MODERN BIOINFORMATICS AS A TOOL TO UNDERSTAND GENOMIC AND TRANSCRIPTOPMIC VARIATION IN LEGUMES

A DISSERTATION SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

JEAN-MICHEL STANISLAS MICHNO

IN PARTIAL FULFULLMENT OF THE REQUIERMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisors
Robert M. Stupar and Chad L. Myers

July of 2018

ACKNOWLEDGMENTS

This work would not have been possible without the help and contributions from all of my collaborators presented here. In particular, I would like to thank Dr. Junqi Liu, Adrian Stec, and Joseph Jeffers for their contributions to my work. Outside of the work presented here, I would like to thank Dr. Tom Kono, Dr. Michael Kantar, Dr. Robert Schaefer, and Dr. Shaun Curtin for providing me with opportunities to expand my knowledge and skillset.

I would also like to express my gratitude to my committee members Dr. Candice Hirsch and Dr. Frank Albert for their insight. They helped me view my research from different perspectives, which has greatly improved my work product.

Most importantly, I would like to thank my two advisors, Dr. Chad Myers and Dr. Robert Stupar. Dr. Myers has provided me with opportunities and invaluable guidance to help me develop my computational skills. His class was my first introduction to bioinformatics. After taking his course, I fell in love with the subject and knew that it was the career I wanted to pursue.

I would also like to credit Dr. Robert Stupar for being the first person to introduce me to research. Giving me that opportunity as an undergraduate student back in 2008 changed my life, and I can never thank you enough for that. From waking up at 4 a.m. to extract RNA as an undergraduate student, to getting the opportunity to manage the soybean transformation pipeline as a lab technician, to getting my first project as a graduate student in your lab, I would have never accomplished everything today without the help and support from you throughout the years.

Finally, I would like to thank my wife, Ana Michno, as well as my parents, Michel and Francoise Michno, for their help during my graduate studies. Without them, I would not have had the support and capacity to pursue a career in science.

Abstract

The research presented here focuses on the deployment of modern bioinformatics to gain a greater understanding of legume genomes and gene functions. While improvement of legume crops still relies on conventional breeding approaches, transgenesis, the introduction of a foreign piece of DNA in a host genome, is becoming increasingly common. Using a transgenic approach, the integration of foreign DNA into the host genome using Agrobacterium-mediated transformation is almost always random and is known to induce mutations at the insertion site, but questions have been raised about the potential for mutagenesis at other loci. While genetic engineering has been widely used for crop improvement, few studies have addressed the genome-wide effects of transgenesis. Chapters two and three of this thesis address this question in the context of Glycine max, a major agricultural crop (soybean). Specifically, chapter two features a reanalysis of data from a previous study that reported a large number of mutations in soybean transgenic plants and describes several factors that led to an overestimation. Chapter three addresses the effects on the genome in a series of soybean plants transformed with CRISPR/Cas9, the most recently developed platform for genome editing. The findings of this work have implications on the frequency and transmission of novel variation resulting from soybean biotechnology. Chapter four focuses on applying transcriptome network analysis for predicting the genes that underlie nodule development variation in the Medicago-Ensifer symbiosis. Co-expression networks were constructed for Medicago truncatula and were integrated with data from genome-wide association analysis to prioritize candidate genes with a high likelihood of causal association with nodule development phenotypes. This approach sheds light on potential new genetic factors underlying an important phenotype, and

more broadly, could be applied to understand genomic and phenotypic variation for a wide range of plant species and traits.

Table of Contents

List of Tables

List of Figures

Chapter 1: The use of sequencing technologies to understand transgenic soybeans and gene expression patterns in agricultural species

The sections involving co-expression uses in agricultural species were written by Jean-Michel Michno as part of the following review paper published in Biochimica et Biophysica Acta (BBA)

Schaefer, R.J., Michno, J.M. and Myers, C.L., 2017. Unraveling gene function in agricultural species using gene co-expression networks. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 1860(1), pp.53-63.

Advances in sequencing technologies

Over the past fifteen years, there has been a huge surge in the adoption of next-generation sequencing (NGS) technologies across a variety of research areas. The introduction of these high-throughput technologies has changed the way researchers think about scientific approaches, similar to the paradigm shift that occurred when the polymerase chain reaction (PCR) was first introduced (Metzker, 2010). NGS technologies generate millions of sequences in parallel fashion rather than 96 at a time through traditional Sanger-based methods (Sanger and Coulson, 1975).

Roche(454), Illumina and SOLiD were the first wave of platforms that were able to generate parallel short reads ranging from 35 - 300 bp in size (Ronaghi et al., 1996; Levene, 2003; Bentley et al., 2008; Mardis, 2008; Turcatti et al., 2008). As these technologies progressed, they became more affordable, allowing researchers to either sequence samples at a deeper coverage or to expand the number of samples analyzed. These data, in turn, were then used to map genomic reads to a reference genome/scaffold and identify single nucleotide polymorphisms (SNPs), insertions/deletions (indels), copy number variants (CNV), and/or structural variants (SVs). Furthermore, NGS data were used to generate reference genomes, quantify expression values, call variants using RNA-seq, and/or analyze epigenetic modifications (Morozova and Marra, 2008).

While short-read technologies have contributed towards addressing a plethora of research questions, they are not without their limitations. Some limitations include the inability to resolve some repetitive regions, differentiate segmental duplications, and resolve complex structural variants. Therefore, long-read technologies (despite carrying

a higher cost per base) are now used to aide these analyses or serve as an alternative (Alkan et al., 2011; Treangen and Salzberg, 2012; Daber et al., 2013). Single molecule technologies have allowed researchers to generate reads as long as 100 kb using PacBio technologies and over 1 Mb using Nanopore MinION (Eid et al., 2009; Schneider and Dekker, 2012; Jiao et al., 2017; Jain et al., 2018). While these technologies are able to generate long-reads, their greatest limitation is their accuracy, which is much lower that short-read technologies (Goodwin et al., 2016). The error rate of PacBio is currently between 13-15% (Weirather et al., 2017) while Nanopore is between 15-30% (Wang et al., 2015; Morisse et al., 2017). Synthetic technologies, which split large DNA fragments into smaller barcoded ones, allow for local assembly using short-read technologies. Illumina synthetic long-reads and 10x Genomics (Voskoboynik et al., 2013; McCoy et al., 2014) have the ability to sequence large fragments up to ~100kb in size using this strategy (Koren and Phillippy, 2015; Goodwin et al., 2016). Although these synthetic long-read technologies are more accurate than single-molecule technologies, they rely on local assembly which can create challenges for distinguishing repeat structures (Koren and Phillippy, 2015).

These advances in short and long-read sequencing technologies have contributed greatly to the fields of structural and functional genomics. The technologies have provided an unprecedented resolution of genomes, transcriptomes, and epigenomes, answering old questions while enabling new questions. One way these technologies have been applied is to understand the variation induced by creating transgenic crops.

Transgenics and sequencing in soybean: Understanding the aftermath

Transgenic crops have played an important role in the agricultural industry. From generating herbicide resistant lines to altering seed composition traits, these methods transformed modern-day agricultural practices. While this technology has been successful in integrating a variety of traits such as herbicide tolerance and pest resistance (Shah et al., 1985; Padgette et al., 1995; Estruch et al., 1997; Vaughn et al., 2005), they are subject to tight regulation (Kessler et al., 1992; Mauro Vigani, 2015; Strauss and Sax, 2016). These regulations serve as a way to ensure that genetically modified organisms (GMOs) are not detrimental to the environment and are safe for consumption (Kessler et al., 1992; Potrykus, 2017).

In soybean, there are two main methods to create stable GMOs: biolistics or *agrobacterium*-based methods. Both methods serve as a way to introduce a foreign piece of DNA, known as a transgene, into random location(s) in the genome. These transgenes can either serve as a gain of function, where the transgene encodes a new functioning protein (Padgette et al., 1995), suppresses a native gene (such as through RNA-interference to knock down expression of a specific gene) (Nunes et al., 2006; Steeves et al., 2006; Flores et al., 2008; Wang and Xu, 2008; Takagi et al., 2011; Wagner et al., 2011; Zhang et al., 2011), or generates precise modifications (through using genome engineering reagents such as CRISPR/Cas9) (Kim et al., 1996; Bibikova et al., 2002; Bibikova et al., 2003; Bhaya et al., 2011; Cermak et al., 2011; Curtin et al., 2011; Sander et al., 2011a; Sander et al., 2011b; Christian et al., 2012; Curtin et al., 2012; Jinek et al., 2012; Li et al., 2012; Belhaj et al., 2013; Feng et al., 2013; Li et al.,

2013b; Puchta and Fauser, 2013; Qi et al., 2013; Schmid-Burgk et al., 2013; Shan et al., 2013a; Canver et al., 2014; Schiml et al., 2014; Baltes and Voytas, 2015; Cai et al., 2015; Jacobs et al., 2015; Michno et al., 2015; Sun et al., 2015; Tang et al., 2016; Cai et al., 2018; Curtin et al., 2018).

Biolistics is a direct gene transfer mechanism that uses high-velocity microprojectiles to introduce foreign DNA into tissues, resulting in non-homologous integration of transgenic DNA into the genome. This process commonly inserts more than one copy of foreign DNA, partial copies of foreign DNA, as well as genomic structural rearrangements (Sanford, 1988; Jackson et al., 2001; Svitashev and Somers, 2001; Makarevitch et al., 2003; Twyman and Christou, 2004; Collier et al., 2017; Jupe et al., 2018). Plants that result from biolistics are more likely to be chimeric which requires further screening in subsequent generations (Sanford, 1988; Sato et al., 1993). Unlike *agrobacterium*-mediated transformation, biolistics efficiency is less dependent on the genotype used during the transformation process and is also considered faster and less laborious (Christou, 1992; Simmonds and Donaldson, 2000; Altpeter et al., 2005; Homrich et al., 2012).

Agrobacterium-mediated transformation uses a disarmed strain of either *Agrobacterium rhizogenes* or *A. tumerfacians* as a means to deliver a vector containing a transgenic cassette (T-DNA) into the soybean host (Hinchee et al., 1988; Chee et al., 1989; Paz et al., 2006; Veena and Taylor, 2007). Although the efficiency of these methods greatly depends on the soybean host genotype, other groups have combatted this issue in monocots (Lowe et al., 2016), which may also be applicable to soybean in the future. Similar to biolistics, *agrobacterium*-mediated transformation is prone to multiple insertion events (Collier et al., 2017; Curtin et al., 2018). Furthermore, its

generation time is longer and its efficiency slightly lower compared to biolistics (Paz et al., 2006; Homrich et al., 2012; Gao and Nielsen, 2013).

Using either biolistics or *agrobacterium*-based delivery methods to deliver transgenic DNA to genomes results in random integration sites. Therefore, it is important to not only screen for the presence of the transgene but to also locate where and how often it has been integrated into the genome.

Using sequencing to detect transgenic insertion events within a genome

Currently, PCR and protein-based methods are the most common form of screening for GMOs (Taylor and Sajan, 2005; Holst-Jensen, 2009; Rosa et al., 2016; Scholtens et al., 2017). The most predominant form of GMO screening involves traditional PCR and qPCR for a segment of a transgene. Other methods such as loop-mediated isothermal amplification and ligase detection reaction have been explored as an alternative screening method to address PCR's limitations (Dong et al., 2008; Morisset et al., 2008) Protein-based methods primarily consist of immunoassays where target proteins are detected by specific antibodies (Holst-Jensen, 2009). Although, these types of screening methods are not as popular as DNA-based methods due to the cost associated with developing antibodies and the amount of effort required to set up a screen. Both DNA and protein-based methods are useful in the detection of the presence of transgenic DNA or expression, but they do not give information to where the event has inserted itself into the genome. With the uncertainty associated with integration sites, it is imperative to not only locate the integration events but to also see if they induced any other mutations throughout the genome.

With advances in next-generation sequencing, groups have gained a better understanding of not only how to detect where a transgene event has inserted, but also

the variation induced from these insertion events. Several methods have been developed to detect transgenic DNA using paired-end sequencing (Kovalic et al., 2012; Wahler et al., 2013; Srivastava et al., 2014; Lambirth et al., 2015b; Pauwels et al., 2015; Anderson et al., 2016; Guttikonda et al., 2016; Willems et al., 2016). These studies have allowed researchers to analyze the various outcomes of T-DNA integration events, such as complete T-DNA insertions, fragmented insertions, insertions within genes, duplicated insertions, multiple insertions, deletions and additions (Anderson et al., 2016; Schouten et al., 2017; Curtin et al., 2018)

While there has been plenty of interest in the detection of transgenic insertion events, relatively few studies have investigated the impacts of transgenesis genome-wide. Although there is a consensus that somaclonal variation can result from certain aspects of the transformation process (e.g., through cell introducion/maintenance in tissue culture), there has been inconsistant reports about the number of mutations induced genome-wide due to transformation (Labra et al., 2004; Jiang et al., 2011; Sabot et al., 2011; Miyao et al., 2012; Kawakatsu et al., 2013; Endo et al., 2015; Kashima et al., 2015). The majority of mutations recorded in these studies were either SNPs or indels, but they all demonstrate that variation occurs not only at the transgene insertion site but also genome-wide. The discrepancies between studies do not just pose an academic debate. Indeed, these findings can be used to inform real-world issues of safety, risk, and regulation of transgenic plants. This is particularly relevant to address whether there may be unintended consequences resulting from the development of a transgenic product, and what standards should be imposed in the process of deregulating a given event.

Current applications of co-expression networks in crops species

RNA-seq technology for expression analysis

The ability for technologies to not only capture but quantify genome-wide expression profiles has matured significantly. Measurement/quantification of these expression profiles across tissues, cultivars, and environments can give insight into their function. RNA-seq has gained traction in the community as a way to use high-throughput systems to quantify the transcriptome (Wang et al., 2009). A standard RNA-seq protocol consists of converting RNA to a library of cDNA fragments where sequencing adaptors are attached on either or both ends of the fragment. These fragments are then sequenced using a form of high-throughput technology and then processed later on by mapping reads to a reference genome then quantifying expression levels as fragments per kilobase per million reads (FPKM) (Wang et al., 2009).

As with any genome-scale technology, RNA-seq has limitations, which must be considered in interpreting the resulting data. Some of the limitations include the underrepresentation of small transcripts in RNA-seq libraries, difficulty in assigning overlapping transcripts between genes, and mapping reads among conserved paralogous genes (Hirsch et al., 2015). This paralogy issue is particularly problematic in plant species, as most crop genomes are either polyploid or have experienced a relatively recent whole genome duplication event and have maintained multiple copies of the same gene.

Co-expression networks in agricultural species

Historically, large scale microarray experiments have extensively been used to characterize co-expression networks in the plant model *Arabidopsis thaliana* (Schena et

al., 1995). Broadly, experiments surveyed genome-wide expression response to abiotic stress (Kilian et al., 2007), plant tissue and development (Schmid et al., 2005) as well as response to different plant hormones (Goda et al., 2008). In general, analysis of patterns of gene expression identified an enrichment for co-expression among genes that are co-annotated in both KEGG terms (Lee, 2003) as well as the Gene Ontology (Schmid et al., 2005). These proof of principle approaches blazed in model systems naturally extend to analogous experiments performed in agricultural species.

To date, gene co-expression networks have been used to rapidly predict gene function in many non-model plant species, many of which have agricultural importance. Most networks were built with the purpose of discovering and characterizing highly connected subnetworks or *modules* to better understand various phenotypes or functions or to provide a general resource to the community. Such studies have now been completed for a variety of agronomic species including soybean, poplar, grape, alfalfa, rice, maize, tomato, and barley (Zhu et al., 2002; Ficklin et al., 2010; Ozaki et al., 2010; Ficklin and Feltus, 2011; Mochida et al., 2011; Fukushima et al., 2012; Swanson-Wagner et al., 2012; Obayashi et al., 2014; Schaefer et al., 2014). A typical study involves using either publicly or self-generated expression data to build a network, looking for functional enrichment within modules, then focuses directly on a module of interest for biological interpretation.

Changes in gene expression can occur from introducing variation stemming from many different sources. As with model species, this variation can be examined in the context of co-expression networks to assess the putative functional impact of different experimental conditions in crop species. Recent studies performed in crops generally survey gene expression variation that arises from several major sources reviewed here:

changes in environmental conditions; developmental and organ based variation; and variation due to population and ecological dynamics.

Co-expression in response to environmental conditions

Co-expression networks can be built with the intent of discovering various modules in response to environmental conditions. Factors such as diseases and/or abiotic stressors can have a very broad impact on the development and phenotype of a plant, causing expression profile changes that differ among individuals with different genetic variation or those exposed to different environmental conditions. Co-expression networks can be built by measuring transcriptomic changes under these conditions to generate environmental specific co-expression networks (Zheng and Zhao, 2013; Sarkar et al., 2014). For example, Mochida *et al.* assembled a global co-expression network from 1,347 experiments surveying both diverse environmental conditions and stresses in barley (Mochida et al., 2011). They discovered functional modules using gene ontology enrichment as well as Triticeae-specific network modules using comparative approaches.

