Genome and Transcriptome Based Characterization of Low Phytate Soybean and *Rsv3*-Type Resistance to *Soybean Mosaic Virus*

Neelam R. Redekar

Dissertation submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Crop and Soil Environmental Sciences

M. A. Saghai Maroof, Chair
Roderick V. Jensen
Richard F. Helm
Takeshi Fukao

July 17, 2015
Blacksburg, Virginia

Keywords: Seed development, Phytic acid, *Soybean mosaic virus* resistance, *Rsv3*, Co-expression network, Nucleotide-binding-leucine rich repeat protein

Genome and Transcriptome Based Characterization of Low Phytate Soybean and *Rsv3*-Type Resistance to *Soybean Mosaic Virus*

Neelam R Redekar

ABSTRACT

Soybean is a dominant oilseed cultivated worldwide for its use in multiple sectors such as food and feed industries, animal husbandry, cosmetics and pharmaceutical sectors, and more recently, in production of biodiesel. Increasing demand of soybean, changing environmental conditions, and evolution of pathogens pose challenges to soybean production in limited acreage. Genetic research is the key to ensure the continued growth in soybean production, with enhanced yield and quality, while reducing the losses due to diseases and pests. This research is focused on the understanding of transcriptional regulation of two economically important agronomic traits of soybean: low seed phytic acid and resistance to *Soybean mosaic virus* (SMV), using the 'transcriptomics' and 'genomics' approaches. The low phytic acid (*lpa*) soybean is more desirable than conventional soybean, as phytic acid is an anti-nutritional component of seed and is associated with phosphorus pollution. Despite the eco-friendly nature of the *lpa* soybean, it shows poor emergence, which reduces soybean yield. This research is mainly focused on addressing the impact of *lpa*-causing mutations on seed development, which is suspected to cause low emergence in *lpa* soybeans. The differences in transcriptome profiles of developing seeds in *lpa* and normal phytic acid soybean are revealed and the biological pathways that may potentially be involved in regulation of seed development are suggested. The second research project is focused on *Rsv3*-type resistance, which is effective against most virulent strains of *Soybean mosaic virus*. The *Rsv3* locus, which maps on to soybean chromosome 14, contains 10 genes including a cluster of coiled coil-nucleotide binding-leucine rich

repeat (CC-NB-LRR) protein-encoding genes. This dissertation employed a comparative sequencing approach to narrow down the list of *Rsv3* gene candidates to the most promising CC-NB-LRR gene. The evidence provided in this study clearly indicates a single CC-NB-LRR gene as the most promising candidate to deliver *Rsv3*-type resistance.

In loving memory of my grandfathers,
Mr. Marutrao Redekar and Mr. Narayanrao Ingawale.

Dedicated to women,
who were not able to get an education, to build a career,
and to pursue their passion!

Acknowledgements

Many people have contributed to the success of my research, and I would like to acknowledge them for their love, inspiration, and support.

I would like to thank my professor, Dr. M. A. Saghai Maroof, for seeing the potential in me and giving me an opportunity to work in his lab. He has been very supportive and has encouraged me to pursue my interests in bioinformatics. His research expertise and experiences have played an important role in shaping my career. I would like to thank my other committee members – Dr. Richard Helm, Dr. Roderick Jensen, and Dr. Takeshi Fukao, who have provided me immense support and expert advice.

I would like to thank Dr. Sue Tolin for being my mentor, guide, and well-wisher. She has trained me and gave me valuable research knowledge. I feel very fortunate for getting an opportunity to learn from her. I would like to thank Dr. Ruslan Biyashev for providing technical expertise in my experiments. Along with practical training, I have received endless support and motivation from him. I would also like to thank Dr. John McDowell, Dr. Guillaume Pilot, Dr. Elizabeth Grabau, and Dr. Glenda Gillaspy for providing me valuable feedback on my research.

I would like to thank my colleagues from the Maroof lab – Dr. Juan J. Ruiz-Rojas, Dr. Christin Kastl, Sandesh Shrestha, Colin Davis, Elizabeth Clevinger, and Lindsay Demers, for providing endless assistance and advice. I have made countless memories with them. Long, tiresome work hours have become enlightened in their presence, transforming work into fun. I would like to thank my colleagues, staff, and friends from the Crop and Soil Environmental Sciences department and Translational Plant Sciences program, who have always wished well for me and assisted me directly and indirectly towards my research.

I would like to thank my family and friends for their never-ending love and support. I would like to thank my parents, grandmother, and siblings who have made sacrifices to support my interest to pursue higher education. I very much appreciate my family for standing by my side against all the people who kept on telling my family that educating a girl was a bad choice over marriage. I would like to thank my best friend, Richard Rodrigues, for keeping me motivated, inspired, and focused throughout the term, and for helping me get through the tough times. I would like to thank Rodrigues family for countless love and support.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

PA: Phytic acid

lpa: low phytic acid

3mlpa: a low phytic acid line with three mutations (*mips1/mrp-l/mrp-n*)

3MWT: a wildtype line, without any mutation (MIPS1/MRP-L/MRP-N)

1mlpa: a low phytic acid line with one mutation (*mips1*)

1MWT: a wildtype line, without any mutation (MIPS1)

DEGs: Differentially Expressed Genes

RIL: Recombinant Inbred Line

MIPS: *myo*-inositol phosphate synthase

MRP: Multi-drug resistance protein

SMV: *Soybean mosaic virus*

Rsv: Resistance to SMV

CC-NB-LRR: Coiled-coil nucleotide-binding leucine-rich repeat

CNV: Copy number variation

GRAS: Gibberellin-Insensitive, Repressor of ga1–3, Scarecrow

ZIM: Zinc-finger protein expressed in Inflorescence Meristem

CAMTA: CAlModulin-binding Transcription Activator

GRF: Growth-Regulating Factor1

ZF-HD: Zinc Finger-Homeodomain

MBF1B: Multiprotein Bridging Factor 1

TCP: Teosinte branched 1, Cycloidea, PCF

ATTRIBUTION

All the manuscripts listed in this dissertation, i.e., Chapters 2, 3, 4, and 5, have multiple authors. Contributions of all co-authors in different chapters is described as follows:

- **Dr. M. A. Saghai Maroof:** Professor, Crop and Soil Environmental Sciences Department at Virginia Tech. He has participated in research proposal preparation, experimental design and coordination and reviewing of all manuscripts.
- **Dr. Elizabeth A Grabau:** Professor and Department Head, Plant Pathology, Physiology, and Weed Science at Virginia Tech. She has participated in research proposal preparation and reviewing of the differential gene expression manuscript.
- **Dr. Richard F. Helm:** Professor, Biochemistry Department at Virginia Tech. He has participated in research proposal preparation and reviewing of the manuscripts.
- **Dr. Roderick V. Jensen:** Professor, Biological Sciences Department at Virginia Tech. He has participated in research proposal preparation, and reviewing of the manuscripts.
- **Dr. Ruslan M Biyashev:** Research Manager, Crop and Soil Environmental Sciences Department at Virginia Tech. He has participated in development of genetic material, designing and conduct of the experiments, and reviewing of the manuscripts.
- **Dr. Song Li:** Assistant Professor, Crop and Soil Environmental Sciences, Virginia Tech. He has participated in experimental design, and review of the network manuscript.
- **Dr. Sue Tolin:** Professor Emerita, Plant Pathology, Physiology, and Weed Science, Virginia Tech. She has participated in experimental design, and review of the *Rsv3* manuscript.
- **Dr. Soon Chun Jeong:** Principal Researcher, Bio-Evaluation Center, Korea Research Institute of Bioscience and Biotechnology, Korea. He has participated in experimental design, and review of the *Rsv3* manuscript.
- **Dr. M. A. Laskar:** Biotechnology Department, St. Anthony's College, Shillong, India. He has participated in experimental design, and review of the *Rsv3* manuscript.

- **Elizabeth Clevinger:** Research Associate, Crop and Soil Environmental Sciences, Virginia Tech. She has participated in conduct of experiment, and review of the *Rsv3* manuscript.

CHAPTER 1

Literature Review

INTRODUCTION

High-throughput sequencing in modern genomics

Evolution of high-throughput sequencing technologies in the last decade has generated billions of short sequence fragments of size ranging between 50-20,000 base pairs (bp), in a matter of days or hours (depending on technology), at an affordable cost. Therefore, use of high-throughput sequencing for research has become commonplace. These sequencing platforms are mostly based on a technology referred to as "sequencing-by-synthesis," for example, "Solexa" technology with reversible termination (Illumina), "454" technology with pyrosequencing (Roche), "SOLiD" technology with ligation (Applied Biosystems), etc. The most recent platform from PacBio, known as "RS II," allows for single molecule real-time sequencing with fragment lengths up to 20,000 bp, and has wide applications in (1) de novo sequencing and re-sequencing (or DNA-Seq) of complex genomes, with their polyploidy and repetitive sequences, (2) transcriptome (or RNA-Seq) and DNA-methylome or epigenome sequencing for single cells, et cetera. These technologies are exploited for multiple biological applications such as ChIP-Seq (mapping DNA-protein interactions using chromatin immunoprecipitation), BS-Seq (mapping DNA-methylation sites using bisulphite treatment), RAD-Seq (mapping restriction site

associated DNA genotyping), GBS (genotyping-by-sequencing), et cetera. Numerous computational tools coupled with sophisticated infrastructure provide assistance for high-throughput sequencing data processing, management, and storage. Several protocols for high-throughput sequencing data analysis are available, enabling biologists or lab-scientists with limited computer expertise to apply this technology in their research [Wilhelm, et al. 2010, Trapnell, et al. 2012, Anders, et al. 2013].

Importance of soybean research

Soybean (*Glycine max* (L.) Merr.) is an economically important legume crop cultivated worldwide mainly for its oilseed properties. The United States ranks first in the world soybean production and export and the crop worth was estimated about 40 billion U.S. dollars in 2014 (SoyStats®2015). Soybean seeds serve as a major source of protein and oil, utilized in the food, cosmetics, animal feed, pharmaceutical and medicinal industries. Recently, soybean has also been employed in production of eco-friendly biodiesel. The demand for soybean and soybean products has been increasing in the past couple of decades. Fulfilling these ever-increasing demands is a challenge for soybean farmers with limited crop acreage. Genetic research is the key to ensure the continued growth in soybean production, with enhanced yield and quality, while reducing the losses due to diseases and pests. Extensive research is being conducted, ranging from basic science to understand the fundamental aspects of soybean development, to more market-driven applied research for crop improvement. The whole genome sequencing of the soybean cultivar 'Williams 82' in 2010 was a breakthrough for soybean genetics research [Schmutz, et al. 2010]. The availability of the soybean reference genome has stimulated remarkable advances in genomics research in recent years. Several soybean information

databases such as “SoyBase” - the USDA-ARS Soybean Genetics and Genomics Database (www.soybase.org), “SoyKB” - the Soybean Knowledge Base (www.soykb.org), “SoyDB” - the Soybean Genome Database (http://casp.rnet.missouri.edu/soydb), “Soy-TFKB” - the Soybean Transcription Factor Database (www.igece.org/Soybean_TF), “SoyTEDB” - the Soybean Transposable Elements Database (www.soybase.org/soytedb), “SoyMetDB” - the Soybean Metabolome Database (www.soymetdb.org), etc. are useful resources for soybean “omics” research.

Soybean functional genomics research

The release of the soybean reference genome sequence placed the the soybean research community in a position to explore the soybean genome diversity, functional structural variations, transcriptome profiles, and genetic basis of complex traits, et cetera [Du, et al. 2012, Anderson, et al. 2014, Coate, et al. 2014]. Two independent studies reported a high-resolution, integrated transcriptome map for diverse soybean tissues, such as roots, leaves, nodules, flowers, pods, seeds, and shoot meristems [Libault, et al. 2010, Severin, et al. 2010]. Libault et al. (2010) compared the soybean expression data with that of two other legumes, viz., *Medicago trancatula* and *Lotus japonicas*, and with *Arabidopsis thaliana.* The global gene expression profiles from these two studies were soon made accessible to the research community at the following websites: “www.soybase.org/soyseq” [Severin, et al. 2010] and “http://digbio.missouri.edu/soybean_atlas” [Libault, et al. 2010], to facilitate soybean research. Thereafter, numerous transcriptome expression studies were conducted to address topics such as drought stress [Martins, et al. 2008, Le, et al. 2012], the causative agent of soybean root and stem rot disease (*Phytophthora sojae*) [Li, et al. 2011], nodule formation [Hayashi, et al. 2012], salt

stress [Fan, et al. 2013]. It is clear that genomics has become an indispensable part of soybean research. The enormous amount of sequencing data generated from each of these studies can be a valuable resource for future genomics studies. In addition, the information can be utilized for meta-analyses to address new research problems.

Transcriptome of developing soybean seeds

The soybean seed developmental processes are directly responsible for the quality of mature seeds, and hence the market value of the product. As most key agronomic traits are related to seeds, such as improving seed nutrition, reducing anti-nutritional elements, engineering the seed developmental process is an important research goal. Several genomic studies have reported on soybean seed development. Using laser capture micro-dissection of seeds, coupled with microarray profiling, Le et al. (2007) have identified sets of genes that are activated in different compartments of developing soybeans [Le, et al. 2007]. Other microarray-based studies have compared the expression profile of genes between different seed developmental stages [Jones, et al. 2010, Asakura, et al. 2012]. The RNA-Seq experiment of Jones and Vodkin (2013) incorporated several seed developmental stages from the microarray-based study of Jones et al. (2010), which revealed an elaborate encounter of the seed development process in soybeans [Jones and Vodkin 2013]. Similarly, Shamimuzzaman and Vodkin (2012) incorporated seed developmental stages from the microarray-based study of Jones et al (2010) to study the regulation of gene expression using degradome sequencing; this involved microRNAs, or non-coding small RNAs associated with soybean seed development. A more elaborate study of regulation of developing seeds was reported more recently, which included 10 stages ranging

from embryo development to seed maturation [Collakova, et al. 2013]. The information available from these studies can be harnessed in future seed development studies.

SOYBEAN LOW PHYTIC ACID TRAIT

Phytic acid, a storage form of phosphorus in seeds

Phytic acid, also known as *myo*-inositol-1, 2, 3, 4, 5, 6-hexa*kis*phosphate or $InsP_6$ (hereafter, PA), is the most abundant inositol phosphate found ubiquitous in eukaryotes [Sasakawa, et al. 1995]. It is found in cereals and legume grains, nuts, oilseeds, tubers, pollen, spores, and organic soils [Cosgrove, et al. 1980, Loewus and Murthy 2000]. PA is a major storage form of phosphorus (P) in seeds, which accounts for 75% of the total seed P [Raboy 1997]. Plants absorb P from soil, which is an essential nutrient involved in photosynthesis, respiration, cell division, and energy transfer. Excess of P is translocated to the developing seed, which accumulates more P than is required for basic cellular functions [Raboy, et al. 2001]. The PA is primarily stored in the form of mixed salts of PA with mineral cations including zinc, iron, potassium, magnesium, and calcium, also known as phytate or phytin. The non-phytic acid form of P or inorganic P represents a minor portion of the total seed P. The PA metabolism during seed development and germination is regulated to maintain inorganic P levels in seeds [Strother 1980]. Since the majority of the absorbed plant nutrient P is diverted to the formation of PA, the seed PA represents the bottleneck in the flux of P in the agricultural ecosystem. In fact, it was estimated that the P content in PA produced globally is more than 60% of the total P applied world-wide in the form of mineral fertilizer [Lott, et al. 2000].

Phytic acid biosynthesis during seed development

The PA can be synthesized in plants during seed development via the phosphatidylinositol (PtdIns)-independent pathway, which is predominant in cereals and legumes as well as the PtdIns-dependent pathway during seed development. The first step involving the conversion of glucose-6-phosphate to *myo*-inositol-3-*mono*phosphate (denoted as $InsP_1$) is common to both pathways and is catalyzed by the *myo*-inositol-3-*mono*phosphate synthase (MIPS) [Loewus and Murthy 2000]. In the PtdIns-independent pathway, $InsP_1$ is sequentially phosphorylated to $InsP_2$, $InsP_3$, $InsP_4$, $InsP_5$, and ultimately $InsP_6$ or PA [Stevenson-Paulik, et al. 2005]. The PtdIns-dependent pathway involves the formation of *myo*-inositol from $InsP_1$ catalyzed by $InsP_1$ monophosphatase; this is followed by the formation of PtdIns and $PtdInsP_2$ catalyzed by PtdIns synthase and PtdIns kinases. The $PtdInsP_2$ is broken down by phospholipase C to form $InsP_3$ and diacylglycerol. The $InsP_3$ formed in this pathway is then sequentially phosphorylated as in PtdIns-independent pathway to form PA [Loewus and Murthy 2000, Raboy, et al. 2001]. Newly synthesized PA is transported to vacuoles for storage in the form of globoids. It is hypothesized that the multidrug resistance-associated protein ATP-binding cassette (ABC) transporters (hence onwards referred to as, MRPs) are involved in PA transport to vacuoles [Shi, et al. 2007, Nagy, et al. 2009, Sparvoli and Cominelli 2014]. The enzymes associated with the PA biosynthesis pathway, or intracellular compartmentalization, transport, and storage of inositol phosphate, can serve as potential targets to generate *lpa* traits in plants [Raboy 2009].

Role of phytic acid in plant growth and development

PA and its biosynthesis pathway intermediates, such as *myo*-inositol, serve as signaling messengers and are associated with numerous developmental and signaling processes:

phosphorous and mineral storage, DNA repair, chromatin remodeling, RNA editing and export, ATP generation, regulation of gene expression, regulation of guard cells, biotic and abiotic stress tolerance, oligosaccharide synthesis, and regulation of cell death. [York, et al. 1999, Hanakahi 2000, Lemtiri-Chlieh, et al. 2000, Lemtiri-Chlieh, et al. 2003, Shen, et al. 2003, Karner, et al. 2004, Taji, et al. 2006, Donahue, et al. 2010]. PA is also utilized in other metabolic pathways, including auxin metabolism and cell-wall polysaccharide biosynthesis [Loewus, et al. 1962, Tan, et al. 2007]. The basal plant defense mechanism-initiating signals require PA for regulating defenses against fungal, viral, and bacterial pathogens [Murphy, et al. 2008]. The PA provides P for ATP generation, therefore serving as the energy source, which is essential for early seed germination. Finally, PA also serves as an antioxidant for germinating seeds, due to its chelating action [Raboy 2003]. During seed germination, phytase, a phosphohydrolase enzyme, breaks down phytate to release inorganic P, minerals, and *myo*-inositol, which are used for seedling growth.

Significance of low phytic acid trait in crops

Despite the importance of PA in plant growth and development, it is considered as an anti-nutritive trait in cereal and legume crops. Seed-derived dietary PA is not effectively digested in humans and non-ruminant animals, such as poultry, swine, and fish, due to the lack of the phytase enzyme in their digestive tract [Brinch-Pedersen, et al. 2002]. PA also chelates important minerals like iron, zinc, calcium, manganese, et cetera, making it undesirable for consumption [Navert, et al. 1985, Hallberg, et al. 1989, Weaver, et al. 1991, Davidsson, et al. 1995]. Low iron and zinc absorption from legume and cereal-based foods is a major cause of the widespread mineral deficiency in infants in developing countries. Such deficiencies in infants can impair

growth and development, with long-term effects on their growth and development [Black 1998, Halterman, et al. 2001]. The undigested PA, released into the environment, can contaminate water bodies as a result of runoff and increase the risk of eutrophication, adversely affecting aquatic ecosystems [Sharpley, et al. 1994]. In order to achieve phytate degradation, microbial phytases are usually added to animal feed, but it is not a very economical alternative [Cromwell, et al. 1995, Raboy 2001]. Breeding of low PA (*lpa*) crops, on the other hand, provides a cost-effective solution to the problem. The *lpa* crops exhibit an increase in inorganic P content as a result of a reduction in PA, hence maintaining total seed P levels. Various animal feeding trial experiments have reported that *lpa* seeds can increase phosphorous availability to animals, satisfying more of their dietary requirements [Hill, et al. 2009]. Human nutrition studies have also shown that iron, zinc, and calcium availability is improved by 35-50% using *lpa* food [Raboy 2007]. An animal nutrition and waste utilization study has reported that P in manure can be reduced to as much as 75% with *lpa* feedstuff. Therefore, crops with reduced PA have gained noteworthy attention in the animal feed industry. The *lpa* crops are highly desirable due to positive effects on nutrition and the environment. The genetic materials, which carry *lpa*-causing mutations, have been used in numerous studies to explore this trait.

Low phytic acid crops and their challenges

Remarkable efforts have been made to develop *lpa* crop plants. Many key enzymes in the PA biosynthetic pathway, such as MIPS, myo-inositol kinases, inositol polyphosphate kinases, and MRPs, have been targeted to engineer the *lpa* trait [Hitz, et al. 2002, Shi, et al. 2003, Shi, et al. 2005, Shi, et al. 2007]. It was observed that *lpa* plants accumulate more inorganic (available) P, and total seed P remains unchanged. This can also alleviate the need for excessive use of

phosphorous fertilizers. The *lpa* crops have been developed in maize (*Zea mays* L.), barley (*Hordeum vulgar-ei* L.), soybean (*Glycine max* L.), wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.) etc., by random mutagenesis and screening for the *lpa* phenotype [Larson, et al. 1998, Larson, et al. 2000, Raboy, et al. 2000a, Wilcox, et al. 2000, Guttieri 2004]. Several studies have often revealed the negative impacts on seed performance resulting from lack of PA metabolism [Raboy, et al. 2000b, Rasmussen and Hatzack 2004, Oltmans, et al. 2005, Guttieri, et al. 2006]. Seeds with *lpa* content show poor emergence potential, resulting into yield reduction. Improving the emergence rate of *lpa* crops can enable their large-scale sector application in agriculture. In contrast to these reports, other studies suggest no effect on seed germination and/or emergence [Bregitzer and Raboy 2006, Yuan, et al. 2007, Dong, et al. 2013]. With transgene expression of myo-inositol methyltransferase, Dong et al. (2013) have induced breakdown of myo-inositol to ononitol, which eventually leads to a reduction in seed phytic acid, without affecting germination [Dong, et al. 2013]. These contrasting germination-related phenotypes can be due to a knockout of a different set of genes and the effect of environmental conditions employed in these studies.

Importance of low phytic acid soybean lines

The *lpa* in soybean has gained huge interest due to the crop demand in the animal feed industry as soybean in the form of soymeal is used as a protein source in animal feed formulations. Soymeal with normal PA content is not completely utilized by monogastric animals, which reduces its economic value. Soymeal with lower PA contents, on the other hand, is shown to improve P digestibility, reducing the excretion of P, which is expected to reduce the impact of P on the environment [Hill, et al. 2009]. Several *lpa* mutants have been characterized in soybean, including ‘LR33’, ‘M766’, ‘M153’, ‘CX-1834’, ‘V99-5089’, ‘*Gm-lpa*-TW-1’, and

'*Gm-lpa*-ZC-2' [Wilcox, et al. 2000, Hitz, et al. 2002, Oltmans, et al. 2004, Walker, et al. 2006, Yuan, et al. 2007, Maroof and Buss 2008, Gillman, et al. 2009, Maroof, et al. 2009].

V99-5089

The *Glycine max* line, **V99-5089**, was developed at Virginia Tech by conventional plant breeding. It produces soybean seeds with low phytate, low stachyose, and high sucrose content [Maroof and Buss 2008]. The quantitative trait locus (QTL) for low phytate/low stachyose/high sucrose seed content maps to chromosome 11 and is due to a point mutation from C to G, in the coding region of the MIPS1 gene within this locus [Maroof, et al. 2009].

CX-1834

Another low phytate *Glycine max* line, **CX-1834-1-6,** was developed at USDA/Purdue University [Wilcox, et al. 2000]. The low phytate phenotype in this line is controlled by two recessive gene alleles, and both alleles must be homozygous for low phytate seeds [Oltmans, et al. 2004]. The QTL for low phytate was later mapped to two epistatically interacting loci on chromosome 19 (Linkage group (LG)-L) and 3 (LG-N) [Walker, et al. 2006]. Single nucleotide mutations from A to T, resulting in the substitution of an Arg residue to a stop codon in a MRP gene on LG N, and a point mutation of G to A, resulting in an amino acid change from arginine to lysine on the 6^{th} exon of the MRP gene on LG-L [Maroof, et al. 2009], were responsible for a low phytate phenotype in CX-1834. The MRP genes on LG L and N will be referred to as MRP-L and MRP-N, respectively.

Background of the genetic material used in this research

The two *lpa* lines described above, V99-5089 and CX-1834, served as the parents for one of the pairs of experimental lines, developed at the Maroof lab in Virginia Tech, for use in this study. The genetic cross of V99-5089 x CX-1834 was made in 2001. Progenies from this population were selfed for 8 generations. Genotype verification was performed at every generation using the allelic discrimination method. A low phytate recombinant inbred line (RIL), *3mlpa* (*mips1/mrp-l/mrp-n*), and a normal phytate RIL, 3MWT (MIPS1/MRP-L/MRP-N), was developed from this population. Harvested in 2009, the $F_{8:9}$ seeds of these RILs were used for this study.

A third parental line, *Glycine max* cultivar "**Essex"** (PI548667), was also developed at Virginia Tech by conventional plant breeding. Essex is recognized for high yield, high stachyose, normal phytate, and sucrose content (USDA-ARC GRIN database, http://www.ars-grin.gov/). The genetic cross of Essex x V99-5089 was made, and progenies were examined every generation for *lpa*-causing *mips1* mutation, phytate and sugar content. The two near isogenic lines with normal phytate–1MWT (MIPS1/MRP-L/MRP-N), and low phytate–*1mlpa* (*mips1*/MRP-L/MRP-N) developed in this process were used as experimental lines. This genetic material is a unique resource for comparative genomic approaches to study the effect of three *lpa* mutations, viz., *mips1, mrp-l,* and *mrp-n,* on the regulation of the seed development process.

SOYBEAN MOSAIC DISEASE RESISTANCE TRAIT

Soybean mosaic disease—symptoms and impact

Soybean mosaic disease, caused by *Soybean mosaic virus* (SMV; Genus *Potyvirus*; Family *Potyviridae*), is one of the most important viral diseases found in all soybean-growing

areas worldwide. Seeds from the diseased plants are mostly responsible for transmission of this disease to the next generation. It is also transmitted via a non-persistent or the stylet-borne mechanism through the aphid, *Aphis glycines* [Hill, et al. 2001, Tolin, et al. 2004]. Disease symptoms include wrinkled, distorted, curled, and rolled leaves with blisters on leaf surfaces, appearing as a green to yellow mosaic pattern. It causes serious seed coat mottling (or discoloration). This reduces the seed quality, number of pod set, seed size, and weight, thereby causing significant yield losses in soybean [Tolin and Lacy 2004]. The yield loss associated with soybean mosaic disease is usually more than 30%, however total yield losses as high as 94% have also been reported [Hartman, et al. 1999]. The symptom severity and the accompanying yield losses largely depend on host genotype, virus strain, growth stage of plants at the time of infection, and environmental conditions [Tolin and Lacy 2004]. The most effective method to prevent the occurrence of soybean mosaic disease is the development of resistant cultivars, which often carry dominant resistance (*R*) genes against SMV.

Soybean mosaic virus—genome structure and strain groups

The SMV genome is a monopartite, single-stranded, positive-sense RNA approximately 9.5 kilobase nucleotides long with a single open reading frame. The genome is translated into a single polyprotein, which is cleaved to form 11 multifunctional proteins viz., P1 (Protein 1), HC-Pro (Helper Component-Protease), P3 (Protein 3), P3-PIPO (Pretty Interesting Potyviruses ORF), 6K1 (first 6KDa peptide), CI (Cylindrical Inclusion), 6K2 (second 6KDa peptide), NIa-Pro (Nuclear Inclusion ‘a’-Protease), NIa-VPg (Nuclear Inclusion ‘a’-viral genome-linked protein), NIb (Nuclear Inclusion ‘b’-replicase), and CP (Coat Protein) [Cui, et al. 2011]. These viral proteins are mainly involved in polyprotein cleavage, viral genome replication, virion

assembly, suppression of the host defense response, cell-to-cell movement, aphid and seed transmission of virus. Numerous SMV strains have been isolated worldwide. In the United States, the SMV strains are classified into seven strain groups (G1–G7) based on their virulence on a series of differential U.S. soybean cultivars [Cho and Goodman 1979]. Several strain groups were later added to this classification as SMV-N, G5H, G7A, G7H, C14, etc. In China, SMV strains are classified based on geographical distribution and response to resistant soybean cultivars [Wang, et al. 2003]. Diverse virulence groups of SMV co-evolved with their hosts show different virulence reactions on different soybean cultivars. The different R proteins in hosts detect the different viral proteins to initiate the host defense response. This SMV-soybean interaction has been extensively reviewed [Maroof, et al. 2008b, Cui, et al. 2011].

Genetic resistance to SMV—*Rsv1*, *Rsv3*, and *Rsv4*

Three independent, dominant resistance loci—*Rsv1*, *Rsv3*, and *Rsv4*—conditioning their resistance to SMV, have been identified by extensive screening of the soybean response to several U.S.-classified SMV strain groups (G1-G7). These loci encode for either single allele or multi-allelic genes, which provide resistance to either few or all SMV strain groups. The *Rsv1* locus encodes for a multi-allelic gene, which confers extreme resistance (symptomless response) to strains G1-G3, while necrotic or mosaic to strains G4-G7 [Kiihl and Hartwig 1979, Chen, et al. 1994]. This locus is mapped to soybean chromosome 13, to a region enclosing a cluster of resistance genes. The *Rsv4* locus encodes for a multi-allelic gene, which confers resistance to all strains of SMV (G1-G7), and is mapped to soybean chromosome 2 [Ma, et al. 1995, Hayes, et al. 2000, Maroof, et al. 2010]. The *Rsv4*-type resistance is associated with late susceptibility or the early resistance symptom characterized by restricted virus movement in infected plants [Ma, et

al. 2002, Gunduz, et al. 2004, Li, et al. 2010, Shakiba, et al. 2013]. The *Rsv3* locus confers resistance to the most virulent strains of SMV (G5-G7), and is mapped to soybean chromosome 14, enclosing a cluster of resistance genes [Jeong, et al. 2002, Jeong and Maroof 2004, Suh, et al. 2011]. The research described in the work reported here in was mainly focused on identifying the best candidate among the R genes enclosed by the *Rsv3* locus.

Rsv3 gene interaction with Soybean mosaic virus

The *Rsv3* locus confers resistance to highly virulent SMV strains G5 through G7. The *Rsv3*-genotype soybeans, however, develop susceptible necrotic mosaic symptoms with SMV strains G1 through G4 [Gunduz, et al. 2002]. The *Rsv3*-mediated resistance involves limiting the viral replication and movement to the inoculated leaves. The viral protein, CI, has been associated with the virulence and occurrence of symptoms in the *Rsv3*-genotype soybean [Zhang, et al. 2007, Zhang, et al. 2009]. Both N- and C-terminal regions of CI are involved in *Rsv3*-mediated resistance, but only the N-terminal region is involved in symptom development [Zhang, et al. 2009]. HC-Pro was also identified as the symptom determinant in *Rsv3*-mediated resistance [Seo, et al. 2011]. The 154-kb *Rsv3* locus on soybean chromosome 14 is comprised of 10 genes, which includes a cluster of five disease resistance genes, encoding coiled-coil nucleotide binding-leucine rich repeat (CC-NB-LRR) proteins [Suh, et al. 2011]. The NB-LRR proteins are disease resistance proteins that initiate a plant defense response once the pathogen has been detected. Therefore, the CC-NB-LRR encoding genes have the potential to be the key *Rsv3* gene. The high sequence similarity among the CC-NB-LRR encoding genes poses challenges to narrow down this gene cluster so as to identify a single dominant *Rsv3* gene. The

current research is focused on identifying the best *Rsv3*-gene candidate by narrowing down the CC-NB-LRR gene cluster from the locus.

RESEARCH OBJECTIVES

(1) Genome-wide transcriptome analysis of developing seeds from low and normal phytic acid soybean lines

(2) Regulation of metabolic process genes in seed development process in low phytic acid soybean lines

(3) Copy number variations in two isogenic-like soybean lines

(4) Candidate gene sequence analyses towards identifying *Rsv3*-type resistance *Soybean mosaic virus*

CHAPTER 2

Genome-wide transcriptome analysis of developing seeds from low and normal phytic acid soybean lines

Neelam R Redekar[1], Ruslan M Biyashev[1], Roderick V Jensen[2], Richard F Helm[3], Elizabeth A Grabau[4], MA Saghai Maroof[1§]

[1]Department of Crop and Soil Environmental Sciences. [2]Department of Biological Sciences. [3]Department of Biochemistry. [4]Department of Plant Pathology, Physiology, and Weed Science, Virginia Tech. [§]Corresponding author: smaroof@vt.edu

This Chapter is to be submitted for publication in *BMC Genomics*.

ABSTRACT

Background: Production of low phytic acid (*lpa*) crops is an eco-friendly alternative to conventional normal phytic acid (PA) crops, which causes poor mineral bioavailability in monogastric animals, resulting in phosphate pollution. The *lpa* crops carry mutations that are directly or indirectly associated with PA biosynthesis and accumulation during early seed development. The *lpa* crops exhibit altered carbohydrate profiles, increased free phosphate, and lower seed emergence which reduce overall crop yield, and limit their large-scale cultivation. In order to improve *lpa* crop yield, it is important to understand the downstream effects of *lpa*-causing mutations on seed development. Here, we present a comprehensive comparison of gene-expression profiles between *lpa* and normal PA developing soybean (*Glycine max*) using RNA-Seq approaches. The *lpa* line used in this study carries a single point mutation in a *myo*-inositol phosphate synthase gene, and in two multidrug-resistance protein ABC transporter genes.

Results: RNA from five seed developmental stages was sequenced in triplicate from both *lpa* and normal PA soybean lines (total of 30 libraries). Gene expression levels were estimated for every annotated gene and compared between *lpa* and the normal PA lines for each developmental stage. Our analyses revealed 4235 unique differentially expressed genes significantly represented in these five comparisons. This included 512 differentially expressed genes encoding for transcription factors. Functional enrichment of differentially expressed genes significant at a 1% false discovery rate represented biological processes including cellular glucan metabolism, apoptosis, photosynthesis, and glycolysis and transcription factor families including GRAS, WRKY, TCP, and CAMTA. Regulation of the raffinose oligosaccharide pathway in developing *lpa* and normal PA seeds was also estimated using quantitative real-time PCR.

Conclusions: The results show that developing soybean seeds in *lpa* mutant have differential gene expression profiles as compared to wildtype. Knowledge of differentially enriched biological processes and transcription factors suggests putative regulatory targets of *lpa* mutations. Overall, this study enhances our understanding of transcriptome-level regulation of developing soybean seeds in the presence of *lpa* mutations.

KEYWORDS

Phytic acid; *myo*-inositol phosphate synthase; multidrug-resistance protein ABC transporter; seed development; transcriptomics; differential gene expression; functional enrichment.

BACKGROUND

Soybean seed is one of the most important agricultural commodities produced worldwide, generating oils, proteins, and carbohydrates [Wilson 2004]. Final seed composition is greatly influenced by both the genotype and environmental factors [Meinke, et al. 1981, Breene, et al. 1988, Fehr, et al. 2003, Bennett and Krishnan 2005]. Breeding programs endeavor to improve the functional properties, and hence the economic value, of soybean by reducing anti-nutritive seed components such as phytic acid. Phytic acid (PA), a major source of phosphorus in seeds, can cause problems such as poor mineral bioavailability and phosphate pollution [Cosgrove D.J. 1980, Raboy 1997]. Low PA (*lpa*) crops are therefore highly desirable to reduce anti-nutritional and environmental effects of conventional crops [Zhou, et al. 1992, Htoo, et al. 2007, Hill, et al. 2009]. Two *lpa* soybean (*Glycine max* (L.) Merr.) lines, viz., 'V99-5089' and 'CX-1834' carry a non-lethal, recessive mutations in a *myo*-inositol phosphate synthase (MIPS) 1 gene and two

multidrug resistant protein (MRP)-type ATP-binding cassette transporter genes respectively [Wilcox, et al. 2000, Hitz, et al. 2002, Shi, et al. 2007, Maroof, et al. 2008, Gillman, et al. 2009, Maroof, et al. 2009]. The MIPS1 gene, expressed during seed development in soybean, is associated with the conversion of glucose-6-phosphate to *myo*-inositol-3-*mono*phosphate. This is the first step in the PA biosynthesis. This loss of the function mutation disrupts the pathway, reducing PA levels [Loewus and Loewus 1983, Chappell, et al. 2006]. The MRP genes are also highly expressed in developing embryos, however the mechanism by which they regulate PA levels in soybean is poorly understood [Shi, et al. 2007]. Combining these *lpa* mutations together ensures a higher reduction in PA content of soybean [Glover *et al.*, unpublished].

The PA biosynthesis pathway plays a vital role in maintaining homeostasis. Several pathway intermediates, such as *myo*-inositol-1, 4, 5-tris-phosphate, act as secondary messengers in signal transduction and are known to regulate growth and developmental processes, such as phosphorus and mineral storage, DNA repair, chromatin remodeling, RNA editing and export, ATP generation, regulation of gene expression, regulation of guard cells, auxin metabolism, and cell-wall polysaccharide biosynthesis [Loewus and Loewus 1983, York, et al. 1999, Hanakahi 2000, Lemtiri-Chlieh, et al. 2000, Loewus and Murthy 2000, Lemtiri-Chlieh, et al. 2003, Shen, et al. 2003]. Numerous studies have reported the effects of low PA on plant growth and development. An RNAi-mediated *mips1* knockdown in soybean was reported to inhibit seed development along with reduced PA content [Nunes, et al. 2006]. Similarly, seed embryo defects were reported for Arabidopsis and common bean (*Phaseolus vulgaris L.*) *mips* mutants [Donahue, et al. 2010, Abid, et al. 2012]. The *mips* mutation down-regulates the raffinose family oligosaccharide pathway, with mutants exhibiting impaired pathogen resistance, programmed cell death in leaves, and polar auxin transport causing deformed cotyledon development

[Murphy, et al. 2008, Abid, et al. 2009, Meng, et al. 2009, Obendorf, et al. 2009, Chen, et al. 2010, Donahue, et al. 2010, Luo, et al. 2011]. The *mrp* mutants are known to exhibit *lpa* phenotypes in soybean, rice, maize, and Arabidopsis [Shi, et al. 2007, Gillman, et al. 2009, Nagy, et al. 2009, Xu, et al. 2009]. MRP knockout studies in Arabidopsis also exhibited phenotypes such as insensitivity to ABA-mediated germination and unresponsive stomata opening, resulting in a reduced transpiration rate and increased drought tolerance; this was rescued by MRP overexpression [Klein, et al. 2003]. Moreover, *lpa* crops are known to show poor seed emergence, resulting in reduced crop yield decreasing the agronomic value of *lpa* crops [Raboy 2001, Gao, et al. 2008, Zhao, et al. 2008].

Despite these diverse physiological responses of different *lpa* mutations, very little is known about the effect of combining *lpa* mutations together on seed development and the underlying regulation of gene expression in soybean. Bowen et al. investigated microarray-based gene expression changes in developing embryos of a barley *lpa* mutant [Bowen, et al. 2007]. This study identified several differentially expressed genes associated with different cellular processes, such as carbohydrate and cell wall metabolism, hormonal signaling, and transport, suggesting the effect of the *lpa* mutation on barley seed development [Bowen, et al. 2007]. Advances in the sequencing technologies have enhanced the scope of genome-wide gene expression studies, permitting a global view of the transcriptome. We utilized mRNA-sequencing (or RNA-Seq) approach to study the combined effect of *lpa*-causing MIPS and MRP mutations on global changes in gene expression profiles of developing soybean seeds. The experimental lines (*lpa* mutant and wildtype) used in this study were developed from a single bi-parental cross [Glover et al., unpublished]. This unique genetic material eliminates the source-background effect, which is known to be a major confounding factor in gene expression studies.

In this study, a total of 30 transcriptome datasets derived from 5 developing seed stages with 3 biological replicates each of *lpa* mutant and wildtype soybean were sequenced and analyzed. To the best of our knowledge, this is the first extensive report describing the gene regulatory effect of MIPS and MRP mutations together. We identified several significantly enriched biological processes and transcription factors that suggest regulation of metabolism in the *lpa* mutant.

MATERIALS AND METHODS

Genetic material and background

The experimental lines of soybean (*Glycine max* (L.) Merr.) used in this study were: (1) a triple mutant line, called "*3mlpa*" (*mips1/mrp-l/mrp-n*) with low PA (hereafter, referred to as *lpa*), and (2) a wildtype line, called "3MWT" (MIPS1/MRP-L/MRP-N) with normal PA (hereafter, referred to as wildtype) (Table 2.1). These lines were developed from a cross of V99-5089 (*lpa*) with CX-1834 (*lpa*) [Maroof, et al. 2009]. The V99-5089 soybean experimental line carries a point mutation in the MIPS1 gene that results in *lpa*, low stachyose, and a high sucrose phenotype [Maroof and Buss 2008]. Similarly, the CX-1834 soybean line carries point mutations in two MRP genes, namely, MRP-L and MRP-N, located on chromosomes 19 and 3, respectively. Both mutations are required to obtain the *lpa* phenotype without any effect on seed stachyose and sucrose contents [Walker, et al. 2006, Gillman, et al. 2009, Maroof, et al. 2009]. Although the experimental line, carrying mutations in both MIPS1 and two MRP genes, shows reduced PA content, the stachyose and sucrose phenotype associated with the MIPS1 mutation is rescued (Table 2.1).

Plant growth and sampling

For each of the two experimental lines, 48 plants were grown in 12 pots (4 plants per pot) containing Metro-Mix® 360 (Sun Gro) soilless media, over-layered with GardenPro ULTRA^{LITE} topsoil. All plants were grown in the same growth chamber unit, with controlled conditions, as follows: 14/10 hours photoperiod, 24/16°C temperatures, light intensities in the range of 300 and 400 μE and the relative humidity 50-60%. About 41-47 plants from each experimental line were used for sampling of developing seeds. Seed length was the criterion for sampling different developmental stages. First, pods were randomly selected, opened and seed length was measured with a scale. Five developmental stages were defined by exclusive range of seed length as: Stage 1 (between 2-4 mm); Stage 2 (between 4-6 mm); Stage 3 (between 6-8 mm); Stage 4 (between 8-10 mm); and Stage 5 (between 10-12 mm) (Figure 2.1). We sampled three biological replicates for each stage, where each replicate sample was represented by a minimum of 10-15 seeds (stages 1-2), and at least 3 seeds (stages 3-5), collected from different pods on separate plants. Sampled seeds were immediately frozen in liquid nitrogen and stored at −70 °C.

RNA extraction, library preparation, and mRNA sequencing

Frozen seeds were ground to a fine powder to extract total RNA using RNeasy Plant Mini Kit, with on column DNase digestion (QIAGEN). Total RNA was diluted in RNase-free water, and the RNA concentration determined by UV spectrophotometry (260 nm, NanoDrop 1000, Thermo Fischer Scientific). RNA concentrations were then normalized to 200 ng/μl and the RNA integrity number (RIN) was measured using Bioanalyzer (Agilent Technologies). Total RNA samples with RIN values ranging between 9.0-10.0 were obtained as an indication of high

quality. High quality total RNA samples (50 µl each) were used for library preparation and mRNA sequencing at the Génome Québec Innovation Centre, Canada. A total of 30 cDNA libraries were generated using the TruSeq RNA sample preparation kit (Illumina) and sequenced on 5 lanes of a HiSeq2000 sequencing system (Illumina) to obtain single-end 100-bp long RNA-Seq reads. Six libraries representing three biological replicates of the single sampling stage from both *lpa* and wildtype were multiplexed together in single lane.

Transcriptomics data processing and analysis

Sequencing data quality control was performed prior to data analysis. Sequencing reads were then mapped/aligned to the well-annotated Williams82 soybean reference genome using the splice-aware mapping tool, TopHat, v2.0.8 [Schmutz, et al. 2010, Goodstein, et al. 2011, Kim, et al. 2013]. Sequence mapping data was used to estimate expression values for annotated genes using HTSeq-count [Anders, et al. 2014]. Differential gene expression analyses were performed using the statistical tool, DESeq, v1.12.1 [Anders and Huber 2010]. Genes with highly significant (less than 1% false discovery rate, FDR) fold change differences between low and normal PA soybean lines at each seed developmental stage were identified. Functional enrichment analysis was performed to identify ontology terms and pathways represented by these significant genes using online AgriGO tool [Du, et al. 2010]. R-script for the statistical hyper-geometric test was used to identify significantly enriched transcription factor families.

Quantitative real-time PCR

RNA-Seq output was validated using quantitative real-time PCR. First strand cDNA was synthesized from 2 µg of high quality total RNA (see above) using the High Capacity RNA-to-

cDNA kit (Applied Biosystems) following the manufacturer's instructions, from a total of 30 samples comprising of three biological replicates for each time point. About 2 µl of stock cDNA from each of the 30 samples were diluted to up to 200 µl using UltraPure™ distilled water (Invitrogen). Quantitative real-time PCR was performed in 20 µl of the total reaction volume, comprising of 4 µl of diluted cDNA, 10 µl of 2×SYBR Green PCR Master Mix (Applied Biosystems), 0.4 µl of 10 mM each of the gene-specific primers, and 3.2 µl of UltraPure™ distilled water. PCR conditions were: 50°C for 2 min, 95°C for 10 min, followed by 40 cycles at 95°C for 20s and finally the primer annealing temperature for 1 min, using the 7500 Real Time PCR system (Applied Biosystems). Melting curve analyses was performed to test primer specificity. Target gene specific primers were designed according to the soybean reference sequence using Primer3.0 for six randomly selected most significant differentially expressed genes [Koressaar and Remm 2007, Untergasser, et al. 2012]. Primer information is as provided in Supplementary Table A1. The ubiquitin 10 (UBQ10) gene was used as a reference gene to normalize the target gene transcript level amongst all samples [Hu, et al. 2009]. For estimating target and reference gene efficiency, an equal volume of diluted cDNA from all samples was pooled together. The standard curve of Ct values was generated using a five-fold serial dilution of this pooled cDNA sample. PCR efficiency was estimated using the equation: $E = 10^{(-1/slope)}$. Relative quantification of target gene transcript was estimated using the Pfaffl (or efficiency correction) method [Pfaffl 2001].

