",]
          #342

length(intersect(Sig_Con_repe$id, Tom$Tom1))
           # 18

###############################################################################
#                                Creat Heatmap                                #
###############################################################################

# creat dataset with multiple factors

#GeneCountMod=cbind(lg13_20153.2,lg13_20153.3,lg13_7552.1,lg13_7552.5)
#GeneCountMod= GeneCountMod[1:(nrow(GeneCountMod)-5),]
#colnames(GeneCountMod) <- c("40_1","40_2","40_3", "42_1", "42_3")
#GeneCountDesignMod=data.frame (
  #row.names = colnames(GeneCountMod),
  #condition = c("40", "40","40", "42", "42"),
  #libType   = c("paired-end","paired-end", "paired-end","paired-end","paired-end")
#)

cdsFull = newCountDataSet(count_matrix, GeneCountDesign)
#cdsFull = newCountDataSet(GeneCountMod, GeneCountDesignMod)

# estimate the size factors and dispersions
cdsFull = na.omit(cdsFull)
cdsFull = estimateSizeFactors( cdsFull )
cdsFull = estimateDispersions( cdsFull )


# heatmap of the count table
cdsFullBlind = estimateDispersions ( cdsFull, method = "blind")
vsdFull = varianceStabilizingTransformation(cdsFullBlind)

library("RColorBrewer")
library("gplots")

select = order(rowMeans(counts(cdsFull)), decreasing = TRUE) [1:30]   # TOP30 highest expressed
genes
hmcol=colorRampPalette(brewer.pal(9, "GnBu"))(100)
```

```
heatmap.2(exprs(vsdFull)[select,], col = hmcol, trace="none", margin=c(10,6))

heatmap.2(counts(cdsFull)[select,], col = hmcol, trace="none", margin=c(10,6))

################################################################################
#                                PCA                                           #
################################################################################

print (plotPCA(vsdFull), intgroup = c("condition", "libType"))

#> resSig_Anno[resSig_Anno$geneid=="Glyma06g21920",]
#           id baseMean baseMeanA baseMeanB foldChange log2FoldChange         pval         padj
#112 c23278_g1 1458.547  2379.334   537.761  0.2260133       -2.14552 1.066371e-06 6.140228e-05
#geneid
#112 Glyma06g21920

################################################################################
#                                                                              #
#                                    DEG Comparison                            #
#                                                                              #
################################################################################
setwd("C:/Users/swang130/Desktop/Research/RNAseq/Intersect")

# load 2 sets of DEGS
trinity   = read.csv("C:/Users/swang130/Desktop/Research/RNAseq/mergedRNAseq Count/DEGs_trinity")
genemodel = read.csv("C:/Users/swang130/Desktop/Research/RNAseq/Gene Count/DEGs_gene_model")
colnames(trinity) = 
c("row.names","id","baseMean","baseMeanA","baseMeanB","foldChange","log2FoldChange","pval","padj"
,"geneid")
colnames(genemodel) = 
c("row.names","id","baseMean","baseMeanA","baseMeanB","foldChange","log2FoldChange","pval","padj")

# remove contigs in trinity table where multiple contig clusters were annotated with the same
# Glyma id
temp <-duplicated(trinity$geneid)
trinity_uniq <- trinity[!temp,]

# check for intersection in ven diagram
Intersect = intersect(trinity_uniq$geneid, genemodel$id)
Trinity_uniq_intersect = trinity_uniq[trinity_uniq$geneid %in% Intersect,]
genemodel_intersect = genemodel[genemodel$id %in% Intersect,]

# order based on gene id
M2A <- Trinity_uniq_intersect[order(Trinity_uniq_intersect$geneid),]
M2G <- genemodel_intersect[order(genemodel_intersect$id),]

# install.packages("VennDiagram")
library(VennDiagram)

# Draw a Venn Diagram with 2 sets
    # low quality
draw.pairwise.venn(nrow(trinity_uniq),nrow(genemodel),length(Intersect),category=c("Assembly","Al
ignment"))
    # high quality
venn.plot = venn.diagram(list(Alignment=1:2292,Assembly=1541:3103),"Venn Diagram of genemodel vs 
trinity.tiff",
                         col = "transparent",
                         fill = c("cornflowerblue", "darkorchid1"),
                         cat.col = c("cornflowerblue", "darkorchid1"))

