PLEIOTROPIC EFFECTS ON THE TRANSCRIPTOME AND GENOME OF TRANSGENIC SOYBEANS RESULTING FROM TRANSGENE INTEGRATION AND EXPRESSION IN SEED TISSUE

by

Kevin Chad Lambirth

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Biology

Charlotte

2015

Approved by:

_____
Dr. Kenneth Bost

_____
Dr. Kenneth Piller

_____
Dr. Ian Marriott

_____
Dr. Daniel Nelson

_____
Dr. Cynthia Gibas

ABSTRACT

KEVIN CHAD LAMBIRTH. Pleiotropic effects on the transcriptome and genome of transgenic soybeans resulting from transgene integration and expression in seed tissue. (Under the direction of DR. KENNETH L. BOST)

The seeds of the common soybean (*Glycine max*) produce and store large amounts of protein, making them an appealing bioreactor for producing valuable recombinant proteins at high levels. However, the effects of accumulating recombinant protein at high levels on bean physiology are not well understood. To address this, we investigated whether gene expression within transgenic soybean seed tissue is significantly altered when large amounts of recombinant proteins are produced and stored in the seeds. Measurable unscripted gene expression changes were detected in the seed transcriptomes of three transgenic soybean lines chosen for analysis, with one line (764) exhibiting extensive gene expression changes. Further investigations revealed nucleotide polymorphism rates in line 764 nearly double that of the other two transgenic lines and wild type controls. In all three lines examined, the transgene insertions did not disrupt any currently annotated soybean genes. These results suggest that recombinant protein expression and accumulation in seed tissue may impact native gene expression, possibly due to chemical attributes of the particular recombinant protein being expressed or effects resulting from transformation mutagenesis rather than heterologous protein expression levels.

DEDICATION

To all people, past, present, and future, who pursue individual tolerance and seek the unaltered truth.

"*Vi veri veniversum vivus vici*" –
"By the power of truth, I, while living, have conquered the universe."

ACKNOWLEDGMENTS

I would like to extend my gratitude first and foremost to my committee members for dedicating their time, expertise, and support throughout my graduate career. Without your constant guidance and advice, none of this would ever have been possible. Dr. Bost's professional advice always ensured my efforts were practical and reminded me to tell a story, Dr. Piller's "pearls" always shaped high standards for great science, Dr. Nelson always made me feel younger than I am, and Dr. Marriott's interactive instruction was pragmatic and refreshing. I would also like to thank my friends and family for your constant encouragement past and present, and Sara Claytor for being my rock during these years. I would also like to recognize funding from the UNCC GASP program and the USDA Kannapolis Scholars training grant for making this work possible.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Selective farming of agricultural food sources has been conducted by modern humans for millennia with the intention of producing more fruitful crops for consumption. Although Mendelian inheritance and selective breeding was not a scientific practice until the mid-1800's, preference for advantageous phenotypes led to increased allelic frequency of favorable traits such as improved germplasm size and quantity, particularly in crop plants of maize, soybean, rice, cereals, and potato [1]. Backcrossing of parents allows for further control over traits seen in progeny, although this process is time consuming and often unpredictable, particularly in species that are not self-pollinating. Segregation of alleles also does not always produce desired functionality in overall phenotype, and due to individual variability, may come at the cost of overall fitness [2].

Classically, artificial selective breeding pressures for favored endogenous traits were the only avenues for agricultural improvement; however modern genetic engineering practices allow for specific controlled traits to be introduced into the desired species that are non-native. This includes not only nutritional and crop yield enhancements such as herbicide and pest resistance, but also the generation and accumulation of trans-species therapeutic and vaccine proteins for medical and industrial use that would otherwise have been impossible to achieve.

Utilizing plants for molecular farming over conventional cell culture systems offers numerous advantages. Namely, the overall cost of a self-replicating system for pharmaceutical protein generation is significantly reduced when compared to traditional cell culture systems, generating therapeutics for fractions of the current cost of many biologics with minimal waste in the process. Furthermore, the stability of biologics targeted to seed tissues removes the requirement for a cold-chain in production, with demonstrated preservation of both structure and function in excess of 8 years at room temperature [3]. Higher order plants used for production of biologics such as tobacco, *Arabidopsis,* soybean and rice are fully capable of eukaryotic post-translational modifications including tertiary folding, glycosylation, and disulphide bonding. This allows the production and accumulation of fully functional mammalian peptides where secondary and tertiary structures are critical for function, such as antibodies, vaccine candidates, interleukins, and hormones [4, 5].

Glycine max, or the modern cultivated soybean, is a legume branched from the wild species *Glycine soja*, and is classified as a diploidized tetraploid (2n=40). Soybean has a relatively lengthy growth cycle (~6 months) and is primarily cultivated for its seed, which is high in both protein and oil content [4]. It is also primarily a self-pollinating dicot, as the anthers and stigma mature together in the same flower of most cultivars preventing cross-pollination with other plants [6]. This makes soybean desirable for several reasons: One is the simplicity of inheritance selections when breeding, and the other is the high protein content (~40% by weight) of the soybean seed itself, which through millions of years of evolution, has been designed to maintain the durability and stability of its internal cargo until optimum conditions are met for germination. Soluble

protein extraction from soy seed is also straightforward, and has proven efficient for the generation and purification of vaccine antigens, therapeutic proteins, and diagnostic peptides [7-9]. This makes soybean an ideal candidate platform as a bioreactor for the generation, accumulation, and long-term storage of plant-based biologics [10], removing the trade-off dichotomy of either high yield (leaf tissue expression) or long-term stability (seed tissue).

1.1 Transformation Practices for Transgenesis

There are two major methods currently utilized for the introduction of foreign genetic material into plant cells: Direct and indirect transformation. Each transformation system has its own benefits and pitfalls, and is more efficiently suited to industrial applications on a per case basis. Direct cell electroporation, direct microinjection, and particle bombardment are direct methods of transformation, resulting in immediate transient expression of the injected gene of interest (GOI). For soybean, particle bombardment is typically the direct transformation method of choice to preserve the highest amount of viable regenerative tissue, of which much is destroyed through electroporative methods. During particle bombardment, heavy metal nanoparticles of gold or tungsten are coated in the desired transferable DNA sequence and are then expelled at high speed using gaseous pressure from a gene gun, forcing the micro-projectiles through the plant cell wall, membrane, and nuclear envelope. Through this process, the DNA is released from the particle and integrates within the genomic DNA sequences of the host, usually with multiple copies of the transferred DNA inserting into chromosomal regions. While expression of the GOI is generally immediate, multiple integrated copies may induce silencing of the transgene, reducing or eliminating

expression altogether [11]. Furthermore, this method is entirely transient as integration does not occur in germ line tissues, and must be repeated for every subsequent generation of progeny [12].

Indirect transformation methods are most commonly mediated by the soil-dwelling gram-negative bacterium *Agrobacterium tumefaciens.* In its wild type strain, *Agrobacterium* infects many species of plants with unique mechanisms involving inter-kingdom DNA transfer of genetic material to induce crown gall disease. *Agro* contains a Ti plasmid with a *vir* region encoding many factors that function to import the transfer DNA (T-DNA) region. Containing numerous oncogenes, the wild type T-DNA once inside the plant cell hijacks internal plant cellular machinery to induce overproduction of plant growth hormones forming a tumor-like projection of plant tissues. In addition, these bacterial genes also induce the production of large quantities of opines, which are concentrated in tumor tissues and serve as carbon and sugar sources by the bacteria for nutrients. In order to transfer desired gene sequences into cultured plant tissues, these tumor-forming genes are removed, and replaced by the GOI, leaving the transfer mechanism encoded by the Ti plasmid's *vir* factors intact. Advantages of *Agrobacterium*-mediated transformation over particle bombardment are the ability to transform germ line cells, allowing expression in all following generations of explants, and also reducing copy numbers of integrated T-DNA molecules, diminishing the possibility of downstream gene silencing and complex segregation patterns.

The first step to transformation via *Agrobacterium* is the recognition of substances released by wounded plant cells, which are the preferable site of infection for the bacterium. Substances such as lignin precursors, phenols and cell membrane

components activate an internal signaling mechanism cascade initiated by the extracellular membrane-bound receptor VirA, dually acting as chemoattractants for the bacterial cell [13]. Subsequent phosphorylation of VirG by VirA initiates expression of Ti virulence factors, as well as three *Agrobacterium* chromosomal genes that facilitate attachment to the host plant cell through extracellular oligosaccharides (*chvA, chvB, pscA).* The actual initial attachment process isn't entirely known [14], but may involve unidentified adhesins prior to cellulose fiber organization and biofilm formation.

Following cellular attachment, the VirB complex together with the VirD4 protein product form a type 4 secretion system [15], allowing conjugation and transfer of the T-DNA sequence, which is located between two 25bp border repeat regions on the Ti plasmid. Virulence factor D2 acts as an endonuclease, nicking three bases into the right and left T-DNA border sequences and binding to the 5' end, generating a single stranded T-strand for transport through the VirB complex [16]. It has also been proposed that cellular attachment may also be permanently established during the formation of the T-DNA transfer pillus from this complex. Virulence factor E2 (VirE2) coats the entire length of the T-strand to prevent nucleolytic degradation through the transfer process, and is transported along with VirF and VirD2 into the cytoplasm of the host cell.

Nuclear direction and import across the nuclear membrane is facilitated by the bacterial virulence factors imported along with the T-strand. VirD2 contains both an amino and carboxy terminal nuclear localization signal (NLS) in addition to the endonuclease domain [17], and VirE2 also contains two NLSs to direct the newly transferred T-strand to the nucleus [18]. Not only does the VirD2 endonuclease domain allow for cleavage of the T-DNA strand prior to transfer to the plant cell cytoplasm, but

is also thought to interact with plant importin α for nuclear import. VirE2 does not interact directly with plant importins, but is still effectively transported to the nucleus with VirD2 and the T-strand. This is achieved through facilitated import and chromatin targeting using plant transcription factors VIP1 and VIP2, which also assist in the integration of the T-DNA into the host chromosome at nucleosome complexes [19]. Indeed, VIP2 has been demonstrated to modify host histone structures, suggesting a potential direct role in T-DNA genomic integration [20].

Prior to the actual integration of the T-strand, *Agrobacterium* virulence products must be removed from the strand to allow for integration. VirF, previously imported along with the VirD2 T-DNA complex, binds VIP1 in the plant cell nucleus to degrade the VirE2 protein coat through ubiquitination processes [21]. Following removal of the coating proteins, it is understood that the actual integration of the T-strand into the host genome is likely mediated almost entirely by host DNA repair machinery, as recent knockout investigations of *Agro* virulence factors in tobacco show negligible changes in transformation efficiencies (for review, see Lacroix *et al*, 2013 [19]).

Agro favorably targets double stranded break (DSB) sites for T-strand integration, which is expected for an organism evolutionarily designed to be opportunistic [22, 23], but does not seem to show preference for gene rich or sparse chromatic regions [24]. It is also not known whether the T-DNA complex becomes a double stranded molecule prior to integration at a DSB site, or if this occurs simultaneously as part of the endogenous break repair mechanism; however recent works suggest that it is likely the former due to the common presence of filler nucleotides and inverted tandem T-DNA multiplexes at the insertion site [25]. Following conversion to a double-stranded molecule, it is understood

that stable integration requires association with DSB repair components *Ku70* and *Ku80* through the non-homologous end joining (NHEJ) DNA repair pathway. *Ku70* and *Ku80* form heterodimers binding the ends of the double stranded T-DNA, while association with the exonuclease *Mre11* and *Lig4* DNA ligase completes the repair complex. While DSB's may be repaired through homologous recombination events (HR), the predominant pathway for this mechanism in higher order plants overwhelmingly favors NHEJ. Although HR events increase upon defects in chromatin assembly factors [26], deletion of both *Ku70* and *Rad52*, a key enzyme in HR, resulted in total elimination of T-DNA integration [27]. Interestingly, knockout mutants of critical NHEJ pathway genes, notably *Ku80*, did not decrease T-DNA insertion rates or the stability of the integration [28]. This indicates that the precise details of the final step of T-DNA integration is still uncertain; likely, there is a yet unknown pathway responsible for integration in addition to HR, NHEJ, and alternative NHEJ mechanisms. See figure 1.1 for a graphical overview of this process.

In light of the ambiguity surrounding the integration process, investigations into histone modifications revealed VIP1 directly associates with all core plant histone proteins [29]. Acetylation of histone H4 in conjunction with phosphorylation of histones surrounding a DSB create epigenetic "markers" that repair machinery use for recognition of a repair site, indicating possible involvement of host chromatin structures in T-DNA integration efficiencies. Confirmation of these possibilities have proven difficult however, as mutants affecting these structures interfere with host transcriptional regulation processes, confounding definitive results [19, 30].

1.2 T-DNA Vector Design

Transformation vectors have improved over time from initial recombination techniques of Ti plasmids. Previously, it was a cumbersome effort to remove opine synthesis reading frames and oncogenes from the Ti plasmids of wild type *Agrobacterium*. A breakthrough in modern *Agro* biotechnology came from the efforts of Hoekema [31, 32] with the discovery that *Agro vir* genes required for transformation can function without residing on the same plasmid as the desired GOI. Thus, the binary vector system was designed, allowing the GOI to exist on its own plasmid, with the *vir* helper elements on a separate plasmid element within the same *Agrobacterium* cell. This greatly simplified the process of cloning the GOI into *Agrobacterium* strains with engineered plasmid vectors containing specific restriction sites, while at the same time allowing specific removal of harmful oncogenes from the Ti plasmid. A variety of clone-ready vectors are available with a multitude of unique restriction sites to facilitate GOI insertion, as well as an assortment of selectable marker genes for both GOI-positive *Agro* via antibiotic resistance ORI's in the vector backbone, and plant selectable markers typically conferring resistance to herbicides. For a list of many available binary vectors, see Lee and Gelvin, 2008 [33].

Initial binary vector systems placed the GOI near the left border with the selection open reading frame oriented towards the right border. Nucleolytic deletion of the 3' end of the T-strand is common during transformation, while the VirD2 cap protects the 5' end of the T-DNA. Extensive deletions from the 3' end would delete or truncate the GOI, leaving the selectable marker intact. To enhance integration efficiency, the GOI is now placed by the right border, eliminating the possibility of nucleotide degradation before the marker gene is removed.

The flexibility of utilizing binary vectors also allows the choice of specific promoter sequences to maximize expression of transgenes, and also to drive tissue-specific protein expression and accumulation. While constitutive plant promoters such as the cauliflower mosaic virus 35S promoter have proven to yield high amounts of total recombinant protein, this accumulation is distributed among all tissues in the transgenic plant. Much of the biomass expressing the recombinant protein may be low-yield, or complicate downstream purification. In the case of soybean, seed-specific promoters such as glycinin 11S and β-conglycinin 7S (which contribute 65-80% of total seed proteins [9, 34]) target expression only to cotyledon tissues, resulting in higher tissue-specific expression in seeds when compared to systemic constitutive promoters [35, 36]. This allows preserved, stable storage within cotyledon tissues and reduces potential biohazardous waste disposal. Likewise, abrupt termination of transcription following the reading of the GOI and marker gene is desirable in the case of incomplete nicking of the left border repeat sequence by *Agrobacterium*. Otherwise, backbone vector sequences, which have the possibility of integrating in these read-through events, may be transcribed generating undesired vector expression. In these cases, constitutive terminator elements such as the 35S cauliflower mosaic virus terminator are desired and effective.

Codon optimization is also a crucial part of transgene design, particularly when the GOI nucleotide sequence from an alternate organism from the host. Although amino acids are coded for by trinucleotide sequences they are degenerate, meaning that more than one codon may produce an identical amino acid; however different organisms use particular codons with higher frequencies than others. Codon matching the GOI sequence to using the preferred codon bias of the driving promoter gene families is

generally understood to maximize protein yields, as lower frequency codons can reduce the transcription rates of polymerase [36].

Beyond codon optimization and specific tissue accumulation sites, subcellular organelles may also be targeted with the use of specific signal peptide sequences designed onto the N-terminal end of the T-DNA segment during GOI design. Depending on the fragility of the recombinant protein to be expressed in the system, the internal environment of different organelles such as pH, ionic and enzymatic presence, may offer a more advantageous setting for peptide stability. Plastids such as chloroplasts and mitochondrial targeting have been demonstrated to harbor high levels of certain recombinant proteins, as well as internal vacuoles, endoplasmic reticulum, apoplast, or the cytoplasm [37]. The inclusion of native signal peptides may direct the translocation of the protein without the addition of an engineered signal peptide, however this may not be entirely desirable depending on the conditions required for the particular peptide to accumulate. In these cases, the native peptide may be removed, or an additional retention signal added to the 3' end of the translated T-DNA sequence (such as an endoplasmic reticulum KDEL sequence) to prevent secretion or vesicle-mediated transport [38]. This may be customized on a case-by-case basis for the attributes of the protein of interest. For a demonstration of signal peptide control of protein localization in soybean seed, see Hudson, *et al*, 2014 [5].

Translational enhancer element sequences are also frequently incorporated before the open reading frame of the GOI to further enhance protein expression levels. Leader sequences like the tobacco etch virus (TEV) and tobacco mosaic virus (TMV) enhancer elements are derived from RNA-based viruses, and increase recruitment of eukaryotic

translation initiation factors to the 5' end of the GOI through secondary hairpins without the need for a 5' untranslated region cap [39, 40]. Using these strategies, heterologous protein expression levels of biologics in soy seed tissues have approached ~3% total soluble protein [41] with no measurable degradation over time in ambient conditions [3], and homogeneity of protein composition between batches [7].

1.3 Improvements in Agronomy and Biological Farming of Transgenic Soybean

Since the mid-1980's when transformation technologies began to climb towards their zenith, agronomical improvements in crop plants grew substantially. Soybean was a specifically targeted crop for many enhancements, as nearly 50% of global soy production is cultivated in the United States, and contributes a valuable source of oil and dietary protein. Naturally, increasing the overall yield was a priority in such an economically viable crop, and was accomplished by selectively breeding varieties overexpressing the *Arabidopsis* BBX32 B-box gene yielding an overall increase in biomass of ~10% [42]. In addition, drought resistance was enhanced with the overexpression of BiP, an ER-lumen binding protein [43], and increases in total oil content has been achieved through two different avenues [44, 45], greatly increasing profitability. Nutritional content manipulations of linolenic and other fatty acids, tocopherols and tocotrienols (vitamins), and dietary amino acids have been addressed. Resistance to bacterial, nematode, viral, insect and fungal infections have been conferred to soybean varieties, as well as tolerance to non-selective herbicides such as glyphosate and glufosinate [4]. Current and forthcoming enhancements in soy are shown in figure 1.2.

As previously described, soybean has also been a highly successful and cost-effective bioreactor for many pharmaceuticals, including antibody cocktails [46], vaccine candidates [5, 41], and homogeneous therapeutic and diagnostic biologics [7, 9]. As soybean is also an edible food crop, the production of vaccine peptides may be targeted to the gut mucosal lymphoid tissues to stimulate adaptive immunity in either a progressive inflammatory fashion, or a suppressive regulatory response for tolerance. Retention signals such as the aforementioned KDEL allows accumulation and packaging in protein bodies, which ensures maximum protection and stability from low pH and digestive enzymes upon delivery to the gut associated lymphoid tissues (GALT) [47].

1.4 Transcriptomics and Genomics Analysis Tools and Methods

Recent advancements in sequencing technologies have revolutionized the range of detection, quality, and throughput of analytical approaches to both genomic and transcriptomic studies. Sequencing platforms such as Illumina, Solexa, and Roche 454 allows generation of reads from DNA or RNA molecules (which are converted to poly-A cDNA libraries prior to sequencing for stability) in a high throughput manner down to single base resolution. The resulting reads may then be either aligned to a reference genome sequence if desired, or assembled *de novo* with no reference input. Termed RNA-seq for RNA sequencing, this recent technology allows for detailed investigations into exon-exon and exon-intron boundaries, in addition to comprehensive and quantitative assessment of gene expression levels [48]. Read lengths have increased substantially since the technology's inception in 2009, and have currently expanded to over 500bp. Recently, Pacific Biosciences' Single Molecule Real-Time (SMRT) sequencing technologies have allowed read lengths to exceed an average of 10,000 bases,

with the potential to reach in excess of 60 kb in length [49, 50]. Final sequencing files are provided in FASTA files of varying size, which depends on the quantity of bases sequenced.

RNA-seq reads may be stranded or unstranded, in which information regarding the transcript origin from the leading or complementary strand is known or unknown, respectively. Reads may also be paired or unpaired; the former in which fragmented cDNA libraries are sequenced from each end by the specified read length, and the remaining bases between the reads remain unsequenced. Fragment size selection can narrow pools of reads to the anticipated fragment sizes from the cDNA library construction, removing possible incomplete or truncated reads. Freely available online tools for quality control analysis of the raw sequencing data will return multiple parameters, including repeated sequences from adapter contamination, per base sequence quality, and read length distribution. cDNA library preparation is adjusted depending on the experimental design; several for strand-specific protocols are described and compared for consistency in Levin *et al.* 2010 [51].

If *de novo* assembly is not desired and a reference genome is available, sequencing reads are aligned to the appropriate assembled genome reference sequence through a chosen mapping program. Here, the alignment program Bowtie [52] is described due to its flexible application and current development and improvements, which also has a sister program Bowtie2 for more efficient long read alignment processing. After downloading the reference genome from a chosen source (Phytozome or NCBI are popular repositories), Bowtie must index the reference in order to utilize mapping algorithms, many of which are already pre-built and available on the Bowtie

website [http://bowtie-bio.sourceforge.net/index.shmtl]. This indexing step uses the Burrows-Wheeler indexing algorithm to allow for rapid referencing of characters in large FASTA files, reducing computational time and memory footprints. Bowtie also uses a greedy alignment algorithm, in which the read aligned to a particular genome location is not always the highest quality or best match read for that location, but Bowtie may be instructed to continue to search for better alignment at expense of speed. Base mismatch tolerance may also be specified, as extremely stringent cutoffs (zero mismatches allowed) may remove true aligning reads that simply contained a base-call sequencing error, and settings too lenient may allow frivolous matches in more repetitive regions. Caution should be used with allowances of mismatches however, as backtracking loops of continued attempted alignments may occur as a result. In addition, single nucleotide polymorphisms (SNPs) may be identified in this manner, so if this is part of an experimental pipeline that includes SNP discovery, alignment parameters may need to be optimized.

TopHat is an extension of the Bowtie architecture, allowing the processing and correct alignment of spliced reads, gene fusions, and insertions/deletions. TopHat2 is the more current successor to the original TopHat, providing increased splice junction detection and further performance enhancements [53]. TopHat2 addresses two important complications in the alignment of RNA-seq data: 1.) Introns are removed from eukaryotic genes during transcription, and depending on the organism, these intron lengths may vary and can be extremely lengthy, complicating accurate alignments, and 2.) Reads that span a splice junction may extend several bases into the neighboring exon, aligning incorrectly to processed pseudogenes if transcripts are generated from the sequence. TopHat then

breaks the reads into segments and aligns them to the genome. If these segments align at locations far from each other, it infers that the read spans a splice junction and then estimates their locations. This results in much higher sensitivity and accuracy in the produced alignment, which even in complex transcriptomes such as human, correctly aligned 96-98% of total transcripts and generates a .sam file as output.

Following read mapping, differential gene expression analysis may be carried out with a multitude of freely available software packages available through many collaborative and bioinformatics resources. For the purpose of this chapter, two common strategies for differential expression (DE) analysis will be described: 1.) Transcript assembly and transcript counts using the Cufflinks suite and 2.) using Bioconductor's edgeR in conjunction with featureCounts in the "R" statistical programming environment.

Determining expression level is directly proportional to a transcript's relative abundance in the transcriptome, however due to alternative splicing, many genes may have alternate isoforms that complicate accurate counts. Furthermore, longer transcripts will likely produce a higher number of aligned reads compared to shorter transcripts based on size alone, and cDNA libraries will inherently vary in size between samples based on original mRNA template content. In order to accurately calculate the expression level of a gene, these two factors must be accounted for through normalization of the read counts from the total reads produced by the sequencing run itself. Typically, normalization is achieved by assessing the transcripts by the fragments per kilobase of mapped transcript per million mapped reads (FPKM). This incorporates both transcript length and total reads in the library in reported expression levels for genes to allow

different samples to be compared across different RNA-seq runs. While FPKM is the most popular method of normalization, other methods exist and are described in Dillies *et al,* 2013 [54].

The Cufflinks pipeline [55] is a multi-layered protocol that incorporates several integrated tools to rigorously assess transcript levels for splice variants and poorly expressed/low coverage transcripts. It accomplishes this in part by incorporating the Cuffmerge integrated module, which merges each transcriptome assembly from each sample performed by Cufflinks, which has already accounted for isoform splice variants in the transcriptome assembly. Cuffmerge also utilizes reference transcripts provided by the reference genome (which is required for Cufflinks to function) in order to produce a more accurate annotation of the sequence fragments. Since potentially undocumented novel splice variants may be within this assembled pool, Cufflinks also contains an integrated tool called Cuffcompare, which on request can compare the Cufflinks assemblies to the available reference gene annotation file. Accuracy of these predictions depend on a variety of factors, such as coverage at the particular discovered locus and sequencing gaps, but may be verified with traditional molecular cloning techniques.

The actual differential gene expression analysis is conducted by the final module in the Cufflinks suite, aptly named Cuffdiff, tabulates expression values across two or more samples in a group-wise fashion and calculates the statistical significance between each measured change. Using multiple replicate groups is recommended, as Cuffdiff is able to adapt to patterns of read count variation between replicates in a defined sample group, thereby removing the majority of previously described biases to RNA-seq data. While RNA-seq is highly replicable with much less technical variation than other assays

for gene expression, Cufflinks and Cuffdiff's adaptive algorithms can adjust for this beyond standard linear modeling statistics [56]. A full description of all statistics involved in differential expression calls and normalization can be found in Trapnell *et al.* 2010 [57].

Conveniently, the most recent incarnation of Cufflinks contains a visualization module named CummeRbund, which produces a variety of quality control, organizational, and gene expression information provided from the output files of Cuffdiff. CummerRbund runs in the "R" programming environment, is consistently being improved and updated with new features, and contains a variety of plotting tools to generate publication ready figures in a simple package [55].

EdgeR (empirical analysis of differential gene expression in R) is part of the Bioconductor bioinformatics developmental project [58] as a general statistical counting tool, designed to analyze changes between multiple groups with replicated measurements. EdgeR uses an over dispersed Poisson distribution for modeling biological and technical variance between samples, and supports multiple group comparisons specified by the user. Input files containing the total table of transcript counts require only two specified factors: The total number of aligned reads, and specification of the experimental and control comparison groups. Using empirical Bayes methods to measure variability across multiple sample groups, edgeR can assess whether dispersions of gene expression transcript counts are significant using a modified Fisher's exact test tailored for over dispersed distributions [59]. For further information on the application of empirical Bayes applications to RNA-seq data, see Robinson and Smyth, 2007 [60].

Because of the flexibility of the R programming environment, edgeR is easily integrated with other tools such as featureCounts [61], which will assign aligned reads to annotated features such as gene models and can be directly inputted into edgeR. This reduces both the computational time and memory imprint on upstream processing events. A description of this implementation has been described by Chen *et al* [62]. For validation and data integrity of the reports described in this manuscript, both Cuffdiff and edgeR were used together in subsequent differential expression analyses. Yendrek *et al* [63] also directly compared edgeR with DESeq, showing highly comparable DE gene sets, although some genes were unique to one tool or the other. Therefore for statistical stringency, utilizing the overlapping DE gene set corroborated by two programs is an efficient way to limit type II statistical errors.

1.5 Functional Analysis of Differentially Expressed Genes

Gene ontology (GO) is a universal descriptor for gene function classification, and an invaluable tool for analysis of patterns in gene sets to apply biological function and significance. With large datasets generated by current next-generation sequencing technologies, it was imperative to create a universal annotation language to describe three major components of large gene list products, including cellular components, molecular functions, and biological processes. Thus, the GO consortium was formed to annotate and represent how gene function relates to biological function in complex and lengthy condition lists. Many tools are available for conducting these GO enrichment analyses, however for the purpose of this work, a GO tool designed specifically for agricultural datasets, AgriGO, as well as a general GO tool, GOseq, will be described.

AgriGO [64] is an integrated web tool that employs an enhanced and improved design of the original agricultural-based GO tool EasyGO, adding several different enrichment tests and post-hoc statistical tools for elimination of false positive results. The default enrichment test, or single enrichment analysis (SEA) compares functional annotations of two gene set lists: One is the target list, and the other is the background reference. Three statistical tests are available to be applied to the provided input lists, including the hypergeometric test, Fisher's exact test (for target lists expected to share many terms with the background reference), and the Chi square test (for large gene lists with few subjects expected to overlap the reference list). For gene lists that contain integrated expression values, a parametric gene set enrichment analysis (PAGE) will report statistical significance in terms of the *z* value to evaluate GO terms associated with significantly altered expression patterns in addition to GO term enrichment. Therefore, AgriGO is adaptable to a multitude of different experimental designs, and allows for direct export of enriched gene lists to visualization tools such as REViGO [65] and through the integrated online bar and flow chart tools.

AgriGO has many background lists available within the web tool for many species of plants, including *Glycine max, Arabidopsis thaliana, Medicago truncatula, Nicotiana tobacum, Phaseolus vulgaris, Zea mays, Sorghum bicolor,* as well as many other monocots, dicots, and a limited number of vertebrates. Currently annotated probe sets from microarray data is compiled to generate background reference lists, and species lacking a published reference genome may use a list from a related species, or a custom list may be uploaded for use. AgriGO is also consistently updated, and contact

information for the laboratory of Dr. Zheng Su from China Agricultural University is available on the AgriGO website for requests for new background references.

Alternative tools such as Bioconductor's GOseq [66] perform the gene enrichment analysis in much the same way, but employ several selection techniques to prevent selection bias in the GO enrichment tasks by accounting for transcript and gene lengths as a function of the likelihood of being called differentially expressed. These comparisons are then transferred to the gene length reported in the reference background GO term list, accounting for any detected bias in the enrichment calls. Normalization steps of RNA-seq data processing before GO analysis likely removes the majority of these biases, and thus this tool is more suitably applied to information obtained via microarrays; albeit the ability to utilize GOseq offline in the R programming environment is a distinct advantage over AgriGO's online exclusive functionality.

1.6 Reproducibility and Accuracy of RNA-seq

Previously, the overwhelming majority of gene expression analyses were conducted using microarray technology, which while extremely useful, is not suitable for massive high throughput applications or non-targeted transcriptome-wide observations. Additionally, in more extensive experimental designs, base-per-base microarray technology is vastly more expensive to conduct than high-throughput RNA-sequencing technologies. Microarrays, as mentioned previously, exhibit more variance and are not ideal for detection of transcripts with low expression values. Many comparison studies have been conducted and have found RNA-seq to be vastly superior in nearly all aspects with the exception of initial cost [67]. As sequencing costs continue to decrease, we

expect to see a much higher rate of adoption of RNA-seq techniques over traditional microarrays.

Using RNA-seq for transcriptome profiling and DE gene analysis is not without its own set of challenges. Although RNA-seq derived raw reads and quality is highly consistent, this precision of these measurements is also an Achilles heel as subtle changes between may alter final results. A simulation study addressing the effects of transcriptome complexity, nucleotide polymorphisms, alternative splicing, errors in sequencing, normalization methods, reference mapping versus *de novo* assembly, and gene annotation has recently been conducted, deducing that transcriptome complexity and DE profiling methods were the most influential on final results, whereas polymorphisms and sequencing errors were negligible [68].

The Sequencing Quality Control Consortium has recently conducted a thorough comparative analysis of RNA-seq data reproducibility on different platforms (Illumina, ABI SOLiD, Roche 454), as well as comparisons of several data processing pipelines including TopHat2 [53], MAGIC [69], Subread [70], Bitseq [71], and r-make [72]. Differential expression gene calls were compared between each pipeline, as well as differences in reported foldchange and gene splice variant detection. Each pipeline was deemed comparable in nearly all aspects, including novel splice junction detection, although TopHat2 had the largest variance of differential gene expression calls when lower numbers of DE genes were reported. Relative expression values reported for all methods including comparisons to qPCR and microarrays were in agreement, indicating that there is no one method or platform that supersedes all others in performance or accuracy. Some inconsistencies were witnessed between identical samples prepared

through identical pipelines that were sequenced on an identical machine for RNA-seq datasets, indicating the importance of multiple replicates per analysis group. Taken together, these results suggest that RNA-seq is a comparable expression profiling method to qPCR [63] and microarrays that is highly repeatable, such that DE analyses derived from different platforms can acceptably be combined [73].

1.7 Advances in *Glycine max* Genomics and Transcriptomics

A considerable leap forward in soybean characterization came after the publication of the first draft soybean genome in 2010 [74], making *Glycine max* the first legume with a fully sequenced genome. As stated previously, having a reference genome for transcriptome alignments greatly improves the speed, accuracy, and flexibility of sequencing studies, as well as annotation improvements and gene characterization. While not the first to publish on soybean genomic sequencing efforts, this work vastly improved the understanding of the structure and organization of the paleopolyploid legume's genome. Primarily, the first assembled version reported the total genome size as 1,115 Mb, and represented approximately 85% of the anticipated genome sequence. Nearly 5,000 SNPs were documented, and paralogous regions of the chromosomes confirmed two large genome duplication events at 60 MYA and 13 MYA. The genome sequence was also found to be highly repetitive, with 57% of the mapped genomic sequence occurring in repeat-rich regions near identified centromeric areas, which are also transposon rich. In total, 59% of the reported genome consists of repetitive sequences encompassing many transposable elements, of which 42% are long terminal repeat retrotransposons. Alternative splicing rates were found to be slightly higher in soybean than *Arabidopsis,* and genes involving lipid signaling and synthesis were up to

3-fold higher in soy. Fatty acid synthesis was also significantly increased over *Arabidopsis,* suggesting a much more complex transcriptome in soy [74].

Following this achievement, many studies followed building upon the precedent with fully sequenced legume genomes of pigeonpea, chickpea, *Medicago truncatula*, *Phaseolus vulgaris,* and *Lotus japonicas.* RNA-seq has now allowed in depth functional analysis of the transcriptomes of these legumes, including stress responses, nitrogen fixation, miRNA's, and ozone responses in soybean [75, 76]. Tissue specific gene expression studies from soy including leaf tissue, roots, seeds, nodules, meristems, flowers, and pods have allowed the development of the soybean RNA-seq gene atlas of the Williams 82 cultivar. Other genomic and transcriptomic databases and repositories specific for soybean include the Soy Knowledge Base (SoyKB), SoyBase, the Soybean Database (SoyDB), and the Soybean Transcription Factor Knowledge Base (Soy-TFKB). For a list of current databases, servers, and a comprehensive list of available content, see the review by Chan *et al.*, 2012 [77].

Transcriptome studies following seed developmental stages R1-R8 have revealed our T-DNA promoters of choice for 7S and 11S seed storage proteins typically reach peak expression in the R6 and R7 stages of seed maturity. Conglycinin appears to peak before glycinin seed storage products at the 200mg and 500mg seed weight stage respectively [78]. Genes overexpressed when the seed tissue is dry and desiccated are expectedly related to ubiquitination processes and proteases for breaking down products no longer needed for desiccation. Interestingly, several studies have indicated genes related to amino acid and flavonoid synthesis, and translational chaperone products. It is suggested that these mRNAs are not utilized during the desiccation process, but are

stored within the seed for rapid activation of transcriptional machinery upon seed germination [79].

1.8 Pleiotropic Effects and Assessments of Substantial Equivalence

Infection of plants by *Agrobacterium* leads to a radical "reprogramming" of host gene expression. Many direct responses involve infection defense and stress responses to the infection, while at the same time T-DNA *vir* genes attempt to suppress the plant immune response in order to facilitate maximal infection and virulence [14]. In addition, left border read-through (skipping) and subsequent integration of backbone T-DNA vector sequences is relatively common in transformation, and can present barriers to commercialization through regulatory agencies demanding cultivars be free of backbone elements due to mutational possibilities in native genes. Instances of backbone sequences have been measured in up to 81% of transformed events in *Arabidopsis* and tobacco, in many cases integrating the entire backbone vector sequence into the host genome [80, 81]. Characterizations in T-DNA integration in creeping bentgrass also demonstrated this phenomenon, with T-DNA vector sequence being commonly carried over linked to the left border end of the transgene and associated with many instances of "filler" DNA at the integration site [82].

Concerns over pleiotropic and unknown effects within genetically modified crops centered on several key mechanisms that could generate potential alterations as a result of transgenesis. Most obviously, because transgene integration by *Agro* has been demonstrated to be a random process and may be distributed along the entire length of the chromosomes, integrations in gene dense regions carry a higher likelihood of disrupting endogenous gene sequences by integrating within genes themselves.

Furthermore, insertions may disrupt untranslated regions and regulatory elements if located nearby but outside gene regions, or generate loss/gain of function genotypes. Indeed, T-DNA insertions have been shown to affect elements over 8 kb away from the integration site in *Arabidopsis* [83, 84]. Incomplete termination of transgene open reading frames may allow neighboring sequences to fall under transcription initiation from transgene promoters, leading to premature stop codons or truncated mRNAs or dsRNAs that could possibly initiate gene silencing. Metabolite levels have also been modified as a result of transformation, and tissue stress responses to regeneration during tissue culture woundings are seen in the initial transformants and the following generation of progeny [85]. Roundup Ready soybeans were demonstrated to exhibit lower levels of phytoestrogens than unmodified specimens, and other varieties of nutritional enhanced soybean demonstrated higher accumulation of free amino acids and protease inhibitors [86]. Glyphosate resistant soybeans were shown in a recent study by Barbosa *et al.* [87] to contain concentrations of malondialdehyde ~30% higher than non-transgenic varieties, indicating higher instances of lipid peroxidation and oxidative stresses in the transgenic plants even without exposure to the herbicide. The plant growth hormone auxin has been reported to be reduced in glyphosate resistant soybeans, which can produce significant detrimental effects on plant size and yield [88]. A consolidated overview of possible pleiotropic effects is shown in figure 1.3.

A thorough examination of current literature within this field has not shown direct comparisons of tissue-specific gene expression changes in transgenic soybeans expressing high amounts of recombinant protein. Although equivalence studies have been conducted on trait-enhanced varieties, such as glyphosate resistant soybeans, the

enzymes expressed in these systems are several orders of magnitude less than transgenics constructed for the accumulation of biologics. In order to advance understanding of possible pleiotropic effects resulting from high-level expression systems, it was imperative that this examination gap be filled with wide data mining approaches. This thesis describes a series of comprehensive evaluations that are summarized in figure 1.4, including whole transcriptome gene expression analysis, exome mutation rates, transgene integration locations and T-DNA structure. The results of these efforts will further elucidate approaches for efficient vector design and preemptive amelioration of potential barriers for downstream deregulation of biopharmaceutical agriculture.