Zheng et al. studied citrus response to one of its most destructive diseases, Huanglongbing, more commonly known as citrus greening (Zheng and Zhao, 2013). Using various transcriptome datasets and a set of genes that are up-regulated in early stages of inoculation, they were able to construct and identify several modules that provided insights into the mechanism of immune response to citrus greening. Sarkar et al. studied rice's response to heat stress discovering various genes/modules that could help provide insight in mechanistic changes in response to stress (Sarkar et al., 2014). Gene expression profiles from rice exposed to two different durations of heat stress were

used to discover several modules consisting of functionally correlated genes, which included previously documented genes involved in heat stress.

Tissue and developmental atlases for agricultural crops

Different developmental time points and/or tissues have also been used to build and study modules within a co-expression network. Just like different environmental conditions, different developmental time points and/or tissues can have varying expression patterns. Expression data from these categories have been applied to a variety of agricultural crops to look for tissue enriched modules (Brady et al., 2007; Fu and Xue, 2010; Childs et al., 2011; Sekhon et al., 2013; Zheng and Zhao, 2013; Schaefer et al., 2014; Cho et al., 2016). As an example, Ozaki et al. built a tomato network using microarray data and discovered that 75 out of their 199 modules had significant functional enrichment (Ozaki et al., 2010). They further investigated a module related to the flavonoid biosynthetic pathway and compared genes within that module to a transgenic cultivar identifying genes that were up-regulated in flavonoid biosynthesis.

Downs *et al.* and Sekhon *et al.* sampled the expression of different tissue types and developmental stages to discover tissue enriched modules (Sekhon et al., 2011; Downs et al., 2013). Downs *et al.* used maize tissue from 50 different developmental time points to build developmental co-expression networks. Using those tissues, they were able to discover 24 modules where subsets of genes were associated with specific tissues or different developmental stages of a tissue. Sekhon *et al.,* similarly, surveyed 60 tissue/time points and found that many genes displayed organ specific expression patterns (Sekhon et al., 2011). Schaefer *et al*. leveraged co-expression networks to identify general functional modules derived from both developmental as well as genotypic diversity. They built networks using publicly available developmental and

genotypic expression data in maize to not only show functional enrichment within modules but also to demonstrate differences across networks as a way to capture biological function.

Comparative based co-expression approaches

Co-expression networks can also be built with the purpose of discovering conserved modules across species. Even though the majority of genes for agronomic species do not have a functional annotation, investigators can use modules composed of well-defined functionally annotated genes from one species to compare and discover similar modules in another. These comparisons are not only limited to using co-expression networks from model species such as Arabidopsis (Ruprecht et al., 2011; Leal et al., 2014; Obertello et al., 2015; Righetti et al., 2015), but can be applied using other species if they are using genes with well-defined functional annotations (Swanson-Wagner et al., 2012; Itkin et al., 2013).

Obertello et al. built rice and Arabidopsis nitrogen regulatory networks to discover genes that are directly related to nitrogen use. They first identified genes that were induced or repressed by nitrogen regulators using both rice and Arabidopsis expression data. Conserved, nitrogen-related gene clusters and predicted transcription factors were identified by creating a cross-species functional network that combined putative protein-protein and regulatory interaction from both rice and Arabidopsis, including data from gene expression, using orthologous genes to share interaction evidence across species. Similarly, Itkin *et al.* used comparative co-expression network analysis to identify genes related to steroidal glycoalkaloids (SGA), a toxic substance found in some tubers and tomatoes (Itkin et al., 2013). Using homologs (SGT1/GAME1) in tomato and potato as seed genes, the authors discovered highly co-expressed genes in each species that are

involved in SGA biosynthesis. Targeting these co-expressed genes, they were able to successfully reduce SGA concentrations in both species using virus-induced gene silencing.

Evolutionarily conserved functional modules can also be discovered by integrating a wide variety of species-specific co-expression networks (Ruprecht et al., 2011). More recently, Ruprecht *et al.* used multiple plant species together to discover a wide variety of conserved modules, with an emphasis on modules related to cell wall formation (Ruprecht et al., 2016). In a similar approach, Leal *et al.* used gene expression data from Arabidopsis, Maize, soybean, rice, tomato, and cassava to identify conserved genes involved in immune response under pathogenic stress (Leal et al., 2014). Using the networks that were built from the expression data, they were able to find functional similarities in the immune response across species.

Co-expression analysis has also been leveraged to gain meaningful insight between a domesticated agricultural species and its crop wild relative. For example, Swanson-Wagner *et al.* used 38 diverse maize genotypes and 24 teosinte genotypes to generate separate co-expression networks for maize and teosinte, its wild relative. They were able to identify gene clusters that were rewired between maize and its crop wild ancestor, suggesting modulated regulation could have played a role during domestication (Swanson-Wagner et al., 2012). The genes identified by this differential co-expression analysis complemented those identified through standard differential expression analysis. Cho et al. used transcriptome and metabolite data to profile sprouts from three evolutionarily divergent potato cultivars to characterize genes involved in anthocyanin production. Comparing these data, they were able to find 119 genes that were strongly correlated with anthocyanin-related metabolites (Cho et al., 2016).

As newer technologies increase the output and precision of both transcriptional as well as other types of functional data, interdisciplinary collaborations will be more important than ever. Furthermore, as the techniques and tools used to develop and analyze co-expression networks mature, publicly available datasets will play an integral role in profiling the functions of genes across many different experimental contexts. These advances will give opportunities to increase agricultural output and address the future demands of the global food supply.

Chapter 2: The importance of genotype identity, genetic heterogeneity, and bioinformatic handling for properly assessing genomic variation in transgenic plants

Preface

Background: The advent of –omics technologies has enabled the resolution of fine molecular differences among individuals within a species. DNA sequence variations, such as single nucleotide polymorphisms or small deletions, can be tabulated for many kinds of genotype comparisons. However, experimental designs and analytical approaches are replete with ways to overestimate the level of variation present within a given sample. Analytical pipelines that do not apply proper thresholds nor assess reproducibility among samples are susceptible to calling false-positive variants. Furthermore, issues with sample genotype identity or failing to account for heterogeneity in reference genotypes may lead to misinterpretations of standing variants as polymorphisms derived *de novo*.

Results: A recent publication that featured the analysis of RNA-sequencing data in three transgenic soybean event series appeared to overestimate the number of sequence variants identified in plants that were exposed to a tissue culture based transformation process. We reanalyzed these data with a stringent set of criteria and demonstrate three different factors that lead to variant overestimation, including issues related to the genetic identity of the background genotype, unaccounted genetic heterogeneity in the reference genome, and insufficient bioinformatics filtering.

Conclusions: This study serves as a cautionary tale to users of genomic and transcriptomic data that wish to assess the molecular variation attributable to tissue culture and transformation processes. Moreover, accounting for the factors that lead to sequence variant overestimation is equally applicable to samples derived from other germplasm sources, including chemical or irradiation mutagenesis and genome engineering (e.g., CRISPR) processes.

This work was published in BMC Biotechnology in June 2018, with full citation information given below. This work was a collaborative effort, with Jean-Michel Michno and Dr. Robert Stupar. JMM and RMS designed the experiment, JMM performed the analysis, and JMM and RMS wrote the manuscript.

Michno, JM and Stupar, RM, 2018. The importance of genotype identity, genetic heterogeneity, and bioinformatic handling for properly assessing genomic variation in transgenic plants. *BMC Biotechnology*, 18(1), p.38.

Background

The process of genetic transformation typically involves inserting DNA sequences originating from one species into the genome of another species. This tool has been used to add traits into crop species, such as herbicide tolerance in soybean and root worm tolerance in corn (Shah et al., 1986; Padgette et al., 1995; Estruch et al., 1997; Vaughn et al., 2005). The commercialization of transgenic products is subject to tight regulation, as transgenic strains must undergo intense safety testing before being brought to market (James and Krattiger, 1996). The testing phase involves confirmation of the intended trait encoded by the transgene, and confirmation that the transgenic plant does not have unintended consequences that may be detrimental to the environment or to the consumer (Kessler et al., 1992). Adverse effects are generally characterized in two categories: effects from the transgene itself, and effects that arise from mutations resulting from gene insertion or the tissue culture process. As a result, safety testing ensures that unintended DNA-level changes are not present in commercialized products (Weber et al., 2012; Glenn et al., 2017).

With the recent revolution in high-throughput sequencing technology, there is now increased interest in understanding the molecular nature of transgenic events, and identifying possible safety implications of unintended molecular changes that may result. This information may be useful in assessing the likelihood that a particular event will express the intended trait(s) without detrimental unintended effects.

Molecular studies have previously characterized the effects of transgenesis in several different plant species, focusing on the sequence changes at transgene integration sites (Nacry et al., 1998; Clark and Krysan, 2010) and/or the sequence

changes genome-wide (Labra et al., 2004; Jiang et al., 2011; Sabot et al., 2011; Miyao et al., 2012; Kawakatsu et al., 2013; Zhang et al., 2014; Endo et al., 2015; Kashima et al., 2015; Schouten et al., 2017). While no clear consensus has emerged, studies utilizing sequence-level resolution have reported a range of possible sequence changes in transgenic plants, including frequent observations (e.g., small deletions occuring adjacent to the integration site) and less frequent occurances (e.g., translocations between chromosomes).

A curious discrepancy in genome-wide sequence polymorphisms has been observed in recent resequencing studies of transgenic soybean. One study, published by our group (Anderson et al., 2016), resequenced two independent transgenic T1 plants, and respectivley found only two and 18 single nucleotide polymorphisms (SNPs) genome-wide (along with deletions adjacent to the integrated transgene, as has been previously observed in other plant transformation studies). In contrast, Lambirth et al. (Lambirth et al., 2015a; Lambirth et al., 2016) reported high rates of molecular variation among transgenic soybean plants, both in terms of transcriptomic changes and DNA sequence changes. The authors analyzed RNA-sequencing (RNA-seq) data on families from three different transgenic events and reported thousands of sequence variants per plant, focusing on SNPs and small insertion-deletion (indel) variants. They reported tens of thousands of sequence variants in these plants, including approximately 1,000 to 7,700 variants that were unique to each of the three event series. This contrast between studies is even more surprising considering that Anderson et al. 2016 searched genome-wide while Lambirth et al. 2016 searched only the transcribed portion of the genome. Both groups were studying the same species (soybean) transformed by similar methods

(*Agrobacterium*-mediated transformation of cotyledonary nodes) (Paz et al., 2006) and resequenced using similar chemistries (Illumina short-read).

Given the importance and real-world relevance of this topic, it is imperative to resolve the discrepancy between the Anderson et al. 2016 and Lambirth et al. 2015 and 2016 studies. We are not aware of any transgenic resequencing studies that have reported mutation rates similar to those published by Lambirth et al. 2016. Therefore, the current study focuses on a reanalysis of the Lambirth et al. 2016 dataset, applying a more stringent analytical pipeline. The outcome of this reanalysis demonstrates that the Lambirth et al. 2015 and 2016 studies overestimated the transcriptional and DNA sequence variation in the transgenic plants. These findings provide insight into the importance of identity preservation of genotypes, awareness of genomic heterogeneity within cultivars, and leveraging bioinformatics filters and replicated data as a way to minimize false positives.

Results and Discussion

Primary source of variation in transgenic event series 764: incorrectly identified genetic background

Lambirth et al. 2015 and 2016 performed RNA-seq analyses of 27 transgenic plants, including nine individuals each selected from three different transgenic series known as ST77, ST111, and 764. They reported that all three of these transgenic series were developed in the genetic background of cultivar ‘Williams 82’. As a control, they also performed RNA-seq on nine individuals of ‘Williams 82’, thus resulting in a total of 36 RNA-seq samples in the study. As ‘Williams 82’ was also the genotype used to develop the soybean reference genome (Schmutz et al., 2010), all of the mutations reported by Lambirth et al. 2016 were identified simply by comparing their transcriptome sequence to the reference genome. The authors reported surprisingly high mutation

frequencies in both the transgenic and control plants, particularly the 764 transgenic event series. As *de novo* mutations caused by the tissue culture or transgenesis pathway are expected to be unique to a given event, the authors calculated the number of unique event-specific mutations in each series compared to the other groups/series in the study (i.e., the number of mutations in one series that is not shared by the other two series of transformants or the control 'Williams 82' plants). They reported a unique polymorphic SNP count of 981 in event ST77, 927 in event ST111, and 7,717 in event 764. This discrepancy matched their earlier analysis of gene expression variation among three series, where series 764 exhibited much greater expression variation as compared to controls than did the other two transgenic groups (Lambirth et al., 2015a).

Two findings in the Lambirth et al. 2016 mutation analysis stand out: (1) The SNP frequencies were much higher than other similar studies of soybean (Anderson et al., 2016) and model plant species wide (Labra et al., 2004; Jiang et al., 2011; Sabot et al., 2011; Miyao et al., 2012; Kawakatsu et al., 2013; Zhang et al., 2014; Endo et al., 2015; Kashima et al., 2015; Schouten et al., 2017), particularly considering that only the transcribed portion of the genome was being analyzed; (2) Even with the generally high mutation rates reported, the 764 series is still an outlier. To cross-validate the findings of this analysis, we downloaded and reanalyzed the raw RNA-seq data from these studies.

Using the GATK Best Practices workflow (DePristo et al., 2011; Van der Auwera et al., 2013), we re-generated polymorphic SNP lists from all 36 samples of RNA-seq data used (Lambirth et al., 2015a; Lambirth et al., 2016). As stated above, *de novo* SNPs generated by tissue culture or transformation would be expected to be unique to each respective transgenic event. Therefore, we focused our analysis on SNPs that were unique to only one of the four groups (e.g., SNPs observed as an alternative base

in one transgenic series, while matching the reference genome sequence in the other two transgenic series and the 'Williams 82' controls). Given that the transgenic plants were self-pollinated for several generations after transformation, the SNPs derived from the tissue culture or transformation process are expected to be predominantly homozygous. Therefore, we filtered our initial lists for homozygous SNPs that are uniquely polymorphic relative to the reference genome, compared to the other transgenic lines and 'Williams 82' controls (Figure S1 in Appendix 1). This analysis and filtering pipeline differed from the Lambirth et al. 2016 pipeline in at least four critical ways: (1) The GATK Best Practices workflow imposed a higher standard for calling variants (see Methods section); (2) we did not include heterozygous calls; (3) we did not include heterogeneous SNPs among the nine samples of any group (the three transgenic series or controls); (4) we required at least six out of the nine samples within each group to exhibit the same homozygous base call.

The analysis and filtering pipeline described above was designed to prevent false-positive SNP calls. Nevertheless, the pipeline was able to detect nearly 10,000 SNPs among the transgenic samples (Table 1). However, the distribution of SNPs among the genotypes was substantially different than what was reported previously (Lambirth et al., 2016). Almost all of the unique SNPs that we identified were found in transgenic series 764 (9,738 out of the 9,884 SNPs). Meanwhile, only 143 and 3 SNPs, respectively, were identified in ST77 and ST111 (Table 1).

We postulated that the discrepancy exhibited by the 764 series might have resulted from experimental error rather than biological factors. To test this, we compared the list of SNPs we generated (Table 1) with a list of pre-ascertained SNPs that were previously used to genotype the entire USDA soybean germplasm collection (Song et

al., 2013). We found that 525 of the SNPs that were unique to series 764 also matched the genome positions on the pre-ascertained SNP list (Table 1). We compared the SNP profile of these 525 SNPs for series 764 with all of the accessions in the USDA collection. One genotype, cultivar 'Thorne' (PI 564718) (McBlain et al., 1993), was a nearly perfect match to series 764 (521 of the 525 SNPs match; Figure 1). The four SNPs that did not match between series 764 and 'Thorne' were clustered together between positions ~4.9 Mb and ~5.9 Mb on chromosome 15. It is likely that this interval on chromosome 15 represents a region of genetic heterogeneity between the individual of 'Thorne' used for transformation in the development of the 764 event and the individual(s) of 'Thorne' sampled for the USDA genotyping effort (Song et al., 2013). While the series 764 profile was a 99.2% match to 'Thorne' across the 525 SNPs, the next closest match was 'Washita' (PI 618809) (Farno et al., 2003), which was only a 74.2% match. Both 'Williams' and 'Williams 82' had a 0% match rate to the 525 SNPs in the 764 series (Figure 1), as would be expected because the reference genome is based on 'Williams 82' and these SNPs were initially identified as polymorphic between the 764 series and the reference genome.

The clear conclusion from this analysis is that series 764 was developed in 'Thorne,' rather than 'Williams 82'. 'Thorne' is commonly used for soybean transformation (e.g., (Paz et al., 2006)). It is clear that the high polymorphism rate reported in event series 764 is not an unintended consequence of tissue culture or transgenesis. Instead, the majority (if not all) of the variation reported in this line is simply standing variation that exists between 'Thorne' and 'Williams 82'. This statement can be applied to all previous reports of variation observed between these plants,

including gene transcription (Lambirth et al., 2015a), mutations (Lambirth et al., 2016), or any other characteristic.