# Compare the expression level of intersect genes at 2 datasets
Intersect_gene_expressions=data.frame("id"=M2A[,2],"M2Ageneid"=M2A[,10],"M2Afoldchange"=M2A[,7],"
M2Ggeneid"=M2G[,2],"M2Gfoldchange"=M2G[,7])
    # Check gene id
    Match=identical(Intersect_gene_expressions$M2Ageneid, Intersect_gene_expressions$M2Ggeneid)
    table(Match)

    # Check if same expression level
    test=c()
    for (i in 1:nrow(Intersect_gene_expressions)){
```

```
   if (Intersect_gene_expressions[i,3]>=0 & Intersect_gene_expressions[i,5]>=0) {
    test[i] = 1
   }
   else if (Intersect_gene_expressions[i,3]<0 & Intersect_gene_expressions[i,5]<0) {
    test[i] = 2
   }
   else test[i] = 0
  }
  table(test)
        #test
        #0  up   Down
        #8 436    308

  # Remove genes with opposite expression level
  as.vector(test)
  Intersect_gene_expressions = data.frame(Intersect_gene_expressions,test)
  Intersect_gene_expressions =
subset(Intersect_gene_expressions,Intersect_gene_expressions$test != 0)

  CV=c()
  for (i in 1:nrow(Intersect_gene_expressions)){
  CV[i] =
sd(c(Intersect_gene_expressions[i,3],Intersect_gene_expressions[i,5]))/mean(c(Intersect_gene_expr
essions[i,3],Intersect_gene_expressions[i,5]))

  }

write.csv(Intersect_gene_expressions,"Intersect_gene_expressions.csv")

savehistory(file="Compare_DEGs.Rhistory")
save.image(file="Compare_DEGs.RData")


# match gene id in trinity against id in gene model
MatchT=match(trinity[,10], genemodel[,2],nomatch=0)
MatchG=match( genemodel[,2], trinity[,10],nomatch=0)

# Creat Hitstogram of genes across chromosomes
par(mfrow=c(1,3), bg="transparent")
 # Gene Model
Chr_DEG_genemodel = substr(genemodel$id,1,7)
Chr_DEG_genemodel = gsub("Glyma","",Chr_DEG_genemodel)
Chr_DEG_genemodel = as.numeric(Chr_DEG_genemodel)
barplot(table(Chr_DEG_genemodel),breaks=20,main="Aligning to Soybean Genome",xlab="Chromosome
Number",ylab="Number of DEGs",
     col=c("white","gray"))

 # trinity
Chr_DEG_trinity = substr(trinity$geneid,1,7)
Chr_DEG_trinity = gsub("Glyma","",Chr_DEG_trinity)
Chr_DEG_trinity = as.numeric(Chr_DEG_trinity)
barplot(table(Chr_DEG_trinity),breaks=20,main="Trinity de novo Assembly",xlab="Chromosome
Number",ylab="Number of DEGs",
     col=c("white","gray"))