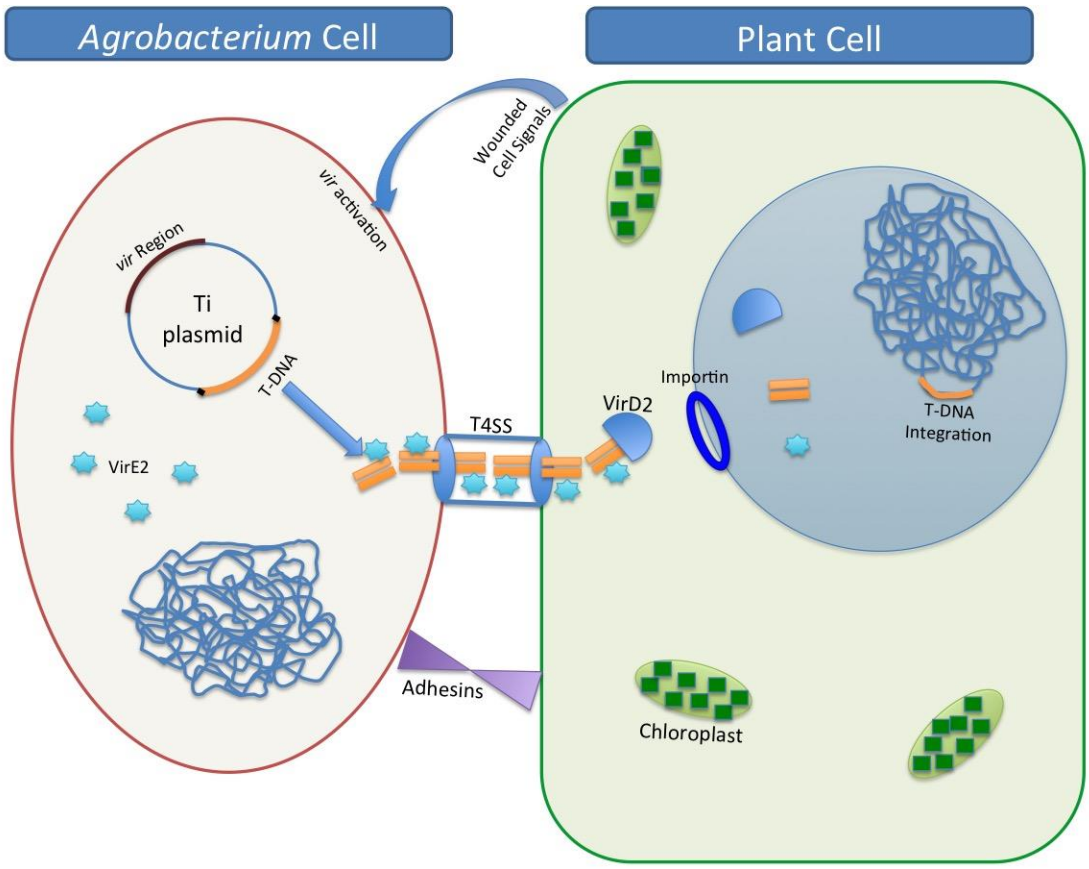


Figure 1.1: Schematic representation of the process of *Agrobacterium* transformation.

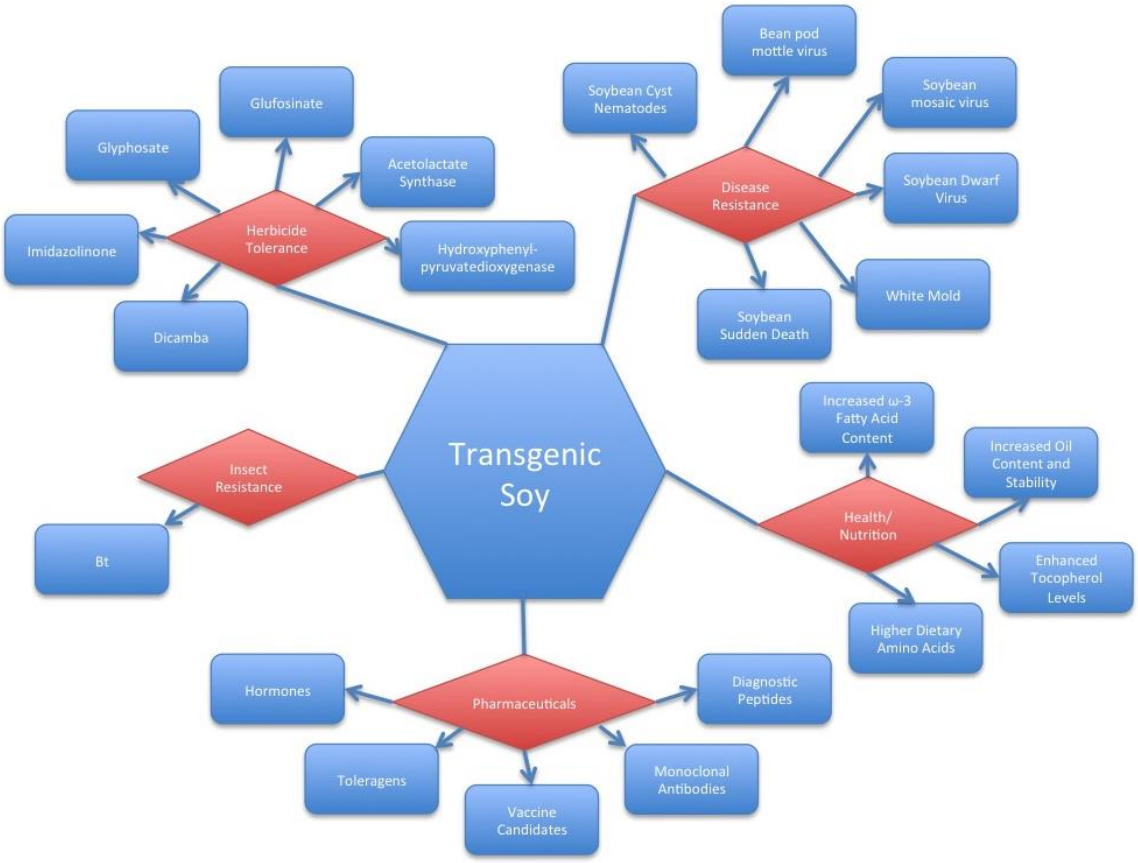


Figure 1.2: Overview of current agronomical improvement efforts in transgenic soybean. Some listed such as Bt and Glyphosate resistant soybean crops are already commercially available, while others are in the developmental pipeline.

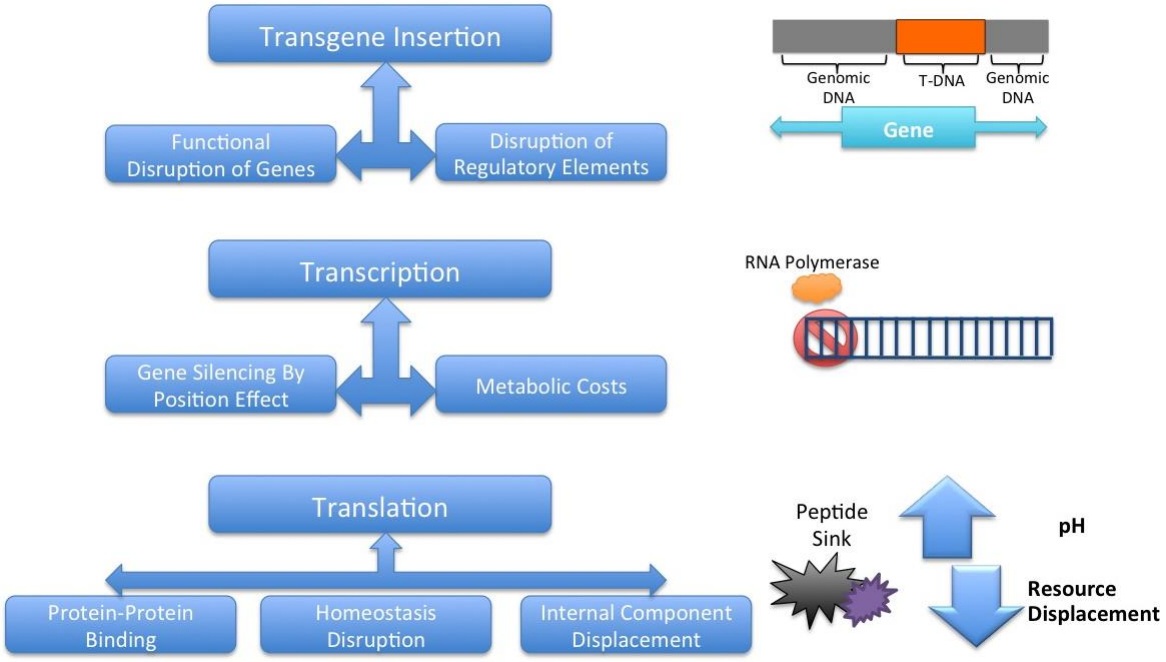


Figure 1.3: Flowchart of the possible mechanisms that could lead to pleiotropic effects in transgenic plants.

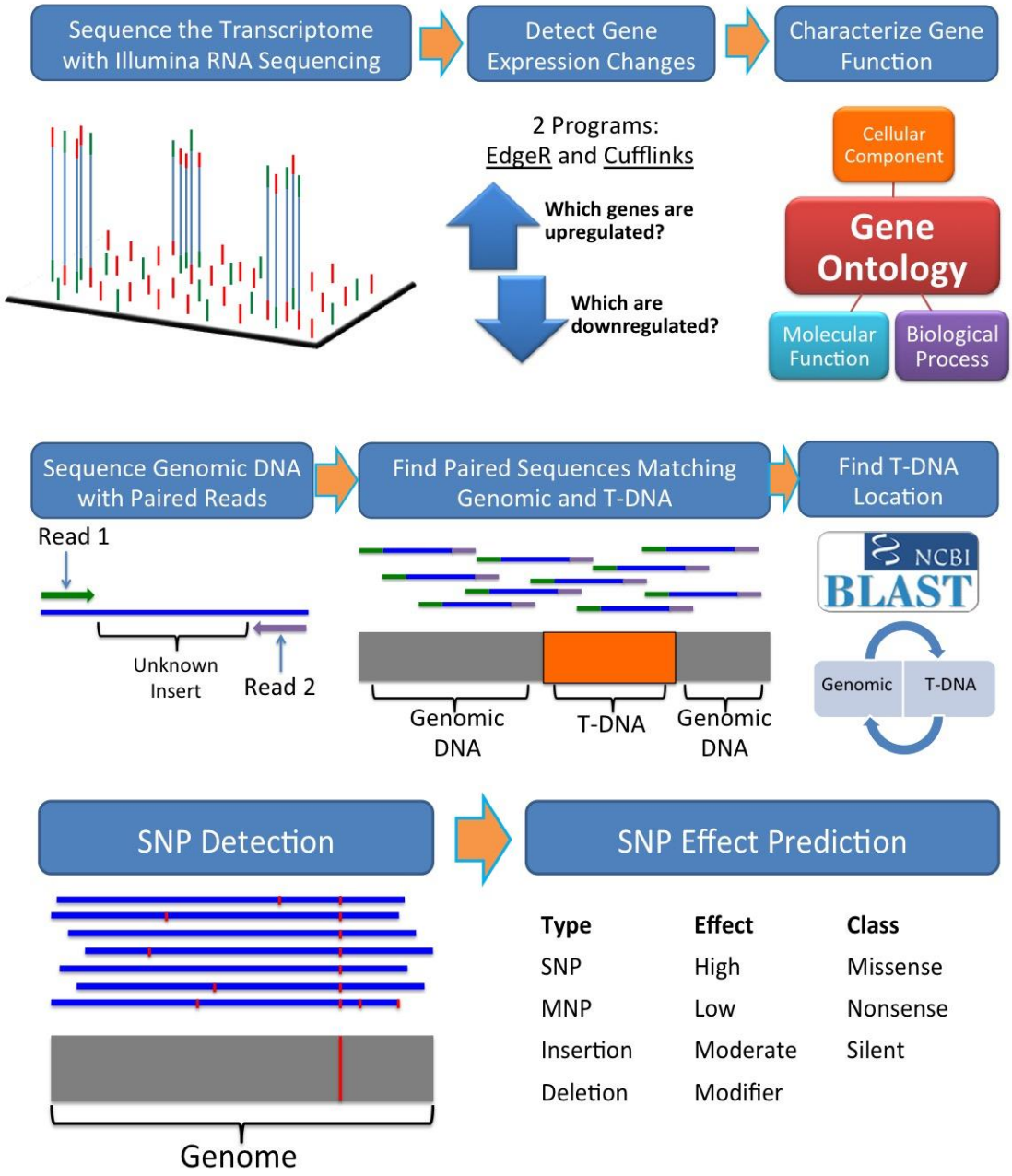


Figure 1.4: Overview of experimental design.

CHAPTER 2: A COMPARISON OF TRANSGENIC AND WILD TYPE SOYBEAN SEEDS: ANALYSIS OF TRANSCRIPTOME PROFILES USING RNA-SEQ

2.1 Introduction

Soybean *(Glycine max)* has been a staple crop and important source of protein worldwide for centuries. The significance of soybean is magnified by the composition of the seed which is naturally rich in protein, oil and linolenic acid [89]. Furthermore, the high protein content of soy (~38% of dry mass) makes this tissue a fitting candidate for targeted expression of recombinant proteins. The first commercial transgenic soybean plants entered the marketplace in 1996 and contained a gene conferring resistance to the herbicide Roundup. Over the past two decades, a variety of transgenes have been introduced into soy to generate soybeans with increased nutritional content as well as resistance to pests and adverse environmental conditions [4].

In recent years, emphasis on biotechnology has directed many efforts to the generation of genetically modified plants, due in part to the increase in their potential for applications in the pharmaceutical industry. With increasing healthcare costs and shortages of medication alternatives, there has been much interest in the development of cost-effective biologics. Proteins have been generated in bulk via bacterially derived methods for years, but limitations in protein size and post-transcriptional modifications have demanded the development and use of other expression systems. Traditional eukaryotic expression systems such as yeast, insect and mammalian cell cultures remedy

many of these issues, but production costs of protein purification and storage usually proves to be expensive [4, 90, 91]. Plant systems have proven to be an economically viable alternative to cell culture systems, despite involving more complex molecular and genetic design phases prior to transformation. Although *Arabidopsis* and tobacco represent heavily utilized model plant systems, they require sizeable quantities of leaf biomass for extracting large quantities of recombinant protein.

Soybeans represent one of the richest natural sources of protein on a per mass basis. Soybean seeds represent a favorable biochemical environment for production of large and complex proteins that are often recalcitrant to expression in traditional systems [9]. Furthermore, transgenic soybeans can be stored as ground powder for years without a need for refrigeration [3, 4, 36]. For these reasons, our laboratory has been interested in developing soybean as a platform for the expression of cost-effective therapeutics [4, 7, 9] that can either be purified or formulated for oral delivery [4, 92]. Although soybean transformation is technically challenging and requires lengthy regeneration times, once transgenic events have been generated and taken to homozygosity they represent a low cost, sustainable solution for production of recombinant protein [10]. Our laboratory has successfully expressed a variety of recombinant proteins in soybean seeds, including subunit vaccines for traditional injection and oral delivery [5, 92, 93], immunogens for treatment of autoimmune disease, and diagnostic reagents for the detection of cancer [7, 9]. The production of these novel soy-based proteins have the potential to address current unmet needs in the healthcare industry and provide novel processing, formulation, and delivery options of therapeutics that are not currently available. Our group and others [35] have reported the expression and accumulation of recombinant proteins in

soybean to levels approaching 3% of total soluble seed protein. These levels equate to >1 mg target protein per seed and represent a significant yield of target protein contained within an environmentally stable package. The production of such large quantities of recombinant protein raises fundamental questions regarding the transcriptional profiles and proteomics in transgenic seeds.

Transgenic plants have been investigated for comparative equivalence to their wild type derivatives prior to deregulation of commercial crops to ensure that the inserted transgene does not negatively impact the quality and nutritional value of seeds and grains [94]. Typical analyses of "substantial equivalence" for transgenic plants stems from the FDA guidelines for inspection, and have traditionally used metabolites, antioxidants, oils, and other molecular compositions as measurements for equivalency [95, 96]. Studies in crop species and other edible plants have determined that compositional variation is typically within the natural range observed through traditional breeding methods [97-101]. While most studies conclude that measured differences are insignificant, some nutritional and metabolic differences have been observed in different transgenic events [95, 102, 103]. Such studies conducted using transgenic soybean have shown only minor fluctuations in metabolites, free amino acids and sugar content, but surprisingly demonstrate that seed protein content remains unchanged [86, 87, 104]. Although acceptable levels of variance have not been clearly defined for specific molecules, significant differences from wild type organisms in the above mentioned studies have not been demonstrated in the examined plants, or shown to have long-term health impacts when used for human consumption [101, 105].

Due to the random nature of the mechanisms associated with plant transformation [14], transgene cassettes could integrate at genomic locations that may positively or negatively impact recombinant protein expression and accumulation [106]. Insertion could also affect the expression of neighboring and downstream genes from the insertion site. Due to the myriad of feedback mechanisms associated with gene expression and regulation, it is possible that disruption of a single exon could alter expression of hundreds or thousands of other genes. Comparative analyses of genetically modified plants has been previously conducted [101], however those studies focused on metabolomics, proteomics and nutritional comparisons. For years genomics and transcriptomics have been recommended as additional evaluation criteria for inclusion in substantial equivalence studies [107]. In this regard, microarrays have been utilized to examine differences between transgenic plants and their wild type equivalents [97] and to detect differentially expressed genes under a variety of environmental conditions. Recent developments in next generation sequencing technologies, in conjunction with the publication of the soybean genome and transcriptome, allows access to more detailed information and refined tools that were not previously available, which in turn can lead to more accurate detection of differentially expressed genes. In this study, we utilized the most recent sequencing technology available on the Illumina platform to conduct whole transcriptome sequencing of seed tissue from three soybean lines developed in our laboratory. These lines express three different recombinant proteins that accumulate to varying levels, with ST77 expressing hTG at 1.61%, 764 expressing mSEB at 0.76%, and ST111 expressing MBP sigma at 0.07% of total soluble protein. The resulting datasets were used for direct transcriptomic comparisons with identically treated wild type seeds.

We found that varying numbers of genes were differentially regulated in all three transgenic soybean lines, with one line having significantly more extensive differences than the others. These results demonstrate the potential for significant transcriptomic variances in transgenic events. To our knowledge, this study represents one of the first to compare the transcriptomes of transgenic soybean seeds with their wild type counterparts using significant statistical power and reproducibility.

2.2 Materials and Methods

2.2.1 Vector Construction and Transformation of Soybean

The binary constructs used to generate the 764 events expressing mSEB protein and ST77 events expressing hTG protein have been previously described by our laboratory [5, 9]. The binary construct used to generate the ST111 events was similar in design to the ST77 binary vector with the exception of the target gene, which encodes a novel fusion protein referred to as hMBP-Sigma. A soybean codon-optimized synthetic gene encoding hMBP-Sigma was synthesized by DNA2.0 (Menlo Park, CA). This gene contained sequences encoding the soybean glycinin signal peptide and full-length myelin basic protein fused to the Reovirus Sigma 1 protein [108]. The hMBP-Sigma fusion protein was engineered with NcoI and XbaI restriction endonuclease sites at the 5' and 3' termini respectively, to facilitate cloning. To generate the ST111 binary vector, the ST77 binary vector was digested with NcoI and XbaI (to release the hTG coding region) and the resulting vector backbone was ligated with the synthesized hMBP-Sigma gene that was also previously digested with NcoI and XbaI. The resulting ST111 binary vector used for soybean transformation contained the 7S β-conglycinin promoter, Tobacco Etch Virus (TEV) translational enhancer, glycinin signal peptide, hMBP-Sigma fusion protein

and the 35s terminator. The ST111 binary vector also contained a selectable marker cassette utilizing the phosphinothricin acetyltransferase (BAR) gene under control of the nopaline synthase (NOS) promoter and terminator sequences. The integrity of ST111 was verified by multiple restriction digest analyses and double-stranded sequencing of the hMBP-Sigma gene (Davis Sequencing, LLC, Davis CA). Transformation of soybean (Williams 82) was performed using the cotyledonary-node *Agrobacterium*-mediated half-seed method previously described [109]. The Williams 82 cultivar of soybean is the same cultivar used for the release of the soybean genome [74]. The declaration of rDNA constructs and propagation of transgenic soybeans was approved by the University of North Carolina at Charlotte Institutional Biosafety Committee.

2.2.2 Soybean Cultivation

Seeds from each transgenic event as well as from wild type were germinated in moistened soil in 6-pack planting trays. Following germination, plants were propagated in Scott's 6-month nutrient Miracle Grow potting mix with 16 hour light (26°C) and 8 hour night cycles (20°C) in controlled growth chambers with ~50% relative humidity. Plants were watered every other day or as needed if the soil was observed to be dry, and were transferred to 4-inch pots (Dillen Greenhouse, 4.00 Square Traditional) at 3 weeks of age and then 1.5 gallon pots (Nursery Supplies Inc. C600) at 6 weeks of age. Light intensities were measured at ~500-550 $\mu E\ m^{-2}sec^{-1}$. Three plants were chosen from each genotype, which were all phenotypically identical to wild type plants with respect to overall size, leaf structure, and approximate seed yield. Dried pods were collected following senescence and fully matured dry seeds at the final R8 stage of development were removed and used for molecular characterization and transcriptome sequencing.

Three seeds from each plant were collected and processed individually, generating three biological replicates from each plant, and three biological replicates from each construct. ST77, 764, and ST111 seeds were obtained from T7, T4, and T3 generation transgenic plants respectively, and were stored in individual seed bags at 23°C and 50% relative humidity until processing. In total, nine seeds were chosen from each transgenic event and from wild type for a total of 36 samples (See Figure 2.1A).

2.2.3 Transgenic Soybean Genomic DNA Extraction and Duplex PCR

Genomic DNA was extracted from seed cotyledon tissue using a Maxwell 16 Instrument and DNA extraction kit (Promega, Madison WI) and cleaned by phenol-chloroform extraction followed by ethanol precipitation. Duplex PCR conditions for ST77 and 764 were described previously [5, 9]. For ST111 duplex PCR, ~1 µg of genomic DNA was mixed with GoTaq Flexi DNA polymerase (Promega, Madison, WI) and buffers provided by the manufacturer with the following primers: hMBP forward (5'-ATGGACCCAAGACTTAGAGAGG-3'), hMBP reverse (5'-CCACATAGACTGTCTGAACCTG-3'), vegetative storage protein (VSP) forward (5'-GCTTCCACACATGGGAGCAG-3'), and VSP reverse (5'-CCACATAGACTGTCTGAACCTG-3'). Following an initial 5-minute denaturation step at 95°C, amplification was performed using 38 cycles of denaturation (95°C for 30 seconds), annealing (50°C for 45 seconds), and extension (72°C for 60 seconds), followed by a final extension step (72°C for 5 minutes). Amplified products were separated and visualized in 1.0% agarose gels.

2.2.4 Transgenic Soybean Seed Protein Extraction and Western Blot Analysis

Seed protein was extracted and quantified as previously described [93]. Briefly, sections of cotyledon tissue from mature seeds were placed in 300 µL of phosphate-buffered saline and sonicated for ~15 seconds. Samples were centrifuged to clarify soluble protein from insoluble debris, and the clarified protein was quantified using a Bradford assay (Bio-Rad, Hercules CA) with bovine serum albumin (BSA) as a standard. Due to the various sizes and inherent properties of each recombinant protein, a variety of different polyacrylamide gel concentrations and buffers were used for the separation of proteins prior to immuno-detection. For analysis of hTG protein, 5 µg of ST77 seed protein extracts were separated in 5% native SDS gels using non-reducing conditions as described previously [9]. For analysis of mSEB protein, 3 µg of 764 seed protein extracts were separated in 10% SDS gels using standard reducing conditions as previously described [5]. For analysis of hMBP-Sigma protein, 20 µg of ST111 total seed protein extract was incubated with non-reducing sample buffer (10 µg of bromophenol blue, 3% SDS, 1.5% glycerol, and 0.025 M Tris-HCl) and separated in 8% SDS-PAGE gels. Following electrophoresis at 100v for ~2 hours, gels were incubated with 1x CAPS buffer (3-[Cyclohexylamino]-1-propanesulfonic acid) in 10% methanol and transferred to nitrocellulose Immobilon P membranes (Millipore, Billerica MA) for 1 hour at 100 v. Membranes containing transferred protein were blocked in 1x PBS containing 5% non-fat milk powder overnight at 4°C, followed by a 3-hour incubation at 23°C with respective primary antibodies. Blots were washed three times for 15 minutes each in 1x PBS/0.1% SDS and incubated with a secondary antibody (HRP-linked goat anti-rabbit IgG) for 1 hour at 23°C. Blots were washed again three times for 15 minutes each in 1x PBS/0.1%

SDS prior to a 5-minute incubation with 10mL of SuperSignal West Pico luminol enhancer solution (Thermo Scientific, Rockford, IL) at 23°C before detection with film.

2.2.5 RNA Extraction

Each of the selected seeds was cut in half along the embryonic axis using an RNase-free razor. To eliminate possible RNA contamination, RNaseOUT (G-Biosciences, St. Louis MO) was used throughout the extraction procedure. Bisected seed halves including the testa, hilum, micropyle, and embryo tissue were flash frozen in liquid nitrogen and ground to a fine powder with a mortar and pestle. Crushed powder was immediately transferred to RNase/DNase-free 1.5 mL spin tubes. Total RNA was extracted and purified using the RNeasy Plant Mini Kit protocol (Qiagen, Germantown MD) for plant cells and filamentous fungi. Buffer RLC was incorporated as recommended by the protocol due to the high concentrations of starch and metabolites in soybean seed tissues. Residual DNA contamination was removed by treating the spin column with 30 units of RNase-free DNase I (Invitrogen, Grand Island NY) for 15 minutes at 23°C prior to RNA elution. RNA concentrations and purity were verified for each sample following elution with a Nanodrop 2000 spectrophotometer (Thermo Scientific, Waltham MA). The 260/280 nm wavelength ratios were ~2.0 for all samples with an RNA concentration ranging from 0.1-1.0 µg/µL. RNA samples were stored at -80°C for up to two weeks until all cDNA libraries were prepared.

2.2.6 Library Construction

cDNA libraries for each sample were generated using the TruSeq RNA Sample Preparation Kit A (Illumina, San Diego CA) according to the recommended low-sample TruSeq RNA Sample Preparation Guide protocol (Illumina, version 2 revision C).

Samples were prepared in four groups, with nine samples per event for a total of 36 libraries. 50 µL of total RNA was loaded into 0.2 mL DNase and RNase-free PCR tubes for use during the purification steps prior to amplification. cDNA was generated through reverse transcriptase PCR using Superscript II reverse transcriptase (Invitrogen, Carlsbad CA). cDNA was bound for purification during the protocol with Agencourt AMPure XP beads (Beckman-Coulter, Pasadena CA). Following ligation of unique adapter sequences, the DNA was enriched by PCR with 15 cycles of amplification according to the TruSeq protocol. Ligation and library integrity was verified using a DNA chip on the Agilent 2100 Bioanalyzer (Agilent, Santa Clara CA) with clean elution profiles at the correct size peak of 261 bp. The Illumina TruSeq kit "A" adapter sequences were ligated to each sample in each group to allow for sequencing multiplexing (see table 2.1). Samples were stored at -20° C for up to 2 weeks until single-end sequencing could be conducted on all samples simultaneously.

2.2.7 Sequencing

Sample libraries ligated with unique adapter sequences were multiplexed six to a lane and were sequenced by the David H. Murdock Research Institute Core lab genomics department (Kannapolis, NC) using Illumina HiSeq 2000 100-cycle, single-end sequencing. Table 2.1 reports ligated adapter sequences and other details of sequencing strategy and multiplexing. Quality control analysis on the resulting fastq sequencing files was performed using FastQC (Babraham Bioinformatics, Cambridgeshire UK). FastQC reports for each sequence file are available in the project repository in the folder named "FastQC".

2.2.8 Sequence Alignment

Sequence reads were aligned onto soybean transcriptome and genome reference sequences using TopHat version 2.0.13 [110] using the maximum intron size (-I) parameter 5000 as recommended for non-mammalian genomes. Gene structure annotations corresponding to the latest annotation release were used to build a transcriptome index and provided to tophat during the alignment step. A copy of the gene annotations was obtained from the Joint Genome Institute (JGI) [111] download site for soybean and is version-controlled in the project repository in the "ExternalDataSets" folder. The reference genome used was version 2.75 [112] supplemented with scaffolds containing transgene sequences. Sequence files are available from the Short Read Archive [113] under accession SRP051659.

2.2.9 Differential Expression Analysis with edgeR

The featureCounts program [61] was used to count the number of reads aligning to annotated soybean genes and the transgenes. The program was invoked three times with different options to enable different treatment of reads with ambiguous genomic mappings. The "sm" (single-map) invocation ran featureCounts with default settings ensuring only single-mapping reads were counted. The "mm" (multi-map) invocation added the option "-M", which counted read alignments for reads with more than one alignment. The "pm" invocation added the option "--primary", which counted just the primary alignments for reads, including reads that mapped multiple times but ignoring alignments not reported as a primary alignment for a read. Files produced by featureCounts, including both outputs and summary reports, are available from the project repository in the "data" subfolder within the "Counts" directory. Since comparing pm and mm files indicated that the results were similar (see the file

CountsComparison.html in the “Counts” folder), only the pm gene counts were used in subsequent differential expression analyses. Expression values in reads per million (RPM) and reads per kilobase transcript per million (RPKM) were calculated for the sm and pm data sets and are also available in the “results” subfolder within the “Counts” directory.

EdgeR [59] version 3.8.5 was used to identify differentially expressed genes. Following the procedures described in the edgeR documentation, read count tables were loaded into R, normalized using the default method for edgeR (trimmed mean of M values, or TMM), and then tested for differential expression using the exactTest method. P values reported by edgeR were used to calculate false discovery rates (FDR) for each gene using the method of Benjamini and Hochberg [114]. Results from differential testing of every gene are available in the results directory of the “DiffExpr” folder in the project git repository. Fold-changes are reported as the log (base 2) of normalized count abundance of the transgenic samples divided by count abundance for the wild type (nontransgenic samples). Samples were clustered by the differentially expressed gene lists using multi-dimensional scaling (MDS) plots (Figure 2.8A-C), and were also grouped according to all detected genes in dendrograms (Figure 2.8D-F).

2.2.10 Differential Expression Analysis Using Cufflinks

Cufflinks version 2.2.1 [55] was used in addition to edgeR as a complementary approach to differential expression analysis. Read mapping was performed as described above and reads were assembled using Cufflinks, including parameters for fragment bias correction and multi-read correction. Scripts used to run Cufflinks are in the project repository in the folder named “src” at the top of the source code tree. The resulting

output was used to create a merged GFF file using cuffmerge, and this merged GFF was used in the differential expression analysis with cuffdiff, again using multi-read and fragment bias correction parameters (see Figure 2.9). Fold-changes are reported as the log (base 2) of normalized read count abundance for the wild type samples divided by the read count abundance of the transgenic samples. Output of Cufflinks in the form of GTF files are available from the Gene Expression Omnibus [115] under accession number [GEO:GSE64620]. Version-controlled data processing and analysis code are available from the project git repository at http://bitbucket.org/lorainelab/soyseq.

2.2.11 Gene Ontology Analysis

As in the differential expression analyses described above, gene ontology (GO) enrichment analysis was conducted twice in parallel for each transgenic line. In both methods, only genes with a FDR 0.01 or smaller were considered for the 764 line. For analysis of ST111 and ST77 lines, a FDR of 0.05 was used. The DE genes list for both GOseq and AgriGO included all DE genes (genes called as DE by edgeR or cuffdiff).

The GOseq package version 1.18.0 [66] was used to identify GO categories with unusually many or unusually few differentially expressed genes in the merged dataset. Categories with unusually many differentially expressed genes represented functions, processes, or cellular components that were affected by the transgene, while categories with unusually few differentially expressed genes represented processes that were resistant to perturbation by the transgene. GOseq was used in order to correct for well-known bias in which differentially expressed genes with larger transcripts are easier to detect. GO annotations for soybean were from the “annotation info” file downloaded from the JGI Web site and version-controlled in the “ExternalDataSets” folder of the

project repository. Code used to run the analysis resides in the folder named "GeneOntologyAnalysis" in the project repository.

Upregulated and downregulated genes from the merged DE edgeR and Cufflinks output were also loaded into the AgriGO web tool [64] to identify enriched GO terms for visualization and to complement the GOseq results. Each list was entered into a single enrichment analysis (SEA) against the current *Glycine max* background reference provided by Phytozome [112] using a Fisher's exact statistical test and Hochberg FDR post-hoc test.

2.3 Results

For this study we chose seed tissue derived from three independent transgenic lines expressing different recombinant proteins at varying levels of accumulation. All three lines were generated using *Agrobacterium*-mediated transformation methods. A summary of the selection process is shown in Figure 2.1A, and the binary vectors used to create these transgenic lines are shown in Figure 2.1B. ST77 is a transgenic line expressing the 330 kDa human thyroglobulin protein (hTG); ST111 is a transgenic line expressing a 75 kDa protein comprising the human myelin basic protein fused in frame to the Reovirus Sigma1 protein (hMBP-Sigma); and 764 is a transgenic line expressing a 28 kDa mutant, nontoxic form of a staphylococcal subunit vaccine protein (mSEB). All three lines are homozygous with a single T-DNA insert at a single genomic locus. While ST111 was originally a complex insertion event containing T-DNA insertions at multiple loci, segregation of loci from multiple generations resulted in the single copy line used for these studies. Southern blot screens were used to characterize complexity of all lines (data not shown). For biological replicates, three plants were chosen from each

transgenic line, and three seeds were selected from each plant (Figure 2.1A). In the same fashion, three seeds from three wild type parents were also chosen for a total of nine individual negative controls.

2.3.1 Molecular analysis and sequencing

Prior to Illumina sequencing, molecular analyses were performed to verify the presence of each respective transgene in each seed as well as expression of the corresponding recombinant protein. To assay for transgene integration, duplex PCR was performed using two sets of primers for simultaneous detection of the transgene and internal control gene (vegetative storage protein). The results of these PCR assays are shown in Figure 2.1C. In all cases, the presence of stably integrated T-DNA in each seed genome was verified.

Western analyses were carried out to demonstrate the stable accumulation of recombinant protein in each of the selected seeds and these results are shown in Figure 2.1D. It should be noted that 3-5 µg of seed protein was sufficient for visualization of hTG and mSEB in lines ST77 and 764, while 20 µg of protein was required for visualization of hMBP-Sigma protein from line ST111. We estimate that recombinant hMBP-Sigma protein accumulates to a level representing 0.07% of total soluble protein (TSP); therefore ST111 was classified as a line expressing a "low" level of recombinant protein. For comparison, ST77 and 764 are classified as lines expressing relatively "high" and "medium" levels of recombinant protein as hTG accumulates to 1.61% TSP [9] while mSEB protein accumulates to 0.76% TSP in the 764 line [5].

Illumina sequencing was performed on libraries prepared from seed cDNA from each set of nine transgenic seeds as well as nine wild type seeds of the same genotype.

Single-end, 100 base sequencing generated between 7 and 12 million reads per library. Reads were aligned to the reference genome and transcriptome and mRNA expression levels for transgenes and native soybean genes were assessed. Normalized expression values per sample for the transgenes are shown in Figure 2.2A. Coverage maps from the highest expressing seed from each transgenic plant are depicted in Figures 2.2B-D. The ST111 line which accumulated the least amount of recombinant protein showed the fewest aligned T-DNA reads, while the ST77 and 764 lines expressed greater levels of recombinant protein and showed a higher number of aligned reads. Note that the transcript levels of ST77 and 764 are similar despite ST77 expressing twice as much recombinant protein by mass as 764. This observation is likely due to the large size of the hTG transgene coupled with fewer aligned reads in the upstream portion of the gene, and can be visualized in the coverage maps (Figure 2.2C-D). Analysis of the coverage data revealed accurate transcript initiation and termination of each transgene. Similarly, accurate initiation and termination of the selectable marker gene transcripts (BAR) was also observed in all cases.

RNA-seq data was analyzed using cuffdiff and edgeR as complementary differential expression analysis methods. Several studies have suggested that combining and comparing outputs from complementary methods such as these can yield more accurate results [116-120]. Using a false discovery rate (FDR) of 0.01, the edgeR-based analysis identified relatively few differentially expressed genes in the ST77 and ST111 lines (52 and 307 respectively), but found ~3,800 total up and downregulated genes in the 764 line. To illustrate differences between the lines, a heat map was constructed using TM4 MeV software [121] showing RPKM expression values for 500 of the most

differentially expressed genes in the 764 versus nontransgenic comparison (Figure 2.3). It should be noted that because ST77 and ST111 only contained 52 and 307 significant differentially expressed genes, transcripts displayed in Figure 2.3 beyond these for ST77 and ST111 are sorted by decreasing average detected logfold change for ease of comparison between the three lines. Expression levels of the top differentially expressed genes in the 764 line were different from wild type, ST77, and ST111 gene expression. Expression differences were consistent within all groups with the exception of one outlier in the ST77 group (ST77 F1). It should be noted that the archived seed of sample ST77 F1 showed visible fungal growth two weeks after sequencing; this growth was not visible during the selection process, however, it may be one explanation for the observed differences in expression. The inclusion of ST77 F1 in the analysis did not alter the conclusion that ST77 was the most similar to wild type.

Cufflinks software was used in addition to edgeR to investigate differential expression and revealed similar differences in gene expression. Cufflinks reported 47 upregulated and 28 downregulated genes in ST77, 744 upregulated and 361 downregulated genes in ST111 and 1249 upregulated and 843 downregulated genes in 764. Volcano plots were constructed from the results and are shown in Figure 2.4. These plots show the relationship between fold change and statistical significance of differentially expressed genes. Note that there is >20-fold difference in the number of differentially expressed (DE) genes between ST77 and 764. Thus, it is clear from both the edgeR and Cufflinks results that while there were significant differences in all three transgenic events, differences were the most substantial in line 764 relative to the wild type controls. The results of these two programs are illustrated in Figure 2.5. The Venn

diagrams (Figure 2.5A-C) indicate the number of up and downregulated genes identified by each program separately and together, while the bar chart (Figure 2.5D) shows the total number of upregulated and downregulated genes as well as the portion of shared genes identified from each program. Five genes were differentially expressed in all three transgenic lines, including Glyma.12G136600 (protein kinase), Glyma.13G171200 (ribosomal RNA protein-7 related), Glyma.01G103100 (branched chain alpha-keto acid decarboxylase E1 beta subunit), and two genes with no functional annotation information (Glyma.07G207000, Glyma.13G011800). Glyma.01G103100 and Glyma.13G171200 showed no commonality between the three events in the direction of altered expression; however Glyma.01G103100, Glyma.07G207000, and Glyma 13.G011800 were upregulated in all three transgenics. Figure 2.5E shows the number of common DE genes shared between each of the three lines based on the edgeR results. A list of all shared differentially expressed genes between all events is available in the git repository file "Diffexpoverlap" under the "DiffExp" directory.

Numbers of DE genes are a function of statistically significant gene calls within groups, but do not illustrate between sample variance. Clustering algorithms integrated in cummeRbund allowed visualization of individual sample similarity and variance in comparison to wild type by generating dendrograms with the "csdendro" command. Dendrograms allow visualization of between sample variance, reflected by their clade distance from others. Based on the differentially expressed gene sets, the ST77 and ST111 samples clustered randomly intermixing with wild type samples, while the 764 samples clustered independently of wild type (Figure 2.6). ST77 and ST111 samples clustered across both their respective biological groups and the wild type group showing

variances were not substantial enough to completely segregate, while all 764 samples were in a distinct clade from wild type.