Source of variation in transgenic event series ST77: genetic heterogeneity between different individuals of 'William 82.'

The relatively lower polymorphism rates found in the reanalysis of S77 and S111 compared to that of 764 (Table 1) indicated that these groups are likely derived from the 'Williams 82' background. However, standing variation can persist within soybean cultivars (Haun et al., 2011), as the breeding process typically involves bulk harvesting of breeding populations prior to full fixation of homozygosity through single seed descent. Therefore, most soybean cultivars are expected to exhibit slight differences from plant to plant (Fasoula and Boerma, 2005; Fasoula and Boerma, 2007), as heterogeneous sub-lines fix different haplotypes within relatively small (but sometimes large) genomic intervals. For example, previous genotyping of four different 'Williams 82' sub-lines revealed specific regions of genomic variation on chromosomes 3, 7, 15 and 20 (Haun et al., 2011).

It is relatively intuitive to identify genomic heterogeneity between sub-lines of a cultivar, as sub-lines will show nearly complete homogeny throughout the genome, interrupted by specific regions with (sometimes dense) clusters of polymorphisms. We investigated whether the 143 SNPs identified in our reanalysis of group ST77 could be explained by this type of standing heterogeneity between the 'Williams 82' controls used in the study and the 'Williams 82' individual that was used for the original ST77 transformation event (Lambirth et al., 2015a; Lambirth et al., 2016). Indeed, 140 of the 143 SNPs and all 16 indels were clustered at a single locus between positions 1.4 Mb and 2.2 Mb on chromosome 15 (Figure 2). This cluster overlaps with a previously reported region of heterogeneity in 'Williams 82' (Haun et al., 2011). These results

suggest that these variants are not associated with transgenesis, but represent natural standing heterogeneity between the 'Williams 82' plant used to generate the ST77 transformation event and the 'Williams 82' individuals used as controls by Lambirth et al. 2016.

Therefore, after filtering for genotype identity and background heterogeneity, we found three SNPs each in S77 and S111 that could not be explained by these factors. Follow-up analysis of S77 revealed one SNP within an intron, one synonymous SNP within an exon, and one non-synonymous SNP within an exon (M to V amino acid change in the sixth exon of Glyma.10G150500). Analysis of S111 revealed two SNPs within introns, and one non-synonymous SNP within an exon (T to G amino acid change in the fourth exon of Glyma.04G134800).

Source of variation in all transgenic series: bioinformatics handling and threshold parameters

The previous two sections addressed our reanalysis of RNA-seq data (Lambirth et al., 2015a; Lambirth et al., 2016), focusing on the subset of unique SNPs and indels within any one transgenic series. However, the majority of the analysis reported, discussed and interpreted in the Lambirth et al. 2016 paper (including the base substitution profile, the predicted effect of each polymorphism, and gene ontology enrichment analysis) used the original full set of SNPs and indels identified, rather than the "unique" subset. Hence it is necessary to focus on the factors that inflated the overall higher number of SNPs and indels discovered by their bioinformatic pipeline. While we would expect the authors to identify polymorphisms due to the reasons outlined in the previous sections (e.g., the 'Thorne' background of series 764 and the genetic heterogeneity between ST77 and the control 'Williams 82' plants), the reported polymorphism counts were unexpectedly high. For example, the plants in the 764 series

averaged 38,188 SNPs and 2,390 indels per plant. Obviously, this number will be higher than the other two transgenic series because it is the 'Thorne' genetic background. However, the ST77 series averaged 21,666 SNPs and 1,829 indels, and the ST111 series averaged 20,208 SNPs and 1,750 indels. Furthermore, the untransformed 'William 82' control plants exhibited counts of 20,707 SNPs and 1,863 indels. Therefore, this section is devoted to addressing the sources of these high estimates.

We retrieved the variant calls for each of the 36 samples used in their analysis (http://de.iplantcollaborative.org/dl/d/533570A3-1EFB-4864-B9A9-9D82F17E09A8/snpeffgenes.zip). Initial analyses of genotype calls revealed that there was a higher number of heterozygous variants than homozygous variants for the alternate allele compared to the reference genome. ST77 and ST111 were respectively advanced to the T8 and T4 generation before sequencing. We can estimate the expected proportion of heterozygous variants in these generations if we assume the following: all of the mutations induced by transgenesis were heterozygous in the T0 generation, the variants are not subject to segregation distortion, and the variants have negligible effects on organismal fitness. Under these assumptions, we would expect approximately 0.39% of the ST77 variants to be heterozygous at the T8 generation, and 6.25% of the ST111 variants to be heterozygous at the T4 generation. However, the retrieved data showed that 50.21% and 48.62% of the variants were called as heterozygous for ST77 and ST111, respectively. The proportion of heterozygous variants were far higher than what was expected, and were most likely false positives resulting from the analysis method.

We further investigated whether the authors filtered their variants for read depth and/or quality. Although read depth alone is not sufficient to determine whether a variant

is real, calls based on low read depth are more likely to be false positives than calls based on higher read depths. False positives can arise from reads that map poorly to the genome, or bases that are of low quality at the site of a polymorphism. When analyzing the depth of variant calls for all 36 samples in the study, 43.2% of variants were called at a depth of one read, and 20.2% of variants were called with a depth of two reads (Figure 3). Similarly, when analyzing the distribution of quality scores across all 36 samples, 55.3% of variant calls had a quality score of 10 or lower (Figure S2 in Appendix 1). A quality score is represented on a log-based Phred scale where, for example, a quality score of 10 indicates that there is a 10% chance of the variant being incorrect and a quality score of 20 indicates that there is a 1% chance of the variant being incorrect. Further investigation into the authors' methods revealed that the variant calls lacked any type of depth or quality filter. This further reinforces the likelihood that a large portion of these variants at low depth and quality are most likely false positives.

The experiments in these studies (Lambirth et al., 2015a; Lambirth et al., 2016) included the sequencing of nine samples per transgenic series (or the 'Williams 82' controls), consisting of three sibling seeds taken from three plants each. As mutations induced by transformation or tissue culture would presumably occur in the T0 generation, one would expect the vast majority of these loci to be fixed as homozygotes by the T4-T8 generations. Therefore, it may be intuitive to exclude any variants that were not observed in all three siblings. While the authors reported on average ~20,000 SNPs and ~1,800 indels per individual plant for ST77, ST111, WT, and ~40,000 SNP's and ~2,400 indels per individual plant for 764 compared to the reference genome, the majority of variants were detected as polymorphic in only one of the 36 samples in the study. Figure 4A shows a comparison of the variants from three selected ST77 plants,

each derived from a different T_7 individual. In this case, over 20,000 variants were called for each plant, but only 2,807 of the variants were common across all three plants (Figure 4A). Similar findings were observed for the ST77 "D" series siblings (all derived from a T_7 plant designated as "D"), in which a relatively small proportion (4,356 out of 64,636) of the variants were in common to all three siblings (Figure 4B). These trends were observed across all sibling groups in the study (Figure S3 in Appendix 1). Series 764 exhibited a greater proportion of variants shared among the siblings, which would be expected for a plant from a different genetic background than 'Williams 82,' i.e., these plants have more "true" sequence variants that can be faithfully detected among the different siblings.

Another indication of the high frequency of false positives called in the Lambirth et al. 2016 study relates to the structure of the indels that were called as polymorphic. Of the 70,486 indels that were called, 52.9% of them were heterozygous and 59.6% of them had a read depth of 3 or less. Interestingly, all of the indels reported in the study exhibited polymorphisms that were either 1 bp insertions (22,809 calls), 2 bp insertions (8,480 calls), 1 bp deletions (13,427 calls) or 2 bp deletions (25,770 calls). The high number of only 1- or 2-bp indels are likely a consequence of the read mapping software and bioinformatics pipeline used (Sun et al., 2016).

Conclusions

In the present study, we re-examined an existing data set that was previously used to report high mutation counts from three transgenic plant series. We identified three major factors that inflated the estimates of molecular variation in the transgenic plants from these studies. These factors included residual heterogeneity, genotype

misidentification, and insufficient data filtering. The issue of genotype identity is obvious and intuitive, but requires caution, both for those handling and maintaining the materials (e.g., seeds, tissue, DNA) and those handling the computational analysis. Errors in genotype identity can be diagnosed using strictly molecular approaches, but situations where the identity of the material has been compromised or misinterpreted can be problematic (see commentaries (Bergelson et al., 2016; Lareau et al., 2018)). The issue of genetic heterogeneity within lines and seed stocks can create more subtle complications in analysis, as has been documented in the soybean line 'Williams 82' (Haun et al., 2011). When properly accounted for, heterogenetity does not disrupt accurate analysis and interpretation. However, when not properly accounted for, this issue may be problematic in assessing genomic, transcriptomic, and other types of variation. Within-line genetic heterogeneity can be an issue in many species, particularly those in which a reference genome is presumed to be perfectly representative of every individual in the seed stock. Lastly, data handling can be a major source of variation leading to inflated variant calls. Informatics pipelines generate large data sets, and users should be aware of quality control measures, and commonly used filtering parameters. Furthermore, experimental designs that provide replicated samples or comparisons among near-isogenic materials (e.g., the sibling lines discussed in this study) can be used to further differentiate the high-confidence and low-confidence variant calls.

While the present reanalysis focused specifically on comparisons between transgenic lines, all the factors addressed in this paper need also be considered when conducting any type of expression and/or genomic comparisons. This includes studies that focus on the effects of mutagenesis, on-target and off-target effects of genome

engineering technologies, assessments of standing/natural variation, or other comparisons of germplasm sources. This is particularly true for experiments on materials within the realm of biotechnology, as the findings may be used to inform regulatory agencies about the intended and unintended consequences of using these technologies. Evaluation for the presence of unintended changes at the DNA level continues to be a part of the safety evaluation for transgenic plants, and whole-genome sequencing has been proposed as a tool for this purpose (Pauwels et al., 2015). However, technical issues may make this problematic in crop species, which have complex, highly variable, and often heavily duplicated genomes. Furthermore, as demonstrated by the present study, the analysis and interpretation of whole-genome sequencing data may be inconsistent among research groups. While Lambirth et al. 2016 reported high rates of mutation in transgenic soybean lines, our reanalysis of their data concluded that there are relatively few sequence variants detected in these lines that might be attributed to the transformation process. It will be difficult to standardize a regulatory methodology that accounts for every complication that will arise across research groups and species (e.g., standing genetic heterogeneity within a parental seed stock) that may be incorrectly attributed to the genetic transformation process.

Methods

Variant and indel detection

RNA-seq from (Lambirth et al., 2015a) was downloaded from the National Center for Biotechnology Information Sequence Read Archive using project number PRJNA271477 and reanalyzed as described below. Sequencing adapters and low-quality bases were removed using Cutadapt with minimum read length set to 40 and

quality cutoff set to 20 (Martin, 2011). Using the GATK Best Practices workflow for RNA-seq (DePristo et al., 2011; Van der Auwera et al., 2013), reads were aligned to assembly version two of the reference genome (Wm82.a2) from www.soybase.org using the STAR aligner (Dobin et al., 2013). Read-group identifications were added and duplicate reads were marked using Picard tools. Reads were then split into exon segments, overhanging intronic segments were hard clipped, and mapping qualities were reassigned using the SplitNCigarRead tool from the GATK Genome Analysis Toolkit with -RMQF set to 255 -RMQT set to 60 and enabling the -U ALLOW_N_CIGAR_READS flag (McKenna et al., 2010). SNPs and indels were called using GATK HaplotypeCaller with the -dontUseSoftClippedBases flag and -stand_call_conf set to 20. The resulting VCF file was then split into separate files for SNPs (Additional File 3) and indels (Additional File 4) and then filtered using VariantFiltrations from the Genome Analysis Toolkit with parameters set to window of 35, cluster of 3, filter parameters of FS > 30, and QD < 2.0 for SNPs. Similar parameters were used for indel filtration, except FS filter was set to > 200 for all 36 samples. Variants that passed filtration were then used for downstream analysis.

Accession identification

Genotype calls from the filtered SNP list were extracted using a custom python script then loaded into R statistical software. The dataset was filtered for homozygous SNPs that are uniquely polymorphic to the reference compared to the other transgenic lines and 'Williams 82' controls. SNPs were removed from the analysis if there was more than 33% missing data for a given line and if there was no consensus genotype call between plants and replicates (Figure S1 in Additional File 1). The resulting SNPs were used to identify positions that overlapped within the SoySNP50k iSelect BeadChip (Song

et al., 2013) VCF file using the Wm82.a2 coordinates downloaded from www.Soybase.org. SNP calls for each of the 20,087 accessions in the 50k dataset were compared to the SNP calls for the 764 series to identify the accession with the highest level of SNP identity.

Analysis of data from previous studies

The Lambirth et al. 2016 supplementary data was downloaded from http://de.iplantcollaborative.org/dl/d/533570A3-1EFB-4864-B9A9-9D82F17E09A8/snpeffgenes.zip, and each of the 36 samples VCF files were parsed for depth, quality, and genotype information using a custom python script.

Software and figures

Parallelization of commands was run using GNU parallel. Data that was generated using R statistical software was plotted using the ggplot2 package (Wickham, 2011). The genome distribution of SNPs was created by using Phenogram (Wolfe et al., 2013).

Data availability

Software versions, options, thresholds, workflow details and custom scripts can be found at https://github.com/MeeshCompBio/The_Other_WPT_Study.

Acknowledgments

The authors are grateful to Drs. Wayne Parrott, Tom Clemente, and Candice Hirsch for helpful suggestions and comments on this manuscript and Peter Morrell and Fernanda Rodriguez for fruitful discussions. The authors are appreciative of the University of Minnesota's Office of Information Technology for providing data storage and the Minnesota Supercomputing Institute for other computational needs.

	764	ST77	ST111	Williams 82
“Unique” SNPs found in the whole RNA-Seq dataset	9738	143	3	0
“Unique” SNPs found in the RNA-Seq dataset that overlap with 50k SNP positions	525	11	0	0

Table 1. Number of SNPs identified as unique for each transgenic line based on reanalysis of the RNA-Seq dataset.

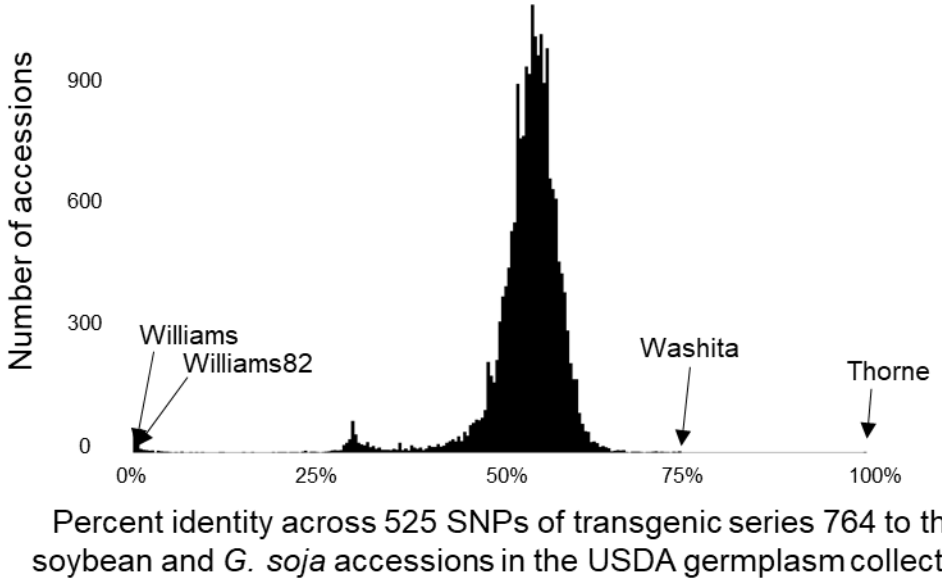


Figure 1. Reanalysis of series 764 reveals that its genetic background comes from genotype 'Thorne' rather than genotype 'Williams 82'. 525 SNPs were identified that met two criteria: (1) they were consistently polymorphic between series 764 plants and the 'Williams 82' reference genome in the RNA-seq dataset; (2) they were previously genotyped across the USDA germplasm (Song et al. 2015). A comparison of these SNPs to the all of the accessions in the USDA soybean accessions revealed 'Thorne' as a near-perfect match (99.2% identity), with a substantial gap to the next closest match (Washita at 74.2%). The reanalysis also confirmed that this panel of SNPs is completely polymorphic between the 764 series and 'Williams 82' (0% match).