 #Intersect
Chr_DEG_Intersect = substr(Intersect_gene_expressions$M2Ageneid,1,7)
Chr_DEG_Intersect = gsub("Glyma","",Chr_DEG_Intersect)
Chr_DEG_Intersect = as.numeric(Chr_DEG_Intersect)
barplot(table(Chr_DEG_Intersect),breaks=20,main="Common DEGs between Alignment
        and de novo Assembly",xlab="Chromosome Number",ylab="Number of DEGs",
     col=c("white","gray"))
```

Appendix D

Functional annotations of assembled contigs from Trinity *de novo* assembly using merged RNA sequencing reads of the DAAL (LG12-7063) and the disomic progeny (LG13-7552) that matched *G. tomentella* sequences. All the annotations were acquired from National Center for Biotechnology Information (NCBI) database using blastx

Contig ID[1]	Annotation from NCBI BLAST[2]	Description[3]	% ID[4]	E-value[5]
c6193_g1_i2	gi|113205396|gb|ABI34377.1|	Polyprotein, putative [Solanum demissum]	41.3	1.E-09
c16714_g1_i1	gi|147845547|emb|CAN78493.1|	hypothetical protein VITISV_037041 [Vitis vinifera]	56.6	6.E-09
c28016_g1_i1	gi|358343207|ref|XP_003635698.1|	hypothetical protein MTR_001s0023 [Medicago truncatula]	46.9	8.E-07
c36149_g2_i1	gi|357140780|ref|XP_003571941.1|	PREDICTED: RNA-directed DNA polymerase from mobile element jockey-like [Brachypodium distachyon]	46	4.E-08
c40414_g1_i1	gi|357445665|ref|XP_003593110.1|	Zinc finger MYM-type protein [Medicago truncatula]	62.8	2.E-09
c40746_g2_i1	gi|147776056|emb|CAN69911.1|	hypothetical protein VITISV_027081 [Vitis vinifera]	55.9	2.E-17
c45590_g2_i1	gi|356544228|ref|XP_003540556.1|	PREDICTED: uncharacterized protein LOC100799395 [Glycine max]	52.3	3.E-07
c46986_g1_i1	gi|357494985|ref|XP_003617781.1|	hypothetical protein MTR_5g095400 [Medicago truncatula]	28.5	7.E-11
c46986_g1_i2	gi|357494991|ref|XP_003617784.1|	hypothetical protein MTR_5g095430 [Medicago truncatula]	29.2	4.E-09
c46986_g1_i4	gi|357494985|ref|XP_003617781.1|	hypothetical protein MTR_5g095400 [Medicago truncatula]	28.5	1.E-10
c54260_g1_i2	gi|356519637|ref|XP_003528477.1|	PREDICTED: glutamate receptor 2.7-like [Glycine max]	52.2	3.E-18
c55555_g4_i1	gi|147772264|emb|CAN71870.1|	hypothetical protein VITISV_044169 [Vitis vinifera]	75.9	2.E-08
c58516_g3_i5	gi|87162498|gb|ABD28293.1|	RNA-directed DNA polymerase (Reverse transcriptase); Zinc finger, CCHC-type; Peptidase aspartic, active site; Retrotransposon gag protein [Medicago truncatula]	80.6	1.E-11
c66267_g1_i1	gi|357498095|ref|XP_003619336.1|	hypothetical protein MTR_6g046770 [Medicago truncatula]	35.5	9.E-08
c68302_g1_i1	gi|396582343|gb|AFN88207.1|	integrase core domain containing protein [Phaseolus vulgaris]	42.8	1.E-24
c71148_g1_i1	gi|113205396|gb|ABI34377.1|	Polyprotein, putative [Solanum demissum]	51.7	3.E-10
c73277_g1_i1	gi|356514878|ref|XP_003526129.1|	PREDICTED: uncharacterized protein LOC100777620 [Glycine max]	32.1	2.E-08
c76286_g1_i1	gi|357503811|ref|XP_003622194.1|	Cellular nucleic acid-binding protein-like protein, partial [Medicago truncatula]	28	1.E-08
c76858_g1_i1	gi|356532924|ref|XP_003535019.1|	PREDICTED: uncharacterized protein LOC100795609 [Glycine max]	33	6.E-07
c78066_g1_i1	gi|77548423|gb|BA91220.1|	retrotransposon protein, putative, unclassified [Oryza sativa Japonica Group]	64.9	5.E-08

[1] Identifier of the assembled contigs from Trinity *de novo* assembly;
[2] Functional annotations of the assembled contigs from NCBI database;
[3] Descriptions of the functional annotations of the assemblyed contigs;
[4] The probability of the functional annotations and the assembled contig sequences were matched by chance;
[5] Percent identity between the functional annotations and the assembled contigs.

Appendix E

Summary of RepeatMasker result