2.3.2 Gene ontology results for 764 using GOseq

We next performed a gene ontology (GO) enrichment analysis using GOseq [66] which accounts for selection biases in RNA-Seq data in which larger, more highly expressed transcripts are preferentially detected as differentially expressed. In line 764 we detected at least one read sequence from ~42,000 of the 56,000 annotated soybean genes, and of these ~42,000 expressed genes, approximately 3,800 (9%) were differentially expressed. The input list consisted of approximately 1500 genes that were considered differentially expressed after combining the lists from both edgeR and Cufflinks. Thus, on average, we expected that approximately 3.5% of genes in any random sample of expressed genes would be differentially expressed. However, there were several GO categories that exceeded this 3.5% threshold and are grouped according to their parent terms in Figure 2.7. A more detailed flowchart of all GO terms can be found in Figure 2.10. Of 16 genes annotated to the term “nuclear pore”, nine were differentially expressed, and all were downregulated. Of the 490 genes annotated to the term “structural constituent of ribosome”, 47 were differentially expressed, and of these 94% were upregulated. All DE genes annotated as protease inhibitors were upregulated, including 8 of 19 genes encoding serine-type endopeptidase inhibitors, and 10 of 60 genes encoding peptidase inhibitor and regulator activity. Intracellular transport also appeared affected in the 764 samples, as 8% annotated to non-membrane intracellular organelles were differentially expressed and most (82%) were upregulated. All 10 DE genes encoding mitochondrial function were also upregulated. In addition, several genes

(5 of 8) annotated with the biological process term “response to wounding” were upregulated. Taken together, these results suggested that protein synthesis was more active in the 764 seeds as compared to the nontransgenic controls. These results also suggest that aspects of intracellular transport and nuclear pore structures may be altered. The annotation of upregulated genes involved in wounding responses and peptidase inhibitors suggests that some aspects of a physical stress response may have been activated.

ST111 enriched GO terms were not as extensive as those found in the 764 line. However, it is of notable mention that 5 out of 29 genes (17.2%) involving photosystem 1 were differentially expressed, all of them being downregulated. In addition, 4 out of 13 genes (31%) were downregulated involving the photosystem 1 reaction center. Four out of 7 (43%) detected phosphorylation genes were also differentially expressed and all were downregulated. In this case, it seems the ST111 line is exhibiting a reduction in metabolism and photosynthetic processes. The ST77 line being the most similar to wild type revealed no significant GO terms.

2.3.3 Gene ontology results for 764 using AgriGO

Lists of all significantly differentially expressed genes as described above were exported to AgriGO for comparison with the *Glycine max* V2.1 GO background gene enrichment reference. Following analysis with AgriGO, the ST77 group again failed to show any highly significant GO term enrichment. ST111 samples show significant enrichment of photosynthesis and nucleic acid binding GO terms as reported by GOseq, while the 764 group did not. Likewise, the 764 group showed enriched terms indicating intracellular protein transport and translational terms, which were absent in the ST111

GO analysis. Overall, the results from GOseq and AgriGO were comparable with minor parent GO term variations. Complete AgriGO flow charts summarizing GO enrichment for the 764 line are shown in Figure 2.10.

2.4 Discussion

In this study, we addressed the possibility of detecting differentially expressed genes resulting from T-DNA insertion in three different transgenic soybean lines. Each line expressed and accumulated varying levels of recombinant protein targeted to seed tissue. Our experimental design allowed the testing of multiple factors that could potentially contribute to gene expression differences, including different recombinant proteins, progeny generation, and recombinant protein expression level. The inclusion of both edgeR and Cufflinks allowed us to detect differentially expressed genes with high stringency while limiting false positives and characterize them using gene ontology enrichment analyses.

Contrary to our expectation that the transgenic line with the highest transgene or protein expression level would show the most drastic changes when compared to wild type, we found that the 764 line with moderate protein expression was the most different compared to wild type. Examination of transcript coverage across the T-DNA constructs shows an abrupt end in transcription before the end of the included terminator sequences, demonstrating tight transcription regulation and absence of non-terminated transcripts which have been reported from other transgenic soybean constructs utilizing the NOS terminator element [122]. In addition, line 764 lacks the tobacco etch virus enhancer element present in the other two lines, eliminating the possibility of downstream gene transcription effects [83]. This suggests that transcriptome alterations may not be

fundamentally based on protein expression levels or insert complexity, but instead could be due to the attributes of the specific recombinant protein being expressed, mutations/disruptions from the insertion of T-DNA, or some combination of both. The hTG, mSEB, and hMBP-Sigma recombinant proteins all have very different physical characteristics, including size, charge, amino acid content and tertiary structure, therefore it is possible that accumulation of each recombinant protein could induce different response mechanisms within the seed. While 764 was not the highest expressing of the three transgenic lines, there may be characteristics of the mSEB protein that contributed to the observed transcriptome effects based on internal tolerance of the seed to this specific recombinant protein. Furthermore, while soybean has a relatively low mutation rate, mutations are commonly seen in plant tissue culture through transplantations and regeneration of tissues [123]. Alterations related to this are likely limited in their effects due to the generational distance of these lines from the initial transformation and the self-crossing nature of soybean limiting allelic variations. However, point mutations are still a possible occurrence that could potentially effect gene expression. Since this study focused on one specific mSEB event, it is unknown whether similar differential expression would be detected in other independent events transformed with the same 764 binary vector, or in 764 seed tissue derived from previous or subsequent generations. Indeed, gene expression responses to physical wounding from tissue culture procedures have been shown to carry over into the first generation of transformants [124], however the 764 line described here was harvested from fourth generation transgenic plants, limiting the potential for this kind of effect to contribute significantly to the extent of

gene expression differences measured. Nonetheless, the possibility of random mutations occurring during propagation cannot be concluded to have no measureable effects.

In our datasets, we observed differential expression of helicase genes, suggesting the potential for DNA-level regulatory processes such as methylation, as well as down-regulation of genes for ribosomal subunits and translational processes in ST111 and 764. Furthermore, genes involved with transcriptional regulation and DNA/RNA binding are also differentially expressed consistent with potential gene silencing processes. Mapping the location of the transgene insert within the nuclear genome will reveal whether T-DNA integration has occurred in a transcriptionally active versus repressed region of the genome, and identify those genes (if any) that may have been disrupted as a result of the insertion, as past characterizations of T-DNA integrations in *Arabidopsis* demonstrated the capability of *Agrobacterium* to induce large deletions in genomic sequences [125, 126]. Information regarding neighboring genes in close proximity to the insert will allow exploration of methylation patterns, euchromatin and heterochromatin content of the integration site. The ST111 line will be of particular interest due to the relatively low expression and nearly absent transcript levels along the transgene open reading frame, suggesting the possibility of transcriptional level gene silencing which can occur in some events through methylation in the promoter region [127].

Post-translational regulation can also be a concern if it impacts recombinant protein turnover since decreased levels of accumulated protein could significantly impact downstream cost margins (e.g. of isolated therapeutics). The 764 line characterized in this study exhibited upregulation of serine-type endopeptidase inhibitors, which have been linked to delaying or reprogramming apoptotic processes [128-130]. Endopeptidase

genes involving serine proteases in soybean seeds are typically upregulated as a response to tissue wounding or plant pathogen infections [130]. Serine proteases are also involved in proteolysis of the soybean β-conglycinin seed storage protein [131] in response to an increased demand for amino acids during translation [34]. The upregulation of genes involving translation and endopeptidase inhibitor activity in the 764 events suggests some induced response to programmed cell death (PCD) unrelated to a pathogenic infection. If recombinant protein accumulation activated endopeptidases as a result of PCD signals, then it is possible that the recombinant protein may become nicked, resulting in fragmented or degraded (e.g. undetectable) protein. We have previously noted endogenous nicking of recombinant mSEB protein in the 764 line [5], and other groups have also noted severe fragmentation of recombinant human growth hormone expressed in soy [132]. It should be noted that the seeds harvested and utilized in this study were fully matured and dried seeds in the R8 stage of maturation. Many genes expressed at this stage have been identified as proteases, ubiquitin and proteasome elements [79]. The products of these genes likely function in the elimination of proteins that are not necessary for seed germination processes. Likewise, mRNA transcripts relating to ribosomal machinery and transcription are upregulated in late seed development for immediate use during seed germination [78, 133]. It is possible that some of the differences in gene expression in line 764 are a result of delayed cessation of protein synthesis due to recombinant protein expression. If expression of the transgenes under control of the 7S or 11S promoters is extended, it may be possible to see the appearance of increased peptidase inhibitors as remnants of the cleanup phase following seed quiescence.

Gene ontology analysis allowed visualization of functional patterns of differentially expressed genes identified by both edgeR and Cufflinks. No significant GO terms were identified from the list of ST77 differentially expressed genes, and only a few genes involved in photosynthesis and thylakoid functions were downregulated in ST111. However, the significant enrichment of GO terms related to translation and intercellular protein packaging and transport in 764 seeds shows a clear pattern in differentially expressed genes. Although the heterologous gene of interest is not present in the reference genome, the expression machinery utilized to synthesize and transport the seed-targeted protein is quantifiable, and is therefore detectable in our differential expression analyses as well as our GO enrichment analyses. Although it is unclear whether such genes are differentially expressed due to the expression of recombinant protein, but this may be one explanation for many of the GO terms observed involving intracellular transport and ribosomal constituents. It is also possible that small transcriptomic disruptions could have activated downstream cascades involved with gene regulation in a signaling type response to the initial disturbance via T-DNA insertion and position effects. Regardless, the upregulation of peptidase inhibitors is of potential concern if such a triggered response resulted from internal apoptotic signaling activated by the presence of high amounts of recombinant protein.

2.5 Conclusions

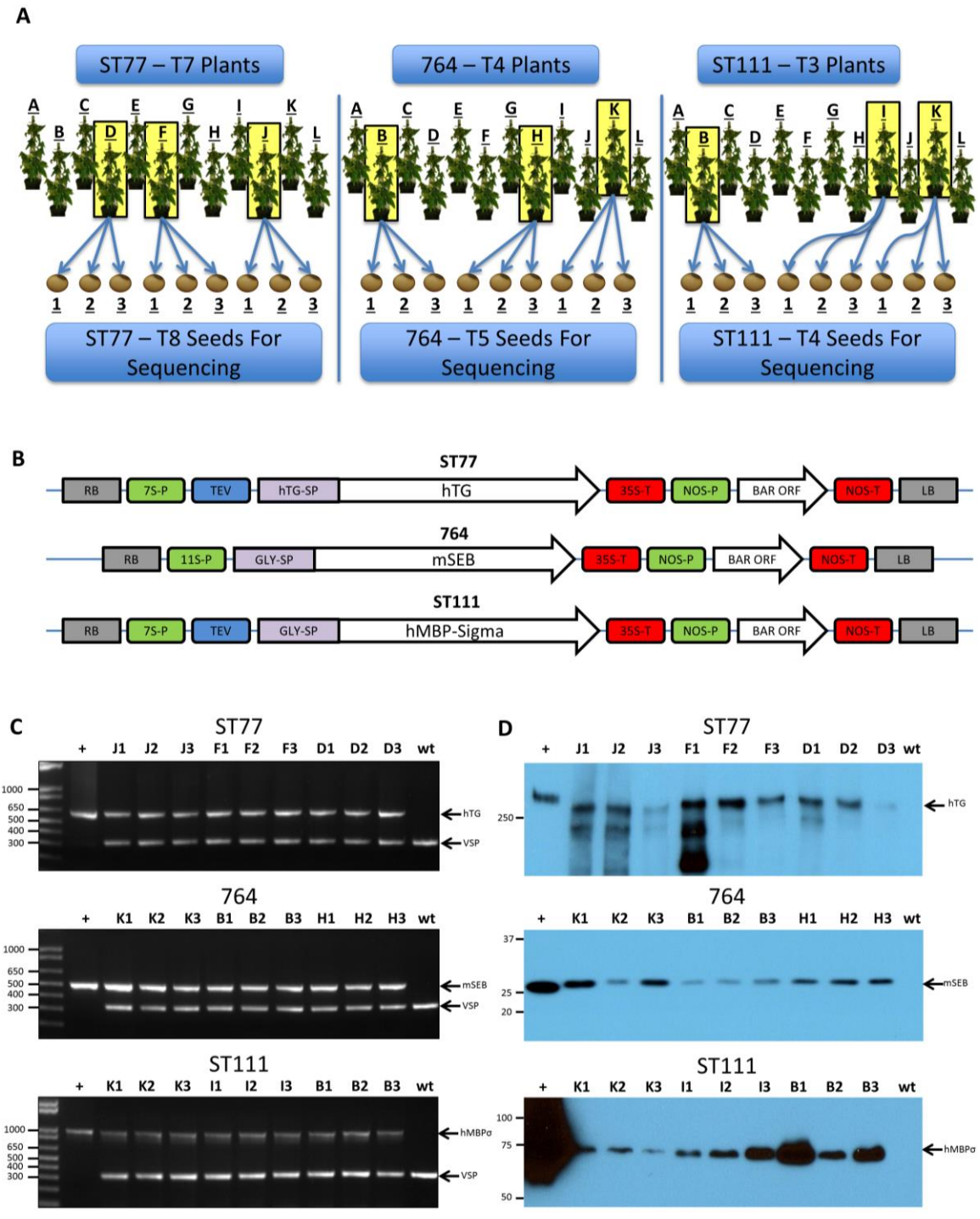
The present study is a comparative analysis of differential gene expression in transgenic soybean seed tissue addressing multiple factors that could potentially induce changes in endogenous gene expression (e.g. transgene expression, protein accumulation, etc.). This study compared three separate transgenic lines expressing different

recombinant proteins at varying levels, and found that all three lines exhibited differences in gene expression with one line (764) being substantially different. In this one line, nearly 10% of the transcriptome was differentially expressed relative to wild type controls. Genes involving responses to wounding, translation, ribosomal constituents, endopeptidase inhibitors, cellular biosynthesis and gene expression were all upregulated while genes involving the nuclear envelope were downregulated. The results from this study suggest that the transcriptomic profiles of transgenic plants can be significantly different than those of wild type controls, and has provided a comprehensive first investigation into gene expression differences resulting from high levels of transgene expression and recombinant protein generation targeted to soybean seed tissues. It is not clear whether altered transcriptome profiles impact other variables traditionally characterized for the determination of substantial equivalence, though current literature suggests nutritional and metabolomic attributes remain comparable to non-transgenic plants. Based on the limited amount of differentially expressed genes shared between all three events, there doesn't seem to be a consistent functional pattern induced based on transformation or recombinant protein expression, indicating each transformation event may respond differently to the inserted T-DNA or the resulting recombinant protein. As high throughput sequencing technologies advance and associated costs decrease, the selection of favorable transgenic lines based on transcriptome profiles could reveal valuable information beyond Mendelian breeding techniques and other methods currently used to characterize transgenic events. The approach proposed here can be utilized to investigate potential detrimental changes resulting from transgene integration and

recombinant protein expression to maximize downstream recombinant protein yields in transgenic plants.

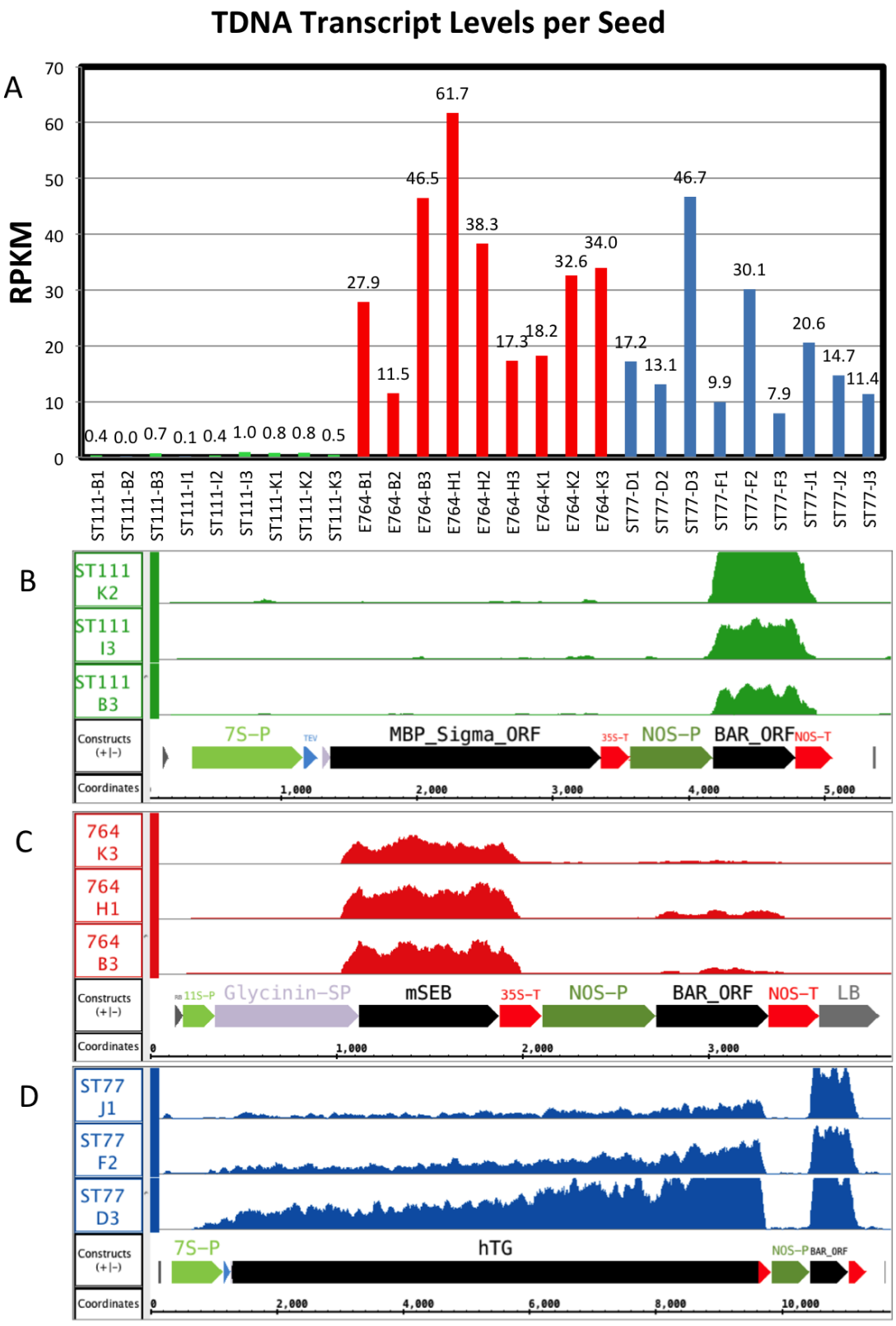
2.6 Availability of Supporting Data

Sequence files supporting the results of this article are available from the NCBI Sequence Read Archive [113] under accession number [SRP051659]. Output of Cufflinks in the form of GTF files are available from the Gene Expression Omnibus [115] under the accession number [GSE64620] available at [http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64620]. Version-controlled data processing and analysis code are available from the project git repository [http://bitbucket.org/lorainelab/soyseq]. Analysis code available in the repository includes shell scripts used to run data processing programs and R Markdown files used to perform statistical analysis. R Markdown output files (with file extension "HTML") documenting the details of analysis are available and can be opened and examined using a web browser. R markdown output files contain version numbers of all R libraries used. Additional instructions for viewing analysis results and data are available on-line at the project repository web site. Alignments from TopHat, coverage graphs, and assembled reads (from Cufflinks) are available for visualization in Integrated Genome Browser [134] from the IGBQuickLoad site [http://igbquickload.org/soy].



(see the following page for figure legend)

Figure 2.1: Experimental design and gene constructs. (A) The selection and propagation process of the plants and seeds used in this study. (B) The binary vectors used for *Agrobacterium*-mediated transformation are shown. The regulatory elements include: 7S-P (7S soybean β-conglycinin promoter), TEV (tobacco etch virus translational enhancer element), hTG (human thyroglobulin gene), hTG-SP (hTG signal peptide), 35S-T (35S cauliflower mosaic virus terminator element), Gly-SP (soybean glycinin signal peptide), hMBP-Sigma (human myelin basic protein fused to Reovirus Sigma 1 protein), 11S-P (soybean 11S glycinin promoter), mSEB (mutant nontoxic staphylococcal enterotoxin B gene), NOS-P (nopaline synthase promoter), BAR (phosphinothricin acetyltransferase gene) and NOS-T (nopaline synthase terminator element). Arrows indicate orientation of cassettes relative to the right border (RB) and left border (LB) sequences. Regulatory elements and genes are not drawn to scale. Molecular characterization of transgenic events. (C) Duplex PCR of the nine progeny seeds from the indicated transformation events. wt: nontransgenic (negative control); +: plasmid DNA (positive control). Arrows indicate amplified DNA fragments derived from the specific gene of interest as well as vegetative storage protein gene (VSP) following separation in agarose gels. Sizes of molecular weight markers are shown in base pairs. (D) Western blots of total seed protein derived from the transgenic progenies shown in (C). Arrows indicate the hTG, mSEB and hMBP-Sigma immunoreactive proteins. Sizes of molecular weight standards are shown as kDa. Positive controls (+) are purified hTG (Cal Biochem), *E. coli*-derived mSEB, and soy-derived hMBP-Sigma from a higher expressing line.



(see the following page for figure legend)

Figure 2.2: T-DNA transcript levels and coverage in transgenic seeds. (A) The number of reads aligned to the T-DNA sequence by TopHat from the reference genome comprising an extra scaffold containing the respective gene of interest cassette. Numbers on the y-axis represent RPKM normalized expression values for the gene of interest in each sample. (B-D) Coverage graphs over the annotated region of each added scaffold are shown for lines ST111 (B), 764 (C), and ST77 (D). The annotated components correspond to those shown in Figure 2.1. The highest expressing seed is shown for each transgenic plant. The y-axis of each coverage graph ranges from 0 to 100.

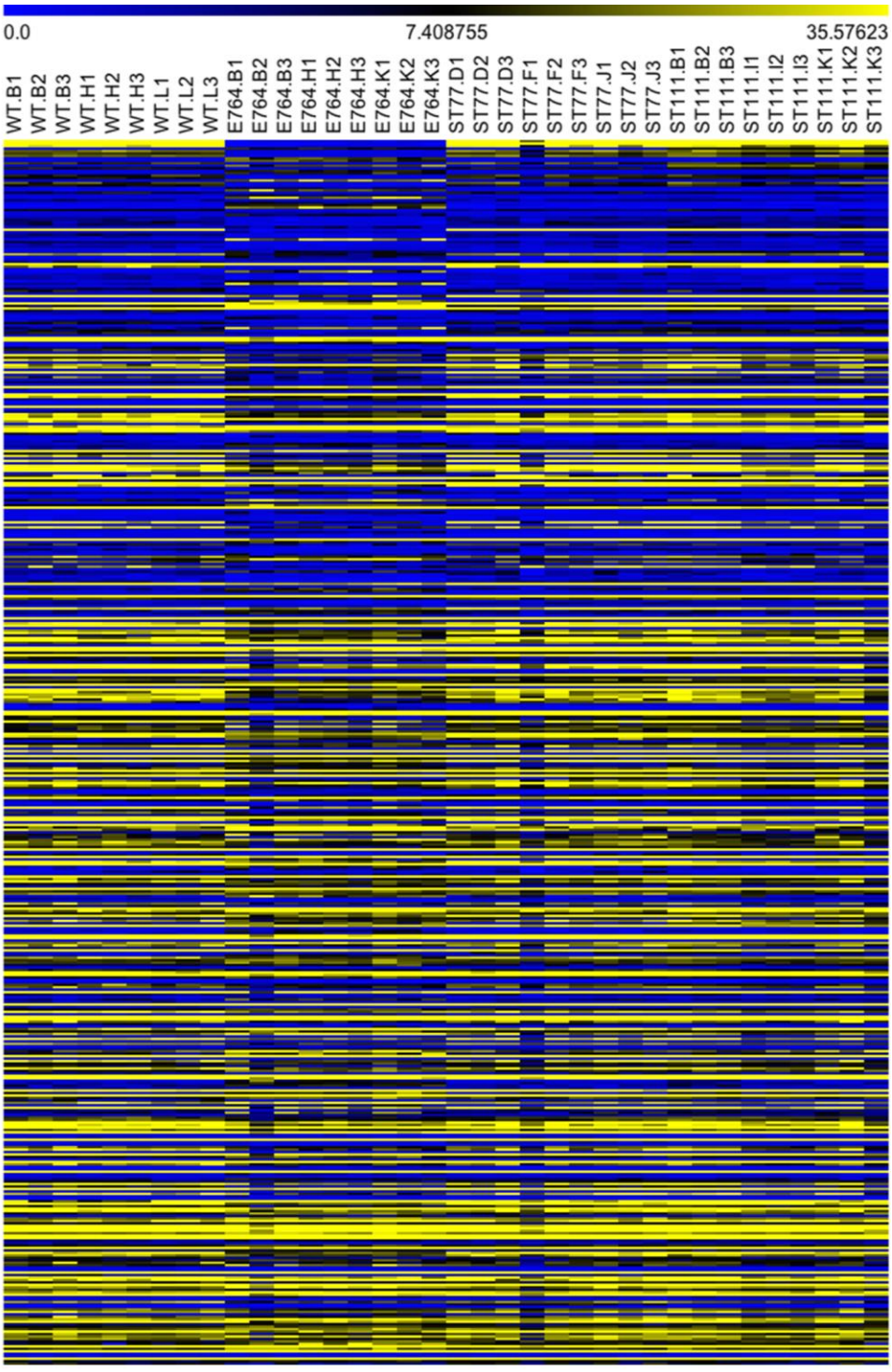


Figure 2.3: Heatmap generated from the top 500 differentially expressed genes as reported by edgeR. Genes are sorted in descending order according to fold-change. The yellow color indicates higher levels of gene expression while blue indicates lower expression by RPKM.

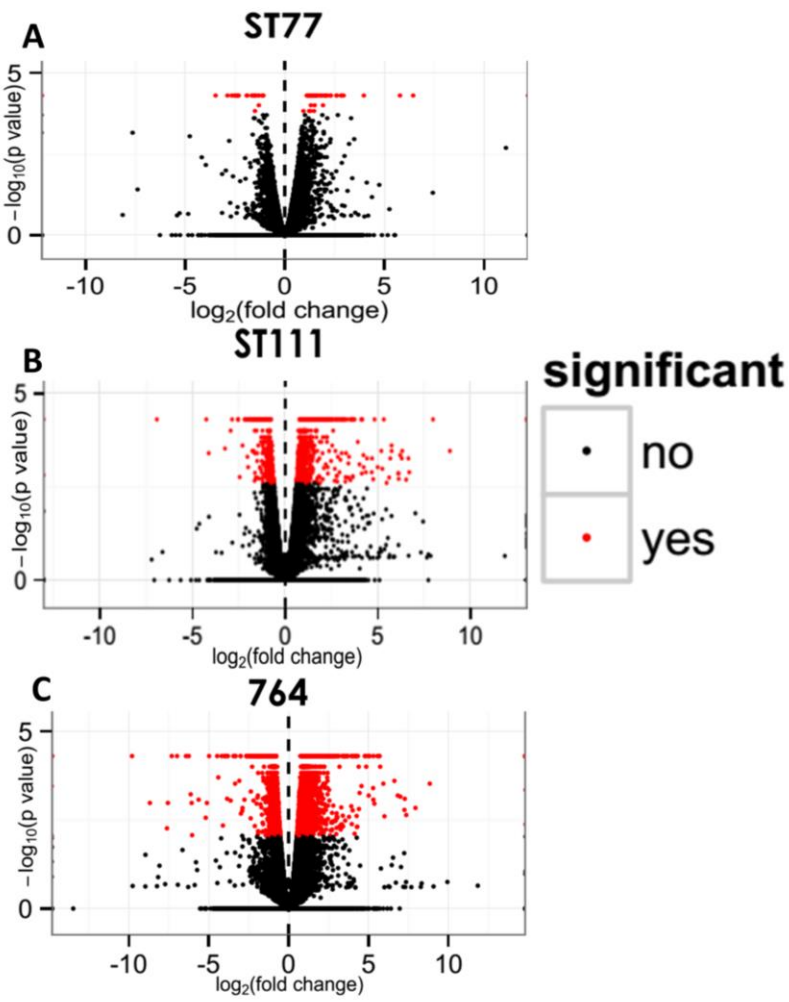


Figure 2.4: Cufflinks plots of differentially expressed genes. (A-C) Cufflinks volcano plots for each transgenic event showing variances in gene expression with respect to fold-change and significance. Each dot represents an individual gene. Black dots represent genes that are not significantly differentially expressed while red dots represent genes that are significantly differentially expressed.

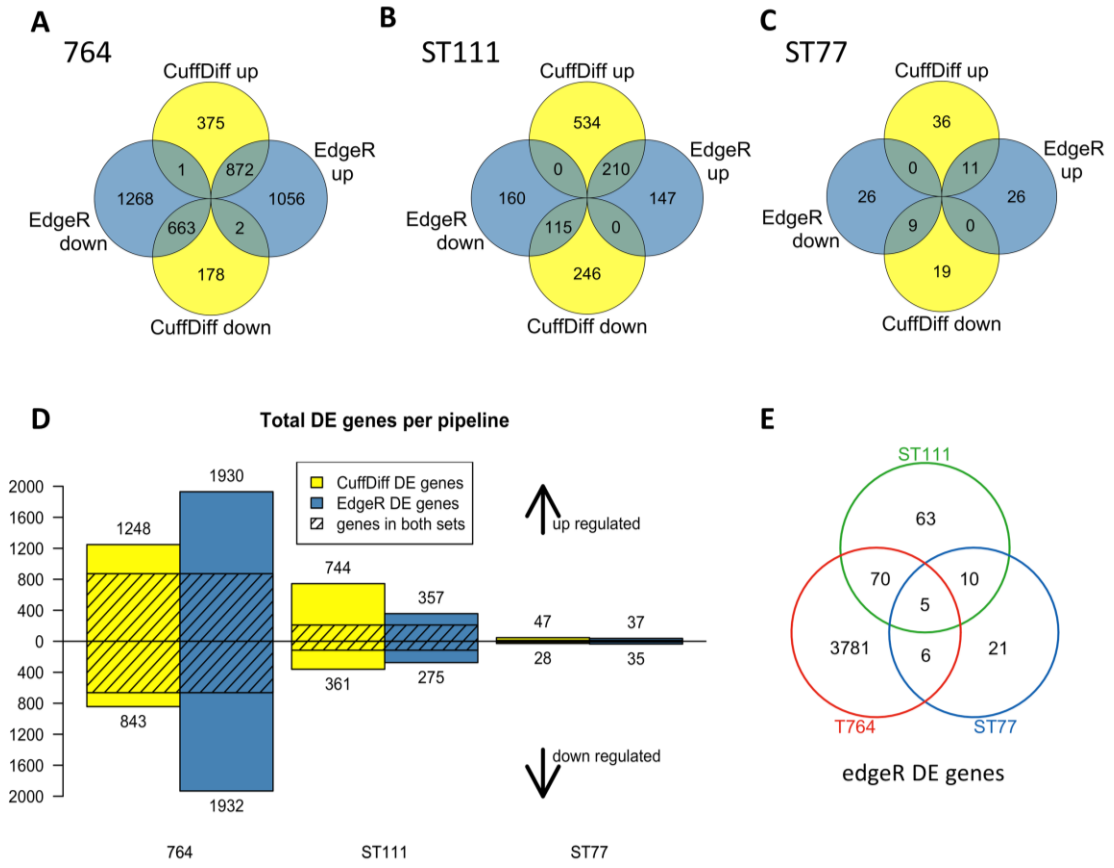


Figure 2.5: Venn diagrams of differentially expressed genes between edgeR and Cufflinks. Numbers of genes that are up and down-regulated in both edgeR and Cufflinks are shown for 764 (A), ST111 (B), and ST77 (C) lines. Total differentially expressed genes for each line determined by each program and the number of shared genes for each is shown in (D). Numbers of differentially expressed (DE) genes shared between each line from the edgeR results are illustrated in (E). Significance was defined by an FDR of 0.01.

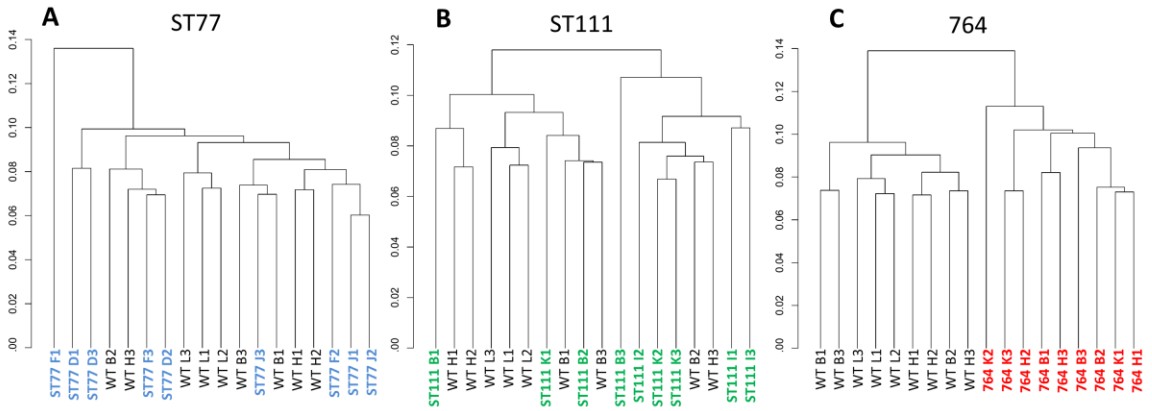


Figure 2.6: Dendrograms prepared by cummeRbund using differentially expressed gene lists. Samples were clustered in their respective groups compared to wild type and plotted based on variance (A-C).

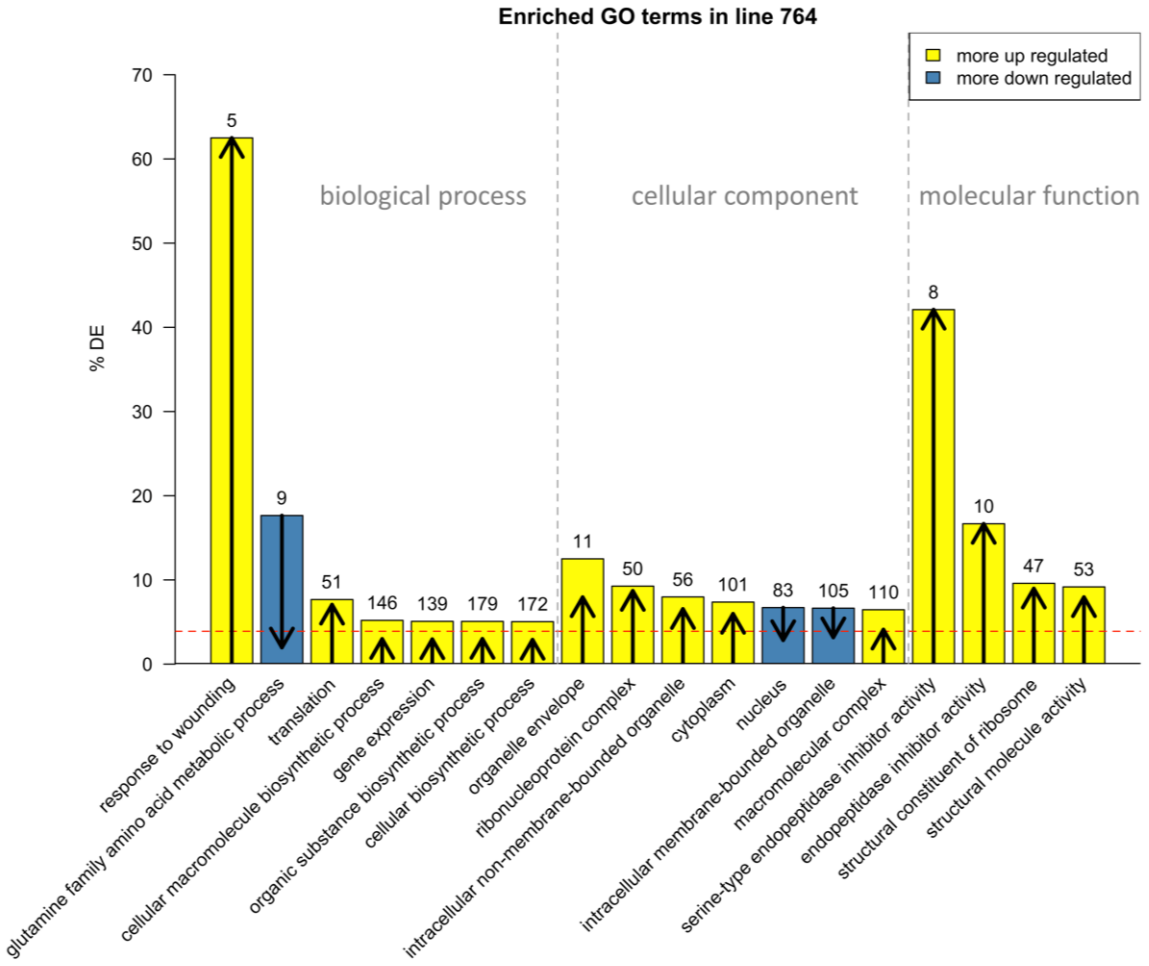


Figure 2.7: GOseq results from the edgeR and Cufflinks merged DE gene list. The numbers over each bar indicate the total number of differentially expressed (DE) genes in each category. The length and direction of the arrow in each bar indicate how many of the genes were up (arrows pointing up) or down (arrows pointing down) regulated in that category. The red dashed line indicates the percentage of all GO-annotated genes that are DE from edgeR and Cufflinks in this line (~4%). All terms shown have an FDR of 0.05 or smaller.

Table 2.1: Samples ligated to Illumina TruSeq adapters and their respective sequences. The specific lanes in which samples were loaded on the Illumina flow cell are indicated, as well as total reads, mapped reads, single and multi-mapping reads, percent mapping reads, and percent multimapping reads per sample.