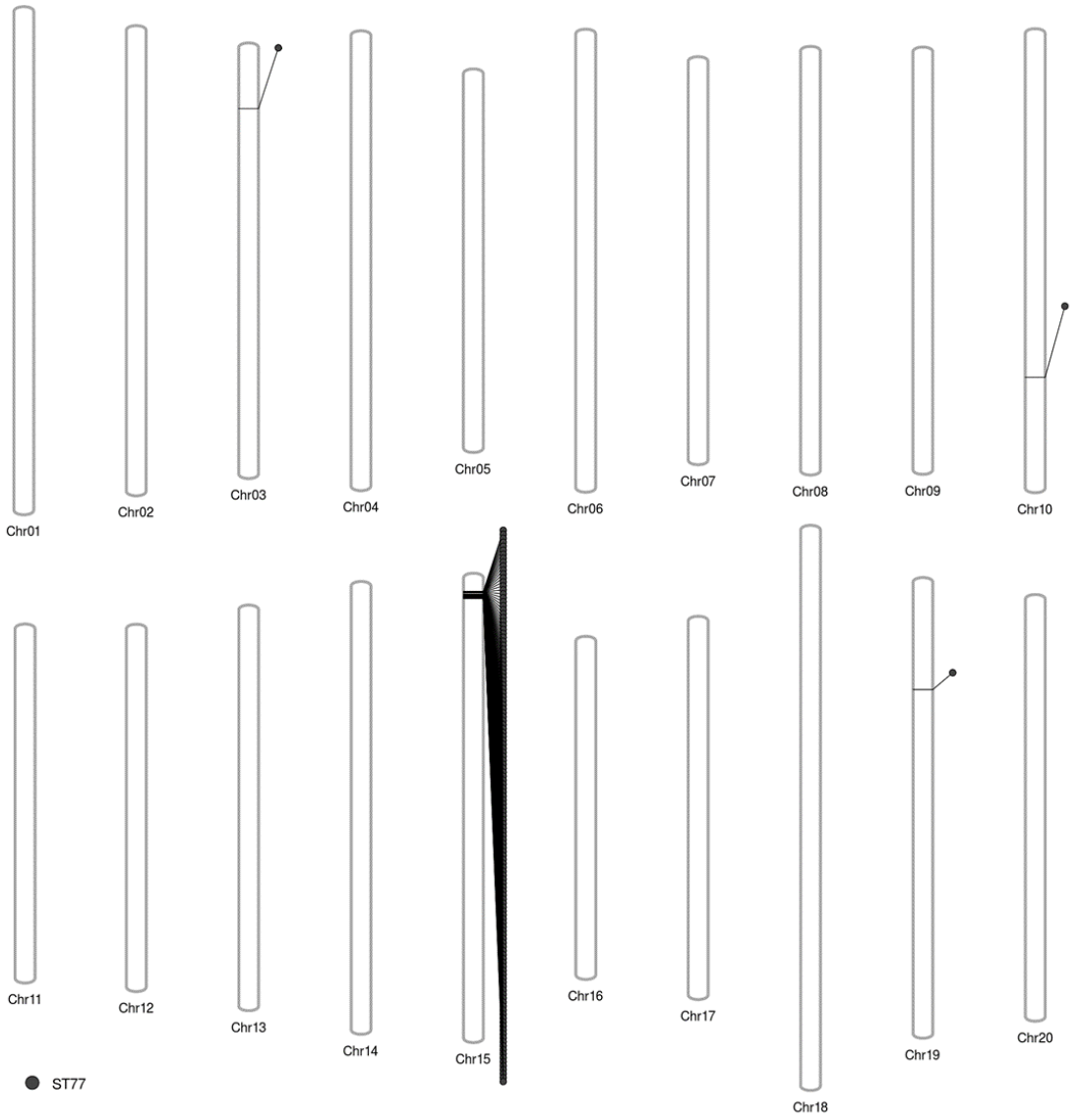


Figure 2. Distribution of unique SNPs across transgenic series ST77. The distribution of the 143 unique SNPs identified in ST77 is shown among the 20 chromosomes. Almost all of the ST77 SNPs (140 out of 143) cluster at a single locus on chromosome 15, which is a typical signature of genetic heterogeneity among the 'Williams 82' parental lines used in this study. The clustering of SNPs at specific, rather than random, positions is indicative of heterogenous standing variation that has previously documented in the 'Williams 82' cultivar (Haun et al., 2011).

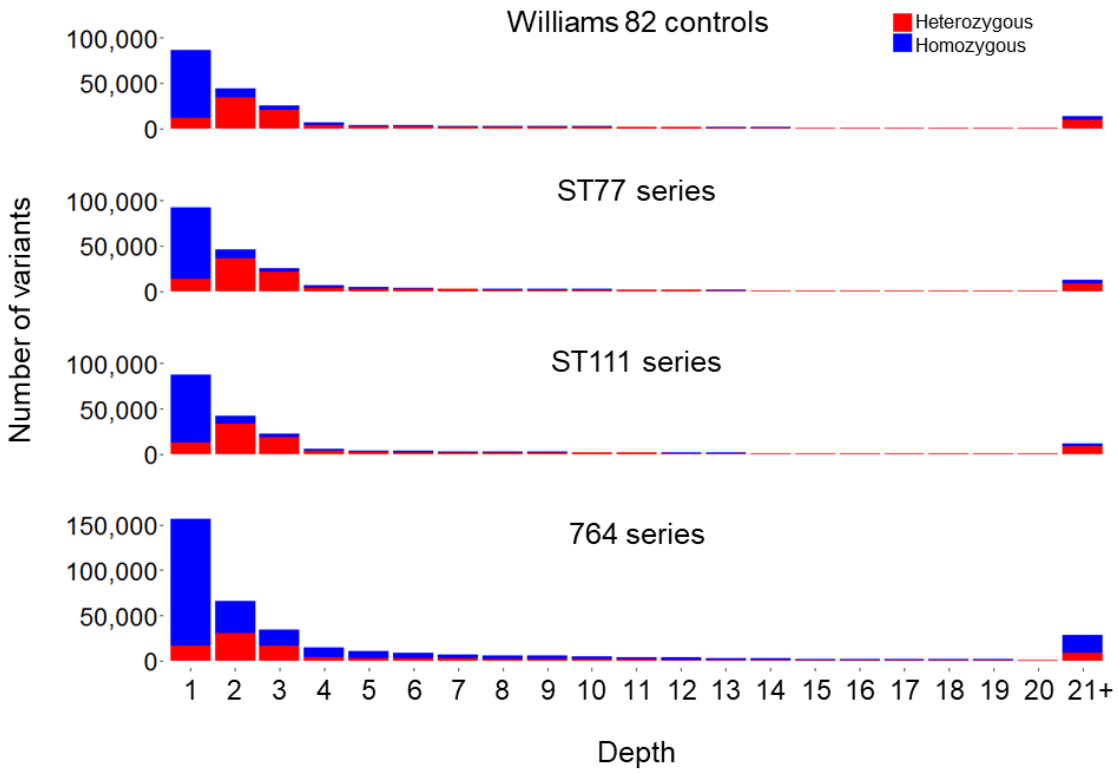


Figure 3. Depth of sequence coverage for all polymorphic variants (SNPs and indels) called in the Lambirth et al. (2016) study. The polymorphic calls shown here were made between each sample and the reference genome 'Williams 82', without consideration for the uniqueness of the call among series or reproducibility among different plants within the series. Homozygous calls are shown in blue and heterozygous calls are shown in red. Each bar sums the number of polymorphisms across the nine plants that were called at each read depth (e.g., we are showing the ~211,448 total variants called in series ST77 across the nine plants; ST77 averaged 23,494 variants per plant). Note the larger peak in the 21+ category for the 764 series; many of these (mostly homozygous) calls likely represent standing variants between lines 'Thorne' and 'Williams 82'. The 21+ peaks in the other three groups (ST77, ST111, and 'Williams 82' controls) may derive from various factors, most obviously the clusters of variants that are found within heterogeneous regions of different sub-lines of 'Williams 82'.

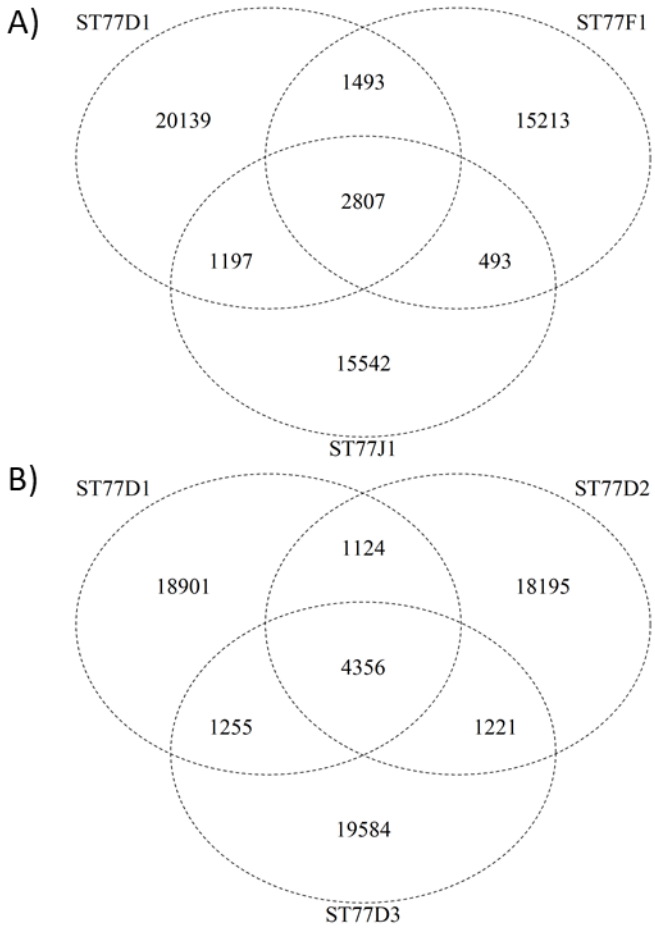


Figure 4. Number of overlapping polymorphisms in the Lambirth et al. (2016) study. A) Venn diagram showing of the number of sequence variants alternate to the reference genome that overlapped between three T8 individuals derived from transgenic event ST77. Heterozygous and homozygous alternate calls are not differentiated in this analysis. B) A similar Venn diagram of the number of polymorphism that overlapped between the T7:8 siblings in the ST77D family.

Chapter 3: Integration, abundance, and transmission of mutations and transgenes in a series of CRISPR/Cas9 soybean lines

Preface

As with many plant species, current genome editing strategies in soybean are initiated by stably transforming a gene that encodes an engineered nuclease into the genome. Expression of the transgene results in a double-stranded break and repair at the targeted locus, oftentimes resulting in mutation(s) at the intended site. As soybean is a self-pollinating species with 20 chromosome pairs, the transgene(s) in the T0 plant are generally expected to be unlinked to the targeted mutation(s), and the transgene(s)/mutation(s) should independently assort into the T1 generation, resulting in Mendelian combinations of transgene presence/absence and allelic states within the segregating family. This prediction, however, is not always consistent with observed results. In this study, we investigated inheritance patterns among three different CRISPR/Cas9 transgenes and their respective induced mutations in segregating soybean families. Next-generation resequencing of four T0 plants and four T1 progeny plants, followed by broader assessments of the segregating families, revealed both expected and unexpected patterns of inheritance among the different lineages. These unexpected patterns included: (1) A family with nearly complete transmission of the CRISPR/Cas9 transgene in the T1 generation, presumably caused by the integration of multiple unlinked transgenes; (2) A family in which the transgene integrated directly into two paralogous CRISPR target break sites, leading to a complete co-segregation of the transgenes and knockout mutations of the target genes; (3) A family in which mutations

were observed and transmitted, but without evidence of transgene integration nor transmission. These patterns and the mechanisms that drive them are discussed.

This work was a collaborative effort, with Jean-Michel Michno, Dr. Kamaldeep Virdi, Adrian O. Stec, Dr. Junqi Liu, Dr. Xiaobo Wang, Yer Xiong and Dr. Robert M. Stupar. JMM and RMS designed the experiment, JMM, JL, XW, and KV created CRISPR constructs, YX transformed constructs into soybean lines, AOS and KV extracted DNA and ran PCR assays, JMM performed all of the bioinformatics, JMM and RMS wrote the manuscript, JMM and KV created figures.

Introduction

Genome editing/engineering provides a toolkit for modifying DNA in a gene-specific manner, allowing researchers, geneticists, and breeders to move beyond the ordinary boundaries of germplasm and genetic variation. In crop plant species, the majority of trait-driven editing applications have focused on creating targeted gene knockouts, with many such efforts using CRISPR/Cas9 editing reagentsCS (Belhaj et al., 2013; Feng et al., 2013; Mao et al., 2013; Shan et al., 2013b; Schiml et al., 2014; Cermak et al., 2017). Oftentimes, this process involves delivering a transgene to the plant genome that encodes the CRISPR gRNA(s) and Cas9 protein. Expression of these reagents in the T0 generation can generate mutation(s), which can be transmitted to subsequent generations. Moreover, the CRISPR/Cas9 transgene will likely not be linked to the mutation(s). Therefore the breeder/geneticist can specifically select for segregating individuals that carry the desired mutated allele and no longer harbor the transgene.

Soybean genes have been successfully modified using CRISPR/Cas approaches in both somatic and germline transmissible cells and for a variety of agronomic traits (Cai et al., 2015; Jacobs et al., 2015; Li et al., 2015; Michno et al., 2015; Sun et al., 2015; Tang et al., 2016; Cai et al., 2018; Curtin et al., 2018). One recent study (Curtin et al., 2018) carefully tracked the transmission of mutations and transgenes from T0 soybean plants to the next generation. In this study, *Agrobacterium* was used to transform CRISPR/Cas9 into whole soybean plants to knockout genes involved in small RNA pathways. Curtin et al. (2018) targeted three genes, GmDrb2a, GmDrb2b and GmDcl3a and generated mutations at each target site in the T0 generation.

The GmDrb2 CRISPR construct used two guide-RNAs that each recognized both GmDrb2ba and GmDrb2b loci. The resulting transformation yielded two T0 plants, WPT590-1 and WPT590-4 that were derived from the same cluster of cells. From these two events, Curtin et al. 2018 detected 1, 4, 7, and 8 bp deletions in both transgenic events at the GmDrb2a locus. Screening of the GmDrb2b locus reveled a 4 and 7 bp deletion shared between transgenic events and a 6 bp deletion unique to WPT 590-1. Using next-generation sequencing, they identified three separate transgenic insertion events on chromosomes 4, 13 and 15 in the same locations for both WPT590-1 and WPT 590-4. After selfing the T0 progeny to the T1 generation, PCR screening for mutations revealed that only the 1 and 7 bp deletions were heritable while the 4 and 8 bp deletions did not transmit. Similarly, when screening T1 progeny at the GmDrb2b locus, only the 4 and 7 bp deletions were heritable while the 6 bp deletion was not. Further analysis of each of the three transgenic insertion events in the T1 generation revealed that each event was heritable.

Meanwhile, a different CRISPR/Cas9 construct was designed to target GmDcl3a. Analysis of the GmDcl3a CRISPR mutations in two separate events WPT527-1 and WPT 527-2, resulted in a 1 and 13 bp deletion for WPT 527-1 and a 4 bp deletion and 1 bp insertion for WPT 527-2. PCR screening and next-generation sequencing of the WPT527-1 GmDcl3a event gave evidence of a transgenic insertion event on chromosome 9. Similar analysis for WPT527-2 did not result in the identification of any transgenic insertion events using sequencing. The authors then analyzed 60 T1 plants from each event and failed to identify any heritable mutations or transgene integration events.

In this study, we expanded upon Curtin et al. (2017) and sequenced four T0 parents and four offspring of transgenic CRISPR/Cas9 lines to study the effects of CRISPR/Cas9 at gRNA target sites, as well as variation induced due to transgenic insertion events into the genome. The transformed lines studied in this experiment demonstrate the potential outcomes of *Agrobacterium*-mediated transgenesis using CRISPR/Cas9.

Results

Identification of CRISPR mutations at target sites in T0 plants

Three separate whole-plant transformation (WPT) series named WPT536, WPT553, and WPT608 were generated using the expression vectors diagramed in Figure 1. Each vector used a constitutive promoter (Gmubi or Califlower mosaic double 35S (Benfey and Chua, 1990; Hernandez-Garcia et al., 2009)), a Cas9 endonucleases (Soybean codon optimized (Michno et al., 2015)) or Arabidopsis codon optimize (Li et al., 2013a)), single or double gRNA cassette (Curtin et al., 2018) driven either by the Arabidopsis U6 or 7sL promoter, and either a Glufosiante (BAR) or Hygromycin plant selectable marker (Figure 1, Table S1 in Appendix 2). Guide-RNA (gRNA) cassettes were constructed and inserted into each WPT destination vector. WPT536 and WPT553 each targeted a single locus on one gene model, Glyma16g12160, and Glyma.18g041100, respectively (Table 1). WPT608 included two gRNAs targeting gene model Glyma.16G209100, which were also a perfect match to its paralog gene model Glyma.09G159900. Each destination vector was transformed into the background BertMN-01, and DNA was extracted from putatively transformed T0 plants.