```
==================================================
filenames: trinity_out_dir.Trinity.fasta
sequences:     133872
total length:  152424681      bp      (152424681     bp excl N/X-runs)
GC level:      39.88   %
bases masked:  10000061       bp      (6.56%)
==================================================
        number of elements *          length occupied               perc of sequence
--------------------------------------------------
Retroelements          16105  6266656 bp     4.11    %
        SINEs: 232     22594  bp     0.01   %
        Penelope       0      0      bp     0       %
        LINEs: 4645    2021126 bp    1.33   %
        CRE/SLACS      0      0      bp     0       %
        L2/CR1/Rex     0      0      bp     0       %
        R1/LOA/Jockey  0      0      bp     0       %
        R2/R4/NeSL     0      0      bp     0       %
        RTE/Bov-B      1472   330609 bp     0.22    %
        L1/CIN4 3176   1690705 bp    1.11   %
        LTR elements:  11228  4222936 bp     2.77    %
        BEL/Pao 0      0      bp     0      %
        Ty1/Copia      6322   2189008 bp     1.44    %
        Gypsy/DIRS1    4615   1933928 bp     1.27    %
        Retroviral     0      0      bp     0       %

DNA transposons               6441   1326545 bp     0.87    %
        hobo-Activator 1552   262180 bp     0.17    %
        Tc1-IS630-Pogo 49     4869   bp     0       %
        En-Spm  0      0      bp     0      %
        MuDR-IS905     0      0      bp     0       %
        PiggyBac       0      0      bp     0       %
        Tourist/Harbinger     582    87609  bp     0.06    %
        "Other (Mirage, P-element, Transib)" 0      0      bp     0      %


Rolling-circles               0      0      bp     0       %

Unclassified:          772    268887 bp     0.18    %

Total Interspersed Repeats:                  7862088 bp     5.16    %


Small RNA:             260    39545  bp     0.03    %

Satellites:            57     4042   bp     0       %
Simple repeats:               46464  1783217 bp     1.17    %
Low complexity:               6889   334417  bp     0.22    %
==================================================

*       most    repeats fragmented     by      insertions     or      deletions
        have    been    counted as     one     element


The     query   species was     assumed to      be      eudicotyledons
RepeatMasker    version open-4.0.5     ","     sensitive      mode

run     with    rmblastn       version 2.2.27+
RepBase Update  "20140131,"    RM      database       version 20140131
```

Appendix F

Multiple sequence alignment using Clustal Omega (1.2.1)

```
c8531_g1_i1               GTCTATATTCAGTAAAATGTTGGTGGAGACGTGCCTTGCGTGATGAACTACCATTTCAAA    60
gi|302129056:26-1555      ------------------------------------------------------------    0
Glyma13g04210.1           ------------------------------------------------------------    0


c8531_g1_i1               CACCACACGACACCACACATCCATTTTAAATATAACCCCATCATAGATATCCCGAATCAT    120
gi|302129056:26-1555      ------------------------------------------------------------    0
Glyma13g04210.1           ------------------------------------------------------------    0


c8531_g1_i1               CAAATTATTACTTCATAGCAACTAGCAAATTAATTAGCTTCACCATGGACTCATTGTTAC    180
gi|302129056:26-1555      --------------------------------------------ATGGACTCATTGTTAC    16
Glyma13g04210.1           --------------------------------------------ATGGACTCATTGTTAC    16
                                                                      ****************

c8531_g1_i1               TTCTAAAAGAAATTGCCACTTCCATTTTGATCTTCTTGATCACTCGTCTCTCCATTCAAA    240
gi|302129056:26-1555      TTCTAAAAGAAATTGCCACTTCCATTTTGATCTTCTTGATCACTCGTCTCTCCATTCAAA    76
Glyma13g04210.1           TTCTAAAAGAAATTGCCACTTCCATTTTGATCTTCTTGATCACTCGTCTCTCCATTCAAA    76
                          ************************************************************