RESULTS AND DISCUSSION

Differential gene expression analyses

Five stages of soybean seed development representing tissue differentiation and nutrient accumulation events from each of the two lines were examined as whole seeds comprising of cotyledons, endosperm, and seed coat (Figure 2.1). According to [Meinke, et al. 1981], our seed stages are defined as early cotyledon to fully-grown cotyledon. We performed RNA-Seq analyses on these five stages of soybean seeds from each experimental line, with three biological replicates per stage. This experiment resulted in high throughput sequencing data with more than 961 million 100-bp long reads generated from 30 mRNA sample libraries (Supplementary Table A2). About 87% (more than 833 million) of the sequencing reads were mapped to the annotated soybean reference, Williams82 genome (Glyma1.1), using TopHat with default settings (Figure 2.2). The read count (number of reads mapping to a given gene) was estimated from sequencing mapping data for all the annotated gene models (total of 54,175). The reads mapping to more than one gene were eliminated while estimating read counts. Read counts were normalized and a differential expression analysis was performed using DESeq.

A principal component analysis (PCA) and sample-to-sample distance clustering variance stabilized log_2 transformed the normalized read count values for 54,175 genes from 30 mRNA libraries, as shown in Figure 2.3. Sample libraries generated from different seed developmental stages were distinctly represented along PC1 in a unidirectional pattern starting from stage 1 to 5 in PCA (Figure 2.3a). This means that developing seed stages are major contributors for variation in the data. Also, the sample libraries generated from the *lpa* line were clearly differentiated from their respective wildtype libraries along PC2. This means that genotype is the second largest contributor for variation in data. At the same time, the three biological replicate

samples of each stage were found clustering together, suggesting low variance between replicates. Thus the total variance in the data is defined by both seed developmental stages and genotype more than biological replicate variance. Moreover, similarities and dissimilarities between individual sample libraries were visualized using a heat map of sample-to-sample distance clustering (Figure 2.3b). This clustering suggests that sample libraries from early (stages 1-2) and late (stages 4-5) seed development stages are dissimilar to each other, while those from stage 3 are partially similar to both groups.

We identified a total of 4235 unique genes with significant differential expression between the *lpa* and wildtype lines at 1% FDR calculated using a P-value adjusted for multiple testing with the Benjamini-Hochberg method (Figure 2.4a-e, Table 2.2)[Benjamini and Hochberg 1995]. Fold-change (FC) ratio was calculated by dividing the mean normalized gene expression value in *lpa* over that of the wildtype. Of these, 2624 (62%) and 2485 (59%) genes identified as up- and down regulated in the *lpa* line, based on positive and negative log_2 (FC) ratios, respectively. These were 174 (4%) and 102 (2%) genes only expressed in either the *lpa* line or wildtype, respectively (Table 2.2). Some of the differentially expressed genes (DEGs) were also observed for more than one stage. Of these, 192 (4.5%) genes were represented in all five stages of seed development (Figure 2.4f). Many of these genes were functionally characterized in our reference genome, indicating diverse metabolic functions.

Functional enrichment analyses

We performed functional enrichment analyses on all the DEGs for each stage using a statistical hyper-geometric test with the Benjamini-Hochberg method for multiple tests to obtain *adjusted*-P-values with the AgriGO tool. These analyses resulted in identification of enriched

gene ontology (GO) terms, a standardized gene function classification system described in three categories: biological process, cellular component, and molecular function (www.geneontology.org). The enriched GO terms identified in this process were further filtered using *adjusted*-P-values <=0.01 (or 1% FDR) to obtain highly significant enriched GO terms. These GO terms can be arranged in a hierarchy, with more specialized terms at the bottom (child) originating from the less specialized terms at the top (parent). We focused on child terms in the GO hierarchy represented by various developing seed stages in our comparison (Table 2.3). Some of the enriched GO terms were found overlapping with more than one stage, while some were stage-specific, but none were common in all 5 stages, which suggest that not a single biological process was represented throughout the seed developmental stages. Hence, in order to simplify the interpretations, we grouped our data into two parts as early phase (stages 1-2) and late phase (stages 4-5) of seed development. Enriched GO terms from stage 3 showed partial overlap with both early and late phases. Most of the differentially expressed genes associated with significantly enriched biological processes in the early phase of seed development were up regulated, whereas the ones in late phase were down-regulated. Hierarchical clustering of the mean normalized gene expression level of DEGs associated with different biological processes is indicated in Figure 2.5.

Regulation of cell wall components in early seed development of *lpa* mutant

The cellular glucan metabolic process (GO:0006073) was found enriched in the early phase of seed development (stages 1-2), with a total of 27 unique genes up regulated and 2 genes down regulated in the *lpa* mutant. The genes associated with this process and the respective log_2 ratios for stage 1 and stage 2 are reported in Supplementary Table A3. The FC ratio of 27 up

regulated genes was in the range of 1.5-60.19 fold for stage 1, and 1.29-18.12 fold for stage 2 (Supplementary Table A3). A higher FC ratio indicates higher gene expression in the *lpa* mutant as compared to the wildtype, and vice-versa. This included genes encoding for cellulose synthase (also, known as glucan synthase, or CESA, EC 2.4.1.12) such as CESA4, CSL-B4, and xyloglucan endotrans-glucosylase/hydrolase (also known as xyloglucan:xyloglucosyl transferase or XET, EC 2.4.1.207) enzymes. The XET enzymatic activity (GO:0016762) was also significantly enriched under the molecular function domain. The CESA and XET enzymes are involved in the synthesis of building units of the cell wall, viz., cellulose, and xyloglucan (hemi-cellulose) chains, respectively. Differential expression of the CESA and XET genes suggest upregulation of processes associated with cell wall synthesis in the *lpa* mutant.

The expression level of genes associated with glucan cell wall metabolism also varied by the seed development stage. The FC ratio of the Glyma06g46450 gene (encoding CSL-B4) dropped from a 60-fold difference at stage 1 to less than 2-fold at stage 2. Similarly, the expression of the XET encoding genes, Glyma09g0707 and Glyma15g18360, with an FC ratio of 24-fold and 8.9-fold at stage 1, respectively, reduced considerably at stage 2 in the *lpa* mutant. In contrast, the XET encoding gene Glyma13g00280, showed an increase in the FC ratio from 6.8-fold in stage 1 to 18-fold in stage 2. These fluctuations in FC ratios between two stages suggest that different sets of genes take part in carbohydrate metabolism during early seed development.

Revealing the contribution of either *mips* or *mrp* mutations towards this response lies out of the scope of this experiment. Nonetheless, *myo*-inositol is associated with cell wall synthesis via an oxidation pathway as *myo*-inositol can be converted to D-glucuronic acid via the action of *myo*-inositol oxygenase (MIOX) [Loewus, et al. 1962]. However, because of the *mips1* mutation, the *lpa* mutant line is expected to have low *myo*-inositol levels. This means that the *myo*-inositol

precursor, glucose-6-phosphaste, in the *lpa* mutant must be getting converted into cell wall polysaccharides, cellulose, and xyloglucan, potentially through the pentose phosphate pathway. Although the carbohydrate metabolism has been previously reported in *lpa* mutants, none of the previous studies on *mips* or *mrp* mutants have reported any association with cell wall components such as cellulose and xyloglucan.

Regulation of defense response in early seed development of *lpa* mutant

The apoptosis process (GO:0006915) was found enriched in early stages of seed development (stages 1-3), with a total of 58 differentially expressed genes. More than 82% of these genes were up regulated in the *lpa* mutant line. The FC and $\log_2$ FC ratio of apoptosis-related differentially expressed genes is reported in Supplementary Table A4. These mainly included genes encoding for LRR (leucine rich repeat) and NB-ARC (nucleotide-binding adaptor shared by APAF-1, R proteins, and CED-4) domain-containing disease resistance proteins, Bcl-2-associated athanogene 1, ADR1-L1 (Activated Disease Resistant 1-like 1), cysteine proteinases, protein kinase and NTP hydrolases. The LRR and NB-ARC domain-containing disease resistance genes are involved in initiating a defense response, such as the formation of reactive oxygen species, induction of plant hormones, such as salicylic acid (SA), leading to apoptosis. Meng et al. (2009), reported that apoptosis in Arabidopsis *mips* mutants is dependent on SA accumulation, and that the mutant plants were rescued from apoptosis by treating them with either myo-inositol or galactinol [Meng, et al. 2009]. *Myo*-inositol abolishes SA-dependent apoptosis, which is triggered by peroxisomal hydrogen peroxide [Chaouch, et al. 2010]. Therefore, induction of defense-related genes in the *lpa* mutant helps to understand the cause of SA accumulation and cell death in *mips* mutants.

We also observed ADR1-L1 genes, which belong to a subgroup of the CNL-A clade of the coiled-coil NBS-LRR gene family [Meyers 2003]. Mutational studies in rice and Arabidopsis have associated ADR1 gene function with the dwarf phenotype [Kato, et al. 2011] and drought tolerance in presence of SA [Chini, et al. 2004]. The Arabidopsis MRP5 gene is known to be drought tolerant. We observed a two-fold higher expression of ADR1-L1 genes (Glyma14g08700 and Glyma17g36420) in *lpa* mutants. Further studies must be conducted to test drought tolerance of the *lpa* mutant in association with ADR1-L1 gene.

Regulation of cellular transport in early seed development of *lpa* mutant

Processes involving oligo-peptide transporters (GO:0006857) and transmembrane (GO:0055085) transporters were enriched in stage 2. This suggests transport processes differed between the *lpa* and wildtype lines. Although several transporter genes were differentially expressed in other stages, the GO term associated with the transporter activity was not significantly enriched in those stages. Together, oligo-peptide and transmembrane transport activity were associated with a total of 79 unique differentially expressed genes and 86% of these genes were found up regulated in the mutant line. About 44% of these genes were encoding for multidrug transporters, including 15 genes from the major facilitator superfamily (MFS), 6 genes from the multidrug and toxic compound extrusion (MATE) efflux carrier superfamily, 5 genes from the multidrug resistance superfamily, 2 genes for the P-glycoprotein (PGP) and 1 gene for the ATP binding cassette (ABC) subfamily (B4) (Supplementary Table A5). These multidrug transporters are mainly involved in the removal of toxic compounds from the cell [Zheleznova 2000]. Recently, an MFS transporter, Zinc-Induced Facilitator-Like 1 (ZIF1), from Arabidopsis was reported to be associated with polar auxin transport in roots, as well as the regulation of

stomata for drought stress tolerance [Remy, et al. 2013]. The ABC B4 transporters are also involved in auxin-gradient dependent polar auxin transport in roots [Kubes, et al. 2012]. PGP transporters are also involved in cellular and long distance transport of auxin [Geisler and Murphy 2006]. Defects in *mips* mutant embryos were previously associated with an impaired endomembrane system and lack of polar auxin transport [Luo, et al. 2011]. The genes encoding for monosaccharide transporters such as the Sugar Transport Protein (STP), Inositol Transporter (INT) and Polyol/Monosaccharide transporter (PMT) were also differentially expressed in our dataset. These genes are involved in transport of sugars such as glucose, fructose, galactose, mannose, xylose, sorbitol, mannitol, xylitol, and epimers and derivatives of myo-inositol [Slewinski 2011]. Other genes encoding for transporters/carriers of cationic amino acids, oligopeptides, potassium, sulfate, nitrates, zinc, chloride, dicarboxylate ions, etc., were also identified in our *lpa* mutant line.

Photosynthesis and glycolysis processes represented in late seed development in the *lpa* mutant

In later stages of seed development, processes involving photosynthesis (GO:0015979, stages 3-5), and glycolysis (GO:0006096, stages 3 and 5) were enriched. In stage 4, we also observed the process of generation of precursor metabolites and energy (GO:0006091), which is the GO ancestral term of glycolysis and photosynthesis. Most of the DEGs associated with these enriched processes were down regulated in the *lpa* mutant. Up to 55 DEGs associated with photosynthesis were down regulated in the *lpa* mutant. These included several genes encoding for different subunits in the photosystem (PS) I (such as, psaD, psaE, psaF, psaG, psaH, psaK, psaL, and psaN), and PS II (such as, psbA, psbE, psbP, psbQ, psbW, psbX, and psbY), light-

harvesting chlorophyll complex proteins from PS I (such as, LHCA1 and LHCA2) and PS II (such as, LHCB1, LHCB4, LHCB5), and a single gene encoding for the magnesium-protoporphyrin IX methyltransferase enzyme (EC: 2.1.1.11, CHLM) (Supplementary Table A6). This enzyme catalyzes the transfer of a methyl group from S-adenosyl methionine to magnesium protoporphyrin IX resulting in the formation of Mg-protoporphyrin IX monomethyl ester. This reaction contributes to porphyrin and chlorophyll metabolism and is involved in photosynthesis and respiration via the tetrapyrrole biosynthesis pathway [Tanaka and Tanaka 2007]. Differential expression of the photosystem complex building subunits suggests less photosynthesis in the *lpa* mutant line. Bowen et al. previously reported two photosynthesis-related genes differentially expressed in the M955 *lpa* mutant of maize at 7 days post-anthesis. One gene was down regulated, while the other gene was up regulated in the M955 *lpa* mutant [Bowen, et al. 2007]. Although *myo*-inositol protects the photosynthesis process in *Mesembryanthemum crystallinum,* the down-regulation of photosynthesis-related genes may or may not have been caused due to the lack of myo-inositol in the *lpa* mutant [Nelson, et al. 1998]. A gradual increase in chlorophyll pigments during the course of seed development can also be one explanation for observing photosynthesis in late stages, however the differential regulation of photosynthesis in *lpa* mutants should be the subject of further study.

Genes associated with glycolysis and Kreb's cycle, such as fructose-bisphosphate aldolase, phosphofructokinase, pyruvate kinase, phosphoglycerate kinase, sugar isomerase were found down regulated in the *lpa* mutant during these developmental stages (Supplementary Table A7). In stage 5, additional biological processes, such as nucleosome assembly, the malate metabolic process (GO:0006108), cellular amino acid biosynthetic process (GO:0008652), GTP

catabolic process (GO:0006184), lipid biosynthetic process (GO:0008610), and translation (GO:0006412) were also enriched.

RNA-Seq Data Validation

Gene expression profiles (*lpa* vs. wildtype) for six randomly selected DEGs associated with the above discussed biological processes (CESA: Glyma12g36570, Cyb559: Glyma15g38110, InsT2: Glyma15g22820, PS2D1: Glyma13g15560, SugT1: Glyma08g06420, and XET6: Glyma13g00280) were validated using quantitative real-time PCR (qPCR) (Figure 2.6). PCR efficiency of each target gene was estimated using the slope of the calibration curve with six 5-fold-dilutions of pooled cDNA as an input. The qPCR analyses suggested a higher fold change difference for CESA, XET, and InsT2 genes in early stages, while Cyb559 and PS2D1 genes saw a higher fold change difference in the late stages of seed development. For SugT1 gene, the fold change difference increased from stage 1 to stage 3, and then dropped until stage 5. Therefore, the differential expression of the CESA, XET, InsT2, SugT1, Cyb559, and PS2D1 genes suggests that differential regulation of biological processes happen in the *lpa* mutant. The qPCR data validates the relative gene expression levels obtained from the RNA-Seq analysis.

Transcription factor analyses

The first assembly soybean genome sequence is annotated to contain 5683 transcription factor (TF) genes from 63 families (Schmutz et al., 2010). Our significant (1% FDR) differentially expressed genes from seed developmental stages 1 to 5 provided a total of 512 genes encoding for TF from 32, 31, 33, 20, and 34 different TF families, respectively. From this

data set, we identified TF families significantly enriched within our dataset using hypergeometric testing with Benjamini-Hochberg adjusted P-values <= 0.01. This analysis resulted in identification of 2, 2, 4, 2, and 2 families significantly enriched in seed developmental stages 1 to 5, respectively. Table 2.4 represents the number of DEGs and adjusted P-values for enriched TF families in all five seed developmental stages. We observed that TF families like GRAS (Gibberellin-Insensitive, Repressor of ga1–3, Scarecrow), WRKY, ZF-HD (Zinc Finger-Homeodomain), and ZIM (Zinc-finger protein expressed in Inflorescence Meristem) were enriched in early stages (stages 1-2), whereas families like CAMTA (CAlModulin-binding Transcription Activator), GRF (Growth-Regulating Factor1), MBF1 (Multiprotein Bridging Factor 1), SNF2, and TCP (Teosinte branched 1, Cycloidea, PCF) were enriched in later stages (stages 3-5) of seed development. Among these enriched TF families, the TCP family was represented in stages 3 and 4, while CAMTA was represented in stages 3-5. All together these TF families represented 53 unique DEGs. The genes belonging to enriched TF families, such as WRKY, GRAS, ZIM, CAMTA, GRF, and SNF2, were up regulated, whereas those belonging to ZF-HD, MBF1B, and TCP were down regulated in the *lpa* mutant.

Regulation of raffinose family oligosaccharide biosynthesis in developing seeds of *lpa* mutant

The *lpa* mutant line carries a mutation in the coding region of the MIPS1 gene that results into a loss of catalytic function and reduced phytic acid levels. The *mips* mutation in V99-5089, one of the parents of the *lpa* mutant, is also known to cause a low stachyose/high sucrose phenotype [Maroof and Buss 2008]. Whereas, the levels of *myo*-inositol, stachyose, and raffinose in the developing seeds of the soybean *lpa* line CX-1834, which is also the source of *mrp-l* and

mrp-n mutations in our experimental lines, remain unchanged [Israel, et al. 2011]. However, the experimental *lpa* mutant line, despite the *mips* mutation, shows normal sucrose and stachyose levels in mature seeds [Glover et al., unpublished]. In an effort to evaluate the effect of *mips1*, *mrp-l*, and *mrp-n* mutations on the raffinose family oligosaccharides (RFOs) pathway, the transcript levels of galactinol synthase (GS), raffinose synthase (RS), and stachyose synthase (SS) encoding genes were evaluated using quantitative real-time PCR. Figure 2.7 represents the RFO pathway, and the relative expression values are indicated as the fold change ratio of mutant over wildtype. In the first step in this pathway, GS catalyzes the formation of galactinol from *myo*-inositol and UDP-galactose. We observed a gradual increase in relative expression levels of the GS (Glyma19g41550) gene with seed development. This suggests that in the *lpa* mutant, GS expression is down regulated in early-, and up regulated in late-stages of seed development. The raffinose synthase 2 (RS2) gene (Glyma06g18890), also known as Rsm1, controls the level of raffinose and stachyose in soybean [Skoneczka, et al. 2009]. The RS2 enzyme adds sucrose to galactinol by releasing a *myo*-inositol molecule. RS2 gene expression was previously observed in developing seeds of the Willams82 soybean reference line [Dierking, et al. 2008]. Similar to GS, we observed up-regulation of the RS2 gene expression in later stages of seed development. However, as opposed to GS and RS2 gene expressions, SS transcript levels were down regulated in the *lpa* mutant in all stages of seed development. Complexity of the inositol pathway and regulatory mechanism makes it difficult to explain the behavior of the stachyose synthase gene at this point. However, these observations suggest that the up regulation of the RFO pathway in the *lpa* mutant as compared to the wildtype.

CONCLUSION

PA biosynthesis pathway intermediates are involved in many growth and developmental processes. The key enzymes regulating this pathway are mutated to obtain *lpa* crops. Therefore, to develop better *lpa* crops, it is necessary to understand the global scenario of regulation of this pathway. We used a transcriptomics approach to identify differential gene expression in *lpa* soybean lines with *mips1, mrp-l,* and *mrp-n* mutations as compared to the wildtype by sequencing cDNAs from 5 different stages of developing soybean seeds. The differential expression and functional enrichment analyses indicated regulation of different biological processes such as glucan synthesis, apoptosis, photosynthesis, etc. We also identified regulated transcription factor families, such as WRKY, CAMTA, GRAS, ZIM, etc. These results delineate the metabolic events associated with regulation of the PA biosynthetic pathway in presence of *lpa* mutations. We also quantified transcript levels of enzymes involved in the raffinose family oligosaccharides pathway, suggesting differential transcript-level regulation in the *lpa* mutant. Overall, these results contribute towards understanding of regulation of metabolism in the *lpa* mutant during seed development.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

NRR carried out the designing and conducting of the experiment, sample preparation, data collection, sequencing data analysis, and drafted the manuscript. RMB participated in the conducting of the experiment. RFH, RVJ, and EAG helped to review the manuscript. MASM

participated in its design and coordination and reviewed the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

This work was funded by USDA-NIFA grant via Bio-design and Bioprocessing Research Center (BBRC) at Virginia Tech. We would like to thank support team of Advanced Research Computing (ARC) server and Translational Plant Sciences' MAGYK server at Virginia Tech. We would like to thank Dr. Victor Raboy, from United States Department of Agriculture's Agricultural Research Service in Aberdeen, Idaho, and Dr. Song Li from Virginia Tech, for providing helpful comments, which improved the quality of this manuscript.

REFERENCES

Abid G, Sassi K, Muhovski Y, Jacquemin J-M, Mingeot D, Tarchoun N, Baudoin J-P: **Comparative Expression and Cellular Localization of Myo-inositol Phosphate Synthase (MIPS) in the Wild Type and in an EMS Mutant During Common Bean (Phaseolus vulgaris L.) Seed Development**. *Plant Mol Biol Rep* 2012, **30**(3):780-793.

Abid G, Silue S, Muhovski Y, Jacquemin JM, Toussaint A, Baudoin JP: **Role of myo-inositol phosphate synthase and sucrose synthase genes in plant seed development**. *Gene* 2009, **439**(1-2):1-10.

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

Anders S, Pyl PT, Huber W: **HTSeq – A Python framework to work with high-throughput sequencing data**. 2014.

Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society Series B (Methodological)* 1995, **57**(1):289-300.

Bennett JO, Krishnan HB: **Long-term study of weather effects on soybean seed composition**. *Korean Journal of Crop Science* 2005, **50**(1):32-38.

Bolle C, Koncz C, Chua N-H: **PAT1, a new member of the GRAS family, is involved in phytochrome A signal transduction**. *Genes & Development* 2000, **14**(10):1269-1278.

Bowen DE, Souza EJ, Guttieri MJ, Raboy V, Fu J: **A Low Phytic Acid Barley Mutation Alters Seed Gene Expression**. *Crop Science* 2007, **47**(S2):S-149.

Breene WM, Lin S, Hardman L, Orf J: **Protein and oil content of soybeans from different geographic locations**. *Journal of the American Oil Chemists' Society* 1988, **65**(12):1927-1931.

Chaouch S, Noctor G: **Myo-inositol abolishes salicylic acid-dependent cell death and pathogen defence responses triggered by peroxisomal hydrogen peroxide**. *The New phytologist* 2010, **188**(3):711-718.

Chappell AS, Scaboo AM, Wu X, Nguyen H, Pantalone VR, Bilyeu KD: **Characterization of the MIPS gene family in Glycine max**. *Plant Breeding* 2006, **125**(5):493-500.

Chen H, Xiong L: **myo-Inositol-1-phosphate synthase is required for polar auxin transport and organ development**. *The Journal of biological chemistry* 2010, **285**(31):24238-24247.

Chini A, Fonseca S, Fernandez G, Adie B, Chico JM, Lorenzo O, Garcia-Casado G, Lopez-Vidriero I, Lozano FM, Ponce MR *et al*: **The JAZ family of repressors is the missing link in jasmonate signalling**. *Nature* 2007, **448**(7154):666-671.

Chini A, Grant JJ, Seki M, Shinozaki K, Loake GJ: **Drought tolerance established by enhanced expression of the CC-NBS-LRR gene, ADR1, requires salicylic acid, EDS1 and ABI1**. *The Plant journal : for cell and molecular biology* 2004, **38**(5):810-822.

Cosgrove D.J. IGCJ: **Inositol phosphates: their chemistry, biochemistry, and physiology**: Elsevier Scientific Pub. Co.; 1980.

Di Mauro MF, Iglesias MJ, Arce DP, Valle EM, Arnold RB, Tsuda K, Yamazaki K, Casalongue CA, Godoy AV: **MBF1s regulate ABA-dependent germination of Arabidopsis seeds**. *Plant Signal Behav* 2012, **7**(2):188-192.

Dierking EC, Bilyeu KD: **Association of a Soybean Raffinose Synthase Gene with Low Raffinose and Stachyose Seed Phenotype**. *The Plant Genome Journal* 2008, **1**(2):135.

Donahue JL, Alford SR, Torabinejad J, Kerwin RE, Nourbakhsh A, Ray WK, Hernick M, Huang X, Lyons BM, Hein PP *et al*: **The Arabidopsis thaliana Myo-inositol 1-phosphate synthase1 gene is required for Myo-inositol synthesis and suppression of cell death**. *The Plant cell* 2010, **22**(3):888-903.

Du L, Ali GS, Simons KA, Hou J, Yang T, Reddy AS, Poovaiah BW: **Ca(2+)/calmodulin regulates salicylic-acid-mediated plant immunity**. *Nature* 2009, **457**(7233):1154-1158.

Du Z, Zhou X, Ling Y, Zhang Z, Su Z: **agriGO: a GO analysis toolkit for the agricultural community**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W64-70.

Eulgem T, Rushton PJ, Robatzek S, Somssich IE: **The WRKY superfamily of plant transcription factors**. *Trends in plant science*, **5**(5):199-206.

Fehr WR, Hoeck JA, Johnson SL, Murphy PA, Nott JD, Padilla GI, Welke GA: **Genotype and Environment Influence on Protein Components of Soybean**. *Crop Science* 2003, **43**(2):511.

Galon Y, Nave R, Boyce JM, Nachmias D, Knight MR, Fromm H: **Calmodulin-binding transcription activator (CAMTA) 3 mediates biotic defense responses in Arabidopsis**. *FEBS Lett* 2008, **582**(6):943-948.

Gao Y, Biyashev RM, Maroof MAS, Glover NM, Tucker DM, Buss GR: **Validation of Low-Phytate QTLs and Evaluation of Seedling Emergence of Low-Phytate Soybeans**. *Crop Science* 2008, **48**(4):1355.

Geisler M, Murphy AS: **The ABC of auxin transport: the role of p-glycoproteins in plant development**. *FEBS Lett* 2006, **580**(4):1094-1102.

Gillman JD, Pantalone VR, Bilyeu K: **The Low Phytic Acid Phenotype in Soybean Line CX1834 Is Due to Mutations in Two Homologs of the Maize Low Phytic Acid Gene**. *The Plant Genome Journal* 2009, **2**(2):179.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N *et al*: **Phytozome: a comparative platform for green plant genomics**. *Nucleic Acids Research* 2011.

Hanakahi L: **Binding of Inositol Phosphate to DNA-PK and Stimulation of Double-Strand Break Repair**. *Cell* 2000, **102**(6):721-729.

Helariutta Y, Fukaki H, Wysocka-Diller J, Nakajima K, Jung J, Sena G, Hauser M-T, Benfey PN: **The SHORT-ROOT Gene Controls Radial Patterning of the Arabidopsis Root through Radial Signaling**. *Cell*, **101**(5):555-567.

Herve C, Dabos P, Bardet C, Jauneau A, Auriac MC, Ramboer A, Lacout F, Tremousaygue D: **In vivo interference with AtTCP20 function induces severe plant growth alterations and deregulates the expression of many genes important for development**. *Plant physiology* 2009, **149**(3):1462-1477.

Hill BE, Sutton AL, Richert BT: **Effects of low-phytic acid corn, low-phytic acid soybean meal, and phytase on nutrient digestibility and excretion in growing pigs**. *Journal of animal science* 2009, **87**(4):1518-1527.

Hitz WD, Carlson TJ, Kerr PS, Sebastian SA: **Biochemical and molecular characterization of a mutation that confers a decreased raffinosaccharide and phytic acid phenotype on soybean seeds**. *Plant physiology* 2002, **128**(2):650-660.

Htoo JK, Sauer WC, Zhang Y, Cervantes M, Liao SF, Araiza BA, Morales A, Torrentera N: **The effect of feeding low-phytate barley-soybean meal diets differing in protein content to**

growing pigs on the excretion of phosphorus and nitrogen. *Journal of animal science* 2007, **85**(3):700-705.

Hu R, Fan C, Li H, Zhang Q, Fu YF: **Evaluation of putative reference genes for gene expression normalization in soybean by quantitative real-time RT-PCR**. *BMC Mol Biol* 2009, **10**:93.

Israel DW, Taliercio E, Kwanyuen P, Burton JW, Dean L: **Inositol Metabolism in Developing Seed of Low and Normal Phytic Acid Soybean Lines.** *Crop Sci* 2011, **51**(1):282-289.

Kato H, Shida T, Komeda Y, Saito T, Kato A: **Overexpression of the Activated Disease Resistance 1-like1 (ADR1-L1) Gene Results in a Dwarf Phenotype and Activation of Defense-Related Gene Expression in Arabidopsis thaliana**. *Journal of Plant Biology* 2011, **54**(3):172-179.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome biology* 2013, **14**(4):R36.

Kim JH, Choi D, Kende H: **The AtGRF family of putative transcription factors is involved in leaf and cotyledon growth in Arabidopsis**. *The Plant Journal* 2003, **36**(1):94-104.

Klein M, Perfus-Barbeoch L, Frelet A, Gaedeke N, Reinhardt D, Mueller-Roeber B, Martinoia E, Forestier C: **The plant multidrug resistance ABC transporter AtMRP5 is involved in guard cell hormonal signalling and water use**. *The Plant Journal* 2003, **33**(1):119-129.

Koressaar T, Remm M: **Enhancements and modifications of primer design program Primer3**. *Bioinformatics* 2007, **23**(10):1289-1291.

Kubes M, Yang H, Richter GL, Cheng Y, Mlodzinska E, Wang X, Blakeslee JJ, Carraro N, Petrasek J, Zazimalova E *et al*: **The Arabidopsis concentration-dependent influx/efflux**

transporter ABCB4 regulates cellular auxin levels in the root epidermis. *The Plant journal : for cell and molecular biology* 2012, **69**(4):640-654.

Lemtiri-Chlieh F, MacRobbie EA, Brearley CA: **Inositol hexakisphosphate is a physiological signal regulating the K+-inward rectifying conductance in guard cells**. *Proc Natl Acad Sci U S A* 2000, **97**(15):8687-8692.

Lemtiri-Chlieh F, MacRobbie EA, Webb AA, Manison NF, Brownlee C, Skepper JN, Chen J, Prestwich GD, Brearley CA: **Inositol hexakisphosphate mobilizes an endomembrane store of calcium in guard cells**. *Proc Natl Acad Sci U S A* 2003, **100**(17):10091-10095.

Li S, Zachgo S: **TCP3 interacts with R2R3-MYB proteins, promotes flavonoid biosynthesis and negatively regulates the auxin response in Arabidopsis thaliana**. *The Plant Journal* 2013, **76**(6):901-913.

Loewus FA, Kelly S, Neufeld EF: **Metabolism of Myo-inositol in Plants: Conversion to Pectin, Hemicellulose, D-xylose, and Sugar Acids**. *Proceedings of the National Academy of Sciences of the United States of America* 1962, **48**(3):421-425.

Loewus FA, Loewus MW: **myo-Inositol:Its Biosynthesis and Metabolism**. *Annual Review of Plant Physiology* 1983, **34**(1):137-161.

Loewus FA, Murthy PPN: **myo-Inositol metabolism in plants**. *Plant Science* 2000, **150**(1):1-19.

Luo Y, Qin G, Zhang J, Liang Y, Song Y, Zhao M, Tsuge T, Aoyama T, Liu J, Gu H *et al*: **D-myo-inositol-3-phosphate affects phosphatidylinositol-mediated endomembrane function in Arabidopsis and is essential for auxin-regulated embryogenesis**. *The Plant cell* 2011, **23**(4):1352-1372.

Maroof AS, Buss GR: **Low phytic acid, low stachyose, high sucrose soybean lines**. In.: Google Patents; 2008.

Maroof MAS, Glover NM, Biyashev RM, Buss GR, Grabau EA: **Genetic Basis of the Low-Phytate Trait in the Soybean Line CX1834**. *Crop Science* 2009, **49**(1):69.

Meinke DW, Chen J, Beachy RN: **Expression of storage-protein genes during soybean seed development**. *Planta* 1981, **153**(2):130-139.

Meng PH, Raynaud C, Tcherkez G, Blanchet S, Massoud K, Domenichini S, Henry Y, Soubigou-Taconnat L, Lelarge-Trouverie C, Saindrenan P *et al*: **Crosstalks between Myo-Inositol Metabolism, Programmed Cell Death and Basal Immunity in *Arabidopsis***. *PLoS ONE* 2009, **4**(10):e7364.

Meyers BC: **Genome-Wide Analysis of NBS-LRR-Encoding Genes in Arabidopsis**. *The Plant Cell Online* 2003, **15**(4):809-834.

Murphy AM, Otto B, Brearley CA, Carr JP, Hanke DE: **A role for inositol hexakisphosphate in the maintenance of basal resistance to plant pathogens**. *The Plant journal : for cell and molecular biology* 2008, **56**(4):638-652.

Nagy R, Grob H, Weder B, Green P, Klein M, Frelet-Barrand A, Schjoerring JK, Brearley C, Martinoia E: **The Arabidopsis ATP-binding cassette protein AtMRP5/AtABCC5 is a high affinity inositol hexakisphosphate transporter involved in guard cell signaling and phytate storage**. *The Journal of biological chemistry* 2009, **284**(48):33614-33622.

Nelson DE, Rammesmayer G, Bohnert HJ: **Regulation of cell-specific inositol metabolism and transport in plant salinity tolerance**. *The Plant cell* 1998, **10**(5):753-764.

Nunes AC, Vianna GR, Cuneo F, Amaya-Farfan J, de Capdeville G, Rech EL, Aragao FJ: **RNAi-mediated silencing of the myo-inositol-1-phosphate synthase gene (GmMIPS1) in**

transgenic soybean inhibited seed development and reduced phytate content. *Planta* 2006, **224**(1):125-132.

Obendorf RL, Zimmerman AD, Zhang Q, Castillo A, Kosina SM, Bryant EG, Sensenig EM, Wu J, Schnebly SR: **Accumulation of Soluble Carbohydrates during Seed Development and Maturation of Low-Raffinose, Low-Stachyose Soybean.** *Crop Sci* 2009, **49**(1):329-341.

Pfaffl MW: **A new mathematical model for relative quantification in real-time RT-PCR**. *Nucleic Acids Research* 2001, **29**(9):45e-45.

Pysh LD, Wysocka-Diller JW, Camilleri C, Bouchez D, Benfey PN: **The GRAS gene family in Arabidopsis: sequence characterization and basic expression analysis of the SCARECROW-LIKE genes**. *The Plant Journal* 1999, **18**(1):111-119.

Raboy V: **Accumulation and storage of phosphate and minerals.** . In: *Cellular and Molecular Biology of Plant Seed Development* Dordrecht, Netherlands: Kluwer Academic Publishers; 1997: 441-477.

Raboy V: **Seeds for a better future: 'low phytate' grains help to overcome malnutrition and reduce pollution**. *Trends in plant science* 2001, **6**(10):458-462.

Remy E, Cabrito TR, Baster P, Batista RA, Teixeira MC, Friml J, Sa-Correia I, Duque P: **A major facilitator superfamily transporter plays a dual role in polar auxin transport and drought stress tolerance in Arabidopsis**. *The Plant cell* 2013, **25**(3):901-926.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J *et al*: **Genome sequence of the palaeopolyploid soybean**. *Nature* 2010, **463**(7278):178-183.

Shaked H, Avivi-Ragolsky N, Levy AA: **Involvement of the Arabidopsis SWI2/SNF2 chromatin remodeling gene family in DNA damage response and recombination**. *Genetics* 2006, **173**(2):985-994.

Shen X, Xiao H, Ranallo R, Wu W-H, Wu C: **Modulation of ATP-Dependent Chromatin-Remodeling Complexes by Inositol Polyphosphates**. *Science* 2003, **299**(5603):112-114.

Shi J, Wang H, Schellin K, Li B, Faller M, Stoop JM, Meeley RB, Ertl DS, Ranch JP, Glassman K: **Embryo-specific silencing of a transporter reduces phytic acid content of maize and soybean seeds**. *Nature biotechnology* 2007, **25**(8):930-937.

Skoneczka JA, Maroof MAS, Shang C, Buss GR: **Identification of Candidate Gene Mutation Associated With Low Stachyose Phenotype in Soybean Line PI200508**. *Crop Science* 2009, **49**(1):247.

Slewinski TL: **Diverse functional roles of monosaccharide transporters and their homologs in vascular plants: a physiological perspective**. *Mol Plant* 2011, **4**(4):641-662.

Tanaka R, Tanaka A: **Tetrapyrrole biosynthesis in higher plants**. *Annu Rev Plant Biol* 2007, **58**:321-346.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG: **Primer3--new capabilities and interfaces**. *Nucleic Acids Res* 2012, **40**(15):e115.

Walker DR, Scaboo AM, Pantalone VR, Wilcox JR, Boerma HR: **Genetic mapping of loci associated with seed phytic acid content in CX1834-1-2 soybean**. *Crop Science* 2006, **46**(1):390-397.

Wang L, Hua D, He J, Duan Y, Chen Z, Hong X, Gong Z: **Auxin Response Factor2 (ARF2) and its regulated homeodomain gene HB33 mediate abscisic acid response in Arabidopsis**. *PLoS Genet* 2011, **7**(7):e1002172.

Wilcox JR, Premachandra GS, Young KA, Raboy V: **Isolation of High Seed Inorganic P, Low-Phytate Soybean Mutants**. *Crop Science* 2000, **40**(6):1601.

Wilson JH: **Seed composition**. In: *Soybeans : improvement, production, and uses.* Edited by Boerma HR, Specht JE, 3rd edn. Madison, Wis.: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America; 2004: 621-668.

Xu XH, Zhao HJ, Liu QL, Frank T, Engel KH, An G, Shu QY: **Mutations of the multi-drug resistance-associated protein ABC transporter gene 5 result in reduction of phytic acid in rice seeds**. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik* 2009, **119**(1):75-83.

York JD, Odom AR, Murphy R, Ives EB, Wente SR: **A Phospholipase C-Dependent Inositol Polyphosphate Kinase Pathway Required for Efficient Messenger RNA Export**. *Science* 1999, **285**(5424):96-100.

Zhang Z-L, Ogawa M, Fleet CM, Zentella R, Hu J, Heo J-O, Lim J, Kamiya Y, Yamaguchi S, Sun T-p: **SCARECROW-LIKE 3 promotes gibberellin signaling by antagonizing master growth repressor DELLA in Arabidopsis**. *Proceedings of the National Academy of Sciences* 2011, **108**(5):2160-2165.

Zhao H-J, Liu Q-L, Fu H-W, Xu X-H, Wu D-X, Shu Q-Y: **Effect of non-lethal low phytic acid mutations on grain yield and seed viability in rice**. *Field Crops Research* 2008, **108**(3):206-211.

Zheleznova E: **A structure-based mechanism for drug binding by multidrug transporters**. *Trends in Biochemical Sciences* 2000, **25**(2):39-43.

Zhou JR, Fordyce EJ, Raboy V, Dickinson DB, Wong MS, Burns RA, Erdman JW, Jr.: **Reduction of phytic acid in soybean products improves zinc bioavailability in rats**. *The Journal of nutrition* 1992, **122**(12):2466-2473.

Table 2.1: Characteristics of experimental lines and their parents.

Cultivars	Genotype‡	Phytate	Emergence	Stachyose	Sucrose
V99-5089	*mips1*/MRP-L/MRP-N	Low	Low	Low	High
CX-1834	MIPS1/*mrp-l/mrp-n*	Low	Low	Normal	Normal
lpa	*mips1/mrp-l/mrp-n*	Low	Low	Normal	Normal
Wildtype	MIPS1/MRP-L/MRP-N	Normal	Normal	Normal	Normal

‡ Text in italics indicate mutations or mutant line.
All experimental lines represented here are homozygous.

Table 2.2: Differential gene expression between *lpa* mutant and wildtype.

Stages	1	2	3	4	5
Differentially expressed genes (DEGs)‡	1526	1791	1348	684	1639
DEGs up regulated in *lpa*	831	1114	788	269	493
DEGs down regulated in *lpa*	695	677	560	415	1146
DEGs only expressed in *lpa*	29	41	39	32	33
DEGs only expressed in wildtype	26	20	17	21	18

‡ Out of total 6988 DEGs, 4235 were unique (counted only once).
Remaining genes were repeatedly identified in more than one stage.

Table 2.3: Enriched gene ontology terms associated with biological processes.

Stage	GO Term	Biological Process	DEG	P-value	FDR
1	GO:0006915	Apoptosis	31	4.0E-04	3.4E-03
	GO:0006073	Cellular Glucan Metabolic Process	17	6.8E-05	7.4E-04
	GO:0006334	Nucleosome Assembly	15	6.9E-06	1.2E-04
	GO:0006412	Translation	71	2.9E-10	6.4E-08
2	GO:0006915	Apoptosis	39	1.1E-05	7.9E-04
	GO:0006073	Cellular Glucan Metabolic Process	18	1.0E-04	3.4E-03
	GO:0006857	Oligopeptide Transport	15	1.6E-04	3.5E-03
	GO:0055114	Oxidation Reduction	140	3.6E-06	7.9E-04
	GO:0055085	Transmembrane Transport	68	3.4E-04	6.3E-03
3	GO:0006915	Apoptosis	42	1.9E-10	1.7E-08
	GO:0006096	Glycolysis	10	1.2E-03	9.8E-03
	GO:0045087	Innate Immune Response	20	2.2E-06	5.1E-05
	GO:0015979	Photosynthesis	20	6.0E-07	2.2E-05
4	GO:0006091	Generation of precursor metabolites and energy (Glycolysis and Photosynthesis)	14	7.0E-06	4.9E-04
	GO:0015979	Photosynthesis	16	4.1E-09	5.6E-07
5	GO:0008652	Cellular Amino acid Biosynthetic Process	18	1.0E-03	4.0E-03
	GO:0006096	Glycolysis	12	7.3E-04	3.8E-03
	GO:0006184	GTP Catabolic Process	7	6.6E-04	3.6E-03
	GO:0008610	Lipid Biosynthetic Process	30	9.2E-04	3.8E-03
	GO:0006108	Malate Metabolic Process	6	8.3E-05	5.7E-04
	GO:0006334	Nucleosome Assembly	20	9.2E-09	2.1E-07
	GO:0015979	Photosynthesis	55	4.4E-30	1.1E-27
	GO:0009765	Photosynthesis, Light Harvesting	14	1.3E-09	6.6E-08
	GO:0006814	Sodium ion Transport	5	8.2E-04	3.8E-03
	GO:0006414	Translational Elongation	9	4.6E-04	2.7E-03
	GO:0006412	Translation	63	1.7E-06	1.9E-05

Table 2.4: Transcription factor families significantly enriched in developing seed stages.

Stages	TF Family	FDR	Genes	DEGs associated with TF family	Function
AB1	WRKY	4.92E-04	WRKY33, WRKY40, WRKY29, WRKY6, WRKY28, WRKY23, WRKY15, WRKY11	Glyma11g29720, Glyma08g23380, Glyma08g02160, Glyma13g44730, Glyma09g00820, Glyma12g10350 Glyma08g08720, Glyma13g38630, Glyma15g11680, Glyma03g37940, Glyma17g18480, Glyma05g20710, Glyma06g08120, Glyma04g08060	Associated with plant defense, senescence, and abiotic stress [Eulgem, et al.]
	GRAS	3.19E-04	SCL1, SCL3, SCL5, SCL14, PAT1, SGR7	Glyma04g28490, Glyma18g09030, Glyma08g43780, Glyma11g14670, Glyma14g27290, Glyma13g09220, Glyma15g04160, Glyma14g01960, Glyma14g01020, Glyma17g17400, Glyma13g02840	Involved in gibberellin signaling, phytochrome A signal transduction, controls radial patterning [Helariutta, et al. , Pysh, et al. 1999, Bolle, et al. 2000, Zhang, et al. 2011]
AB2	ZIM	8.60E-03	JAZ6, JAZ12	Glyma17g04850, Glyma16g01220, Glyma07g04630	Repressor of jasmonate responses [Chini, et al. 2007]
	ZF-HD	7.44E-03	HB22, HB24, HB33	Glyma06g09970, Glyma04g09910, Glyma08g06120, Glyma11g07360, Glyma14g35770	Regulator of ABA-response [Wang, et al. 2011]
AB3	TCP	3.23E-05	TCP3, TCP20	Glyma09g42140, Glyma16g05840, Glyma19g26560, Glyma12g35720, Glyma06g34330, Glyma13g34690, Glyma03g02090, Glyma09g42120	Controls cell expansion and morphogenesis, negatively regulates auxin response, and promotes flavonoid synthesis [Herve, et al. 2009, Li and Zachgo 2013]
	SNF2	1.63E-03	EDA16, RGD3, PIE1, SYD, CHR5	Glyma07g31180, Glyma09g36910, Glyma02g29380, Glyma17g02540, Glyma13g25310, Glyma02g45000	Embryo sac development, repressor of flowering, chromatin remodeling, gravitrophism [Shaked, et al. 2006]
	GRF	8.68E-03		Glyma19g28010	Cotyledon growth [Kim, et al. 2003]
	CAMTA	4.18E-04	CAMTA3, SR1	Glyma05g31190, Glyma08g14370, Glyma17g04310	Negative regulator of plant immunity [Galon, et al. 2008, Du, et al. 2009]
AB4	TCP	9.25E-03	TCP3	Glyma13g34690, Glyma09g42120, Glyma09g42140	Negatively regulates auxin response, and promotes flavonoid synthesis [Li and Zachgo 2013]
	CAMTA	8.14E-04	CAMTA3, SR1	Glyma05g31190, Glyma08g14370	Negative regulator of plant immunity [Galon, et al. 2008, Du, et al. 2009]
AB5	MBF1	5.27E-03	MBF1B	Glyma06g42890	Negative regulator of ABA-dependent inhibition of germination [Di Mauro, et al. 2012]
	CAMTA	9.57E-03	CAMTA3, SR1	Glyma08g11080, Glyma08g14370	Negative regulator of plant immunity [Galon, et al. 2008, Du, et al. 2009]

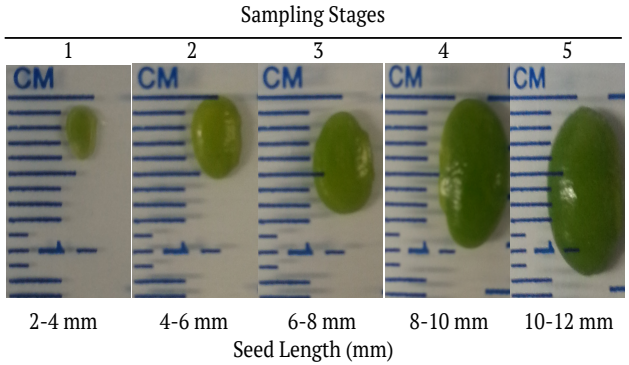


Figure 2.1: Seed developmental stages for sampling. Three biological replicates sampled per stage for both *lpa* and wildtype.