PCR-based gel assays (see Methods for details) were used to screen for mutations at the intended sites for each T0 plant. Four T0 plants were identified with putative mutations, one each from the WPT536 (individual WPT536-2) and WPT553 (individual WPT553-6) series, and two from the WPT608 series (individuals WPT608-1 and WPT608-3). Sequencing of PCR amplicons at each of the target sites for these four T0 plants revealed mutations (details are provided in the sections below). These four plants and some of their progeny were tracked for the inheritance of the targeted mutations and transgene integration loci.

WPT536-2: Expected transmission and segregation patterns from single transgene and mutation events

WPT536-2 was a T0 plant transformed with a Gmubi-driven *Glycine max* codon-optimized Cas9 and a single gRNA targeting Glyma16g12160 (herein known as GmRin4b). PCR confirmed the presence of the Cas9 and plant-selectable marker (Figure 2A), indicating successful transformation of the construct. Sequencing of a PCR amplicon from the gRNA target site revealed a 2 bp deletion. WGS of the T0 plant confirmed the previously identified 2 bp deletion along with evidence of a 1 bp insertion at the target site (Figure 2B, Figure S1 in Appendix 2). Furthermore, WGS revealed a single CRISPR/Cas9 transgene integration site localized to an interval on chromosome 11 (Figure 2C, Figure S2 in Appendix 2). The interval had a 35 bp hemizygous deletion and within it, a 4 bp and 17 bp addition flanking the transgenic insertion (Table 2). The reads spanning the genome into the transgene indicate that a complete cassette between the RB to LB was inserted within the deleted region. Given the presence of both a transgene and mutation, the generation of this plant was renamed T0/M0.

Screening of GmRin4b mutations in the segregating T1/M1 and T2/M2 generations revealed germline transmission of the transgene (Figure 2A). WGS was performed on two progeny plants that were identified from the PCR assay as no longer carrying the transgene. To further validate that there was no trace of transgenic DNA, reads from WGS were mapped directly to the transgene for each sequenced plant (Figure S3 in Appendix 2). Only the T0 parent had consistent coverage across the transgene, while the progeny plants lacked any reads mapping to the transgene, except for the Gmubi promoter where similar sequences are located on chromosomes 10 and 20 of the genome. Furthermore, the WGS revealed that WPT536-2-13-16 retained the 2 bp mutation at the CRISPR target site while WPT536-2-13-15 segregated back to homozygosity for the wild-type allele. Given these findings, it was determined that plant WPT536-2-13-16 is a simple M2 generation plant (contains a mutation, but no transgene), while plant WPT536-2-13-15 is neither a transgenic or mutant individual. This segregation represents expected Mendelian patterns, wherein the respective transgenic and mutated loci could be selected for or against in subsequent generations.

WPT608-1: T0 transgenes and mutations were not transmitted to progeny

Gene model Glyma.16G209100 and Glyma.09G159900 was targeted by CRISPR Cas9 using a nearly-identical construct to that used by Curtin et al. (2018), with the only modification being the gRNA target site. PCR screening revealed that two lines, WPT608-1, and WPT608-3, had evidence for mutations at identical recognition sites on Chromosomes 9 and 16 from a single gRNA, as well as evidence of transgene integrations into the genome. WGS of 608-1 confirmed the presence of a 1 bp insertion and two different 4 bp deletions as seen by PCR (Figure 3A). Furthermore, an additional

target site on the paralogous gene model Glyma.09G159900, which has an identical gRNA recognition site, also showed some evidence for mutations (Figure 3A).

WGS identified a single transgene integration site on chromosome 17 for WPT608-1 (Figure 3B). The T-DNA segment induced a 1 bp deletion at the transgene integration site with a 9 bp insertion flanking the transgenic segment (Table 2). Reads that were spanning the genomic-transgene junction revealed that a portion of the right border inserted itself into that location. The transgenic sequence at the left junction was undetectable due to the lack of any chimeric reads aligning to that segment of the genome.

PCR assays could not detect any presence of mutations or the transgene in the T1 generation among 22 tested plants, suggesting that neither the mutations nor the transgenic insertion event were germline transmissible (Figure S6 A and B in Appendix 2). Therefore, the WPT608-1 event appears to be an instance where the reagents may have been delivered and expressed transiently, or the T0 plant was chimeric, and the transgenic sector did not produce seeds. In either case, mutations may have been produced in some somatic cells of the T0 plant but did not reach the germline.

WPT608-3: Mutations and transgene integrations at the CRISPR target sites

WGS of 608-3 revealed four separate transgene insertion events on chromosomes 6, 9, 16, and 18 (Figure 3C in Appendix 2). The event on chromosome 6 induced an 8 bp deletion in the host genome while inserting 3 and 20 bp addition on either side of the transgene integration site (Table 2). Analysis of the reads spanning the genomic/transgene junctions suggests that there was a partial insert of half of the transgene from the RB to halfway through the cassette. The transgenic insertion event

on chromosome 18 deleted 3 bp of the host genome and created a more complex transgenic insertion event. The transgenic sequence detected on the left junction was in the antisense orientation while the sequence on the right junction was in the sense orientation, suggesting that there were multiple insertions/rearrangements of the transgene at that location (Figure 3C).

The transgene integration site on chromosome 16 was observed within the CRISPR gRNA target site on gene model Glyma.16G209100. The sequenced regions flanking the transgene integration site indicated that 1 bp of the host genome was deleted while inserting a full transgene cassette. Furthermore, the transgene integration site on chromosome 9 was also observed within a CRISPR gRNA target site on the paralogous gene model Glyma.09G159900, except that it created a 10 bp deletion in the host genome. There was also an 11 bp insertion flanking the sequence of one end of the chromosome 9 transgene integration site (Table 2). Reads spanning the junctions of both the chromosome 9 and chromosome 16 events suggest that a full transgene cassette was inserted into both locations.

Due to the transgene integrating itself into the gRNA target site for Glyma.16G209100, three of the six WPT608-3 T1 progeny were homozygous for the transgene integration event and were unable to be amplified for screening (Figure 7A in Appendix 2). PCR assays for two of T1/M1 progeny from WPT603-1 confirmed germline transmission of the 1 bp insertion on Glyma.16G209100 (Figure S7 B and C in Appendix 2). To detect transmission of the transgene in the T1 generation, primers were developed using genomic coordinates of each transgenic insertion event and confirmed the heritability of each transgene across six T1/M1 progeny (Figure S8 in Appendix 2).

In summary, WPT608-3 represents a unique T0 plant in which two of the four transgene integration sites were located at the gRNA target site. Presumably, this was caused by CRISPR/Cas9 induction of double-stranded breaks at the paralogous target sites that were repaired by transgene integration during the transformation process.

WPT553-6: Unresolved transgene inheritance in line with germline mutations

The CRISPR/Cas9 construct targeting Glyma.18g041100 (herein known as GS1) was developed as a result of a previous study and shown to be effective at generating mutations in soybean somatic hairy root tissues (Michno et al., 2015). We used the same construct in whole-plant transformation to generate the WPT553 series of plants for the present study. PCR screening and WGS of the WPT553-6 T0 plant revealed two different 7 bp deletions at the target site (Figure 4A, Figure S9 in Appendix 2). Sequencing of the progeny plants 553-6-8 and 553-6-11 identified a 2 bp and a 6 bp mutation in the respecitve plants. Neither of these mutated alleles were identified in the T0 parental plant (Figure S9 in Appendix 2). Furthermore, the plant-selectable marker and the Cas9 were not detected by PCR in the 553-6-8 and 553-6-11 plants, nor were these transgene components detected in any of the 31 putative T1/M1 offspring (Figure 5B). Aside from the 553-6-8 and 553-6-11 individuals, none of these plants showed evidence for mutations at the target site.

To help detect chimeric transgenic or mutation events, leaf tissue was pooled from different parts of plant WPT553-6, and DNA was prepared for WGS. Similar pooling strategies were also applied within each of the 553-6-8 and 553-6-11 offsping plants. WGS analyses were not able to identify any transgene integration sites in the WPT553-6 T0 plant nor the 553-6-8 and 553-6-11 offspring. When mapping the DNA of each plant

directly to the transgene (Figure S10 in Appendix 2), only the WPT553-6 T0 plant had reads that consistently mapped to the transgene. However, the average read coverage for the transgene was far below the WPT plants described in previous sections that exhibited heritable transgenic insertion events. WGS mapping of reads to the transgene sequence for 553-6-8 and 553-6-11 respectively yielded only 7 and 1 reads that mapped (Figure S10 in Appendix 2). This extremely low mapping coverage may be better explained by trace levels of sample contamination rather than the presence of a stably integrated transgene. Therefore, we speculate that the initial mutagenesis observed in the WPT553-6 T0 plant may have been derived from a non-integrated CRISPR/Cas9 transgene, which may explain the transmission of mutated alleles with minimal evidence for transmission of any transgene components.

Discussion

The results described above highlight the range of outcomes one might expect from strategies that rely on stable transformation of a DNA editing construct. Such experiments can be complicated, as they typically require a minimum of two loci of interest, the transgene integration site(s) and the targeted region(s). This quickly becomes more complex when there are multiple unlinked transgene integrations and when there are multiple gene editing targets.

Resequencing of four T0 plants and selected progeny provided a high-resolution view of transgene integration structures and gene editing events. The four T0 plants each exhibited a different outcome, though each outcome parallels similar findings in the recent crop genome editing literature. Plant WPT536-2 exhibited the most straight-forward scenario, in which a single transgene integration produces frameshift mutations

at a single target site. The transgene and mutations transmitted and segregated in the progeny, as is generally the desired outcome for the majority of such experiments and has often been reported in previous studies (Feng et al., 2014; Wang et al., 2014; Xing et al., 2014; Butler et al., 2015; Chandrasekaran et al., 2016; Osakabe et al., 2016; Pyott et al., 2016; Yan et al., 2016; Zhang et al., 2016b; Yin et al., 2017; Cai et al., 2018; Curtin et al., 2018).

Plant WPT608-1 exhibited evidence for a single transgene integration and targeted mutations at two paralogous loci. However, neither the transgenes nor mutations were recovered in the progeny. This type of negative result may be commonplace in genome editing projects, but it is an undesirable outcome for most projects and is likely to be unreported in scientific articles (Curtin et al., 2018). There are different mechanisms that may explain this result, including the possibility that WPT608-1 was a chimeric plant in which the transgene and mutations were part of a sector that did not produce seeds. It is noteworthy that the DNA used to resequence plant WPT608-1 was pooled from five different leaflets growing on different branches of the plant. Perhaps only one or two branches harbored the transgene and mutations, and these failed to produce seeds. Alternatively, the transgene integration and/or mutations may have disrupted a critical process for gametophyte or early sporophyte survival, thus purifying the progeny into only wild-type segregants. While these hypotheses remain untested, there are many other such hypotheses that could be suggested to explain the observed result.

Plant WPT608-3 exhibited an unexpected phenomenon in which two paralogous CRISPR target sites were each found to harbor CRISPR/Cas9 transgenes. The process to create such loci is somewhat analogous to a previously described non-homologous

end-joining strategy used to insert a specific T-DNA segment into a specific genomic locus (Bortesi and Fischer, 2015). In this strategy, the editing reagent (e.g., the CRISPR/Cas9) is designed to simultaneously cut both the intended T-DNA segment from the transgene and the genomic target where the T-DNA is to be inserted. In effect, the released T-DNA segment acts as a donor molecule that can be integrated into the genomic target site during the double-stranded break repair. In the case of plant WPT608-3, it appears that when the full transgene was delivered to the cell it generated double-stranded breaks at the intended paralogous loci, and then copies of the transgene were used to repair the targeted double-stranded breaks. This phenomenon has been previously reported in the literature (Chilton, 2003; Tzfira, 2003; Cai et al., 2009; D'Halluin et al., 2013). However, it is not common and we are unaware of any examples in which two unlinked (in this case, paralogous) target sites acted as transgene integration loci in a single cell. Importantly, all four transgenic loci in the T0 plant were shown to segregate in subsequent generations. Furthermore, a simple frameshift allele for gene model Glyma.16G209100 was also shown to segregate in these generations. Therefore, a researcher could select for progeny that specifically carry the frameshift allele and no longer harbor the transgenes, if such an outcome is desired.

Plant WPT553-6 exhibited a unique outcome in which the T0 plant exhibited the presence of mutations at the targeted locus (Glyma.18g041100). However, resequencing data could not confirm integration of the CRISPR/Cas9 transgene. Analysis of progeny indicated that a small number of plants (two out of 31) carried mutations, while none of the plants harbored the transgene. On the surface, this appears to be a highly favorable outcome, as transmissible mutations were recovered in an

apparently non-transgenic background. However, this may be a difficult result to reproduce, as it would require transient expression of the transgene without integrating into the host genome, thereby generating mutations in a non-transgenic background. Zhang et al. (2016) reported a purposeful identification of such plants in wheat, wherein the authors specifically screened plants bombarded with CRISPR/Cas9 constructs for individuals carrying mutations and no transgenes (Zhang et al., 2016a). This process was able to identify plants of this type but required extensive screening of large populations to identify these rare events. In the case of WPT553-6, it is also possible that the transgene did insert stably into the genome but was located in a region of the genome difficult to map and/or a structurally rearranged T-DNA was inserted such that it was not detected by PCR or resequencing. Alternatively, as discussed for WPT608-1 above, it is possible that the WPT553-6 transgene integration may have disrupted a critical process for gametophyte or early sporophyte survival and was thus not able to be recovered in the progeny. This would not entirely explain the inability to identify the transgene integration site in the T0 plant but would provide an explanation for the failure to transmit the transgene to progeny.

Despite the complications of working with these complex plants, there is a high probability to recover a desired product using the CRISPR/Cas9 technology in soybean. In this study, we used two different Cas9 endonucleases that yielded similar mutation profiles between events. While the size of mutations seen were all under 7 bp in size, all but one mutation induced at a gRNA target site created a frame-shift mutation, most likely knocking out the function of the target gene. In the case of multiple transgene insertions, it may be difficult to completely segregate away from all copies. However, additional backcrosses or outcrosses can be used to remove these loci, as

demonstrated by Curtin et al. (2018). This is a relatively minor inconvenience, given the capacity to generate vast and novel allelic diversity for so many loci.

Material and Methods

Generation of whole plant transformant expression vectors

Plant expression vectors were created using three different binary vectors; PMDC123, PMDC32, and pNB96 (Curtis, 2003; Curtin et al., 2011). The expression vector used to create WPT536 was a modified version of the Cas9 MDC123 found on addgene.org (https://www.addgene.org/59184/). The vector was modified by replacing the 2x35S Cas9 promoter with a *Glycine max* ubiquitin promotor (Hernandez-Garcia et al., 2009) and adding the Rin4b (Glyma16g12160) gRNA recognition sites. The WPT553 expression vector, MDC32/GUS/GmCas9, was originally developed and used in (Michno et al., 2015). WPT 608-1 and 608-3 used the same pSC218GG construct used in (Curtin et al., 2018), except with different gRNA recognitions sites for the Glyma.16G209100 (and Glyma.09G159900) target sites.

Identification of CRISPR/Cas9 target sites

CRISPR target sites were identified using a soybean CRISPR design website (http://stuparcrispr.cfans.umn.edu/CRISPR/) (Michno et al., 2015). Glyma numbers from version two of the soybean assembly were input into the webtool, and target-sites were screened for unique restriction sites designed to cut 3-5 bp upstream of the proto-spacer adjacent motif.

Delivery of expression vectors to soybean whole-plants

Constructs were delivered to the Bert-MN-01 background using 18r12, a disarmed k599 *Agrobacterium rhizogenes* strain (Veena and Taylor, 2007). Methods for delivery and growth of whole-plant transformants were performed as previously described (Curtin et al., 2011).

DNA extraction and identification of transgene insertion sites using next-generation sequencing

Leaf tissue was harvested from five different soybean branches for each whole-plant transformant and extracted with a Qiagen DNeasy plant kit (item 69106). DNA samples were sent to the University of Minnesota Genomics Center for sequencing using an Illumina HiSeq2500 with v4 chemistry to generate 125bp parried-end reads. Reads were checked for initial quality using Fastqc version 0.11.5 and Illumina Truseq adapters were trimmed using cutadapt version 1.8.1 with a minimum read length set to 40bp and quality cutoff set to a phread score of 20 (Andrews, 2010; Martin, 2011). To map reads to the soybean reference genome (Wm82.a2.v1), we used bwa version 0.7.12 with band width set to 100, mark shorter splits as secondary, and penalty for mismatch set to 6 (Li and Durbin, 2009). Samtools version 1.6 was used to convert SAM file format to BAM format, sort, and index files (Li et al., 2009). Identification of transgene insertion sites was performed in a manner similar to Srivastava et al. 2014 (Srivastava et al., 2014). Fasta files were created using the transgene cassette with 100 bp flanking backbone sequence to serve as our reference genome. Sequenced reads were then mapped to transgene reference using the same programs and parameters used to map reads to the reference genome. Orphaned reads were extracted using a modified version of extract_unmapped_mates.pl (Srivastava et al., 2014), to accept bam files as

input. Orphaned reads we then mapped to the Wm82.a2.v1 reference using bowtie2 version 2.2.4 using -- local -- very-sensitive-local (Langmead et al., 2009). SAM files were then converted to BAM file format, sorted and indexed in the same manner mentioned above. Orphaned reads that mapped to the reference were further investigated upon using IGV version 2.3.90 (Robinson et al., 2011). Orphaned read mapping was then compared to read mapping to the soybean reference and the parental line (Bert-MN-01) as a control. Deletions were investigated using IGV at each CRISPR site throughout the genome. To automate this process, a custom bash script was created called TransGeneMap (https://github.com/MeeshCompBio/Soybean_Scripts) that allows users to input only the forward and reverse reads, indexed reference genome, and transgene sequence to automate the analysis.