c8531_g1_i1               CATTCCTCAAAAGCTATCGCCAGAAACTCCCACCGGGGCCAAAAGGGTGGCCAGTTGTGG    300
gi|302129056:26-1555      CATTCCTCAAAAGCTATCGCCAGAAACTCCCACCGGGGCCAAAAGGGTGGCCAGTTGTGG    136
Glyma13g04210.1           CATTCCTCAAAAGCTATCGCCAGAAACTCCCACCGGGGCCAAAAGGGTGGCCAGTTGTGG    136
                          ************************************************************

c8531_g1_i1               GTGCACTCCCTCTCATGGGAAGCATGCCTCATGTCACCTTAGCAAAGATGGCAAAAAAAT    360
gi|302129056:26-1555      GTGCACTCCCTCTCATGGGAAGCATGCCTCATGTCACCTTAGCAAAGATGGCAAAAAAAT    196
Glyma13g04210.1           GTGCACTCCCTCTCATGGGAAGCATGCCTCATGTCACCTTAGCAAAGATGGCAAAAAAAT    196
                          ************************************************************

c8531_g1_i1               ATGGACCTATAATGTACCTCAAAATGGGCACTAACAACATGGTTGTGGCCTCTACTCCAG    420
gi|302129056:26-1555      ATGGACCTATAATGTACCTCAAAATGGGCACTAACAACATGGTTGTGGCCTCTACTCCAG    256
Glyma13g04210.1           ATGGACCTATAATGTACCTCAAAATGGGCACTAACAACATGGTTGTGGCCTCTACTCCAG    256
                          ************************************************************

c8531_g1_i1               CTGCTGCTCGTGCCTTCCTCAAAACCCTTGATCAAAACTTTTCAAACCGGCCCTCCAATG    480
gi|302129056:26-1555      CTGCTGCTCGTGCCTTCCTCAAAACCCTTGATCAAAACTTTTCAAACCGGCCCTCCAATG    316
Glyma13g04210.1           CTGCTGCTCGTGCCTTCCTCAAAACCCTTGATCAAAACTTTTCAAACCGGCCCTCCAATG    316
                          ************************************************************

c8531_g1_i1               CTGGTGCAACCCATTTGGCTTATGATGCACGGGATATGGTGTTTGCTCATTACGGATCAC    540
gi|302129056:26-1555      CTGGTGCAACCCATTTGGCTTATGATGCACGGGATATGGTGTTTGCTCATTACGGATCAC    376
Glyma13g04210.1           CTGGTGCAACCCATTTGGCTTATGATGCACGGGATATGGTGTTTGCTCATTACGGATCAC    376
                          ************************************************************

c8531_g1_i1               GGTGGAAGTTGCTAAGAAAACTAAGTAACTTGCACATGCTTGGAGGAAAGGCACTTGATG    600
gi|302129056:26-1555      GGTGGAAGTTGCTAAGAAAACTAAGTAACTTGCACATGCTTGGAGGAAAGGCACTTGATG    436
Glyma13g04210.1           GGTGGAAGTTGCTAAGAAAACTAAGTAACTTGCACATGCTTGGAGGAAAGGCACTTGATG    436
                          ************************************************************

c8531_g1_i1               ATTGGGCCCAAATTCGAGATGAAGAGATGGGGCACATGCTTGGTGCAATGTACGATTGTA    660
gi|302129056:26-1555      ATTGGGCCCAAATTCGAGATGAAGAGATGGGGCACATGCTTGGTGCAATGTACGATTGTA    496
Glyma13g04210.1           ATTGGGCCCAAATTCGAGATGAAGAGATGGGGCACATGCTTGGTGCAATGTACGATTGTA    496
                          ************************************************************

c8531_g1_i1               ACAAGAGGGATGAGGCTGTGGTGGTGGCGGAGATGTTGACATATTCAATGGCCAACATGA    720
gi|302129056:26-1555      ACAAGAGGGATGAGGCTGTGGTGGTGGCGGAGATGTTGACATATTCAATGGCCAACATGA    556
Glyma13g04210.1           ACAAGAGGGATGAGGCTGTGGTGGTGGCGGAGATGTTGACATATTCAATGGCCAACATGA    556
                          ************************************************************