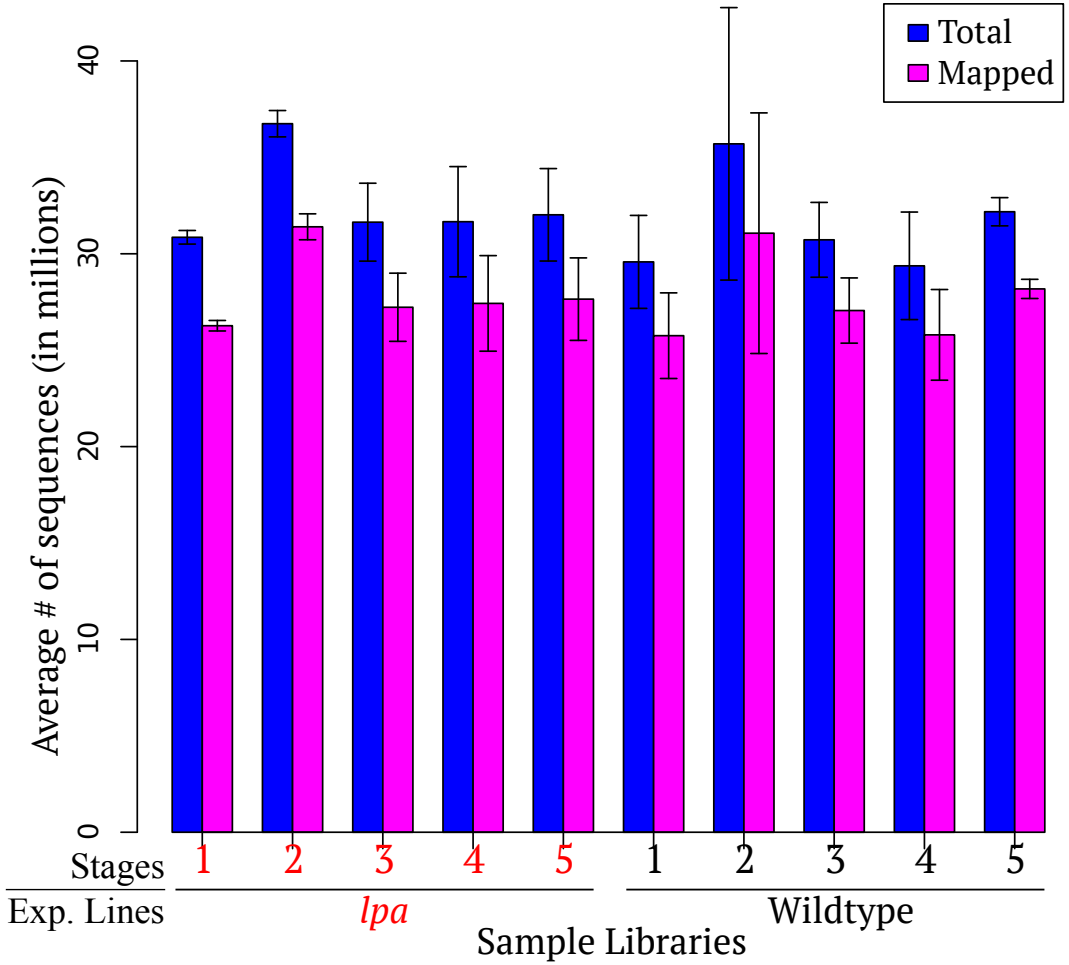
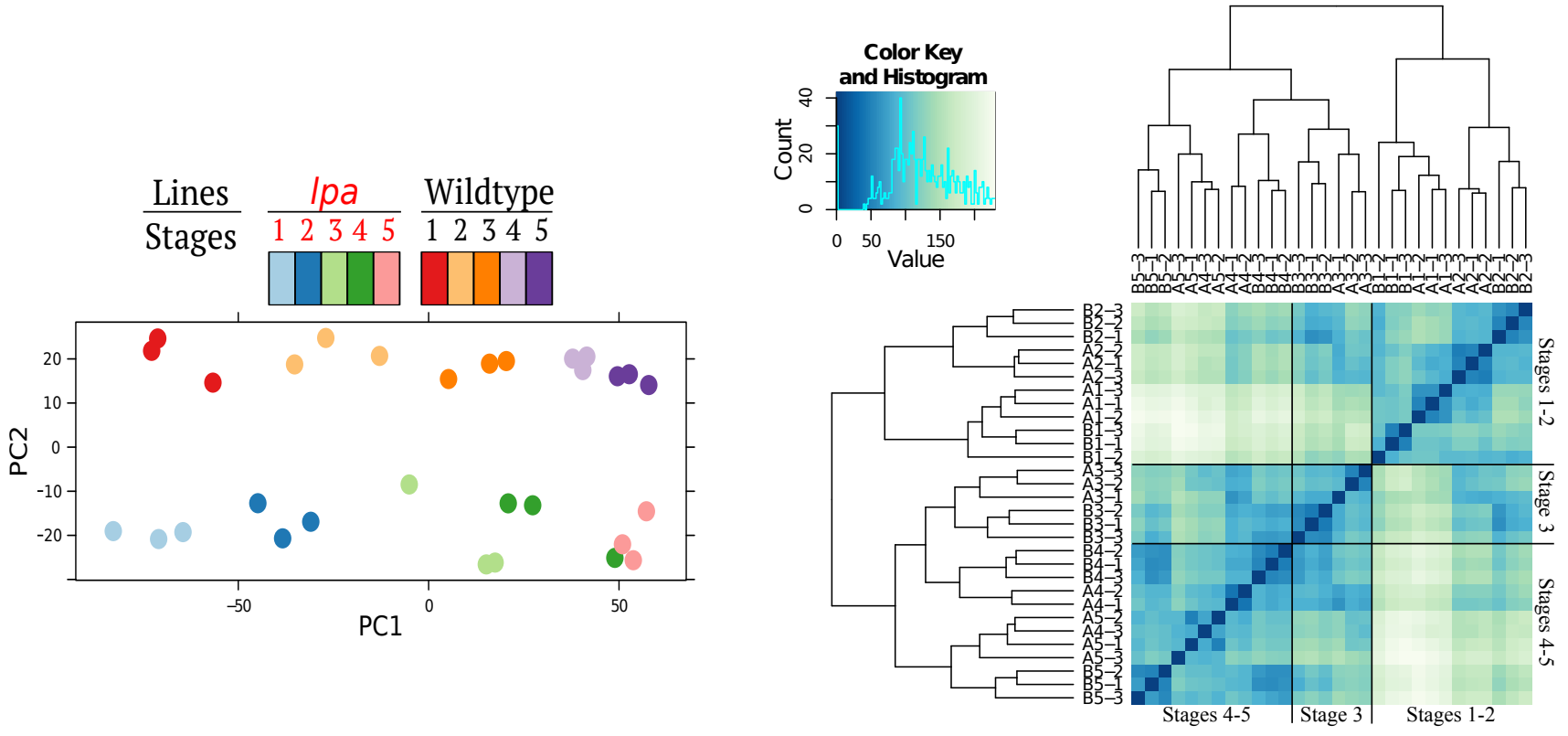


Figure 2.2: Alignment statistics. Average number of sequences generated and mapped to reference genome for each library. Total of 30 libraries were sequenced for this experiment. Error bars indicate standard error for biological replicates.



(a) (b)

Figure 2.3: Biological sample variability. (a) Principle component analysis (PCA) plot explains the variance in gene expression data from biological sample libraries along PC1 or X-axis and PC2 or Y-axis. (b) Sample clustering heat map representing sample-to-sample distance. Blue color suggests similarity between sample libraries. Samples A and B correspond to *lpa* mutant and wildtype, respectively, e.g. A3-2, means *lpa*-stage3-replicate2 (Refer Supplementary Table A1 for more information).

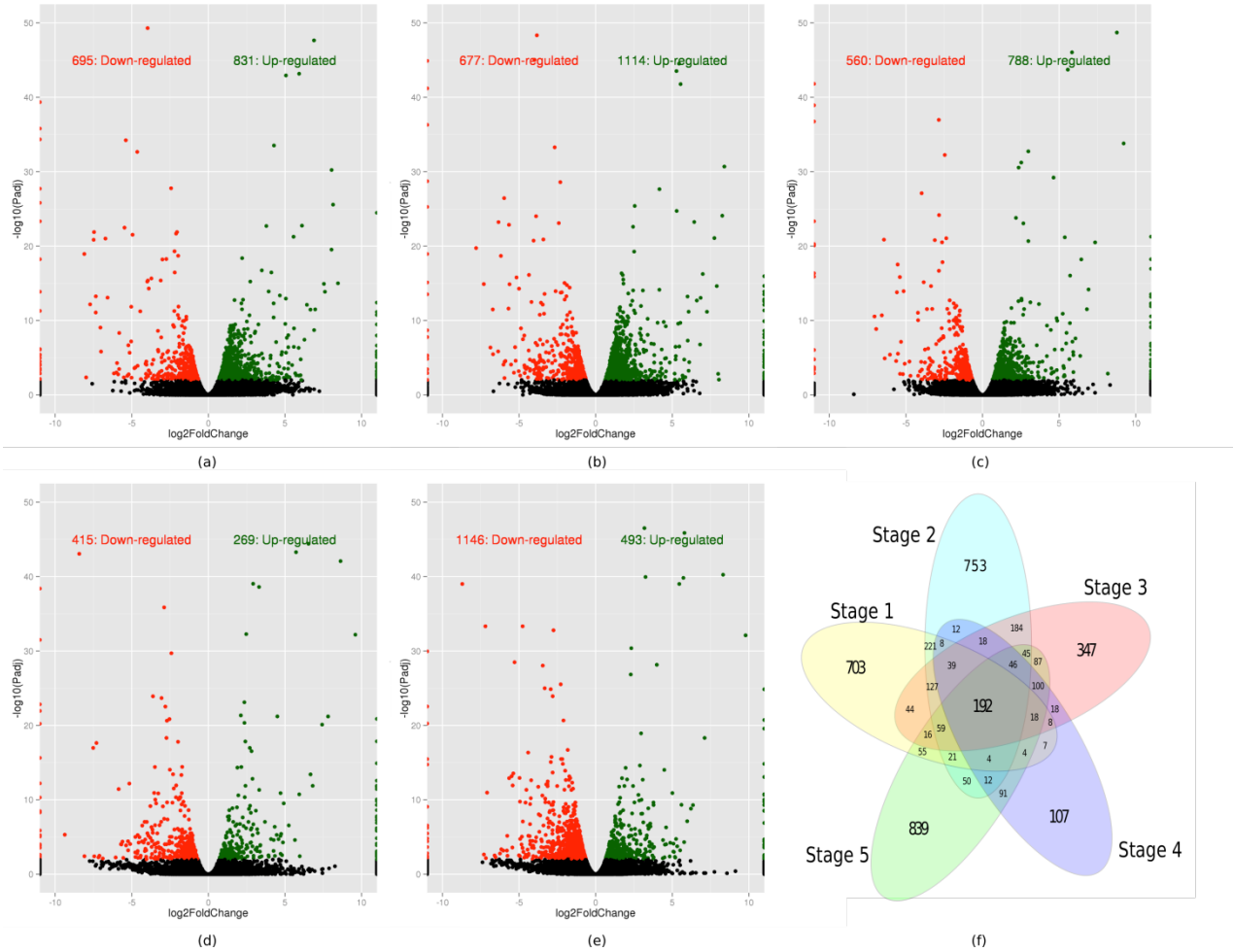


Figure 2.4: Differential gene expression. (a)-(e) Significant differentially expressed genes at 1% FDR in stages 1-5, respectively. X-axis represents log2FoldChange ratio, and Y-axis represents $-\log_{10}$(adjusted-P-value). Red indicates down regulated genes, while green indicates up regulated genes in *lpa.* (f) Overlap of differentially expressed genes between different stages.

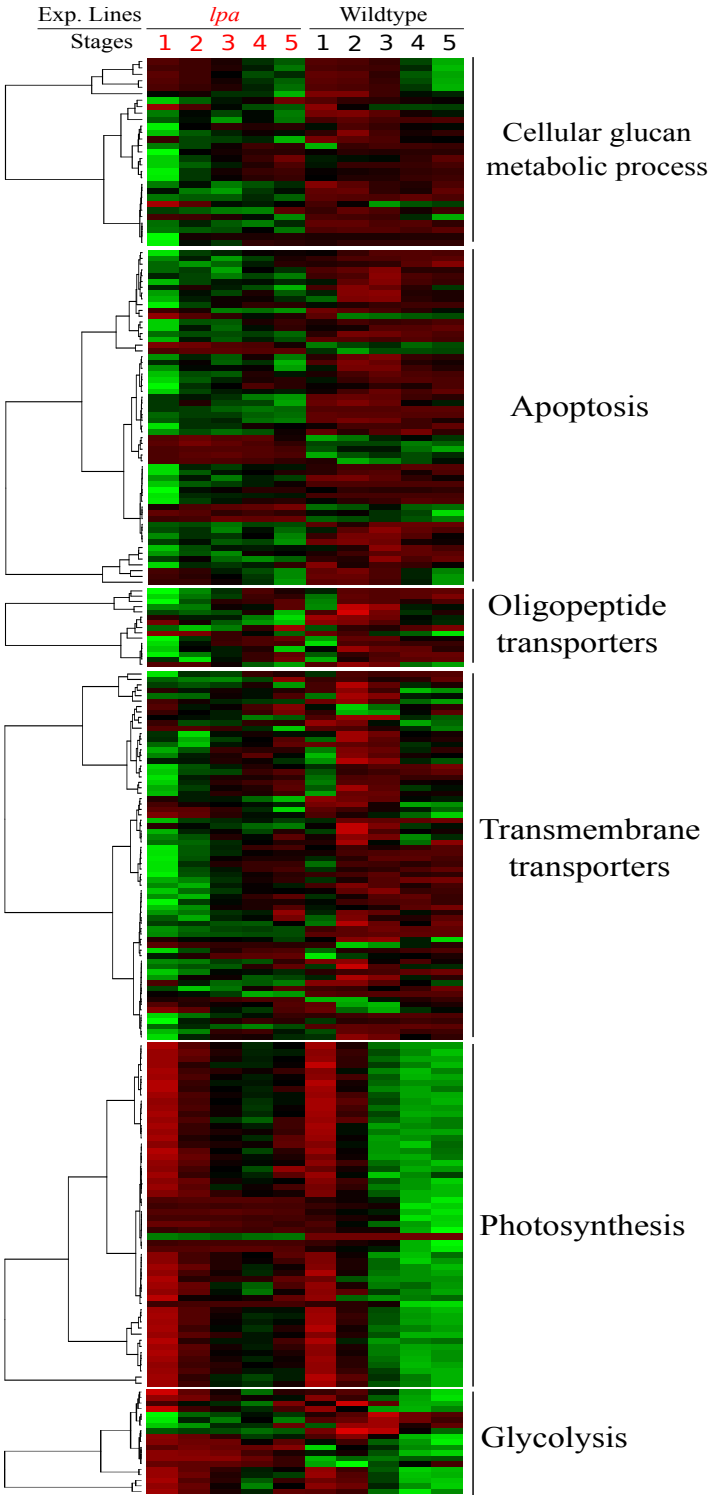
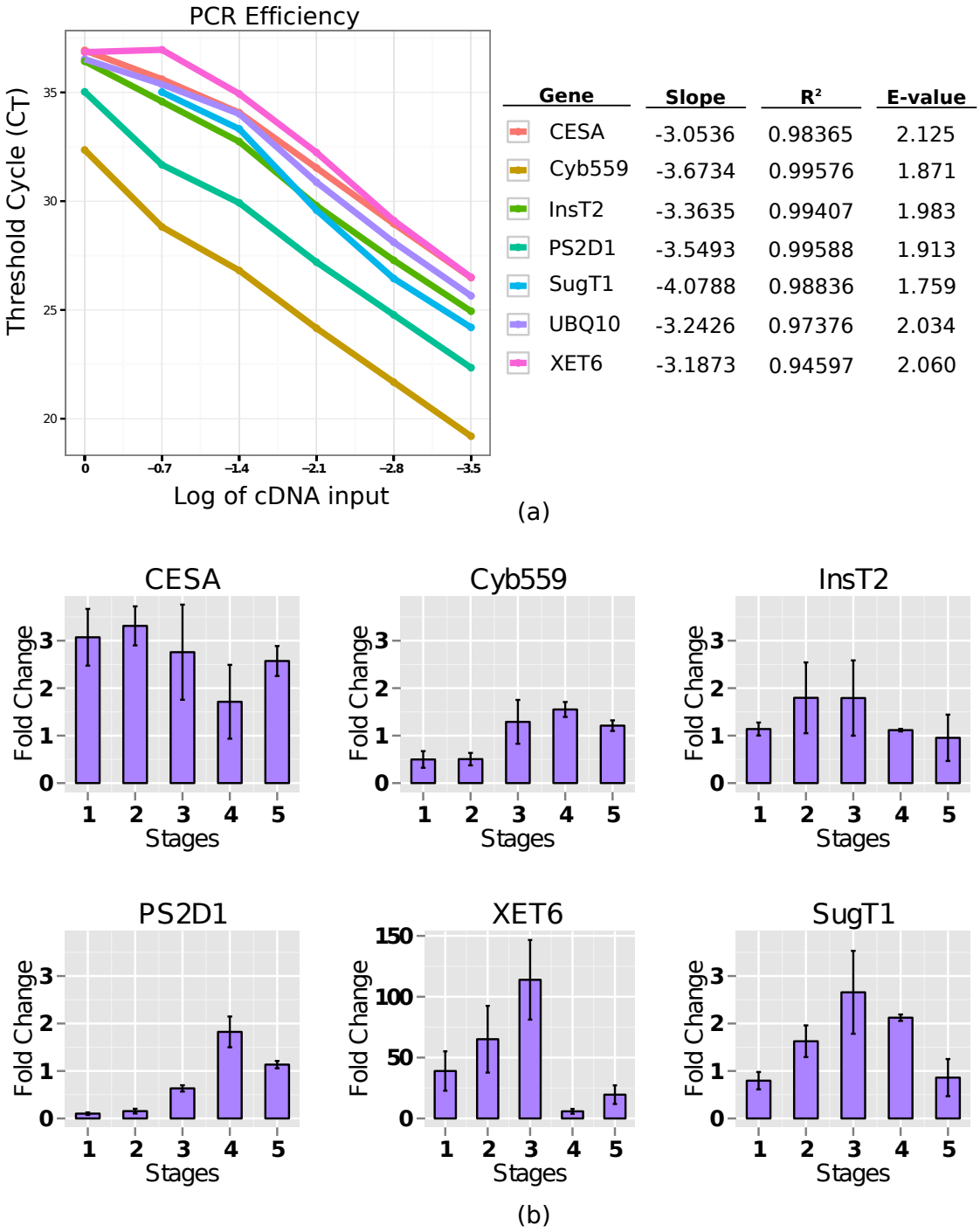


Figure 2.5: Mean normalized gene expression profiles of DEGs associated with different biological processes. Hierarchical clustering of mean normalized gene expression values based on euclidean distance between seed developmental stages of *lpa* and wildtype. Rows represent genes, while columns represent samples. Green color indicates higher gene expression values.
.



Gene	Slope	R^2	E-value
CESA	-3.0536	0.98365	2.125
Cyb559	-3.6734	0.99576	1.871
InsT2	-3.3635	0.99407	1.983
PS2D1	-3.5493	0.99588	1.913
SugT1	-4.0788	0.98836	1.759
UBQ10	-3.2426	0.97376	2.034
XET6	-3.1873	0.94597	2.060

Figure 2.6: Relative gene expression of DEGs for RNA-Seq data validation. (a) Standard calibration curve for estimating PCR efficiency; (b) Fold change in gene expression levels between mutant and wildtype at respective seed developmental stages estimated using quantitative real-time PCR. Housekeeping gene, UBQ10 was used as a reference. Data was analyzed using Pffafl method, and is represented as a fold change expression in mutant as compared to wildtype at respective stages.

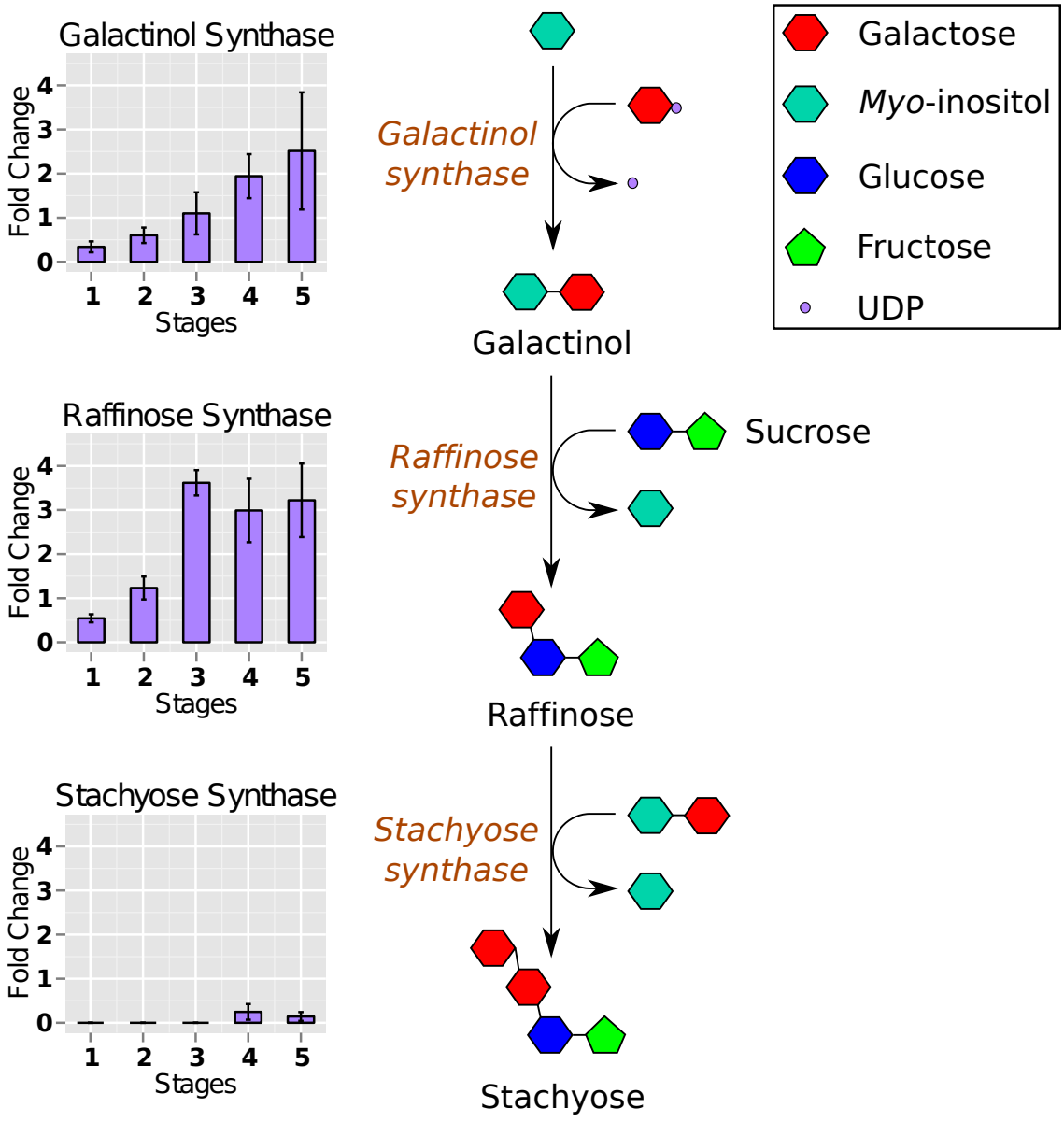


Figure 2.7: Raffinose Family Oligosaccharide biosynthesis pathway. Bar graphs indicate fold change ratio of normalized gene expression values from *lpa* mutant over wildtype for 5 stages of seed development.

CHAPTER 3

Regulation of metabolic genes in seed development process of low phytic acid soybeans

Neelam R Redekar, Song Li, M. A. Saghai Maroof[§]

Department of Crop and Soil Environmental Sciences, Virginia Tech. [§]Corresponding author: smaroof@vt.edu

This Chapter is to be submitted for publication in *BMC Genomics*.

ABSTRACT

The seed development is a complex metabolic process, which involves both biosynthesis of storage seed reserves and catabolism of energy reserves for growth and differentiation of the embryo. Low phytic acid-causing mutations disturb the seed development, ultimately resulting in low emergence. It is therefore important to identify the regulatory genes that affect seed development in low phytic acid crops. We performed a comparative metabolic process gene co-expression network analysis with two different low phytic acid soybean mutants and their respective wildtypes. We identified co-expression modules whose gene expression profiles were significantly correlated with the progression of seed development. The positively correlated modules in wildtype networks were enriched with genes associated with lipid biosynthesis, and amino acid synthesis, whereas the negatively correlated modules in mutant networks were enriched with genes related to glutamate metabolism. The network comparisons between mutants and wildtypes identified the regulatory nodes that are associated with lipid and amino acid metabolism, sugar signaling, and biosynthesis of secondary metabolites. Two genes associated with peroxisomal beta-oxidation pathway: Glyma05g28390, encoding acyl-activating enzyme 16 (AAE16) and Glyma03g27360, encoding dienoyl-CoA-isomerase 1 (DCI1), were represented in the co-expression networks of both mutants, but not in their respective wildtypes. It is highly likely that AAE16 and DCI1 could be potential targets of low phytic acid-causing mutations in regulating the seed storage reserves during seed development.

KEYWORDS

Co-expression network, metabolic genes, seed development, beta-oxidation pathway, phytic acid, *myo*-inositol phosphate synthase, multidrug-resistance protein ABC transporter

INTRODUCTION

Seed development is mainly comprised of three phases: (1) cell division phase (post fertilization), when tissue differentiates to form embryo axis and cotyledons, (2) cell expansion phase, when seed storage reserves start accumulating, and (3) maturation phase, when all storage reserves are immobilized. The seed development is a complex metabolic process, which involves both synthesis and breakdown of macromolecules for growth and maintenance of the embryo [Weber, et al. 2005, Le, et al. 2007]. The metabolic activities that result in the accumulation of seed storage reserves, such as proteins and lipids, are also crucial for seed viability and germination [Weber, et al. 2005]. During germination, before the onset of photosynthesis, seeds are completely dependent on storage reserves for nutrition. Transcription of metabolic genes plays a major role in determining the seed composition, and hence its germinability. Therefore, the expression of genes that are involved in these metabolic activities are tightly regulated by synergistic action of several transcription factors and other regulatory genes [Weber, et al. 2005, Le, et al. 2007]. Several studies have explored the transcriptome status for seed development in Arabidopsis [Ruuska, et al. 2002], barley [Watson and Henry 2005], soybean [Severin, et al. 2010, Song, et al. 2011, Shamimuzzaman and Vodkin 2012, Collakova, et al. 2013, Jones and Vodkin 2013, Li, et al. 2015], and *Brassica napus* [Li, et al. 2005, Fei, et al. 2007]. Findings from these studies have enabled us to more fully understand the transcriptional regulation of the seed developmental processes.

The phytic acid biosynthesis is one of the key regulators of seed development [Raboy 1997]. In this pathway, glucose-6-phosphate is converted to *myo*-inositol, an intracellular signaling molecule, which is hexa-phosphorylated to form phytic acid [Raboy 1997]. Mutations

that block this pathway have shown to alter the seed metabolite levels in soybean, rice, maize, et cetera [Wilcox, et al. 2000, Shi, et al. 2003, Shi, et al. 2005, Stevenson-Paulik, et al. 2005, Raboy 2007, Glover 2011, McClellan, et al. 2012]. For example, a mutation in the gene encoding *myo*-inositol phosphate synthase (MIPS) enzyme results in reduced phytic acid, stachyose, raffinose, elevated sucrose, and low seed emergence in soybean [Maroof and Buss 2008]. Other non-pathway genes, such as the multi-drug resistance protein (MRP) genes encoding ATP binding cassette transporters, that are believed to be involved in transport of phytic acid to storage vacuoles, are also known to regulate metabolite levels during seed development and affect seed emergence [Shi, et al. 2007, Maroof, et al. 2009, Nagy, et al. 2009, Xu, et al. 2009, Kastl 2014]. The seeds with reduced phytic acid content, on the other hand, are commercially more valuable, as their consumption tends to reduce both mineral deficiencies in monogastric animals and phosphorus pollution (Raboy 2007). Hence, deciphering the regulation of seed development in low phytic acid (*lpa*) crops has gained more attention. Two independent studies, one with a barley *lpa* mutant, and another with a soybean *mips1/mrp-l/mrp-n* triple mutant have reported the effect of *lpa* mutations on the transcriptome profiles of developing seeds [Bowen, et al. 2007, Redekar, et al., unpublished]. Redekar et al. reported substantial gene expression turnover of transcription factor families, suggesting a complex regulatory mechanism associated with the phytic acid biosynthesis pathway [Redekar, et al., unpublished]. Discovery of genes that play roles in such regulatory mechanisms requires more sophisticated approaches than differential gene expression determination. Study of co-expression gene networks may help to elucidate the regulated processes in developing *lpa* seeds.

In this study, we harness transcriptome information to understand the effect of low phytic acid mutations on the co-regulation of genes involved in metabolic processes during seed

development. We used two *lpa* mutants, including a *mips1/mrp-l/mrp-n* triple mutant (Redekar et al unpublished) and a *mips1* single mutant and compared them with the respective wildtypes. The Gene Ontology defines "Metabolic Process" (GO:0008152) as: "*The chemical reactions and pathways, including anabolism and catabolism, by which living organisms transform chemical substances. Metabolic processes typically transform small molecules, but also include macromolecular processes such as DNA repair and replication, and protein synthesis and degradation"* [www.geneontology.org]. We constructed weighted gene correlation networks to identify the co-expressed genes involved in the metabolic processes. Functional enrichment, and hub gene analysis were performed to identify enriched metabolic processes and top regulatory nodes. The regulatory genes involved in lipid and amino acid metabolism, sugar signaling, and biosynthesis of secondary metabolites were found highly co-expressed in our data. We identified key genes involved in peroxisomal beta-oxidation of fatty acids, which were represented in the co-expression networks of developing seeds in both mutants, as opposed to both wildtypes.

MATERIAL AND METHODS

Genetic material, sampling and sequencing of developing seeds

Four soybean experimental lines designated as: (i) *3mlpa*, (ii) 3MWT, (iii) *1mlpa*, and (iv) 1MWT were used in this study. The *lpa* mutant line*,* "*3mlpa*" carrying three mutations *mips1/mrp-l/mrp-n,* and its corresponding normal phytic acid line, "3MWT" were derived from the crossing of 'CX-1834' (*lpa* line with two *mpr-l/mrp-n* mutations) with 'V99-5089' (*lpa* line with single *mips1* mutation) [Maroof, et al. 2009]. Another *lpa* line, "*1mlpa*" carrying a single *mips1* mutation, and its corresponding normal phytic acid line, "1MWT" were derived from the crossing of 'Essex' (a normal phytic acid line with no mutations) with V99-5089 [Glover 2011].

The experimental lines, *1mlpa* and 1MWT were grown under the controlled growth chamber environment as described in Redekar et al. [Redekar, et al., unpublished]. Developing seeds were sampled in triplicates for each experimental line based on seed lengths: Stage 1 (between 2-4 mm), Stage2 (between 4-6 mm), Stage3 (between 6-8 mm), Stage4 (between 8-10 mm), and Stage 5 (between 10-12 mm). The total RNA was extracted from these developing seed tissues as described in Redekar et al. [Redekar, et al., unpublished]. A total of 30 mRNA libraries were prepared and sequenced as 100SE using HiSeq2000. The transcriptome sequencing data for 3mlpa and 3MWT was obtained from Redekar et al. [Redekar, et al., unpublished]. Each dataset was comprised of sequencing reads generated from 15 libraries, representing 5 stages of developing seed.

Sequencing data analysis

The mRNA sequencing reads generated from 5 developing seed stages with varying seed lengths for four experimental lines were mapped to the soybean reference, 'Williams82' genome (Wm82.a1.v1.1), using TopHat2 [Kim, et al. 2013]. The read count, which is a measure of gene expression, was estimated using htseq-count on "Union" mode [Anders, et al. 2014]. This gene expression data set was further normalized across all the samples using the DESeq normalization method [Anders and Huber 2010]. The 1679 genes from the soybean reference genome are classified under the gene ontology (GO) category "Metabolic Process" (GO:0008152), and were used for generating a co-expression network of the metabolic genes. Hereafter, these genes will be referred to as "metabolic process genes." Log-transformed normalized expression profiles of these genes were arranged in the form of a matrix, where each row corresponds to a gene and a

column corresponds to a sequencing library sample. This expression data matrix was used as an input for co-expression analysis.

Weighted gene co-expression network analysis (WGCNA)

The co-expression analysis was performed using the WGCNA package (v1.46) installed in R (v3.2.6 – "Full of Ingredients") [Langfelder and Horvath 2008]. The input data was tested to remove genes with zero or no expression and/or any outliers in the sample libraries that were used for network construction. One undirected weighted co-expression gene network with power (soft thresholding) $\beta = 12$ was constructed for each experimental line. A node within these networks represented a gene, while an edge represented the correlation between two genes. Highly interconnected genes exhibiting similar expression profiles were clustered together to form modules. These gene co-expression modules were further correlated to the seed developmental stages to search for biologically interesting modules. The modules whose eigengene expression profiles significantly increased or decreased with respect to seed development were used for further analysis. The hub-genes (or highly connected genes) within each significant module were also identified. The co-expression gene network was visualized using the VisANT (v5.25) tool [Hu, et al. 2008]. The genes present within these significant modules were then used for functional enrichment with a hypergeometric test at 1% false discovery rate (FDR), using the Benjamini Hochberg method to identify biologically meaningful metabolic processes using AgriGO tools [Du, et al. 2010].

RESULTS AND DISCUSSION

Metabolic gene co-expression network in developing soybean seeds

In this study, we have performed a comparative metabolic gene co-expression network analysis in developing soybean seeds from two mutants (*3mlpa* and *1mlpa*), with *lpa*-causing mutations, and the respective wildtypes (3WT and 1WT) with no mutations. Towards this goal, we sampled five developing seed stages based on whole seed length from four experimental lines—*3mlpa*, *1mlpa*, 3MWT, and 1MWT. The seed developmental phase that best describes our sampling stages is the cell expansion phase, when cotyledons are producing storage nutrient reserves and metabolic activities are at their peak. The mRNA sequencing data from these experimental lines was analyzed to obtain a normalized read count for genes, as a measure of gene expression. Summary statistics for the sequence alignment is as shown in Supplementary Table B1. More than 82% of the mRNA sequencing reads per library were mapped to the reference genome, Williams 82. The gene expression values were estimated based on the number of reads that uniquely mapped to the gene. For generating co-expression networks, only the normalized gene expression data from each mRNA sample library of each experimental line, representing the three biological replicates of 5 seed developmental stages, was used. This only included 1679 soybean genes, which were categorized as "metabolic process genes." The principal component (PC) analysis was performed on this data. At stage 1, all samples are clustered together, indicating highly similar expression profiles (Figure 3.1). From stage 2 onwards, the expression profiles start to vary. Interestingly, at stage 3 of seed development, samples originating from both mutant lines (*3mlpa* and *1mlpa*) cluster together, indicating similar metabolic gene expression profiles. Samples were clustered as per seed developmental stages along PC1, explaining more than 95% variation in this data (Figure 3.1). This means that

the major fluctuations in the metabolic process gene expression are due to progression of seed development, as expected. We also observed that the samples originating from different experimental lines gain more variation in metabolic process gene expression profiles starting from the second stages of seed development, which explains more than 3% of the variation in the data along PC2.

We used two strategies for co-expression network analysis. First, we generated a separate network for each of the four experimental lines and identified co-expressed genes that are shared by two mutant networks, as opposed to two wildtype networks (hereafter, this strategy will be referred to as "individual network"). For construction of these networks, we used metabolic process gene expression data from 15 sample libraries generated for each experimental line. Second, we generated a network combining two mutant lines (*3mlpa* + *1mlpa*), and another network combining two wildtype lines (3MWT + 1MWT), and identified genes that are unique to the mutant network, as opposed to the wildtype network (hereafter, this strategy will be referred to as "combined network"). For construction of these combined networks, we used metabolic process gene expression data from 30 sample libraries generated from two experimental lines. Prior to network construction, we eliminated the outlier genes, resulting in more than 1545 (92%) metabolic process genes being available for analysis in each case (Table 3.1). The log-transformed normalized gene expression profiles of these metabolic genes were used to construct an undirected weighted gene co-expression network using correlations between all pairs of genes as a measure of co-expression. We identified co-expression modules, comprised of groups of genes that all share similar expression patterns (Table 3.1). In individual networks, we identified 4, 6, 3, and 3 co-expression gene modules for *3mlpa*, 3MWT, *1mlpa*, and 1MWT, respectively; whereas as in combined networks, we identified 4 and 5 co-expression

gene modules for mutants (*3mlpa* + *1mlpa*) and wildtypes (3MWT + 1MWT), respectively (Table 3.1). Based on a decreasing number of genes, the modules are referred to as turquoise, blue, brown, yellow, green, and red. So for each network, the module with the maximum number of genes is referred to as the "turquoise" module. The module-specific eigengene summarizes the expression profile of genes within a module (Supplementary Figure C1). Three types of co-expression trends were prominently observed for modules in all networks: (1) co-expressed genes *increase* their expression, (2) co-expressed genes *decrease* their expression, and (3) co-expressed genes *increase and then decrease* their expression (like a parabola) during seed development (Table 3.1). For this paper, we will focus only on the co-expression modules, which show the highest significant positive (trend #1) and negative (trend #2) correlation with seed development. These positively correlated modules are brown, blue, blue, brown, brown, blue; and the negatively correlated modules are turquoise, turquoise, turquoise, blue, turquoise, and turquoise for *3mlpa*, 3MWT, *1mlpa*, 1MWT, *3mlpa* + *1mlpa*, and 3MWT + 1MWT, respectively (Supplementary Figure C1). The gene expression profiles of these significantly correlated modules are as shown in Figure 3.3 and Supplementary Figure C2.

Functional enrichment of modules

The functional enrichment analysis of the co-expressed gene modules resulted in several enriched metabolic processes significant at 1% FDR (Table 3.2). Since all the negatively correlated modules consisted of more genes than the positively correlated modules, the number of enriched metabolic processes was higher in negatively correlated modules. The positively and negatively correlated gene modules showed 2 and 14 enriched metabolic processes, respectively (Table 3.2). The lipid biosynthetic process (GO:0008610) was represented in positively

correlated modules of wildtype experimental lines (3WT and 1WT), but not in mutant lines (*3mlpa* and *1mlpa*) (Table 3.2). In other words, the lipid biosynthesis-related gene co-expression profiles increase with seed development in both wildtypes (3WT and 1WT), but not in mutants (*3mlpa* and *1mlpa*). Similarly, the cellular amino acid biosynthesis process (GO:0008652) was associated with the positively correlated module of 3MWT + 1MWT wildtypes combined network, but not *3mlpa* + *1mlpa* mutant network (Table 3.2). The glutamate metabolic process (GO:0006536), on the other hand, was represented in negatively correlated modules of mutant (*3mlpa* and *1mlpa*) networks, but not in wildtype (3WT and 1WT) networks (Tables 3.2). Table 3.3 lists all the genes associated with the lipid biosynthetic process, cellular amino acid biosynthesis process, and glutamate metabolic process. Since the seed developmental stages in this study correspond to the phase of accumulating seed storage reserves, such as lipids and proteins, it is obvious to expect for lipid and amino acid metabolism. However, the enrichment of these metabolic processes in the wildtype, but not in the mutant modules, suggests that different sets of genes are being co-expressed. For example, NADH-dependent glutamate synthase 1 (GLT1) encoding genes are common to both the cellular amino acid biosynthesis process, and glutamate metabolic process GO categories. However, the set of GLT1-encoding genes associated with positively correlated modules of 3MWT + 1MWT wildtype network were different than the ones associated with negatively correlated modules of mutant (*3mlpa* and *1mlpa*) networks (Table 3.3). These observations suggest that the *lpa*-causing mutations appear to regulate glutamate metabolism, lipid- and amino acid-biosynthesis during seed development.

Regulatory nodes of co-expression module

The most important regulatory nodes are usually highly connected [Langfelder and Horvath 2008]. A hub is a gene connected to the majority of the nodes within a module, suggesting its role in regulating different biological processes at the same time. Therefore, the knowledge of hub genes and their neighboring nodes in the co-expression module can facilitate identification of pathways that regulate metabolic processes involved in the seed development. We identified highly co-expressed hub genes for each module (Table 3.4). The hub genes encoding for formate dehydrogenase (FDH), aldehyde dehydrogenase 5F1 (ALDH5F1), threonine synthase, hydroxysteroid dehydrogenase 1 (HSD1), and phosphoribulokinase (PRK) were identified in positively correlated modules; while those encoding for cysteine desulfurase (NFS1), UDP-glucose 4-epimerase 1 (UGE1), dienoyl-CoA isomerase (DCI1), and calcium transporting ATPase (ACA1) were identified in negatively correlated modules. These hub genes are mainly involved in carbon metabolism, steroid biosynthesis, and amino acid metabolism (www.kegg.jp). Several of these proteins have been previously studied in context with seed development. For example, FDH is involved in the removal of toxic compounds associated with the stress response, and is known to be expressed during seed development [Li, et al. 2002, David, et al. 2010, Hajduch, et al. 2010]. DCI1 is involved in degradation of unsaturated fatty acids via beta-oxidation and is associated with seed germination [Goepfert, et al. 2005], while HSD1 is expressed during the seed filling stage and is associated with seed dormancy and germination [Li, et al. 2007, Baud, et al. 2009]. UGE1 is involved in cell wall carbohydrate biosynthesis [Rosti, et al. 2007], while other hub genes are required for growth and development of plants. Most hub genes were distinct between mutant and wildtype network modules, except for FDH, suggesting different regulation pathways. Two distinct genes, Glyma13g23790 and

Glyma19g01210, encoding for FDH were hubs in positively correlated modules of 3mlpa and 3MWT. This does not necessarily mean that FDH is connected to the same set of neighboring nodes in mutants and wildtype.

We therefore selected the top 30 nodes, including the hub gene based on connectivity and edge weights for each module. Several of these top regulatory nodes in positively and negatively associated modules were shared between mutants and wildtypes. Exactly 104 distinct nodes were identified as top regulatory nodes combining modules from all networks (Supplementary Table B2). These nodes represent the key regulatory factors that influence metabolic processes during seed development. Since our interest lies in observing the effect of low phytic acid-causing mutations on this regulatory pathway, we focused on the nodes that are present in either the mutant or wildtype network (Table 3.5). In case of individual network modules, we identified the nodes that are present in both mutants and absent in both wildtypes, and vice versa (Table 3.5). In individual network modules, *3mlpa* and *1mlpa* mutants represented regulatory nodes encoding for phospholipase D delta (PLDDELTA, Glyma11g08640), acyl-activating enzyme 16 (AAE16, Glyma05g28390), and DCI1 (Glyma03g27360); whereas wildtypes indicated the haloacid dehalogenase-like hydrolase superfamily protein (Glyma10g27980) (Table 3.5). In modules of combined networks, 30 distinct regulatory nodes were represented. These included several regulatory nodes encoding for proteins involved in amino acid and lipid metabolism, and sugar signaling in networks for both mutants and wildtypes. Within lipid metabolism, the nodes were specifically associated with fatty acid degradation via the beta-oxidation pathway, fatty acid elongation, glycerophospholipid metabolism, and translocation of phospholipids. The node encoding for PLDDELTA was observed in the mutant and wildtype for individual and combined networks, respectively, and therefore may not be a critical node (Table 3.5). On the contrary,

two regulatory nodes encoding for AAE16 and DCI1 were represented by mutants in both individual and combined networks, but not in respective wildtypes (Table 3.5). Both of these enzymes are involved in beta-oxidation of unsaturated fatty acids. Beta-oxidation of fatty acids takes place in peroxisomes and is required for embryo development and seed germination [Rylott, et al. 2003, Cassin-Ross and Hu 2014]. During seed development, fatty acids derived from maternal tissue are broken down via beta-oxidation to obtain energy for embryo development. When seeds reach the filling stage, the beta-oxidation activity decreases and triglycerides are accumulating as nutrient reserves. Storage lipids then undergo beta-oxidation to support seed germination [Goepfert, et al. 2005]. Presence of beta-oxidation nodes in both mutants, as opposed to wildtypes, suggests this as the mode of regulation of seed development in *lpa* soybeans. Although further studies are required to validate these findings, the genes identified in this study can serve as prime candidates for physiological studies on elaborating our understanding of regulatory pathways associated with *lpa* mutations.