Plant number	Transgene integration	Target gene(s)	Target 1	Target 2
536-2	Chr11	Glyma16g12160	2 bp Δ , 1bp +	NA
536-2-13-15	NA	Glyma16g12160	wt	NA
536-2-13-16	NA	Glyma16g12160	2 bp Δ	NA
553-6	NA	Glyma.18g041100	7 bp Δ, 7 bp Δ	NA
553-6-8	NA	Glyma.18g041100	2 bp Δ	NA
553-6-11	NA	Glyma.18g041100	6 bp Δ	NA
608-1	Chr17	Glyma.16G209100, Glyma.09G159900	4 bp Δ , 4bp Δ, 1 bp +	4 bp Δ
608-3	Chr06, 09, 16, 18	Glyma.16G209100, Glyma.09G159900	3bp Δ, 1 bp +, TGI*	TGI*

*TGI: Transgene integration

Table 1. Mutation profiles induced by CRISPR/Cas9 and number of transgene insertions for each transgenic series

Plant	Integration	Type	Genic	Flanking insert sizes
536-2	Chr11:2,511,324-2,511,349	35bp Δ	+	4 bp, 17 bp
608-1	Chr17:37,687,748	1bp Δ	-	missing, 9 bp
608-3	Chr06:3,498,485-3,498,492	8bp	+	3 bp, 20 bp
608-3	Chr09:38,390,575-38,390,586	10bp Δ *	+	0bp, 11 bp
608-3	Chr16:36,848,517	1bp Δ *	+	NA
608-3	Chr18:55,616,603-55,616,607	3bp Δ	-	NA

*TGI: Transgene integration

Table 2. Types of variation induced for each transgene insertion event

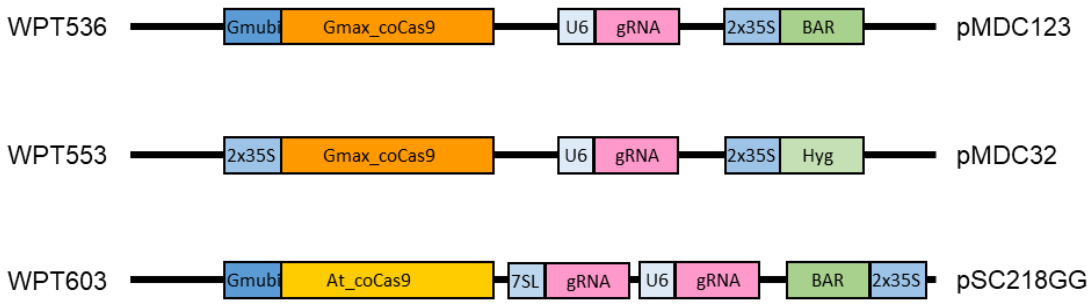
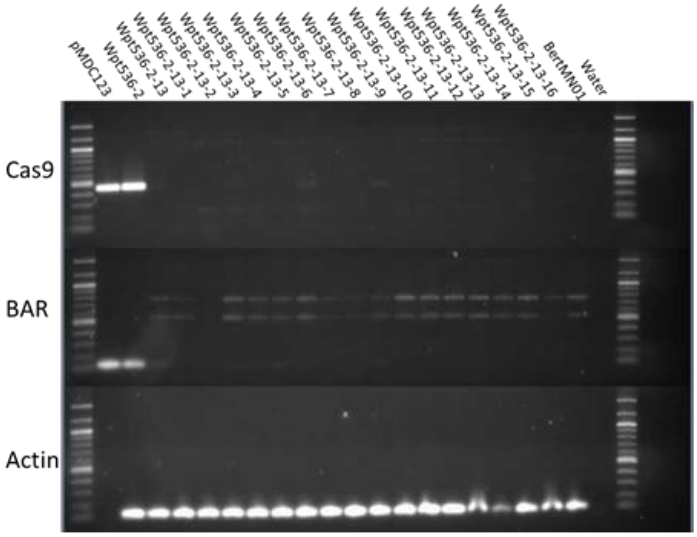


Figure 1. Transformation vectors used in whole-plant transformations. Plant expression cassettes used to integrate transgenic DNA through *agrobacterium*-based whole-plant transformation methods. Boxes in shades of blue represent promoters, boxes in shades of orange represent Cas9 endonucleases, boxes in shades of green represent plant-selectable markers, and boxes in shades of pink represent guide RNA's.

A)



B)

BertMN01	GGTTCTGCTGATAACCGTAGAAGACATTCAACGCAAAGTACTGGG	100 %
WPT536-2	GGTTCTGCTGATAACCGTAGAAGACATTCAACGCAAAGTACTGGG	79 %
WPT536-2	GGTTCTGCTGATAACCGTAGAAGACATTCAACGCAAAGTGACTGGG	7 %
WPT536-2	GGTTCTGCTGATAACCGTAGAAGACATTCAACGCAAA--ACTGGG	14 %
WPT536-2-13-15	GGTTCTGCTGATAACCGTAGAAGACATTCAACGCAAAGTACTGGG	100 %
WPT536-2-13-16	GGTTCTGCTGATAACCGTAGAAGACATTCAACGCAAA--ACTGGG	100 %

C)

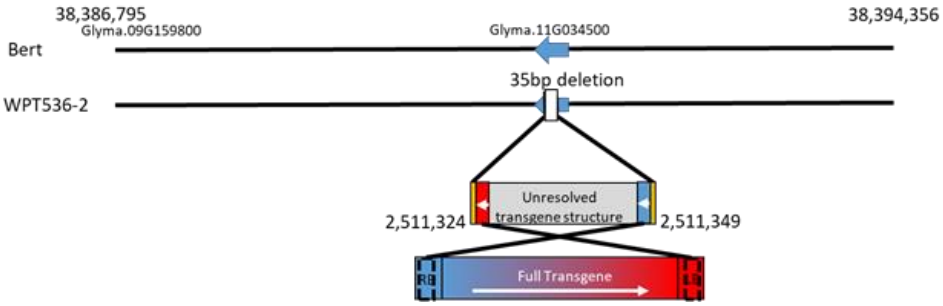


Figure 2. Screening of markers and mutations in the Rin4b transgenic series. A) PCR assay screening for presence absence or the Cas9 endonuclease, BAR plant-selectable marker, and actin control for the transformation vector, Rin4b T0 parent and the Rin4b T2 offspring. B) Sequence of transgenic plants and BertMN-01 control at gRNA target site. Dashes represent a deletion within a sequence, green text represents and insertion. Percentages on the right of sequences signify the proportion of reads representing the allele in the transgenic plant. C) Diagram depicting WGS detection of the transgene insertion event and the variation induced at the insertion site.

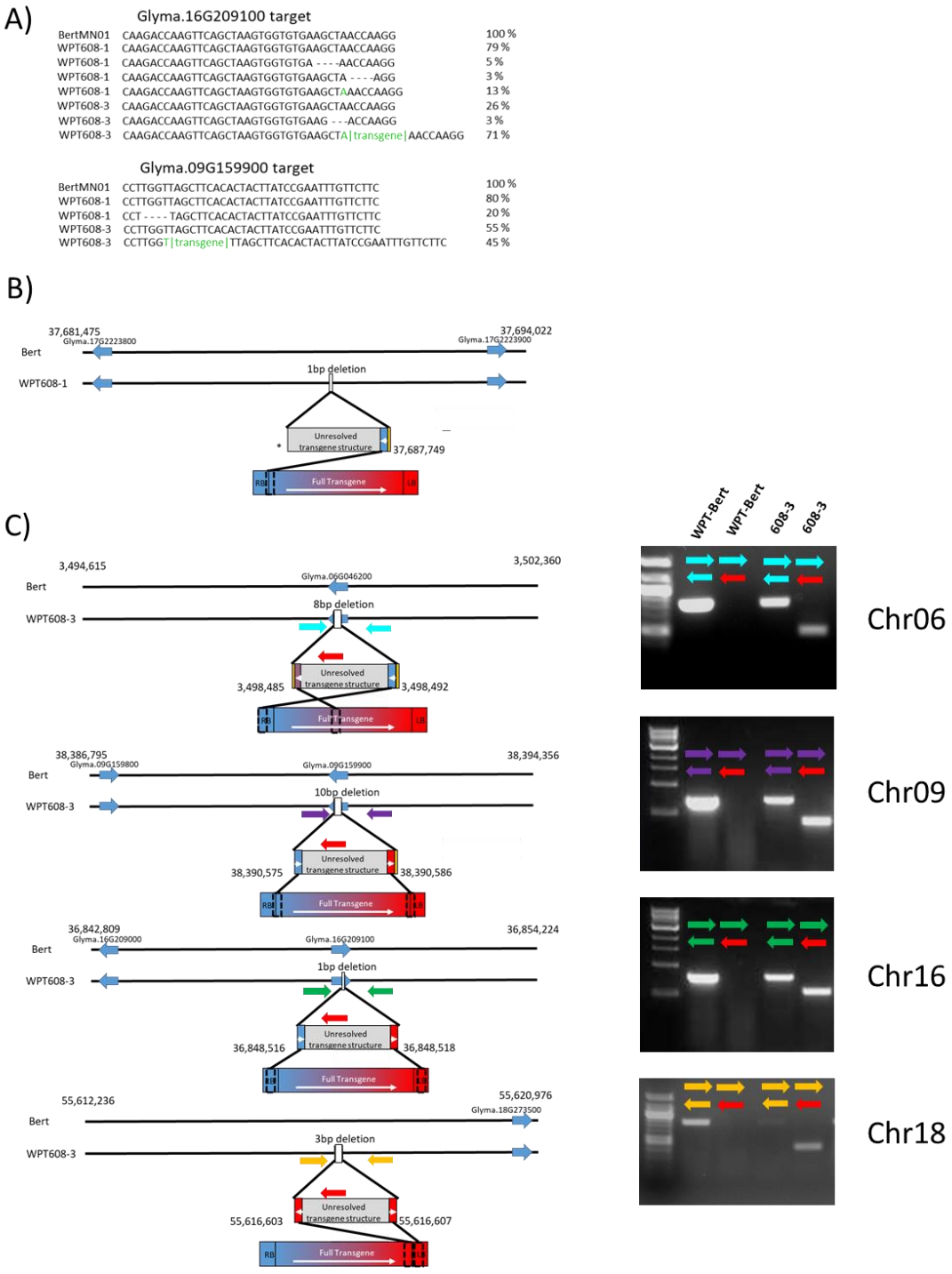


Figure 3. Screening of mutations and transgene insertions in the Glyma.16G209100 transgenic series. A) Sequence of transgenic plants and BertMN-01 control at Glyma.16G209100 and Glyma.09G159900 gRNA target sites. Dashes represent a deletion within a sequence, green text represents and insertion. Percentages represent

the proportion of reads representative of the sequence. The Glyma.16G209100 sequence targets the sense strand while Glyma.09G159900 target the antisense strand. B) Diagram depicting WGS detection of the transgene insertion event and the variation induced at the insertion site for WPT 608-1. C) Diagram depicting WGS detection of the transgene insertion events as well as PCR detection of primers flanking the transgene insertion event/junction for WPT608-3. Colored bars in each PCR, represent the primers used and their respective locations.

A)

BertMN01	ATCAAAACCGCCATTGAGAAGTTGGGGAAGAGACACAAGG	100 %
WPT553-6	ATCAAAACCGCCATTGAGAAGTTGGGGAAGAGACACAAGG	92 %
WPT553-6	ATCAAAACCGCCATTGAGAAGTTGGGGAA - - - - - - - AAGG	4 %
WPT553-6	ATCAAAACCGCCATTGAGAAGTTGGGG - - - - - - - ACAAGG	4 %
WPT553-6-8	ATCAAAACCGCCATTGAGAAGTTGGGGAAGAGACACAAGG	50 %
WPT553-6-8	ATCAAAACCGCCATTGAGAAGTTGGGGA - - AGACACAAGG	50 %
WPT553-6-11	ATCAAAACCGCCATTGAGAAGTTGGGGAAGAGACACAAGG	50 %
WPT553-6-11	ATCAAAACCGCCATTGAGAAGTTGGGG - - - - - - CACAAGG	50 %

B)

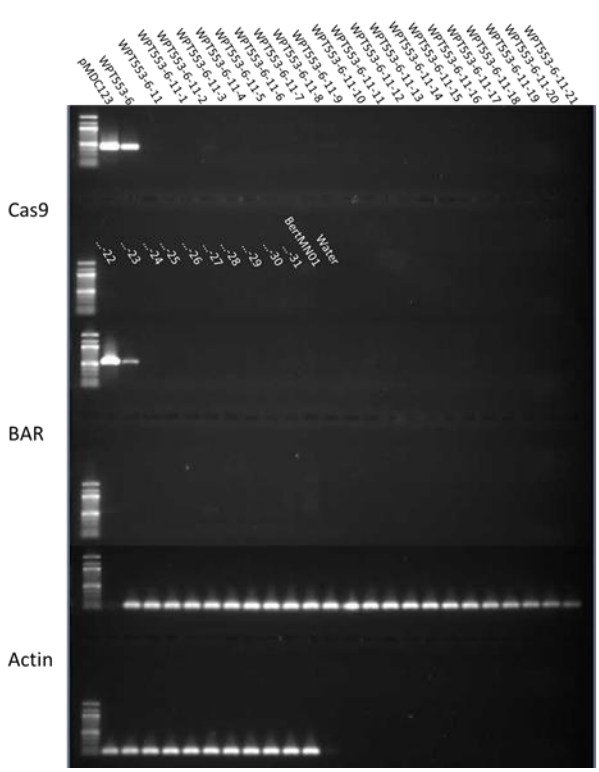


Figure 4. Screening of markers and mutations in the GS1 transgenic series. A) Sequence of transgenic plants and BertMN-01 control at gRNA target site. Dashes represent a deletion within a sequence. Percentages represent the proportion of reads representative of the sequence. B) PCR assay screening for presence absence or the Cas9 endonuclease, BAR plant-selectable marker, and actin control for the transformation vector, GS1 T0 parent and the GS1 T2 offspring

Chapter 4: Identification of candidate genes underlying nodulation-specific phenotypes in *Medicago truncatula* through integration of genome-wide association studies and co-expression networks

Preface

Genome-wide association studies (GWAS) have proven to be a valuable approach for identifying genetic intervals associated with phenotypic variation in *Medicago truncatula*. These intervals can vary in size, depending on the historical local recombination near each significant interval. Typically, significant intervals span numerous gene models, limiting the ability to resolve high-confidence candidate genes underlying the trait of interest. Additional genomic data, including gene co-expression networks, can be combined with the genetic mapping information to successfully identify candidate genes. Co-expression network analysis provides information about the functional relationship of each gene through its similarity of expression patterns to other well-defined clusters of genes. In this study, we integrated data from GWAS and co-expression networks to pinpoint candidate genes that may be associated with nodule-related phenotypes in *Medicago truncatula*. We further investigated a subset of these genes and confirmed that several had existing evidence linking them nodulation, including MEDTR2G101090 (PEN3-like), a previously validated gene associated with nodule number.

This work was a collaborative effort, with Jean-Michel Michno, Dr. Liana Burghardt, Dr. Junqi Liu, Joseph R. Jeffers, Dr. Peter Tiffin, Dr. Robert M. Stupar, and Dr. Chad L. Myers. JMM and CLM designed the experiment, LB, JL and PT grew and extracted RNA from Medicago tissue, JMM analyzed sequencing data and performed all bioinformatics analysis, JMM and JRJ generated figures, JMM, RMS, and CLM wrote the manuscript.

Introduction

The ability to fix nitrogen into the soil makes legumes an integral part of the plant ecosystem. Unfortunately, the expected increase in human population size by the year 2050 will require a higher amount of nitrogen than current legume cropping systems can fulfill (Smil, 1999). This increase in demand requires researchers to better understand and improve nitrogen fixation in current legume species. One species in particular, Medicago truncatula, is widely considered a model species for understanding nitrogen fixation due to its diploid nature, seed to seed generation time, small genome size, and the vast amount of genomic resources (Young and Udvardi, 2009). Although previous studies have identified genes associated with nodulation (Oldroyd et al., 2001; VandenBosch, 2003; Elise et al., 2005; Combier et al., 2006; Wasson, 2006; Curtin et al., 2017), the trait is highly polygenic, and a large number of genes involved in nodulation remain to be discovered. One way researcher have tried to overcome this obstacle is through the use of Genome-wide association studies (GWAS).