c8531_g1_i1               TTGGCCAAGTTATATTGAGTCGTCGAGTGTTTGAGACAAAGGGTTCGGAGTCTAACGAGT    780
```

```
gi|302129056:26-1555      TTGGCCAAGTTATATTGAGTCGTCGAGTGTTTGAGACAAAGGGTTCGGAGTCTAACGAGT    616
Glyma13g04210.1           TTGGCCAAGTTATATTGAGTCGTCGAGTGTTTGAGACAAAGGGTTCGGAGTCTAACGAGT    616
                          ************************************************************

c8531_g1_i1               TCAAGGACATGGTGGTTGAGCTCATGACCGTTGCTGGTTACTTCAACATTGGTGACTTCA    840
gi|302129056:26-1555      TCAAGGACATGGTGGTTGAGCTCATGACCGTTGCTGGTTACTTCAACATTGGTGACTTCA    676
Glyma13g04210.1           TCAAGGACATGGTGGTTGAGCTCATGACCGTTGCTGGTTACTTCAACATTGGTGACTTCA    676
                          ************************************************************

c8531_g1_i1               TACCCTTTTTGGCCAAGTTGGACTTGCAAGGCATAGAGCGTGGCATGAAGAAGTTGCACA    900
gi|302129056:26-1555      TACCCTTTTTGGCCAAGTTGGACTTGCAAGGCATAGAGCGTGGCATGAAGAAGTTGCACA    736
Glyma13g04210.1           TACCCTTTTTGGCCAAGTTGGACTTGCAAGGCATAGAGCGTGGCATGAAGAAGTTGCACA    736
                          ************************************************************

c8531_g1_i1               AGAAGTTTGATGCGTTGTTAACGAGCATGATTGAGGAGCATGTTGCTTCTAGTCACAAGA    960
gi|302129056:26-1555      AGAAGTTTGATGCGTTGTTAACGAGCATGATTGAGGAGCATGTTGCTTCTAGTCACAAGA    796
Glyma13g04210.1           AGAAGTTTGATGCGTTGTTAACGAGCATGATTGAGGAGCATGTTGCTTCTAGTCACAAGA    796
                          ************************************************************

c8531_g1_i1               GAAAGGGCAAGCCCGATTTCTTAGACATGGTAATGGCTCATCATAGTGAGAACTCCGATG    1020
gi|302129056:26-1555      GAAAGGGCAAGCCCGATTTCTTAGACATGGTAATGGCTCATCATAGTGAGAACTCCGATG    856
Glyma13g04210.1           GAAAGGGCAAGCCCGATTTCTTAGACATGGTAATGGCTCATCATAGTGAGAACTCCGATG    856
                          ************************************************************

c8531_g1_i1               GGGAGGAACTATCGCTCACCAACATCAAGGCACTACTCTTGAACCTATTCACCGCAGGCA    1080
gi|302129056:26-1555      GGGAGGAACTATCGCTCACCAACATCAAGGCACTACTCTTGAACCTATTCACCGCAGGCA    916
Glyma13g04210.1           GGGAGGAACTATCGCTCACCAACATCAAGGCACTACTCTTGAACCTATTCACCGCAGGCA    916
                          ************************************************************

c8531_g1_i1               CCGATACATCTTCAAGTATAATAGAGTGGTCCTTAGCCGAGATGTTGAAGAAGCCCAGCA    1140
gi|302129056:26-1555      CCGATACATCTTCAAGTATAATAGAGTGGTCCTTAGCCGAGATGTTGAAGAAGCCCAGCA    976
Glyma13g04210.1           CCGATACATCTTCAAGTATAATAGAGTGGTCCTTAGCCGAGATGTTGAAGAAGCCCAGCA    976
                          ************************************************************

c8531_g1_i1               TAATGAAGAAGGCTCATGAAGAAATGGACCAAGTCATAGGAAGGGATCGCCGTCTCAAAG    1200
gi|302129056:26-1555      TAATGAAGAAGGCTCATGAAGAAATGGACCAAGTCATAGGAAGGGATCGCCGTCTCAAAG    1036
Glyma13g04210.1           TAATGAAGAAGGCTCATGAAGAAATGGACCAAGTCATAGGAAGGGATCGCCGTCTCAAAG    1036
                          ************************************************************