CONCLUSION

Seed development is a complex metabolic process where nutrients required for establishment of progeny are produced and stored. Low phytic acid mutations tend to affect seed metabolite levels and reduce seedling emergence. In order to study the effect of *lpa*-causing mutations on the regulation of seed development, we constructed weighted gene co-expression networks for two different *lpa* mutants (*3mlpa* and *1mlpa*), and their respective wildtype lines. The metabolic processes, such as cellular amino acid biosynthesis and lipid biosynthesis, were found enriched in two wildtypes; whereas the glutamate metabolism was enriched in two mutants. Comparison of top regulatory nodes between the mutants' and wildtypes' network also

identified key genes that potentially play a role in regulation of seed development with respect to *lpa* mutations in soybean.

COMPETING INTERESTS

Authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

NRR performed the experiment, analyzed the data. MASM, SL, and NR contributed to preparation of manuscript. MASM and SL guided the process and reviewed the manuscript.

ACKNOWLEDGEMENTS

This work was funded by USDA-NIFA grant via Bio-design and Bioprocessing Research Center (BBRC) at Virginia Tech. We would like to thank support team of Advanced Research Computing (ARC) server and Translational Plant Sciences' MAGYK server at Virginia Tech.

REFERENCES

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

Anders S, Pyl PT, Huber W: **HTSeq – A Python framework to work with high-throughput sequencing data**. 2014.

Baud S, Dichow NR, Kelemen Z, d'Andrea S, To A, Berger N, Canonge M, Kronenberger J, Viterbo D, Dubreucq B *et al*: **Regulation of HSD1 in seeds of Arabidopsis thaliana**. *Plant Cell Physiol* 2009, **50**(8):1463-1478.

Bowen DE, Souza EJ, Guttieri MJ, Raboy V, Fu J: **A Low Phytic Acid Barley Mutation Alters Seed Gene Expression**. *Crop Science* 2007, **47**(S2):S-149.

Cassin-Ross G, Hu J: **Systematic phenotypic screen of Arabidopsis peroxisomal mutants identifies proteins involved in beta-oxidation**. *Plant physiology* 2014, **166**(3):1546-1559.

Collakova E, Aghamirzaie D, Fang Y, Klumas C, Tabataba F, Kakumanu A, Myers E, Heath LS, Grene R: **Metabolic and Transcriptional Reprogramming in Developing Soybean (Glycine max) Embryos**. *Metabolites* 2013, **3**(2):347-372.

David P, des Francs-Small CC, Sevignac M, Thareau V, Macadre C, Langin T, Geffroy V: **Three highly similar formate dehydrogenase genes located in the vicinity of the B4 resistance gene cluster are differentially expressed under biotic and abiotic stresses in Phaseolus vulgaris**. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik* 2010, **121**(1):87-103.

Du Z, Zhou X, Ling Y, Zhang Z, Su Z: **agriGO: a GO analysis toolkit for the agricultural community**. *Nucleic acids research* 2010, **38**(Web Server issue):W64-70.

Fei H, Tsang E, Cutler AJ: **Gene expression during seed maturation in Brassica napus in relation to the induction of secondary dormancy**. *Genomics* 2007, **89**(3):419-428.

Glover N: **The Genetic Basis of Phytate, Oligosaccharide Content, and Emergence in Soybean**. Blacksburg, VA: Virginia Tech; 2011.

Goepfert S, Vidoudez C, Rezzonico E, Hiltunen JK, Poirier Y: **Molecular identification and characterization of the Arabidopsis delta(3,5),delta(2,4)-dienoyl-coenzyme A isomerase, a peroxisomal enzyme participating in the beta-oxidation cycle of unsaturated fatty acids**. *Plant physiology* 2005, **138**(4):1947-1956.

Hajduch M, Hearne LB, Miernyk JA, Casteel JE, Joshi T, Agrawal GK, Song Z, Zhou M, Xu D, Thelen JJ: **Systems analysis of seed filling in Arabidopsis: using general linear modeling to assess concordance of transcript and protein expression**. *Plant physiology* 2010, **152**(4):2078-2087.

Hu Z, Snitkin ES, DeLisi C: **VisANT: an integrative framework for networks in systems biology**. *Brief Bioinform* 2008, **9**(4):317-325.

Jones SI, Vodkin LO: **Using RNA-Seq to profile soybean seed development from fertilization to maturity**. *PLoS One* 2013, **8**(3):e59270.

Kastl C: **Metabolomic Discrimination of Near Isogenic Low and High Phytate Soybean [*GLYCINE MAX (L.) MERR.*] Lines**. Blacksburg, VA: Virginia Tech; 2014.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome biology* 2013, **14**(4):R36.

Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC Bioinformatics* 2008, **9**:559.

Le BH, Wagmaister JA, Kawashima T, Bui AQ, Harada JJ, Goldberg RB: **Using genomics to study legume seed development**. *Plant physiology* 2007, **144**(2):562-574.

Li F, Asami T, Wu X, Tsang EW, Cutler AJ: **A putative hydroxysteroid dehydrogenase involved in regulating plant growth and development**. *Plant physiology* 2007, **145**(1):87-97.

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

Li L, Hur M, Lee JY, Zhou W, Song Z, Ransom N, Demirkale CY, Nettleton D, Westgate M, Arendsee Z *et al*: **A systems biology approach toward understanding seed composition in soybean**. *BMC Genomics* 2015, **16 Suppl 3**:S9.

Li R, Moore M, Bonham-Smith PC, King J: **Overexpression of formate dehydrogenase inArabidopsis thaliana resulted in plants tolerant to high concentrations of formate**. *Journal of Plant Physiology* 2002, **159**(10):1069-1076.

Maroof AS, Buss GR: **Low phytic acid, low stachyose, high sucrose soybean lines**. In.: Google Patents; 2008.

Maroof MAS, Glover NM, Biyashev RM, Buss GR, Grabau EA: **Genetic Basis of the Low-Phytate Trait in the Soybean Line CX1834**. *Crop Science* 2009, **49**(1):69.

McClellan MS, Domier LL, Bailey RC: **Label-free virus detection using silicon photonic microring resonators**. *Biosensors & Bioelectronics* 2012, **31**(1):388-392.

Nagy R, Grob H, Weder B, Green P, Klein M, Frelet-Barrand A, Schjoerring JK, Brearley C, Martinoia E: **The Arabidopsis ATP-binding cassette protein AtMRP5/AtABCC5 is a high affinity inositol hexakisphosphate transporter involved in guard cell signaling and phytate storage**. *The Journal of biological chemistry* 2009, **284**(48):33614-33622.

Raboy V: **Accumulation and storage of phosphate and minerals.** . In: *Cellular and Molecular Biology of Plant Seed Development* Dordrecht, Netherlands: Kluwer Academic Publishers; 1997: 441-477.

Raboy V: **The ABCs of low-phytate crops**. *Nat Biotechnol* 2007, **25**(8):874-875.

Redekar N, Biyashev R, Jensen R, Helm R, Grabau E, Maroof SMA: **Genome-wide transcriptome analysis of developing seeds from low and normal phytic acid soybean lines**. In.; unpublished.

Rosti J, Barton CJ, Albrecht S, Dupree P, Pauly M, Findlay K, Roberts K, Seifert GJ: **UDP-glucose 4-epimerase isoforms UGE2 and UGE4 cooperate in providing UDP-galactose for cell wall biosynthesis and growth of Arabidopsis thaliana**. *The Plant cell* 2007, **19**(5):1565-1579.

Ruuska SA, Girke T, Benning C, Ohlrogge JB: **Contrapuntal Networks of Gene Expression during Arabidopsis Seed Filling**. *The Plant cell* 2002, **14**(6):1191-1206.

Rylott EL, Rogers CA, Gilday AD, Edgell T, Larson TR, Graham IA: **Arabidopsis mutants in short- and medium-chain acyl-CoA oxidase activities accumulate acyl-CoAs and reveal that fatty acid beta-oxidation is essential for embryo development**. *The Journal of biological chemistry* 2003, **278**(24):21370-21377.

Severin AJ, Woody JL, Bolon YT, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE *et al*: **RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome**. *BMC Plant Biol* 2010, **10**:160.

Shamimuzzaman M, Vodkin L: **Identification of soybean seed developmental stage-specific and tissue-specific miRNA targets by degradome sequencing**. *BMC Genomics* 2012, **13**:310.

Shi J, Wang H, Hazebroek J, Ertl DS, Harp T: **The maize low-phytic acid 3 encodes a myo-inositol kinase that plays a role in phytic acid biosynthesis in developing seeds**. *The Plant journal : for cell and molecular biology* 2005, **42**(5):708-719.

Shi J, Wang H, Wu Y, Hazebroek J, Meeley RB, Ertl DS: **The maize low-phytic acid mutant lpa2 is caused by mutation in an inositol phosphate kinase gene**. *Plant physiology* 2003, **131**(2):507-515.

Shi JR, Wang HY, Schellin K, Li BL, Faller M, Stoop JM, Meeley RB, Ertl DS, Ranch JP, Glassman K: **Embryo-specific silencing of a transporter reduces phytic acid content of maize and soybean seeds**. *Nature biotechnology* 2007, **25**(8):930-937.

Song QX, Liu YF, Hu XY, Zhang WK, Ma B, Chen SY, Zhang JS: **Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing**. *BMC Plant Biol* 2011, **11**:5.

Stevenson-Paulik J, Bastidas RJ, Chiou S-T, Frye RA, York JD: **Generation of phytate-free seeds in Arabidopsis through disruption of inositol polyphosphate kinases**. *PNAS* 2005, **102**(35):12612-12617.

Watson L, Henry R: **Microarray analysis of gene expression in germinating barley embryos (Hordeum vulgare L.)**. *Funct Integr Genomics* 2005, **5**(3):155-162.

Weber H, Borisjuk L, Wobus U: **Molecular physiology of legume seed development**. *Annual review of plant biology* 2005, **56**:253-279.

Wilcox JR, Premachandra GS, Young KA, Raboy V: **Isolation of High Seed Inorganic P, Low-Phytate Soybean Mutants**. *Crop Science* 2000, **40**(6):1601.

Xu XH, Zhao HJ, Liu QL, Frank T, Engel KH, An G, Shu QY: **Mutations of the multi-drug resistance-associated protein ABC transporter gene 5 result in reduction of phytic**

acid in rice seeds. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik* 2009, **119**(1):75-83.

Table 3.1: Summary of network construction.

Strategy	Metabolic genes used for network	Co-expression Gene Modules+			
		Total	Expression increase with seed development	Expression decrease with seed development	Expression parabolic to seed development
***Individual network*:**					
a. 3mlpa	1566	4	Brown	Turquoise, Yellow	Blue
b. 3MWT	1555	6	Blue	Turquoise	Brown
c. 1mlpa	1551	3	Blue	Turquoise	Brown
d. 1MWT	1546	3	Brown	Blue	Turquoise
***Combined network*:**					
a. *3mlpa + 1mlpa*	1587	4	Brown	Turquoise	Blue, Yellow
b. 3MWT + 1MWT	1583	5	Blue	Turquoise	Green, Brown

+ indicates modules with significant association with seed development.

Table 3.2: Enriched metabolic processes in co-expressed gene modules.

GO terms	Metabolic Processes (Significant at 1% FDR)	Networks*					
		3mlpa	3WT	*1mlpa*	1WT	*3mlpa* + *1mlpa*	3MWT + 1MWT
GO:0006754	ATP biosynthetic process	N	N	N	N	N	N
GO:0008652	Cellular amino acid biosynthetic process ➔	N	-	N	N	N	N, P
GO:0006812	Cation transport	N	N	N	N	N	N
GO:0045226	Extracellular polysaccharide biosynthesis process	N	N	N	N	N	N
GO:0006012	Galactose metabolic process	-	N	N	N	N	N
GO:0006536	Glutamate metabolic process ➔	N	-	N	-	N	N
GO:0030259	Lipid glycosylation	N	-	N	N	N	N
GO:0008610	Lipid biosynthetic process ➔	-	P	-	P	-	-
GO:0006564	L-serine biosynthetic process	-	N	N	-	-	-
GO:0006479	Protein amino acid methylation	N	N	N	-	N	N
GO:0036260	RNA capping	-	N	N	-	N	N
GO:0000154	rRNA modification	N	N	N	-	N	N
GO:0006694	Steroid biosynthesis process	N, P	N, P	N, P	N, P	N, P	N, P
GO:0006400	tRNA modification	N	N	N	-	N	N

* N and P correspond to negatively and positively associated modules, respectively.
Arrows (➔) indicates the metabolic processes that were enriched in either mutants or wildtypes.

Table 3.3: Genes associated with differentially enriched metabolic processes.

Metabolic processes differentially enriched in mutant and wildtype	Genes IDs		
(A) Lipid biosynthetic process (GO:0008610) ---- enriched in wildtypes			
High chlorophyll fluorescence phenotype 173 (HCF173)	Glyma07g00240		
Dihydroflavonol 4-reductase (DFR, M318, TT3, BEN1)	Glyma13g27390		
UDP-D-glucose/UDP-D-galactose 4-epimerase 1 (UGE1)	Glyma05g38120		
NmrA-like negative transcriptional regulator family protein	Glyma01g37840	Glyma01g37850	Glyma11g07490
NAD(P)-binding Rossmann-fold superfamily protein	Glyma06g41520	Glyma11g29460	Glyma12g02250
(B) Cellular amino acid biosynthesis process (GO:0008652) ---- enriched in wildtypes			
Delta 1-pyrroline-5-carboYylate synthase 2 (P5CS2)	Glyma02g41850		
N-acetyl-l-glutamate synthase 1 (NAGS1)	Glyma01g40230		
NADH-dependent glutamate synthase 1 (GLT1)	Glyma04g41540	Glyma14g32500	Glyma19g16486
Chloroplastic NIFS-like cysteine desulfurase (NSF2, CPNIFS, SUFS)	Glyma15g13350		
Branched-chain aminotransferase 3 (BCAT-3)	Glyma11g04870		
Branched-chain amino acid transaminase 2 (BCAT-2)	Glyma08g06766		
Aspartate kinase-homoserine dehydrogenase ii (AK-HSDH II)	Glyma05g28380		
D-aminoacid aminotransferase-like PLP-dependent enzymes superfamily protein	Glyma07g30500		
Fumarylacetoacetase, putative	Glyma09g01270		
(C) Glutamate metabolic process (GO:0006536) ---- enriched in mutants			
Glutamate synthase 1 (FD-GOGAT, GLS1, GLU1, GLUS)	Glyma03g28410	Glyma19g31120	
Glutamate decarboxylase 1 (GAD1)	Glyma11g33280		
Glutamate decarboxylase 4 (GAD4)	Glyma18g04940		
Glutamate decarboxylase 5 (GAD5)	Glyma09g29900		
NADH-dependent glutamate synthase 1 (GLT1)	Glyma06g13280		

Table 3.4: Gene co-expression modules significantly correlated to soybean seed development.

Network	Module		*N*	%	Hub gene	*r**	Hub gene annotation	Significance#
3mlpa	Brown	+	136	8.1	Glyma13g23790	0.92	Formate dehydrogenase	5.26E-06
	Turquoise	-	998	59.4	Glyma11g04800	-0.88	Cysteine desulfurase (nitrogen fixation 1 homolog)	8.74E-05
3MWT	Blue	+	165	9.8	Glyma19g01210	0.89	Format dehydrogenase	1.70E-05
	Turquoise	-	986	58.7	Glyma15g14433	-0.81	UDP-glucose 4-epimerase	2.44E-05
1mlpa	Blue	+	216	12.8	Glyma17g05980	0.78	Threonine synthase	2.10E-03
	Turquoise	-	1000	59.5	Glyma03g27360	-0.93	Dienoyl-CoA hydratase/isomerase	2.67E-08
1MWT	Brown	+	216	12.8	Glyma15g41690	0.90	Aldehyde dehydrogenase 5F1	3.26E-06
	Blue	-	661	39.3	Glyma06g04900	-0.84	Autoinhibited Calcium ATPase 1	1.67E-04
3mlpa + 1mlpa	Brown	+	132	7.8	Glyma08g01390	0.89	Hydroxysteroid dehydrogenase 1	1.00E-10
	Turquoise	-	1006	59.9	Glyma03g27360	-0.89	Dienoyl-CoA hydratase/isomerase	3.30E-11
3MWT + 1MWT	Blue	+	202	12.03	Glyma09g34410	0.86	Phosphoribulokinase	1.77E-08
	Turquoise	-	866	51.5	Glyma06g04900	-0.77	Autoinhibited Calcium ATPase 1	4.33E-09

N corresponds to total number of genes within a module.
% indicates the percentage metabolic process genes present in a module
* indicates the correlation coefficient of module relationship with seed development process.
# indicates the p-value of hub gene significance for correlation of gene expression profile with seed development process.

Table 3.5: Regulatory nodes in developing seeds of mutants and wildtype lines.

Regulatory nodes common to *3mlpa* and *1mlpa*	Network	Module
Fatty acid metabolism:		
Long chain acyl-CoA synthetase 8 (LACS8)	Combined	Positive
Phospholipase D delta (PLDDELTA)	Individual	Positive
Acyl-activating enzyme (AAE16)	Both	Negative
Delta(3,5)-delta(2,4)-dienoyl-CoA-isomerase 1 (DCI1)	Both	Negative
Lysophosphatidyl acyl transferase 2 (LPAT2)	Combined	Negative
Beta-ketoacyl reductase 1 (KCR1)	Combined	Negative
D-isomer-specific 2-hydroxyacid dehydrogenase (HADH)	Combined	Negative
Aminophospholipid ATPase 3 (ALA3)	Combined	Negative
Heavy metal ATPase 1 (HMA1)	Combined	Negative
Alpha/beta hydrolases superfamily (WAV2)	Combined	Negative
Sucrose signaling:		
UDP glucose pyrophosphorylase 2 (UGP2)	Combined	Positive
UDP-glycosyl transferase 74E2 (UGT74E2)	Combined	Positive
Amino acid metabolism:		
NADH dependent glutamate synthase 1 (GLT1)	Combined	Positive
Alkaline-phosphatase-like family protein (Apase)	Combined	Negative
Aminophospholipid ATPase 3 (ALA3)	Combined	Negative
Secondary metabolites biosynthesis:		
S-adenosyl-Methionine (SAM)-dependent methyltransferases	Combined	Positive Negative
Regulatory nodes common to 3MWT and 1MWT		
Fatty acid metabolism:		
Phospholipase D delta (PLDDELTA)	Combined	Positive
Short chain dehydrogenase reductase B (DECR, SDRB)	Combined	Negative
Phospholipid/glycerol acyltransferase family protein (ACT1)	Combined	Negative
Acetoacetyl-CoA thiolase 2 (ACAT2)	Combined	Negative
ATP-dependent caseinolytic protease/crotonase family protein	Combined	Negative
Sucrose signaling:		
UDP-D-glucoronate 4-epimerase 3 (GAE3)	Combined	Negative
Aldehyde dehydrogenase 2B4 (ALDH2A)	Combined	Positive
UDP-glycosyl transferase 78D2 (UGT78D2)	Combined	Positive
Amino acid metabolism:		
Aspartate kinase homoserine dehydrogenase II (AK-HSDH2 II)	Combined	Positive
N-acetyl-l-glutamate synthase 1 (NAGS1)	Combined	Positive
Aspartate kinase 1 (AK-LYS1)	Combined	Negative
ACT-domain containing protein (ACTP)	Combined	Negative
Peptidase M20/M25/M40 family protein	Combined	Positive
Pyridoxal-5'-phosphate dependent transferase superfamily protein	Combined	Negative
Others:		
H^+−ATPase 11 (AHA11)	Combined	Negative
Haloacid dehalogenase (HAD)-like hydrolase superfamily protein	Individual	Positive

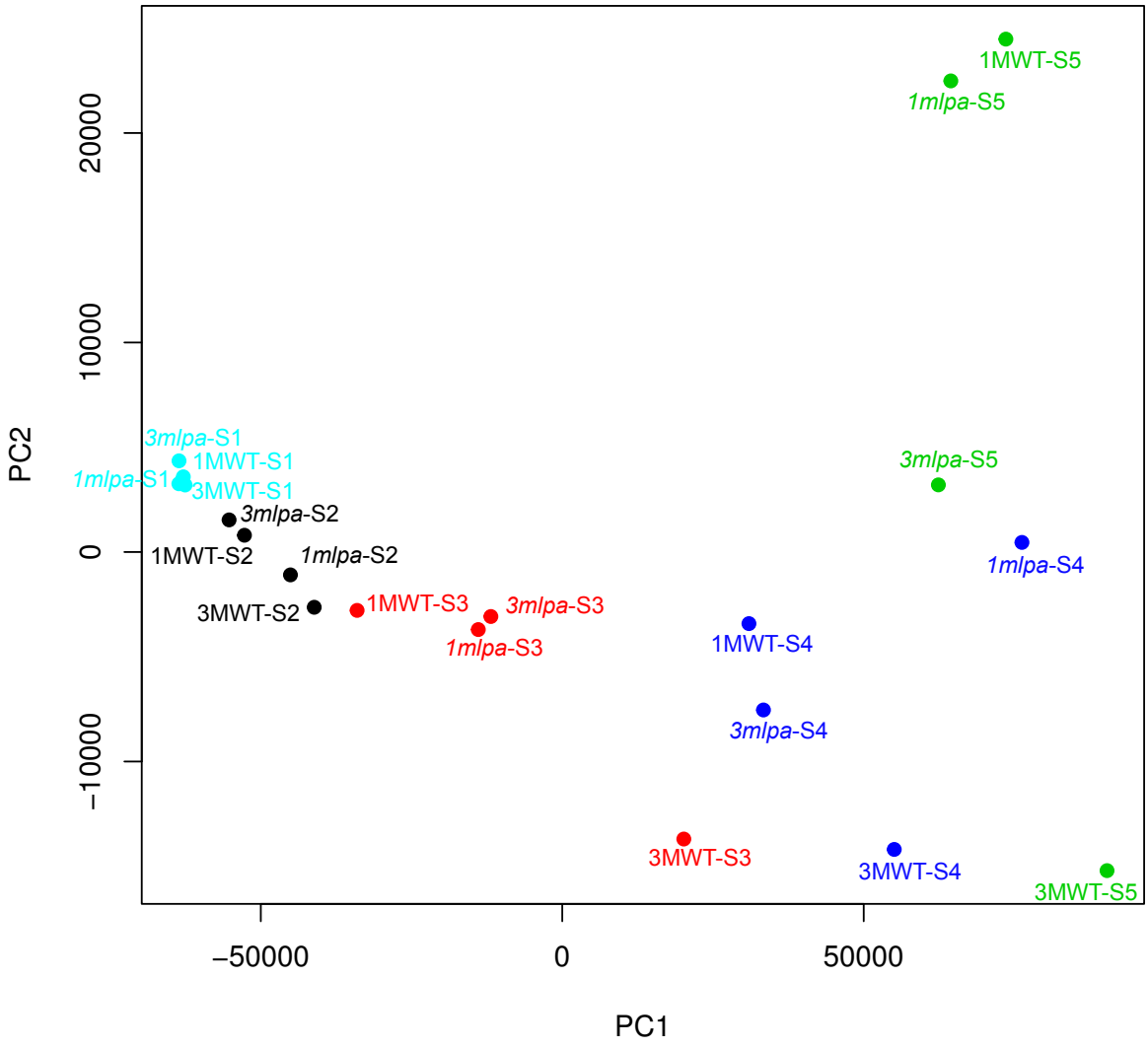
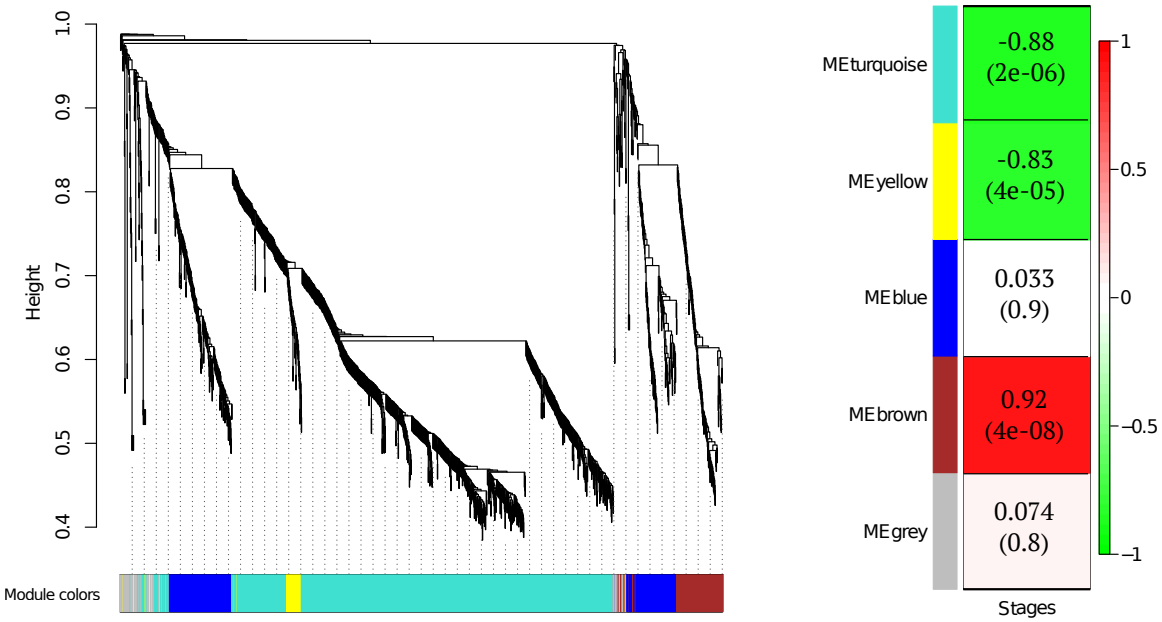


Figure 3.1: Principal component analysis. PC1 and PC2 explained more than 95% and 3% variation in the normalized expression profiles of metabolic process genes within our samples respectively. Samples are clustered according to the seed developmental stages along PC1. Samples originating from stage 1 to 5, designated by S1 to S5, are represented in cyan, black, red, blue, and green colors, respectively. Samples at seed developmental stage 1 show no variation among the experimental lines; however, this variation increases with progression of seed development.



(a) (b)

Figure 3.2: *3mlpa* co-expression gene network modules. (a) Dendrogram for co-expressed gene modules. (b) Module eigengenes relationship with seed development stages. Module colors are indicated on left. Heat map indicates the correlation of module eigengenes with the seed development. Green and red colors indicate negative and positive correlations. Correlation values and p-value in the parenthesis are indicated for each module.

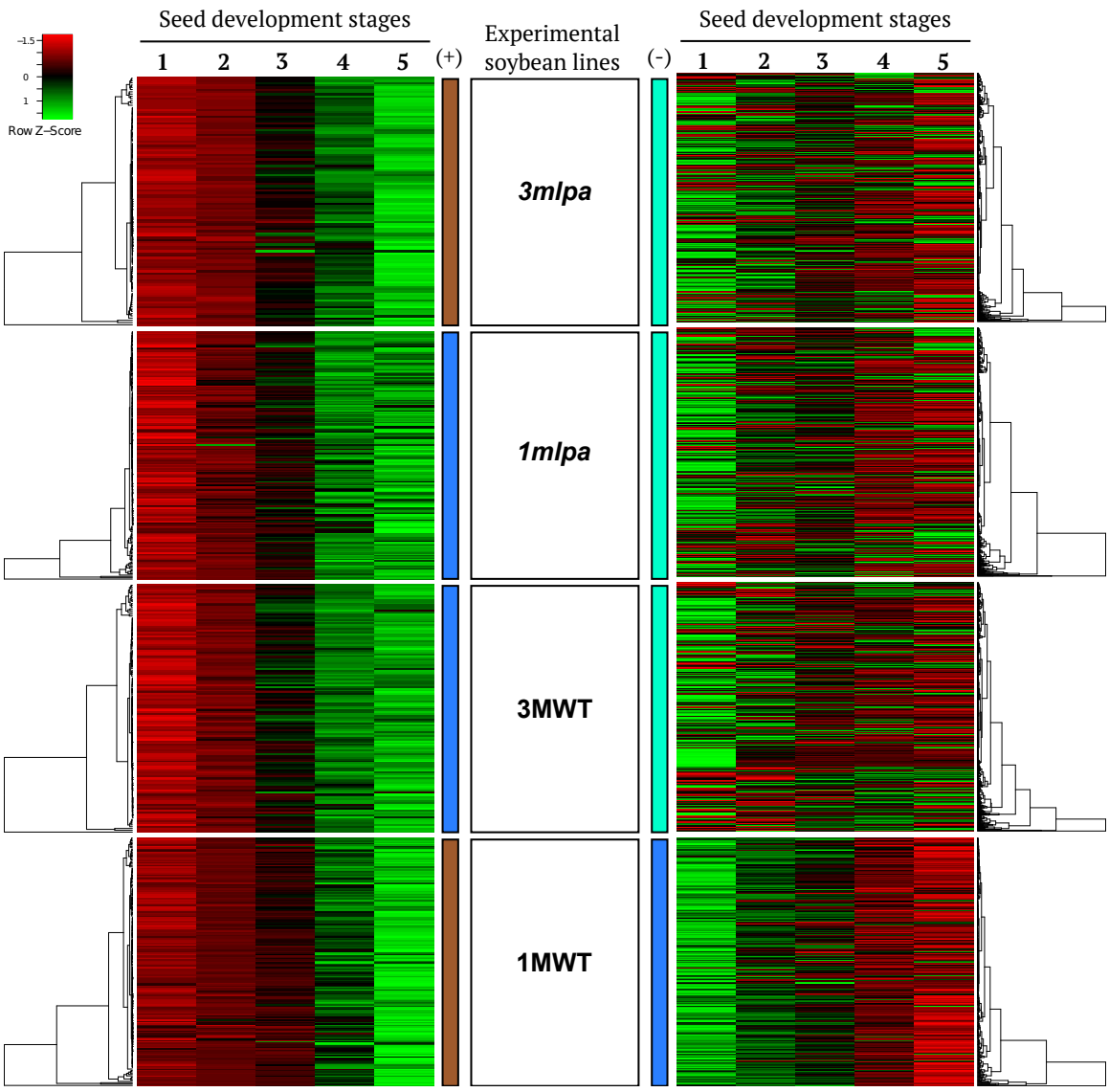


Figure 3.3: Expression profiles of co-expressed genes within positively and negatively correlated modules of individual networks. Heatmap on left and right indicates positively and negatively correlated gene modules respectively. The respective module colors are indicated in color bars right next to heat maps. Red and green bars indicate low and high gene expression levels, respectively.

CHAPTER 4

Copy Number Variations in Near Isogenic Soybean Lines

Neelam R Redekar, Ruslan Biyashev, M A Saghai Maroof[§]

Department of Crop and Soil Environmental Sciences, Virginia Tech. [§]Corresponding author: smaroof@vt.edu

ABSTRACT

Copy number variations (CNVs) play a substantial role in the evolution of complex organisms such as plants. These variations are stable over generations and can arise in response to environmental stimuli, resulting in an increase or decrease in the genomic content. Changes in the genome content can be falsely interpreted as the differences in transcriptome expression levels. In order to study the effect of copy number variations on transcriptome expression levels, we sequenced two isogenic soybean siblings that originated from selfing of a single seed heterozygous for a mutation in the *myo*-inositol phosphate synthase gene. We estimated over 6000 significant genomic positions that showed copy number variations with 76 of these were associated with one gene each. We also used transcriptome sequencing data from five seed developmental stages to estimate the differential gene expression between the siblings at each stage. We found no overlap between CNV-associated genes and the differentially expressed genes, suggesting that the CNVs between the isogenic-like siblings have no effect on differential gene expression.

KEYWORDS

Copy number variation, near isogenic lines, CNV-seq

INTRODUCTION

Copy number variations (CNVs), a form of structural polymorphism, usually range between a few kilo- to mega-bases in length. The CNVs are considered to be one of the key players in amending genetic diversity, as they are heritable and can be introduced

de novo, resulting in duplication by insertion, deletion, and translocation of the genomic segments. Several mechanisms associated with homologous and non-homologous recombination have been proposed to change the copy number [Hastings, et al. 2009, Liu, et al. 2012, Lu, et al. 2012]. The "recurrent" CNVs are formed as a result of homologous recombination between two repeated segments, whereas the "non-recurrent" ones are formed as a result of non-homologous recombination [Hastings, et al. 2009, Liu, et al. 2012]. Current sequencing technologies enable the detection of CNVs at a genome-wide scale. Occurrence of CNVs in plant genomes is a common phenomenon. Several studies have reported genome scale CNVs in Arabidopsis, rice (*Oryza sativa*), maize (*Zea mays*), barley (*Hordeum vulgare*), and soybean (*Glycine max*), providing insights into domestication, speciation, and biotic-abiotic stress responses [Springer, et al. 2009, DeBolt 2010, Hurwitz, et al. 2010, Cook, et al. 2012, Lu, et al. 2012, McHale, et al. 2012, Munoz-Amatriain, et al. 2013, Anderson, et al. 2014, Zmienko, et al. 2014]. Intraspecific CNVs, meaning the variations between two individuals of the same species, are formed as a result of crossing between individuals, followed by segregation between non-allelic homologs [Springer, et al. 2009, Liu, et al. 2012]. Inter-lineage CNVs were also observed in 5th generation siblings, which originated from a common ancestor in *Arabidopsis thaliana* [DeBolt 2010]. Intraspecific and inter-lineage CNVs can be useful resources for modern crop breeding.

Soybean (*Glycine max* (L.) Merr.), a self-pollinating species, with a natural outcrossing rate less than 1%, has undergone two-rounds of duplication approximately 59 and 13 million years ago [Wilcox 1987, Schmutz, et al. 2010]. Genomes of self-pollinating species, like soybean, are more stable than outcrossing plant species, such as

maize. However, recent studies on structural variations in soybean have changed this perspective. A study with 41 different soybean accessions discussed the impact of CNVs on soybean genome diversity, and on the expression of underlying genes [Anderson, et al. 2014]. The CNVs estimated in a pair of near-isogenic lines of soybean, exhibiting contrasting trait phenotypes, have been associated with introgression mapping [Stec, et al. 2013]. The soybean cyst nematode resistance gene *Rhg1* was associated with CNVs in a multi-gene locus, suggesting CNV implications in adaptive traits [Cook, et al. 2012].

In this study, we employ a comparative genomics approaches to identify the CNVs in a pair of near isogenic lines, originated by selfing of a single soybean seed that was known to be heterozygous for the *myo*-inositol phosphate synthase 1 (MIPS1) gene locus. One of the near isogenic lines has a low phytic acid-causing mutation, *mips1*, while the other does not. These near isogenic lines are expected to be genetically identical, which makes them an excellent resource for low phytic acid research. We previously used these isogenic-like siblings for the co-expression network study described in previous Chapter 3. However, existence of CNVs between these siblings can confound the gene expression research findings. It is therefore necessary to validate the presence of CNVs between these experimental lines and explore their impact on expression of CNV-associated genes.

MATERIAL AND METHODS

Genetic material background

Two soybean (*Glycine max* (L.) Merr.) isogenic-like sibling lines used in this study primarily originate from a cross of a low phytic acid line ‘V99-5089’ with a normal

phytic acid line 'Essex'. The V99-5089 soybean line carries a point mutation in the MIPS1 gene that results in a low phytic acid, low stachyose, low emergence and high sucrose phenotype [Maroof and Buss 2008]. Progenies from this cross were advanced in the field for several generations to develop a recombinant inbred line (RIL) population. Subsequently, a single heterozygous (MIPS1/*mips1*) RIL individual plant was identified [Glover 2011] based on a marker assay. We refer to this RIL plant generation, as "H_0" This heterozygous plant was selfed to obtain two homozygous individuals: a mutant line (*mips1*/*mips1*), and a wildtype line (MIPS1/MIPS1). This generation will be referred to as "H_1" These homozygous individuals were advanced for two additional generations to obtain the H_3 generation seeds, which were used for this study. These two near isogenic lines will be referred to as: (1) "*1mlpa*," for *mips1/mips1* mutant and, (2) "1MWT," for MIPS1/MIPS1 wildtype. The seed phenotype in terms of PA, sucrose, raffinose, and stachyose content of the *1mlpa* mutant and 1MWT wildtype lines is similar to the original parents, V99-5089 and Essex, respectively (Figure 1).

Sample preparation for sequencing

Each experimental line, *1mlpa* and 1MWT, was grown in two 7-inch diameter pots (4 seeds per pot), containing Metro-Mix® 360 (Sun Gro) soilless media and GardenPro ULTRA^{LITE} topsoil, under controlled growth chamber conditions: 14h light/10h dark photoperiod and 24/16°C temperatures. Genomic DNA was extracted from the young trifoliolate leaves of individual plants [Yu, et al. 1994]. High quality genomic DNA samples bulked from different pots were used for DNA library preparation and sequencing at the Génome Québec Innovation Centre, Canada. The DNA-Seq libraries

were prepared from 1MWT and *1mlpa* genomic DNA with barcodes-index IDs "ACAGTG" and "GCCAAT," respectively. Two libraries were then multiplexed together and this mixture was sequenced on 2 different lanes of the HiSeq2000 sequencer as 150 bp paired-end (150PE) sequences to obtain roughly 22.5x coverage.

Estimating copy number variations

Four data files, or files generated from two lanes for each line, comprising of 150PE sequencing reads will be referred to as—MWT-001, 1MWT-002, *1mlpa*-001, and *1mlpa*-002. Supplementary Table 1 indicates the number of sequences in these data files. The sequencing data was analyzed on MAGYK-MPS server. The barcode diversity in the sequencing data was estimated using basic UNIX programming, followed by the selection of the sequencing reads with library-specific barcode indices (Supplementary Table 2). The quality of these barcode-selected sequencing reads was estimated using FastQC [Andrews]. These sequences were aligned to the soybean reference 'Williams 82' genome sequence using default parameters of BWA-MEM program [www.soybase.org]. The sequencing read alignment data were further sorted with coordinate position, and optical duplicates were removed using PicardTools [Wysoker, et al. 2013]. The alignment data obtained from single library sequences were merged together and read mapping positions were estimated using SAMtools [Li, et al. 2009]. A perl script "cnv-seq.pl" provided by the CNV-Seq package, was used to divide the reference genome into sliding windows and to count mapped reads within each sliding window [Xie and Tammi 2009]. Based on the soybean genome sequence size (including chromosomes and scaffolds) of 973344380 bases, CNV-Seq estimated the sliding window size to be equal to 1269 bases.

Each sliding window was overlapping with its immediate neighbor, with the step size equal to half of the window size. The total number of reads mapped within each sliding window of 1MWT and *1mlpa* were used to calculate *log2* (1MWT/*1mlpa*) count ratios and estimate CNVs using the R package "cnv" provided by the CNV-Seq package [Xie and Tammi 2009]. Final CNV calls were obtained by filtering data at 0.01% false discovery rate (FDR) using the Benjamini Hochberg method [Benjamini and Hochberg 1995].

Differential gene expression

The transcriptome sequencing for five developing seed stages of *1mlpa* and 1MWT is as described in Chapter 3. For determination of differential gene expression, the sequencing reads were mapped to the reference 'Williams 82' genome sequence using TopHat2 (v2.0.8) [Kim, et al. 2013]. The read count data was estimated for genes as a measure of gene expression using HTSeq-count [Anders, et al. 2014]. The gene expression data was normalized and differential expression analysis was performed comparing *1mlpa* and 1MWT for 5 stages using DESeq (v1.12.1) [Anders and Huber 2010]. Genes with significant fold change difference between *1mlpa* and 1MWT were identified at 1% FDR.

RESULTS AND DISCUSSION

Copy number variations in two near isogenic lines

The two homozygous near isogenic lines of soybean: "*1mlpa*" (*mips1/mips1* mutant) and "1MWT" (MIPS1/MIPS1 wildtype) are derived from selfing of a single

heterozygous (*mips1*/MIPS1) RIL individual (Figure 1). The *mips1* mutation causes a reduction in phytic acid, stachyose, and seedling emergence, while increasing seed sucrose levels in the *1mlpa* line. On the contrary, the 1MWT line has normal seed phytic acid, stachyose, emergence, and sucrose. Given the genetic history of these near isogenic lines, they are expected to be very similar to each other, except at the MIPS1 gene locus. These contrasting phenotypes of these near isogenic lines make them a perfect specimen for the CNV study. The *log2* (1MWT/*1mlpa*) ratios, indicative of a presence of CNVs, were estimated by comparing the mapped read count between 1MWT and *1mlpa* in overlapping sliding windows. A positive *log2* (1MWT/*1mlpa*) ratio is indicative of copies higher in 1MWT, and vice-versa. Total of 6646 CNV events were identified by the CNV-Seq tool. The summary distribution of CNVs in the genome is shown in Figure 2 (top panel). The last three panels of Figure 2 are the zoomed-in view of the largest CNV event on chromosome 14. Exactly 6632 (i.e., more than 99%) CNV events were found significant with 0.01% FDR post Benjamini-Hochberg correction. These included CNV events with 35 unique sizes, ranging between the smallest CNV with 2537 bp to the largest CNV with 59597 bp (average size 3215 bp) (Figure 3). More than 92% of these significant CNVs were less than 5000 bp in size (Figure 3). The *log2* (1MWT/*1mlpa*) ratios of significant CNV events were in the range of -3.2 to 3.5. We observed that 6275 and 375 CNV events showed a negative and positive log2 (1MWT/*1mlpa*) ratio, respectively. One CNV locus, CNVR_4219, at position Gm13, 36579582..36582118, was found overlapping the duplicated synteny region. A positive log2 ratio indicates the higher copies for this locus in 1MWT. The validation experiments are required to confirm the results using quantitative real time methods. To conclude, presence of CNVs between

1MWT and *1mlpa* suggest near isogenic lines are quite different from each other. These CNV-associated genetic loci can be a valuable resource for characterizing the low phytic trait in these near isogenic lines.

Effect of CNVs on differential gene expression in *1mlpa* and 1MWT

About 76 CNV loci were identified in the coding regions, containing one gene per locus (Table 1). It is very likely that the CNVs associated with these genes can alter the read count-based expression profiles between these near isogenic lines, resulting in false interpretations. Gene dosage was previously shown to affect the gene expression in Drosophila [Zhou, et al. 2011]. Annotations indicated that these CNV-associated genes belonged to several functional categories such as transport, response to biotic stimulus and stress, signal transduction, flower development, carbohydrate metabolism, and translation. Many of these functional categories were observed in the previous chapter, where we compared gene co-expression networks between these near isogenic lines. In order to check the effect of these CNVs on transcriptome expression, we used the transcriptome sequencing data of five seed developmental stages, sampled based on seed length (Refer to Chapters 2 and 3). First we checked if CNV-associated genes are expressed in our experimental condition. About 17 were expressed in none, 38 were expressed in all, while remaining genes were expressed in fewer sample libraries. Since CNVs can be misread as an increase in the transcript count by sequencing-based expression profiling methods, we decided to find an overlap of CNV-associated genes and differentially expressed genes between *1mlpa* and 1MWT at each experimental condition. Highly significant differentially expressed genes between *1mlpa* and 1MWT

for these five stages were identified at 1% FDR. About 127, 11, 37 172, and 0 genes were found differentially expressed between *1mlpa* and 1MWT at stages 1 to 5, respectively, and a total of 324 unique genes were differentially expressed when all samples were combined. The CNV-associated genes and differentially expressed genes between *1mlpa* and 1MWT showed no overlap, suggesting no effect of CNVs on gene expression. Further investigations are required to obtain a detailed understanding of CNVs.

CONCLUSION

Copy number variation analysis is important to determine the gene dosage effect on the expression. To investigate the effect of CNVs on gene expression, we employed third-generation isogenic-like siblings, which descended by selfing of a single soybean plant. We identified 6636 highly significant loci with CNVs between these siblings. About 76 loci were associated with coding regions, encoding for signal transduction, carbohydrate metabolism, transport, response to stress, etc. We also identified highly significant genes that are differentially expressed at 5 seed developmental stages between the two siblings. There was no overlap between the CNV-associated and differentially expressed genes.

COMPETING INTERESTS

Authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

MASM and RB developed the near isogenic lines. MASM, RB, and NR participated in designing, detailing and conducting of the experiment, analyzing the sequencing data, and drafting of the manuscript. This work needs further experiments, such as estimating copy numbers using second software and comparing the result to find more biologically meaningful variations, before this manuscript is ready for publication.

ACKNOWLEDGEMENTS

This work was funded by USDA-NIFA grant via Bio-design and Bioprocessing Research Center (BBRC) at Virginia Tech. We would like to thank support team of Advanced Research Computing (ARC) server and Translational Plant Sciences' MAGYK server at Virginia Tech.

REFERENCES

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

Anders S, Pyl PT, Huber W: **HTSeq – A Python framework to work with high-throughput sequencing data**. 2014.

Anderson JE, Kantar MB, Kono TY, Fu F, Stec AO, Song Q, Cregan PB, Specht JE, Diers BW, Cannon SB *et al*: **A roadmap for functional structural variants in the soybean genome**. *G3 (Bethesda)* 2014, **4**(7):1307-1318.

FastQC A Quality Control tool for High Throughput Sequence Data [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/]

Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society Series B (Methodological)* 1995, **57**(1):289-300.

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE *et al*: **Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean**. *Science* 2012, **338**(6111):1206-1209.

DeBolt S: **Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales**. *Genome Biol Evol* 2010, **2**:441-453.

Glover N: **The Genetic Basis of Phytate, Oligosaccharide Content, and Emergence in Soybean**. Blacksburg, VA: Virginia Tech; 2011.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G: **Mechanisms of change in gene copy number**. *Nat Rev Genet* 2009, **10**(8):551-564.

Hurwitz BL, Kudrna D, Yu Y, Sebastian A, Zuccolo A, Jackson SA, Ware D, Wing RA, Stein L: **Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus Oryza**. *Plant J* 2010, **63**(6):990-1003.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome biology* 2013, **14**(4):R36.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

Liu S, Ying K, Yeh CT, Yang J, Swanson-Wagner R, Wu W, Richmond T, Gerhardt DJ, Lai J, Springer N *et al*: **Changes in genome content generated via segregation of non-allelic homologs**. *Plant J* 2012, **72**(3):390-399.

Lu P, Han X, Qi J, Yang J, Wijeratne AJ, Li T, Ma H: **Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis**. *Genome Res* 2012, **22**(3):508-518.