Sample Name	Adapter	Sequence	Sequencing Lane	Fastq Reads	Mapped Reads	Single Mapping	Multi-Mapping	Percent Mapping	Percent MultiMapping
764-B1	2	CGATGT(A)	5	12188286	11430971	10249695	1181276	93.8	10.3
764-B2	4	TGACCA(A)	5	7539899	6898887	5696854	1202033	91.5	17.4
764-B3	5	ACAGTG(A)	5	13671914	12718266	11381904	1336362	93	10.5
764-H1	6	GCCAAT(A)	6	10905665	10276969	9364695	912274	94.2	8.9
764-H2	7	CAGATC(A)	6	17197458	16233695	14213752	2019943	94.4	12.4
764-H3	12	CTTGTA(A)	6	11657551	10803863	9046988	1756875	92.7	16.3
764-K1	13	AGTCAA(C)	6	13756687	12953387	11771484	1181903	94.2	9.1
764-K2	14	AGTTCC(G)	6	10104578	9419037	7533337	1885700	93.2	20
764-K3	15	ATGTCA(G)	6	13899704	12948837	11781113	1167724	93.2	9
ST111-B1	2	CGATGT(A)	4	12042216	11233063	10048614	1184449	93.3	10.5
ST111-B2	4	TGACCA(A)	4	12414573	11593382	10531103	1062279	93.4	9.2
ST111-B3	5	ACAGTG(A)	4	13866087	12741378	11634463	1106915	91.9	8.7
ST111-I1	6	GCCAAT(A)	4	9542865	8949278	8129392	819886	93.8	9.2
ST111-I2	7	CAGATC(A)	4	10833212	10236538	9248908	987630	94.5	9.6
ST111-I3	12	CTTGTA(A)	4	17240688	15787106	14408108	1378998	91.6	8.7
ST111-K1	13	AGTCAA(C)	5	8323147	7735687	6508842	1226845	92.9	15.9
ST111-K2	14	AGTTCC(G)	5	20006526	19125508	17682323	1443185	95.6	7.5
ST111-K3	15	ATGTCA(G)	5	6172327	5584459	4984882	599577	90.5	10.7
ST77-D1	13	AGTCAA(C)	3	15684558	14829092	13516676	1312416	94.5	8.9
ST77-D2	14	AGTTCC(G)	3	17957527	17067844	15263944	1803900	95	10.6
ST77-D3	15	ATGTCA(G)	3	14409042	13668166	12577710	1090456	94.9	8
ST77-F1	6	GCCAAT(A)	3	14081020	13325701	12422831	902870	94.6	6.8
ST77-F2	7	CAGATC(A)	3	8788020	8030031	7302092	727939	91.4	9.1
ST77-F3	12	CTTGTA(A)	3	18165328	17380007	15710166	1669841	95.7	9.6
ST77-J1	2	CGATGT(A)	2	5987939	5590605	5148010	442595	93.4	7.9
ST77-J2	4	TGACCA(A)	2	13776866	12995073	11747869	1247204	94.3	9.6
ST77-J3	5	ACAGTG(A)	2	7832717	7315857	6430098	885759	93.4	12.1
WT-B1	6	GCCAAT(A)	1	13432916	12896401	12057902	838499	96	6.5
WT-B2	14	AGTTCC(G)	2	15088733	14476530	13489239	987291	95.9	6.8
WT-B3	4	TGACCA(A)	1	13684295	13049162	11974273	1074889	95.4	8.2
WT-H1	12	CTTGTA(A)	1	14480179	13815997	12936889	879108	95.4	6.4
WT-H2	7	CAGATC(A)	1	15040106	14390591	11283722	3106869	95.7	21.6
WT-H3	15	ATGTCA(G)	2	13869661	13114740	12125983	988757	94.6	7.5
WT-L1	5	ACAGTG(A)	1	9889969	9303677	8328909	974768	94.1	10.5
WT-L2	2	CGATGT(A)	1	10389189	9902433	9149384	753049	95.3	7.6
WT-L3	13	AGTCAA(C)	2	9121134	8681226	7914912	766314	95.2	8.8

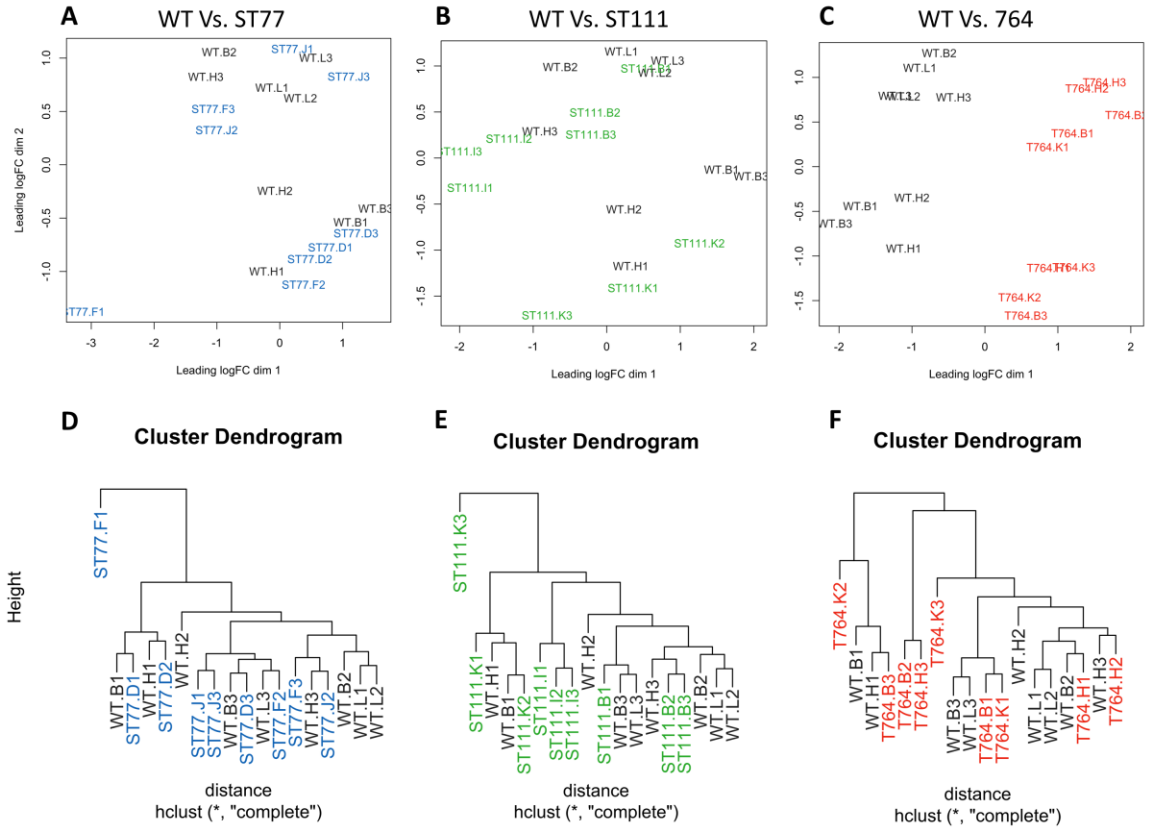


Figure 2.8: Multi-dimensional scaling plots of variance between samples from edgeR. Sample variance between ST77 (A), ST111 (B), and 764 (C) versus wild type are plotted based on differentially expressed gene number and fold change. The cluster dendrograms include all expressed genes for ST77 (D), ST111 (E) and 764 (F), showing the Euclidean distance between each sample based on overall gene expression.

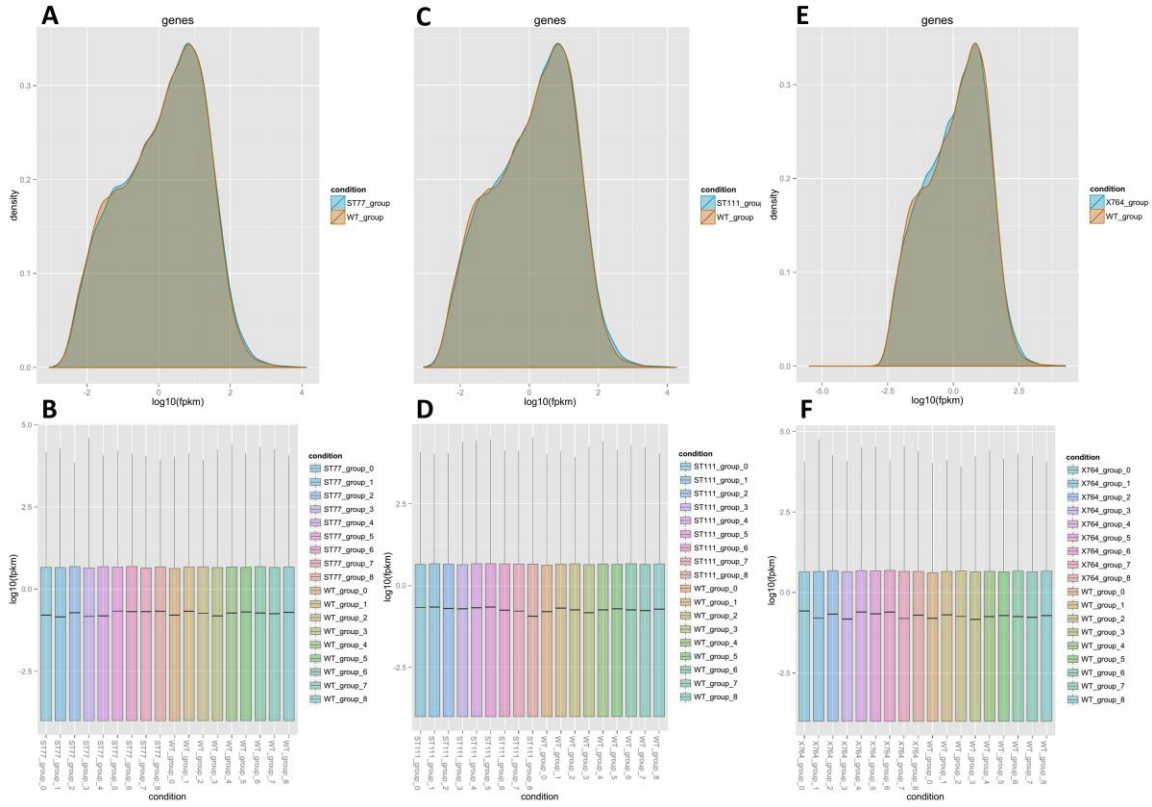


Figure 2.9: Normalization curves of gene density from cummeRbund of each transgenic event versus wild type (A, C, E) and each sample (B, D, F).

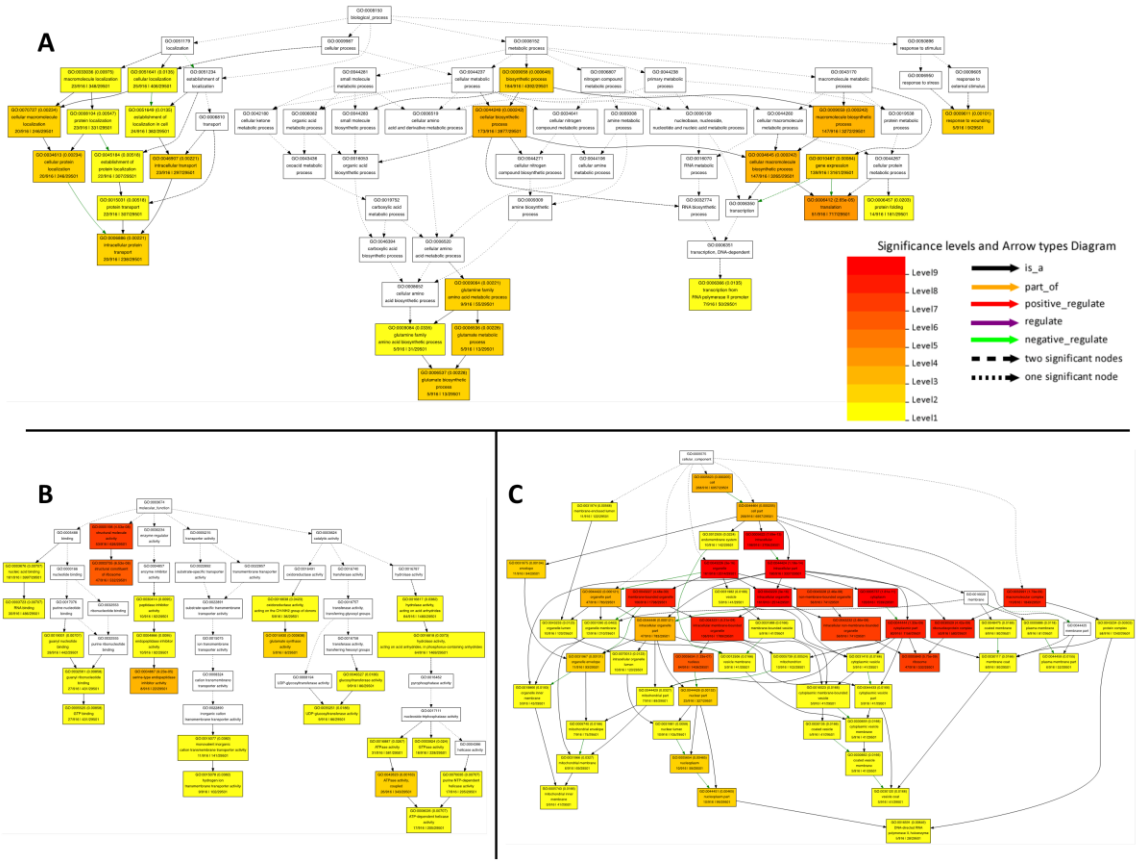


Figure 2.10: AgriGO single enrichment analysis results for the 764 event. (A) Enriched biological process GO terms (A), cellular component GO terms (B), and molecular component GO terms (C) for 764. The gene list used as input was a merged list of all genes considered differentially expressed (DE) by edgeR and Cufflinks with an FDR of 0.05.

CHAPTER 3: CONTRAILS: A TOOL FOR RAPID IDENTIFICATION OF TRANSGENE INTEGRATION SITES IN COMPLEX, REPETITIVE GENOMES USING LOW-COVERAGE PAIRED-END SEQUENCING

3.1 Introduction

Over the past two decades transgenic crops and foods have become integrated into worldwide agriculture, greatly increasing yields and easing cultivation labors through value added traits. *Agrobacterium*-mediated transformation and particle bombardment are common methods for creating crops to achieve this advancement. Current understanding indicates transfer DNA (T-DNA) integration into the host's genome is a random process that has been reviewed extensively [19, 24, 81, 135]. Characterization of integration sites is of great interest, particularly if the host is to be deregulated for human consumption or for commercial applications to assess potential pleiotropic effects resulting from transformation and evaluate the potential for inadvertent mutagenesis [85, 106, 136].

T-DNA inserts have been reported in both transcriptionally active and repressed regions of chromatin [14, 19, 24, 30, 135]. Additionally, in some instances T-DNA sequences have been detected within host endogenous genes, including promoter and regulatory regions [24, 137]. Transgenic plants containing multiple copies of T-DNA sequences have also been reported, and these complex events can lead to silencing of the gene of interest [11] emphasizing the favorable selection of simple, single T-DNA insertion events. Single insertion events in transgenic plants can be

generated through multi-generation propagation and are traditionally screened for complexity using Southern blots. While Southern blotting has been proven to be a reliable method for identifying copy numbers, no information regarding T-DNA insertion orientation, random DNA insertions or deletions at the insertion site, or the genomic location of the insert is revealed using this method. Furthermore, Southern blots can require extensive troubleshooting, may require radioactive materials, and can produce ambiguous results if the restriction enzymes exhibit star activity or digested genomic DNA products containing the transgene are similar in size. Thus, many alternative methods to estimate T-DNA copy number have been utilized but aren't without certain shortcomings.

Quantitative PCR analyses of transgene expression levels can be correlated with transgene copy numbers [138-140], although results from these methods are not always reliable due to other factors that could alter transgene expression independent of zygosity, such as gene silencing and truncation. Visualization methods such as Fluorescent In-Situ Hybridization (FISH) have been implemented for years to identify insertion regions on specific chromosomes [141-144], however this is a relatively expensive visual technique and confers no information about the surrounding sequence of the insertion region, or if tandem insertions have occurred. FISH must be coupled with targeted PCR amplification of sequences spanning the observed integration region, followed by sequencing to identify more precise integration points.

PCR techniques designed for transposon characterization, such as splinkerette PCR and inverse PCR [145-147], can reveal detailed integration information and have proven accurate for transgene insertion characterization due to reliance on sequence

specific initiation. Consequently, the presence of multiple or complex insertions, truncated transgene sequences, and highly repetitive genomes of host organisms can prevent: a) adequate detection, b) generate non-specific products, or c) fail to amplify products if primer targets are missing. Specialty restriction enzymes may also be required depending on the T-DNA fragment sequence (e.g.: methylation sensitivity, star activity, etc.), and a larger amount of genomic DNA is needed in order to visually verify digestion and ligation at each step. Genome walking has been employed effectively with universal primers [148], however as with the other PCR-based techniques, highly complex insertion events and repetitive genomic regions can potentially confound the results. In addition, larger T-DNA insertion sequences (e.g.: >10kb) are difficult to fully amplify in their entirety due to the limits of traditional polymerase activity; specialized polymerase varieties for longer amplification are available, but are more expensive than traditional polymerase, are subject to PCR-based assay complications, and can only extend amplification reliably to ~20-30kb.

In order to address these limitations, many groups have utilized next-generation sequencing (NGS) to identify and validate transgene insertion events [149-153]. Within the past 10 years, sequencing costs have been significantly reduced, while throughput and efficiency have greatly increased. NGS has already proven to be a reliable and accurate method for rapid identification of transposon insertion locations [154]. In addition, further analyses may be conducted on the resulting stored datasets in future genomic studies, such as genome-wide single nucleotide polymorphism (SNP) profiling, updated gene models and fusions, and complete sequencing of the transgene fragment for verification of the insert's integrity. Recently, several reports have successfully used

NGS to identify transgene insertion locations in various organisms [150, 151, 155], even at relatively low coverage (2-5X). The short turn-around time, coupled with the absence of a need for pre-experimental troubleshooting makes this a very attractive and cost-effective option for reliably identifying random transgene insertions. Furthermore, reference genomes for many species have been fully sequenced and are available for use, removing the need for complete genome *de novo* assembly of the resulting sequencing reads. This allows effective use of short read sequences in large and complex genomes, as efficient and accurate algorithms for such large *de novo* assemblies do not currently exist.

Here, we present and demonstrate CONTRAILS (Characterization of Transgene Insertion Locations with Sequencing): a pipeline using existing bioinformatics tools and paired-end Illumina next-generation genomic sequencing to identify and characterize transgene insertion locations in the highly complex and repetitive genome of the legume *Glycine max* (Figure 3.1). Paired-end reads spanning the T-DNA insertion junction allow for one read to map to the reference genome, and the other to map to the transgene sequence. Using short insert ($\leq$500b.p.) paired-end reads allows the user to narrow the insertion site to a genomic region of 500b.p. or less, provided assembly is assisted with an established reference genome. In some cases, it is possible for a single read to span both genomic and T-DNA sequences at the transgene insertion junction, giving immediate confirmation of insert location and neighboring sequences at single base resolution. However if this is not achieved, the matched paired-end reads will disclose the location well within conventional PCR amplification range for rapid characterization of the T-DNA junction sites. Using this technique, we have identified and characterized

a single T-DNA insert site in a transgenic line expressing recombinant hTG protein [9] to single-base resolution. These results are consistent with previous Southern blot and western blot screens, confirming the findings of the NGS analysis. Using this pipeline in conjunction with event-specific PCR assays, we were able to fully characterize flanking genomic sequences surrounding the T-DNA location.

3.2 Methods

3.2.1 Genomic DNA Extraction and Preparation

Whole-seed genomic DNA was extracted from chips of cotyledon tissue using a Maxwell 16 instrument and DNA extraction kit (Promega, Madison WI). Extracts were cleaned by phenol-chloroform and precipitated with 100% ethanol. DNA concentrations and purity were assessed with a Nanodrop 2000 spectrophotometer (Thermo Scientific, Waltham MA) and agarose gels to ensure optimal quality and concentration (260/280 absorbance ratio 1.8-2.0, greater than 1ug total DNA).

3.2.2 Illumina HiSeq 2000 Library Preparation, Sequencing, and Quality Control

Library generation was conducted at the David H. Murdock Research Institute genomics department according to the Illumina (San Diego, CA) HiSeq protocol, generating reported insert sizes of 350b.p. after quality control analysis. Paired-end sequencing was conducted on the Illumina HiSeq 2000 system. The soy sample ST77-KP2 characterized in this study was one of two pooled soy samples on a single lane sequenced to ~5x theoretical genome-wide coverage with 100 base-pair reads. Low-quality reads were filtered out using in-house Illumina software and validated with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/*fastqc)*, showing the remaining read basecall quality scores all greater than 30.

3.2.3 Reference Genome Construction and Read Alignment

The soybean reference genome sequence version 2.75 was obtained from Phytozome [112] and amended with an extra chromosome scaffold containing the T-DNA sequence located between the left and right border repeat regions (Figure 3.2) [9]. Paired sequence reads from the previously described seed genomic DNA sequencing were aligned to the constructed reference using Bowtie (ver. 2.2.1) [52] with parameters *--un-conc* to specify discordant read output. Default Bowtie search methods were used with zero allowed mismatches to limit ambiguous alignments due to the abundance of highly repetitive and homologous endogenous sequences, and in global mode to not trim read ends to enhance alignment scores.

3.2.4 Identification of the Transgene Insertion Site

Fragments in which one read aligned to known genomic reference sequence and the other read aligned to T-DNA sequence were flagged and separated from reads that aligned strictly to the known soybean reference sequence. Each enriched discordant read sequence were aligned against both the *Glycine max* reference genome using the "refseq_genomic" function in BLAST [156] and the T-DNA sequence, and matching mates were selected for further characterization. Reads matching the endogenous 7S glycinin promoter were detected in the filtered output and were excluded as illegitimate insertion sites. The genomic read furthest upstream and downstream from the T-DNA read pairs were selected for PCR amplification of the insert junctions to ensure the anticipated fragment was included within the selected genomic region.

3.2.5 Validation of T-DNA Insertion via PCR

Primers were designed to generate an amplicon that spans the genomic region and into both the right and left border sequences: genomic right border forward (5'-AGGATGACCCGACATGTCTCTTAG-3'), T-DNA right border reverse (5'-CAAATGAAGGGCATGGATCCTGC-3'), T-DNA left border forward (5'-CGGTTTGCGTATTGGCTAGAGC-3'), and genomic left border reverse (5'-GCCCGTCCTGAGCCTAAAATTG-3'). PCR amplification of the right and left border junction sequences consisted of an initial 5 minute denaturation step (95°C), followed by 35 cycles of 95°C for 30 seconds, 54°C for 30 seconds, and 72°C for 1 minute and a final extension at 72°C for 5 minutes. Wild type soy DNA was used as a negative control in reactions containing both border primer pairs, as well as with the right border forward and left border reverse primers as a positive control to amplify the native genomic locus. Amplified products were separated and visualized on 1% agarose gels stained with ethidium bromide.

3.2.6 Sequencing of Border and Junction Sequences

PCR reactions were cleaned in preparation for sequencing with phenol chloroform/3M sodium acetate containing glycogen as a carrier and precipitated with 100% ethanol at -80°C for 1 hour. Extracts were spun at 21,000xg for 15 minutes, washed twice with 70% ethanol and air dried for 10 minutes. Cleaned precipitated DNA pellets were re-suspended in molecular grade water and quantified with a Nanodrop 2000 spectrophotometer (Thermo Scientific, Waltham MA). Concentrations of PCR product were adjusted based on product size according to the recommendations provided by the University of California, Davis sequencing center (2ng/uL/100bp). Based on estimated sizes from migration in 1% agarose gels, the right border product was supplied at 8ng/uL

(~400 bases) and the left border product was supplied at 11ng/uL (~550 bases), giving both forward and reverse strand sequences for each junction. Primers were provided at a concentration of 3uM for sequencing.

3.2.7 Sequence alignment and Characterization

Sequences obtained from Davis showed a right border junction product of 373 bases and a left border junction product of 530 bases. Both sequences were BLASTed against the soybean reference genome, the T-DNA construct sequence for ST77, and against each other. Aligned regions were then extrapolated and examined for overlap by viewing the genomic sequence using the Integrated Genome Browser (IGB) [134] and the T-DNA sequence with SnapGene software (from GSL Biotech; available at snapgene.com).

3.3 Results

Illumina sequencing for the ST77-KP2 hTG sample generated 27,983,663 reads after quality filtering. Bowtie mapped 96.01% of the paired reads to the soybean reference sequence generating a theoretical whole-genome coverage of ~5x, establishing 8 total discordant read pairs mapping across the right and left border ends of the T-DNA sequence. Reads mapping to the genomic reference corresponded to sequences at a single locus on chromosome 3, with upstream reads beginning at bases 44,332,446 and 44,332,559 paired with reads 187 and 226 base pairs into the right border, respectively. Likewise, two reads within the left border region of the T-DNA at bases 11,269 and 11,433 paired with reads in downstream genomic sequence at bases 44,332,927 and 44,332,928 respectively. All discordant reads and respective information is shown in Table 3.1. This indicates a narrow region where the insertion occurred (base 44,332,659-

44,332,927 shown in Figure 3.3A), and illustrates that the right border of the T-DNA is oriented towards upstream genomic sequences in the 5' to 3' direction. Once this narrowed region was identified, primer design for genomic upstream and downstream sequences were facilitated utilizing the most recent *Glycine max* reference genome build in conjunction with visualization in IGB to achieve products within range for normal PCR amplification.

Junction sites were amplified for both the left and right border sequences generating products of ~400 bases and ~550 bases respectively. Sequencing results identified products of 373 and 530 bases for the right and left border PCR amplicons, respectively. The primers used for amplification, their attributes and the sequences generated are shown in Figure 3.3B. Alignments of these sequences to both the soybean genome reference and the T-DNA sequence identified the insertion site to single-base resolution at base 44,332,733. Furthermore, alignments revealed a 40 base pair deletion at the insertion locus on chromosome 3 as shown in Figure 3.4A. This deleted sequence was not part of an existing regulatory region, exon, or gene. In addition, 159 bases were deleted from the 5' end of the right border region from the T-DNA, but left the 7S promoter intact. From the junction sequencing data, we constructed a consensus sequence of the insert relative to the genome which is illustrated in Figure 3.4B.

3.4 Discussion

Previously we have demonstrated the efficacy of transgenic *Glycine max* as a cost-effective expression and storage system for recombinant proteins that are expensive to manufacture and/or difficult to generate in traditional systems [5, 7, 9, 92]. Until now, we have determined zygosity of transgenic events based on Mendelian inheritance,

western and Southern blotting. However these techniques reveal no characteristics of the T-DNA genomic insertion site, potential disruptions of endogenous genes, or truncation of the transgene and/or border sequences. Due to the highly repetitive nature of the soybean genome, our previous characterization attempts of the T-DNA using PCR-based techniques have failed to produce verifiable products.

Next-generation sequencing technologies offer a multitude of advantages when compared to traditional molecular characterization techniques, including rapid results, precise datasets that can be repurposed, exceptional consistency, and little experimental troubleshooting. Furthermore, sequencing costs are consistently decreasing every year making NGS methods more accessible to a larger range of investigators. In this study, low coverage paired-end genomic sequencing using the Illumina HiSeq 2000 platform was able to locate and identify a single copy transgene insertion in a highly complex and repetitive genome. The ability to use lower coverage is assisted with the existence of a reference genome to facilitate alignments. The absence of such a reference in a different organism would likely require higher coverage for confidence in the resulting assembly, however further optimization of *de novo* assembly algorithms will be the more likely technical bottleneck. The ability to pool multiple samples together on a single lane drastically reduces sequencing costs; however caution must be used to not dilute potential reads too extensively to avoid the possibility of a large coverage gap over the insert location, especially in organisms with particularly sizeable or complex genomes.

Insertion site identification exemplifies many properties of the transgene structure and can identify problematic or non-desirable transgenic events early in a production pipeline. Locations within interspersed repeat regions, or regions of heavy methylation

and dense chromatin may exhibit lower than expected expression of the transgene. Likewise, transformation events containing multiple copies of the T-DNA may show promise in molecular characterizations (e.g.: increased expression of the transgene), but are not ideal for the generation of homozygous events. Verification of the insertion site with PCR is rapid and straightforward to design using the genomic sequencing information, and can be used to screen other siblings from a particular event to assist in assessing zygosity for each specific locus. Quality control following library generation will report the total fragment size for each library, which can be used in conjunction with the genomic locations of the discordant reads to predict the size of the PCR products from the junction sites. Deviations from the reported insert size are not uncommon; extreme variances in the size of the amplified product may reveal a genomic deletion or insertion in the insert region that would otherwise remain undetected. The actual T-DNA sequence transferred to the host is contained between the right border and left border repeat regions, which act as cleavage signals for internal virulence factors in *Agrobacterium*. While designing PCR primers, it is prudent to choose sites well within the border boundaries to create a margin of safety against nucleolytic truncation and potential deletion of primer annealing sites. In addition, the raw aligned reads from the sequencing output may be consulted to verify the integration of these sequences and bolster confidence in the presence of primer annealing locations.

In some instances, illegitimate insert locations may be reported in the discordant read output if the T-DNA sequence contains promoter regions or other elements that are native in the target host (e.g.: glycinin promoters in soybean). In these cases, it would be beneficial to know the genomic location of these elements in the host genome prior to

designing PCR assays for junction sequencing, as this will aid in the selection of read pairs representing true insertion locations and prevent attempts to amplify an absent sequence.

In the case of multiple T-DNA copy events, screens for tandem T-DNA inserts are easily implemented using the same forward primers designed for amplification of left border junction sequences in conjunction with the reverse primers used for right border amplification. Likewise, reversed and inverted tandem insert junctions with the left border integrated in the 5' direction should form self-amplified products utilizing only the left border forward primer in the PCR reaction. Reversed tandem inserts may require an additional primer annealing to the lagging strand of the left border for amplification. Information pertaining to the orientation of the T-DNA at the identified locus is easily evaluated by comparing which region of the T-DNA is paired with the upstream and/or downstream genomic reads.

Assembly of reads spanning the T-DNA sequence can also be aligned to the reference construct sequence to assess total insert integrity without the use of step-wise PCR techniques. Fragmented or truncated inserts are easily identified in this way, preventing propagation of incomplete or partially transformed events. In addition, it is a relatively common occurrence for *Agrobacterium* to incompletely nick the T-DNA leading to read-through at the left border, possibly integrating vector features into the host [80]. The inclusion of vector backbone sequences in the T-DNA scaffold supplemented in the reference genome will allow for their detection as an integrated step.

Native endogenous gene disruption is moderately prevalent following *Agrobacterium* transformation via base inserts/deletions at the integration site, or direct

insertion of the transgene into native exons. Gene disruption can induce pleiotropic effects on the host, many of which may cause adverse effects that might not be phenotypically identifiable. Identification of these modifications as a result of integration breakpoints is a crucial advantage of CONTRAILS in candidate products for commercialization.

An indirect advantage to NGS-based approaches is the generation of permanent datasets containing extensive genomic sequence information. Soft data results are easily and rapidly referable, non-consumable, and are preserved indefinitely unlike biological samples. Collaborative efforts and the interpretation of results greatly benefit from shared digital datasets on cloud-based storage, and current organism-specific databases (e.g.: Soybase, the Soy Knowledge Base, Wormbase, etc.) and public repositories welcome the addition of new data. Further expansion of these freely accessible databases as genomics studies advance is crucial, and will serve as invaluable references for current and future genetic and molecular investigations.

3.5 Conclusions

Here we have demonstrated a cost-effective, rapid method for identification and characterization of transgene insertion locations in the complex, repetitive genome of transgenic *Glycine max*. Utilizing next-generation genomic sequencing and conventional PCR verification techniques, this method may be employed for many applications and genomes of varying complexity, with little to no time required for laboratory troubleshooting, using benchtop computational power in a straightforward pipeline. Considerable time savings from a universally applicable process, in conjunction with the

generation of extensive genomic datasets for future analyses, make this a valuable resource for genomics analysis of all organisms containing DNA insertions.

3.6 Availability of Supporting Data

Genomic sequencing files associated with the ST77 transgenic event described herein are available at the NCBI Short Read Archive under the Biosample accession number SRR2180176.

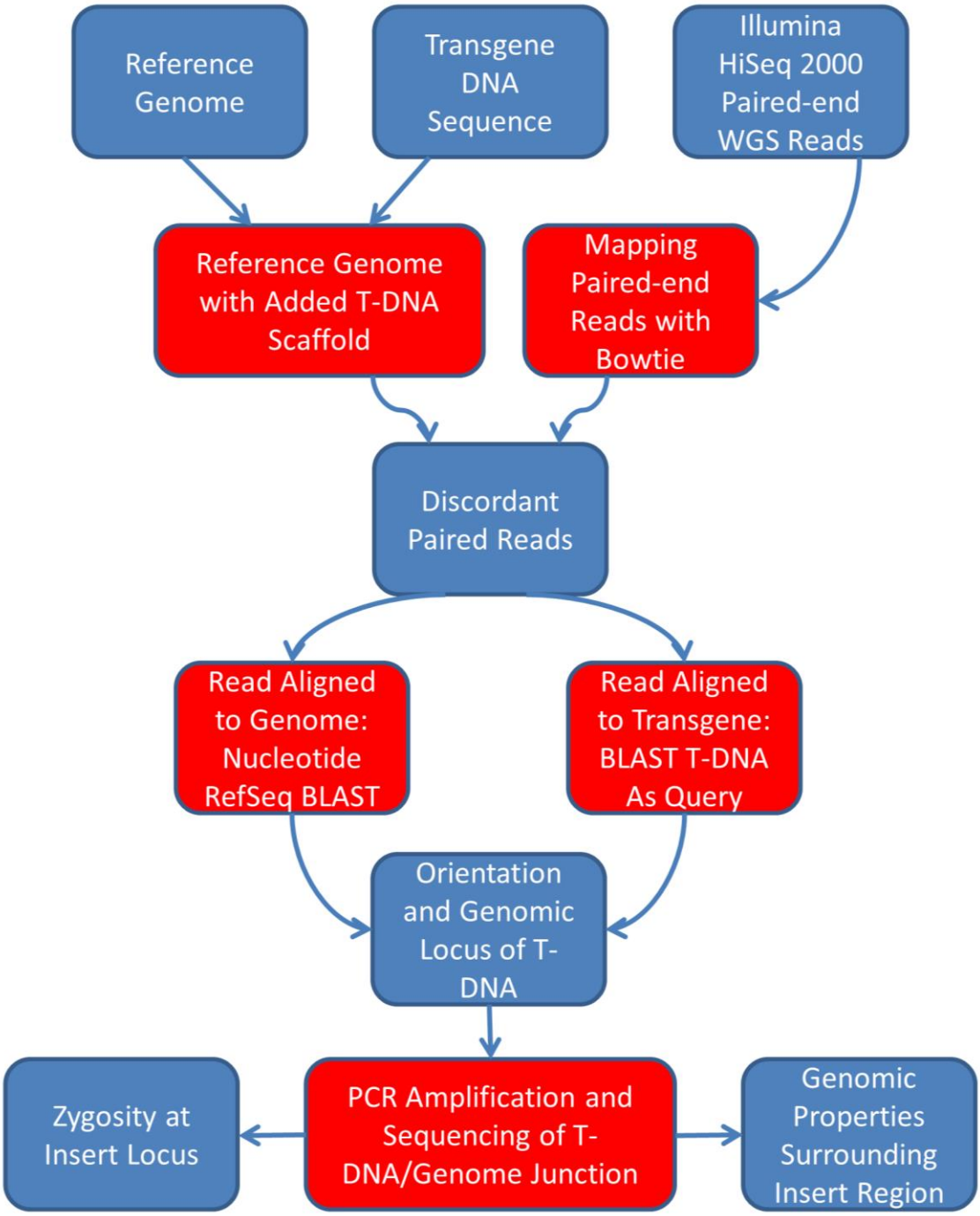


Figure 3.1: Experimental Pipeline. Flowchart detailing each major step in the pipeline, from DNA extraction and sequencing to alignment to the reference genome and T-DNA sequence.

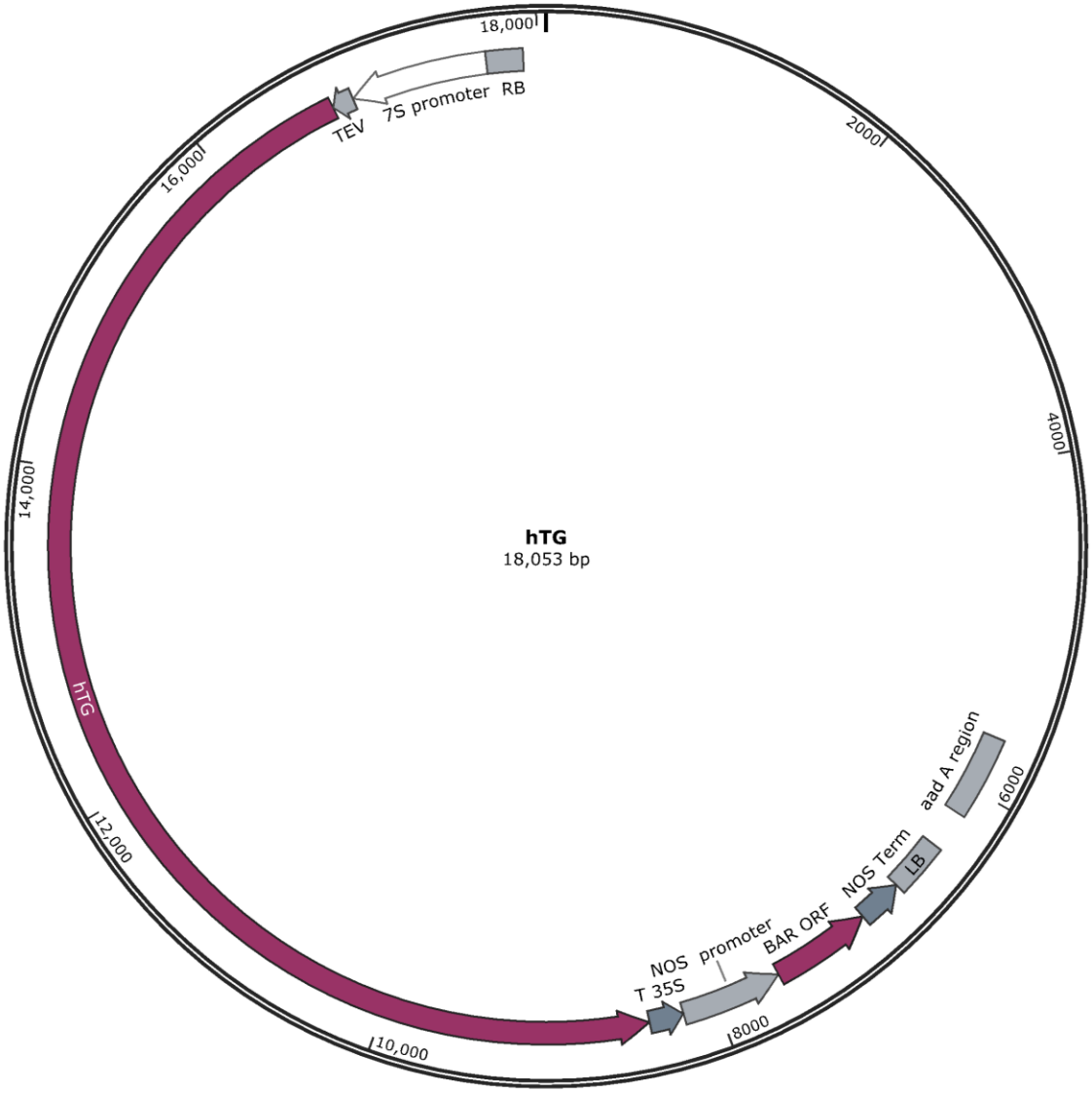
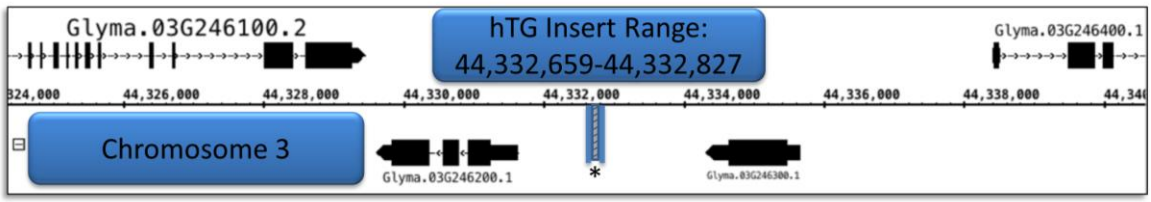


Figure 3.2: Plasmid map of hTG construct. The hTG plasmid map shows all regions included in the transformation plasmid utilized in the Agrobacterium transformation of the original ST77 event. The T-DNA construct contains the soybean β-conglycinin promoter (7S), tobacco etch virus translational enhancer element (TEV), human thyroglobulin gene (hTG), cauliflower mosaic virus terminator element (T35S) followed by the selectable marker cassette comprised of the nopaline synthase promoter (NOS promoter), phosphinothricin acetyltransferase gene (BAR ORF), and nopaline synthase terminator element (NOS Term). The aad A region of the vector confers antibiotic resistance to spectinomycin and streptomycin for selection of *Agrobacterium*.

Table 3.1: Discordant read pairs and sequences. All discordant read pairs for ST77 and the position of the start of the read are shown, as well as their mated sequence and pair relationship.