Genetic analysis performed on standing collections of diverse lines or accessions reveals the locations of historical recombination that differentiate each genotype. GWAS leverage this information to discover associations between genetic markers and a phenotype of interest that exhibits variation within the population. However, these strong associations typically implicate genomic regions that are too large to allow for the identification of the specific gene that underlies this variation (Flint-Garcia et al., 2005; Visscher et al., 2012; Breseghello and Coelho, 2013). In most cases, further investigation is required to identify genes surrounding each marker that may be associated with the phenotype. Furthermore, it is possible that numerous markers truly

associated with the trait are not identified as significant in GWAS, due to stringent statistical cutoffs (Storey and Tibshirani, 2003; Johnson et al., 2010; Sham and Purcell, 2014). Conversely, lowering the statistical threshold introduces false positives that are problematic for further analysis (Korte and Farlow, 2013).

Advances in next-generation sequencing technologies have allowed researchers to generate numerous reference genomes for a variety of plant species. However, many of the genes within these species remain functionally uncharacterized, limiting the amount of biological information available to interpret a gene's effect on a specific phenotype. Using technologies such as RNA-seq and microarrays, it is possible to measure quantitative levels of expression throughout the genome across multiple samples. Using these large-scale genomic datasets, it is possible to develop a functional network where one can infer a gene's function using "guilt by association." More specifically, it is possible to use transcription-based expression data from various tissues, species, and environments to capture expression profiles of genes (Stuart, 2003; Usadel et al., 2009) and then calculate a similarity metric between pairs of genes to identify if they are co-expressed with each other. Eventually, a network can be developed from these relationships where each node is a gene, and each edge would represent how well the pair of genes are co-expressed with one another (Aoki et al., 2007).

Most co-expression networks were built with the purpose of discovering and characterizing highly connected subnetworks or modules to better understand various phenotypes to provide a general resource to the community (Aoki et al., 2007; Benedito et al., 2008; Mao et al., 2009; Childs et al., 2011; Swanson-Wagner et al., 2012; Schaefer et al., 2017). Networks can also be built for multiple different purposes, such

as capturing gene function in response to environmental changes or stresses (Mochida et al., 2011; Zheng and Zhao, 2013), using a developmental network with the intent of understanding expression of various biological processes during plant development (Brady et al., 2007; Fu and Xue, 2010; Sekhon et al., 2011; Downs et al., 2013; Schaefer et al., 2014; Cho et al., 2016), or to compare and contrast networks from different species to identify conserved modules (Movahedi et al., 2012).

A recent study described a new framework to integrate co-expression networks with GWAS as a means to identify candidate genes (Schaefer et al., 2018). In Maize, Schaefer et al. (2018) ran a GWAS to identify a panel of SNPs associated with elemental accumulation in seeds. Although they were able to identify significant makers associated with regions of the genome, they would have had to manually sift through candidate genes for prioritization unless they integrated a separate source of data. They further built three co-expression networks, two from publicly available data and one from root tissue designed to represent the phenotype measured in the respective GWAS. By using the guilt by association principle from clusters of genes within a co-expression network as well as a panel of significant markers from their GWAS, Schaefer et al. 2018 combined these two datasets using their Camoco framework to identify and better prioritize candidate genes associated with elemental accumulation. We apply this framework to *Medicago truncatula* using publicly available expression datasets, and markers from a previously published GWAS focused on nodulation traits aiming to provide a functional context to the networks and further identify candidate genes associated with nodulation.

Results and Discussion

Integration of nodule focused genome-wide association study with co-expression networks

To identify candidate genes associated with nodulation traits, we used a previously published GWAS (Stanton-Geddes et al., 2013) as well as two publicly available RNA-seq datasets. The GWAS consisted of 226 *M. truncatula* accessions that were previously grown in replicate and phenotyped for five different nodulation traits as well as flowering time, trichrome density and height. By manually sifting through their most significant 50-200 SNPs based on p-value rank, the authors discovered sets of genes near significant SNP's that were previously associated with nodulation traits (Stanton-Geddes et al., 2013). Similar to other GWAS studies, the authors focused on genes that either contained or were directly adjacent to significant markers even though other genes may also be plausible candidates given their linkage to the significant markers (Branca et al., 2011). We selected a subset of these traits and markers from the study to serve as input for the GWAS/co-expression pipeline using Camoco (https://github.com/LinkageIO/Camoco) (Schaefer et al., 2018) (Table S1 in Appendix 3).

To measure the similarity of expression profiles between genes across different tissues and treatments, we used two publicly available RNA-seq expression data sets to assemble co-expression networks. The data consisted of 138 samples consisting of three different genotypes, three different tissues, four different rhizobium treatments, and presence-absence of nitrogen (Table S2 in Appendix 3). We then built six different co-expression networks using Camoco (Schaefer et al., 2018). Four of the six networks were constructed from a single tissue type (Leaf, Root, Nodule, JQL_Nodule), and the

other two networks (referred to as the “General” network and “JQL” network) were constructed from a combination of different tissue types (Table S3 in Appendix 3). The diversity of tissue types within each co-expression network allows for the detection of signals corresponding to different biological processes that may have remained undiscovered if all samples were combined into one large network (Schaefer et al., 2014; Schaefer et al., 2018).

The total number of genes that passed the co-expression network construction phase was relatively consistent among the four networks, with the general network consisting of roughly 22,000 genes each (Table S3 in Appendix 3). Genes that were excluded from each network were either not expressed, or there was not enough variation in expression between samples to robustly measure covariation. The smaller number of genes within the nodule-specific network was expected, as fewer genes are expressed in nodule tissue relative to other tissues (Benedito et al., 2008).

To test whether each network was capturing biologically meaningful relationships, we measured each network for functional enrichment. Using sets of genes coannotated to the same Gene Ontology (GO) term, the relative density (how highly an established set of functionally related genes are co-expressed with each other) was measured and compared to density values of randomly sampled gene sets of the same size. All six networks demonstrated functional enrichment of at least ten-fold within each network (Figure S1 in Appendix 3), indicating many more GO terms exhibited evidence of co-expression than expected by chance for all six networks.

Using the six co-expression networks and selected GWAS markers, we applied the Camoco pipeline to prioritize candidate causal genes. Briefly, Camoco, which was

originally described in Schaefer et al. 2018, evaluates candidate genes linked to significant GWAS marker on the basis of their co-expression with genes linked to other significant GWAS marker based on the assumption that some causal genes should exhibit strong co-expression relationships with other genes associated with the trait. Camoco is depicted in Figure 1, and the details of this analysis are provided in the Methods section. Any genes reported by Camoco with an FDR < 0.35 in the resulting analysis were considered candidate genes and included in further analysis.

The results of the Camoco framework yielded 489 discoverable genes across all GWAS trait and network combinations. Analysis of the Nod_A trait (strain occupancy in the top 5 cm of roots) with the Mt_JQL_Nodule network combination, revealed a high amount of network connectivity between genes (Z-score > 2.5). Further analysis of one of these genes, MEDTR2G101090, demonstrates how discoverable genes are identified (Figure 3). The supporting evidence that MEDTR2G101090 underlies a GWAS peak as well as being highly co-expressed with other genes across chromosomes that also underlie different GWAS peaks strongly suggests that these sets of genes are functionally related to each other and that the Camoco framework is discovering meaningful relationships.

Importance of trait and tissue specificity in co-expression networks

The number of high-confidence candidate genes discovered by Camoco varied significantly across different combinations of traits, networks, and parameters (Figure 2). It is interesting that the nodule based Mt_JQL_Nodule co-expression network discovered a consistent number of candidate genes when using a nodule focused GWAS trait, Nod_A. While this result produced the most consistent number of significant genes across parameters out of all of the network and trait combinations, this combination also

made the most biological sense since we would expect nodule tissue co-expression interactions to be representative of the GWAS trait.

Surprisingly, the root based network performed the worst when we would expect there to be biological relevance to nodulation based traits. It was the poorest performer across all GWAS traits, only exhibiting few candidate genes for the Nod_B trait (strain occupancy below the top 5 cm of roots). This result could possibly be due to the timepoint in which RNA was extracted from the roots. If RNA was extracted at an earlier timepoint when nodules were still early in development, there might have been different expression patterns within the samples allowing for the discovery candidate genes.

The leaf network was the only network that consistently identified candidates for the height trait. While this is biologically unsurprising, it is important to point out that it discovered significant genes for a few nodulation traits, suggesting that there may be some form of connection between the top of the plant and bottom of the plant when it comes to nodulation signaling. The General network which consisted of the largest number of samples and tissue types only generated consistent candidates for the Nod_B phenotype although the number of candidates was low.

These results suggest that the type of network, as well as the GWAS phenotype play an important role in discovering significant interactions. Combining many different types of tissue into one large network does not perform well as a smaller, more concise network. One reason for this is that combining expression data from very different contexts introduces more variation across each gene's profile but that variation essentially results in very generic modules that represent whole tissues. What is needed instead is a highly specific tissue, especially one that's relevant for the phenotype in the

GWAS. The use of highly specific tissue(s) will result in more subtle variation that reveals more specific functional relationships that would otherwise be lost in larger networks.

GWAS marker significance and proximity to genes are variable when integrating co-expression analysis

A common approach to interpreting GWAS studies is to manually inspect the most significant markers and look for candidates that are closest in proximity to the marker of interest. Unfortunately, the closest genes to GWAS markers may not always be the ones that are causally driving the association with the phenotype. When looking at the height trait in the leaf network, we see an increase in signal (i.e., number of Camoco-identified high-confidence candidate genes) as we increase the number of flanking genes surrounding each marker (Figure S2 in Appendix 3). When the window size is increased from 10 kb to 20kb, we see that the signal drastically increases, indicating that there are genes further out from the marker that are highly co-expressed with a subset of these genes. However, when an even larger 50kb window is used, no high-confidence genes are reported. The loss of signal at the largest interval (50kb) is expected as the number of potential candidate genes per locus increases sharply (the large majority of them being false positive as one considers candidates further from the locus peak). Ultimately, this large number of false candidate genes obscures the identification of co-expression relationships among true causal genes, and the approach no longer works. This analysis suggests that several of the GWAS loci implicated for these traits are likely driven by causal genes that are not directly adjacent to the GWAS peaks.

Similarly, the constraint of only focusing on the most significant markers leaves other candidates that are truly associated with the phenotype neglected. The Camoco framework is better able to differentiate false negatives and false positives at a lower significance threshold, by integrating associations from two different data sources. For instance, if we used the common GWAS p-value cutoff of 5×10^{-8} (Barsh et al., 2012; Panagiotou and Ioannidis, 2012; Fadista et al., 2016) we would only consider analyzing two GWAS markers from the entire Nod_A phenotype, instead, we are able to use 292 SNPs (p-value 3.00E-05 or lower) using the 10kb window parameter (Table S1 in Appendix 3) and discover candidate genes that would otherwise be ignored. Furthermore, the number of markers to include as input to identifying genes can also influence the analysis. If the trait is not highly quantitative, then it will prove difficult to find multiple genes showing similar expression profiles if many markers are included. Even quantifying the number of markers to include in the analysis can prove to be challenging. While a conservative number of markers may seem like the obvious choice, it is important to consider the false negative markers that would be excluded from a stringent cutoff could drive co-expression between sets of genes (Park et al., 2010; Visscher et al., 2012). On the other hand, if too many markers are included in the analysis, there is potential that the signal will be masked by incorporating too many spurious genes.

Identification of nodulation-related genes using co-expression and GWAS

To identify a small set of the most promising high confidence candidate genes for more investigation, we further narrowed candidate genes lists for the Nod_A trait by identifying genes that were consistently discovered across different parameter settings. Using the JQL_Nodule network, we narrowed the candidate gene lists by identifying

candidate genes that appeared in at least three out of the nine 10kb, 20kb, 50kb by 1,2,5 flanking gene combinations of parameter settings (Figure S1 in Appendix 3); this process resulted in 25 genes for further investigation (Table1). When viewing the strength of co-expression between these 25 genes (Z-score of 2.5 or higher) within the nodule network, it was observed that the majority of the genes were connected and formed a single module (Figure 4).

Interestingly, among those 25 candidate genes from the Nod_A analysis, was PEN3-like (MEDTR2G101090; Table 1), a gene that was associated with the most significant GWAS marker for the Nod_A trait (Stanton-Geddes et al., 2013). Functional validation of PEN3-like using CRISPR and Tnt1-mutated plants previously confirmed that loss-of-function of this gene resulted in decreased nodule number (Curtin et al., 2017). Another strong candidate within the module was the hub gene (gene with the highest number of connections), MEDTR7G109130, which is annotated as a P-loop nucleoside triphosphate hydrolase superfamily protein and is known to play a role in nodulation (Jayaraman et al., 2017).

Because multiple co-ex networks were able to support the discovery of strong candidate genes for Nod_B, we defined a short list of high-confidence candidates by requiring high confidence genes to be consistently considered candidates across all networks in the Nod_B trait instead of a single network, as was used in the previous example. Using every parameter that had candidate genes in the Nod_B trait, we looked for any genes that were a candidate across more than four parameters (Table 2). One promising gene, MEDTR1G012530, appeared as a candidate for 9 out of the 20 parameter settings that resulted in at least one candidate gene discovery. This gene is annotated as a TPX2 (targeting protein for Xklp2) family protein and has been shown to

be highly expressed during nodule formation (Jardinaud et al., 2016). Another promising candidate, MEDTR4G073400, which also appeared as a candidate nine times, is annotated as Synaptotagmins-1-related, which play a role in the formation of root nodules (Gavrin et al., 2017).

Overall, these results demonstrate that the co-expression/GWAS integration was able to discover genes putatively associated with nodulation processes. The genes that are directly connected to PEN3-like would serve as valuable candidates for follow-up studies due to their similarity in expression profiles across tissues. Another approach would be to look at the significance of each marker associated with a candidate gene as a way to prioritize candidate genes resulting from the Camoco analysis (Table 1). Since the Camoco pipeline generates candidate based of GWAS marker locations and density scores, candidates can be further prioritized based on their associated GWAS marker's significance value. For instance, the gene associated with the marker with the highest significance was the PEN3-like gene while our P-loop nucleoside triphosphate hydrolase superfamily protein hub gene was ranked 270 out of 523 based on the number of markers input into the analysis.

Conclusions

Using an *M. truncatula* GWAS focused on nodulation traits as well as expression data from different tissues, rhizobium strains, nitrogen treatments and accessions, we were able to identify a subset of genes surrounding GWAS markers that are highly co-expressed with one another. From these lists, we discovered a previously validated nodulation gene PEN3-like as well as several other genes whose annotations are

associated with nodulation. Uncharacterized genes within our high-confidence lists are worthy of more in-depth follow-up studies using Tnt1 or CRISPR knockouts.

Schaefer et al. 2018 developed the Camoco framework and integrated co-expression networks and GWAS in maize in order to capture variation associated with elemental uptake in seeds. Our current study used a higher-density GWAS that focused on a different phenotype, different plant species, and an expression data set that was not explicitly created for this study. One common theme between the studies is that the choice of the co-expression network matters; specifically, tissue-relevant networks derived from expression variation across diverse genotypes appear to perform the best in ranking candidate genes. This was true in maize, and we report here that this is also true in Medicago. We believe this result is likely to generalize to many other contexts, and it suggests as a community, more emphasis in the generation genotype-focused networks would be worthwhile if we hope to build resources for functional interpretation of phenotype-associated variants. It is also important to mention that we were able to generate a panel of high confidence candidate genes using two independent datasets that were not generated for this study. Having the ability to combine independent public datasets and detect novel candidate genes can prove to be resourceful to the research community.

The majority of candidate genes discovered in this analysis would have most likely been neglected by traditional GWAS analyses unless they were under the most significant markers. By combining co-expression networks with GWAS, the functional relationship between genes related to the GWAS phenotype are more likely to be discovered. The Camoco framework also demonstrated that the nearest gene to a marker is not always the one associated with phenotype. Camoco is better able to

differentiate which genes are associated with the phenotype through the use the information generated from co-expression networks. By leveraging the information obtained from a second source of data independent of the GWAS, candidates generated through this method are worth investigating due to the support generated from co-expression. Our results imply that the methods developed by Schaefer et al. 2018 can be used on a variety of GWAS. Through these findings, we believe that this whole approach developed by Schaefer et al. 2018 can be generalized to many different contexts, and it is worth applying to a wide variety of species and traits.

Acknowledgments

We would like to thank the University of Minnesota Office of Information Technology for accommodating our data storage needs and the Department of Computer Science at the University of Minnesota for server maintenance and support.