Maroof AS, Buss GR: **Low phytic acid, low stachyose, high sucrose soybean lines**. In.: Google Patents; 2008.

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM: **Structural variants in the soybean genome localize to clusters of biotic stress-response genes**. *Plant Physiol* 2012, **159**(4):1295-1308.

Munoz-Amatriain M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, Spannagl M, Nussbaumer T *et al*: **Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome**. *Genome biology* 2013, **14**(6):R58.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J *et al*: **Genome sequence of the palaeopolyploid soybean**. *Nature* 2010, **463**(7278):178-183.

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H *et al*: **Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content**. *PLoS Genet* 2009, **5**(11):e1000734.

Stec AO, Bhaskar PB, Bolon YT, Nolan R, Shoemaker RC, Vance CP, Stupar RM: **Genomic heterogeneity and structural variation in soybean near isogenic lines**. *Front Plant Sci* 2013, **4**:104.

Wilcox JR: **Soybeans: improvement, production and uses**: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America; 1987.

Picard tools version 1.90

Xie C, Tammi MT: **CNV-seq, a new method to detect copy number variation using high-throughput sequencing**. *BMC Bioinformatics* 2009, **10**:80.

Yu YG, Maroof MAS, Buss GR, Maughan PJ, Tolin SA: **Rflp and Microsatellite Mapping of a Gene for Soybean Mosaic-Virus Resistance**. *Phytopathology* 1994, **84**(1):60-64.

Zhou J, Lemos B, Dopman EB, Hartl DL: **Copy-number variation: the balance between gene dosage and expression in Drosophila melanogaster**. *Genome Biol Evol* 2011, **3**:1014-1024.

Table 4.1: Gene-inclusive CNV events between *1mlpa* and 1MWT.

CNV event*	Locus	Start (Mb)	Size (bp)	Log2 ratio	Adjusted P-value	Genes within CNV events
CNVR_5765	Gm01	49.59	2537	0.74	4.9E-24	Glyma01g37200
CNVR_2446	Gm02	3.22	2537	0.80	1.6E-27	Glyma02g04010
CNVR_2450	Gm02	44.03	2537	0.72	4.5E-23	Glyma02g38673
CNVR_2451	Gm02	50.40	2537	0.65	4.1E-19	Glyma02g46450
CNVR_2453	Gm02	1.20	2537	-0.73	2.8E-22	Glyma02g01660
CNVR_17	Gm03	9.40	4439	0.99	1.2E-58	Glyma03g08478
CNVR_28	Gm03	14.44	2537	1.55	8.2E-66	Glyma03g11820
CNVR_40	Gm03	34.23	2537	0.69	5.4E-19	Glyma03g26750
CNVR_41	Gm03	37.63	2537	0.76	4.5E-22	Glyma03g29610
CNVR_42	Gm03	39.30	2537	0.65	5.6E-17	Glyma03g31390
CNVR_65	Gm03	5.91	2537	-1.89	6.1E-79	Glyma03g05590
CNVR_66	Gm03	6.50	2537	-1.21	3.1E-44	Glyma03g06300
CNVR_371	Gm03	31.66	2537	-1.39	6.8E-54	Glyma03g24730
CNVR_4211	Gm04	47.71	3171	-0.83	2.8E-32	Glyma04g41891
CNVR_3919	Gm04	19.07	2537	-0.94	4.6E-32	Glyma04g17600
CNVR_3875	Gm04	13.18	2537	-1.00	2.8E-35	Glyma04g13531
CNVR_3859	Gm04	1.03	2537	-1.26	4.2E-50	Glyma04g01550
CNVR_3858	Gm04	47.58	3171	0.66	7.3E-23	Glyma04g41720
CNVR_3856	Gm04	46.64	2537	0.78	1.5E-24	Glyma04g40650
CNVR_3855	Gm04	38.19	2537	0.72	1.3E-21	Glyma04g32940
CNVR_3841	Gm04	14.96	2537	0.86	6.0E-29	Glyma04g14770
CNVR_3839	Gm04	8.98	2537	0.71	2.7E-21	Glyma04g10740
CNVR_1204	Gm05	8.67	3805	-0.91	1.4E-47	Glyma05g08781
CNVR_1181	Gm05	40.58	2537	0.68	5.9E-21	Glyma05g36820
CNVR_1177	Gm05	23.45	3171	0.75	9.6E-31	Glyma05g19497
CNVR_5026	Gm06	18.76	2537	-0.80	1.3E-24	Glyma06g22030
CNVR_5021	Gm06	46.62	2537	0.69	1.2E-19	Glyma06g43620
CNVR_5015	Gm06	10.99	2537	0.76	1.7E-23	Glyma06g13910
CNVR_1778	Gm07	42.09	2537	0.76	5.0E-24	Glyma07g36805
CNVR_1777	Gm07	30.56	2537	0.68	7.0E-20	Glyma07g27420
CNVR_1771	Gm07	8.79	3171	1.16	1.0E-59	Glyma07g10508
CNVR_1770	Gm07	5.98	2537	0.73	1.5E-22	Glyma07g07270
CNVR_892	Gm08	11.83	2537	0.67	4.5E-19	Glyma08g16210
CNVR_4643	Gm09	30.68	3805	-1.23	9.0E-73	Glyma09g24790
CNVR_4393	Gm09	45.87	2537	0.71	9.0E-22	Glyma09g41130
CNVR_4392	Gm09	35.15	2537	0.77	8.5E-25	Glyma09g28140
CNVR_4380	Gm09	10.42	2537	0.74	1.7E-23	Glyma09g10361
CNVR_4997	Gm10	38.74	2537	-0.78	1.2E-23	Glyma10g30040
CNVR_4670	Gm10	5.63	2537	-0.68	9.8E-19	Glyma10g06910
CNVR_4667	Gm10	44.60	2537	0.68	1.1E-19	Glyma10g36460
CNVR_646	Gm11	10.97	2537	-0.72	1.8E-21	Glyma11g15340
CNVR_645	Gm11	10.25	2537	-1.07	4.0E-40	Glyma11g14290
CNVR_639	Gm11	0.10	2537	0.76	6.0E-24	Glyma11g00410
CNVR_1559	Gm12	19.52	2537	-1.70	4.3E-81	Glyma12g18835
CNVR_1481	Gm12	4.17	2537	1.10	2.4E-46	Glyma12g06111
CNVR_4287	Gm13	14.88	2537	-1.21	3.8E-52	Glyma13g11931
CNVR_2376	Gm14	37.22	2537	-0.92	2.0E-32	Glyma14g30490
CNVR_2125	Gm14	13.81	3171	-0.83	8.1E-34	Glyma14g14050
CNVR_2083	Gm14	29.73	2537	0.83	1.1E-28	Glyma14g24710
CNVR_5689	Gm15	46.14	2537	-2.21	3.0E-101	Glyma15g39490
CNVR_5628	Gm15	37.60	2537	-0.71	4.3E-21	Glyma15g33761

CNVR_5337	Gm15	35.88	2537	1.11	1.3E-44	Glyma15g32330
CNVR_5331	Gm15	10.20	2537	0.76	5.9E-24	Glyma15g13600
CNVR_638	Gm16	37.12	2537	-0.77	3.4E-24	Glyma16g34480
CNVR_628	Gm16	31.35	2537	-1.79	1.2E-81	Glyma16g27340
CNVR_410	Gm16	36.69	2537	0.65	1.1E-18	Glyma16g33950
CNVR_407	Gm16	25.93	2537	0.83	3.1E-28	Glyma16g22460
CNVR_404	Gm16	6.32	3171	0.71	7.2E-27	Glyma16g07010
CNVR_403	Gm16	3.70	2537	0.90	2.9E-32	Glyma16g04380
CNVR_2812	Gm17	13.82	2537	-1.00	2.4E-35	Glyma17g16990
CNVR_2809	Gm17	8.74	3171	-1.17	1.8E-55	Glyma17g11640
CNVR_2803	Gm17	19.06	2537	0.73	4.1E-22	Glyma17g20320
CNVR_6360	Gm18	28.51	2537	-1.25	7.8E-50	Glyma18g24740
CNVR_6288	Gm18	23.46	3171	-0.78	8.8E-30	Glyma18g20840
CNVR_6192	Gm18	4.93	3171	-0.74	2.8E-27	Glyma18g06350
CNVR_6190	Gm18	55.22	2537	1.16	2.6E-47	Glyma18g45471
CNVR_6162	Gm18	8.70	3805	2.46	4.5E-176	Glyma18g09812
CNVR_3492	Gm19	35.80	3805	-1.19	4.3E-65	Glyma19g28364
CNVR_3488	Gm19	33.96	2537	-1.10	2.3E-39	Glyma19g26850
CNVR_3104	Gm19	6.12	2537	-0.84	7.1E-26	Glyma19g05590
CNVR_3087	Gm19	44.47	2537	0.68	3.1E-19	Glyma19g37270
CNVR_3086	Gm19	40.10	2537	0.79	1.4E-24	Glyma19g32350
CNVR_3835	Gm20	43.38	2537	-1.41	9.9E-61	Glyma20g35095
CNVR_3834	Gm20	43.31	2537	-0.83	4.3E-27	Glyma20g35020
CNVR_3758	Gm20	22.38	2537	-0.81	4.4E-26	Glyma20g16130
CNVR_3503	Gm20	24.96	5073	1.49	4.4E-137	Glyma20g17941

* CNVR stands for copy number variation region

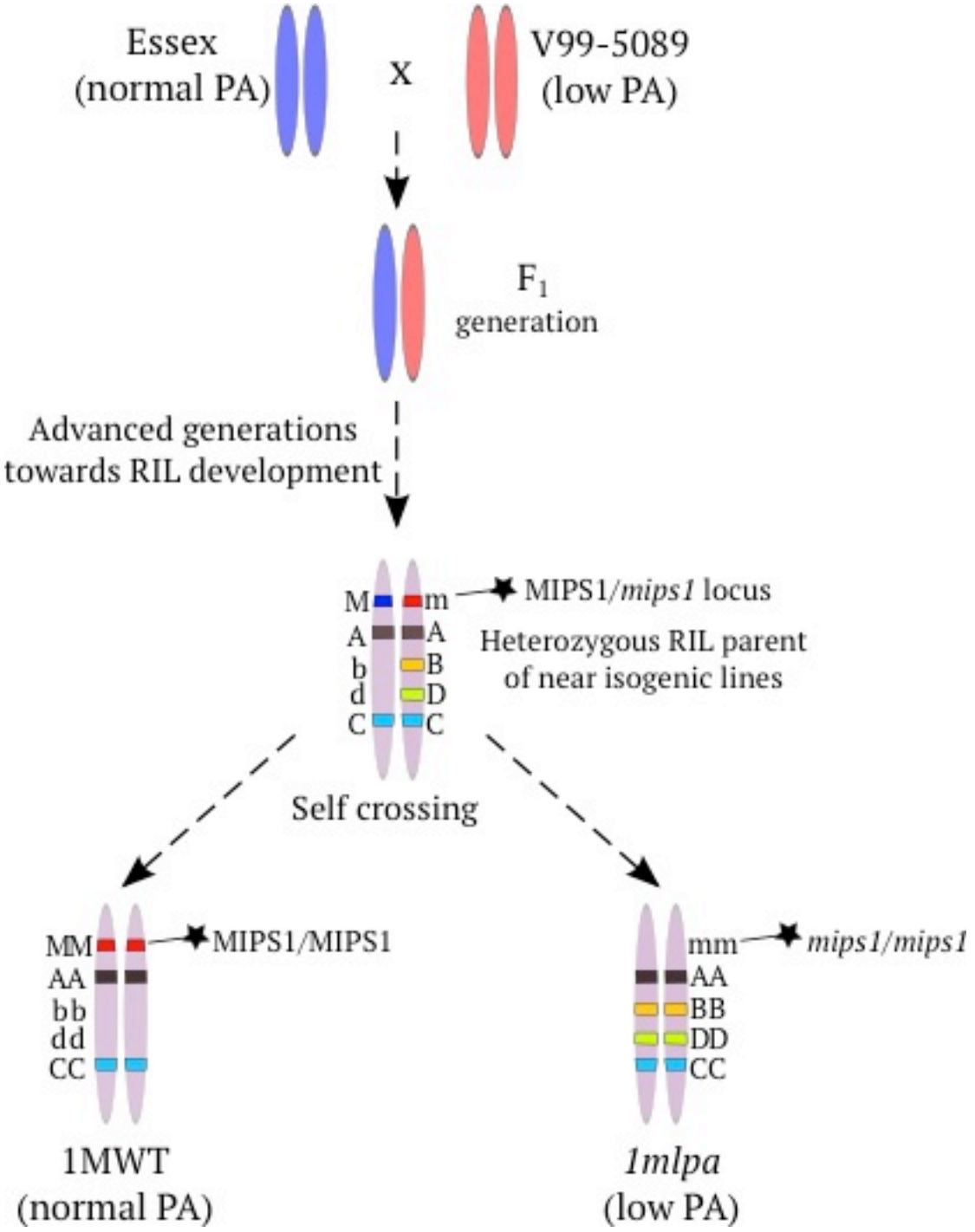


Figure 4.1: Breeding scheme of near isogenic lines, *1mlpa,* and 1MWT. In the text, the heterozygous RIL parent of near isogenic lines is referred to as "Ho" generation. This heterozygous parent was select based on genetic screening for mutation in the MIPS1 locus (represented as M locus). PA stands for phytic acid. Note, A, B, C, and D are hypothetical locus.

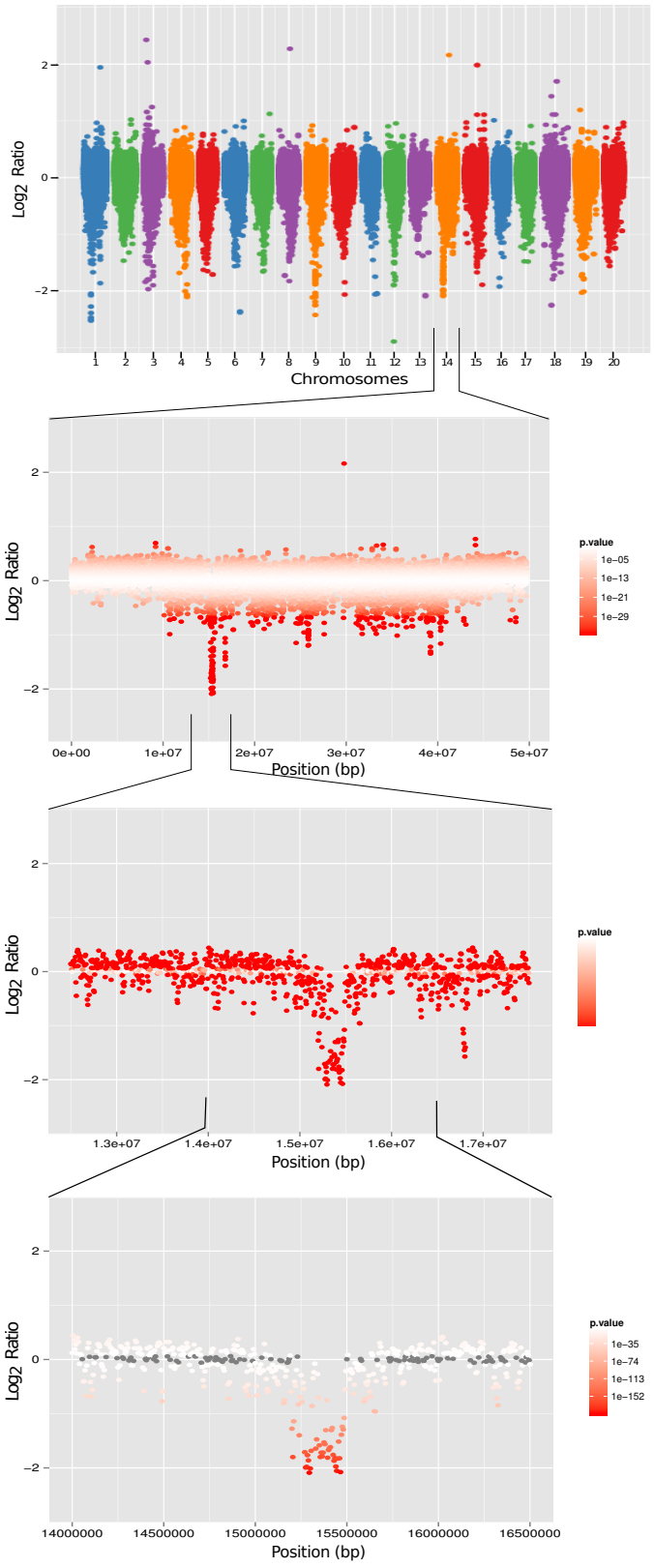


Figure 4.2: Summary of copy number variation analysis using CNV-Seq tool. Each data point (dot) represents log2 ratio for each sliding window. Top panel: Manhattan plot indicating the "$\log_2$ ratio" of read counts from 1MWT and *1mlpa* for 20 soybean chromosomes. Last three panels indicate zoomed-in $\log_2$ ratio for chain of CNV events on chromosome 14.

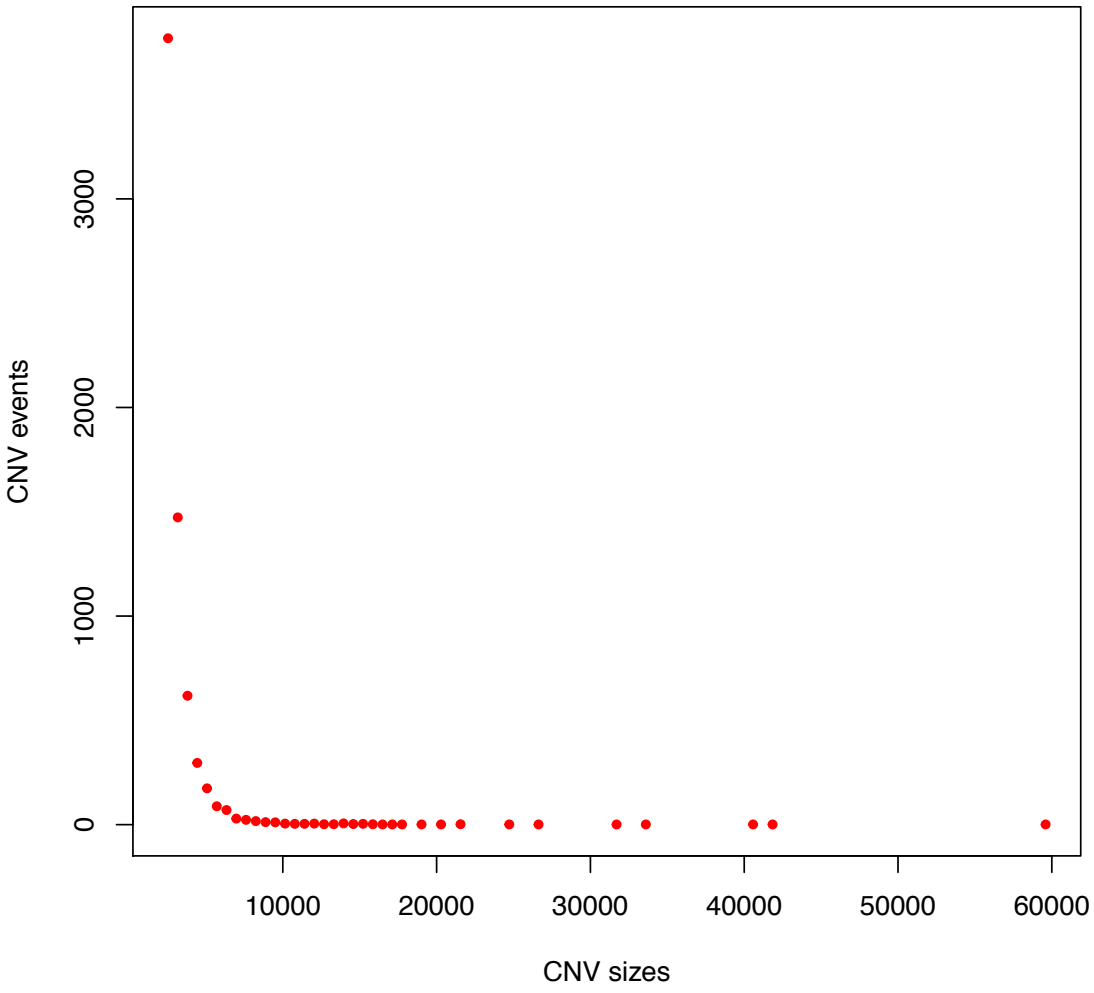


Figure 4.3: Distribution of CNV event sizes. X-axis indicates size or length of CNV events, whereas Y-axis represents the CNV count. More then 92% of the CNV events were less 5000 bp in size.

CHAPTER 5

Candidate gene sequence analyses towards identifying *Rsv3*-type resistance to *Soybean mosaic virus*

Neelam Redekar[1], Elizabeth Clevinger[1], M. A. Laskar[2], Ruslan Biyashev[1], Roderick V. Jensen[3], Soon Chun Jeong[4], Sue Tolin[5], M. A. Saghai Maroof[1]§

[1]Department of Crop and Soil Environmental Sciences, Virginia Tech. [2]Biotechnology Department, St. Anthony's College, Shillong, India. [3]Department of Biological Sciences, Virginia Tech. [4]Korea Research Institute of Bioscience and Biotechnology, Korea. [5]Department of Plant Pathology, Physiology, and Weed Sciences, Virginia Tech. §Corresponding author: smaroof@vt.edu

This Chapter is to be submitted for publication in *The Plant Genome*.

ABSTRACT

Soybean mosaic virus (SMV) is a seed-borne disease ubiquitously found in most of the soybean growing fields worldwide, causing considerable yield reduction and poor seed quality. Use of durable genetically resistant cultivars is the most promising solution to manage this disease. Three genetic loci, each containing a single dominant resistance gene, viz., *Rsv1*, *Rsv3*, and *Rsv4*, are known to confer strain-specific resistance to SMV. The *Rsv3* locus confers resistance to the most virulent strain groups of SMV (G5-G7). This *Rsv3* gene locus is comprised of a cluster of 5 CC-NB-LRR resistance genes. High sequence similarity between these *Rsv3*-candidates poses challenges to further narrow down the *Rsv3* locus. We performed comparative sequence analyses with *Rsv3*-candidate NB-LRR genes to compare *Rsv3*-candidate gene sequences between the susceptible Williams82 line and a group of three resistant lines, including L29, Hwangkeum, and RRR. We identified about 150 single nucleotide polymorphisms and five insertion-deletion type polymorphisms within a single candidate NB-LRR gene (Glyma14g38533). The coding sequence of this gene was highly conserved in the three resistant lines. Several insertion-deletion type polymorphisms were found conserved in 18 resistant lines and six susceptible lines, suggesting that the Glyma14g38533 gene allele in resistant lines is considerably different than that in susceptible lines. About 95% of these polymorphic sites in the Glyma14g38533 gene were accumulated in the LRR domain encoding region, which is essential for pathogen recognition via protein-protein interactions. These findings suggest Glyma14g38533 gene is the most likely candidate for *Rsv3*.

KEYWORDS

Soybean mosaic virus; Rsv3; NB-LRR; LRR domain

INTRODUCTION

Soybean mosaic virus (SMV) is a seed-borne disease found in the majority of soybean fields, affecting seed quality and yield. Deploying durable disease-resistant soybean varieties can prevent economic losses associated with this disease. The SMV strains identified in the United States are classified into seven strain groups, G1 to G7, based on their virulence [Cho and Goodman, 1979]. Three resistant gene loci, *Rsv1* on chromosome 13, *Rsv3* on chromosome 14, and *Rsv4* on chromosome 2, have been identified in soybean [*Glycine max* (L.) Merr.], which exhibit resistance to different SMV strains [Kiihl and Hartwig 1979, Buzzell and Tu 1984, Buss, et al. 1989, Buss, et al. 1997, Hayes, et al. 2000, Liao, et al. 2002]. All three *Rsv* loci possess a single dominant resistant gene that helps combat SMV infection. The *Rsv1* alleles tend to confer extreme resistance (symptomless) to lower numbered SMV strains G1-G4, but condition a necrotic or mosaic reaction to higher numbered SMV strains G5-G7 [Chen, et al. 1991, Tucker, et al. 2009]. The *Rsv3* gene, on the contrary, confers extreme resistance to higher numbered SMV strains G5-G7 [Jeong, et al. 2002, Ma, et al. 2002]. *Rsv4* gene provides early resistance to all SMV strains; however, the virus overcomes this resistance, resulting in late susceptibility symptoms [Chen, et al. 1993, Ma, et al. 1995, Hayes, et al. 2000].

Several studies have been conducted to understand the SMV-soybean interactions. Babu et al. (2007) reported global transcriptional changes in SMV-susceptible cultivar 'Williams82' at 7, 14, and 21 hours post inoculation (hpi) with the SMV-G2 strain using microarray technology. The transcripts related to the defense response were found down regulated in early stages of infection, suggesting the cause of symptom development in Williams 82 [Babu, et al. 2008]. In contrast to this, Yang et al. (2010) identified several proteins that accumulate in the SMV-resistant cultivar 'Kefeng No. 1' at 4 hpi with the SMV-JN17 strain using a proteomics

approach. Among these were proteins belonging to several functional categories, including the defense response [Yang, et al. 2011]. This means that the SMV-soybean interactions in susceptible and resistant hosts are known to trigger different biological responses [Babu, et al. 2008, Yang, et al. 2011]. A recent study by Seo et al. (2014) reported the association of type 2C protein phosphatase genes with *Rsv3*-mediated extreme resistance (ER) response in SMV-resistant cultivar 'L29' at 8 hpi with SMV-G5H inoculation. The SMV-soybean interaction at the *Rsv3* loci activates ABA signaling, which regulates type 2C protein phosphatase and callose biosynthesis to prevent infection [Seo, et al. 2014]. However, the gene at the *Rsv3* locus that initiates these defense responses is still unknown.

The most characterized disease Resistance (*R*) gene family encodes for nucleotide-binding leucine-rich repeat (NB-LRR) proteins. The *R* gene-mediated defense response is initiated when pathogen effectors (or *avr* proteins) or the co-factors that bind to these effectors are detected in the host [Bent and Mackey 2007]. This detection is facilitated by the binding specificity of the LRR domain within the NB-LRR protein. Both the *Rsv1* and *Rsv3* loci are associated with a cluster of genes that encode for NB-LRR proteins with an N-terminal coiled-coil domain [Jeong, et al. 2002, Hayes, et al. 2004, Suh, et al. 2011, Wang, et al. 2011]. In particular, there are five NB-LRR genes viz., Glyma14g38500, Glyma14g38516, Glyma14g38533, Glyma14g38561, and Glyma14g38586 within the 154-Kb *Rsv3* locus [Suh, et al. 2011]. Fine mapping of this locus, to identify a single NB-LRR gene that encodes *Rsv3*-type resistance is challenging, due to the high sequence similarity of these genes. However, due to NB-LRR gene evolution under selective pressure, the disease resistant and susceptible lines are expected to have different functional forms (active and non-active) of NB-LRR genes.

The current study is focused on downsizing the number of *Rsv3*-candidate NB-LRR genes within a 154-kb span encompassing the *Rsv3* locus. Towards this goal, we employed a comparative sequencing approach to compare the *Rsv3*-candidate gene sequences between susceptible Williams82 against three resistant soybean lines, 'Hwangkeum', L29, and 'RRR'. The *Rsv3*-candidate, Glyma14g38533, showed a conserved gene sequence in these three resistant lines. Several genetic features, in the form of insertion/deletion and single nucleotide polymorphisms, identified in this gene were further validated in 18 resistant and 6 susceptible lines. The genetic features conserved in 18 resistant lines were distinct from those conserved in susceptible lines. This suggests there are two allelic forms of the Glyma14g38533 gene; one is conserved in resistant lines, and the other in susceptible lines. The majority of these genetic features, polymorphic between resistant and susceptible lines, were found within the leucine rich repeat domain, a region often involved in pathogen recognition. Therefore, it appears that the Glyma14g38533 gene is the most likely candidate for *Rsv3*.

MATERIAL AND METHODS

Plant growth, virus inoculation, and tissue sampling

A viable SMV-G7 strain was maintained on the *Rsv3*-genotype susceptible soybean line, 'Hutcheson' and/or 'Essex' as a source of inoculum. Inoculum was prepared as 1:10 (w/v) of SMV-G7 infected leaf tissue in inoculation buffer, which was composed of 1% Celite545 in 0.01M sodium phosphate buffer, pH 7.0 (Fischer Scientific Inc., Waltham, MA). The *Rsv3*-genotype resistant soybean line, L29, was planted in the greenhouse and grown until first trifoliolates were fully expanded. These trifoliolate leaves were then inoculated with virus inoculum by gently rubbing the SMV-G7 inoculum with a pestle. Another set of plants was

inoculated in a similar way with inoculation buffer (without the virus). Inoculated trifoliolate leaves were sampled after 0, 1, 2, 3, and 24 hpi and stored at -80°C until further processing. Total RNA was extracted from these samples using the RNeasy Plant Mini Kit, with on-column DNase digestion (QIAGEN Inc., Valencia, CA). We used this total RNA for two purposes: (1) the RNA-Seq experiment, which included total RNA from only virus-inoculated 1-, 2-, and 3-hpi samples; and (2) quantitative real-time PCR, which included total RNA from all samples.

RNA sequencing and data analysis

Total RNA from virus-inoculated 1-, 2-, and 3-hpi samples was used for RNA-Seq data generation. RNA Integrity (RIN) value was estimated for these total RNA samples using an Agilent 2100 Bioanalyzer system (Agilent Technologies Inc., Santa Clara, CA). Samples with an RIN value above 8.0 were further processed at Virginia Bioinformatics Institute, Virginia Tech, to construct mRNA libraries. Two technical replicates per library were sequenced with 50PE cycles using an Illumina HiSeq1000 sequencer (Illumina Inc., San Diego, CA). Quality control steps were performed on the sequencing data to obtain high quality sequencing reads using FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit). Initially, the sequencing reads with 70% of the bases falling below the threshold quality score (Q=25) were filtered out. These quality filter passing sequencing reads were further processed to remove existing adapter sequences and ambiguous bases. The pre-processed sequencing reads were used for two purposes: (1) estimate transcript abundance for *Rsv3*-candidate genes; and (2) identify polymorphisms in *Rsv3*-candidate gene sequence assemblies.

Estimate transcript abundance of *Rsv3*-candidate genes

The pre-processed reads were aligned to the reference Williams82 genome sequence assembly 1.1 and annotation Glyma1.1 using TopHat2 (v2.0.8), a splice-aware aligner tool [Kim, et al. 2013]. Based on these sequence alignments, read counts (number of reads mapped to a gene) for each gene were estimated using an HTSeq-count [Anders, et al. 2015]. The reads mapping to multiple sites within the reference genome were eliminated, while estimating the number of reads mapped to a given gene. This count data set was normalized into reads per kilobase per million (RPKM), which is the quantitative approximation of the transcript abundance. The RPKM value was estimated as follows: RPKM = reads mapped to a transcript per unit length of transcript (kb)/total reads mapped (millions). Transcript lengths were obtained from Phytozome 9.0v (www.phytozome.net).

Identification of polymorphisms in *Rsv3*-candidate gene sequence assemblies

For this task, sequencing data from three resistant lines was used: (1) pre-processed sequencing reads from the L29-RNASeq experiment described above; (2) Hwangkeum genomic sequencing reads obtained from Korea Research Institute of Bioscience and Biotechnology, Korea [Chung, et al. 2014]; and (3) RRR genomic sequencing data [Maroof, et al. 2008, Chung, et al. 2014]. The RRR genome was sequenced with 101PE cycles using an Illumina HiSeq1000 sequencer at Virginia Bioinformatics Institute, Virginia Tech (Illumina Inc., San Diego, CA). The sequencing reads from these three resistant lines were separately aligned to the *Rsv3*-locus sequence from the reference Williams82 genome assembly 1.1 using Geneious Pro 5.7.7 software (www.geneious.com). The *Rsv3*-locus spans a 154-kb region between nucleotide positions: 47621000-47738999 on chromosome 14 [Suh, et al. 2011]. Reads that mapped to the

Rsv3-locus were further used to assemble transcript sequences of all five *Rsv3*-candidate genes (www.geneious.com). The *Rsv3*-candidate transcript assembly sequences were further pairwise aligned with those of the reference Williams 82, annotation version Glyma1.1 using ClustalW, to identify the nucleotide positions polymorphic between the resistant and susceptible Williams82 line [Larkin, et al. 2007]. These polymorphic sites were compared between different resistant lines to identify conserved alleles.

Validating genetic polymorphisms

Eighteen *Rsv3*-genotype (resistant) and 6 *rsv3*-genotype (susceptible) soybean [*Glycine max* (L.) Merr.] lines, originated from more than 5 different countries, were used for validation of genetic polymorphisms from *Rsv3*-candidate genes, as reported in Supplementary Table E1 [Li, et al. 2010]. Genomic DNA was extracted from different soybean cultivars using the CTAB method [Yu, et al. 1994]. Primers to amplify polymorphic positions from *Rsv3*-candidate genes were designed using Primer3 software and are reported in Supplementary Table E2 [Koressaar and Remm 2007, Untergasser, et al. 2012]. PCR was performed in 50 µl reaction volume, comprising of 5 µl 10X reaction buffer, 1.5 µl 50 mM $MgCl_2$, 8 µl 2mM dNTPs, 0.5 µl 10 mM primers each, and 0.5 µl Taq polymerase (Invitrogen Inc., Waltham, MA). Amplification products were purified using a QIAquick PCR purification kit (QIAGEN Inc., Valencia, CA). Purified PCR products were sequenced and analyzed using Geneious Pro software to identify genotypes at polymorphic sites of *Rsv3*-candidate genes (www.geneious.com).

RESULTS

Relative gene expression of *Rsv3*-candidate NB-LRR transcripts

Many NB-LRR genes are expressed at a minimal level at any given time to facilitate detection of pathogens or pathogen-derived factors. However, many studies have reported an increase in their expression levels upon infection [Kang, et al. 2012, Kar, et al. 2013, Li, et al. 2013]. Therefore, in order to compare the relative abundance of candidate NB-LRR transcripts from the *Rsv3*-locus, we performed sequencing of three cDNA libraries prepared from SMV-G7 strain-inoculated first trifoliolate leaves, sampled 1-, 2-, and 3- hpi from the resistant soybean line, L29. Each sample library was sequenced twice on separate sequencer lanes, resulting in about 1.97 billion paired-end read sequences in total (Table 5.1). This sequencing data set was preprocessed to obtain more than 1.32 billion high quality reads, without any ambiguous bases or adapter sequences. These high-quality sequences included more than 921 million paired-end and 399 million single-end reads. These paired- and single-end reads were analyzed separately, but in a similar manner. Reads were first mapped to the soybean reference genome, Williams82, using a splice-aware alignment tool, TopHat2 [Kim, et al. 2013]. Altogether, over 877 million paired- and 382 million single-reads were mapped. We estimated RPKM as a measure of transcript abundance for five *Rsv3*-candidate NB-LRR genes: Glyma14g38500, Glyma14g38516, Glyma14g38533, Glyma14g38561, and Glyma14g38586. Figure 5.1 shows the relative transcript abundance of *Rsv3*-candidate NB-LRR genes at 1-, 2-, and 3-hpi with the SMV-G7 strain of the virus. We observed that SMV inoculation regulates the expression of NB-LRR genes in L29. The transcript abundance of all *Rsv3*-candidate genes, except Glyma14g38500 was significantly different at three sampling stages. At 3 hpi, the transcript levels declined in Glyma14g38500, Glyma14g38516, and Glyma14g38586, as opposed to increased levels of Glyma14g38533 and

Glyma14g38561. The Glyma14g38533 gene showed highest transcript abundance at each time point, with a total of a 3.3 fold (highest) increase at three hours post-inoculation. Other *Rsv3*-candidate genes showed lower transcript abundance as compared to Glyma14g38533 gene. It is possible that this higher transcript abundance of Glyma14g38533 gene is associated with or required for *Rsv3*-type resistance. However, at this point, the gene expression profiles do not seem to provide any information that will be useful to eliminate any *Rsv3*-candidates. Moreover, the experimental design does not allow for the study of gene expression profiles of these candidates in susceptible lines.

Comparative NB-LRR gene sequence analyses

NB-LRR genes are present in both susceptible and resistant lines, usually in clusters, with high sequence similarity among the members within each cluster. Despite high sequence similarity, allelic forms of these genes differ between susceptible and resistant cultivars. Any allelic features shared between multiple resistant lines may contribute to the resistance phenotype. Therefore, in order to study allelic features in *Rsv3*-candidates, we compared NB-LRR gene sequences between susceptible (Williams82) and resistant (L29, Hwangkeum, and RRR) soybean lines. First, we used genome re-sequencing data of Hwangkeum cultivar, obtained from Korea Research Institute of Bioscience and Biotechnology, Korea [Chung, et al. 2014]. The Hwangkeum line, which originated from 'Suweon97', contains both the *Rsv1* and *Rsv3* resistance genes [Chen, et al. 2002, Yu, et al. 2008, Jeong and Jeong 2014]. The *Rsv3*-candidate NB-LRR gene sequences of the Hwangkeum cultivar were assembled based on annotated gene model (version 1.1) sequences from the reference Williams82 genome. *Rsv3*-candidate gene names for the genome assembly 1 annotations are as shown in Supplementary Table E3. We

identified several polymorphisms in the form of single nucleotide polymorphisms (SNPs) and insertion-deletions (INDELs) in the five *Rsv3*-candidate genes, as summarized in Figure 5.2. Table 5.2 represents 17 non-synonymous SNPs identified in four *Rsv3*-candidates (Glyma14g38500, Glyma14g38561, Glyma14g38516, and Glyma14g38586). All SNPs identified in these candidate genes were present in only a single exon, i.e., SNPs in Glyma14g38500, Glyma14g38561, Glyma14g38516, and Glyma14g38586 genes were present in exon 1, 2, 2, and 3, respectively (Figure 5.2). We did not observe any INDELs for these candidate genes.

Nearly 146 SNPs and several INDELs were observed for the Glyma14g38533 gene sequence comparison between Hwangkeum (resistant) and Williams82 (susceptible) (Supplementary Table E4). This suggested that Glyma14g38533 gene alleles of Hwangkeum and Williams82 must differ considerably. To further explore this candidate gene, we employed transcript sequences from the L29 (resistant) RNA-Seq experiment, and genomic re-sequencing data from the RRR line (resistant). We de novo assembled Glyma14g38533 coding sequence from the L29 and RRR genomes and identified about 140 and 139 SNPs, respectively, by comparing these assembled sequences with that of Williams82. Altogether, we identified a total of 150 unique single nucleotide positions that were polymorphic between the resistant lines (Hwangkeum, L29, and RRR) and susceptible Williams82 line (Supplementary Table E4). Of these 150 polymorphic sites, 136 SNPs were non-synonymous (Table 5.3; Supplementary Table E4). Figure 5.3 summarizes the SNP overlap between Williams82, L29, Hwangkeum, and RRR. Over 131 of the SNPs were a perfect match between the three resistant lines, suggesting substantial gene sequence homology. In addition to SNPs, four in-frame deletions (sizes: 3, 9, 12, and 39 bp), and two in-frame insertions (sizes: 6, and 21 bp) were identified in

Glyma14g38533 coding sequence of resistant lines, predominantly in exons-1 and -2 (Table 5.4; Figure 5.2; Supplementary Table E4). No polymorphisms were identified in exon-3, which is merely 8 bp long. These observations are consistent with the presence of two allelic forms of the Glyma14g38533 gene; one of these alleles is conserved in the three resistant lines L29, Hwangkeum, and RRR, while the other is observed in susceptible Williams 82.

Validation of genetic polymorphisms

The genetic alleles that confer resistance are usually conserved in several resistant lines. This means, if any polymorphic allele (from previous comparative sequence analyses) is conserved between different resistant lines, it is more likely to be associated with resistance. Therefore, we tested the genotype of several polymorphic sites in additional resistant (known to contain *Rsv3*) and susceptible soybean lines. Disease response and genotype information for the soybean lines used for this analysis are reported in Supplementary Table E1. SNP verification results for Glyma14g38500, Glyma14g38561, Glyma14g38516, and Glyma14g38586 are reported in Table 5.5. For the Glyma14g38500 gene, all susceptible and resistant lines showed distinct SNP genotypes, except 'Archer', whose SNP genotypes were the same as that of resistant lines. To re-verify Archer SNP genotypes, we mapped Archer exome-sequencing reads from McHale et al. (2012) to the reference Williams82 Glyma14g38500 gene sequence [McHale, et al. 2012]. All 5 SNPs in the Glyma14g38500 gene, except the one at position 47632978, were perfect matches between our experimental Archer and the one from the McHale et al. (2012) paper. Although some genotype-phenotype correlation exists for the Glyma14g38500 gene (considering Archer as an outlier), the expression level does not change with SMV-G7 inoculation, as shown in Figure 5.1. For the Glyma14g38561 gene, we observed multiple

variations of SNP genotypes, while for Glyma14g38586, all susceptible and resistant lines showed the same SNP genotype, except Williams 82. No perfect genotype-phenotype correlation was observed. Therefore, none of the SNPs from the Glyma14g38561 and Glyma14g38586 genes should be associated with resistance. On the other hand, for the Glyma14g38533 gene, INDELs could distinguish resistant and susceptible lines. Eighteen resistance lines, including L29, 'Touson140' Suweon97, RRR, 'Hourei', 'Harosoy', 'Columbia', 'Hardee', 'VIR5532', 'Paoting', 'PLSO-70', 'PLSO-63', 'PI323555', 'PI323556', 'OCB-81', 'PI91346', 'PI61947' and 'Graine Jaune Uni' showed the presence of 39-bp, 3-bp, and 12-bp deletions when compared to six susceptible lines, viz., Williams82, 'Lee68', Essex, Archer, Hutcheson, and 'York'. The majority of these resistant soybean lines were reported in Li et al. (2010), as originate from more than 5 different countries and they possess the *Rsv3* gene [Li, et al. 2010]. Figure 5.4 shows the 39-bp deletion region from the six susceptible and 18 resistant soybean lines tested. This 39-bp deletion is co-segregating with the *Rsv3* gene in the L29 (*Rsv3*) x Lee68 (*rsv3*) population and L29 (*Rsv3*) x Sowon (*rsv3*) population from Suh et al. (2011) (Supplementary Figure F1). SNPs surrounding this deletion were exactly identical in all 18 resistant lines. In summary, conservation of a Glyma14g38533 allele in multiple resistant lines implies that this gene is the most promising candidate to encode *Rsv3*.

Comparative functional protein domain analysis

NB-LRR proteins contain three domains: a variable N-terminal domain (either Toll/Interleukin-1 Receptor (TIR) or Coiled-Coil (CC)), a central NB-ARC domain with a nucleotide-binding (NB) site, and a C-terminal leucine-rich repeat (LRR) domain [Heil and Baldwin 2002, McHale, et al. 2006]. The *Rsv3*-candidate genes encode for CC-NB-LRR proteins

[Suh, et al. 2011]. Previous studies have suggested that the N-terminal CC domain, along with the NB-ARC domain are involved in protein conformation change and defense signal transduction upon pathogen perception by the LRR domain [McHale, et al. 2006]. To investigate the possible impact of the SNPs/INDELs described above on the Glyma14g38533 gene function, we compared the functional domains from resistant (L29 and Hwangkeum) and susceptible (Williams82) lines. The N-terminal CC and NB-ARC domains of Glyma14g38533 contained less than 5% of the total SNPs identified, while the remaining 95% were present within the LRR domain. This is often involved in pathogen effector recognition via protein-protein interactions. The large number of polymorphisms associated with the LRR domain of Glyma14g38533 is consistent with the alleles found in the resistant/susceptible lines differing in their specificity with respect to pathogen effectors.

To further explore this theory, we compared the LRR domain sequences of the Glyma14g38533 alleles found in resistant-L29 and susceptible-Williams82 lines. We identified 11-residue and 12-residue variants of the LRR core repeats region with consensus sequences of LxxLxLxxNxL, and LxxLxLxxCxxL, respectively, where "L" is Leu (L), Ile (I), Val (V), or Phe (F); "N" is Asn (N), Thr (T), Ser (S), or Cys (C); and "C" is Cys (C) or Ser (S), and "x" is any amino acid [Ohyanagi and Matsushima 1997, Kajava 1998, Kobe 2001, Kajava, et al. 2008, Matsushima, et al. 2010]. We allowed for one mismatch within conserved residues of the consensus pattern so as to identify LRR variants, if any. A total of 22 LRR repeat motifs were identified in the Glyma14g38533 protein (Table 5.6, Figure 5.5, Supplementary Table E5). The INDELs between the resistant and susceptible forms of the Glyma14g38533 gene occurred between LRR repeats, and the number and location of LRR repeats were identical for both L29 and Williams 82. Among the 22 LRR repeat motifs, we identified several 11-residue

(LxxLxLxxLxL) and 12-residue (LxxLxLxxLxxL) LRR variant motifs that were conserved between L29 and Williams82. A total of 18 and 15 LRR repeats from L29 and Williams82, respectively, were a perfect match to the consensus pattern. The LRR1, LRR2, LRR3, LRR14, and LRR15 had one mismatch, while LRR17 had two mismatches within the conserved residues of the consensus LRR motif sequence in Williams82; in L29 only LRR3, LRR15, and LRR17 had one mismatch within consensus residues. All LRRs, except LRR15, displayed several non-conserved residues polymorphic between L29 and Williams82. It is possible that these LRR variations result into different pathogen recognition specificity of the Glyma14g38533 encoded NB-LRR protein in L29 and Williams82. Since RRR contains the *Rsv3* gene from L29, all the LRR repeats were identical between L29 and RRR. However, in the case of Hwangkeum, LRR repeat number-7 and -11, exhibited different "x" or non-conserved residues of the consensus sequence than that in L29 (Supplementary Table E6). This suggests that the conserved sequences of LRR motifs are shared between L29, Hwangkeum, and RRR. In summary, based on the distribution of polymorphisms described above, the pathogen recognition specificity of the LRR domain is likely different in the NB-LRR proteins encoded by the susceptible and resistant lines.