Read Origin	Start Base	Mate Pair Relationship	Read Sequence
Chr03	**44332446**	**mate1; other read matches reverse reference; this read is one of a pair**	**ATTAGGATGACCCGACATGTCTCTTAGAATGAGTAACATAAAACTTAGAATTATGGAAATTAGAATATTTCAAGAGCCTTTCACTTCAACTGATTATAAG**
scaffold_ST77	187	mate2; this read matches reverse reference; this read is one of a pair	AGTCACGACGTTGTAAAACGACGGCCAGTGCCAAGCTTGCATGCCTGCAGGATCCATGCCCTTCATTTGCCGCTTATTAATTAATTTGGTAACAGTCCGT
scaffold_ST77	**11433**	**mate1; other read matches reverse reference; this read is one of a pair**	**CGGCGTTAATTCAGTACATTAAAAACGTCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACAATATATCCTGTTCAACATTCAACA**
Chr03	44332928	mate2; this read matches reverse reference; this read is one of a pair	TAATAATAAAACAAGTAGTCCTTGGCTAGTTGGCTTACTTTTCATGTTTTAAGGAAACAAGTTGAGGAAGGGAAAAAATGTTGATACTGCTGCTCGTACG
Chr03	**44332927**	**mate1; this read matches reverse reference; this read is one of a pair**	**ATAATAATAAAACAAGTAGTCCTTGGCTAGTTGGCTTACTTTTCATGTTTTAAGGAAACAAGTTGAGGAAGGGAAAAAATGTTGATACTACTACTCGTAC**
scaffold_ST77	11269	mate2; other read matches reverse reference; this read is one of a pair	AAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGC
scaffold_ST77	**226**	**mate1; this read matches reverse reference; this read is one of a pair**	**CATGCCTGCAGGATCCATGCCCTTCATTTGCCGCTTATTAATTAATTTGGTAACAGTCCGTACTAATCAGTTACTTATCCTTCCTGCATCATAATTAATC**
Chr03	44332559	mate2; other read matches reverse reference; this read is one of a pair	ATTTAGTTAATACAACGTGGATGAAGAAAGGAAAGACATTAGAGAAAGAGTAAGCAAATAACGCACTCGATTTGTTATCTAATTAGTATGCTGTTGTACC

A



B

Primer Name	Sequence	Tm (°C)	GC Content
Forward 1	5'-aggatgacccgacatgtctcttag-3'	59	50%
Reverse 1	5'-caaatgaagggcatggatcctgc-3'	60	50%
Forward 2	5'-cggtttgcgtattggctagagc-3'	60	55%
Reverse 2	5'-gcccgtcctgagcctaaaattg-3'	60	55%

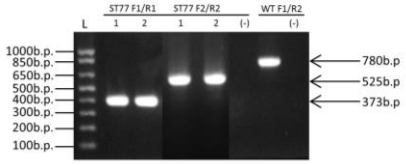


Figure 3.3: Insert location range and PCR verification. (A) The established maximum range of the location of the T-DNA insert based on discordant paired-end read mates. The discordant paired read reported farthest upstream began at base 44,332,659. The discordant paired read reported farthest downstream began at base 44,332,827. (B) Primer sequences and attributes used in the amplification of right and left border T-DNA junction sequences. The resulting products and their sizes are shown for the transgenic sample analyzed in duplicate, including a wild-type control using primers F1 and R2 to amplify the genomic insert locus in the absence of the hTG T-DNA.

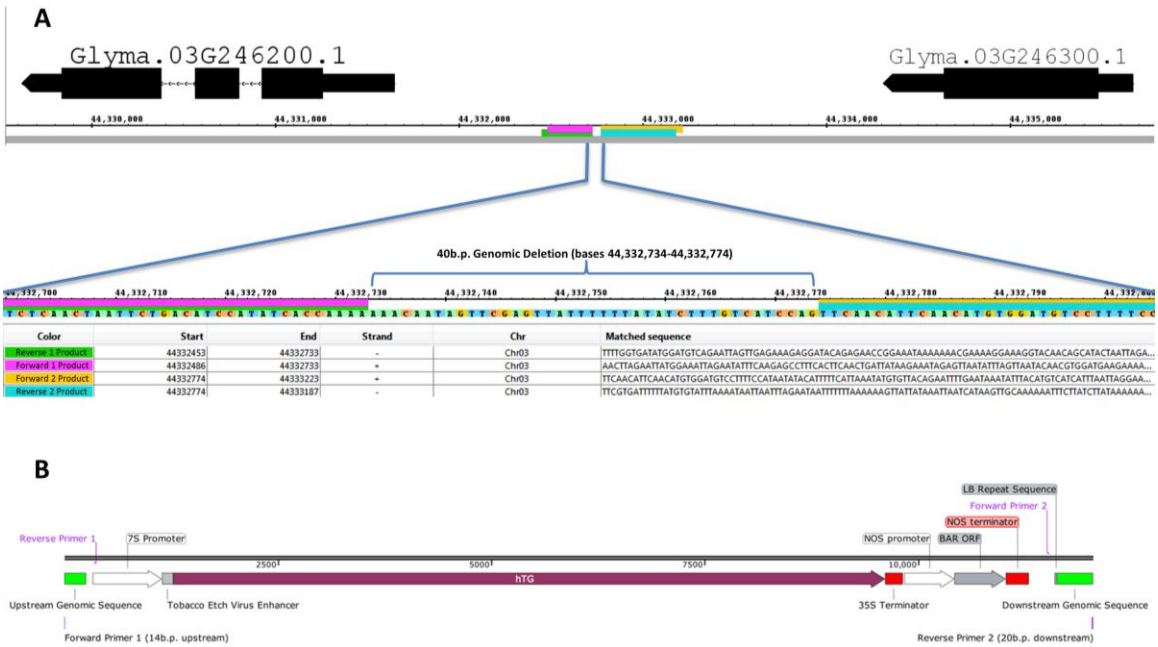


Figure 3.4: Aligned sequenced PCR products and insert layout. (A) Section of the insert location between two soybean genes. Colored bars represent sequences from the PCR amplicons of the junction sites that aligned to the soybean reference genome on chromosome 3. Purple is the product from primer F1, green from primer R1, yellow from primer F2, and blue from primer R2. 40 bases of genomic DNA have been deleted as a result of the insertion, shown as the uncolored region between the primer products. Start bases for each primer product are shown, as well as their alignment to either the sense or antisense DNA strand. (B) Illustration of the constructed consensus sequence of the T-DNA insert locus, showing the location of the primers used for junction characterization, flanking genomic DNA sequences, and inserted T-DNA elements.

CHAPTER 4: ENDOGENOUS SPLICING OF TRANSGENES, POLYMORPHISM RATES, AND T-DNA INSERTION LOCATIONS IN TRANSGENIC SOY

4.1 Introduction

Over the past three decades, transgenic plant biotechnology has become integrated into the agriculture of nearly all societies worldwide. Genetically modified food crops have revolutionized conventional farming methods by increasing yields [157], bolstering resistance to pests and herbicides [158, 159] , and enhancing resistance to environmental stresses [43]. Furthermore, plant systems have been employed for cost-effective production of biological therapeutics such as oral vaccine candidates [92], monoclonal antibody production [160], and accumulation of therapeutic and diagnostic proteins [7, 9]. Coupled with self-regenerating properties and the natural ability to stably store proteins in varied environmental conditions, plants offer a multitude of advantages over traditional cell culture systems for protein generation [8]. In the past 10 years, our lab has focused on *Glycine max* as a model system for the generation and storage of recombinant proteins that are currently expensive to manufacture, difficult to procure or synthesize in bacterial and cell-based cultures, and as oral vaccine candidates by targeting gut-associated lymphoid tissues (GALT) for immune stimulation or suppression [3, 5, 7, 92, 93, 161]. Advantages for using soybean as an expression system include high protein content and yield, self-fertilization simplifies

the generation of homozygous transformants, and stability of proteins in seed tissues, which we have previously reviewed [10].

Generation of transgenic plant specimens typically involves one of two procedures to incorporate foreign DNA: 1) Physical direct transformation via particle bombardment, electroporation, microinjection, or chemical treatment, or 2) Indirect transformation by *Agrobacterium tumefaciens*, a soil-dwelling rod-shaped plant pathogenic bacterium, via interkingdom horizontal gene transfer. While particle bombardment and other direct methods of integration maintain the advantage of being relatively quick, expression and incorporation is transient and limited to the current generation, as germ-line incorporation does not typically occur. For low complexity and stable integration of transfer DNA (T-DNA) containing the gene of interest to subsequent generations, *Agrobacterium*-mediated transformation is the preferred method.

In order for *Agrobacterium*-mediated transformation to occur, several genes required for the bacterium's virulence must be activated. These *vir* genes are located on the virulence plasmid outside of the T-DNA border regions, and are induced in response to signals of plant cell damage such as phenols and lignin precursors [162], which also function as chemoattractants for the bacteria. The products of these, namely *VirA* and *VirG* genes, indirectly allow for the attachment to the host cell and the activation of internal T-DNA transfer machinery. Virulence factors *VirB* and *VirD4* construct a type IV secretion system breaching the bacterial and plant cell membranes, effectively creating a conjugative bridge between the two cells in order to facilitate transfer of the T-DNA sequence [15]. *VirD2* then nicks the T-DNA at both right and left border sequences and attaches itself to the 5' end of the strand capping the sequence, protecting from

nucleolytic attack during transfer and also serving as a nuclear localization sequence (NLS) for nuclear targeted import of the T-DNA [17]. *VirE2* also coats the entire strand of the T-DNA during transfer and contains further NLSs, although these are suspected to be non-functional [18]. The actual transport of the single-stranded T-DNA molecule across the membrane has not been fully characterized; however it is speculated that host motor proteins and cytoskeletal rearrangements play a crucial role in this mechanism [163].

The process of *Agrobacterium*-mediated transformation has been investigated and reviewed extensively [14, 19, 28, 30]. However, information pertaining to the specific insertion process of the T-DNA strand into host nuclear chromatin after transport into the cell is not well understood, although it has been demonstrated to be a seemingly random event targeted to double-stranded breaks (DSBs) [25].

Recently, we investigated the pleiotropic effects on the transcriptome of soybean seed tissues expressing and accumulating high levels of recombinant protein [164]. Results revealed no correlation between transgene or recombinant protein expression level and the quantity of differentially regulated genes, although one of the transgenic lines contained over 3000 differentially expressed genes. We concluded that the gene expression differences observed may have been due to the specific properties of the recombinant proteins themselves on the homeostatic environment of the seed, or due to random mutagenesis; however characteristics of the transgene integration site and related molecular characterizations remained unknown. Information pertaining to T-DNA influence on adjacent gene expression and structure has been investigated before in *Arabidopsis* [84, 125], as well as junction sequences between the integrated cassette and

genomic DNA in *Arabidopsis*, rice and tobacco [24, 81, 165]. Due to the previously described random process of T-DNA integration utilizing *Agrobacterium*, and the possibility of endogenous gene disruption and genomic modification within the host due to transgenesis, it was imperative to identify both the transgene insertion location and junction sequences in all three transgenic events as well as any possible alterations to the native genome sequence. To accomplish this, we used our in-house CONTRAILS pipeline [166] to identify the T-DNA insertion sites in all samples of the three transgenic lines, and amplified the junction sequences of each with conventional PCR. In addition, we utilized the publicly available RNA-seq datasets generated by our previous transcriptome sequencing study to assess internal T-DNA alternative splicing, insertions/deletions (INDELs) and single nucleotide polymorphism (SNP) rates in the transgenic samples. This allowed us to investigate endogenous processing of our specific transgenes within soybean, as well as genomic variances exclusive to the transgenic plants.

4.2 Materials and Methods

4.2.1 T-DNA Alternative Splicing Analysis

To identify possible alternative splice sites within the non-endogenous T-DNA sequences, TopHat2 aligned all reads to the soybean genome version 2.1 containing each T-DNA genomic scaffold from the previous RNA-seq analysis. Resulting output files were subsequently loaded in the Integrated Genome Browser (IGB) [134] for visualization of possible splice sites using the "Find Junctions" to create a junction feature track from the alignment tracks. Following identification of the spliced read sequence location in the ST77 line, the entire genomic sequence of the transgene between

the right and left border repeats was analyzed for predicted donor and acceptor splice sites using NetPlantGene version 2.4 [167] to cross-reference with the observed junction site. The resulting spliced and unspliced consensus sequences reported by the reads were then loaded into the ExPASY SWISS-MODEL protein structure prediction workspace [168] to observe possible structural variations between the protein products resulting from each sequence variation.

4.2.2 Exon SNP/INDEL Analysis

In order to facilitate exome SNP calls, Samtools version 1.2 [169, 170] was used to index the soybean reference genome version 2.75 sequence file obtained from Phytozome [112] amended to contain scaffolds of all three T-DNA sequences using the *-faidx* command. Alignment files previously generated by TopHat [55] for all previously reported samples [164] in .bam format were converted from to .bcf files using the *-mpileup* command with *-g* and *-f* parameters to specify the output format and to use the indexed reference fasta file. The bcftools *call* command was subsequently used on the indexed .bcf files with the *-c* parameter to invoke the original consensus calling method enabling SNP and INDEL identification. The bcftools *stat* and *plot-vcfstats* commands were used to generate statistical summaries for each sample. Total SNP counts for each sample were averaged to calculate the standard deviation and standard error, and unpaired one-tailed t-tests were used to compare each transgenic group with wild type. Individual nucleotide base changes, transition and transversion rates, INDELS, single and multi-allele SNPs were also recorded and compared between groups.

To predict any possible translational effects resulting from detected SNPs and INDELS, snpEff version 4.1i [171] was utilized on the resulting variance call files

generated by bcftools. A custom database for snpEff was constructed using the *–build* command consisting of the soybean genome FASTA reference file described previously containing our T-DNA sequences, as well as the gene model .gff3 file from the Cufflinks output from our previous RNA-seq data [164] obtained from Phytozome. The .gff3 file provided a reference index for gene positions and identifiers, as well as intron/exon models and untranslated regions. No codon table configuration was necessary as *Glycine max* utilizes standard codon triplets allowing snpEff to run with the default parameters, employing the SNP/INDEL call .vcf file from bcftools as input. Multi-threaded processing using the *-t* option was not used, as this removes statistical calculations and the resulting reports from the output file. Statistical comparisons of SNP rates and effects between each group were conducted using Microsoft Excel. Functional annotation of genes containing variants was accomplished by loading the gene output list from snpEff into the agricultural gene ontology (GO) enrichment tool AgriGO [64] using the integrated single enrichment analysis tool.

4.2.3 Seed Genomic DNA Extraction and Preparation

Whole-seed genomic DNA from two representatives of each transgenic genotype was extracted from chips of cotyledon tissue using a Maxwell 16 instrument and DNA extraction kit (Promega, Madison WI). Extracts were cleaned with phenol chloroform/3M sodium acetate containing glycogen as a carrier and precipitated with 100% ethanol at -80°C for 1 hour. Extracts were spun at 21,000xg for 15 minutes, washed twice with 70% ethanol and air dried for 10 minutes. Cleaned precipitated DNA pellets were re-suspended in molecular grade water and quantified with a Nanodrop 2000 spectrophotometer (Thermo Scientific, Waltham MA) and 1% agarose gels to ensure

optimal quality and concentration (260/280 absorbance ratio 1.8-2.0, greater than 1ug total DNA).

4.2.4 Illumina HiSeq 2000 Genomic Library Preparation, Sequencing, and Quality Control

Library generation was conducted at the David H. Murdock Research Institute genomics department according to the Illumina (San Diego, CA) HiSeq protocol, generating reported insert sizes of 350b.p. after quality control analysis. Paired-end sequencing was conducted on the Illumina HiSeq 2000 system and two soy samples were pooled together on each lane and sequenced to ~5x theoretical genome-wide coverage with 100 base-pair reads. Low-quality reads were filtered out using in-house Illumina software and validated with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc), showing the remaining read basecall quality scores all greater than 30.

4.2.5 Identification of the T-DNA Insert Locations Using CONTRAILS

Paired sequence reads from the previously described seed genomic DNA sequencing were aligned to the previously described constructed reference using Bowtie (ver. 2.2.1) [52] with parameters --un-conc to specify discordant read output. Default Bowtie search methods were used with zero allowed mismatches to limit ambiguous alignments due to the abundance of highly repetitive and homologous endogenous sequences, and in global mode to not trim read ends to enhance alignment scores.

Fragments in which one read aligned to known genomic reference sequence and the other read aligned to T-DNA sequence were flagged and separated from reads that aligned strictly to the known soybean reference sequence. Each enriched discordant read

sequence was BLASTed against both the *Glycine max* reference genome (using the "refseq_genomic" function) and the T-DNA sequence, and matching mates were selected for further characterization. Reads matching the endogenous 7S glycinin promoter were detected in the filtered output and were excluded as illegitimate insertion sites. The genomic read furthest upstream and downstream from the T-DNA read pairs were selected for PCR amplification of the insert junctions to ensure the anticipated fragment was included within the selected genomic region.

4.2.6 Primer Design and PCR of the T-DNA Junction Sequences

Primers were designed for each transgenic line to generate an amplicon that spanned the genomic junction and into both the right and left border sequences based on the discordant reads identified above. Primers utilized for the ST77 transgenic line: genomic right border forward (5'-AGGATGACCCGACATGTCTCTTAG-3'), T-DNA right border reverse (5'-CAAATGAAGGGCATGGATCCTGC-3'), T-DNA left border forward (5'- CGGTTTGCGTATTGGCTAGAGC-3'), and genomic left border reverse (5'- GCCCGTCCTGAGCCTAAAATTG-3'). PCR amplification of the right and left border junction sequences for ST77 consisted of an initial 5 minute denaturation step (95°C), followed by 35 cycles of 95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 1 minute and a final extension at 72°C for 5 minutes. For the 764 line, primers for genomic right border forward (5'-GTGCCGTGTTTCAGAACATCTCG-3'), T-DNA right border reverse (5'-CTTAGGCTAGGATCCTGCAGGC-3'), T-DNA left border forward (5'- CCAGCTGCATTAATGAATCGGCC-3'), and genomic left border reverse (5'- GGATGGCAAGGCAAGTAGACTC-3') were used. The 764 PCR reaction steps consisted of an initial 5 minute denaturation step (95°C), followed by 35 cycles of 95°C

for 30 seconds, 54°C for 30 seconds, and 72°C for 1 minute and a final extension at 72°C for 5 minutes. Primers for ST111 right border amplification consisted of genomic right border forward (5'- GCAAGAACAAAATGTCCCTGCGG-3') and T-DNA right border reverse (5'-TGGCCGTCGTTTTACAACGTCG-3'), with an initial denaturation of 95°C, followed by 35 cycles of 95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 1 minute and a final extension at 72°C for 5 minutes. Each primer set was used to screen the junction sites of all nine individuals from each transgenic line. Wild type soy DNA was used as a negative control in all reactions containing both border primer pairs, as well as with the right border forward and left border reverse primers as a positive control to amplify the native genomic locus. For all reactions, ~50 ng of genomic DNA was mixed with GoTaq Flexi DNA polymerase (Promega, Madison, WI) and buffers provided by the manufacturer. Amplified products were separated and visualized on 1% agarose gels stained with ethidium bromide.

4.2.7 Sequencing and Alignment of Amplified Junctions

PCR reactions were cleaned in preparation for sequencing with phenol chloroform/3M sodium acetate containing glycogen as a carrier and precipitated with 100% ethanol at -80°C for 1 hour. Extracts were spun at 21,000xg for 15 minutes, washed twice with 70% ethanol and air dried for 10 minutes. Cleaned precipitated DNA pellets were re-suspended in molecular grade water and quantified with a Nanodrop 2000 spectrophotometer (Thermo Scientific, Waltham MA). Concentrations of PCR product were adjusted based on product size according to the recommendations provided by the University of California, Davis sequencing center (2ng/uL/100bp). Based on estimated sizes from migration in 1% agarose gels, the border products were supplied at 2ng/uL for

every 100 bases of sequence. Twelve microliters of each product (6uL per reaction) was supplied to conduct the sequencing reactions for forward and reverse primers, giving both forward and reverse strand sequences for each junction. Primer volumes provided were 4uL for each reaction at a concentration of 3uM. Sequences received were BLASTed against both the reference genome, the vector sequence for each transgenic event, and each other to discover overlapping bases and establish a consensus sequence for the junctions.

4.3 Results

4.3.1 ST77 hTG Transgenes Are Alternatively Spliced in Soybean Seed

Alternative splicing analysis of all transgenic individuals revealed splice sites within the hTG transgene in ST77-D and ST77-J siblings. A 98 b.p. intronic segment from bases 6893-6990 encoding 32 amino acid residues was removed in 9 reads (~11%) spanning the splice site in ST77-D3 with an average sequencing depth of ~80x across the junction. ST77-D1, ST77-D2, ST77-J1 and ST77-J2 also reported the same splice junction, although with lower counts of spliced reads likely due to lower sequencing depth. The junction initiated at an exon to intron motif on the direct strand between two guanine residues (5'- TCTCAACCAG^GTGATCGTTA-3'). No splicing was detected in the transgenes of any individuals in the 764 or ST111 events. Coverage of the splice junction and the sequences of the region are shown in figure 4.1.

NetPlantGene running from the NetGene2 server revealed 15 possible predicted donor splice sites across all 11521 nucleotides of the hTG open reading frame's leading strand, while the compliment strand produced 34 predicted donor sites. Two leading strand donor sites were predicted with greater than 90% confidence, neither resulting in

any detectable splicing; however the actual splice site at base 6893 was indeed predicted with a 50% confidence. Three high confidence sites were predicted in the complement strand approximately 6300 bases in from the 3' sequence end, none of which produced detectable splice junctions from the RNA-seq transcripts; however the detected downstream acceptor site was indeed predicted with a low confidence of 15%. The NetGene2 predicted donor sites and the matching acceptor site for the ST77 splice junction are shown in table 4.2.

SWISS-MODEL was used to predict tertiary structure changes as a result of the splicing event, which revealed several possible superficial alterations to external motifs of the hTG protein structure. One predicted beta sheet motif at Arginine 2691 and Alanine 2692 was removed in the spliced sequence, as well as orienting the Arginine 2676 loop further inward towards the core of the peptide, while overall core structure appears to remain relatively unchanged. Protein models of the predicted spliced and unspliced peptides as well as the spliced nucleotide and amino acid sequences are shown in figure 4.2.

4.3.2 SNP Rates in Transgenic Soybean Exons

Total SNPs detected by samtools and bcftools across the wild type control group reported a mean of 20,707 SNPs per sample, with a mean SNP rate of one polymorphism for every 42,561 nucleotides. Of these, 48% on average are single allele polymorphisms with 35 multi-allele sites containing 23 multi-allelic SNPs. An overall transition (a purine base changed to another purine base or pyrimidine base changed to another pyrimidine base) to transversion (a purine base to pyrimidine base or pyrimidine base to purine base alteration) ratio (Ts/Tv) of 1.62 and an average of 1862 INDELS were

identified per sample in the wild type group. Mean base changes detected were as follows: 888 A→C, 3458 A→G, 1428 A→T, 793 C→A, 769 C→G, 2977 C→T, 2941 G→A, 787 G→C, 818 G→T, 1495 T→A, 3433 T→C, and 944 T→G.

The ST111 experimental group reported an average of 20,208 SNPs per individual sample, with a polymorphism rate of one SNP for every 44,323 nucleotides. Similar to wild type, 48% of the SNPs in the ST111 line were single allele alterations, with an average of 34 multiallele sites and 25 multiallele SNPs. The Ts/Tv ratio was again very similar to wild type at 1.63 with an average of 1750 INDELS per sample. Per base changes reported for the ST111 group were 933 A→C, 3457 A→G, 1382 A→T, 747 C→A, 710 C→G, 2868 C→T, 2830 G→A, 754 G→C, 779 G→T, 1420 T→A, 3415 T→C, and 968 T→G. ST77 was also comparable to wild type with 21,666 average SNPs per sample at a rate of 41,225 nucleotides for every polymorphism. Forty-nine percent of the detected SNPs were reported as singletons, with 33 identified multiallele SNP sites and 28 multiallele SNPs. The Ts/Tv ratio for the ST77 group was slightly lower than wild type at 1.56, indicating a slightly higher transversion rate in the ST77D and ST77F siblings. INDELS were nearly identical in number and size to the wild type control group with an average total of 1829 per sample. Nucleotide base changes include 1119 A→C, 3650 A→G, 1468 A→T, 837 C→A, 772 C→G, 3015 C→T, 2952 G→A, 796 G→C, 842 G→T, 1520 T→A, 3589 T→C, and 1135 T→G. Lastly, the 764 event contained more substantial nucleotide alterations than the previous three experimental groups revealing an average SNP count of 38,188 per individual; nearly double the quantity of any other group. SNPs were also encountered on average at nearly double the frequency of the other samples, with one SNP recorded every 24,281 bases. In this

group, only 27% of SNPs were reported as singletons, with 78 multiallele sites and 44 multiallele SNPs. The Ts/Tv ratio was the lowest of all four groups at 1.53, indicating the 764 group had the highest overall rate of nucleotide transversions. INDELS also increased to 2390 with a larger deviation spread between samples. Base changes included 1972 A→C, 6094 A→G, 2596 A→T, 1603 C→A, 1386 C→G, 5445 C→T, 5461 G→A, 1417 G→C, 1639 G→T, 2576 T→A, 6104 T→C, and 1940 T→G. Average total SNPs, INDELS, SNP rates, and transition/transversion ratios for each group are shown in figure 4.3, and individual base change rates for each group are shown in figure 4.4A-D.

4.3.3 Classification of Possible SNP Effects and Impacts

Following SNP and INDEL detection with samtools, snpEff version 4.1i [171] was used to evaluate potential alterations to the exome resulting from these changes. SnpEff annotates variants based on their genomic locations, including introns, exons, upstream or downstream, splice sites, and untranslated regions at 5' and 3' sequence ends. Effects were grouped and sorted according to the variant type, potential level of effect impact, functional class, type and region. No multi-nucleotide polymorphisms were detected in any individuals of all four experimental groups, and modifications were limited to SNPs, insertions, and deletions of bases. Across the wild type and all three transgenic groups, polymorphisms defined as low effect average ~10%, moderate effects average ~12%, and high effects average the lowest of the general groups at ~2%. The largest category of impact classification is the genomic modifier category, comprising 72-78% of detected polymorphisms in all groups (see snpEff reports in additional data). Missense mutations are highly prevalent in all four groups, averaging ~60% in wild type,

ST111, and ST77 events with a missense to silent mutation ratio of ~1.6. The 764 group exhibited a slightly lower missense percentage of ~53%, however it also contained a higher percentage of silent mutations at ~45% compared to the other three groups (~37%) generating a missense/silent mutation ratio of ~1.20. Nonsense mutations comprised ~3% of the total reported SNP effects for the wild type, ST111, and ST77 groups, and ~1.8% in the 764 group (figure 4.4E-H).

All four experimental groups showed a high prevalence of SNPs of common types, namely resulting in downstream gene variants (~24%), intron variants (~25%), missense variants (~11%), synonymous variants (~7%), and upstream gene variants (~16%). The most common locations for these variants are intron regions (~25%), downstream (~23%), exons (~23%), introns (~23%), and upstream (~16%). The 764 group distribution of polymorphism types and affected regions seem to mirror those of the wild type, ST111 and ST77 groups, with a slight reduction of SNPs in exonic regions. All individuals also reported variants within 5' and 3' untranslated regions (~4% and ~6% respectively), as well as minimal detection of frameshifts, stop, and start variants (<1%). Regions and predicted effects of detected polymorphisms are shown in figure 4.4E-H.

Transition base changes of A←→G and C←→T were most prevalent in all four experimental groups, while the transversion A←→T was ~50% more common than A←→C, C←→G, or G←→T transversions. Codon changes expectedly varied most commonly at the third wobble base position, typically resulting from a transition in both wild type and transgenic samples. Following SNP patterns previously described, the prevailing modified codons appeared commonly shared between all individuals

measured. Base changes in the first and second positions of the codon were extensively less frequent across all groups. Heatmaps of codon base changes for each experimental group are shown in figure 4.9. Predicted amino acid changes resulting from codon alterations predominantly consist of synonymous substitutions, however other frequently altered residues include valine to alanine, alanine to valine, leucine to phenylalanine and proline, glutamate to glycine and lycine, serine to proline, and phenylalanine to leucine. figure 4.10 shows the distribution heatmaps of detected amino acid substitutions from all events.

Minor changes were also detected within the additional genomic scaffolds containing the transgene sequences. Specifically, snpEff reported SNPs in the transgenes of samples 764B3, 764H3, and 764K1 at position 80, ST77D2 at position 1400, ST77F1 at positions 500 and 700, ST77J3 at position 600, ST111B3 at position 800, ST111I3 at position 190, and ST111K1, ST111B1, and ST111K2 at position 1010. Read alignments were unable to verify all called variants in the transgenes, likely due to areas of low coverage, however one area was corroborated as a consistent variant in the filler DNA of all constructs between the right border and promoter region (figure 4.11), which was included as a control. Table 4.3 summarizes all SNP calls and between group statistical tests. Summaries of all SNP/INDEL calls from bcftools and effects from snpEff, along with chromosomal distribution plots and additional statistics are available from the iPlant collaborative Discovery Environment [172] from the links provided in the supporting data section of this manuscript.

4.3.4 Gene Ontology Analysis of Genes Containing SNPs

Complete gene lists containing SNPs provided by the output of snpEff for the wild type (36,959 transcripts), ST77 (38,691 transcripts), ST111 (34,142 transcripts), and 764 (43,426 transcripts) lines were loaded into the AgriGO online gene ontology analysis tool and subjected to a single enrichment analysis with multiple corrections. Out of the total *Glycine maxV2.1* background set of 29,501 GO terms, wild type matched 11,245, ST77 matched 11,660, ST111 matched 10,326, and 764 matched 13,321. GO categories were similar between groups with minor variations, with the majority of terms putatively annotated to translational or RNA processes. Intracellular transport also constituted a large portion of the identified categories (~12% of total annotations), and in the case of the 764 transgenic line, was the only other significant GO term following RNA processing and binding. Three categories including ATP-dependent helicase activity, ubiquitin-dependent protein catabolic processes, and ncRNA metabolic processes were all unique to the ST77 and ST111 lines. GTP binding was exclusive to the ST111 line, comprising 21% of the total significant GO terms for the event. Specific GO terms for each group are shown in figure 4.5, and complete AgriGO feature relationship trees are shown in figure 4.12.

4.3.5 Transgene Integration Sites and Structure

Illumina sequencing for the ST77 hTG samples F3 and J2 generated 27,983,663 and 30,278,254 total reads, samples 764-B1 and 764-K1 generated 27,731,188 and 30,406,940 total reads, while ST111-I1 and ST111-K2 returned 31,734,725 and 27,673,527 total reads respectively after quality filtering. Bowtie aligned ~96% of all reads to the soybean reference genome across all samples, generating a theoretical sequencing depth of ~5x. Discordant read mismatches for each sample were compiled

from read relationships in which one read of a pair mapped to one soybean chromosome and the other read mapped to one of the T-DNA scaffolds, thereby isolating the T-DNA junction sites [173].

ST77-F3 sequencing reported two sets of discordant read pairs, each pair spanning a border junction. The upstream genomic junction read began at base 44,332,527 on chromosome 3 matching the sense DNA strand, while the T-DNA read matched the reverse reference sequence beginning at base 155 into the T-DNA sequence at the right border. The second mate pair spanned the left border region of the T-DNA beginning at base 11,421 and extending downstream, matched to the genomic downstream read beginning at base 44,333,017 towards the left border side of the T-DNA. This narrowed the insert location between bases 44,332,627 and 44,442,917, as each 100bp genomic read contained no vector sequence. The second ST77 sample, J2, confirmed this finding with a left border mate pair beginning at base 11,369 of the T-DNA facing downstream together with a downstream reverse genomic read facing the left border beginning at base 44,332,827, further narrowing the insert site to bases 44,332,627-44,332,827 (figure 4.6D).

PCR primers to amplify both the left and right borders were designed upstream and downstream of the closest aligned genomic reads to insure all junction sequences were incorporated in the amplified product. Sequencing of the amplified regions of the right and left borders generated products of 373 and 530 bases respectively. Alignments to the soy genome and T-DNA sequence revealed the T-DNA insertion began precisely at base 44,332,733 between genes Glyma.03G251500 and Glyma.03251600, neither of which were differentially expressed nor perturbed by the insertion. The right border side

of the T-DNA sequence inserted first, and 159 bases were deleted from the 5' end of the sequence removing the right border repeat entirely from the insert, although the 7S promoter region remained completely intact. The left border amplicon sequence revealed only a 3bp deletion from the border repeat sequence, but 40 bases (44,332,734-44,332,774) were deleted from the genome at the insertion site. PCR reactions for all 9 transgenic ST77 samples returned a single product of identical size for both right and left border sequences, indicating a single, stable, homozygous integration event (figure 4.6A).

764-K1 returned three mate pairs: two for the right border junction and one for the left border junction. The upstream read on the sense strand began at base 3,007,378 on chromosome 11 oriented towards the right border read pair, which matched the reverse reference of the 764 scaffold beginning at base 385 towards the upstream genomic sequence. The second mate pair began at base 3,007,466 on chromosome 11 also oriented towards the right border T-DNA sequence, matched to the right border mate beginning at base 398 in the T-DNA facing the upstream genomic sequence. This indicated that the right border integrated first in the 5' to 3' orientation in the same fashion as ST77 above. The left border read pair began at base 3790 in the T-DNA in the forward orientation towards the left border repeat, and was mated to the reverse complement genomic read beginning at base 3,009,429 on chromosome 11. The second sample 764-B1 also confirmed this junction with reads at base 3742 and 3,009,286 in the T-DNA and downstream genomic sequences respectively (figure 4.6E).

As described previously, the 764 PCR primers to amplify both junction sequences were designed from the farthest upstream and downstream genomic reads to ensure the junction sequences were encompassed in the PCR products. Sequencing of the junction

amplicons generated products of 444 nucleotides for the right border junction and 357 nucleotides for the left border junction. Alignments of the resulting sequences to the soybean genome indicated the right border inserted first at base 3,007,579 between genes Glyma.11G041200 and Glyma.11G041300, neither of which were differentially expressed or disturbed by the insert. Alignments to the T-DNA sequence indicated 6 bases were deleted from the 5' end behind the right border repeat sequence, significantly less than the ST77 event described above. On the 3' end of the leading strand, 3 bases were deleted just before the left border repeat, again removing the border sequence from the T-DNA insertion. Four non-matching bases of ACAT were located at the end of the transition between the T-DNA product and genomic product junction that did not match either reference sequence. The insertion also resulted in the deletion of 1625 bases from the soybean genome (bases 3,007,580-3,009,205), the longest detected in all three transgenic lines. All nine 764 transgenic seeds were screened using the same primers, generating identical product sizes for both the right and left border primer sets in all samples confirming homozygosity (figure 4.6B).

ST111-K2 returned two mate pair sets, both spanning the right border T-DNA junction, while ST111-I1 returned only illegitimate pairs to chromosome 10 matching to the endogenous promoter sequence. No mates were reported for the left border junction. The farthest upstream read began at base 12,332,747 on chromosome 6 facing the T-DNA insert, with a mated read beginning at base 537 into the 7S promoter of the ST111 scaffold facing the upstream genomic sequence. The second mate pair began at base 12,332,898 in the upstream genomic region facing the insert, with the mated read matching the reverse sequence of the ST111 scaffold beginning at base 347 near the T-

DNA right border, again indicating that the right border inserted first in the same fashion as the other two events. One read spanned the genomic to T-DNA junction, revealing the exact insert base at position 12,332,979 (figure 4.6F).

Primers were designed according to the description provided above for the ST111 genomic reads to amplify the right border junction sequence. Sequencing of the PCR product for the right border integration site yielded a 255 b.p. product which following alignment, revealed the ST111 integration site precisely at base 12,332,976 on chromosome 6 validating the original sequencing read that spanned this junction. The integration site was located between soybean genes Glyma.06G151200 and Glyma.06G151300, neither of which were differentially expressed, and analysis of surrounding genomic reads report no nucleotide deletions from adjacent exons. A total of 29 b.p. was deleted from the 5' leading strand following the right border repeat sequence, deleting the right border repeat from the transferred DNA in the same fashion as the other two events. Junction sequences for the left border were unable to be isolated by neither sequencing nor LongAMP PCR with multiple primer sets due to complications with the highly repetitive properties of the insert region, even though existing reads aligned across the T-DNA to within 90 bases of the left border repeat sequence. This prevented accurate evaluation of genomic base deletions observed in the previous two transgenic events, however downstream reads of the insertion site (base 12,332,976) re-established alignments at genomic base 12,332,993 on chromosome 6, indicating a maximum possible genomic deletion of 17 bases. Right border primers were used in PCR homozygosity screening reactions for all 9 ST111 individuals, with all returning identical size bands (figure 4.6C). Primers used in all junction PCR reactions are described in

table 4.1. Alignments of the junction sequences for each event, including the products from each PCR primer sequencing reaction, are illustrated in figure 4.7.

Reconstructed alignments with the *Agrobacterium* genome and vector backbone sequences added to the reference as additional scaffolds yielded several read matches to both vector and *Agro* sequence in ST111, indicating the possible presence of further uncharacterized and unknown inserted sequences in the genome as opposed to a deletion at the insert site. Twelve total reads matched the *Agrobacterium* genome on chromosome 6 in ST111, although none were within or near the region where the T-DNA was identified. All attempts to obtain paired reads across the ST111 left border junction, including *de novo* assembly and long amplification PCR of the region, were unsuccessful.

4.4 Discussion

Soybean is one of the richest natural sources of protein known, which accounts for ~40% of seed weight. Herbicide resistant varieties have expanded to 94% of total soybean cultivation in the United States totaling nearly 30 million hectares, and some countries grow these varieties exclusively [174]. While many metabolomic and proteomic studies had been conducted on Glycine max, bioinformatics analyses were limited prior to the sequencing and publication of the soybean genome in 2010 [74]. Linkage mapping and phylogenetic analysis coupled with new data from the soybean genome indicate multiple whole genome duplication events at ~59MYA and ~13MYA [74, 175-177], classifying *Glycine max* as an ancient tetraploid. Prior to our previous investigation, transcriptomic alterations in a plant system expressing high amounts of recombinant protein had not been previously documented. Here, we expanded on our original work, reporting internal transgene alternative splicing post transformation,

transgene insertion locations and the surrounding genomic properties, as well as polymorphism rates and the possible resulting exome-wide translational changes.

Alternative splicing is an important mechanism for introducing molecular diversity in gene products of eukaryotes, and can occur within coding and non-coding regions of gene sequences. These post-transcriptional alterations to pre-mRNA can occur through retention of introns, skipping of exons, or alternative 5' and 3' junction sites [178]. The frequency of alternative splicing events varies greatly between organisms, and has been demonstrated to be dependent on many factors including gene structure, GC content, intron number, exon length, histone modifications, tissue developmental age and gene transcript levels [179]. Over the past decade as next generation sequencing technologies have expanded, numerous studies have revealed details of splicing tendencies in plants, indicating that intron retention is the most common mechanism of alternative splicing in higher order plants such as *Arabidopsis* [180] and soybean [181]. Furthermore, recent studies have demonstrated that ~60% of genes containing intron sequences are alternatively spliced in plants [182], with 63% of soybean genes containing multiple exons being alternatively spliced [181].

Upstream steps in transgenic plant biotechnology aim to maximize expression and transgene tolerance in the host organism by codon optimization. This aims to prevent CpG methylation-triggered gene silencing and to reduce the possibility of the gene of interest containing unintentional splicing signals. The Tophat2 pipeline automatically detects splice sites, indels, and possible fusion points in transcripts by examining flanking sequences of flagged low quality or unmapped reads, which are then concatenated to produce a novel transcriptome for Bowtie2 to re-align possible spliced sequences [53].

This allows reads crossing splice junctions that previously could not be aligned, or were incorrectly aligned to adjacent introns that were subsequently removed during splicing.