c8531_g1_i1               AATCTGACATACCAAAGCTTCCCTACTTCCAAGCCATTTGCAAAGAGACCTATAGAAAGC    1260
gi|302129056:26-1555      AATCTGACATACCAAAGCTTCCATACTTCCAAGCCATTTGCAAAGAGACCTATAGAAAGC    1096
Glyma13g04210.1           AATCTGACATACCAAAGCTTCCCTACTTCCAAGCCATTTGCAAAGAGACCTATAGAAAGC    1096
                          ********************** *************************************

c8531_g1_i1               ACCCTTCAACACCCCTAAACCTGCCTCGAATCTCATCTGAACCGTGCCAAGTGAATGGTT    1320
gi|302129056:26-1555      ACCCTTCAACACCCCTAAACCTGCCTCGAATCTCATCTGAACCGTGCCAAGTGAATGGTT    1156
Glyma13g04210.1           ACCCTTCAACACCCCTAAACCTGCCTCGAATCTCATCTGAACCGTGCCAAGTGAATGGTT    1156
                          ************************************************************

c8531_g1_i1               ACTACATTCCCGAGAACACTAGGCTGAATGTGAACATTTGGGCCATAGGAAGAGACCCTG    1380
gi|302129056:26-1555      ACTACATTCCCGAGAACACTAGGCTGAATGTGAACATTTGGGCCATAGGAAGAGACCCTG    1216
Glyma13g04210.1           ACTACATTCCCGAGAACACTAGGCTGAATGTGAACATTTGGGCCATAGGAAGAGACCCTG    1216
                          ************************************************************

c8531_g1_i1               ATGTGTGGAACAATCCTTTGGAGTTTATGCCCGAGAGGTTTTTGAGTGGGAAGAATGCCA    1440
gi|302129056:26-1555      ATGTGTGGAACAATCCTTTGGAGTTTATGCCCGAGAGGTTTTTGAGTGGGAAGAATGCCA    1276
Glyma13g04210.1           ATGTGTGGAACAATCCTTTGGAGTTTATGCCCGAGAGGTTTTTGAGTGGGAAGAATGCCA    1276
                          ************************************************************

c8531_g1_i1               AAATTGACCCACGTGGGAATGATTTTGAGCTTATTCCATTTGGTTCACTACATTTTGGGC    1500
gi|302129056:26-1555      AAATTGACCCACGTGGGAATGATTTTGAGCTTATTCCATTTGGTGCTGGGAGGAGGATTT    1336
Glyma13g04210.1           AAATTGACCCACGTGGGAATGATTTTGAGCTTATTCCATTTGGTGCTGGGAGGAGGATTT    1336
                          ******************************************** *

c8531_g1_i1               TTATTCCAT---------------------------------------------------    1509
gi|302129056:26-1555      GTGCAGGGACTAGGATGGGGATTGTGTTGGTT------CACTACATTTTGGGCACTTTGG    1390
Glyma13g04210.1           GTGCAGGGACTAGGATTTTGAGCTTATTCCATTTGGTTCACTACATTTTGGGCTTATTCC    1396
                           *

c8531_g1_i1               ------------------------------------------------------------    1509
gi|302129056:26-1555      TGCATTCGTTTGATTGGAAGCTACCCAATGGGGAGAGGGAGTTAGACATGGAGGAGTCCT    1450
```

```
Glyma13g04210.1          ATTTTTGA----------------------------------------------------    1404


c8531_g1_i1              ------------------------------------------------------------    1509
gi|302129056:26-1555     TTGGGCTTGCCTTGCAAAAAAAGGTTCCACTTGCTGCTTTGGTTACCCCTAGGTTGAATC    1510
Glyma13g04210.1          ------------------------------------------------------------    1404


c8531_g1_i1              --------------------      1509
gi|302129056:26-1555     CAAGTGCTTACATTTCTTAG      1530
Glyma13g04210.1          --------------------      1404
```