DISCUSSION

The *Rsv3* locus, conferring resistance against highly virulent strains of the *Soybean mosaic virus,* likely contains a single dominant R gene that initiates the defense response. Based on genetic mapping studies and the soybean genome sequence, Suh et al. identified the presence of five CC-NB-LRR genes within the *Rsv3* locus [Suh, et al. 2011]. These genes belong to the most characterized R gene family in plants. The NB-LRR proteins mediate pathogens and trigger R gene-mediated defense responses. The NB-LRR genes are usually expressed at minimal levels

under non-stress conditions to avoid the fitness cost of constitutive R gene activation [Heil 2002, Heil and Baldwin 2002]. Pathogen infection and several components of basal defense, such as salicylic acid, are known to induce the R gene expression, allowing robust activation of R gene-mediated defense [Shirano, et al. 2002, Xiao, et al. 2003, Gu, et al. 2005, Nandety, et al. 2013]. Several studies have reported an increase in NB-LRR gene expression levels upon infection [Kang, et al. 2012, Kar, et al. 2013, Li, et al. 2013].

In order to estimate gene expression profiles of *Rsv3*-candidate NB-LRR genes in the presence of the SMV infection, we performed RNA-Seq experiments with trifoliolate leaves of resistant soybean L29 taken 1-, 2-, and 3-hpi. We observed that the rate of transcription of *Rsv3*-candidate genes was changed significantly from 1-hpi to 2-hpi and from 1-hpi to 3-hpi. *Rsv3*-candidate Glyma14g38533 gene expression was increased 3.3 fold during the first 3 hours of inoculation, and was the most abundant transcript of the *Rsv3*-candidates. Although our current experimental setup fails to provide a comparison of *Rsv3*-candidate gene expression profiles between susceptible and resistant lines, a significant increase in Glyma14g38533 transcript abundance upon infection cannot be ignored, as it may be associated with enhancing the resistance reaction. These observations are contrary to the gene expression assay in 'Dabaima', which contains the *Rsc4* gene locus that coincides with the *Rsv3*-locus [Wang, et al. 2011]. Given the overlap of the *Rsv3* and *Rsc4* locus within the soybean genome, it is more likely that a single gene within the locus encodes for both *Rsv3* and *Rsc4*. The *Rsc4* locus (<100-kb), conferring resistance to the SMV strain SC4 from China, contains three NB-LRR genes, viz., Glyma14g38500, Glyma14g38516, Glyma14g38533 [Wang, et al. 2011]. They reported a decrease in expression of Glyma14g38510 and Glyma14g38533 genes in both the resistant

Dabaima and susceptible 'Nannong1138-2' lines from 1-hpi to 2-hpi with the SMV SC4 strain. The Glyma14g38500 gene expression was not even detected in this study.

NB-LRR genes show high sequence similarity and they co-evolve with pathogens to accumulate distinct haplotypes, or genetic features [Bent and Mackey 2007, Marone, et al. 2013]. The genetic features conferring fitness (in this case, resistance) become fixed during the course of evolution [Marone, et al. 2013]. Therefore, any genetic features distinct between resistant and susceptible forms of the NB-LRR gene have the potential to produce a differential disease response. We performed comparative sequence analyses on such genetic features by comparing the *Rsv3*-candidates NB-LRR gene sequences, which were 83-92% similar to each other, from resistant and susceptible lines. Four *Rsv3*-candidates (Glyma14g38500, Glyma14g38561, Glyma14g38516, and Glyma14g38586) showed about 17 SNP-type differences between a susceptible and a resistant line. All these SNP positions were genotyped in several other SMV-susceptible and -resistant lines; however, none of the SNP alleles were able to distinguish between the two groups. The *Rsv3*-candidate Glyma14g38533 gene, on the other hand, showed 150 SNP- and 6 INDEL-type differences when comparing its gene sequence from a susceptible line with that from three resistant lines. More than 87% of the genetic features of this gene were conserved between the three resistant lines. Several INDELs and SNPs surrounding these INDELS were found conserved in about 18 tested resistant lines. This suggested that the Glyma14g38533 gene is considerably different in resistant lines as compared to the susceptible line. At the same time, the Glyma14g38533 gene sequence is highly conserved between different resistant lines.

The *Rsv3*-candidate genes encode for NB-LRR proteins with coiled-coil N-terminal domains. The N-terminal region of the Glyma14g38533 protein, comprising of the CC and NB-

ARC domains, is likely involved in protein activation and signal transduction post-pathogen recognition, whereas the LRR domain is likely involved in recognizing the pathogen via a protein-protein interaction [McHale, et al. 2006]. The nucleotides encoding LRR domains tend to have a higher rate of variability than those encoding NBS regions, so as to recognize evolving pathogens [Mondragon-Palomino, et al. 2002, Kuang, et al. 2004, Marone, et al. 2013]. About 95% of the polymorphic genetic features identified in Glyma14g38533 were found within the LRR domain, with an increased ratio of non-synonymous to synonymous nucleotide substitutions (Supplementary Table E4). These observations are consistent with diversifying selection maintaining variation in the solvent-exposed residues within the LRR domain [McDowell, et al. 1998, Michelmore and Meyers 1998]. LRR domain variations also included all of the INDELs polymorphic between resistant and susceptible lines. Dissimilarities in LRR domain length, along with other polymorphisms, may lead to differential pathogen recognition and further affect resistance specificity in susceptible versus resistant lines.

The LRR domain in NB-LRR genes consists of 20-30 LRR repeating units, which are frequently interrupted by non-LRR island regions (IRs) [Matsushima, et al. 2010]. Each LRR repeating unit is defined by a highly conserved motif, which is the 11-12 residues LRR signature sequence and a variable sequence [Kajava 1998, Matsushima, et al. 2010]. We identified 22 LRR repeating units within the Glyma14g38533 protein with few variant repeating units. Despite observation of several polymorphisms between the resistant and susceptible forms of Glyma14g38533, the total number of LRR repeating units remained unchanged. More than 57% of the SNPs identified in the Glyma14g38533 gene were accumulated across 21 LRR repeating units. This means that only one repeating unit matched between susceptible and resistant lines.

This supports our conclusion that the majority of the polymorphisms identified in Glyma14g38533 tend to modify the specificity of the LRR domain.

In summary, using comparative sequence analysis, we were able to eliminate four *Rsv3*-candidate genes, as these genes showed no unique polymorphisms between susceptible and resistant lines. In contrast, the SNP genotypes at polymorphic sites within the Glyma14g38533 gene were different between groups of susceptible and resistant lines, while being highly conserved within each group. The Glyma14g38533 gene also showed highest transcript abundance that increased upon the SMV-G7 infection period. The highly conserved gene sequence in resistant lines indicates that the Glyma14g38533 gene is most likely the candidate for *Rsv3*. Further experiments are required to functionally validate the identity of *Rsv3*. Most of the polymorphisms within the *Rsv3* candidate gene are within the LRR domain, consistent with this region conferring pathogen recognition specificity.

AUTHORS' CONTRIBUTIONS

ST, MASM, SCJ, and RVJ contributed towards the conception of the project. All authors contributed to design of the experiments. RB, EC, MAL, and NR performed the experiments. All authors have contributed towards the analyses and interpretation of the data, and drafting of the manuscript.

ACKNOWLEDGEMENTS

This work was funded by the Virginia Soybean Board. We thank Dr. Tom Ashfield, and Dr. John McDowell for their constructive comments on earlier versions of drafts. We would like to thank

support team of Advanced Research Computing (ARC) server and Translational Plant Sciences' MAGYK server at Virginia Tech.

REFERENCES

Anders S, Pyl PT, Huber W: **HTSeq-a Python framework to work with high-throughput sequencing data**. *Bioinformatics* 2015, **31**(2):166-169.

Babu M, Gagarinova AG, Brandle JE, Wang AM: **Association of the transcriptional response of soybean plants with soybean mosaic virus systemic infection**. *Journal of General Virology* 2008, **89**:1069-1080.

Bent AF, Mackey D: **Elicitors, effectors, and R genes: the new paradigm and a lifetime supply of questions**. *Annu Rev Phytopathol* 2007, **45**:399-436.

Bernard RL, Bernard RL, Nelson RL, Cremeens CR: **USDA soybean genetic collection: isoline collection**. *Soybean Genet Newsl* 1991, **18**:27-57.

Buss GR, Ma G, Chen P, Tolin SA: **Registration of V94-5152 soybean germplasm resistant to soybean mosaic potyvirus**. *Crop Science* 1997, **37**(6):1987-1988.

Buss GR, Roane CW, Tolin SA, Chen P: **Inheritance of Reaction to Soybean Mosaic-Virus in 2 Soybean Cultivars**. *Crop Science* 1989, **29**(6):1439-1441.

Buzzell RI, Tu JC: **Inheritance of Soybean Resistance to Soybean Mosaic-Virus**. *Journal of Heredity* 1984, **75**(1):82-82.

Chen P, Buss GR, Roane CW, Tolin SA: **Allelism among Genes for Resistance to Soybean Mosaic-Virus in Strain-Differential Soybean Cultivars**. *Crop Science* 1991, **31**(2):305-309.

Chen P, Buss GR, Tolin SA: **Resistance to Soybean Mosaic-Virus Conferred by 2 Independent Dominant Genes in Pl-486355**. *Journal of Heredity* 1993, **84**(1):25-28.

Chen PY, Buss GR, Tolin SA, Gunduz I, Cicek M: **A valuable gene in Suweon 97 soybean for resistance to soybean mosaic virus**. *Crop Science* 2002, **42**(2):333-337.

Cho EK, Goodman RM: **Strains of Soybean Mosaic-Virus - Classification Based on Virulence in Resistant Soybean Cultivars**. *Phytopathology* 1979, **69**(5):467-470.

Chung WH, Jeong N, Kim J, Lee WK, Lee YG, Lee SH, Yoon W, Kim JH, Choi IY, Choi HK *et al*: **Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes**. *DNA Res* 2014, **21**(2):153-167.

Gu K, Yang B, Tian D, Wu L, Wang D, Sreekala C, Yang F, Chu Z, Wang GL, White FF *et al*: **R gene expression induced by a type-III effector triggers disease resistance in rice**. *Nature* 2005, **435**(7045):1122-1125.

Gunduz I, Buss GR, Chen P, Tolin SA: **Characterization of SMV resistance genes in Tousan 140 and Hourei soybean**. *Crop Science* 2002, **42**(1):90-95.

Hayes AJ, Jeong SC, Gore MA, Yu YG, Buss GR, Tolin SA, Maroof MA: **Recombination within a nucleotide-binding-site/leucine-rich-repeat gene cluster produces new variants conditioning resistance to soybean mosaic virus in soybeans**. *Genetics* 2004, **166**(1):493-503.

Hayes AJ, Ma GR, Buss GR, Maroof MAS: **Molecular marker mapping of RSV4, a gene conferring resistance to all known strains of soybean mosaic virus**. *Crop Science* 2000, **40**(5):1434-1437.

Heil M: **Fitness costs of induced resistance: emerging experimental support for a slippery concept**. *Trends in Plant Science* 2002, **7**(2):61-67.

Heil M, Baldwin IT: **Fitness costs of induced resistance: emerging experimental support for a slippery concept**. *Trends in Plant Science* 2002, **7**(2):61-67.

Jeong N, Jeong S-C: **Multiple genes confer resistance to soybean mosaic virus in the soybean cultivar Hwangkeum**. *Plant Genetic Resources* 2014, **12**(S1):S41-S44.

Jeong SC, Kristipati S, Hayes AJ, Maughan PJ, Noffsinger SL, Gunduz I, Buss GR, Maroof MAS: **Genetic and sequence analysis of markers tightly linked to the soybean mosaic virus resistance gene, Rsv3**. *Crop Science* 2002, **42**(1):265-270.

Kajava A: **Structural diversity of leucine-rich repeat proteins**. *Journal of molecular biology* 1998, **277**(3):519-527.

Kajava AV, Anisimova M, Peeters N: **Origin and evolution of GALA-LRR, a new member of the CC-LRR subfamily: from plants to bacteria?** *PLoS One* 2008, **3**(2):e1694.

Kang YJ, Kim KH, Shim S, Yoon MY, Sun S, Kim MY, Van K, Lee SH: **Genome-wide mapping of NBS-LRR genes and their association with disease resistance in soybean**. *BMC Plant Biol* 2012, **12**:139.

Kar B, Nanda S, Nayak PK, Nayak S, Joshi RK: **Molecular characterization and functional analysis of CzR1, a coiled-coil-nucleotide-binding-site-leucine-rich repeat R-gene from Curcuma zedoaria Loeb. that confers resistance to Pythium aphanidermatum**. *Physiological and Molecular Plant Pathology* 2013, **83**:59-68.

Kiihl RAS, Hartwig EE: **Inheritance of Reaction to Soybean Mosaic-Virus in Soybeans**. *Crop Science* 1979, **19**(3):372-375.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome Biol* 2013, **14**(4):R36.

Kobe B: **The leucine-rich repeat as a protein recognition motif**. *Current Opinion in Structural Biology* 2001, **11**(6):725-732.

Koressaar T, Remm M: **Enhancements and modifications of primer design program Primer3**. *Bioinformatics* 2007, **23**(10):1289-1291.

Kuang H, Woo S-S, Meyers BC, Nevo E, Michelmore RW: **Multiple Genetic Processes Result in Heterogeneous Rates of Evolution within the Major Cluster Disease Resistance Genes in Lettuce**. *The Plant Cell Online* 2004, **16**(11):2870-2894.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**(21):2947-2948.

Li D, Chen P, Alloatti J, Shi A, Chen YF: **Identification of New Alleles for Resistance to Soybean Mosaic Virus in Soybean**. *Crop Science* 2010, **50**(2):649-655.

Li J, Zhang QY, Gao ZH, Wang F, Duan K, Ye ZW, Gao QH: **Genome-wide identification and comparative expression analysis of NBS-LRR-encoding genes upon Colletotrichum gloeosporioides infection in two ecotypes of Fragaria vesca**. *Gene* 2013, **527**(1):215-227.

Liao L, Chen P, Buss GR, Yang Q, Tolin SA: **Inheritance and allelism of resistance to soybean mosaic virus in Zao18 soybean from China**. *Journal of Heredity* 2002, **93**(6):447-452.

Ma G, Chen P, Buss GR, Tolin SA: **Genetic-Characteristics of 2 Genes for Resistance to Soybean Mosaic-Virus in P1486355 Soybean**. *Theoretical and Applied Genetics* 1995, **91**(6-7):907-914.

Ma G, Chen P, Buss GR, Tolin SA: **Complementary action of two independent dominant genes in Columbia soybean for resistance to soybean mosaic virus**. *Journal of Heredity* 2002, **93**(3):179-184.

Marone D, Russo MA, Laido G, De Leonardis AM, Mastrangelo AM: **Plant Nucleotide Binding Site-Leucine-Rich Repeat (NBS-LRR) Genes: Active Guardians in Host Defense Responses**. *Int J Mol Sci* 2013, **14**(4):7302-7326.

Maroof MAS, Jeong SC, Gunduz I, Tucker DM, Buss GR, Tolin SA: **Pyramiding of soybean mosaic virus resistance genes by marker-assisted selection**. *Crop Science* 2008, **48**(2):517-526.

Matsushima N, Miyashita H, Mikami T, Kuroki Y: **A nested leucine rich repeat (LRR) domain: the precursor of LRRs is a ten or eleven residue motif**. *BMC Microbiol* 2010, **10**:235.

McDowell JM, Dhandaydham M, Long TA, Aarts MGM, Goff S, Holub EB, Dangl JL: **Intragenic Recombination and Diversifying Selection Contribute to the Evolution of Downy Mildew Resistance at the RPP8 Locus of Arabidopsis**. *The Plant Cell Online* 1998, **10**(11):1861-1874.

McHale L, Tan X, Koehl P, Michelmore RW: **Plant NBS-LRR proteins: adaptable guards**. *Genome Biol* 2006, **7**(4):212.

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM: **Structural variants in the soybean genome localize to clusters of biotic stress-response genes**. *Plant Physiol* 2012, **159**(4):1295-1308.

Michelmore RW, Meyers BC: **Clusters of Resistance Genes in Plants Evolve by Divergent Selection and a Birth-and-Death Process**. *Genome Research* 1998, **8**(11):1113-1130.

Mondragon-Palomino M, Meyers BC, Michelmore RW, Gaut BS: **Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana**. *Genome Res* 2002, **12**(9):1305-1315.

Nandety RS, Caplan JL, Cavanaugh K, Perroud B, Wroblewski T, Michelmore RW, Meyers BC: **The Role of TIR-NBS and TIR-X Proteins in Plant Basal Defense Responses**. *Plant Physiology* 2013, **162**(3):1459-1472.

Ohyanagi T, Matsushima N: **Classification of tandem leucine-rich repeats within a great variety of proteins**. *FASEB JOURNAL* 1997, **11**(9):A949-A949.

Seo JK, Kwon SJ, Cho WK, Choi HS, Kim KH: **Type 2C protein phosphatase is a key regulator of antiviral extreme resistance limiting virus spread**. *Sci Rep* 2014, **4**:5905.

Shakiba E, Chen P, Shi A, Li D, Dong D, Brye K: **Two Novel Alleles at the Rsv 3 Locus for Resistance to Soybean Mosaic Virus in PI 399091 and PI 61947 Soybeans**. *Crop Science* 2012, **52**(6):2587-2594.

Shirano Y, Kachroo P, Shah J, Klessig DF: **A Gain-of-Function Mutation in an Arabidopsis Toll Interleukin1 Receptor–Nucleotide Binding Site–Leucine-Rich Repeat Type R Gene Triggers Defense Responses and Results in Enhanced Disease Resistance**. *The Plant Cell Online* 2002, **14**(12):3149-3162.

Suh SJ, Bowman BC, Jeong N, Yang K, Kastl C, Tolin SA, Maroof MAS, Jeong SC: **The Rsv3 Locus Conferring Resistance to Soybean Mosaic Virus is Associated with a Cluster of Coiled-Coil Nucleotide-Binding Leucine-Rich Repeat Genes**. *Plant Genome* 2011, **4**(1):55-64.

Tucker DM, Maroof MAS, Jeong SC, Buss GR, Tolin SA: **Validation and Interaction of the Soybean Mosaic Virus Lethal Necrosis Allele, Rsv1-n, in PI 507389**. *Crop Science* 2009, **49**(4):1277-1283.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG: **Primer3--new capabilities and interfaces**. *Nucleic Acids Res* 2012, **40**(15):e115.

Wan H, Yuan W, Ye Q, Wang R, Ruan M, Li Z, Zhou G, Yao Z, Zhao J, Liu S *et al*: **Analysis of TIR- and non-TIR-NBS-LRR disease resistance gene analogous in pepper:**

characterization, genetic variation, functional divergence and expression patterns. *BMC Genomics* 2012, **13**:502.

Wang DG, Ma Y, Liu N, Yang ZL, Zheng GJ, Zhi HJ: **Fine mapping and identification of the soybean R-SC4 resistance candidate gene to soybean mosaic virus**. *Plant Breeding* 2011, **130**(6):653-659.

Xiao S, Brown S, Patrick E, Brearley C, Turner JG: **Enhanced Transcription of the Arabidopsis Disease Resistance Genes RPW8.1 and RPW8.2 via a Salicylic Acid–Dependent Amplification Circuit Is Required for Hypersensitive Cell Death**. *The Plant Cell Online* 2003, **15**(1):33-45.

Yang H, Huang YP, Zhi HJ, Yu DY: **Proteomics-based analysis of novel genes involved in response toward soybean mosaic virus infection**. *Molecular Biology Reports* 2011, **38**(1):511-521.

Yu Y, Saghai Maroof M, Buss G, Maughan P, Tolin S: **RFLP and microsatellite mapping of a gene for soybean mosaic virus resistance**. *Phytopathology* 1994, **84**(1):60-64.

Yu YH, Yu H, Jeong J, Park H, Song D, Kim C, Kim S, Lee Y, Park J, Park K: **A General Survey of Korean Legume Cultivars (in Korea)**. In. Edited by Science NIoC. Suwon, Korea; 2008.

Table 5.1: Statistics for RNA-Seq data analysis.

Samples	Sequencing read count	Reads count post quality filter	Reads count post adapter trimming	Number of reads aligned†
1 hpi-Rep1	490880336	454809690	390097986	368758920
1 hpi-Rep2	473892132	440169545	320889168	302463056
2 hpi-Rep1	257016864	209134234	159307798	152415533
2 hpi-Rep2	217588610	174531825	134746744	130043860
3 hpi-Rep1	253942718	201354074	182489490	129535359
3 hpi-Rep2	275170502	230876258	133325573	177277195

† Total number of reads aligned calculated from sum of paired-end and single-end reads.

Table 5.2: Non-synonymous mutations in *Rsv3* candidate NB-LRR genes.

Gene ID	Physical Positions†	Williams82		Hwangkeum	
		Base	Codon	Base	Codon
Glyma14g38500	47632717	C	Pro	T	Leu
	47632741	T	Phe	C	Ser
	47632942	T	Phe	G	Cys
	47632978	T	Leu	A	Gln
	47633100	G	Gly	A	Arg
Glyma14g38516	47648236	G	Glu	C	Gln
	47648733	G	Glu	C	Asp
	47648875	A	Arg	G	Gly
	47648989	A	Thr	G	Ala
	47649926	T	Leu	C	Ser
	47650126	G	Val	A	Met
	47650357‡	A	Lys	C	Gln
	47650390‡	A	Lys	G	Glu
Glyma14g38561	47694814	A	Lys	G	Glu
	47694820	GC	Ala	CG	Arg
	47694826	AT	Ile	TA	Tyr
Glyma14g38586	47728863	C	Lys	A	Asn

† Physical positions (start) correspond to the genome assembly version Wm82.a1.

Table 5.3: Summary of SNPs identified in Glyma14g38533 coding sequence.

NBLRR-3	Length (bp)	SNPs	Synonymous	Non-Synonymous
Exon 1	3153	98	9	89
Exon 2	763	52	5	47
Exon 3	8	0	0	0
Total	3924	150	14	136

Table 5.4: INDELs in the coding sequence of the Glyma14g38533 gene from Hwangkeum, L29, and RRR.

Physical Positions†		INDEL genotypes	Size (bp)
		Deletions	
47671691	‡	AGCTATAATTTCCTTAGGCGTGAATTGAACAAGGCATGT	39
47672110	‡	CTG	3
47672119	‡	ATGATGGAAGGC	12
47672378		GATCTGCTG	9
		Insertions	
47672947		ATTCACATCAATTTCCTTAAT	21
47675349		TAAAAG	6

† Physical positions (start) correspond to the genome assembly version Wm82.a1.
‡ INDELs were verified by PCR assay from more than 12 different resistant lines.

Table 5.5: SNP validation from different SMV resistant and susceptible soybean lines.

Cultivars	Type†	Genetic positions and Genotypes‡																
		Glyma14g38500					Glyma14g38516								Glyma14g38561			Glyma14g38586
		47632717	47632741	47632942	47632978	47633100	47648236	47648733	47648875	47648989	47649926	47650126	47650357	47650390	47694814	47694820	47694826	47728863
Williams82	S	C	T	T	T	G	G	G	A	A	T	G	A	A	A	GC	AT	C
Lee68	S	C	T	T	T	G	G	G	A	A	T	-	-	-	A	GC	AT	A
Essex	S	C	T	T	T	G	-	G	A	A	T	-	-	-	A	GC	AT	A
Archer §	S	T	C	G	A	A	G	-	-	-	C	-	-	-	G	GC	AT	A
Hutcheson	S	C	C	T	T	G	G	G	G	G	T	G	-	-	G	CT	TA	A
York	S	C	C	T	T	G	G	G	G	G	T	G	-	-	A	GC	AT	A
L29	R	T	C	G	A	A	G	C	G	G	C	T	-	-	G	GC	AT	A
Suweon97	R	T	C	G	A	A	C	C	G	G	C	A	C	G	G	CG	TA	A
RRR	R	T	C	G	A	A	G	-	-	-	C	T	-	-	G	CC	AT	A
Tousan140	R	T	C	G	A	A	G	C	G	G	C	A	-	-	G	GC	AT	A
Harosoy	R	T	C	G	A	A	-	C	G	G	C	T	-	-	G	CA	AT	A
Columbia	R	T	C	G	A	A	G	C	G	G	C	T	-	-	G	CA	AT	A
Hourei	R	T	C	G	A	A	G	C	G	G	C	A	-	-	G	CG	TA	A
Hardee	R	T	C	G	A	A	G	C	G	G	C	T	-	-	G	CA	AT	A
PI91346	R	T	C	G	A	A	-	-	-	-	C	-	-	-	A	CC	AT	A
PI61947	R	T	C	G	A	A	-	C	G	G	C	A	-	-	G	GG	TT	A
VIR5532	R	T	C	G	A	A	-	-	-	-	C	-	-	-	G	CA	AT	A
Paoting	R	T	C	G	A	A	-	-	-	-	C	-	-	-	G	CG	TA	A
PI323555	R	T	C	G	A	A	-	-	-	-	C	-	-	-	G	CG	TA	A
PI323556	R	T	C	G	A	A	-	-	-	-	C	-	-	-	G	CG	TA	A
PLSO-63	R	T	C	G	A	A	-	-	-	-	C	-	-	-	G	CC	AT	A
PLSO-70	R	T	C	G	A	A	-	C	G	G	C	A	-	-	G	CC	AT	A
OCB81	R	T	C	G	A	A	-	C	G	G	C	A	-	-	G	CA	AT	-
Shirome choutan	R	T	C	G	A	A	-	-	-	-	C	-	-	-	G	CG	TA	A
Graine Jaune Unie	R	T	C	G	A	A	-	-	-	-	C	-	-	-	A	CC	AT	A

† Type: "S" stands for susceptible (*rsv3*), and "R" stands for resistance (*Rsv3*) phenotypic response to SMV.
‡ Genotypic alleles similar to that of Williams 82 are indicated in red color.
§ Archer sequence from McHale et al. (2012) indicated base "T" at position 47632978 in Glyma14g38500 gene.

Table 5.6: Leucine-rich repeat motifs identified in Glyma14g38533 protein sequence from L29 and Williams82.

Repeat	Sequence†		Consensus‡
LRR1	**L29:**	LEILLFHSTEVD	LxxLxLxxCxxL
	Wm82:	LEILLFHS**P**EVD	
LRR2	**L29:**	IKILAILTSSL	LxxLxLxxNxL
	Wm82:	IKILAILTSS**Y**	
LRR3	**L29:**	LHTLCLRG**H**IL	LxxLxLxxNxL
	Wm82:	LHTLCLRG**Y**EL	
LRR4	**L29:**	LEVLDLRNSSFI	LxxLxLxxCxxL
	Wm82:	LEVLDLRGSSFI	
LRR5	**L29:**	LKLLDLFNCVI	LxxLxLxxNxL
	Wm82:	LKLLDLFHCSI	
LRR6	**L29:**	LTFLILEDCPEI	LxxLxLxxCxxL
	Wm82:	LIFLILHDCPEI	
LRR7	**L29:**	LVILRLYELDNL	LxxLxLxxLxxL
	Wm82:	LVILSLYGLDNL	
LRR8	**L29:**	LEELSIESCRQL	LxxLxLxxCxxL
	Wm82:	LEELTIERCRQL	
LRR9	**L29:**	LKFLTIDHCPML	LxxLxLxxCxxL
	Wm82:	LKSLTIRDCPML	
LRR10	**L29:**	LEQVTISDCFEL	LxxLxLxxCxxL
	Wm82:	LEQVRISECYEL	
LRR11	**L29:**	LRTLTILRCHSL	LxxLxLxxCxxL
	Wm82:	LRTLTIRGCRSL	

Repeat	Sequence†		Consensus‡
LRR12	**L29:**	LEELNIGFCDKL	LxxLxLxxCxxL
	Wm82:	LEKLMVERCDKL	
LRR13	**L29:**	LETLRLTELPNL	LxxLxLxxLxxL
	Wm82:	LETLRLTQLPNL	
LRR14	**L29:**	VRRVMIIDSDL	LxxLxLxxNxL
	Wm82:	VRR**G**MIIDSDL	
LRR15	**L29:**	LCSVT**T**TFNQL	LxxLxLxxNxL
	Wm82:	LCSVT**T**TFNQL	
LRR16	**L29:**	LRHLQLYGLGV	LxxLxLxxLxL
	Wm82:	LRELGLSGVGV	
LRR17	**L29:**	LAPLNLDLIY**A**	LxxLxLxxLxL
	Wm82:	LAPLNLDL**T**H**A**	
LRR18	**L29:**	LDVIYVNRCPKL	LxxLxLxxCxxL
	Wm82:	LDVINVNRCPKL	
LRR19	**L29:**	LRTLEITHCEEL	LxxLxLxxCxxL
	Wm82:	LGRLQIIDCEEL	
LRR20	**L29:**	LHYICVEKCNKL	LxxLxLxxCxxL
	Wm82:	LYYISVKKCNKL	
LRR21	**L29:**	LIALEIKDCSQL	LxxLxLxxCxxL
	Wm82:	LSKLEIEDCSEL	
LRR22	**L29:**	LLRIRLSRLPNF	LxxLxLxxLxxL
	Wm82:	LLYITLSSLPNF	

† Amino acid differences between L29 and Williams82 (Wm82) are colored red, while amino acids that fail to match the LRR motif consensus sequence are represented as **bold** (underlined) letters.

‡ LRR motif consensus sequences types: (1) with 11-residues (LxxLxLxxNxL), and (2) with 12-residues (LxxLxLxxCxxL). Other LRR variant motifs such as, LxxLxLxxLxL and LxxLxLxxLxxL were also identified in this gene.

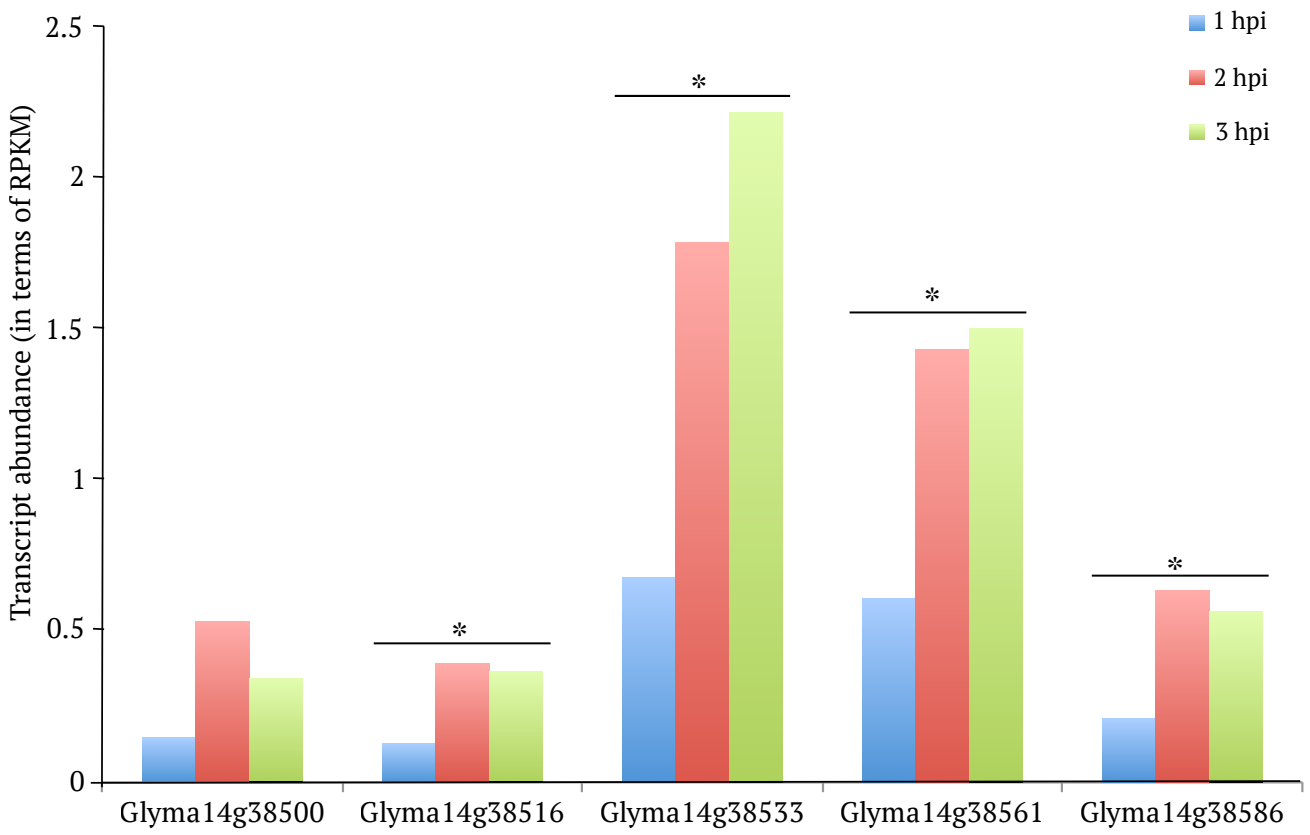


Figure 5.1: Relative transcript abundance of *Rsv3*-candidate NB-LRR genes. RPKM is estimated from normalized gene counts. Error bars are calculated over two technical replicate samples. One-way ANOVA was performed to identify significant change in gene expression levels between 1-hpi, 2-hpi, and 3-hpi. One asterisk (*) denotes *p-value* less than 0.05.

.

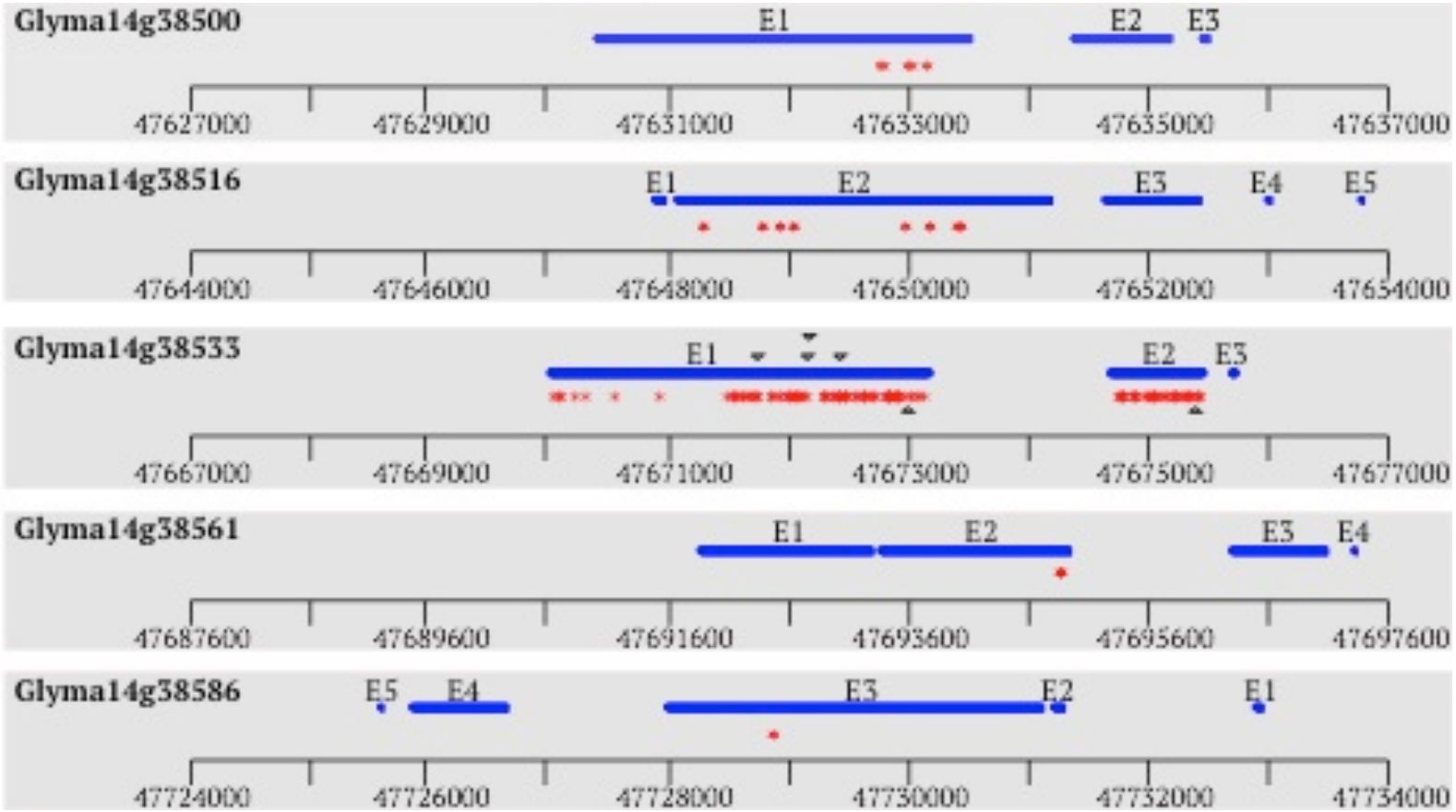


Figure 5.2: Summary of polymorphisms identified in 5 *Rsv3*-candidate genes. Each grey box summarizes *Rsv3*-candidate gene within an inclusive 10kbp region. Blue bars indicate coding regions as per Glyma1.1 gene models. Individual exons are labeled as E1, E2, and so on. Red asterisks indicate SNP polymorphic sites, while the black arrowheads indicate INDEL sites within the coding regions. Down arrowheads represent deletions, while up arrowheads represents insertions within coding region of resistant lines.

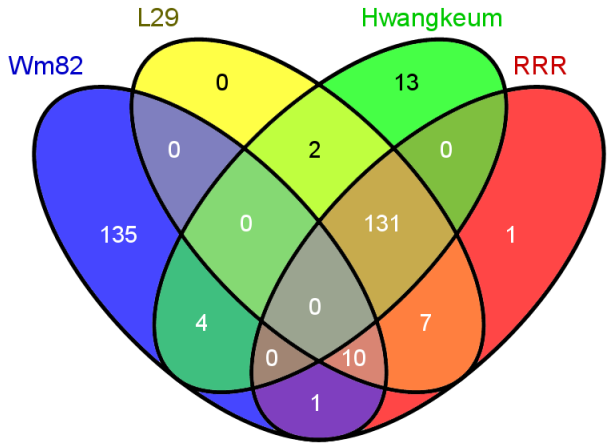


Figure 5.3: Overlap of Glyma14g38533 SNPs between Williams82, L29, Hwangkeum, and RRR soybean lines. Total of 146, 140 and 139 nucleotide positions within Glyma14g38533 gene were polymorphic in Hwangkeum, L29, and RRR when compared against Williams82, respectively. 'Wm82' stands for Williams82.

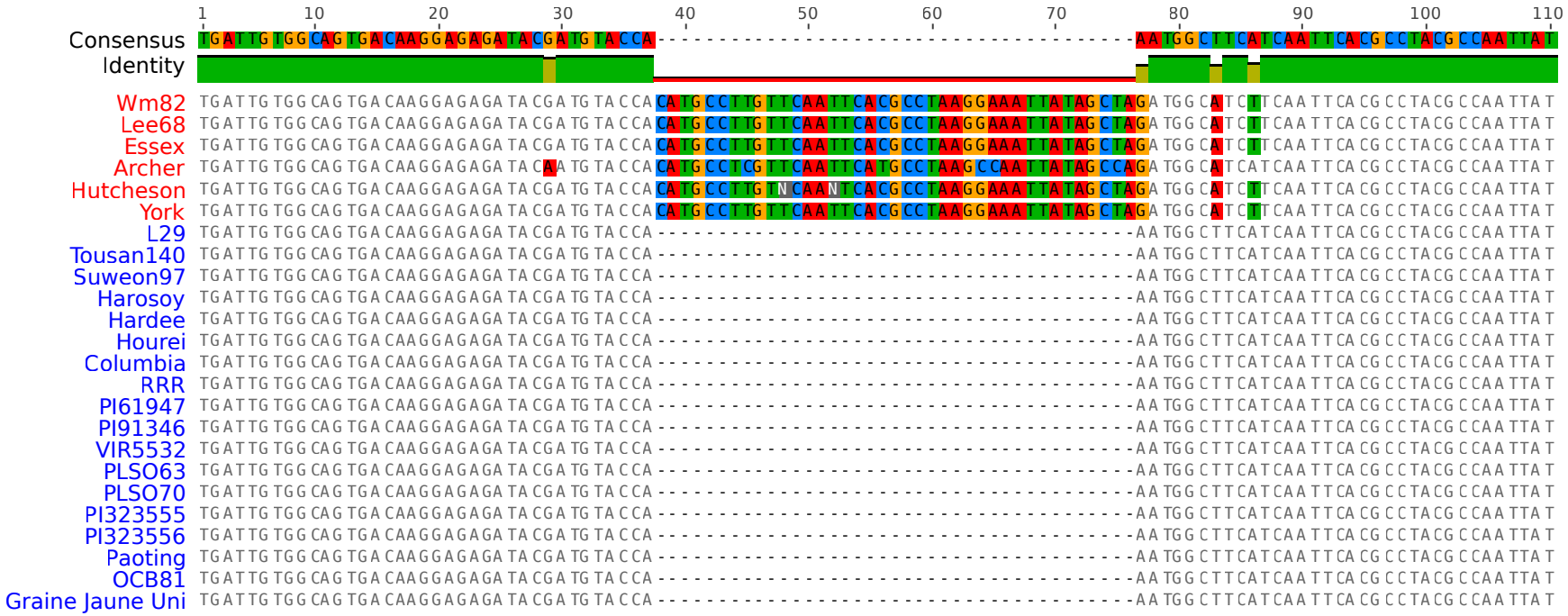


Figure 5.4: Validation of 39-bp deletion region within Glyma14g38533 gene from multiple resistant soybean lines. Only resistant lines, indicated in blue color, showed this deletion. Susceptible lines, indicated in red color, do not have this deletion.

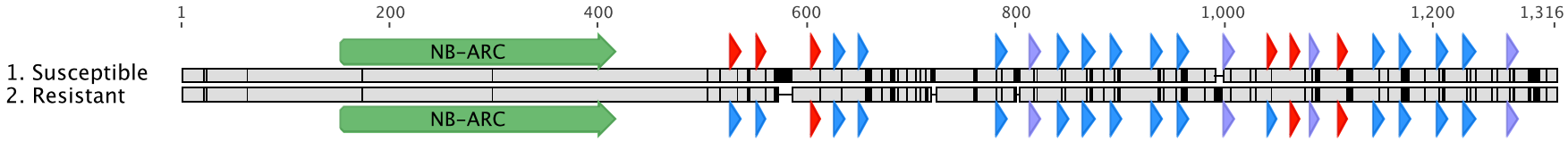


Figure 5.5: Differences in Glyma14g38533 protein domains from susceptible Williams 82 and resistant L29 line. Blue arrows indicate LRR repeats, which are exact match to 11- and 12-residue consensus sequence; red arrows indicate LRR motifs with a mismatch with the consensus sequence; purple arrows indicate LRR variant motifs.

CHAPTER 6

Conclusion

SUMMARY OF RESEARCH

This dissertation is focused on the understanding of transcriptional regulation, genome content, and genome composition for two economically important agronomic traits: low seed phytic acid (PA), and resistance to *Soybean mosaic virus* (SMV), using the latest "omics" tools such as "transcriptomics" and "genomics."

Seed development is an important growth stage for legumes, such as soybean, with respect to economic standpoint. Seed PA is considered an anti-nutritional component for soybean, as it causes mineral deficiency in monogastric animals and phosphorus pollution due to the accumulation of undigested waste. Breeding for low PA (*lpa*) soybeans is an eco-friendly solution to avoid problems associated with normal PA soybeans, however, low emergence phenotype of these *lpa* soybeans is preventing their commercial use. It is possible that the low emergence phenotype is due to reprogramming of the metabolism in developing seeds by *lpa*-causing mutations. The research in this dissertation is mainly focused on addressing the impact of *lpa*-causing mutations on seed development via transcriptome profiling. Transcriptomic regulation can work in multiple modes, such as differential gene expression, gene co-expression,

transcription factor binding, methylation, etc. The focus of this dissertation is on regulation via differential gene expression (Chapter 2) and co-expression (Chapter 3).

Chapter 2 describes the differential expression for five seed developmental stages compared between a triple mutant *lpa* line, "*3mlpa*"-(*mips1/mrp-l/mrp-n*) and a normal PA line "3MWT"-(MIPS1/MRP-L/MRP-N). This is the first global transcriptome profiling study that has described the combined effect of three *lpa*-causing mutations on the regulation of seed development. Over 4000 differentially expressed genes were identified, which, upon functional enrichment, showed significant biological processes associated with cell wall polysaccharide metabolism, regulation of cell death, photosynthesis, and glycolysis. It also revealed the enrichment of transcription factor families such as WRKY, GRAS, ZIM, CAMTA, GRF, SNF2, ZF-HD, MBF1B, and TCP. The differentially expressed genes, biological processes, and transcription factor families identified in this study suggest regulation of metabolism in the *lpa* mutant. Interestingly, the raffinose family oligosaccharide biosynthesis pathway genes were also regulated over the course of seed development in the *lpa* mutant.

Chapter 3 describes the co-expression metabolic network for five seed developmental stages compared among two *lpa* mutants, "*3mlpa*"-(*mips1/mrp-l/mrp-n*) and "*1mlpa*"-(*mips1*), and two normal PA lines, "3MWT"-(MIPS1/MRP-L/MRP-N) and "1MWT"-(MIPS1). This study identified co-expressed genes whose expression profiles change with respect to the seed developmental stages. These co-expressed genes represented lipid biosynthesis, amino acid (especially glutamate) biosynthesis, sugar signaling, and secondary metabolite biosynthesis pathways. Moreover, it identified two peroxisomal beta-oxidation pathway genes encoding for the acyl-activating enzyme 16 and dienoyl-CoA isomerase 1, to be co-expressed in mutants' co-expression network, but not in wildtypes. The differences in co-expression networks of mutants

and wildtypes support the hypothesis that the *lpa*-causing mutations regulate the metabolic pathways in seed development.