In our three experimental groups of transgenic soybean, Tophat2 identified spliced RNA-seq reads of the transgene sequence only in the ST77 group. Specifically, ST77-D and ST77-J progeny exhibited the splice junction, while none of the three ST77-F progeny contained the spliced reads. Interestingly, ST77-D events were the highest average expressors of the hTG transgene, supporting previous evidence that gene expression levels can possibly instigate higher tendencies for alternative splicing in plant tissues [181]. In addition, the splice junction occurred between guanine residues, which has been reported to be an uncommon splice variant in plants. Recent investigations of alternative splicing in soybean report ~97% of splice sites occur at GT^AG events, followed by 2.29% at GC^AG sequences, 0.23% at AT^AC sequences, and 0.31% of splicing events occurring from other types of junctions [181] such as the AG^GT site described here. Furthermore, intron retention has been identified as the primary mechanism for alternative splicing in plants [179], however the hTG junction appears to operate in the traditional sense of alternative splicing by removing a predicted intron segment.

While the predicted overall structure of the polypeptide was not dramatically altered between spliced and unspliced sequence variants, functional consequences of this effect remain unknown. Although the majority of exposed epitopes should remain constant, the predicted external locations of these modifications could impact antibody-based detection and quantification assays. Cases in which plant systems are expressing functional recombinant signaling proteins such as hormones, cytokines, or antibodies,

unpredicted small deletions or unintended splicing may alter the structure of the final peptide product substantially enough to render it non-functional or inhibitory in its intended signaling cascade.

Although recent next generation sequencing studies have revealed much about alternative splicing mechanisms in plants, rare and unexpected splice sites may still be present in optimized transgene vector sequences designed for expression in a particular plant system. As vector design and prediction software algorithms constructed for maximizing transgene expression in specific hosts improve, we can expect that these instances will be further reduced. However, the investigation described here demonstrates that these processes are still imperfect with the presence of predicted low probability splice sites and the absence of junctions predicted to be assured. It should also be noted that all transcriptome sequences used in this study were isolated from R8 stage cotyledon seed tissue, which contains ~37,000 genes [183]. Tissue-specific soybean transcript analyses show evidence that alternative splicing occurs more frequently and with a higher frequency in rapidly developing tissues [181], which would likely classify dried seed tissue as a low splicing frequency candidate. ST77 was the highest expressing transgenic event, and indeed the only one to exhibit alternative splicing of the transgene. However, the 764 line was very similar in total transgene expression levels, and no splicing was detected in any reads across the transgene region. Higher stress levels have also been shown to increase alternative splicing frequency [184], however the ST77 line was the only transgenic event to exhibit transgene alternative splicing, and was the most similar to wild type of all three transgenic events through our examinations. The 764 line, which exhibited stark differences to wild type in

all analyses, contained no splicing of the transgene. Recent soybean splicing investigations have suggested that genes with longer introns, a higher number of exons, and overall longer lengths exhibit more frequent instances of alternative splicing [181], suggesting that the splicing event observed in the ST77 hTG transgene may have been due to the sheer size and complexity of the transgene open reading frame itself. Global alternative splicing rates across all detected expressed genes in all transgenic events may reveal significant differences in splice types and frequency, and will be examined in future studies. For a recent comprehensive review on alternative splicing mechanisms in plants, see Reddy et. al [182].

SNP rates have been evaluated in soybean previously using multiple fragment analysis on several genotypes of cultivated varieties, including Lincoln, Mandarin, Peking, Richland, and others [185, 186]. From this fragment analysis, transition and transversion rates were reported nearly identical at 48% and 52% respectively. Furthermore, nucleotide diversity rates of cultivated soybean varieties were estimated to be 5-8 fold less than the wild variety *Glycine soja* and occurring at an even lower frequency than the highly characterized *Arabidopsis thaliana* self-crossing model [185]. More recent investigations of SNPs and INDELS in soybean using next generation re-sequencing have revealed genome-wide polymorphisms in efforts to identify disease resistant and favorable trait loci [187, 188]. Further studies demonstrate the extensive narrowing of soybean genomic variation due to domestication selection pressures for more valued traits, such as increased seed mass and oil content quantitative trait loci [189], as well as soybean's self-pollinating nature. SNPs detected in our datasets were divulged from seed transcriptome sequences, which represent ~6.5% of the total soybean

genome, and 65% of total genomic protein coding sequences. Due to this focused targeted approach, SNPs demonstrated here are not considered to be representative of the frequencies that may be detected in other tissues, or in different developmental stages of these specific transgenic plants. Although RNA-seq is focused on the functional segments of the genome and therefore doesn't always capture regulatory regions such as promoters or non-transcribed regions (e.g.: methylated bases), the resulting effects of polymorphisms in these regions can be directly witnessed through examination of gene expression levels. Because the soybean expression system described previously by our lab [5, 7, 9, 10, 36, 92] specifically targets seed tissue for recombinant protein accumulation, it was of great interest to identify possible sequence alterations that may have resulted in gene expression or protein structure variations in seeds. Although whole-genome sequencing was conducted from soybean seed tissues as part of this study, the intended purpose was solely for transgene location identification, and thus the resulting genomic datasets were not used for this purpose, as they did not produce adequate coverage for us to confidently predict SNP calls.

Base changes for wild type, ST77, and ST111 events were all comparable, while the 764 event consistently reported SNP rates nearly double that of the other groups (1 detected every ~22,000 bases), which is still well below the previously reported SNP rate of 1 SNP per ~1,400 nucleotides in soybean seeds [190]. Nevertheless, SNP base changes appeared to follow the same pattern of commonality, with all groups demonstrating high percentages of transition base changes with relatively low transversion counts. The 764 line did demonstrate a lower Ts/Tv ratio than the other groups, although progeny from two independent parents in the ST77 line (ST77D and

ST77F) exhibited lower ratios as well, indicating that this alone cannot reliably indicate internal stresses or overall divergence from our controls. The majority of detected base changes were located in the third base of the codon, likely generating synonymous (silent) polymorphisms. Less common changes occurred in the first or second bases of the codon, generating a non-synonymous or missense mutation likely resulting in an amino acid change. Missense to silent ratios detected in all experimental groups appear higher than previously published results for soybean indicate [190], although the higher nonsynonymous to synonymous mutations may be due to high linkage disequilibrium in soy [77]. Interestingly, the 764 line had the lowest overall missense to silent polymorphism ratio, demonstrating that while the 764 line contained the highest overall number of SNPs, a higher percentage of the total polymorphisms were silent mutations (~45%) compared to the other three experimental groups. This reveals the possibility that SNPs and INDELS detected here had originated from the original transformation event and have been highly conserved through self-crossed generations of progeny, particularly considering how remarkably well conserved the attributes of detected SNPs were between different seeds of the same transformation event. Without sequencing data from prior generations of each parental line tested here for comparison, this cannot be confirmed with complete confidence; however this does pose an interesting inquiry that even silent SNPs may have the ability to alter gene expression and peptide structure. Indeed, it has been suggested that although synonymous polymorphisms code for identical amino acids, subtle changes resulting from the utilization of non-optimal synonymous codons can modulate transcription and translation rates [191], thereby altering the folding conformations of the final peptides.

While many SNP studies have excluded synonymous mutations and those present in non-coding regions, recent works have examined these sequences in light of their possible epigenetic effects on transcription and translation; however more advanced tools for their effect prediction are still in development and require further testing for reliable forecasts [192]. Non-synonymous polymorphisms appear at similar rates across the experimental groups, varying between 50-60% of total functional SNP classifications. While non-synonymous mutations nearly always alter amino acid sequences that can potentially yield non-functional protein products, this phenomenon may produce a neutral effect on the organism or protein, or activate signaling cascades in redundant systems [192]. The effects of these detected non-synonymous polymorphisms in each transgenic event is not known or identifiable without extensive proteomic investigations into the altered downstream transcript products. The 764 line, which previously displayed the largest degree of differential gene expression, exhibited the lowest percentage of missense and nonsense polymorphisms; therefore these are not likely to be the root cause of observed changes in this transgenic line.

Spontaneous mutations may also occur as a result of a previously occurring stressful event. Epigenetic alterations and nucleotide transpositions have been detected up to five generations forward from a stressful occurrence in *Arabidopsis* [193], however all seeds examined here are bred to or beyond the 5^{th} generation of progeny. Segregation of SNPs from existing parental heterozygous loci is expected in progeny, however all samples examined here were remarkably consistent in SNP numbers, distributions and types within their experimental groups, further demonstrating the genomic stability of these events. This further solidifies the possibility that the detected variations in all

transgenic lines arose from an early parental event, occurring during the original *Agrobacterium* transformation or planting event, that have been stably integrated and carried forward to the current generations described here.

Detected alterations occurred in a "bowtie"-shaped distribution across the 20 chromosomes with lower instances of SNPs near the centromeres and the highest incidence of SNPs near the telomeric ends of the chromosomal arms where gene density is the highest, which has been illustrated previously by several works conducted on soybean polymorphism rates [187, 189, 194]. Our snpEff dataset summaries available in the supporting data section of this manuscript show the same distribution pattern, with no significant alterations of SNP rates on the chromosomes containing the T-DNA inserts in any of the transgenic lines compared to wild type. Without a specific focus for targeted SNPs in this work, future directions can invoke functional characterizations to cross-reference detected polymorphisms with possible connections to differentially expressed genes and also to improve existing annotated Williams 82 cultivar SNPs. In order to fully deduce the origin of SNP variations with high confidency, multiple generation analyses will need to be conducted. Approximately 70% of the genes that were differentially expressed also contained SNPs or INDELS, however due to the extreme disparity in size between each gene list (a maximum of ~3,000 and ~40,000 genes respectively), this is likely just due to chance and does not hold biological significance.

Gene ontology terms from SNP gene lists of all experimental groups were very similar, returning many enriched terms regarding protein transport and localization as well as RNA and transcriptional processes even in the wild type group. Interestingly, these relationships are highly similar to the enriched GO terms derived from the

differentially expressed gene sets of the transgenic lines, indicating the majority of functional relationships between these polymorphisms are likely arising from intercultivar variations and are not specifically related to transgenesis effects.

Identification of the T-DNA genomic insertion locations revealed no disruption of endogenous gene sequences currently annotated in the latest Phytozome records, and all three independent transformation events were located in gene rich euchromatic regions (see figure 4.8). Although *Agro* has been shown to preferentially target gene-rich areas to maximize expression of virulence factors present on the wild type Ti plasmid, it is currently regarded as a completely random process due to recent investigations revealing inadvertent selective pressures placed on higher expression of vector marker genes in previous reports [24]. Indeed, these selected transgenic events were carried forward through multiple generations after being screened for adequate expression of the BAR selectable marker cassette conferring an herbicide resistance trait and fully intact transgenes. Therefore, previously discarded transgenic progeny showing questionable resistance to herbicide may have contained transgenes integrated in regions with lower overall expression that were removed through segregation.

With the exception of the left border junction sequence of the ST111 transgenic line, all junction sequences were identified revealing intact transgene inserts in all events. Because the left border sequence could not be identified, step-wise PCR analysis of the ST111 hMBP transgene revealed a complete insert with no truncation of the transgene from the gene of interest through to the selectable marker open reading frame (data not shown). The particular locus at which the ST111 transgene was located was highly repetitive and "A/T" rich at the inferred left border junction, with many stretches of poly-

A and poly-T repetitive sequences in excess of 15 nucleotides. With the lack of paired read information for the 3' end of the insert, it was unknown as to what genomic modifications had occurred at the junction site with reference to the ability of *Agro* to insert randomly sized sequences from its own chromosome or the soybean genome in addition to sequence deletions of varying length. RNA-seq reads from the entire vector alignment for ST111 did indicate low expression of some backbone vector and aada region sequences, suggesting that incomplete cleavage of the left border repeat sequence had occurred and unknown lengths of vector and genomic DNA may have integrated at this bridge point preventing adequate read pairing by Bowtie and successful amplification with conventional PCR. *De novo* genomic assembly was also unable to bridge this site, presumably due to lower read coverage and the high complexity and redundancy of the soybean genome. Long amplification PCR was also unsuccessful despite designing many primer sets upstream, downstream, and within the T-DNA vector. Many non-specific products were generated and unable to be resolved despite multiple optimization attempts, indicating possible complex secondary structures or lengthy repetitive insertions of unknown origin. In addition, one recovered genomic read contained a short segment of vector sequence which was inverted and reversed, immediately leading into downstream genomic sequence in the proper 5' to 3' orientation of the leading strand revealing possible complex genomic rearrangements at this junction site.

Based on previous reports, bacterial chromosomal sequence insertion from *Agrobacterium* can possibly integrate into the plant genome along with the T-DNA [195]. Subsequent alignments to the ST111 reference genome supplemented with the complete *Agrobacterium* genome sequence yielded hundreds of matches dispersed throughout the

genome on every soy chromosome with the exception of chromosome 14. All matches from chromosome 6 where the gene of interest had integrated were not located in the same vicinity, indicating that chromosomal insertion of *Agro* DNA was likely not contributing to the inability to amplify the left border junction. Furthermore, *Agro* chromosomal DNA is remarkably similar to the chloroplast genome sequence of *Glycine max*, sharing more than 95% sequence identity in some instances, thereby making unique alignment predictions difficult. In order to fully characterize the structure and sequence of this site, advanced genome walking techniques may be required.

4.5 Conclusions

Previous equivalence studies on transgenic crops have discussed many pleiotropic nutritional and molecular alterations that have been detected across a multitude of different species [8, 86, 136, 196-202]. Transgenesis has the potential to induce possible perturbations through multiple avenues, including internal gene disruption, gene expression regulation through truncated transcripts, transgenic protein interactions with endogenous peptides, or disruptions of internal homeostatic balance due to molecular properties of the expressed transgenic protein. Pleiotropic effects from transgenesis are of particular concern for agricultural biotechnology, especially if there is the potential for complicating future deregulation attempts or negatively impacting the accumulation of pharmaceutical proteins in the plant system. The results discussed here expanded on our previous work, examining three different independent transgenic soybean events, one of which exhibited significant gene expression changes when compared to wild type. While the insertion of foreign genetic material has the capability of inducing unintentional consequences, we are unable to directly infer what caused the observed gene expression

and polymorphism changes in the 764 transgenic line. However, none of the transgenes disrupted any currently annotated soybean genes, and SNP effect predictions show very similar patterns between all three events, although at different rates. The 764 line also exhibited the largest genomic deletion of 1,625 nucleotides at the insertion site. Unexpected splice sites were also detected in the transgene of one of the events despite pre-transformation codon optimization of the vector, which can have profound implications in plant systems producing biologics. The provided datasets from these series of investigations can help to assist in correcting errors in the soybean reference exome, optimize the design of transgene vector sequences for soybean, as well as providing valuable insight into *Agrobacterium* T-DNA integration characteristics and possible perturbations of native sequences resulting from transgenesis.

4.6 Availability of Supporting Data

The RNA sequencing data described herein may be accessed at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64620 through NCBI's Gene Expression Omnibus under accession number GSE64620. Genomic sequencing FASTQ files for each transgenic sample are available through NCBI's Sequence Read Archive under the ST77-KP2 experiment accession number SRX1143641. Summary files for bcftools and snpEff outputs encompassing variant calls and predicted variant effects are available from the iPlant collaborative Discovery Environment directory available at http://de.iplantcollaborative.org/dl/d/B4D75710-BA97-4CE1-A12C-FAEE129FF2A4/snpeffsummaries.zip and http://de.iplantcollaborative.org/dl/d/AD464040-9169-4E52-87D5-DEF7466A2C1E/variantsummaries.zip. Lists of genes containing detected variants from

all samples are available from iPlant in a compressed .zip archive available at http://de.iplantcollaborative.org/dl/d/E173CCFB-D7E6-41C1-90E4-244FBC47648B/snpeffgenes.zip.

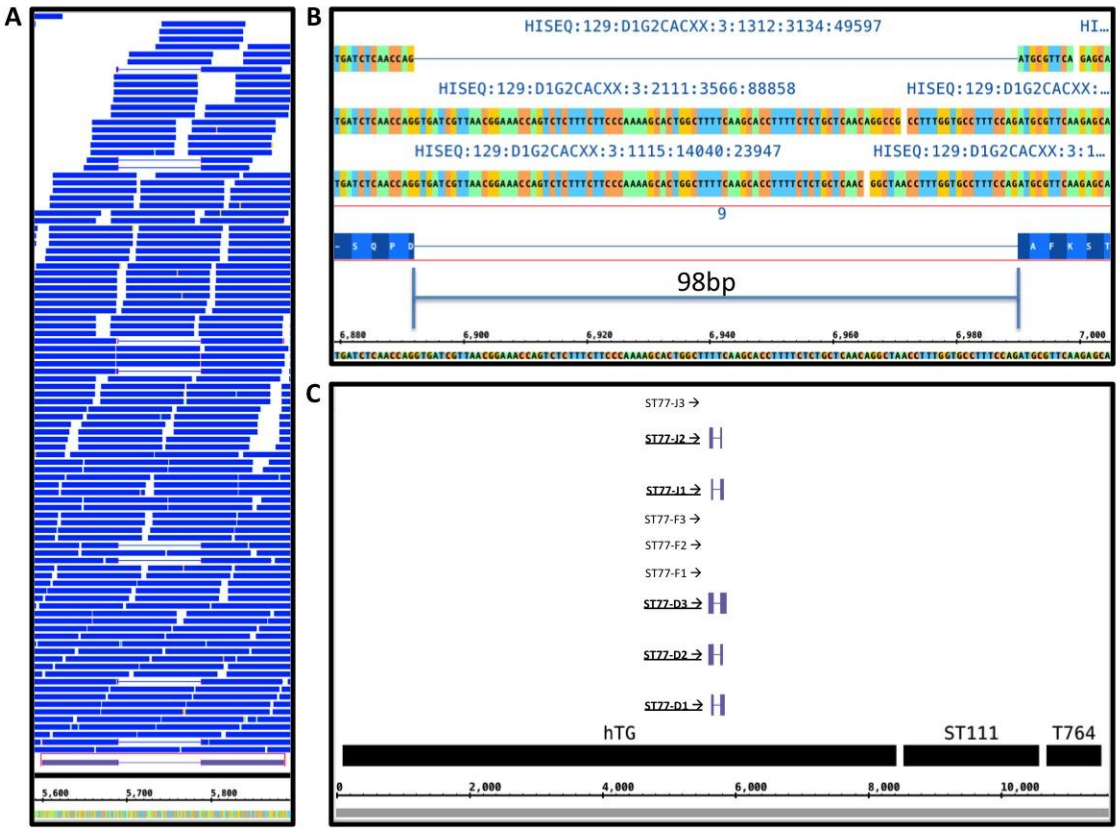


Figure 4.1: Coverage graph and alternative splicing site for ST77. (A) Coverage of the RNA-seq reads for the ST77D3 transgenic sample show the slice junction detected by TopHat, illustrating the high coverage of this area of the transgene. (B) shows an expanded view of the actual splice junction, containing spliced and unspliced reads. The 98 base pairs removed during the splicing event are shown, as well as the amino acid sequence generated as a result of the splice. Alignments showing alternative splice sites for all three transgenic constructs are shown in panel (C), in which the ST77 transgene was the only one to demonstrate alternative splice junctions in two parental lines, ST77D and ST77J.

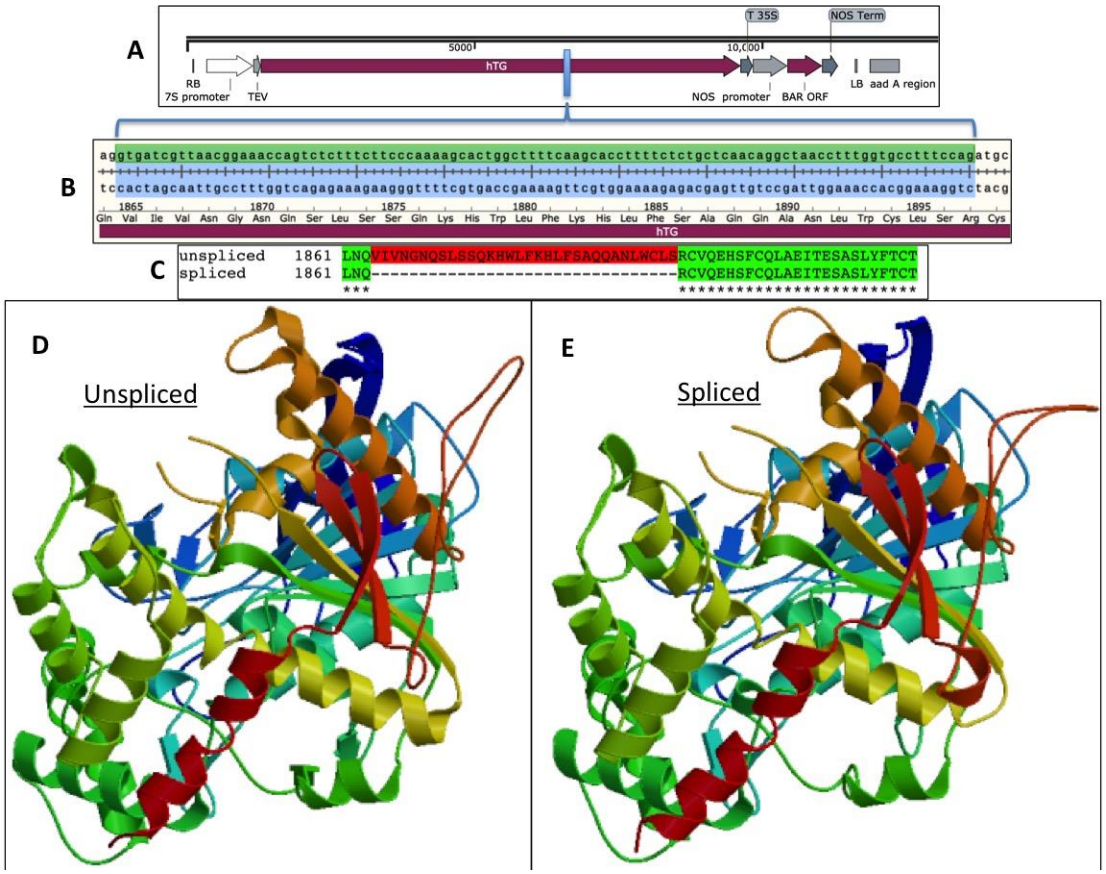


Figure 4.2: Map of ST77 T-DNA deletion site and predicted protein structure changes. (A) The ST77 hTG T-DNA construct with the indicated splice region in blue. Bases deleted from the splicing excision for both the leading strand (green) and complimentary strand (blue) are shown in (B). The consensus amino acid sequence is shown in (C), with both the unspliced native and spliced amino acid sequence aligned together with the spliced out region highlighted in red. ExPASY SWISS-MODEL structural predictions of the native unspliced (D) and alternatively spliced (E) thyroglobulin proteins based on the amino acid sequence.

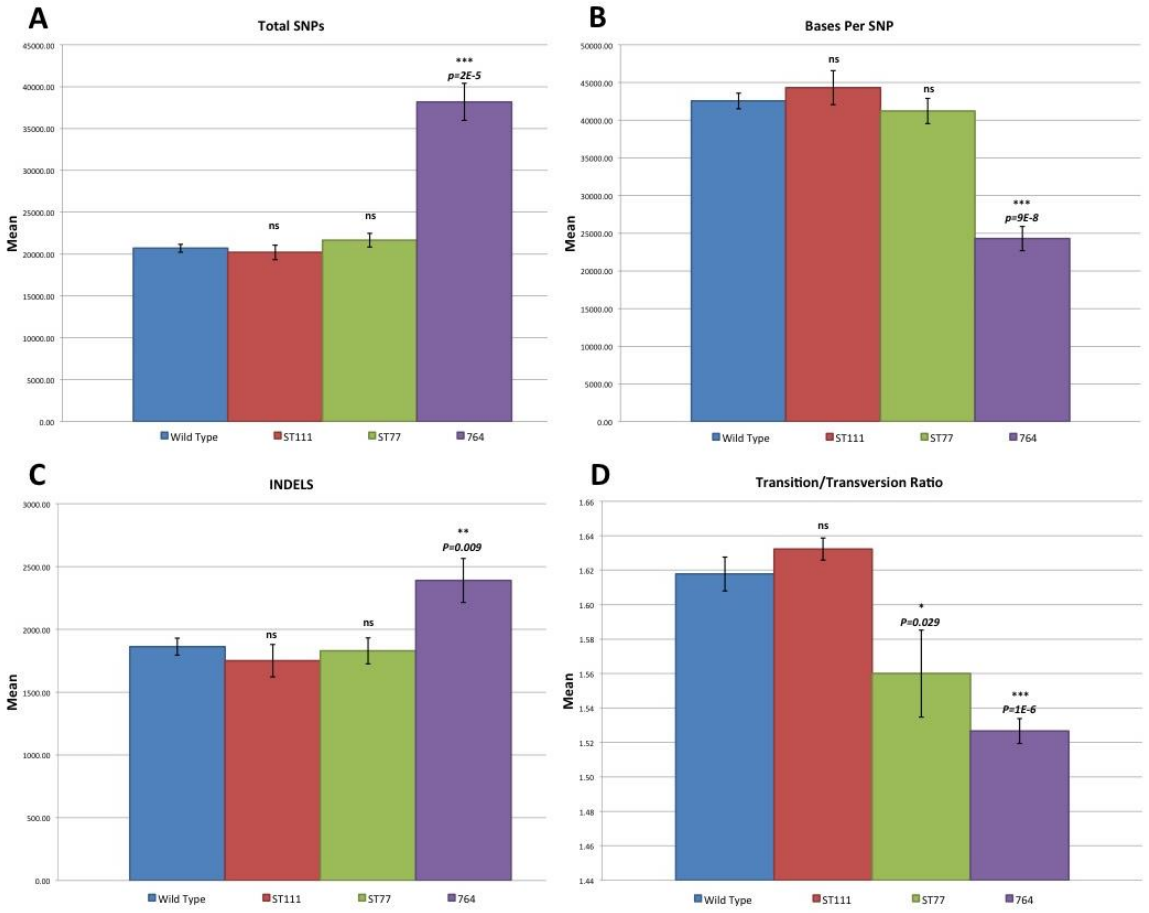


Figure 4.3: Summary of SNP and INDEL counts. Means of total polymorphisms (A), polymorphism rates (B), insertions/deletions (C), and transition to transversion ratios (D) were calculated all four experimental groups. Standard error bars as well as statistical results from unpaired *t*-tests between each group and wild type are shown. Groups marked with an asterisk indicate significance ($p<0.05$=*, $p<0.01$=**, $p<0.001$=***). Wild type is shown in blue, ST111 in red, ST77 in green, and 764 in purple.

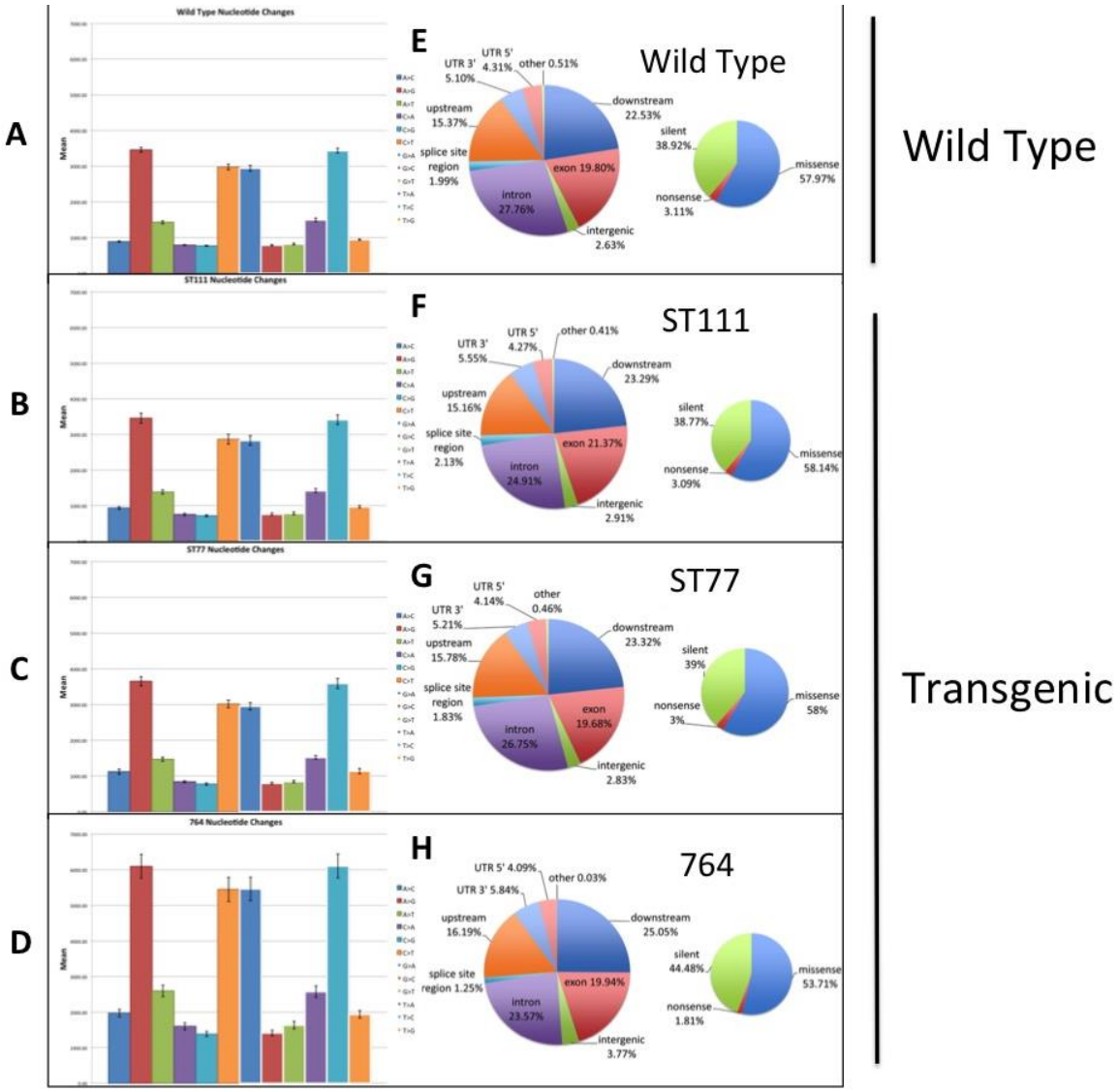


Figure 4.4: Polymorphism nucleotide base changes and predicted effects. Mean rates of change for each nucleotide base are shown for wild type (A), ST111 (B), ST77 (C), and 764 (D), with bars indicating the standard error of the mean. Regions containing detected polymorphisms and their functional classifications are shown for wild type (E), ST111 (F), ST77 (G), and 764 (H).

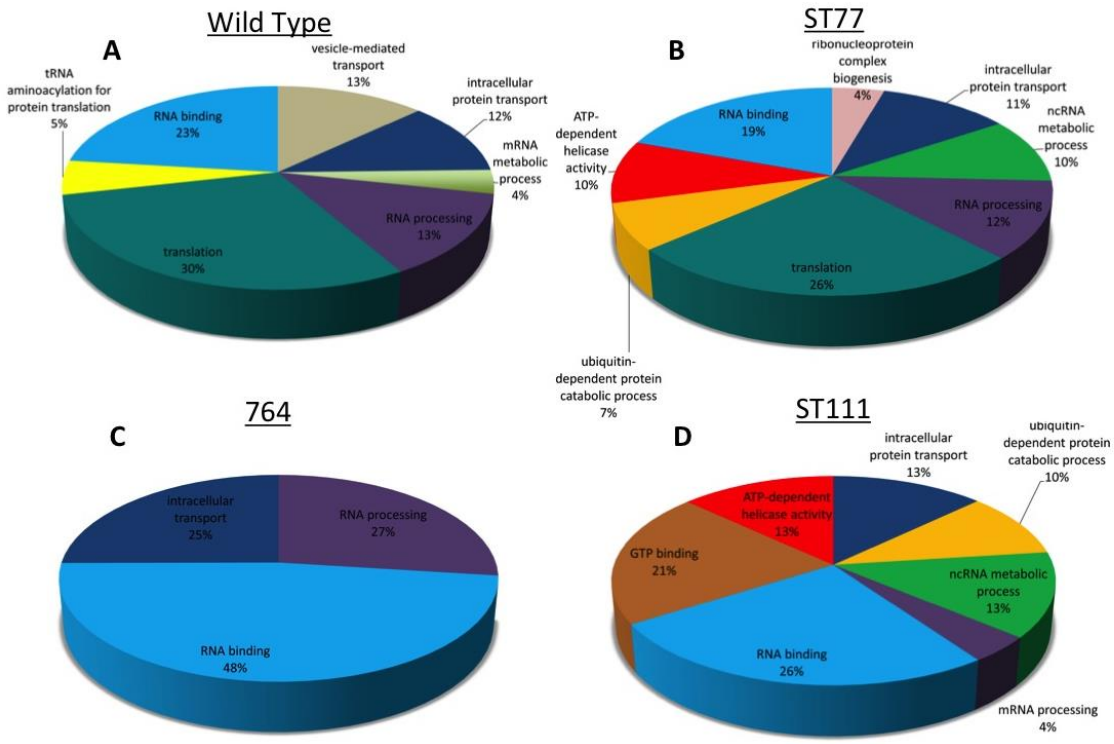


Figure 4.5: AgriGO enriched gene ontology categories of SNP-containing genes. Single enrichment analyses for all genes containing SNPs are shown for wild type (A), ST77 (B), 764 (C), and ST111 (D). Percentages are reflective of how many terms matched the indicated GO group out of the total matches to the background reference.

Table 4.1: Primer sequences, product sizes, and their attributes used for T-DNA junction amplification.

Primer	Sequence	Length	Tm (°C)	GC Content	Product Size
ST77 F1	5'-aggatgacccgacatgtctcttag-3'	24bp	59	50%	373bp
ST77 R1	5'-caaatgaagggcatggatcctgc-3'	22bp	60	50%	373bp
ST77 F2	5'-cggtttgcgtattggctagagc-3'	22bp	60	55%	530bp
ST77 R2	5'-gcccgtcctgagcctaaaattg-3'	22bp	60	55%	525bp
764 F1	5'-gtgccgtgtttcagaacatctcg-3'	23bp	58	52%	443bp
764 R1	5'-cttaggctaggatcctgcaggc-3'	22bp	59	59%	443bp
764 F2	5'-ccagctgcattaatgaatcggcc-3'	23bp	59	52%	444bp
764 R2	5'-ggatggcaaggcaagtagactc-3'	22bp	58	55%	444bp
ST111 F1	5'-gcaagaacaaaatgtccctgcgg-3'	23bp	59	52%	255bp
ST111 R1	5'-tggccgtcgttttacaacgtcg-3'	22bp	60	55%	255bp

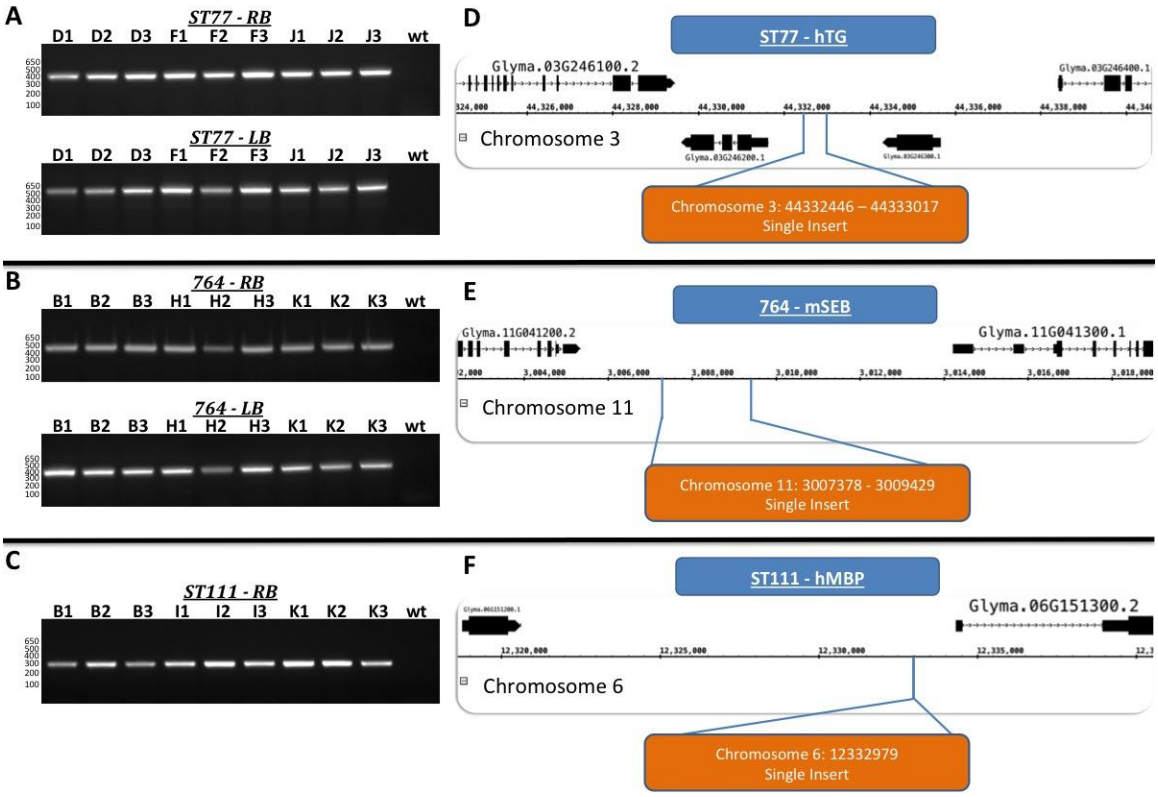


Figure 4.6: PCR of T-DNA junction sequences and insertion regions identified by paired-end sequencing. PCR products from the ST77 right border and left border (A), 764 right and left borders (B), and ST111 right and left borders (C) are shown. Design of the primers for each amplicon were derived from aligned discordant genomic reads, which narrowed the insertion sites for each transformant as shown in (D-F).

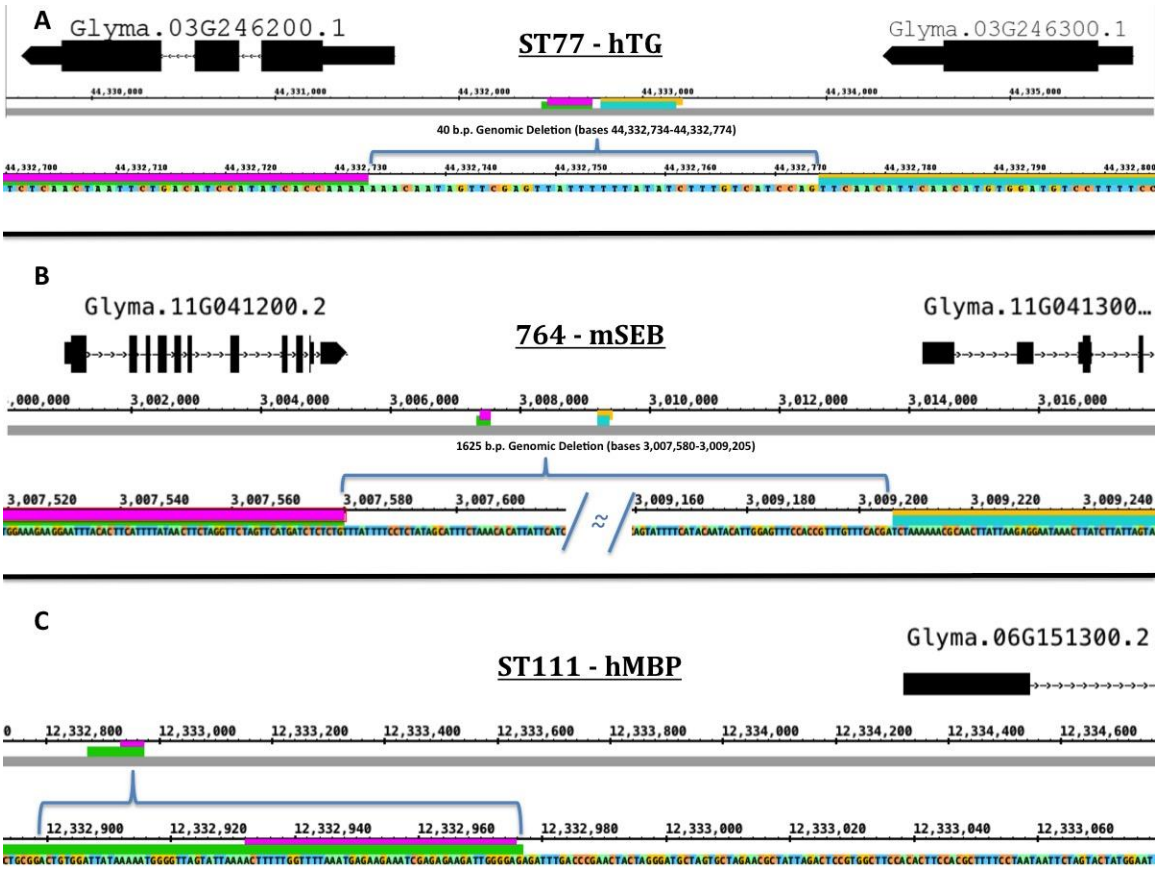


Figure 4.7: Aligned sequenced PCR products and insert layout for each transgenic line. (A) Colored bars represent sequences from the PCR amplicons of the junction sites that aligned to the soybean reference genome on each chromosome harboring the T-DNA insert. Purple is the product from primer F1, green from primer R1, yellow from primer F2, and blue from primer R2. (A) 40 bases of genomic DNA have been deleted as a result of the insertion in ST77, shown as the uncolored region between the primer products. 764 (B) had 1625 bases removed from the genome on chromosome 11. (C) ST111 exhibited a maximum deletion of 17 bases after the left border based on downstream genomic sequencing reads.