Findings from Chapters 2 and 3 provide potential targets of regulation associated with *lpa*-causing mutations. Moreover, the global perspective of the transcriptome is essential to understand the genetic basis of the low emergence phenotype in the *lpa* soybean, and also to engineer new agronomic traits in soybean. It is a well-known fact that genomic DNA contamination can interfere with the transcriptome sequencing, leading to false interpretations. Similarly, the gene dosage (or number of copies of a gene) can influence the sequencing-based transcriptome profiling results. Therefore, to be confident of our transcriptome profiling results, we asked the question: Does gene dosage have any influence on the transcriptome profiling in our experimental lines, "*1mlpa*" and "1MWT." This resulted in defining our next research objective of estimating copy number variations (CNVs) in isogenic lines, "*1mlpa*" and "1MWT."

Chapter 4 describes the CNVs identified by comparing genomic DNA sequences of "*1mlpa*" and "1MWT." This study identified more than 6000 loci that could potentially have CNVs, 76 of which were coding regions. We also identified the differentially expressed genes between "*1mlpa*" and "1MWT" at five seed developmental stages. None of the CNV-associated genes were differentially expressed, suggesting no impact of CNV on transcriptional profiling of these experimental lines. This data is not sufficient at the moment for publication; however, it provides more confidence in our transcriptome sequencing results.

The second important soybean trait studied in this dissertation is Soybean mosaic virus disease resistance. The genetically resistant soybean cultivars provide an environment-friendly solution to prevent this disease. Of the three virus disease resistance loci (*Rsv1, Rsv3,* and *Rsv4*), the *Rsv3* locus, on soybean chromosome 14, containing five nucleotide-binding leucine-rich

repeat (NB-LRR) encoding genes, conditions resistance to most virulent SMV strain groups (G5-G7). These NB-LRR genes are potential candidates for the *Rsv3* gene. In this research, gene sequence comparisons were performed between SMV-resistant and SMV-susceptible soybean cultivars, to narrow down the *Rsv3* candidates to one most promising gene, i.e., Glyma14g38533. This study involved the use of transcriptome and genome sequencing data to assemble the Glyma14g38533 transcript sequence from the SMV-resistant soybean. This dissertation provides an extensive account of the sequence features identified for this promising candidate resistance gene. This study involved the use of transcriptome and genome sequencing data to assemble the Glyma14g38533 transcript sequence from the SMV-resistant soybean L29. Although functional validation of the Glyma14g38533 gene still awaits, the evidence provided in this study clearly indicated the potential of this gene to deliver *Rsv3*-type resistance.

FUTURE DIRECTIONS

Follow up experiments suggested for the phytate project:

Global perspective of proteome and metabolome of seed development in *lpa* soybean lines:

The transcriptome profiling study of this dissertation has revealed the set of genes that could potentially be involved in the regulation of seed development in *lpa* soybean. It is possible that not all genes might result in the change in phenotype. Therefore, it is important to understand how proteome and metabolome of these seeds has been regulated in response to *lpa*-causing mutations. For proteome and metabolome profiling of seed developmental stages 2-5 and 3-5 might provide more insights on: (1) how transcriptome-level regulation has affected the proteome and (2) how both transcriptome- and proteome-level regulations have affected metabolome of seed development in *lpa* soybean lines. The common pathways represented the

transcriptome; the proteome and metabolome in *lpa* mutants are more likely to be involved in regulation of seed development. These regulatory pathways can be the potential subjects for future physiological studies.

Follow up experiments suggested for the CNV project:

Estimating CNVs between near isogenic lines, *1mlpa* and 1MWT using different analysis software, and validating the biologically meaningful CNV events:

Different software packages use different statistical calculations to call CNVs. One can expect that the CNV calls shared by two or more packages are more likely to be real and biologically meaningful. Therefore, another read-depth based software, such as "rdxplorer" should be used for CNV calling. The results must be compared with the "cnv-seq" software used in this study to find common CNV calls. These shared or common CNV calls should be validated.

Follow up experiments suggested for the *Rsv3* project:

Functional validation of Glyma14g38533 gene, via transient gene expression assays:

With the availability of sequence polymorphisms in Glyma14g38533 gene alleles from SMV-resistant and -susceptible soybean lines, one can easily synthesize Glyma14g38533 gene allelic constructs for functional validation experiments. The transient gene expression assays would require both susceptible and resistant forms of the Glyma14g38533 gene. Induction of a hypersensitive response by resistant form of the Glyma14g38533 gene allele in the presence of SMV is an indication that the Glyma14g38533 gene is encoding for *Rsv3*-type resistance.

Functional validation of Glyma14g38533 gene, via gene complementation:

The Glyma14g38533 complementation study would need only the resistant form of the gene. This could be done by two ways: (1) overexpression of the resistant form of the gene in a susceptible soybean lines, and observe the resistance phenotype upon virus inoculation; and (2) silencing of the gene in the resistant soybean line and again observing the phenotype upon virus inoculation. In these cases, gene complementation can be achieved via virus-induced stable expression or virus induced gene-silencing methods, respectively.

REFERENCES

Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD: **Count-based differential expression analysis of RNA sequencing data using R and Bioconductor**. *Nat Protoc* 2013, **8**(9):1765-1786.

Anderson JE, Kantar MB, Kono TY, Fu F, Stec AO, Song Q, Cregan PB, Specht JE, Diers BW, Cannon SB *et al*: **A roadmap for functional structural variants in the soybean genome**. *G3 (Bethesda)* 2014, **4**(7):1307-1318.

Asakura T, Tamura T, Terauchi K, Narikawa T, Yagasaki K, Ishimaru Y, Abe K: **Global gene expression profiles in developing soybean seeds**. *Plant Physiol Biochem* 2012, **52**:147-153.

Black RE: **Therapeutic and preventive effects of zinc on serious childhood infectious diseases in developing countries**. *The American Journal of Clinical Nutrition* 1998, **68**(2):476S-479S.

Bregitzer P, Raboy V: **Effects of Four Independent Low-Phytate Mutations on Barley Agronomic Performance**. *Crop Science* 2006, **46**(3):1318.

Brinch-Pedersen H, Sørensen LD, Holm PB: **Engineering crop plants: getting a handle on phosphate**. *Trends in Plant Science* 2002, **7**(3):118-125.

Chen P, Buss GR, Roane CW, Tolin SA: **Inheritance in Soybean of Resistant and Necrotic Reactions to Soybean Mosaic-Virus Strains**. *Crop Science* 1994, **34**(2):414-422.

Cho EK, Goodman RM: **Strains of Soybean Mosaic-Virus - Classification Based on Virulence in Resistant Soybean Cultivars**. *Phytopathology* 1979, **69**(5):467-470.

Coate JE, Bar H, Doyle JJ: **Extensive translational regulation of gene expression in an allopolyploid (Glycine dolichocarpa)**. *The Plant cell* 2014, **26**(1):136-150.

Collakova E, Aghamirzaie D, Fang Y, Klumas C, Tabataba F, Kakumanu A, Myers E, Heath LS, Grene R: **Metabolic and Transcriptional Reprogramming in Developing Soybean (Glycine max) Embryos**. *Metabolites* 2013, **3**(2):347-372.

Cosgrove DJ, Irving GCJUhbgcbiCkA: **Inositol phosphates: their chemistry, biochemistry, and physiology**: Elsevier Scientific Pub. Co.; 1980.

Cromwell GL, Coffey RD, Parker GR, Monegue HJ, Randolph JH: **Efficacy of a recombinant-derived phytase in improving the bioavailability of phosphorus in corn-soybean meal diets for pigs**. *Journal of animal science* 1995, **73**(7):2000-2008.

Cui X, Chen X, Wang A: **Detection, understanding and control of Soybean mosaic virus**. *Soybean-molecular aspects of breeding* 2011:335-354.

Davidsson L, Almgren A, Juillerat MA, Hurrell RF: **Manganese absorption in humans: the effect of phytic acid and ascorbic acid in soy formula**. *The American Journal of Clinical Nutrition* 1995, **62**(5):984-987.

Donahue JL, Alford SR, Torabinejad J, Kerwin RE, Nourbakhsh A, Ray WK, Hernick M, Huang X, Lyons BM, Hein PP *et al*: **The Arabidopsis thaliana Myo-inositol 1-phosphate synthase1 gene is required for Myo-inositol synthesis and suppression of cell death**. *The Plant cell* 2010, **22**(3):888-903.

Dong J, Yan W, Bock C, Nokhrina K, Keller W, Georges F: **Perturbing the metabolic dynamics of myo-inositol in developing Brassica napus seeds through in vivo methylation impacts its utilization as phytate precursor and affects downstream metabolic pathways**. *BMC Plant Biol* 2013, **13**:84.

Du J, Tian Z, Sui Y, Zhao M, Song Q, Cannon SB, Cregan P, Ma J: **Pericentromeric effects shape the patterns of divergence, retention, and expression of duplicated genes in the paleopolyploid soybean**. *The Plant cell* 2012, **24**(1):21-32.

Fan XD, Wang JQ, Yang N, Dong YY, Liu L, Wang FW, Wang N, Chen H, Liu WC, Sun YP *et al*: **Gene expression profiling of soybean leaves and roots under salt, saline-alkali and drought stress by high-throughput Illumina sequencing**. *Gene* 2013, **512**(2):392-402.

Gillman JD, Pantalone VR, Bilyeu K: **The Low Phytic Acid Phenotype in Soybean Line CX1834 Is Due to Mutations in Two Homologs of the Maize Low Phytic Acid Gene**. *The Plant Genome Journal* 2009, **2**(2):179.

Gunduz I, Buss GR, Chen P, Tolin SA: **Characterization of SMV Resistance Genes in Tousan 140 and Hourei Soybean**. *Crop Science* 2002, **42**(1):90-95.

Gunduz I, Buss GR, Chen PY, Tolin SA: **Genetic and phenotypic analysis of Soybean mosaic virus resistance in PI 88788 soybean**. *Phytopathology* 2004, **94**(7):687-692.

Guttieri MB, David; Dorsch, John A; Raboy, Victor; Souza, Edward: **Identification and Characterization of a Low Phytic Acid Wheat**. *Crop Science* 2004, **44**(2):418-424.

Guttieri MJ, Peterson KM, Souza EJ: **Agronomic Performance of Low Phytic Acid Wheat**. *Crop Science* 2006, **46**(6):2623.

Hallberg L, Brune M, Rossander L: **Iron absorption in man: ascorbic acid and dose-dependent inhibition by phytate**. *The American Journal of Clinical Nutrition* 1989, **49**(1):140-144.

Halterman JS, Kaczorowski JM, Aligne CA, Auinger P, Szilagyi PG: **Iron Deficiency and Cognitive Achievement Among School-Aged Children and Adolescents in the United States**. *Pediatrics* 2001, **107**(6):1381-1386.

Hanakahi L: **Binding of Inositol Phosphate to DNA-PK and Stimulation of Double-Strand Break Repair**. *Cell* 2000, **102**(6):721-729.

Hartman GL, Sinclair JB, Rupe JC: **Compendium of soybean diseases**. St. Paul: American Phytopathological Society (APS Press); 1999.

Hayashi S, Reid DE, Lorenc MT, Stiller J, Edwards D, Gresshoff PM, Ferguson BJ: **Transient Nod factor-dependent gene expression in the nodulation-competent zone of soybean (Glycine max [L.] Merr.) roots**. *Plant Biotechnology Journal* 2012, **10**(8):995-1010.

Hayes AJ, Ma GR, Buss GR, Maroof MAS: **Molecular marker mapping of RSV4, a gene conferring resistance to all known strains of soybean mosaic virus**. *Crop Science* 2000, **40**(5):1434-1437.

Hill BE, Sutton AL, Richert BT: **Effects of low-phytic acid corn, low-phytic acid soybean meal, and phytase on nutrient digestibility and excretion in growing pigs**. *Journal of animal science* 2009, **87**(4):1518-1527.

Hill JH, Alleman R, Hogg DB, Grau CR: **First Report of Transmission of Soybean mosaic virus and Alfalfa mosaic virus by Aphis glycines in the New World**. *Plant Disease* 2001, **85**(5):561-561.

Hitz WD, Carlson TJ, Kerr PS, Sebastian SA: **Biochemical and molecular characterization of a mutation that confers a decreased raffinosaccharide and phytic acid phenotype on soybean seeds**. *Plant physiology* 2002, **128**(2):650-660.

Jeong SC, Kristipati S, Hayes AJ, Maughan PJ, Noffsinger SL, Gunduz I, Buss GR, Maroof MAS: **Genetic and sequence analysis of markers tightly linked to the soybean mosaic virus resistance gene, Rsv3**. *Crop Science* 2002, **42**(1):265-270.

Jeong SC, Maroof MAS: **Detection and genotyping of SNPs tightly linked to two disease resistance loci, Rsv1 and Rsv3, of soybean**. *Plant Breeding* 2004, **123**(4):305-310.

Jones SI, Gonzalez DO, Vodkin LO: **Flux of transcript patterns during soybean seed development**. *BMC Genomics* 2010, **11**:136.

Jones SI, Vodkin LO: **Using RNA-Seq to profile soybean seed development from fertilization to maturity**. *PLoS One* 2013, **8**(3):e59270.

Karner U, Peterbauer T, Raboy V, Jones DA, Hedley CL, Richter A: **myo-Inositol and sucrose concentrations affect the accumulation of raffinose family oligosaccharides in seeds**. *Journal of experimental botany* 2004, **55**(405):1981-1987.

Kiihl RAS, Hartwig EE: **Inheritance of Reaction to Soybean Mosaic-Virus in Soybeans**. *Crop Science* 1979, **19**(3):372-375.

Larson SR, Rutger JN, Young KA, Raboy V: **Isolation and Genetic Mapping of a Non-Lethal Rice (L.) Mutation**. *Crop Science* 2000, **40**(5):1397.

Larson SR, Young KA, Cook A, Blake TK, Raboy V: **Linkage mapping of two mutations that reduce phytic acid content of barley grain**. *Theoretical and Applied Genetics* 1998, **97**(1-2):141-146.

Le BH, Wagmaister JA, Kawashima T, Bui AQ, Harada JJ, Goldberg RB: **Using genomics to study legume seed development**. *Plant physiology* 2007, **144**(2):562-574.

Le DT, Nishiyama R, Watanabe Y, Tanaka M, Seki M, Ham le H, Yamaguchi-Shinozaki K, Shinozaki K, Tran LS: **Differential gene expression in soybean leaf tissues at late developmental stages under drought stress revealed by genome-wide transcriptome analysis**. *PLoS One* 2012, **7**(11):e49522.

Lemtiri-Chlieh F, MacRobbie EA, Brearley CA: **Inositol hexakisphosphate is a physiological signal regulating the K+-inward rectifying conductance in guard cells**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(15):8687-8692.

Lemtiri-Chlieh F, MacRobbie EA, Webb AA, Manison NF, Brownlee C, Skepper JN, Chen J, Prestwich GD, Brearley CA: **Inositol hexakisphosphate mobilizes an endomembrane store of calcium in guard cells**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(17):10091-10095.

Li K, Yang QH, Zhi HJ, Gai JY: **Identification and Distribution of Soybean mosaic virus Strains in Southern China**. *Plant Disease* 2010, **94**(3):351-357.

Li Y-g, Yang M-x, Li Y, Liu W-w, Wen J-z, Li Y-h: **Differential Gene and Protein Expression in Soybean at Early Stages of Incompatible Interaction with Phytophthora sojae**. *Agricultural Sciences in China* 2011, **10**(6):902-910.

Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G, Stacey G: **An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants**. *The Plant Journal* 2010, **63**(1):86-99.

Loewus FA, Kelly S, Neufeld EF: **Metabolism of Myo-inositol in Plants: Conversion to Pectin, Hemicellulose, D-xylose, and Sugar Acids**. *Proceedings of the National Academy of Sciences of the United States of America* 1962, **48**(3):421-425.

Loewus FA, Murthy PPN: **myo-Inositol metabolism in plants**. *Plant Science* 2000, **150**(1):1-19.

Lott JNA, Ockenden I, Raboy V, Batten GD: **Phytic acid and phosphorus in crop seeds and fruits: a global estimate**. *Seed Science Research* 2000, **10**(1):11-33.

Ma G, Chen P, Buss GR, Tolin SA: **Genetic-Characteristics of 2 Genes for Resistance to Soybean Mosaic-Virus in P1486355 Soybean**. *Theoretical and Applied Genetics* 1995, **91**(6-7):907-914.

Ma G, Chen P, Buss GR, Tolin SA: **Complementary action of two independent dominant genes in Columbia soybean for resistance to soybean mosaic virus**. *J Hered* 2002, **93**(3):179-184.

Maroof AS, Buss GR: **Low phytic acid, low stachyose, high sucrose soybean lines**. In.: Google Patents; 2008.

Maroof MAS, Glover NM, Biyashev RM, Buss GR, Grabau EA: **Genetic Basis of the Low-Phytate Trait in the Soybean Line CX1834**. *Crop Science* 2009, **49**(1):69.

Maroof MAS, Tucker D, Tolin S: **Genomics of Viral–Soybean Interactions**. In: *Genetics and Genomics of Soybean.* Edited by Stacey G, vol. 2: Springer New York; 2008: 293-319.

Maroof MAS, Tucker DM, Skoneczka JA, Bowman BC, Tripathy S, Tolin SA: **Fine Mapping and Candidate Gene Discovery of the Soybean Mosaic Virus Resistance Gene, Rsv4**. *Plant Genome-Us* 2010, **3**(1):14-22.

Martins PK, Jordão BQ, Yamanaka N, Farias JRB, Beneventi MA, Binneck E, Fuganti R, Stolf R, Nepomuceno AL: **Differential gene expression and mitotic cell analysis of the drought tolerant soybean (Glycine max L. Merrill Fabales, Fabaceae) cultivar MG/BR46 (Conquista) under two water deficit induction systems**. *Genetics and Molecular Biology* 2008, **31**:512-521.

Murphy AM, Otto B, Brearley CA, Carr JP, Hanke DE: **A role for inositol hexakisphosphate in the maintenance of basal resistance to plant pathogens**. *The Plant journal : for cell and molecular biology* 2008, **56**(4):638-652.

Nagy R, Grob H, Weder B, Green P, Klein M, Frelet-Barrand A, Schjoerring JK, Brearley C, Martinoia E: **The Arabidopsis ATP-binding cassette protein AtMRP5/AtABCC5 is a high affinity inositol hexakisphosphate transporter involved in guard cell signaling and phytate storage**. *The Journal of biological chemistry* 2009, **284**(48):33614-33622.

Navert B, Sandstrom B, Cederblad A: **Reduction of the phytate content of bran by leavening in bread and its effect on zinc absorption in man**. *The British journal of nutrition* 1985, **53**(1):47-53.

Oltmans SE, Fehr WR, Welke GA, Cianzio SR: **Inheritance of Low-Phytate Phosphorus in Soybean**. *Crop Science* 2004, **44**(2):433.

Oltmans SE, Fehr WR, Welke GA, Raboy V, Peterson KL: **Agronomic and seed traits of soybean lines with low-phytate phosphorus**. *Crop Science* 2005, **45**(2):593-598.

Raboy V: **Accumulation and storage of phosphate and minerals.** . In: *Cellular and Molecular Biology of Plant Seed Development* Dordrecht, Netherlands: Kluwer Academic Publishers; 1997: 441-477.

Raboy V: **Seeds for a better future: 'low phytate' grains help to overcome malnutrition and reduce pollution**. *Trends in plant science* 2001, **6**(10):458-462.

Raboy V: **myo-Inositol-1,2,3,4,5,6-hexakisphosphate**. *Phytochemistry* 2003, **64**(6):1033-1043.

Raboy V: **Seed phosphorus and the development of low-phytate crops**. In. Wallingford: CABI; 2007: 111-132.

Raboy V: **Approaches and challenges to engineering seed phytate and total phosphorus**. *Plant Science* 2009, **177**(4):281-296.

Raboy V, Gerbasi PF, Young KA, Stoneberg SD, Pickett SG, Bauman AT, Murthy PP, Sheridan WF, Ertl DS: **Origin and seed phenotype of maize low phytic acid 1-1 and low phytic acid 2-1**. *Plant physiology* 2000a, **124**(1):355-368.

Raboy V, Gerbasi PF, Young KA, Stoneberg SD, Pickett SG, Bauman AT, Murthy PPN, Sheridan WF, Ertl DS: **Origin and Seed Phenotype of Maize low phytic acid 1-1 and low phytic acid 2-1**. *Plant physiology* 2000b, **124**(1):355-368.

Raboy V, Young KA, Dorsch JA, Cook A: **Genetics and breeding of seed phosphorus and phytic acid**. *Journal of plant physiology* 2001, **158**(4):489-497.

Rasmussen SK, Hatzack F: **Identification of two Low-Phytate Barley (Hordeum Vulgare l.) Grain Mutants by TLC and Genetic Analysis**. *Hereditas* 2004, **129**(2):107-112.

Sasakawa N, Sharif M, Hanley MR: **Metabolism and biological activities of inositol pentakisphosphate and inositol hexakisphosphate**. *Biochemical pharmacology* 1995, **50**(2):137-146.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J *et al*: **Genome sequence of the palaeopolyploid soybean**. *Nature* 2010, **463**(7278):178-183.

Seo J, Sohn S, Kim K: **A single amino acid change in HC-Pro of soybean mosaic virus alters symptom expression in a soybean cultivar carrying Rsv1 and Rsv3**. *Archives of Virology* 2011, **156**(1):135-141.

Severin AJ, Woody JL, Bolon YT, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE *et al*: **RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome**. *BMC Plant Biol* 2010, **10**:160.

Shakiba E, Chen P, Shi A, Li D, Dong D, Brye K: **Inheritance and Allelic Relationships of Resistance Genes for in 'Corsica' and 'Beeson' Soybean**. *Crop Science* 2013, **53**(4):1455.

Sharpley AN, Chapra SC, Wedepohl R, Sims JT, Daniel TC, Reddy KR: **Managing Agricultural Phosphorus for Protection of Surface Waters: Issues and Options**. *J Environ Qual* 1994, **23**(3):437-451.

Shen X, Xiao H, Ranallo R, Wu W-H, Wu C: **Modulation of ATP-Dependent Chromatin-Remodeling Complexes by Inositol Polyphosphates**. *Science* 2003, **299**(5603):112-114.

Shi J, Wang H, Hazebroek J, Ertl DS, Harp T: **The maize low-phytic acid 3 encodes a myo-inositol kinase that plays a role in phytic acid biosynthesis in developing seeds**. *The Plant journal : for cell and molecular biology* 2005, **42**(5):708-719.

Shi J, Wang H, Wu Y, Hazebroek J, Meeley RB, Ertl DS: **The maize low-phytic acid mutant lpa2 is caused by mutation in an inositol phosphate kinase gene**. *Plant physiology* 2003, **131**(2):507-515.

Shi JR, Wang HY, Schellin K, Li BL, Faller M, Stoop JM, Meeley RB, Ertl DS, Ranch JP, Glassman K: **Embryo-specific silencing of a transporter reduces phytic acid content of maize and soybean seeds**. *Nature biotechnology* 2007, **25**(8):930-937.

Sparvoli F, Cominelli E: **Phytate Transport by MRPs**. 2014, **22**:19-38.

Stevenson-Paulik J, Bastidas RJ, Chiou S-T, Frye RA, York JD: **Generation of phytate-free seeds in Arabidopsis through disruption of inositol polyphosphate kinases**. *PNAS* 2005, **102**(35):12612-12617.

Strother S: **Homeostasis in germinating seeds**. *Annals of Botany* 1980, **45**(2):217-218.

Suh SJ, Bowman BC, Jeong N, Yang K, Kastl C, Tolin SA, Maroof MAS, Jeong SC: **The Rsv3 Locus Conferring Resistance to Soybean Mosaic Virus is Associated with a Cluster of**

Coiled-Coil Nucleotide-Binding Leucine-Rich Repeat Genes. *Plant Genome-Us* 2011, **4**(1):55-64.

Taji T, Takahashi S, Shinozaki K: **Inositols and Their Metabolites in Abiotic and Biotic Stress Responses**. 2006, **39**:239-264.

Tan X, Calderon-Villalobos LI, Sharon M, Zheng C, Robinson CV, Estelle M, Zheng N: **Mechanism of auxin perception by the TIR1 ubiquitin ligase**. *Nature* 2007, **446**(7136):640-645.

Tolin SA, Lacy GH: **Viral, Bacterial, and Phytoplasmal Diseases of Soybean**. In: *Soybeans: Improvement, Production, and Uses*. Edited by Boerma HR, Specht JE. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America; 2004: 765-819.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nat Protoc* 2012, **7**(3):562-578.

Walker DR, Scaboo AM, Pantalone VR, Wilcox JR, Boerma HR: **Genetic mapping of loci associated with seed phytic acid content in CX1834-1-2 soybean**. *Crop Science* 2006, **46**(1):390-397.

Wang X, Gai J, Pu Z: **Classification and distribution of strains of *Soybean mosaic virus* in middle and lower Huanghuai and Changjiang river valleys**. *Soybean Science* 2003, **22**:102-107.

Weaver CM, Heaney RP, Martin BR, Fitzsimmons ML: **Human Calcium Absorption from Whole-Wheat Products**. *The Journal of nutrition* 1991, **121**(11):1769-1775.

Wilcox JR, Premachandra GS, Young KA, Raboy V: **Isolation of High Seed Inorganic P, Low-Phytate Soybean Mutants**. *Crop Science* 2000, **40**(6):1601.

Wilhelm BT, Marguerat S, Goodhead I, Bahler J: **Defining transcribed regions using RNA-seq**. *Nat Protoc* 2010, **5**(2):255-266.

York JD, Odom AR, Murphy R, Ives EB, Wente SR: **A Phospholipase C-Dependent Inositol Polyphosphate Kinase Pathway Required for Efficient Messenger RNA Export**. *Science* 1999, **285**(5424):96-100.

Yuan FJ, Zhao HJ, Ren XL, Zhu SL, Fu XJ, Shu QY: **Generation and characterization of two novel low phytate mutations in soybean (Glycine max L. Merr.)**. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik* 2007, **115**(7):945-957.

Zhang C, Eggenberger AL, Hajimorad MR, Tsang S, Hill JH: **The N terminal Soybean mosaic virus (SMV) CI is required for SMV virulence and is a symptom determinant on Rsv3 genotype soybean**. *Phytopathology* 2007, **97**(7):S129-S129.

Zhang CQ, Hajimorad MR, Eggenberger AL, Tsang S, Whitham SA, Hill JH: **Cytoplasmic inclusion cistron of Soybean mosaic virus serves as a virulence determinant on Rsv3-genotype soybean and a symptom determinant**. *Virology* 2009, **391**(2):240-248.

APPENDICES

Appendix A. Supplementary Tables from Chapter 2

Supplementary Table A1: Primers used for quantitative real-time PCR.

Gene ID	Primer name	Sequences (5'-3')	Tm	PCR size	Details
Glyma12g36570	CESA-F	GCTTATGTGAACACCACCATCTATC	60	90	Cellulose synthase (CESA) family protein
	CESA-R	CTTATTAGTCAGCAGACAGACAGCA			
Glyma13g00280	XET6-F	CGTAGCTAGGCACCACTGCTA	60	103	Xyloglucan endotrans-glycosylase 6 (XET6)
	XET6-R	AGTCTTGGTTGAAGTTACCAGCA			
Glyma15g22820	InsT2-F	TTCTATGGAGGAAGTGGAGAAGATG	60	90	Inositol transporter 2 (InsT2)
	InsT2-R	CAGTGCTTCTCAGAGCCAGAAT			
Glyma08g06420	SugT1-F	AGATCCGTTCGGCTGCTC	60	169	Sugar transporter 1 (SugT1)
	SugT1-R	TCAGGCAAGAAGAAGTAGATGAAGA			
Glyma13g15560	PS2D1-F	CAATCCGTAGTTGATAGTCAAGGTC	60	100	P680 reaction center D1 (PS2D1)
	PS2D1-R	GAGCATTACGTTCATGCATTACTTC			
Glyma15g38110	Cyb559-F	TTACAGTATGATGGTTGGCTGTTC	58	91	Cytochrome b559 (Cyb559)
	Cyb559-R	TATCGTTGGATGAACTGCATTG			
Glyma19g41550	GalS1-F	TTCACATCGGCACTATTGGA	58	69	Galactinol synthase
	GalS1-R	AAGCATATTAGGCAGCAGAAGG			
Glyma06g18890	RS2-F	CTAGGGCCATCTCTGGTGGA	60	129	Raffinose synthase
	RS2-R	CGTGTGGGGAGTGCATAGTG			
Glyma19g40550	StaSyn1-F	GCTCCCTGATGGAAAACTC	60	65	Stachyose synthase
	StaSyn1-R	CGGTAGAACCACAGGACAAAA			
Glyma07g32020	UBQ10-F	TCCCACCAGACCAGCAGAG	60	117	Ubiquitin
	UBQ10-R	CACGAAGACGCAACACAAGG			

Supplementary Table A2: Sequencing data from developing seed tissue of soybean.

Sample ID	Sample Description	Total Reads	Unmapped	Mapped	Alignment (%)
A1-1	*lpa*-Stage1-Rep1	30508048	4381792	26126256	85.64
A1-2	*lpa*-Stage1-Rep2	30495654	4611700	25883954	84.88
A1-3	*lpa*-Stage1-Rep3	31561685	4766818	26794867	84.90
A2-1	*lpa*-Stage2-Rep1	37229511	5590360	31639151	84.98
A2-2	*lpa*-Stage2-Rep2	35397167	5264937	30132230	85.13
A2-3	*lpa*-Stage2-Rep3	37610941	5185337	32425604	86.21
A3-1	*lpa*-Stage3-Rep1	30305469	3875114	26430355	87.21
A3-2	*lpa*-Stage3-Rep2	35608709	5001128	30607581	85.96
A3-3	*lpa*-Stage3-Rep3	29003679	4363410	24640269	84.96
A4-1	*lpa*-Stage4-Rep1	37309806	4991336	32318470	86.62
A4-2	*lpa*-Stage4-Rep2	29610791	3947507	25663284	86.67
A4-3	*lpa*-Stage4-Rep3	28069153	3772244	24296909	86.56
A5-1	*lpa*-Stage5-Rep1	27492422	3856713	23635709	85.97
A5-2	*lpa*-Stage5-Rep2	32923070	4560694	28362376	86.15
A5-3	*lpa*-Stage5-Rep3	35648802	4705183	30943619	86.80
B1-1	WT-Stage1-Rep1	34398344	4224556	30173788	87.72
B1-2	WT-Stage1-Rep2	27280339	4104073	23176266	84.96
B1-3	WT-Stage1-Rep3	27057606	3150803	23906803	88.36
B2-1	WT-Stage2-Rep1	28287355	3619969	24667386	87.20
B2-2	WT-Stage2-Rep2	28989292	3999738	24989554	86.20
B2-3	WT-Stage2-Rep3	49825212	6280253	43544959	87.40
B3-1	WT-Stage3-Rep1	27413933	3351343	24062590	87.78
B3-2	WT-Stage3-Rep2	34132769	4213944	29918825	87.65
B3-3	WT-Stage3-Rep3	30625627	3438743	27186884	88.77
B4-1	WT-Stage4-Rep1	28535089	3412578	25122511	88.04
B4-2	WT-Stage4-Rep2	34569391	4407111	30162280	87.25
B4-3	WT-Stage4-Rep3	25021423	2920698	22100725	88.33
B5-1	WT-Stage5-Rep1	30823215	3603439	27219776	88.31
B5-2	WT-Stage5-Rep2	32384355	3963707	28420648	87.76
B5-3	WT-Stage5-Rep3	33332196	4438964	28893232	86.68
Total		**961451053**	**128004192**	**833446861**	**86.69**

Supplementary Table A3: DEGs associated with cellular glucan metabolism process.

	Genes	Stage 1			Stage 2		
		FC	log_2 ratio*	FDR	FC	log_2 ratio*	FDR
A. Cellulose Synthase							
1	Glyma06g46450	60.196	5.912	2.52E-03	1.300	0.378	1.00E+00
2	Glyma18g13105	15.789	3.981	1.16E-03	14.496	3.858	6.79E-04
3	Glyma05g29240	3.282	1.715	5.36E-02	3.220	1.687	3.92E-03
4	Glyma13g24270	2.792	1.481	5.75E-03	3.560	1.832	1.63E-04
5	Glyma09g05630	1.908	0.932	1.52E-01	3.878	1.955	1.99E-04
6	Glyma13g27250	1.868	0.902	1.69E-04	1.500	0.585	2.44E-02
7	Glyma06g06870	1.854	0.891	2.44E-01	3.139	1.650	7.10E-03
8	Glyma04g23530	1.802	0.850	7.14E-02	2.700	1.433	6.52E-05
9	Glyma06g30860	1.625	0.701	1.47E-01	2.954	1.563	3.62E-06
10	Glyma12g36570	1.623	0.699	6.10E-03	1.480	0.566	2.78E-02
11	Glyma15g16900	1.528	0.612	5.72E-01	4.418	2.144	1.16E-04
12	Glyma02g47076 †	0.521	-0.940	2.43E-02	0.576	-0.795	6.04E-03
13	Glyma04g43470 †	0.191	-2.387	1.44E-03	0.541	-0.887	4.56E-01
B. Xyloglucan endotransglucosylase/hydrolase							
1	Glyma09g07070	24.169	4.595	4.40E-03	0.000	Inf	1.24E-03
2	Glyma17g07220	9.609	3.264	3.34E-10	6.058	2.599	1.93E-07
3	Glyma15g18360	8.962	3.164	2.30E-01	0.000	Inf	2.63E-03
4	Glyma13g00280	6.804	2.766	2.46E-02	18.128	4.180	8.01E-04
5	Glyma13g01110	5.731	2.519	6.27E-04	3.886	1.958	2.11E-02
6	Glyma05g28310	3.973	1.990	4.95E-09	1.978	0.984	2.07E-02
7	Glyma08g11300	3.805	1.928	7.68E-04	2.821	1.496	1.91E-02
8	Glyma19g28220	3.189	1.673	3.77E-03	1.621	0.697	4.25E-01
9	Glyma09g34140	2.296	1.199	6.35E-07	3.020	1.595	1.61E-12
10	Glyma18g18931	2.240	1.164	2.19E-02	4.374	2.129	1.68E-06
11	Glyma06g45860	2.059	1.042	4.90E-03	1.668	0.738	7.47E-02
12	Glyma13g38040	1.997	0.998	4.44E-05	1.265	0.339	4.53E-01
13	Glyma01g01770	1.954	0.966	9.63E-05	2.237	1.161	3.12E-07
14	Glyma08g12800	1.822	0.866	1.47E-01	2.685	1.425	1.31E-03
15	Glyma05g26960	1.623	0.699	6.38E-03	1.290	0.367	3.64E-01
16	Glyma13g39710 †	0.795	-0.331	6.31E-01	3.239	1.695	6.77E-05

* Positive values of *log2* fold change (FC) ratio suggest up regulation of gene expression in *lpa* mutant. FC is calculated by dividing mean normalized gene expression value in *lpa* over that in wildtype. The *log2* FC ratio greater than 1 suggest that FC is greater than 2.

† Genes down regulated in *lpa* mutant are indicated by negative values of *log2* ratio.

Supplementary Table A4: DEGs associated with apoptosis process.

Genes	Stage 1			Stage 2			Stage 3		
	FC	log_2 ratio*	FDR	FC	log_2 ratio*	FDR	FC	log_2 ratio*	FDR
(1) NB-ARC domain-containing disease resistance protein									
Glyma14g38561	10.972	3.456	2.27E-03	12.795	3.678	9.17E-04	12.255	3.615	8.93E-04
Glyma14g38700	3.976	1.991	1.91E-02	1.859	0.895	3.13E-01	5.452	2.447	2.84E-03
Glyma19g31674	2.316	1.212	3.34E-04	2.156	1.108	1.08E-03	2.501	1.323	3.69E-04
Glyma16g03550	1.739	0.798	2.29E-02	2.101	1.071	1.38E-04	1.912	0.935	3.35E-03
Glyma07g07075	1.633	0.708	1.37E-01	1.829	0.871	2.64E-02	2.207	1.142	4.41E-03
Glyma07g06866	1.517	0.601	7.87E-02	1.835	0.876	1.48E-03	1.772	0.825	5.61E-03
Glyma09g02401	1.356	0.439	3.35E-01	1.816	0.861	8.63E-03	2.202	1.139	7.89E-04
Glyma14g38586	1.182	0.241	9.62E-01	2.002	1.001	3.35E-01	4.304	2.106	6.52E-03
Glyma08g43170	1.173	0.230	7.21E-01	1.755	0.812	9.78E-03	1.614	0.691	6.47E-02
Glyma19g31544	0.416	-1.265	1.67E-01	0.225	-2.152	7.59E-03	0.216	-2.209	2.57E-02
Glyma14g38533	0.116	-3.111	4.20E-16	0.214	-2.222	6.85E-09	0.362	-1.467	4.12E-04
Glyma01g01420	0.105	-3.256	3.37E-03	0.270	-1.890	1.07E-01	0.375	-1.416	2.12E-01
Glyma18g09340	0	-Inf	1.47E-03	0	-Inf	2.17E-02	0	-Inf	6.55E-02
Glyma18g09720	0	-Inf	1.52E-06	0	-Inf	8.19E-06	0	-Inf	1.54E-04
(A) LRR and NB-ARC domains-containing disease resistance protein									
Glyma03g04030	11.226	3.489	1.82E-17	7.770	2.958	1.12E-12	8.713	3.123	3.58E-13
Glyma03g04300	4.717	2.238	2.15E-03	5.296	2.405	8.54E-04	4.531	2.180	3.66E-03
Glyma03g04590	4.306	2.107	6.31E-05	3.993	1.997	4.67E-04	4.679	2.226	5.98E-05
Glyma03g14904	1.801	0.849	2.47E-02	1.813	0.858	8.29E-03	2.723	1.445	1.19E-05
Glyma20g12720	1.792	0.841	8.25E-02	1.804	0.851	7.47E-02	2.450	1.293	3.85E-03
Glyma11g33251	0.481	-1.057	1.99E-02	0.474	-1.078	8.08E-03	0.667	-0.585	2.25E-01
Glyma19g31528	0.325	-1.619	7.30E-03	0.324	-1.627	6.27E-03	0.188	-2.412	2.76E-04
Glyma15g37310	0.295	-1.763	3.28E-06	0.336	-1.574	1.10E-05	0.357	-1.487	2.37E-04
(I) Disease_resistance_protein_(TIR-NBS-LRR_class)_family									
Glyma06g41866	262.970	8.039	2.96E-20	127.032	6.989	5.62E-17	0	-Inf	6.56E-14
Glyma16g33681	9.540	3.254	3.51E-04	6.731	2.751	7.48E-03	5.524	2.466	1.28E-02
Glyma01g05710	5.993	2.583	2.08E-02	0.000	-Inf	2.37E-03	7.184	2.845	1.76E-01
Glyma09g08850	5.585	2.482	3.38E-03	9.504	3.249	6.85E-04	9.858	3.301	5.67E-03
Glyma06g41290	5.154	2.366	1.65E-03	3.786	1.921	3.49E-02	5.210	2.381	1.42E-02
Glyma0220s50	4.294	2.102	1.07E-02	5.256	2.394	2.64E-03	4.075	2.027	2.11E-02
Glyma12g16450	3.376	1.755	6.33E-05	1.260	0.333	8.04E-01	1.645	0.718	2.36E-01
Glyma16g10020	2.796	1.483	4.48E-05	2.719	1.443	1.83E-04	3.943	1.979	4.73E-07

Glyma11g21361	2.775	1.472	6.30E-05	2.611	1.385	2.61E-04	3.281	1.714	1.57E-05
Glyma08g41560	2.766	1.468	5.22E-04	2.521	1.334	3.23E-03	2.470	1.304	5.97E-03
Glyma06g41380	2.617	1.388	6.66E-03	2.163	1.113	4.83E-02	3.664	1.873	9.14E-05
Glyma20g10833	2.546	1.348	5.70E-03	2.204	1.140	2.79E-02	2.774	1.472	2.22E-03
Glyma16g23790	2.499	1.321	2.25E-02	1.311	0.391	8.31E-01	3.065	1.616	4.45E-03
Glyma12g15860	2.499	1.321	1.48E-02	3.645	1.866	4.42E-04	3.614	1.853	1.27E-03
Glyma16g27560	2.220	1.150	5.75E-03	1.446	0.532	2.06E-01	1.909	0.933	5.13E-02
Glyma08g41270	1.986	0.990	1.29E-02	2.293	1.197	1.27E-03	2.308	1.206	4.08E-03
Glyma06g41714	1.748	0.805	4.86E-01	4.821	2.269	1.08E-03	4.688	2.229	2.46E-03
Glyma16g34070	1.746	0.804	8.18E-03	1.283	0.359	3.06E-01	1.416	0.502	1.88E-01
Glyma06g15123	1.652	0.724	5.49E-02	1.178	0.237	6.94E-01	2.308	1.207	1.94E-04
Glyma03g22071	1.625	0.700	5.64E-02	1.849	0.887	1.03E-02	1.898	0.924	8.94E-03
Glyma16g10080	1.490	0.575	5.61E-02	1.366	0.450	1.36E-01	1.861	0.896	9.46E-04
Glyma15g37283	1.371	0.456	3.33E-01	1.541	0.624	8.90E-02	2.343	1.228	7.16E-05
Glyma06g40980	1.314	0.394	2.65E-01	1.082	0.113	9.72E-01	1.938	0.955	3.84E-04
Glyma03g14888	1.239	0.309	5.31E-01	1.308	0.388	4.12E-01	1.892	0.920	3.36E-03
Glyma06g40710	0.482	-1.051	1.91E-01	0.097	-3.364	1.22E-11	0.139	-2.850	5.13E-08
Glyma06g40950	0.388	-1.364	1.91E-02	0.260	-1.942	2.60E-04	0.339	-1.560	6.33E-03
Glyma06g40780	0.127	-2.974	3.41E-06	0.028	-5.165	1.65E-16	0.054	-4.219	9.06E-12
Glyma06g41896	0.118	-3.078	1.86E-08	0.142	-2.814	1.06E-06	0.123	-3.026	1.46E-07
(II) Disease resistance protein (CC-NBS-LRR class) family									
Glyma17g20860	1.857	0.893	3.13E-03	1.837	0.878	5.88E-03	2.516	1.331	3.21E-06
(2) Cysteine proteinases superfamily protein									
Glyma18g51741	3.505	1.809	2.41E-03	3.485	1.801	3.77E-03	3.583	1.841	2.43E-02
Glyma18g51715	3.230	1.691	6.71E-07	3.058	1.613	3.24E-06	2.304	1.204	5.56E-03
(3) BCL-2-associated athanogene 1									
Glyma16g32280	2.235	1.160	1.02E-05	3.056	1.612	3.32E-11	2.181	1.125	1.13E-05
Glyma16g32290	1.729	0.790	8.09E-03	2.082	1.058	1.95E-05	1.713	0.776	5.07E-03
Glyma09g27100	1.280	0.356	5.03E-01	1.733	0.794	9.84E-03	1.748	0.806	1.87E-02
(4) ADR1-like 1									
Glyma14g08700	2.567	1.360	5.68E-05	1.442	0.528	2.92E-01	1.377	0.462	4.24E-01
Glyma17g36420	2.177	1.122	2.05E-05	1.986	0.990	3.03E-04	1.864	0.898	1.50E-03
(5) Protein kinase superfamily protein									
Glyma08g13040	1.615	0.692	5.42E-02	2.210	1.144	9.58E-05	2.448	1.291	9.71E-05
(6) P-loop containing dNTP hydrolases superfamily protein									
Glyma04g12660	0.654	-0.612	1.56E-01	0.454	-1.138	4.43E-05	0.559	-0.839	1.00E-02

* Positive values of the *log2* fold change (FC) ratio suggest up-regulation of gene expression in *lpa* mutant. FC is calculated by dividing mean normalized gene expression value in *lpa* over that in Wildtype.

Supplementary Table A5: Enriched transmembrane multidrug transporter genes up regulated in *lpa* mutant.

	Genes	Stage 2	
		log_2 ratio*	FDR
A. Major facilitator superfamily			
1	Glyma01g44930	1.566	2.2E-08
2	Glyma03g40121	2.271	8.4E-09
3	Glyma04g03850	1.348	9.9E-03
4	Glyma06g45000	1.070	3.7E-04
5	Glyma07g40250	1.194	2.8E-05
6	Glyma08g21810	1.233	4.4E-04
7	Glyma10g42340	1.261	3.6E-03
8	Glyma12g12290	0.947	5.8E-03
9	Glyma13g26760	1.101	6.0E-03
10	Glyma13g40450	2.503	3.0E-04
11	Glyma14g34750	1.009	9.4E-05
12	Glyma17g00550	2.513	5.5E-20
13	Glyma18g11640	3.819	4.5E-05
14	Glyma18g53710	1.120	3.6E-04
15	Glyma19g36940	1.981	3.0E-03
B. MATE efflux carrier superfamily			
1	Glyma02g09941	1.259	1.5E-03
2	Glyma12g32010	0.785	3.1E-03
3	Glyma15g43021	1.063	8.1E-04
4	Glyma17g14550	2.970	3.3E-08
5	Glyma18g43740	1.731	3.1E-04
6	Glyma04g34560	1.461	8.1E-03
C. Multidrug resistance-associated superfamily			
1	Glyma04g15306	1.309	1.2E-03
2	Glyma10g02370	1.140	2.5E-03
3	Glyma13g18960	0.758	3.1E-03
4	Glyma18g09000	0.821	4.1E-03
5	Glyma20g30490	1.162	6.5E-03
D. P-glycoprotein superfamily			
1	Glyma09g33880	0.813	6.1E-03
2	Glyma17g37860	0.913	2.8E-03
E. ATP binding cassette subfamily B4			
1	Glyma02g01100	0.996	6.7E-05

* Positive values of *log2* fold change (FC) ratio suggest up regulation of gene expression in *lpa* mutant. FC is calculated by dividing mean normalized gene expression value in *lpa* over that in wildtype.

Supplementary Table A6: DEGs associated with photosynthesis process.

Gene Description†	log_2 ratio*		
	Stage 3	Stage 4	Stage 5
PS I subunit II (psaD):			
Glyma10g39460	-0.589	-0.707	-0.703
Glyma20g28300	-0.957	-0.949	-1.053
PS I subunit IV (psaE):			
Glyma13g19190	-0.413	-0.798	-1.282
PS I subunit III (psaF):			
Glyma05g00620	-1.050	-0.863	-1.087
PS I subunit V (psaG):			
Glyma04g12510	-0.629	-0.708	-0.797
Glyma06g48030	-0.404	-0.463	-0.760
PS I subunit VI (psaH):			
Glyma07g02240	-0.666	-0.450	-0.789
Glyma08g21900	-0.589	-0.376	-0.790
Glyma13g43370	-0.821	-0.725	-0.951
Glyma15g01940	-0.734	-0.765	-1.051
PS I subunit X (psaK):			
Glyma15g22780	-0.652	-0.531	-0.773
PS I subunit XI (psaL):			
Glyma18g47710	-0.833	-0.902	-0.978
PS I subunit PsaN:			
Glyma06g15540	-0.096	-0.340	-0.697
PS II P680 reaction center D1 protein (psbA):			
Glyma13g15560	-3.775	-5.099	-4.841
PS II cytochrome b559 subunit alpha (psbE):			
Glyma12g36132	-5.063	-7.445	-7.293
Glyma15g38110	-3.860	-5.192	-6.533
PS II oxygen-evolving enhancer protein 2 (psbP):			
Glyma08g41660	-0.700	-0.435	-0.917
Glyma14g03560	-0.492	-0.645	-0.827
Glyma18g14410	-0.639	-0.373	-0.964
Glyma18g52340	-0.272	-0.265	-1.087
PS II oxygen-evolving enhancer protein 3 (psbQ):			
Glyma01g06110	-2.43	-2.37	-2.27
Glyma03g26740	-0.85	-0.69	-1.25
Glyma07g14340	-0.63	-0.72	-0.87
PS II PsbW protein:			
Glyma03g31580	-0.541	-0.516	-0.812
Glyma19g34410	-0.887	-0.859	-1.277
PS II PsbX protein:			
Glyma09g07880	-0.961	-0.601	-1.148
Glyma09g08630	-0.795	-0.858	-1.186
Glyma15g19100	-0.795	-0.590	-0.855
PS II PsbY protein:			
Glyma05g30600	-0.456	-0.552	-0.728
Glyma11g37580	-1.026	-1.178	-1.548
LHC I chlorophyll a/b binding protein 1 (LHCA1):			
Glyma02g07180	-0.626	-0.568	-0.973

Glyma16g26130	-0.795	-0.687	-0.897
LHC I chlorophyll a/b binding protein 2 (LHCA2):			
Glyma03g42310	-0.569	-0.646	-0.762
Glyma15g19810	-0.793	-0.299	-0.951
LHC II chlorophyll a/b binding protein 1 (LHCB1):			
Glyma05g25810	-0.307	-0.398	-1.026
Glyma08g08770	-0.077	-0.235	-0.635
Glyma09g07310	-0.523	-0.043	-1.782
Glyma16g28030	-0.360	-0.048	-1.216
Glyma16g28070	-0.267	-0.243	-1.378
LHC II chlorophyll a/b binding protein 4 (LHCB4):			
Glyma01g28810	-0.497	-0.574	-0.773
Glyma03g08280	-0.525	-0.499	-0.663
LHC II chlorophyll a/b binding protein 5 (LHCB5):			
Glyma09g28200	-0.581	-0.464	-0.983
Glyma10g32080	-0.374	-0.545	-0.658
Glyma16g33030	-0.472	-0.337	-0.894
Magnesium protoporphyrin IX methyltransferase (CHLM):			
Glyma08g01060	-0.794	-0.417	-0.931

† PS=Photosystem; LHC=Light harvesting complex

* Positive values of *log2* fold change (FC) ratio suggest up regulation of gene expression in *lpa* mutant. FC is calculated by dividing mean normalized gene expression value in *lpa* over that in wildtype.

Supplementary Table A7: DEGs associated with enriched glycolysis process.

Gene ID	Stage 3			Stage 4			Stage 5		
	FC	*log2* ratio*	FDR	FC	*log2* ratio*	FDR	FC	*log2* ratio*	FDR
Phosphofructokinase family protein									
Glyma17g01361	1.947	0.961	7.31E-03	1.308	0.387	7.06E-01	1.471	0.557	2.03E-01
Glyma01g00870	2.166	1.115	1.85E-05	1.276	0.352	4.69E-01	0.971	-0.043	9.85E-01
Pyruvate kinase family protein									
Glyma02g24921	0.635	-0.656	2.01E-02	0.740	-0.435	3.39E-01	0.531	-0.914	3.73E-04
Glyma02g28212	0.836	-0.259	7.66E-01	0.694	-0.528	3.77E-01	0.541	-0.885	9.86E-03
Glyma05g09310	0.829	-0.27	6.75E-01	0.975	-0.036	1.00E+00	0.532	-0.911	6.61E-03
Glyma03g34740	1.676	0.745	2.99E-03	1.173	0.23	6.87E-01	0.977	-0.033	1.00E+00
Glyma19g37420	1.369	0.453	2.36E-01	0.918	-0.124	1.00E+00	0.543	-0.88	7.01E-03
Glyma20g33060	0.874	-0.194	7.36E-01	0.921	-0.119	1.00E+00	0.550	-0.863	4.91E-04
Glyma10g37210	0.655	-0.611	2.68E-02	0.748	-0.418	4.16E-01	0.592	-0.757	2.32E-03
Aldolase superfamily protein									
Glyma02g38730	0.744	-0.427	3.57E-01	0.808	-0.307	8.38E-01	0.529	-0.918	3.44E-03
Glyma12g24190	0.024	-5.398	1.45E-16	0.179	-2.481	9.77E-06	0.384	-1.382	1.08E-01
Glyma14g36850	0.660	-0.599	2.50E-01	0.624	-0.68	3.56E-01	0.460	-1.119	3.41E-03
Glyma11g11870	0.450	-1.151	1.17E-05	0.451	-1.15	1.13E-05	0.443	-1.176	3.69E-06
Glyma12g04150	0.607	-0.721	4.10E-03	0.563	-0.83	6.41E-04	0.660	-0.599	2.10E-02
Phosphoglycerate mutase family protein									
Glyma04g11380	0.636	-0.652	1.61E-01	0.511	-0.969	2.74E-02	0.452	-1.146	3.54E-03
Glyma08g15280	0.028	-5.152	1.15E-14	0.171	-2.546	3.75E-05	1.313	0.393	7.32E-01
Glyma15g41540	0.550	-0.862	3.55E-04	0.573	-0.804	2.31E-03	0.519	-0.945	1.93E-05
Sugar isomerase (SIS) family protein									
Glyma06g03560	0.115	-3.119	1.56E-21	0.233	-2.1	1.18E-12	0.326	-1.619	7.26E-07
Pentatricopeptide repeat (PPR) superfamily protein									
Glyma06g09785	1.695	0.761	3.90E-03	1.272	0.347	5.18E-01	1.294	0.372	3.31E-01

* Positive values of *log2* fold change (FC) ratio suggest up regulation of gene expression in *lpa* mutant. FC is calculated by dividing mean normalized gene expression value in *lpa* over that in wildtype.

Appendix B. Supplementary Tables from Chapter 3

Supplementary Table B1: Summary statistics for sequencing data analysis.

Experimental lines	Stages	Rep	Raw Reads*	Mapped Reads#	Percent Alignment	Experimental lines	Stages	Rep	Raw Reads*	Mapped Reads#	Percent Alignment
	1	1	30.51	26.13	86%		1	1	23.34	19.82	85%
		2	30.50	25.88	85%			2	29.91	25.02	84%
		3	31.56	26.79	85%			3	21.55	18.57	86%
	2	1	37.23	31.64	85%		2	1	30.91	25.45	82%
		2	35.40	30.13	85%			2	28.51	23.25	82%
		3	37.61	32.43	86%			3	26.22	21.89	83%
	3	1	30.31	26.43	87%		3	1	27.21	22.36	82%
3mlpa		2	35.61	30.61	86%	**1MWT**		2	32.01	27.17	85%
		3	29.00	24.64	85%			3	49.59	42.10	85%
	4	1	37.31	32.32	87%		4	1	27.98	24.29	87%
		2	29.61	25.66	87%			2	28.73	24.96	87%
		3	28.07	24.30	87%			3	29.04	25.14	87%
	5	1	27.49	23.64	86%		5	1	37.19	32.75	88%
		2	32.92	28.36	86%			2	30.33	26.41	87%
		3	35.65	30.94	87%			3	33.45	28.99	87%
	1	1	34.40	30.17	88%		1	1	26.49	22.36	84%
		2	27.28	23.18	85%			2	32.45	27.40	84%
		3	27.06	23.91	88%			3	31.58	27.39	87%
	2	1	28.29	24.67	87%		2	1	33.17	27.38	83%
		2	28.99	24.99	86%			2	30.08	25.95	86%
		3	49.83	43.54	87%			3	32.55	27.89	86%
3MWT	3	1	27.41	24.06	88%	***1mlpa***	3	1	28.34	24.29	86%
		2	34.13	29.92	88%			2	27.94	24.11	86%
		3	30.63	27.19	89%			3	26.17	22.77	87%
	4	1	28.54	25.12	88%		4	1	32.91	28.53	87%
		2	34.57	30.16	87%			2	27.65	24.01	87%
		3	25.02	22.10	88%			3	38.09	32.94	86%
	5	1	30.82	27.22	88%		5	1	21.47	18.57	87%

		2	32.38	28.42	88%			2	28.84	25.29	88%
		3	33.33	28.89	87%			3	33.04	28.94	88%

* indicates raw sequencing reads generated for five seed developmental stages for each of the four experimental lines.
# indicates the total number of reads mapped to the soybean reference Williams 82 genome sequence using TopHat2.

Supplementary Table B2: The regulatory nodes from co-expression network.

Nodes*	3MWT +	1MWT +	*3mlpa* +	*1mlpa* +		3MWT −	1MWT −	*3mlpa* −	*1mlpa* −		3MWT + 1MWT +	*3mlpa* + *1mlpa* +		3MWT + 1MWT −	*3mlpa* + *1mlpa* −	
AAATL	Y	Y	Y	Y							Y	Y				
AAE16								Y	Y	*					Y	*
AAE18							Y	Y								
ABA2				Y												
ABC33		Y	Y	Y							Y	Y				
ABH	Y															
ACA1						Y	Y	Y	Y					Y	Y	
ACA4								Y						Y	Y	
ACAT2						Y								Y		*
ACO3								Y								
ACR8		Y														
ACT							Y							Y		*
ACTP	Y	Y	Y	Y		Y	Y	Y	Y		Y	Y		Y		*
AHA11									Y					Y		*
AHA2						Y			Y							
AIM1		Y														
AK-HSDH2		Y									Y		*			
AK-LYS1						Y		Y						Y		*
ALA3								Y							Y	*
ALDH11A3			Y													
ALDH1A		Y														
ALDH2A	Y		Y	Y			Y				Y		*			
ALDH5F1	Y	Y	Y	Y							Y	Y				
AMPSL						Y										
Apase															Y	*
AR401								Y							Y	*
ECI3									Y							
NADK1						Y			Y					Y	Y	
ATPE1E2		Y					Y									
PRMT11						Y										
PRMT3						Y										
PRMT6						Y										
LPAT1								Y								
UPRT1							Y		Y							
CER8							Y									
Clp							Y	Y	Y					Y		*
CYSC1						Y		Y								
CYSD1		Y														
DCI1								Y	Y	*					Y	*
DECR														Y		*
ECA4								Y						Y	Y	
Ester Hydrolase									Y						Y	*
FDH	Y	Y	Y	Y							Y	Y				
Folic Transferase	Y										Y		*			
GAE3														Y		*
GID1C							Y									
GlcNAc1pUT1	Y															
GLT1		Y	Y									Y	*			
GPAT8								Y								
HAD	Y	Y			*	Y		Y								
HADH															Y	*
HCF173		Y	Y	Y							Y	Y				
HMA1									Y						Y	*

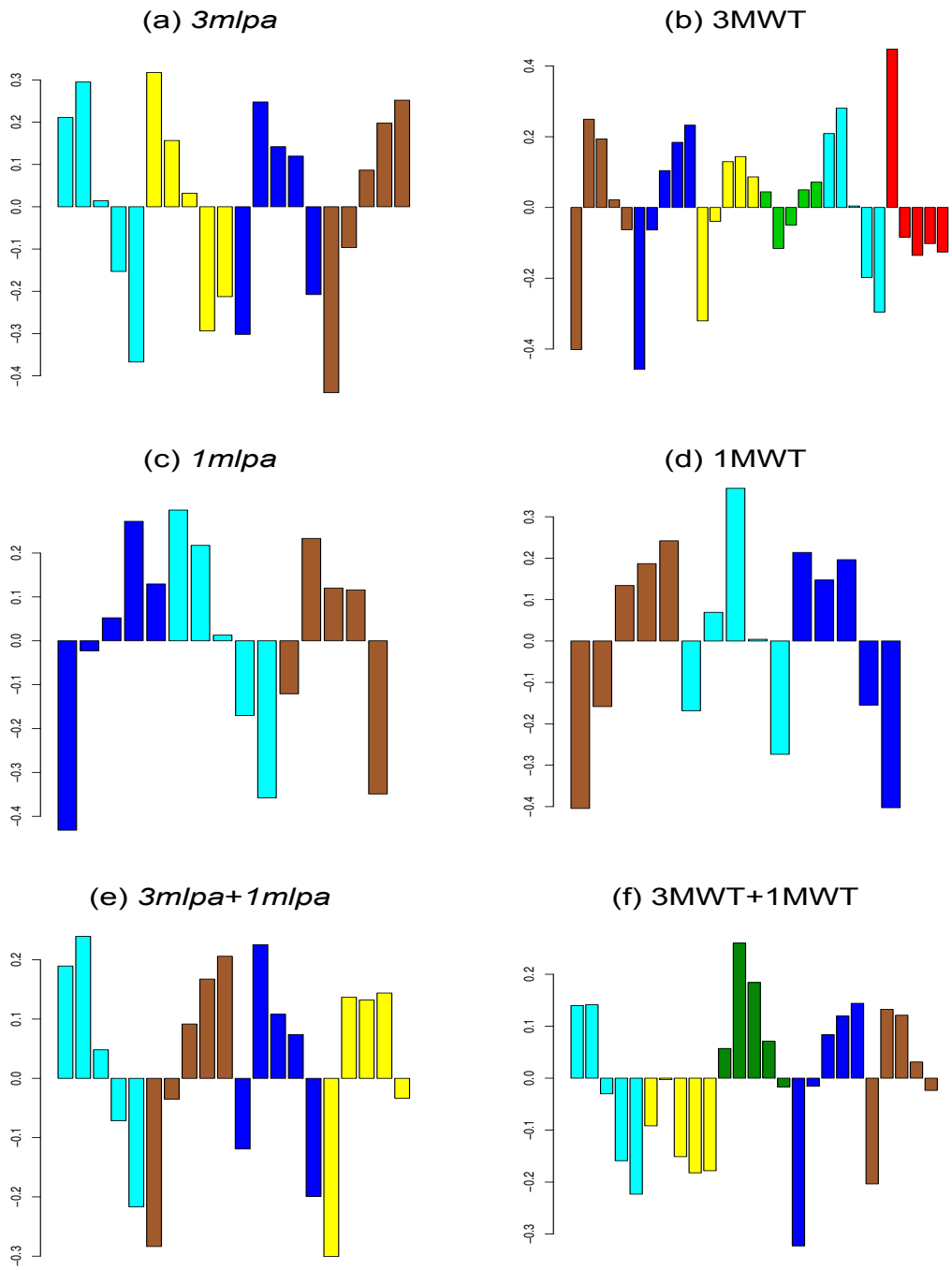
HMA5							Y									
HSD1	Y	Y	Y	Y							Y	Y				
HSD5	Y	Y	Y	Y							Y	Y				
IAR3				Y							Y		*			
KCR1															Y	*
KCR2				Y												
LACS7								Y						Y	Y	
LACS8		Y	Y									Y	*			
LPAT2								Y							Y	*
LPAT5						Y			Y							
MCAP									Y							
MENG	Y		Y			Y	Y	Y	Y			Y	*	Y	Y	
MIA						Y		Y						Y	Y	
MMP							Y									
MTO2	Y	Y		Y							Y	Y				
NADP-ROSS	Y	Y	Y	Y		Y			Y		Y	Y		Y	Y	
NAGS1				Y							Y		*			
NFS1								Y								
NIC1						Y	Y		Y					Y	Y	
NIC2						Y										
NMT1							Y									
OSU1							Y		Y						Y	*
Oxidoreductase								Y							Y	*
P5CS2							Y		Y							
PGDH							Y									
PLDBETA1								Y								
PLDDELTA			Y	Y	*						Y		*			
PLPT							Y		Y					Y		*
PORA	Y	Y		Y							Y	Y				
PRIBO									Y							
PRK	Y	Y	Y								Y	Y				
QUL2						Y								Y	Y	
RimM						Y										
rRNA						Y										
SHM7									Y							
UGE1														Y	Y	
UGP2		Y	Y									Y	*			
UGT72B1	Y	Y	Y	Y			Y	Y			Y	Y				
UGT73B1			Y													
UGT73B3	Y		Y	Y							Y	Y				
UGT73B4	Y		Y	Y							Y	Y				
UGT73B5	Y			Y							Y	Y				
UGT74E2	Y		Y	Y								Y	*			
UGT78D2		Y		Y							Y		*			
UGT80B1							Y	Y								
UGT84B1	Y															
UGT85A1			Y													
UGT85A3	Y															
UGT88A1				Y							Y	Y				
WAV2								Y							Y	*

Y – indicates whether the regulatory node was represented in the module. The nodes that differ between mutant and wildtype are highlighted with light blue color.

* – Protein names are the gene symbols of Arabidopsis genes that share sequence similarity with soybean genes.

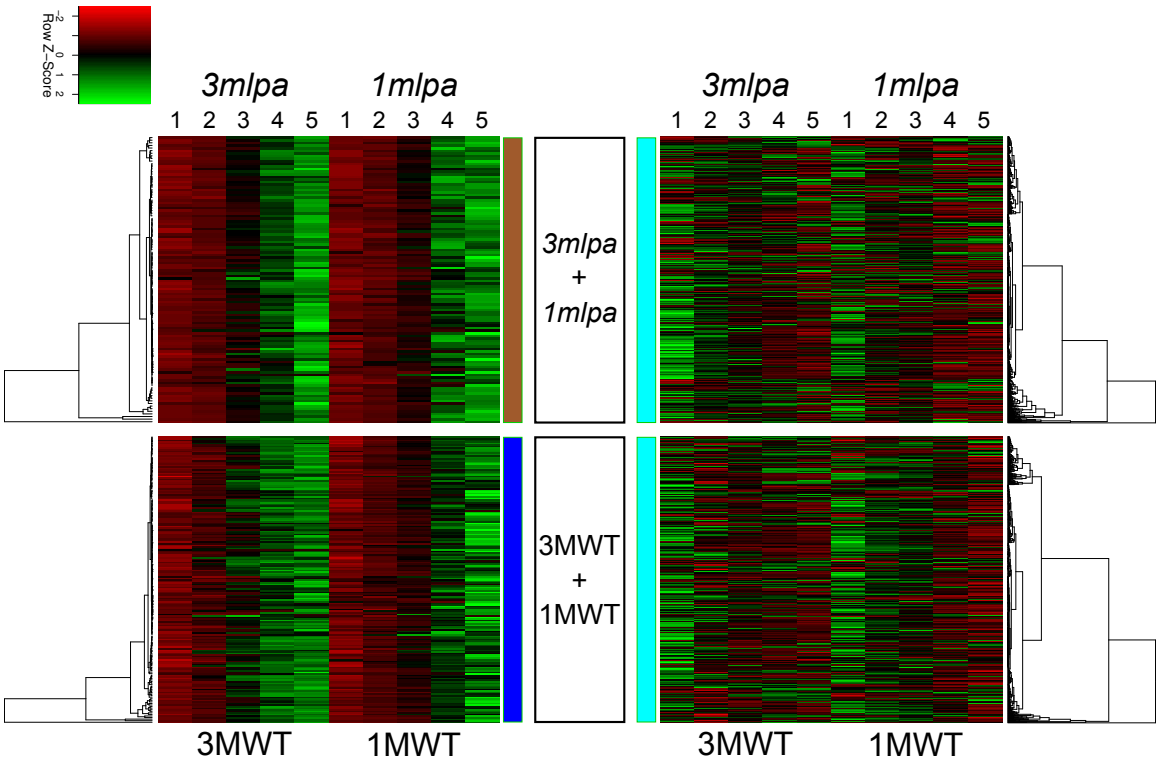
Appendix C. Supplementary Figures from Chapter 3

Supplementary Figure C1: Expression profiles of module-specific eigengenes.



Supplementary figure C1: Expression profiles of module-specific eigengenes. For each experimental line, there is a module-specific eigengene, whose values represent the summary profile of co-expressed genes at seed developmental stages. These eigengene profiles are represented as bar graphs with module specific colors. The five bars indicate five seed developmental stages from left to right.

Supplementary figure C2: Expression profiles of co-expressed genes within positively and negatively correlated modules of combined network.



Supplementary figure C2: Expression profiles of co-expressed genes within positively (left) and negatively (right) correlated modules of combined network. The respective module colors are indicated in color bars right next to heat map. Red and green bars indicate low and high gene expression levels, respectively.

Appendix D. Supplementary Tables from Chapter 4

Supplementary Table D1: Number of genomic sequencing reads.

Sample ID	Raw reads	Bases in raw reads	Barcode selected reads*
1MWT-001	38165768	11449730400	37168939 (97.38%)
1MWT-002	38532749	11559824700	37543938 (97.43%)
1mlpa-001	41123479	12337043700	40669182 (98.89%)
1mlpa -002	41606451	12481935300	41160521 (98.90%)

* indicates the number of reads with library-specific barcode index. It is also represented as the percentage of raw reads in brackets.

Supplementary Table D2: Barcode index diversity in the genomic sequencing data.

#	Barcode	1MWT-001	1MWT-002	Barcode	*1mlpa*-001	*1mlpa*-002
1	AAAGTG	3664	3397	ACCAAT	85307	83902
2	ACAATG	15537	15650	CCCAAT	6142	6236
3	ACACTG	2469	2528	GACAAT	12051	11556
4	ACAGAG	60786	60964	GCAAAT	22678	24280
5	ACAGCG	719531	722575	GCCAAA	41146	41436
6	ACAGGG	6161	6260	GCCAAC	10168	10154
7	ACAGNG	1244	2887	GCCAAG	15226	17689
8	ACAGTA	43478	44208	GCCAAN	709	393
9	ACAGTC	4956	5072	GCCAAT*	40669182	41160521
10	ACAGTG*	37168939	37543938	GCCACT	68581	68826
11	ACAGTN	628	362	GCCAGT	9478	10013
12	ACAGTT	29582	30152	GCCANT	1311	3029
13	ACANTG	2782	489	GCCATT	2054	2024
14	ACATTG	17344	17454	GCCCAT	15316	16382
15	ACCGTG	17058	17551	GCCGAT	18418	18905
16	ACGGTG	7713	7730	GCCNAT	3160	498
17	ACNGTG	148	112	GCCTAT	5257	5398
18	ACTGTG	2352	2395	GCGAAT	4171	4474
19	AGAGTG	2519	2134	GCNAAT	140	128
20	ATAGTG	22078	20145	GCTAAT	22973	23328
21	CCAGTG	4613	4965	GGCAAT	4242	3976
22	GCAGTG	9265	9436	GTCAAT	36303	34986
23	NCAGTG	19603	8867	NCCAAT	21451	9365
24	TCAGTG	3318	3478	TCCAAT	48015	48952

* indicates the barcode index used for genomic DNA library preparation.

Appendix E. Supplementary Tables from Chapter 5

Supplementary Table E1: Soybean cultivars used in this study.

#	Genotype†	Reaction to SMV strains							Resistance genes
		G1	G2	G3	G4	G5	G6	G7	
1	Williams 82	S	S	S	S	S	S	S	*rsv*
2	Lee68	S	S	S	S	S	S	S	*rsv*
3	Essex	S	S	S	S	S	S	S	*rsv*
4	Archer	-	-	-	-	S	S	S	*rsv*
5	Hutcheson	R	R	R	N	S	S	S	*rsv*
6	York	R	R	R	N	S	S	S	*rsv*
7	L29	S	S	S	S	R	R	R	*Rsv3*
8	Tousan140	N	N	R	NA	R	R	R	*Rsv3*
9	Suweon 97	R	R	R	R	R	R	R	*Rsv1, Rsv3*
10	Hardee	S	-	-	-	-	-	R	*Rsv3*
11	Harosoy	S	S	S	S	R	R	R	*Rsv3*
12	Hourei	N	R	R	NA	R	R	R	*Rsv1, Rsv3*
13	Columbia	R	R	R	N	R	R	R	*Rsv3, Rsv4*
14	RRR	R	R	R	R	R	R	R	*Rsv1, Rsv3, Rsv4*
15	Paoting	S	S	S	NA	R	R	R	*Rsv3*
16	PI323555	S	S	S	NA	R	R	R	*Rsv3*
17	PI323556	S	S	S	NA	R	R	R	*Rsv3*
18	VIR 5532	S	S	S	NA	R	R	R	*Rsv3*
19	PLSO-63	S	S	S	NA	R	R	R	*Rsv3*
20	PLSO-70	S	S	S	NA	R	R	R	*Rsv3*
21	OCB 81	S	S	S	NA	R	R	R	*Rsv3*
22	PI91346	S	S	S	NA	R	R	S	*Rsv3-new*
23	PI61947	N/S	N/S	R/N	NA	R	R	R	*Rsv3-h*
24	PI399091	S	S	R	NA	ER	S	ER	*Rsv3-c*
25	Tousan101	R	R	R	NA	R	R	R	*Rsv3, Rsv4*
26	Enoki	S	S	S	NA	R	R	R	*Rsv3*
27	Tej sen da baj pi	S/R	S	R/S	NA	R	R	R/N	*Rsv3-new*
28	Kolhida4	S/R	S	S	NA	S	S	N	*Rsv3-new*
29	Shirome choutan	R	R	R	NA	R	R	R	*Rsv3, Rsv4*
30	Graine Jaune Unie	S	S	S	NA	R	R	R	*Rsv3*

† **Additional information on selected soybean cultivars:** (1) L29 is an isoline of Williams82, with resistant gene from Hardee [Bernard, et al. 1991]; (2) Suweon 97 was designated as Hwangkeum (or, Hwang-Kum) in Korea [Chen, et al. 2002, Yu, et al. 2008, Jeong, et al. 2014]; (3) *Rsv3* locus in RRR is derived from L29 [Maroof, et al. 2008]. (3) Disease response for these cultivars was collected from following citations [Chen, et al. 1991, Gunduz, et al. 2002, Ma, et al. 2002, Li, et al. 2010, Shakiba, et al. 2012].

Supplementary Table E2: Primers used in this study.

Description	Primer Name	Sequence (5' - 3')
SNP validation in Glyma14g38500	NR500snp-F	TCCACAAGGCATGAATCACT
	NR500snp2-R	GGAGCTCACCCAAATAGAGC
SNP validation in Glyma14g38510	NR510snp1-F	ATCTCTTAGCTGGCAATGAAGC
	NR510snp1-R	TTCAGCTGCAAACATCTTCCTA
	NR510snp2-F	TACTTTGGCAAAAGAGGTTGGT
	NR510snp2-R	ACATCTTCGTAACCTCCCTCAA
	NR510snp4-F	CTTGACTTGCGTGGTTCTACTTTTA
	NR510snp4-R	GAGAGGCATTTTATCTCTGGACAAT
	NR510snp5-F	GTACTCATTCATCATTAGGACTCTGC
	NR510snp5-R	AGATAGTCTTTGTGCTCTAGCTTCC
	NR510snp6-F	GCAAAGTTAGCAGAATATACAGTGG
	NR510snp6-R	GTCTCAGATAGTCTTTGTGCTCTAGCTT
39 bp deletion validation in Glyma14g38540	Del540-F	ATAATTGGCGTAGGCGTGAATT
	Del540R2	AATGACGCGATTCCATTAGG
3 bp and 12 bp insertion validation in deletions in Glyma14g38540	NRDels533-F	GGAATTTCCGCATAATATCTCCT
	NRDels533-R	TAGGCTGGAGAACGCATCTT
21 bp insertion validation in Glyma14g38540 gene	NRInser533-F	GAGCTGCTAGAACAAGTGAGAATATC
	NRInser533-R	CGAATTTGGTAAATGTGGATCA
SNP validation in Glyma14g38560	Rsv560F	GGTTGAAGGAGTTGCCGAAT
	Rsv560BR	CATGTTCACAACTCACGGTTC
SNP validation in Glyma14g38590	590N-F	GACTAGGCTGGAGAACGCAT
	590N-R	ATGGACCCCTCATACTGGAG

Supplementary Table E3: *Rsv3*-candidate gene annotations.

Annotation 1.0	Annotation 1.1
Glyma14g38500	Glyma14g38500
Glyma14g38520	Glyma14g38516
Glyma14g38540	Glyma14g38533
Glyma14g38560	Glyma14g38561
Glyma14g38590	Glyma14g38586

Supplementary Table E4: Total polymorphisms identified in Glyma14g38533 gene.

#	Genomic Position	Coding for	Wm82†	L29†	Hwangkeum†	RRR†	Domains‡	Synonymous SNPs	Verified in other lines
1	47669986	Exon1	C	T	T	T		N	
2	47670031	Exon1	T	A	A	A		N	
3	47670038	Exon1	C	G	G	G			
4	47670048	Exon1	A	G	G	G			
5	47670165	Exon1	A	T	T	T			
6	47670253	Exon1	C	A	A	A		N	
7	47670494	Exon1	A	G	A	G	NB-ARC		
8	47670870	Exon1	T	A	A	A	NB-ARC		
9	47671432	Exon1	T	C	C	C		N	
10	47671483	Exon1	A	G	G	G		N	
11	47671485	Exon1	T	C	C	C			
12	47671495	Exon1	G	A	A	A		N	
13	47671523	Exon1	A	G	G	G			
14	47671524	Exon1	T	A	A	A			
15	47671573	Exon1	T	A	A	A	LRR1		
16	47671574	Exon1	C	A	A	A	LRR1		
17	47671601	Exon1	A	G	G	G			
18	47671609	Exon1	A	T	T	T			
19	47671656	Exon1	A	T	T	T	LRR2		
20	47671657	Exon1	T	A	A	A	LRR2		
21	47671680	Exon1	A	T	T	T			*
22	47671683	Exon1	T	A	A	A			*
23	47671689	Exon1	C	T	T	T			*
24	47671691	Exon1	AGCTATAATTTCCTTAGGCGTGAATTGAACAAGGCATGT	-	-	-			*
25	47671808	Exon1	T	C	C	C	LRR3		*
26	47671811	Exon1	G	A	A	A	LRR3		*
27	47671812	Exon1	A	T	T	T	LRR3		*
28	47671871	Exon1	G	A	A	A	LRR4		*
29	47671872	Exon1	G	A	A	A	LRR4		*
30	47671940	Exon1	C	A	A	A	LRR5		*

31	47671946	Exon1	T	G	G	G	LRR5		*
32	47671947	Exon1	C	T	T	T	LRR5		*
33	47671953	Exon1	A	G	G	G	LRR5		
34	47671958	Exon1	A	C	C	C			
35	47671988	Exon1	A	T	T	T			
36	47672013	Exon1	C	G	G	G			
37	47672018	Exon1	C	T	T	T			
38	47672019	Exon1	C	A	A	A			
39	47672022	Exon1	C	T	T	T			
40	47672025	Exon1	A	G	G	G			
41	47672030	Exon1	A	T	T	T			
42	47672059	Exon1	C	G	G	G			
43	47672085	Exon1	T	A	A	A			*
44	47672100	Exon1	C	G	G	A			*
45	47672110	Exon1	CTG	-	-	-			*
46	47672123	Exon1	ATGATGGAAGGC	-	-	-			*
47	47672242	Exon1	T	C	C	C		N	*
48	47672252	Exon1	G	A	A	A			*
49	47672253	Exon1	A	G	G	G			*
50	47672255	Exon1	A	G	G	G			*
51	47672261	Exon1	G	A	A	A			*
52	47672319	Exon1	T	C	C	C	LRR6		*
53	47672333	Exon1	C	G	G	G	LRR6		*
54	47672335	Exon1	T	A	A	A	LRR6		*
55	47672366	Exon1	A	G	G	G			*
56	47672373	Exon1	A	C	A	C			
57	47672374	Exon1	T	A	A	A			
58	47672375	Exon1	G	A	A	A			
59	47672377	Exon1	T	G	G	G			
60	47672378	Exon1	GATCTGCTG	-	-	-			
61	47672423	Exon1	A	C	A	C	LRR7		
62	47672425	Exon1	C	C	G	C	LRR7		
63	47672429	Exon1	T	T	C	T	LRR7		
64	47672430	Exon1	A	A	G	A	LRR7		
65	47672433	Exon1	G	A	T	A	LRR7		
66	47672505	Exon1	C	G	G	G	LRR8		
67	47672515	Exon1	G	T	T	T	LRR8		
68	47672518	Exon1	T	C	C	C	LRR8	N	

69	47672574	Exon1	C	T	T	T	LRR9		
70	47672585	Exon1	A	G	G	G	LRR9		
71	47672586	Exon1	G	A	A	A	LRR9		
72	47672587	Exon1	A	T	T	T	LRR9		
73	47672588	Exon1	G	C	C	C	LRR9		
74	47672625	Exon1	C	T	T	T			
75	47672658	Exon1	G	C	C	C	LRR10		
76	47672668	Exon1	A	C	C	C	LRR10		
77	47672673	Exon1	A	T	T	T	LRR10		
78	47672755	Exon1	C	T	C	T		N	
79	47672769	Exon1	C	C	T	C			
80	47672780	Exon1	C	C	G	C			
81	47672781	Exon1	G	T	A	T	LRR11		
82	47672783	Exon1	G	C	A	C	LRR11		
83	47672789	Exon1	C	C	G	C	LRR11		
84	47672790	Exon1	G	A	A	A	LRR11		
85	47672792	Exon1	A	A	G	A	LRR11		
86	47672794	Exon1	C	C	G	C	LRR11		
87	47672798	Exon1	A	G	G	G			
88	47672835	Exon1	T	C	T	C			
89	47672846	Exon1	A	G	G	G	LRR12		
90	47672853	Exon1	T	A	A	A	LRR12		
91	47672854	Exon1	G	C	C	C	LRR12		
92	47672855	Exon1	G	A	A	A	LRR12		
93	47672859	Exon1	A	G	G	G	LRR12		
94	47672861	Exon1	C	T	T	T	LRR12		
95	47672862	Exon1	G	T	T	T	LRR12		
96	47672863	Exon1	C	T	T	T	LRR12		
97	47672913	Exon1	C	G	G	G	LRR12		
98	47672947	Exon1	-	ATTCACAT CAATTTCC TTAAT	ATTCACAT CAATTTCC TTAAT	ATTCACAT CAATTTCCT TAAT			
99	47672968	Exon1	T	G	G	G	LRR13		
100	47672969	Exon1	C	G	G	G	LRR13		
101	47673027	Exon1	C	T	T	T			
102	47673040	Exon1	C	A	A	A			
103	47673084	Exon1	G	T	T	T	LRR14		
104	47674705	Exon2	T	A	A	A			
105	47674715	Exon2	C	T	T	T		N	

106	47674728	Exon2	G	C	C	C	LRR16		
107	47674730	Exon2	A	T	T	T	LRR16		
108	47674734	Exon2	G	C	C	C	LRR16		
109	47674735	Exon2	G	A	A	A	LRR16		
110	47674741	Exon2	C	A	A	A	LRR16		
111	47674746	Exon2	G	C	C	C			
112	47674797	Exon2	C	C	G	C			
113	47674828	Exon2	C	T	T	T	LRR17		
114	47674830	Exon2	C	T	T	T	LRR17		
115	47674838	Exon2	C	A	A	A			
116	47674842	Exon2	T	A	A	A			
117	47674843	Exon2	G	G	T	G			
118	47674844	Exon2	G	T	T	T			
119	47674917	Exon2	A	T	T	T	LRR18		
120	47674944	Exon2	A	G	G	G			
121	47674945	Exon2	C	T	T	T			
122	47674986	Exon2	G	A	A	A	LRR19		
123	47674990	Exon2	G	C	C	C	LRR19		
124	47674995	Exon2	C	G	G	G	LRR19		
125	47675002	Exon2	T	C	C	C	LRR19		
126	47675003	Exon2	T	C	C	C	LRR19		
127	47675004	Exon2	G	C	C	C	LRR19		
128	47675018	Exon2	A	G	G	G	LRR19		
129	47675050	Exon2	G	C	C	C		N	
130	47675091	Exon2	T	C	C	C	LRR20		
131	47675100	Exon2	A	T	T	T	LRR20		
132	47675106	Exon2	A	G	G	G	LRR20		
133	47675169	Exon2	T	A	A	A	LRR21		
134	47675170	Exon2	C	T	T	T	LRR21		
135	47675172	Exon2	A	G	G	G	LRR21		
136	47675173	Exon2	A	C	C	C	LRR21		
137	47675180	Exon2	G	A	A	A	LRR21	N	
138	47675184	Exon2	G	A	A	A	LRR21		
139	47675196	Exon2	G	C	C	C	LRR21		
140	47675242	Exon2	G	A	A	G			
141	47675249	Exon2	G	A	A	A		N	
142	47675260	Exon2	A	T	T	T			
143	47675292	Exon2	T	C	C	C	LRR22	N	
144	47675295	Exon2	T	C	C	C	LRR22		

145	47675296	Exon2	A	G	G	G	LRR22		
146	47675302	Exon2	C	G	G	G	LRR22		
147	47675310	Exon2	A	C	C	C	LRR22		
148	47675349	Exon2	-	TAAAAG	TAAAAG	TAAAAG			
149	47675358	Exon2	A	G	G	G			
150	47675359	Exon2	T	A	A	A			
151	47675360	Exon2	G	A	A	A			
152	47675361	Exon2	C	G	G	G			
153	47675367	Exon2	G	A	A	A			
154	47675368	Exon2	A	T	T	T			
155	47675374	Exon2	C	A	A	A			
156	47675392	Exon2	C	A	A	A			

† Glyma14g38533 gene sequence from resistant lines Hwangkeum, L29, and RRR were compared with that from a susceptible line Williams82 to identify all these genetic polymorphisms.

‡ SNP positions within nucleotide binding NB-ARC domain and leucine-rich repeat (LRR) domain. Region downstream to NB-ARC domain represents LRR domain, and individual LR-repeats are represented as LRR1, LRR2, and so on.

Supplementary Table E5: Domain structure of Glyma14g38533 protein from SMV-resistant L29 line.

1	**CC domain:**	MGDIVLSIVAKLAEYTVGPILDHARYLCCFNNIAGNLPNAKEELELTRNSVKERVE EAIMRTEIIEPAVEKWLKDVEKVLEEVHMLQGRISEVSKSYFRRQFQYFLTKEIARK IEKMAQLNHNSKFEPFSKIAELPGMKYYSSKDFVRFKSR
2	**NB-ARC domain:** †	ESTYENLLEALKDKSACTIGLVGLGGSGKTTLAKEVGKKAEELKLFEKVVMATVS QTPNITSIQMQIADKLGLKFEEKTEEGRAQRLSERLRTGTTLLILDDVWEKLEFEAI GIPYNENNKGCGVILTTRSREVCISMQCQTIIELNLLAGNEAWDLFKLNANITDESP YALKGVATKIVDECKGLAIAIVTVGSTLKGKTVKEWELALSRLKDSEPLDIPKGLR SPYACLGLSYDNLTNELAKSLFLLCSIFPEDHEIDLEDLFR
3	**NL-Linker:**	FGKGMGLPGTFGTMEKARREMQIAVSILIDCYLLLEASKKERVKMHDMVRDVAL WIASKTGKAILASTGMDPRMLLEDETIKDKRAISLWDLKNGQLLDDDQLNCPS
4	**LR repeats: (n_1)** 6 members	LEILLFHSTEVDFDVSNACFERLKM IKILAILTSSLNWRRRELMKPFGTSYLSLSLPQSMESLQN LHTLCLRGHILGDISILESLQA LEVLDLRNSSFIELPNGIASLKK LKLLDLFNCVIREHNAYEVIGR CLQLNELYLCIYLCAYEEFPHNISLSRLERYVLNFKMYSQSWTDMMEEHRPCRALC INGF
5	**Inter-LRR island (IR):**	NASVQSFISLPIKDFFQKAEYLHLRDLKGGYENVIPSMVPQGMNH
6	**LR repeats: (n_2)** 17 members	LTFLILEDCPEIKCVFDGTTMQTEDAFSS LVILRLYELDNLEEVFNDPSSRCSLKS LEELSIESCRQLYNISFPKNSKLCH LKFLTIDHCPMLTCIFKPSIVQTLEL LEQVTISDCFELKQIIEEVEEGSVDYVSSQSHTSLMLPK LRTLTILRCHSLEYIFPMCYAHGLAS LEELNIGFCDKLKYVFGSEKEHDLRVYQHQSHPQTNIHINFLN LETLRLTELPNLVEIWPKYFDPHLPNLKELQCIDCPRLPDSW VRRVMIIDSDLQQDSTTTEKEL LCSVTTTFNQLSDQVLSSK LRHLQLYGLGVKGLFQFQIREHGSNTE LAPLNLDLIYAELSDLPELEFIWKGPTNFLSLQM LDVIYVNRCPKLKVIFSPTIVRSLPM LRTLEITHCEELEQIFDSGDAQTLYTCSQQVCFPN LHYICVEKCNKLKYLFHNFVAGHFHN LIALEIKDCSQLQKVFAFECETDDDDQEGIVMDGEKVLLRN LLRIRLSRLPNFKEIHHGFKLKDDVEEHIINDCPKYYPSLYLHTEE

† NB-ARC domain comprise of P-loop/ Kinase-1a, RNBS-A, Kinase-2, RNBS-B/Kinase-3, RNBS-C, GLPL, RNBS-D sub-domains [Wan, et al. 2012]. These sub-domain sequences are indicated in red color.

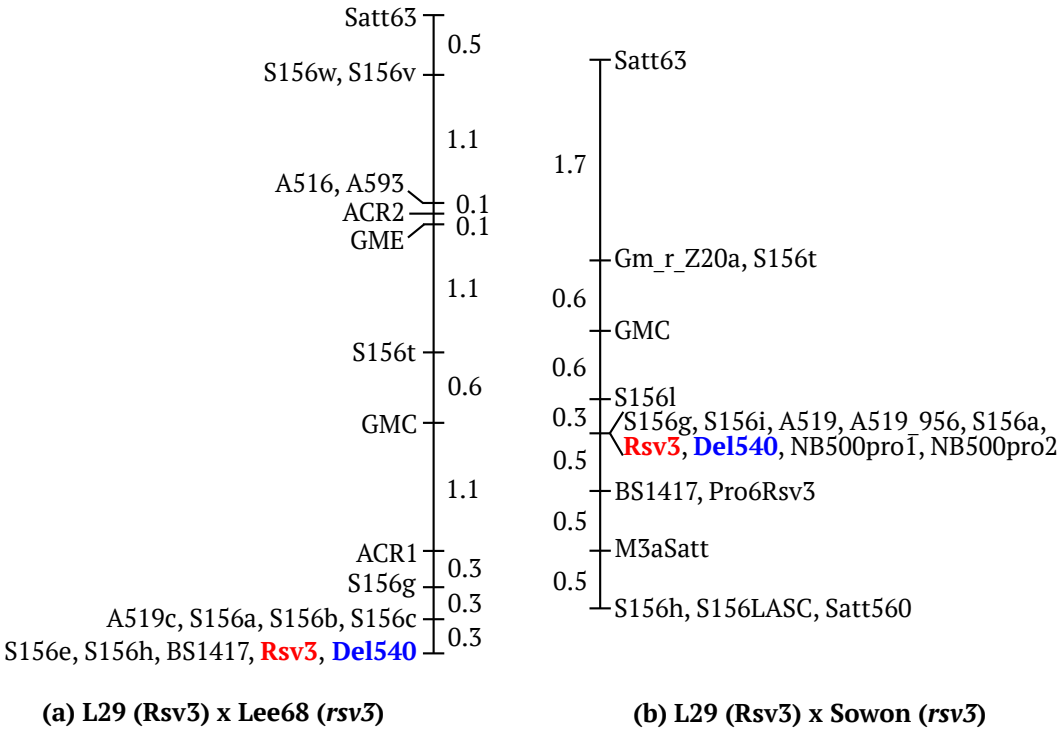
Supplementary Table E6: Differences in LRR motif sequences of (a) LRR7, and (b) LRR11 in Glyma14g38533 gene of Hwangkeum and L29.

(a) LRR7	**L**	**x**	**x**	**L**	**x**	**L**	**x**	**x**	**L**	**x**	**x**	**L**
Wm82:	L	V	I	L	S	L	Y	G	L	D	N	L
L29:	L	V	I	L	R	L	Y	E	L	D	N	L
Hwangkeum:	L	V	I	L	R	L	R	V	L	D	N	L

(b) LRR11	**L**	**x**	**x**	**L**	**x**	**L**	**x**	**x**	**C**	**x**	**x**	**L**
Wm82:	L	R	T	L	T	I	R	G	C	R	S	L
L29:	L	R	T	L	T	I	L	R	C	H	S	L
Hwangkeum:	L	R	T	L	T	I	H	R	C	D	G	L

Appendix F. Supplementary Figures from Chapter 5

Supplementary Figure F1: Genetic mapping of two populations segregating for *Rsv3*-type resistance.



Supplementary Figure F1: Genetic mapping of Del540 marker in two F_2 populations segregating for *Rsv3*-type resistance. The genetic mapping data for other markers in "L29 x Lee68" and "L29 x Sowon" populations was obtained from Seo et al. (2011).