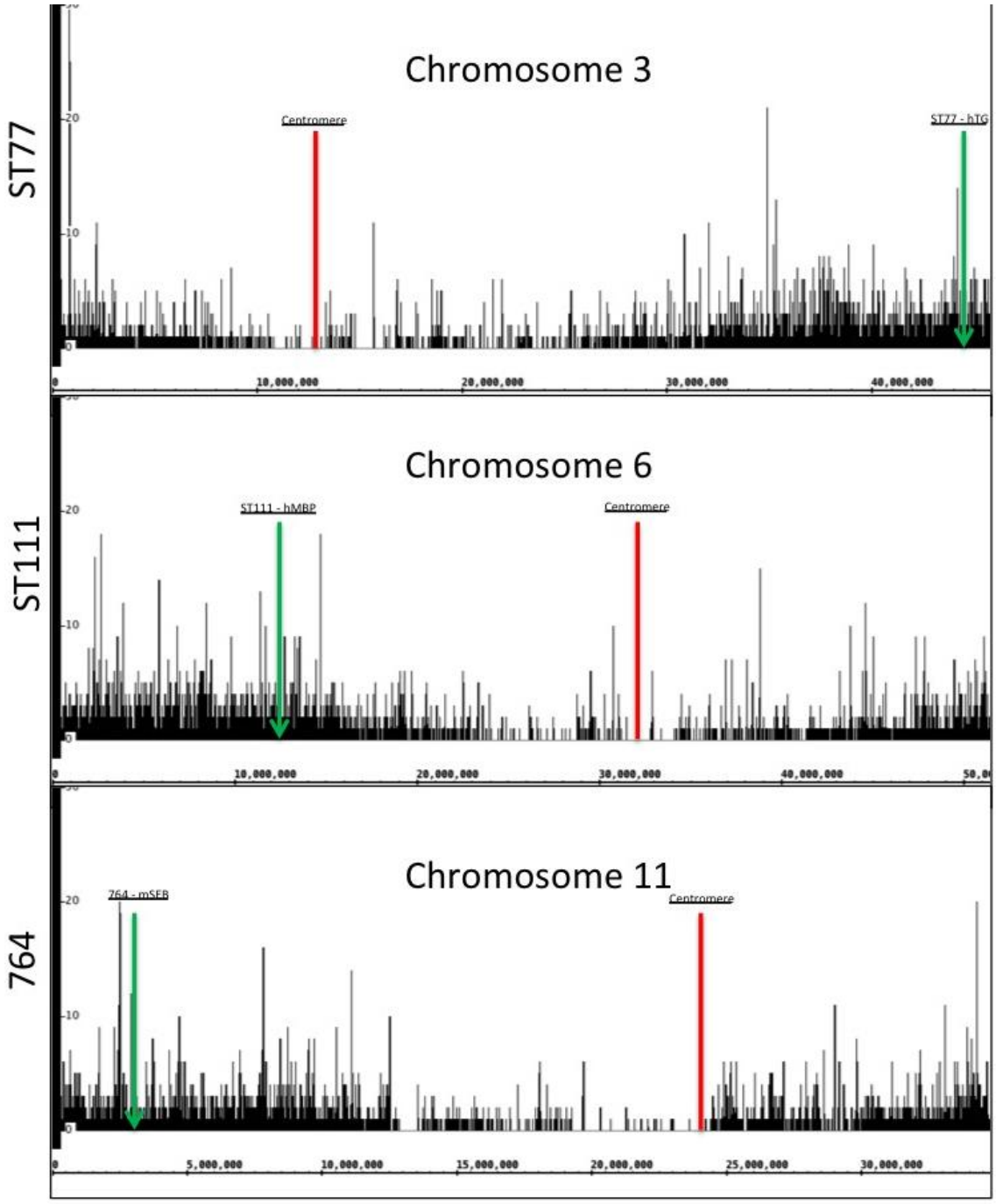


Figure 4.8: Gene density plots of chromosomes containing T-DNA inserts. Centromereic sequences are shown in red, and transgene insertion locations for each transgenic line are shown in green.

Table 4.2: Predicted leading strand donor splice sites and the matching acceptor site from NetGene2 in ST77.

Predicted Donor Splice Sites

pos 5'->3'	phase strand	confidence	5' exon intron 3'
1381	+	0.55	GATCTTGCTG^GTGTGTTGGA
1428	+	0.71	AGGCAACCAG^GTAGACCAGT
1585	+	0.55	TTCAATGCTG^GTGCGTTGAT
2226	+	0.54	AGGCAACAAG^GTGAACCACC
4318	+	0.9	GACACTGTTG^GTGCGTGGAT
4543	+	0.37	CTGGAACATG^GTGTGTTGAC
4735	+	0.62	GATCTTGTTG^GTGCGTGATG
5171	+	0.55	CCAGATCCAG^GTTAAGACTT
5358	+	0.55	GATCTTCTTG^GTAGGTTCAC
6758	**+**	**0.5**	**TCTCAACCAG^GTGATCGTTA**
9571	+	0.35	AGAATAATGT^GTGAGTAGTT
9862	+	0.94	GAGAAAGCAG^GTAGCTTGCA
9953	+	0.87	GCGCCCTCTG^GTAAGGTTGG
10855	+	0.7	CAGCCTGCCG^GTACCGCCCC
10915	+	0.31	GGATCCCCGG^GTACCGAGCT

Matching Acceptor Splice Site

pos 5'->3'	phase strand	confidence	5' intron exon 3'
6855	**+**	**0.15**	**GCCTTTCCAG^ATGCGTTCAA**

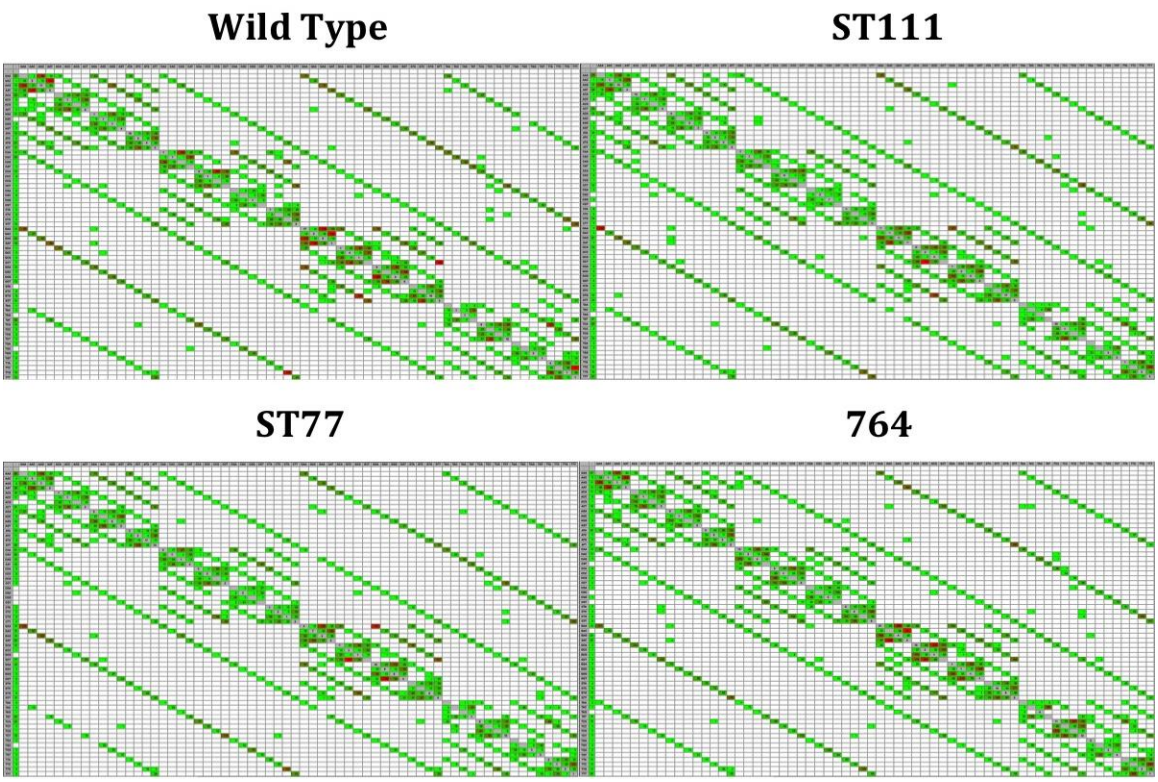


Figure 4.9: Heatmap of detected codon changes across wild type and all three transgenic groups. Green color denotes less frequent occurrence, and red denotes a higher occurrence.

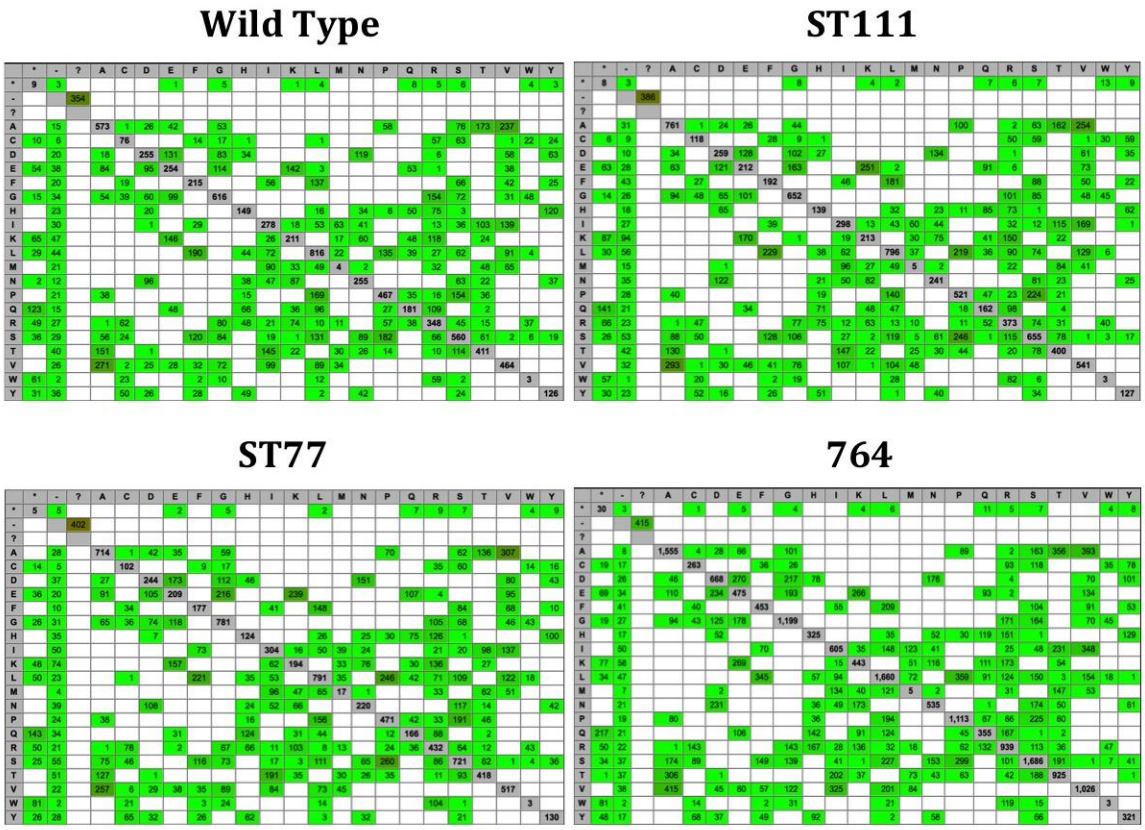


Figure 4.10: Amino acid change heatmap for wild type and the three transgenic groups. Darker colors represent a higher detected instance of the respective amino acid alteration.

Table 4.3: Raw SNP, base changes and deviation values for each sample.

Sample	SNPs	TS/TV Ratio	Indels	singleton snps	singleton ts/tv	singleton indels	multiallele sites	multiallele snps	SNP Rate AVG	A>C	A>G	A>T	C>A	C>G	C>T	G>A	G>C	G>T	T>A	T>C	T>G
WTB1	21096	1.61	1964	50.80%	1.7	57.00%	37	23	41479	889	3609	1497	811	782	3005	2973	816	814	1545	3442	936
WTB2	20388	1.63	1914	48.70%	1.72	57.01%	34	24	42874	859	3412	1435	794	727	2978	2883	748	818	1509	3377	872
WTB3	23047	1.61	2225	52.20%	1.68	58.30%	44	27	37826	972	3827	1661	825	840	3271	3348	876	902	1756	3782	1014
WTH1	21650	1.62	1957	48.10%	1.67	57.70%	43	25	40496	932	3464	1448	843	833	3194	3160	806	891	1530	3576	998
WTH2	20605	1.57	1786	46.70%	1.65	55.20%	31	19	42715	902	3361	1417	813	759	2972	2903	808	876	1493	3377	943
WTH3	21760	1.6	1928	46.20%	1.7	55.70%	41	29	40362	931	3601	1513	828	793	3139	3057	817	846	1612	3620	1032
WTL1	18643	1.65	1527	46.30%	1.74	55.10%	27	13	47416	820	3194	1230	747	699	2561	2607	725	697	1233	3245	898
WTL2	18640	1.6	1622	47.20%	1.72	56.40%	31	26	47179	823	3165	1245	716	709	2642	2603	718	723	1328	3088	906
WTL3	20536	1.67	1843	47.50%	1.74	55.80%	30	22	42704	864	3493	1404	757	776	3031	2938	767	794	1449	3391	894
Mean	20707.22	1.62	1862.89	0.48	1.70	0.56	35.33	23.11	42561.22	888.00	3458.44	1427.78	792.67	768.67	2977.00	2941.33	786.78	817.89	1495.00	3433.11	943.67
StdDev	1427.37	0.03	204.39	0.02	0.03	0.01	6.18	4.78	3117.09	51.60	208.49	132.34	43.03	50.37	236.85	239.45	51.22	71.41	151.82	206.07	57.95
StdErr	475.79	0.01	68.13	0.01	0.01	0.00	2.06	1.59	1039.03	17.20	69.50	44.11	14.34	16.79	78.95	79.82	17.07	23.80	50.61	68.69	19.32
ST111B1	21301	1.64	1984	49.40%	1.67	56.70%	40	26	41064	961	3577	1467	812	739	3121	2968	833	803	1500	3569	978
ST111B2	22297	1.61	1997	49.20%	1.68	59.20%	44	34	39331	1034	3676	1553	840	787	3228	3172	808	882	1562	3714	1076
ST111B3	20363	1.64	1909	49.00%	1.71	57.20%	40	30	42417	911	3487	1470	767	743	2970	2959	772	828	1386	3434	940
ST111I1	19005	1.65	1510	44.70%	1.77	56.90%	26	21	46636	877	3253	1271	705	667	2739	2724	689	724	1348	3143	888
ST111I2	18356	1.64	1530	45.70%	1.79	54.50%	32	26	48097	821	3185	1266	671	662	2556	2592	684	662	1328	3079	876
ST111I3	21702	1.66	1981	44.70%	1.68	56.00%	38	29	40444	997	3899	1471	788	715	2911	2952	779	857	1556	3805	1002
ST111K1	19709	1.63	1450	47.20%	1.73	56.10%	24	20	45232	965	3386	1355	706	715	2661	2688	731	709	1334	3487	993
ST111K2	24097	1.62	2326	52.90%	1.68	59.10%	45	26	36242	1095	4056	1596	898	827	3534	3436	931	984	1703	3907	1156
ST111K3	15044	1.6	1065	45.40%	1.69	56.50%	14	11	59443	732	2597	996	540	538	2090	1976	562	563	1065	2597	800
Mean	20208.22	1.63	1750.22	0.48	1.71	0.57	33.67	24.78	44322.89	932.56	3457.33	1382.78	747.44	710.33	2867.78	2829.67	754.33	779.11	1420.22	3415.00	967.67
StdDev	2615.63	0.02	385.91	0.03	0.04	0.01	10.46	6.72	6787.46	111.19	429.78	184.87	105.96	83.17	420.05	411.12	104.85	127.46	184.86	413.65	107.68
StdErr	871.88	0.01	128.64	0.01	0.01	0.00	3.49	2.24	2262.49	37.06	143.26	61.62	35.32	27.72	140.02	137.04	34.95	42.49	61.62	137.88	35.89
ST77D1	23491	1.5	2145	51.60%	1.54	58.40%	45	32	37285	1305	3891	1640	897	821	3304	3084	860	903	1544	3841	1434
ST77D2	22880	1.52	2016	49.80%	1.55	58.40%	46	29	38410	1282	3898	1531	845	808	3070	2974	808	899	1605	3894	1295
ST77D3	24214	1.51	2202	52.50%	1.55	56.20%	43	28	36220	1345	4058	1653	964	859	3323	3242	877	957	1703	3959	1304
ST77F1	18610	1.51	1396	46.00%	1.56	56.40%	28	21	47788	1045	3125	1244	740	644	2569	2476	692	729	1279	3029	1061
ST77F2	20530	1.51	1624	49.50%	1.56	57.80%	26	16	43163	1112	3446	1406	798	701	2789	2748	784	809	1482	3388	1085
ST77F3	24879	1.51	2110	50.10%	1.56	56.50%	52	37	35419	1362	4066	1688	948	919	3421	3369	915	983	1752	4130	1365
ST77J1	18520	1.64	1519	48.90%	1.76	54.90%	23	19	47715	806	3150	1257	707	688	2655	2660	699	714	1358	3057	791
ST77J2	22737	1.68	1906	47.10%	1.74	58.20%	45	31	38769	969	3923	1496	865	823	3244	3279	818	890	1619	3841	1003
ST77J3	19129	1.66	1540	49.40%	1.76	58.60%	33	28	46259	842	3292	1300	770	681	2758	2735	708	696	1342	3159	876
Mean	21665.56	1.56	1828.67	0.49	1.62	0.57	37.89	26.78	41225.33	1118.67	3649.89	1468.33	837.11	771.56	3014.78	2951.89	795.67	842.22	1520.44	3588.67	1134.89
StdDev	2493.25	0.08	309.83	0.02	0.10	0.01	10.47	6.78	5031.52	216.22	391.96	174.54	90.29	95.02	324.87	312.92	81.82	108.36	166.72	428.88	225.63
StdErr	831.08	0.03	103.28	0.01	0.03	0.00	3.49	2.26	1677.17	72.07	130.65	58.18	30.10	31.67	108.29	104.31	27.27	36.12	55.57	142.96	75.21
764B1	46578	1.55	3098	26.10%	1.67	41.90%	69	40	19267	2256	7393	3220	2012	1625	6736	6737	1662	2036	3198	7502	2241
764B2	26386	1.54	1566	27.30%	1.68	41.60%	56	32	34217	1345	4318	1752	1075	980	3677	3729	999	1109	1784	4309	1342
764B3	40747	1.55	2662	26.20%	1.67	43.40%	83	34	22032	1958	6508	2792	1694	1465	5922	5820	1582	1780	2755	6513	1992
764H1	35176	1.5	2144	26.50%	1.55	41.80%	67	39	25360	1835	5588	2383	1501	1356	4924	5009	1327	1494	2360	5627	1813
764H2	44328	1.49	2869	26.20%	1.56	41.60%	104	55	20267	2303	6989	3053	1904	1630	6385	6321	1646	1908	3019	6890	2336
764H3	35356	1.53	2027	26.60%	1.58	41.10%	86	56	25572	1948	5795	2379	1455	1241	4941	4953	1328	1473	2389	5724	1787
764K1	44693	1.51	2889	26.80%	1.61	42.80%	91	52	20090	2366	7049	3021	1872	1627	6360	6419	1634	1955	3066	7110	2267
764K2	31736	1.54	1855	27.90%	1.62	42.70%	68	45	28445	1712	5070	2119	1302	1122	4505	4539	1189	1330	2045	5161	1687
764K3	38690	1.53	2400	25.40%	1.6	40.90%	82	45	23277	2022	6132	2643	1615	1431	5552	5624	1390	1666	2566	6097	1998
Mean	38187.78	1.53	2390.00	0.27	1.62	0.42	78.44	44.22	24280.78	1971.67	6093.56	2595.78	1603.33	1386.33	5444.67	5461.22	1417.44	1639.00	2575.78	6103.67	1940.33
StdDev	6663.47	0.02	525.75	0.01	0.05	0.01	14.72	8.77	4792.92	321.76	1007.56	481.12	304.21	234.38	1010.03	984.46	231.90	311.33	480.39	1014.57	320.46
StdErr	2221.16	0.01	175.25	0.00	0.02	0.00	4.91	2.92	1597.64	107.25	335.85	160.37	101.40	78.13	336.68	328.15	77.30	103.78	160.13	338.19	106.82

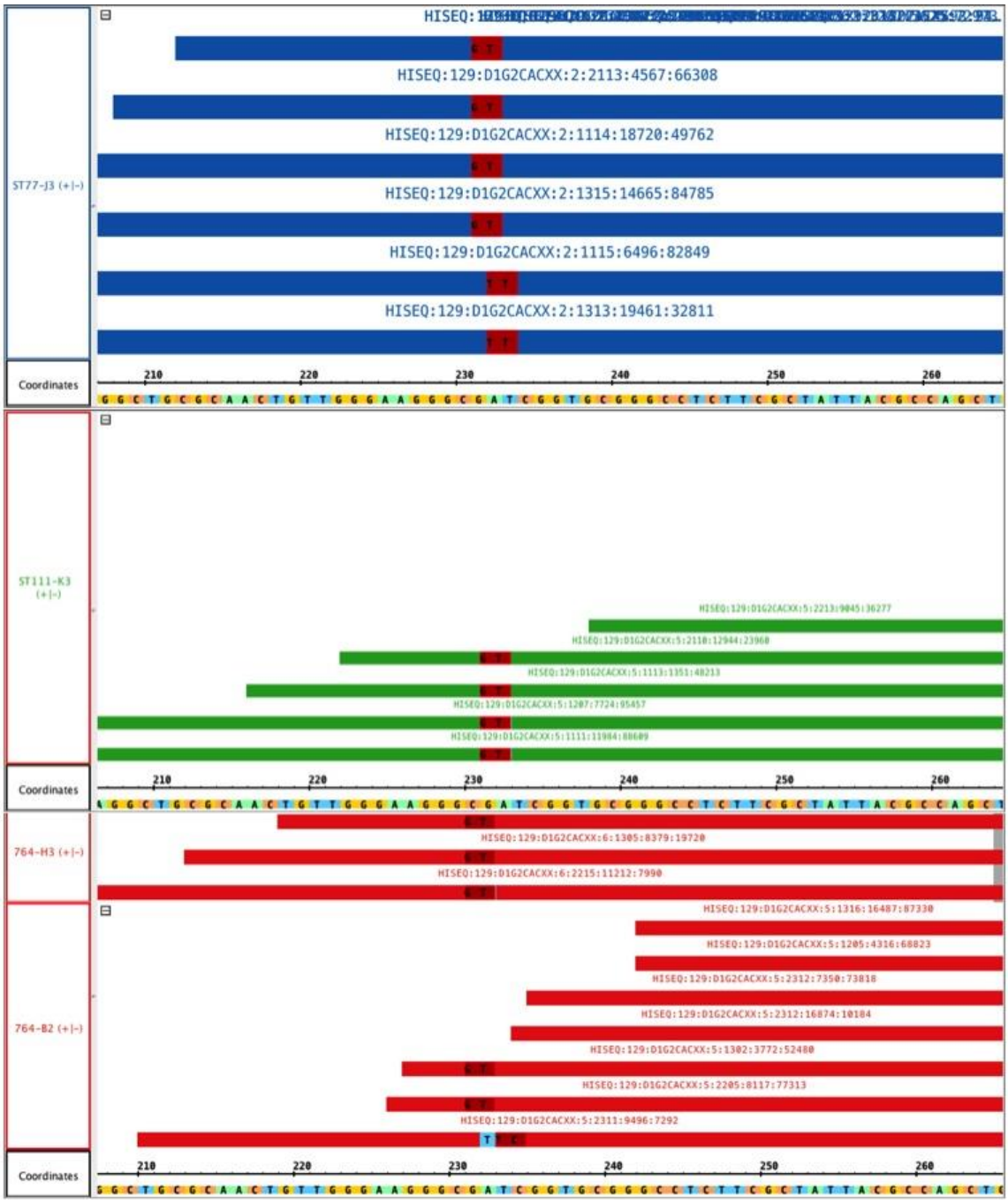


Figure 4.11: Base change controls in transgenes. Variants located at bases 232 and 233 were located in the padded sequence after the right border, and were detected consistently in all events with transcripts across this region, demonstrating the repeatability and consistency of the SNP calls.

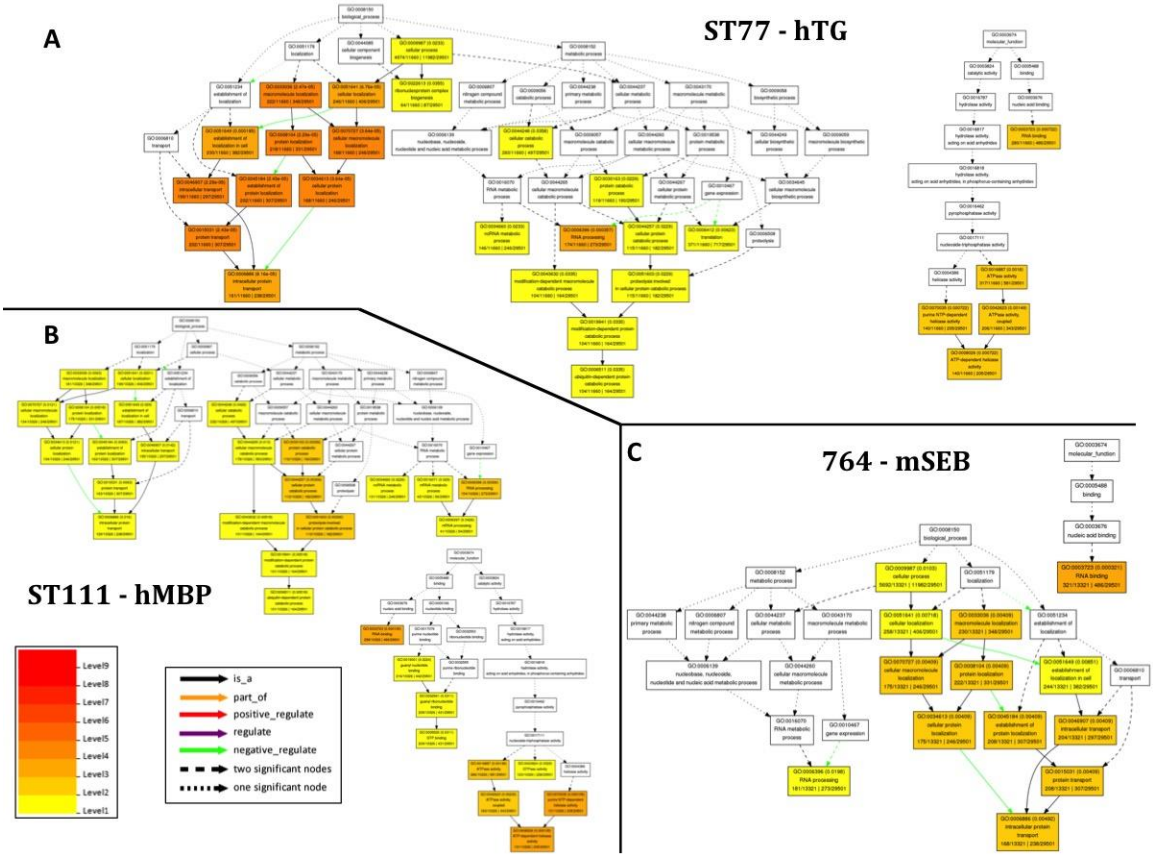


Figure 4.12: AgriGO single enrichment analysis results of the genes containing effectual SNPs in transgenic events. Darker colors indicate a higher level of significance for each node.

CHAPTER 5: DISSERTATION SUMMARY AND CONCLUSIONS

Over the past two decades, equivalence studies for transgenic plants have become quite prevalent and have covered many aspects of comparison. In nearly all cases, differences detected were within variation ranges for unmodified cultivars regarding gene expression, protein content, and metabolomic quantifications. Furthermore, based on variances detected from mutation and allele shifts that occur at a natural rate from generation to generation, it is quite unrealistic by the very nature of transformation to expect no modifications to be induced following transgenesis [198]. Various methods of transformation fluctuate in the levels of induced genomic changes, with particle bombardment typically being the least disruptive during mutagenesis [85]. However as mentioned previously, particle bombardment typically results in complex, multi-insertion events, complicating segregation during propagation that may induce post-transcriptional gene silencing.

Agrobacterium transformation-induced stress responses during regeneration typically do not persist further than one generation, making it the preferred method for applications that do not require transient expression of the T-DNA. Additionally, it must be noted that although the transgene may originate from an alternative organism, it is utilizing native processes by design to activate transcription, fold and assemble the final protein product. The recruitment and application of these factors in the generation of these peptides may cause cascading effects that although are

not harmful to the host, could be differential when compared to non-transformed varieties [203-205].

Several herbicide resistant varieties of crops (including triazine, phenoxy-carboxylic acids, nitriles, dinitroanilines, aryloxyphenoxypropionates, cyclohexanediones, sulfonylureas, imidazolinones, phosphinic acids, and glycines) in both the presence and absence of their respective compounds showed no specific pattern of variance. Approximately 50% of the tested varieties in corn, *Arabidopsis, Brassica rapa*, and others show a reduction in certain aspects of total crop yield; however many showed no difference, or in some instances, exhibited an increase in advantageous properties. Typical yield loss was between 15-20% depending on the variety, suggesting mutation and natural variations attributed to the measured variances, which does not factor in the potential increase of total crop yield resulting from additional fitness conferred by the transformed trait itself [136]. Independent corroborations of these findings reveal less than 1% of comparisons yield a significant measured difference of >20% across any measured criteria, showing more variation due to growing location and handling practices than transgenesis [206]. Health assessments from animal feeding trials to address concerns over the safety of human consumption have been rigorously reviewed by José L. Domingo describing transgenic potatoes, maize, corn, soybean, rice, cucumber, tomatoes, sweet peppers, peas, and canola plants [207].

Allergenicity assessments of transgenic edibles were of concern due to the possible formation of novel peptide structures to exacerbate existing allergic reactions. With the exception of the transgene product, all peptides generated in transgenic food crops are endogenous and native to that species. Therefore, this poses no public health

risk as any individual who was originally allergic to the product will likely be allergic to the transgenic alternative, and would already avoid those particular foodstuffs. Random interactions of recombinant proteins with endogenous ones in such a way to induce an allergic reaction is also an extremely remote possibility, as is transformation-dependent mutagenesis resulting in the upregulation of allergen producing genes [94]. Moreover, transgenic soybeans were recently described to contain very minimal novel peptides exceeding the World Health Organization's (WHO) threshold for potential allergenicity, the majority of which match currently documented allergens [208]. On the other hand, if a transgenic variety is engineered to express a potential allergenic protein, such as peanut, soy, or wheat products, proper sensitivity assays should be conducted to prevent inadvertent exposures to the allergic for food safety [209, 210].

Transcriptome comparisons in transgenic *Arabidopsis* also demonstrate transcriptional disequilibrium is less extreme than natural variance [211]. While some transcriptome changes seemed consistently linked to the transgenic species, unique variances for each cultivar make it impossible to distinguish between transgene exclusive effects and natural variations, further advocating for comparative equivalence between them [106]. Cheng *et al.* in 2008 used microarray transcriptome profiling of the first trifoliate leaves of five different transgenic and non-transgenic soybean cultivars, including Bayfield (University of Guelph), S03-W4 (Syngenta), 2601-R (First Line), PS46RR (First Line), and Mandarin (Ottawa). While this was specifically targeted to leaf tissues, it was the first exome comparison of transgenic and non-transgenic soybean cultivars. Between them, a very small number of differentially expressed genes were detected, mirroring several of the functional categories we observed in our transcriptome

study including genes relating to binding, protease inhibitors, and transport [97]. In addition, differences between cultivar varieties far exceeded those between the transgenic and wild type specimens. These studies were conducted with microarray technologies prior to the mainstream availability of next-generation sequencing, which makes our investigations an important continuation of preceding works.

Measurable unscripted gene expression changes were detected in the seed transcriptomes of all three transgenic soybean lines we chose for analysis, with line 764 being substantially altered beyond natural variance rates. Differences detected at the transcript level may be due to T-DNA insert locations, random mutations following transformation or direct effects of the recombinant protein itself, or a combination of these, of which the physiological consequences of such changes remain unknown. It is unclear when these alterations occurred, as our transcriptomic quantifications were focused on only one generation of progeny, and parallel independent transformed lines expressing the identical recombinant protein were not included in these analyses. Future analyses will need to be conducted on prior generations of seeds, coupled with metabolomic and proteomic assays, to fully investigate the extent and origin of detected variations.

Significant SNPs were detected in all three transgenic lines, of which all but line 764 were comparable to SNP rates seen in the sequenced wild type Williams 82 cultivar. All transgenic lines seemed to keep constant polymorphism rates among all group replicates, indicating these were all likely from an earlier event, conserved to future progeny and were not arising spontaneously from environmental pressures. This not only

reinforces the consistency of the replicates in each experimental group, but also the genomic stability of each.

Detection of alternative splicing in some ST77 hTG transgenic progeny was quite surprising, as no previous screens of ST77 displayed variance in the final protein size. Upon further investigation, the intron segment removed was relatively small (98bp) compared to the entire hTG gene segment (~8kb); a difference miniscule enough to would not be detectable in typical acrylamide gels. The high coverage RNA-sequencing datasets revealed no transgene splicing in the other two transgenic events characterized; however, since the ST77 event was not entirely homogeneous in the progeny exhibiting that particular splice junction, it is possible that other events not included in these studies display a similar trend. Expression of human growth hormone in multiple species of transgenic animals display many different alternative splicing patterns, leading to host-specific pre-mRNA processing of identical gene sequences [212]. Therefore, as soybean is a eukaryotic higher order plant capable of many complex transcriptional and translational modifications, it is realistic to expect that the same phenomenon will also hold true for different plant hosts.

Although significant transcriptomic and genomic changes were detected in the 764 line, phenotypical properties such as seed yield and size, plant height, and maturation time were unaltered. Compositional changes such as protein content, metabolites, and antioxidant content require further experimentation to elucidate any variances that may be present; however based on observed phenotypes we expect no significant displacement of these components to have occurred. Literature reviews on pleiotropic changes in particular genetically modified crops are sparse, and many seem to only

address one aspect of measured equivalence. The World Health Organization (WHO), Organization for Economic Co-operation and Development (OECD), and the United States Department of Agriculture Animal and Plant Health Inspection Service (USDA-APHIS) all have criteria for evaluation of safety of modified foods, which broadly mentions toxicity, allergenicity, nutritional and toxic properties, transgene stability, and any unintended effects resulting directly from the transgene insertion. Because of this broad definition of "substantial equivalence", examinations unambiguously confirming or refuting comparisons with irrefutable and all-encompassing datasets are impossible to generate, and targeted approaches severely limit detections of unknown or novel elements.

All together, some alterations are expected to occur following any transformation process. However, upstream design decisions can limit unintended downstream effects in plants utilized as bioreactors for pharmaceuticals to minimize balancing cost and benefits. Because characterized gene expression changes in transgenics, including those described here, are all part of endogenous pathways and processes existing even in the absence of the transgene, the consequences of these alterations possibly only impact overall plant robustness and yield with no added risk to human health. In fact, the widely consumed cultivated sweet potato has serendipitously been discovered to contain T-DNA sequences from *Agrobacterium* species, effectively becoming a naturally occurring transgenic food [213]. Likewise, it is extremely important to address the fact that unintended effects also result from conventional breeding techniques, and instances where composition varies outside an expected norm does not indicate a health hazard [214]. In molecular farming instances, substantial equivalence is not a measure of safety; rather, it is an approach to

identify variations in characteristics that deviate from the norm, and to address potential considerations for improving the cost-effectiveness of plant bioreactors and soybean as an advantageous platform for biopharmaceuticals.

REFERENCES

[1]. Tilman D, Cassman KG, Matson PA, Naylor R, Polasky S: Agricultural sustainability and intensive production practices. *Nature* 2002, 418(6898):671-677.

[2]. De La Fuente GN, Frei UK, Lubberstedt T: Accelerating plant breeding. *Trends Plant Sci* 2013, 18(12):667-672.

[3]. Oakes JL, Bost KL, Piller KJ: Stability of a soybean seed-derived vaccine antigen following long-term storage, processing and transport in the absence of a cold chain. *J Sci Food Agric* 2009, 89(13):2191-2199.

[4]. Hudson LC, Lambirth KC, Bost KL, Piller KJ: Advancements in Transgenic Soy: From Field to Bedside; 2013.

[5]. Hudson LC, Garg R, Bost KL, Piller KJ: Soybean seeds: a practical host for the production of functional subunit vaccines. In: *Biomed Res Int.* vol. 2014; 2014: 340804.

[6]. Ray JD, Kilen TC, Abel CA, Paris RL: Soybean natural cross-pollination rates under field conditions. *Environ Biosafety Res* 2003, 2(2):133-138.

[7]. Bost KL, Lambirth KC, Hudson LC, Piller KJ: Soybean-Derived Thyroglobulin As an Analyte Specific Reagent for *In Vitro* Diagnostic Tests and Devices. *Advances in Medicine and Biology* 2014, 80.

[8]. Abiri R, Valdiani A, Maziah M, Shaharuddin NA, Sahebi M, Yusof ZY, Atabaki N, Talei D: A Critical Review of the Concept of Transgenic Plants: Insights into Pharmaceutical Biotechnology and Molecular Farming. *Curr Issues Mol Biol* 2015, 18:21-42.

[9]. Powell R, Hudson LC, Lambirth KC, Luth D, Wang K, Bost KL, Piller KJ: Recombinant expression of homodimeric 660 kDa human thyroglobulin in soybean seeds: an alternative source of human thyroglobulin. *Plant Cell Rep* 2011, 30(7):1327-1338.

[10]. Bost KL, Piller KJ: Protein Expression Systems: Why Soybean Seeds? In: *Soybean - Molecular Aspects of Breeding.* Edited by Sudaric A. Intech Open: Intech; 2011.

[11]. Reddy MS, Dinkins RD, Collins GB: Gene silencing in transgenic soybean plants transformed via particle bombardment. *Plant Cell Rep* 2003, 21(7):676-683.

[12]. Kikkert JR, Vidal JR, Reisch BI: Stable transformation of plant cells by particle bombardment/biolistics. *Methods Mol Biol* 2005, 286:61-78.

[13]. Chang CH, Zhu J, Winans SC: Pleiotropic phenotypes caused by genetic ablation of the receiver module of the Agrobacterium tumefaciens VirA protein. *J Bacteriol* 1996, 178(15):4710-4716.

[14]. Pitzschke A, Hirt H: New insights into an old story: Agrobacterium-induced tumour formation in plants by plant transformation. *EMBO J* 2010, 29(6):1021-1032.

[15]. Cascales E, Atmakuri K, Sarkar MK, Christie PJ: DNA substrate-induced activation of the Agrobacterium VirB/VirD4 type IV secretion system. *J Bacteriol* 2013, 195(11):2691-2704.

[16]. Wang K, Stachel SE, Timmerman B, M VANM, Zambryski PC: Site-Specific Nick in the T-DNA Border Sequence as a Result of Agrobacterium vir Gene Expression. *Science* 1987, 235(4788):587-591.

[17]. Mysore KS, Bassuner B, Deng XB, Darbinian NS, Motchoulski A, Ream W, Gelvin SB: Role of the Agrobacterium tumefaciens VirD2 protein in T-DNA transfer and integration. *Mol Plant Microbe Interact* 1998, 11(7):668-683.

[18]. Ziemienowicz A, Merkle T, Schoumacher F, Hohn B, Rossi L: Import of Agrobacterium T-DNA into plant nuclei: two distinct functions of VirD2 and VirE2 proteins. *Plant Cell* 2001, 13(2):369-383.

[19]. Lacroix B, Citovsky V: The roles of bacterial and host plant factors in Agrobacterium-mediated genetic transformation. *Int J Dev Biol* 2013, 57(6-8):467-481.

[20]. Anand A, Vaghchhipawala Z, Ryu CM, Kang L, Wang K, del-Pozo O, Martin GB, Mysore KS: Identification and characterization of plant genes involved in Agrobacterium-mediated plant transformation by virus-induced gene silencing. *Mol Plant Microbe Interact* 2007, 20(1):41-52.

[21]. Tzfira T, Vaidya M, Citovsky V: Involvement of targeted proteolysis in plant genetic transformation by Agrobacterium. *Nature* 2004, 431(7004):87-92.

[22]. Puchta H, Fauser F: Gene targeting in plants: 25 years later. *Int J Dev Biol* 2013, 57(6-8):629-637.

[23]. Salomon S, Puchta H: Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. *EMBO J* 1998, 17(20):6086-6095.

[24]. Kim SI, Veena, Gelvin SB: Genome-wide analysis of Agrobacterium T-DNA integration sites in the Arabidopsis genome generated under non-selective conditions. *Plant J* 2007, 51(5):779-791.

[25]. Tzfira T, Frankman LR, Vaidya M, Citovsky V: Site-specific integration of Agrobacterium tumefaciens T-DNA via double-stranded intermediates. *Plant Physiol* 2003, 133(3):1011-1023.

[26]. Endo M, Ishikawa Y, Osakabe K, Nakayama S, Kaya H, Araki T, Shibahara K, Abe K, Ichikawa H, Valentine L *et al*: Increased frequency of homologous recombination and T-DNA integration in Arabidopsis CAF-1 mutants. *EMBO J* 2006, 25(23):5579-5590.

[27]. van Attikum H, Hooykaas PJ: Genetic requirements for the targeted integration of Agrobacterium T-DNA in Saccharomyces cerevisiae. *Nucleic Acids Res* 2003, 31(3):826-832.

[28]. Park SY, Vaghchhipawala Z, Vasudevan B, Lee LY, Shen Y, Singer K, Waterworth WM, Zhang ZJ, West CE, Mysore KS *et al*: Agrobacterium T-DNA integration into the plant genome can occur without the activity of key non-homologous end-joining proteins. *Plant J* 2015, 81(6):934-946.

[29]. Lacroix B, Loyter A, Citovsky V: Association of the Agrobacterium T-DNA-protein complex with plant nucleosomes. *Proc Natl Acad Sci U S A* 2008, 105(40):15429-15434.

[30]. Magori S, Citovsky V: Epigenetic control of Agrobacterium T-DNA integration. *Biochim Biophys Acta* 2011, 1809(8):388-394.

[31]. Hoekema A, van Haaren MJ, Fellinger AJ, Hooykaas PJ, Schilperoort RA: Non-oncogenic plant vectors for use in the agrobacterium binary system. *Plant Mol Biol* 1985, 5(2):85-89.

[32]. Hoekema A, Hirsch PR, Hooykaas PJJ, Schilperoort RA: A binary plant vector strategy based on separation of vir- and T-region of the Agrobacterium tumefaciens Ti-plasmid. *Nature* 1983, 303(5913):179-180.

[33]. Lee LY, Gelvin SB: T-DNA binary vectors and systems. *Plant Physiol* 2008, 146(2):325-332.

[34]. Singh A, Meena M, Kumar D, Dubey AK, Hassan MI: Structural and functional analysis of various globulin proteins from soy seed. *Crit Rev Food Sci Nutr* 2015, 55(11):1491-1502.

[35]. Ding SH, Huang LY, Wang YD, Sun HC, Xiang ZH: High-level expression of basic fibroblast growth factor in transgenic soybean seeds and characterization of its biological activity. *Biotechnol Lett* 2006, 28(12):869-875.

[36]. LC H, KL B, KJ P: Optimizing Recombinant Protein Expression in Soybean. *Soybean - Molecular Aspects of Breeding* 2011.

[37]. Voges MJ, Silver PA, Way JC, Mattozzi MD: Targeting a heterologous protein to multiple plant organelles via rationally designed 5′ mRNA tags. *Journal of Biological Engineering* 2013, 7:20-20.

[38]. Owen MRL, Pen J: Transgenic Plants: A Production System for Industrial and Pharmaceutical Proteins: Wiley; 1996.

[39]. Zeenko V, Gallie DR: Cap-independent translation of tobacco etch virus is conferred by an RNA pseudoknot in the 5'-leader. *J Biol Chem* 2005, 280(29):26813-26824.

[40]. Gallie DR: The 5′-leader of tobacco mosaic virus promotes translation through enhanced recruitment of eIF4F. *Nucleic Acids Res* 2002, 30(15):3401-3411.

[41]. Moravec T, Schmidt MA, Herman EM, Woodford-Thomas T: Production of Escherichia coli heat labile toxin (LT) B subunit in soybean seed and analysis of its immunogenicity as an oral vaccine. *Vaccine* 2007, 25(9):1647-1657.

[42]. Preuss SB, Meister R, Xu Q, Urwin CP, Tripodi FA, Screen SE, Anil VS, Zhu S, Morrell JA, Liu G *et al*: Expression of the Arabidopsis thaliana BBX32 gene in soybean increases grain yield. *PloS one* 2012, 7(2):e30717.

[43]. Valente MA, Faria JA, Soares-Ramos JR, Reis PA, Pinheiro GL, Piovesan ND, Morais AT, Menezes CC, Cano MA, Fietto LG *et al*: The ER luminal binding protein (BiP) mediates an increase in drought tolerance in soybean and delays drought-induced leaf senescence in soybean and tobacco. *J Exp Bot* 2009, 60(2):533-546.

[44]. Lardizabal K, Effertz R, Levering C, Mai J, Pedroso MC, Jury T, Aasen E, Gruys K, Bennett K: Expression of Umbelopsis ramanniana DGAT2A in seed increases oil in soybean. *Plant Physiol* 2008, 148(1):89-96.

[45]. Rao SS, Hildebrand D: Changes in oil content of transgenic soybeans expressing the yeast SLC1 gene. *Lipids* 2009, 44(10):945-951.

[46]. Zeitlin L, Olmsted SS, Moench TR, Co MS, Martinell BJ, Paradkar VM, Russell DR, Queen C, Cone RA, Whaley KJ: A humanized monoclonal antibody produced in transgenic plants for immunoprotection of the vagina against genital herpes. *Nat Biotechnol* 1998, 16(13):1361-1364.

[47]. Takaiwa F: Seed-based oral vaccines as allergen-specific immunotherapies. *Hum Vaccin* 2011, 7(3):357-366.

[48]. Wang Z, Gerstein M, Snyder M: RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009, 10(1):57-63.

[49]. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al*: Real-time DNA sequencing from single polymerase molecules. *Science* 2009, 323(5910):133-138.

[50]. Roberts RJ, Carneiro MO, Schatz MC: The advantages of SMRT sequencing. *Genome Biol* 2013, 14(7):405.

[51]. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A: Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* 2010, 7(9):709-715.

[52]. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009, 10(3):R25.

[53]. 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.

[54]. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J *et al*: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013, 14(6):671-683.

[55]. 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.

[56]. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L: Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* 2011, 12(3):R22.

[57]. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010, 28(5):511-515.

[58]. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al*: Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004, 5(10):R80.

[59]. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010, 26(1):139-140.

[60]. Robinson MD, Smyth GK: Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 2007, 23(21):2881-2887.

[61]. Liao Y, Smyth GK, Shi W: featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014, 30(7):923-930.

[62]. Chen Y, Lun AL, Smyth G: Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR. In: *Statistical Analysis of Next Generation Sequencing Data.* Edited by Datta S, Nettleton D: Springer International Publishing; 2014: 51-74.

[63]. Yendrek CR, Ainsworth EA, Thimmapuram J: The bench scientist's guide to statistical analysis of RNA-Seq data. *BMC Res Notes* 2012, 5:506.

[64]. 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.

[65]. Supek F, Bosnjak M, Skunca N, Smuc T: REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one* 2011, 6(7):e21800.

[66]. Young MD, Wakefield MJ, Smyth GK, Oshlack A: Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010, 11(2):R14.

[67]. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X: Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PloS one* 2014, 9(1):e78644.

[68]. Vijay N, Poelstra JW, Kunstner A, Wolf JB: Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol Ecol* 2013, 22(3):620-634.

[69]. Thierry-Mieg D, Thierry-Mieg J: AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* 2006, 7 Suppl 1:S12 11-14.

[70]. Liao Y, Smyth GK, Shi W: The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* 2013, 41(10):e108.

[71]. Glaus P, Honkela A, Rattray M: Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* 2012, 28(13):1721-1728.

[72]. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013, 29(1):15-21.

[73]. Consortium SM-I: A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* 2014, 32(9):903-914.

[74]. 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.

[75]. O'Rourke JA, Bolon YT, Bucciarelli B, Vance CP: Legume genomics: understanding biology through DNA and RNA sequencing. *Ann Bot* 2014, 113(7):1107-1120.

[76]. Whaley A, Sheridan J, Safari S, Burton A, Burkey K, Schlueter J: RNA-seq analysis reveals genetic response and tolerance mechanisms to ozone exposure in soybean. *BMC Genomics* 2015, 16:426.

[77]. Chan C, Qi X, Li MW, Wong FL, Lam HM: Recent developments of genomic research in soybean. *J Genet Genomics* 2012, 39(7):317-324.

[78]. Jones SI, Vodkin LO: Using RNA-Seq to profile soybean seed development from fertilization to maturity. *PloS one* 2013, 8(3):e59270.

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

[80]. De Buck S: T-DNA vector backbone sequences are frequently integrated into the genome of transgenic plants obtained by Agrobacterium-mediated transformation. *Mol Breed* 2000, 6(5):459-468.

[81]. Kononov ME, Bassuner B, Gelvin SB: Integration of T-DNA binary vector 'backbone' sequences into the tobacco genome: evidence for multiple complex patterns of integration. *Plant J* 1997, 11(5):945-957.

[82]. Fu D, St. Amand PC, Xiao Y, Muthukrishnan S, Liang GH: Characterization of T-DNA integration in creeping bentgrass. *Plant Sci* 2006, 170(2):225-237.

[83]. Ichikawa T, Nakazawa M, Kawashima M, Muto S, Gohda K, Suzuki K, Ishikawa A, Kobayashi H, Yoshizumi T, Tsumoto Y *et al*: Sequence database of 1172 T-DNA insertion sites in Arabidopsis activation-tagging lines that showed phenotypes in T1 generation. *Plant J* 2003, 36(3):421-429.

[84]. Kadam U, Moeller CA, Irudayaraj J, Schulz B: Effect of T-DNA insertions on mRNA transcript copy numbers upstream and downstream of the insertion site in Arabidopsis thaliana explored by surface enhanced Raman spectroscopy. *Plant Biotechnol J* 2014, 12(5):568-577.

[85]. Filipecki M, Malepszy S: Unintended consequences of plant transformation: a molecular insight. *J Appl Genet* 2006, 47(4):277-286.

[86]. Natarajan S, Luthria D, Bae H, Lakshman D, Mitra A: Transgenic soybeans and soybean protein analysis: an overview. *J Agric Food Chem* 2013, 61(48):11736-11743.

[87]. Barbosa HS, Arruda SC, Azevedo RA, Arruda MA: New insights on proteomics of transgenic soybean seeds: evaluation of differential expressions of enzymes and proteins. *Anal Bioanal Chem* 2012, 402(1):299-314.

[88]. Zhu J, Patzoldt WL, Shealy RT, Vodkin LO, Clough SJ, Tranel PJ: Transcriptome response to glyphosate in sensitive and resistant soybean. *J Agric Food Chem* 2008, 56(15):6355-6363.

[89]. Liu K: Chemistry and Nutritional Value of Soybean Components. In: *Soybeans.* Springer US; 1997: 25-113.

[90]. Bazalo GR, Joshi AV, Germak J: Comparison of human growth hormone products' cost in pediatric and adult patients. A budgetary impact model. *Manag Care* 2007, 16(9):45-51.

[91]. Franklin SL, Geffner ME: Growth hormone: the expansion of available products and indications. *Endocrinol Metab Clin North Am* 2009, 38(3):587-611.

[92]. Hudson LC, Seabolt BS, Odle J, Bost KL, Stahl CH, Piller KJ: Sublethal staphylococcal enterotoxin B challenge model in pigs to evaluate protection following immunization with a soybean-derived vaccine. *Clin Vaccine Immunol* 2013, 20(1):24-32.

[93]. Piller KJ, Clemente TE, Jun SM, Petty CC, Sato S, Pascual DW, Bost KL: Expression and immunogenicity of an Escherichia coli K99 fimbriae subunit antigen in soybean. *Planta* 2005, 222(1):6-18.

[94]. Herman RA, Ladics GS: Endogenous allergen upregulation: transgenic vs. traditionally bred crops. *Food Chem Toxicol* 2011, 49(10):2667-2669.

[95]. Catchpole GS, Beckmann M, Enot DP, Mondhe M, Zywicki B, Taylor J, Hardy N, Smith A, King RD, Kell DB *et al*: Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc Natl Acad Sci U S A* 2005, 102(40):14458-14462.

[96]. Zhang X, Zhao P, Wu K, Zhang Y, Peng M, Liu Z: Compositional equivalency of RNAi-mediated virus-resistant transgenic soybean and its nontransgenic counterpart. *J Agric Food Chem* 2014, 62(19):4475-4479.

[97]. Cheng KC, Beaulieu J, Iquira E, Belzile FJ, Fortin MG, Stromvik MV: Effect of transgenes on global gene expression in soybean is within the natural range of variation of conventional cultivars. *J Agric Food Chem* 2008, 56(9):3057-3067.

[98]. Beale MH, Ward JL, Baker JM: Establishing substantial equivalence: metabolomics. *Methods Mol Biol* 2009, 478:289-303.

[99]. Batista R, Saibo N, Lourenco T, Oliveira MM: Microarray analyses reveal that plant mutagenesis may induce more transcriptomic changes than transgene insertion. *Proc Natl Acad Sci U S A* 2008, 105(9):3640-3645.

[100]. Baudo MM, Lyons R, Powers S, Pastori GM, Edwards KJ, Holdsworth MJ, Shewry PR: Transgenesis has less impact on the transcriptome of wheat grain than conventional breeding. *Plant Biotechnol J* 2006, 4(4):369-380.

[101]. Ricroch AE, Berge JB, Kuntz M: Evaluation of genetically engineered crops using transcriptomic, proteomic, and metabolomic profiling techniques. *Plant Physiol* 2011, 155(4):1752-1761.

[102]. Baker JM, Hawkins ND, Ward JL, Lovegrove A, Napier JA, Shewry PR, Beale MH: A metabolomic study of substantial equivalence of field-grown genetically modified wheat. *Plant Biotechnol J* 2006, 4(4):381-392.

[103]. Kusano M, Redestig H, Hirai T, Oikawa A, Matsuda F, Fukushima A, Arita M, Watanabe S, Yano M, Hiwasa-Tanase K *et al*: Covering chemical diversity of genetically-modified tomatoes using metabolomics for objective substantial equivalence assessment. *PloS one* 2011, 6(2):e16989.

[104]. Lepping MD, Herman RA, Potts BL: Compositional equivalence of DAS-444O6-6 (AAD-12 + 2mEPSPS + PAT) herbicide-tolerant soybean and nontransgenic soybean. *J Agric Food Chem* 2013, 61(46):11180-11190.

[105]. Snell C, Bernheim A, Berge JB, Kuntz M, Pascal G, Paris A, Ricroch AE: Assessment of the health impact of GM plant diets in long-term and multigenerational animal feeding trials: a literature review. *Food Chem Toxicol* 2012, 50(3-4):1134-1148.

[106]. Houshyani B, van der Krol AR, Bino RJ, Bouwmeester HJ: Assessment of pleiotropic transcriptome perturbations in Arabidopsis engineered for indirect insect defence. *BMC Plant Biol* 2014, 14:170.

[107]. Kuiper HA, Kok EJ, Engel KH: Exploitation of molecular profiling techniques for GM food safety assessment. *Curr Opin Biotechnol* 2003, 14(2):238-243.

[108]. Rynda-Apple A, Huarte E, Maddaloni M, Callis G, Skyberg JA, Pascual DW: Active immunization using a single dose immunotherapeutic abates established EAE via IL-10 and regulatory T cells. *Eur J Immunol* 2011, 41(2):313-323.

[109]. Paz MM, Martinez JC, Kalvig AB, Fonger TM, Wang K: Improved cotyledonary node method using an alternative explant derived from mature seed for efficient Agrobacterium-mediated soybean transformation. *Plant Cell Rep* 2006, 25(3):206-213.

[110]. Trapnell C, Pachter L, Salzberg SL: TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009, 25(9):1105-1111.

[111]. Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I: The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res* 2014, 42(Database issue):D26-31.

[112]. 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 Res* 2012, 40(Database issue):D1178-1186.

[113]. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C: The sequence read archive. *Nucleic Acids Res* 2011, 39(Database issue):D19-21.

[114]. 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.

[115]. Edgar R, Domrachev M, Lash AE: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002, 30(1):207-210.

[116]. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlen M, Nielsen J: A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res* 2012, 40(20):10084-10097.

[117]. Kvam VM, Liu P, Si Y: A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am J Bot* 2012, 99(2):248-256.

[118]. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D: Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 2013, 14(9):R95.

[119]. Seyednasrollah F, Laiho A, Elo LL: Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform* 2015, 16(1):59-70.

[120]. Soneson C, Delorenzi M: A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 2013, 14:91.

[121]. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M *et al*: TM4: a free, open-source system for microarray data management and analysis. *BioTechniques* 2003, 34(2):374-378.

[122]. Rang A, Linke B, Jansen B: Detection of RNA variants transcribed from the transgene in Roundup Ready soybean. *Eur Food Res Technol* 2005, 220(3-4):438-443.

[123]. Schoen DJ, David JL, Bataillon TM: Deleterious mutation accumulation and the regeneration of genetic resources. *Proc Natl Acad Sci U S A* 1998, 95(1):394-399.

[124]. Molinier J, Ries G, Zipfel C, Hohn B: Transgeneration memory of stress in plants. *Nature* 2006, 442(7106):1046-1049.

[125]. Forsbach A, Schubert D, Lechtenberg B, Gils M, Schmidt R: A comprehensive characterization of single-copy T-DNA insertions in the Arabidopsis thaliana genome. *Plant Mol Biol* 2003, 52(1):161-176.

[126]. Latham JR, Wilson AK, Steinbrecher RA: The mutational consequences of plant transformation. *J Biomed Biotechnol* 2006, 2006(2):25376.

[127]. Vaucheret H, Beclin C, Elmayan T, Feuerbach F, Godon C, Morel JB, Mourrain P, Palauqui JC, Vernhettes S: Transgene-induced gene silencing in plants. *Plant J* 1998, 16(6):651-659.

[128]. Beers E, Woffenden B, Zhao C: Plant proteolytic enzymes: possible roles during programmed cell death. In: *Programmed Cell Death in Higher Plants.* Edited by Lam E, Fukuda H, Greenberg J: Springer Netherlands; 2000: 155-171.

[129]. Solomon M, Belenghi B, Delledonne M, Menachem E, Levine A: The involvement of cysteine proteases and protease inhibitor genes in the regulation of programmed cell death in plants. *Plant Cell* 1999, 11(3):431-444.

[130]. Botella MA, Xu Y, Prabha TN, Zhao Y, Narasimhan ML, Wilson KA, Nielsen SS, Bressan RA, Hasegawa PM: Differential expression of soybean cysteine proteinase inhibitor genes during development and in response to wounding and methyl jasmonate. *Plant Physiol* 1996, 112(3):1201-1210.

[131]. Antao CM, Malcata FX: Plant serine proteases: biochemical, physiological and molecular features. *Plant Physiol Biochem* 2005, 43(7):637-650.

[132]. Russell DA, Spatola LA, Dian T, Paradkar VM, Dufield DR, Carroll JA, Schlittler MR: Host limits to accurate human growth hormone production in multiple plant systems. *Biotechnol Bioeng* 2005, 89(7):775-782.

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

[134]. Nicol JW, Helt GA, Blanchard SG, Jr., Raja A, Loraine AE: The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* 2009, 25(20):2730-2731.

[135]. Gheysen G, Montagu MV, Zambryski P: Integration of Agrobacterium tumefaciens transfer DNA (T-DNA) involves rearrangements of target plant DNA sequences. *Proc Natl Acad Sci U S A* 1987, 84(17):6169-6173.

[136]. Darmency H: Pleiotropic effects of herbicide-resistance genes on crop yield: a review. *Pest Manag Sci* 2013, 69(8):897-904.

[137]. Krysan PJ, Young JC, Tax F, Sussman MR: Identification of transferred DNA insertions within Arabidopsis genes involved in signal transduction and ion transport. *Proceedings of the National Academy of Sciences of the United States of America* 1996, 93(15):8145-8150.

[138]. Mason G, Provero P, Vaira AM, Accotto GP: Estimating the number of integrations in transformed plants by quantitative real-time PCR. *BMC Biotechnol* 2002, 2:20.

[139]. Ingham DJ, Beer S, Money S, Hansen G: Quantitative real-time PCR assay for determining transgene copy number in transformed plants. *BioTechniques* 2001, 31(1):132-134, 136-140.

[140]. Honda M, Muramoto Y, Kuzuguchi T, Sawano S, Machida M, Koyama H: Determination of gene copy number and genotype of transgenic Arabidopsis thaliana by competitive PCR. *J Exp Bot* 2002, 53(373):1515-1520.

[141]. Lattenmayer C, Loeschel M, Steinfellner W, Trummer E, Mueller D, Schriebl K, Vorauer-Uhl K, Katinger H, Kunert R: Identification of transgene integration loci of different highly expressing recombinant CHO cell lines by FISH. *Cytotechnology* 2006, 51(3):171-182.

[142]. Kulnane LS, Lehman EJ, Hock BJ, Tsuchiya KD, Lamb BT: Rapid and efficient detection of transgene homozygosity by FISH of mouse fibroblasts. *Mamm Genome* 2002, 13(4):223-226.

[143]. Nakanishi T, Kuroiwa A, Yamada S, Isotani A, Yamashita A, Tairaka A, Hayashi T, Takagi T, Ikawa M, Matsuda Y *et al*: FISH analysis of 142 EGFP transgene integration sites into the mouse genome. *Genomics* 2002, 80(6):564-574.

[144]. Moscone EA, Matzke MA, Matzke AJ: The use of combined FISH/GISH in conjunction with DAPI counterstaining to identify chromosomes containing transgene inserts in amphidiploid tobacco. *Chromosoma* 1996, 105(4):231-236.

[145]. Potter CJ, Luo L: Splinkerette PCR for mapping transposable elements in Drosophila. *PloS one* 2010, 5(4):e10168.

[146]. Pavlopoulos A: Identification of DNA sequences that flank a known region by inverse PCR. *Methods Mol Biol* 2011, 772:267-275.

[147]. Zhang B, Huang JQ, Wei ZM: [A quick method to estimate the T-DNA copy number in transgenic rice using inverse PCR (IPCR)]. *Shi Yan Sheng Wu Xue Bao* 1999, 32(2):207-211.

[148]. Leoni C, Gallerani R, Ceci LR: A genome walking strategy for the identification of eukaryotic nucleotide sequences adjacent to known regions. *BioTechniques* 2008, 44(2):229, 232-225.

[149]. Chambers K, Lowe R, Howlett B, Zander M, Batley J, Van de Wouw A, Elliott C: Next-generation genome sequencing can be used to rapidly characterise sequences flanking T-DNA insertions in random insertional mutants of Leptosphaeria maculans. *Fungal Biology and Biotechnology* 2014, 1(1):10.

[150]. Srivastava A, Philip VM, Greenstein I, Rowe LB, Barter M, Lutz C, Reinholdt LG: Discovery of transgene insertion sites by high throughput sequencing of mate pair libraries. *BMC Genomics* 2014, 15:367.

[151]. Ji Y, Abrams N, Zhu W, Salinas E, Yu Z, Palmer DC, Jailwala P, Franco Z, Roychoudhuri R, Stahlberg E *et al*: Identification of the genomic insertion site of Pmel-1 TCR alpha and beta transgenes by next-generation sequencing. *PloS one* 2014, 9(5):e96650.

[152]. Lepage E, Zampini E, Boyle B, Brisson N: Time- and cost-efficient identification of T-DNA insertion sites through targeted genomic sequencing. *PloS one* 2013, 8(8):e70912.

[153]. Kovalic D, Garnaat C, Guo L, Yan YP, Groat J, Silvanovich A, Ralston L, Huang MY, Tian Q, Christian A *et al*: The Use of Next Generation Sequencing and Junction Sequence Analysis Bioinformatics to Achieve Molecular Characterization of Crops Improved Through Modern Biotechnology. *Plant Genome* 2012, 5(3):149-163.

[154]. Jiang C, Chen C, Huang Z, Liu R, Verdier J: ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC Bioinformatics* 2015, 16(1):72.

[155]. Zhang R, Yin Y, Zhang Y, Li K, Zhu H, Gong Q, Wang J, Hu X, Li N: Molecular characterization of transgene integration by next-generation sequencing in transgenic cattle. *PloS one* 2012, 7(11):e50348.

[156]. Zhang Z, Schwartz S, Wagner L, Miller W: A greedy algorithm for aligning DNA sequences. *J Comput Biol* 2000, 7(1-2):203-214.

[157]. Ziemienowicz A: Agrobacterium-mediated plant transformation: Factors, applications and recent advances. *Biocatalysis and Agricultural Biotechnology* 2014, 3(4):95-102.

[158]. Koch MS, Ward JM, Levine SL, Baum JA, Vicini JL, Hammond BG: The food and environmental safety of Bt crops. *Front Plant Sci* 2015, 6:283.

[159]. Dill GM, Cajacob CA, Padgette SR: Glyphosate-resistant crops: adoption, use and future considerations. *Pest Manag Sci* 2008, 64(4):326-331.

[160]. Zhang Y, Li D, Jin X, Huang Z: Fighting Ebola with ZMapp: spotlight on plant-made antibody. *Sci China Life Sci* 2014, 57(10):987-988.

[161]. Garg R, Tolbert M, Oakes JL, Clemente TE, Bost KL, Piller KJ: Chloroplast targeting of FanC, the major antigenic subunit of Escherichia coli K99 fimbriae, in transgenic soybean. *Plant Cell Rep* 2007, 26(7):1011-1023.

[162]. Shimoda N, Toyoda-Yamamoto A, Nagamine J, Usami S, Katayama M, Sakagami Y, Machida Y: Control of expression of Agrobacterium vir genes by synergistic actions of phenolic signal molecules and monosaccharides. *Proc Natl Acad Sci U S A* 1990, 87(17):6684-6688.

[163]. Salman H, Abu-Arish A, Oliel S, Loyter A, Klafter J, Granek R, Elbaum M: Nuclear localization signal peptides induce molecular delivery along microtubules. *Biophys J* 2005, 89(3):2134-2145.

[164]. Lambirth K, Whaley A, Blakley I, Schlueter J, Bost K, Loraine A, Piller K: A Comparison of transgenic and wild type soybean seeds: analysis of transcriptome profiles using RNA-Seq. *BMC Biotechnol* 2015, 15(1):89.

[165]. Fang J, Zhai WX, Wang WM, Li SW, Zhu LH: [Amplification and analysis of T-DNA flanking sequences in transgenic rice]. *Yi Chuan Xue Bao* 2001, 28(4):345-351.

[166]. Lambirth KC, Whaley AM, Schlueter JA, Bost KL, Piller KJ: CONTRAILS: A tool for rapid identification of transgene integration sites in complex, repetitive genomes using low-coverage paired-end sequencing. *Genomics Data* 2015, 6:175-181.

[167]. Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S: Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res* 1996, 24(17):3439-3452.

[168]. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L *et al*: SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res* 2014, 42(Web Server issue):W252-258.

[169]. Li H: A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011, 27(21):2987-2993.

[170]. 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.

[171]. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012, 6(2):80-92.

[172]. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A *et al*: The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Front Plant Sci* 2011, 2:34.

[173]. Lambirth KC, Whaley AM, Schlueter JA, Bost KL, Piller KJ: CONTRAILS: A tool for rapid identification of transgene integration sites in complex, repetitive genomes using low-coverage paired-end sequencing. *Genomics Data*.

[174]. James C: Global Status of Commercialized Biotech/GM Crops: 2014. In: *ISAAA Brief*. ISAAA: Ithaca, NY.; 2014.

[175]. Shoemaker RC, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP *et al*: Genome duplication in soybean (Glycine subgenus soja). *Genetics* 1996, 144(1):329-338.

[176]. Schlueter JA, Lin JY, Schlueter SD, Vasylenko-Sanders IF, Deshpande S, Yi J, O'Bleness M, Roe BA, Nelson RT, Scheffler BE *et al*: Gene duplication and paleopolyploidy in soybean and the implications for whole genome sequencing. *BMC Genomics* 2007, 8:330.

[177]. Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ: Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. *Syst Biol* 2005, 54(3):441-454.

[178]. Black DL: Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 2003, 72:291-336.

[179]. Syed NH, Kalyna M, Marquez Y, Barta A, Brown JWS: Alternative splicing in plants – coming of age. *Trends Plant Sci* 2012, 17(10-10):616-623.

[180]. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M: Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res* 2012, 22(6):1184-1195.

[181]. Shen Y, Zhou Z, Wang Z, Li W, Fang C, Wu M, Ma Y, Liu T, Kong LA, Peng DL *et al*: Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell* 2014, 26(3):996-1008.

[182]. Reddy AS, Marquez Y, Kalyna M, Barta A: Complexity of the alternative splicing landscape in plants. *Plant Cell* 2013, 25(10):3657-3683.

[183]. Wang L, Cao C, Ma Q, Zeng Q, Wang H, Cheng Z, Zhu G, Qi J, Ma H, Nian H *et al*: RNA-seq analyses of multiple meristems of soybean: novel and alternative transcripts, evolutionary and functional implications. *BMC Plant Biol* 2014, 14(1):169.

[184]. Staiger D, Brown JW: Alternative splicing at the intersection of biological timing, development, and stress responses. *Plant Cell* 2013, 25(10):3640-3656.

[185]. Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB: Single-nucleotide polymorphisms in soybean. *Genetics* 2003, 163(3):1123-1134.

[186]. Shu Y, Li Y, Zhu Z, Bai X, Cai H, Ji W, Guo D, Zhu Y: SNPs discovery and CAPS marker conversion in soybean. *Mol Biol Rep* 2011, 38(3):1841-1846.

[187]. Yadav CB, Bhareti P, Muthamilarasan M, Mukherjee M, Khan Y, Rathi P, Prasad M: Genome-wide SNP identification and characterization in two soybean cultivars with contrasting Mungbean Yellow Mosaic India Virus disease resistance traits. *PloS one* 2015, 10(4):e0123897.

[188]. Wu X, Ren C, Joshi T, Vuong T, Xu D, Nguyen HT: SNP discovery by high-throughput sequencing in soybean. *BMC Genomics* 2010, 11:469.

[189]. Li YH, Reif JC, Jackson SA, Ma YS, Chang RZ, Qiu LJ: Detecting SNPs underlying domestication-related traits in soybean. *BMC Plant Biol* 2014, 14:251.

[190]. Goettel W, Xia E, Upchurch R, Wang ML, Chen P, An YQ: Identification and characterization of transcript polymorphisms in soybean lines varying in oil composition and content. *BMC Genomics* 2014, 15:299.

[191]. Komar AA: Genetics. SNPs, silent but not invisible. *Science* 2007, 315(5811):466-467.

[192]. Katsonis P, Koire A, Wilson SJ, Hsu TK, Lua RC, Wilkins AD, Lichtarge O: Single nucleotide variations: biological impact and theoretical interpretation. *Protein Sci* 2014, 23(12):1650-1666.

[193]. Haun WJ, Hyten DL, Xu WW, Gerhardt DJ, Albert TJ, Richmond T, Jeddeloh JA, Jia G, Springer NM, Vance CP *et al*: The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol* 2011, 155(2):645-655.

[194]. Lee YG, Jeong N, Kim JH, Lee K, Kim KH, Pirani A, Ha BK, Kang ST, Park BS, Moon JK *et al*: Development, validation and genetic analysis of a large soybean SNP genotyping array. *Plant J* 2015, 81(4):625-636.

[195]. Ulker B, Li Y, Rosso MG, Logemann E, Somssich IE, Weisshaar B: T-DNA-mediated transfer of Agrobacterium tumefaciens chromosomal DNA into plants. *Nat Biotechnol* 2008, 26(9):1015-1017.

[196]. Simo C, Ibanez C, Valdes A, Cifuentes A, Garcia-Canas V: Metabolomics of genetically modified crops. *Int J Mol Sci* 2014, 15(10):18941-18966.

[197]. Ammann K: Genomic misconception: a fresh look at the biosafety of transgenic and conventional crops. A plea for a process agnostic regulation. *N Biotechnol* 2014, 31(1):1-17.

[198]. Herman RA, Price WD: Unintended compositional changes in genetically modified (GM) crops: 20 years of research. *J Agric Food Chem* 2013, 61(48):11695-11701.

[199]. Bawa AS, Anilakumar KR: Genetically modified foods: safety, risks and public concerns-a review. *J Food Sci Technol* 2013, 50(6):1035-1046.

[200]. Schauzu M, Poting A, Rubin D, Lampen A: [Assessment of allergenicity of genetically modified food crops]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2012, 55(3):402-407.

[201]. Domingo JL, Gine Bordonaba J: A literature review on the safety assessment of genetically modified plants. *Environ Int* 2011, 37(4):734-742.

[202]. Trials EGPWGoAF: Safety and nutritional assessment of GM plants and derived food and feed: the role of animal feeding trials. *Food Chem Toxicol* 2008, 46 Suppl 1:S2-70.

[203]. Nelson KA, Renner KA: Soybean Growth and Development as Affected by Glyphosate and Postemergence Herbicide Tank Mixtures. *Agron J* 2001, 93(2):428-434.

[204]. Stark JD, Chen XD, Johnson CS: Effects of herbicides on Behr's metalmark butterfly, a surrogate species for the endangered butterfly, Lange's metalmark. *Environ Pollut* 2012, 164:24-27.

[205]. Duke SO: Taking stock of herbicide-resistant crops ten years after introduction. *Pest Manag Sci* 2005, 61(3):211-218.

[206]. Harrigan GG, Lundry D, Drury S, Berman K, Riordan SG, Nemeth MA, Ridley WP, Glenn KC: Natural variation in crop composition and the impact of transgenesis. *Nat Biotechnol* 2010, 28(5):402-404.

[207]. Domingo JL: Toxicity studies of genetically modified plants: a review of the published literature. *Crit Rev Food Sci Nutr* 2007, 47(8):721-733.

[208]. Young GJ, Zhang S, Mirsky HP, Cressman RF, Cong B, Ladics GS, Zhong CX: Assessment of possible allergenicity of hypothetical ORFs in common food crops using current bioinformatic guidelines and its implications for the safety assessment of GM crops. *Food Chem Toxicol* 2012, 50(10):3741-3751.

[209]. Goodman RE, Tetteh AO: Suggested improvements for the allergenicity assessment of genetically modified plants used in foods. *Curr Allergy Asthma Rep* 2011, 11(4):317-324.

[210]. Goodman RE, Panda R, Ariyarathna H: Evaluation of endogenous allergens for the safety evaluation of genetically engineered food crops: review of potential risks, test methods, examples and relevance. *J Agric Food Chem* 2013, 61(35):8317-8332.

[211]. El Ouakfaoui S, Miki B: The stability of the Arabidopsis transcriptome in transgenic plants expressing the marker genes nptII and uidA. *Plant J* 2005, 41(6):791-800.

[212]. Aigner B, Pambalk K, Reichart U, Besenfelder U, Bosze Z, Renner M, Gunzburg WH, Wolf E, Muller M, Brem G: Species-specific alternative splicing of transgenic RNA in the mammary glands of pigs, rabbits, and mice. *Biochem Biophys Res Commun* 1999, 257(3):843-850.

[213]. Kyndt T, Quispe D, Zhai H, Jarret R, Ghislain M, Liu Q, Gheysen G, Kreuze JF: The genome of cultivated sweet potato contains Agrobacterium T-DNAs with expressed genes: An example of a naturally transgenic food crop. *Proc Natl Acad Sci U S A* 2015, 112(18):5844-5849.

[214]. Kuiper HA, Kleter GA, Noteborn HPJM, Kok EJ: Substantial equivalence—an appropriate paradigm for the safety assessment of genetically modified foods? *Toxicology* 2002, 181–182:427-431.

PUBLICATIONS

1. **Lambirth K**, Whaley A, Blakley I, Schlueter J, Bost K, Loraine A et al. A Comparison of transgenic and wild type soybean seeds: analysis of transcriptome profiles using RNA-Seq. BMC Biotechnology. 2015;15(1):89.

2. **Lambirth KC**, Whaley AM, Schlueter JA, Bost KL, Piller KJ. CONTRAILS: A tool for rapid identification of transgene integration sites in complex, repetitive genomes using low-coverage paired-end sequencing. Genomics Data. 2015;6:175-81. doi:http://dx.doi.org/10.1016/j.gdata.2015.09.001.

3. Bost KL, **Lambirth KC**, Hudson LC, Piller KJ. Soybean-Derived Thyroglobulin As an Analyte Specific Reagent for In Vitro Diagnostic Tests and Devices. Advances in Medicine and Biology. 2014;80.

4. Hudson LC, **Lambirth KC**, Bost KL, Piller KJ. Advancements in Transgenic Soy: From Field to Bedside. A Comprehensive Survey of International Soybean Research - Genetics, Physiology, Agronomy and Nitrogen Relationships. 2013.

5. Powell R, Hudson LC, **Lambirth KC**, Luth D, Wang K, Bost KL et al. Recombinant expression of homodimeric 660 kDa human thyroglobulin in soybean seeds: an alternative source of human thyroglobulin. Plant Cell Rep. 2011;30(7):1327-38. doi:10.1007/s00299-011-1044-8.