Material and Methods

Medicago experimental design and sample extraction

Three accessions from the Medicago HapMap project (HM56, HM101, HM340) were grown in greenhouse conditions. Rhizobium strains S. meliloti (KH46c) and S. medicae (WSM419), as well as nitrogen, were applied to the soil shortly after planting. Tissues was harvested and frozen in liquid nitrogen 31 days after planting. RNA was extracted using the Qiagen RNeasy Plant mini kit (Product ID: 74903). Individual nodules we pooled and extracted as a single sample for each plant.

Generation of expression data

RNA from 138 samples were sent to the University of Minnesota's Genomic Center for sequencing using Illumina HiSeq2500 100bp single-end reads. One sample required resequencing (L88) which resulted in 125bp reads. Samples were barcoded and multiplexed using Illumina TruSeq HT adapters. Fastq files were checked with Fastqc version 0.11.5 and adapters were trimmed using cutadapt version 1.8.1 with non-default parameters -m 40 and -q 30 (Andrews, 2010; Martin, 2011). Reads were then aligned to Mt_4.0 gene models and reference (http://jcvi.org/medicago/) using STAR 2.5.3a (Dobin et al., 2013), then filtered based off unique mapping scores, sorted and indexed using samtools version 1.6 (Li et al., 2009). FPKM values were generated using Cufflinks version 2.2.1 using non-default parameters of -I 20000 and --min-intron-length 5. Raw sequencing files are publicly available at (PRJNA449544).

Co-expression network construction and genome-wide association study integration

Methods used were similar to those in the previously mention co-expression GWAS integration study (Schaefer 2017). Briefly, Camoco takes a set of SNP's as input and uses their location within a genome as well the number of genes flanking a marker within a given window size to extract genes lists for testing (Figure 2). If there are multiple SNPs overlapping within the same window size, then all but the most significant SNP is discarded, varying the number of tested SNPs for each window size (Table S1 in Appendix 3). Once genes are selected for testing; each gene is then measured to see how well it is co-expressed with other genes selected within a given network. Once a density a score is generated, Camoco will resample a random set of genes equal in size to the testing set and compare the density score of the resample genes verses the

observed. This process was iterated 1000 times to generate an FDR for a single gene then repeated for each gene, GWAS trait, and network combination. To account for the varying amount of linkage disequilibrium within different regions of the population, we used multiple different window sizes and number of flanking genes (Stanton-Geddes et al., 2013). Any gene that had an FDR < 0.35 was called “candidate” and included in further analysis.

FPKM expression tables were used as input into Camoco (https://github.com/LinkageIO/Camoco) using the Mt_4.1 reference genome. Non-default parameters used to build each network included rawtype='RNASEQ', max_gene_missing_data=0.5, max_accession_missing_data=0.5, min_single_sample_expr=1, min_expr=0.001, quantile=False, max_val=300, sep=',’. Network health statistics were generated using, GO terms from (http://jcvi.org/medicago) and 1000 bootstraps. SNPs were integrated into camoco using built-in functions, and per gene, density measurements were run with 1,000 bootstraps, 10kb, 20kb and 50kb window sizes and 1, 2, and 5 flanking genes. Figures were created using ggplot2 (Wickham, 2006).

Gene	Number of connections (Z-score 2.5 or higher)	SNP_position	GWAS - log10(p.val)	Rank (out of 523)	Annotation
MEDTR2G101090	8	chr2:43448968	7.591607	1	drug resistance transporter-like ABC domain protein
MEDTR8G074920	4	chr8:31665171	6.753532	11	receptor-like kinase theseus protein
MEDTR2G100280	4	chr2:43061039	6.743592	12	RNA exonuclease-like protein
MEDTR4G018770	4	chr4:5776217	6.509395	19	GDP-mannose transporter GONST3
MEDTR3G026650	6	chr3:8183997	6.177657	53	GDP-fucose protein O-fucosyltransferase
MEDTR4G059870	4	chr4:22091245	5.827601	114	C2H2 and C2HC zinc finger protein, putative
MEDTR4G019910	4	chr4:6362962	5.7494	139	SnoaL-like domain protein
MEDTR5G076270	1	chr5:32504251	5.707181	156	auxin response factor 2
MEDTR6G084440	2	chr6:31605458	5.678609	161	DUF1666 family protein
MEDTR2G090960	9	chr2:39088095	5.657328	171	TCP family transcription factor
MEDTR4G104350	2	chr4:43099392	5.512627	210	DNA polymerase III subunit gamma/tau
MEDTR7G102310	6	chr7:41285876	5.493289	220	rhodanese/cell cycle control phosphatase superfamily protein
MEDTR5G093580	5	chr5:40860194	5.415629	252	co-factor for nitrate, reductase and xanthine dehydrogenase
MEDTR3G019490	5	chr3:5482913	5.410043	257	S-locus lectin kinase family protein
MEDTR7G109130	16	chr7:44591633	5.381151	270	P-loop nucleoside triphosphate hydrolase superfamily protein
MEDTR8G027385	1	chr8:9668134	5.239786	350	Endomembrame Family Protein
MEDTR4G126160	11	chr4:52449376	5.231223	358	cytokinin oxidase/dehydrogenase-like protein
MEDTR7G076250	5	chr7:28686036	5.221541	366	zinc finger, C3HC4 type (RING finger) protein
MEDTR4G058970	10	chr4:21744831	5.102555	448	homeodomain leucine zipper protein
MEDTR7G075580	13	chr7:28296141	5.067043	470	cytochrome P450 family protein
MEDTR1G075610	5	chr1:33462984	5.06158	474	cyclin-dependent kinase
MEDTR2G096950	8	chr2:41430755	5.050944	485	kinase 1B
MEDTR1G070455	9	chr1:31235133	5.044264	491	WRKY transcription factor
MEDTR3G111650	10	chr3:52196531	5.019337	507	hypothetical protein
MEDTR1G080690	0	chr1:35874811	5.009149	517	TPX2 (targeting protein for Xklp2) family protein

Table 1: List of genes that were discoverable across all six parameters (10kb, 20kb and 1,2,5 flanking genes) for the Nod_A phenotype using the Mt_JQL Nodule GWAS.

Gene	Numer of hits across parameters and terms	Annotation
MEDTR4G027195	10	N/A
MEDTR4G035980	10	pectinesterase/pectinesterase inhibitor
MEDTR1G012530	9	TPX2 (targeting protein for Xklp2) family protein
MEDTR4G073400	9	Synaptotagmin-1-related
MEDTR2G073540	8	cysteine-rich RLK (receptor-like kinase) protein
MEDTR1G028960	6	glycolipid transfer protein (GLTP) family protein
MEDTR1G037520	5	N/A
MEDTR1G040105	5	methylenetetrahydrofolate reductase
MEDTR2G048855	5	pentatricopeptide (PPR) repeat protein
MEDTR2G090960	5	TCP family transcription factor
MEDTR2G450720	5	SAM domain (sterile alpha motif) protein, putative
MEDTR3G088820	5	PPR containing plant-like protein
MEDTR4G087510	5	O-acetylserine (thiol) lyase
MEDTR5G053950	5	allene oxide cyclase
MEDTR5G065080	5	purine permease
MEDTR5G094290	5	tubulin folding cofactor A
MEDTR6G023600	5	short-chain dehydrogenase/reductase
MEDTR6G048290	5	PPPDE thiol peptidase family protein, putative
MEDTR7G039370	5	origin recognition complex subunit 6
MEDTR8G432620	5	methyltransferase

Table2: List of genes that were discoverable for at least 5 different parameters across all networks for the Nod_B trait

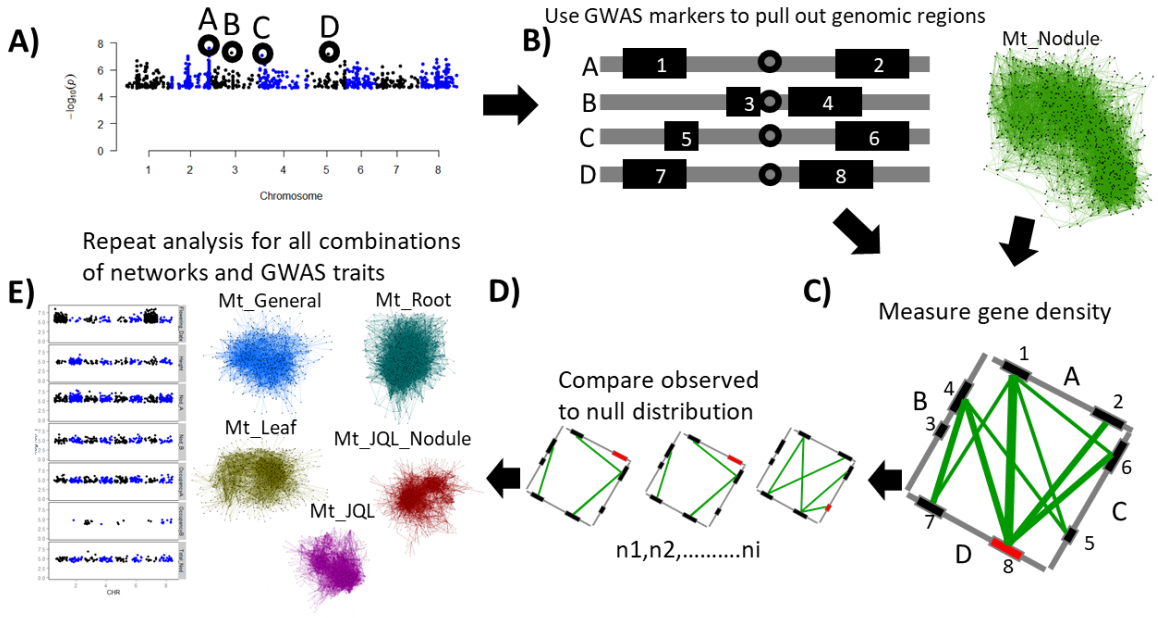


Figure 1. GWAS and co-expression pipeline.
GWAS/co-expression pipeline using Camoco. A) Manhattan plot represents DNA markers used as input for Camoco, bold black circles represent a subset of markers used for illustrative purposes. B) Regions along a chromosome from previously selected markers are represented as grey bars, genes are represented as black rectangles. C) Genes from previously identified intervals are then selected from the co-expression network for per-gene density measurements. Colored lines represent the strength of co-expression between two genes in a co-expression network. Wider lines, represent genes that are more strongly co-expressed. The red box represents the current gene being measured for density. D) Per-gene density measurement of random sub-networks equal in size to the testing set. E) Other GWAS traits and networks used for analysis.

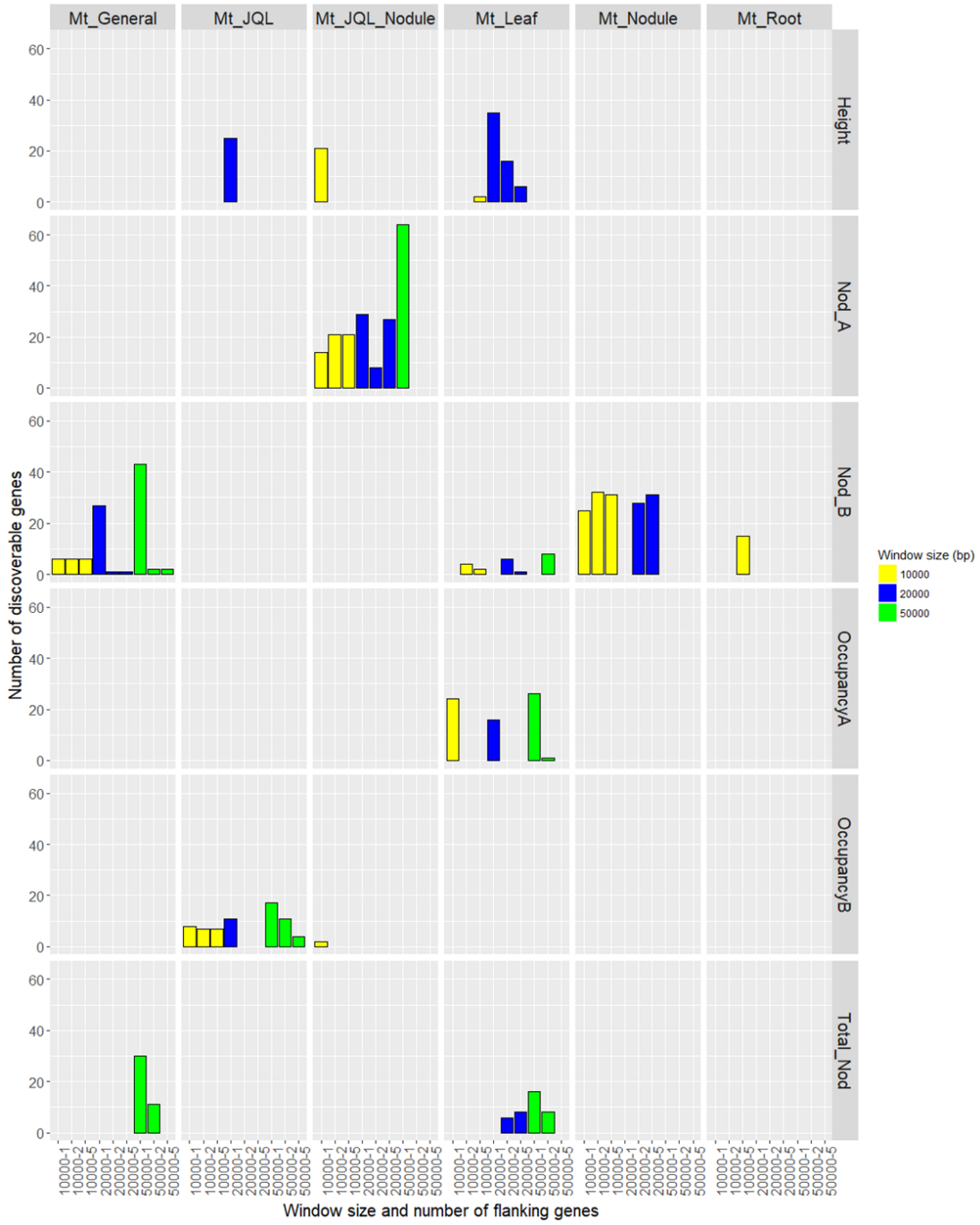


Figure 2. Co-expression/GWAS discoverable gene summary.
Number of discoverable genes (FDR < 0.35) obtained from co-expression/GWAS integration. Colors represent the window size parameters used for our analysis.

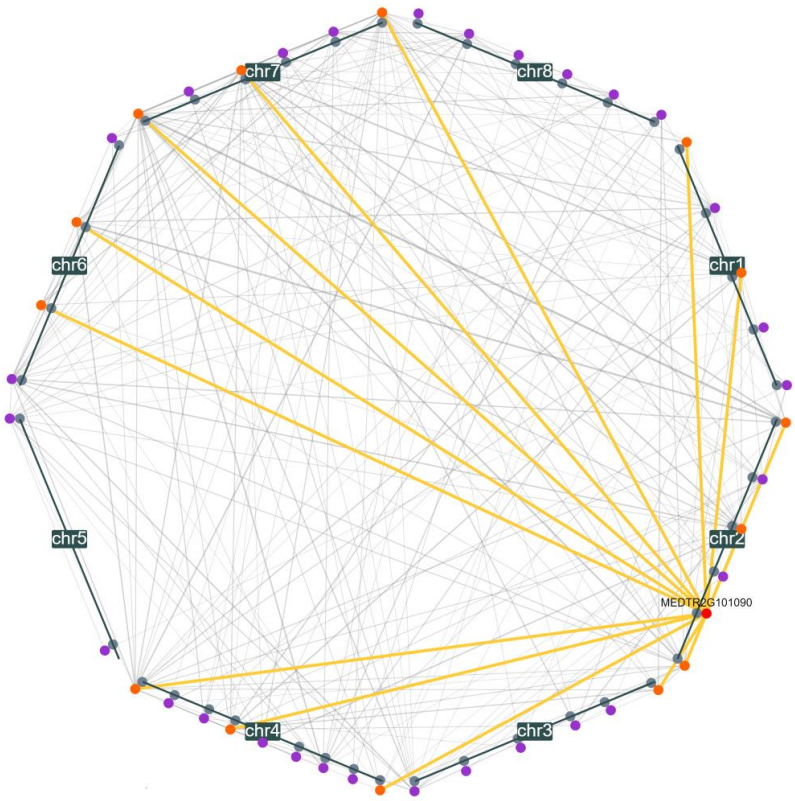


Figure 3. Nodule_A discoverable genes in the Mt_JQL_Nodule network. Polywas diagram of the connectivity of discoverable genes (FDR < 0.35) to MEDTR2G101090 within the JQL_nodule network for the Nod_A trait. Grey circles represent GWAS markers, colored circles represent genes, with MEDTR2G101090 in red, its first neighbors in orange, and other discoverable genes in purple. Grey lines represent co-expression between genes (minimum Z-Score of 2.5); the wider the line, the stronger the co-expression between genes.

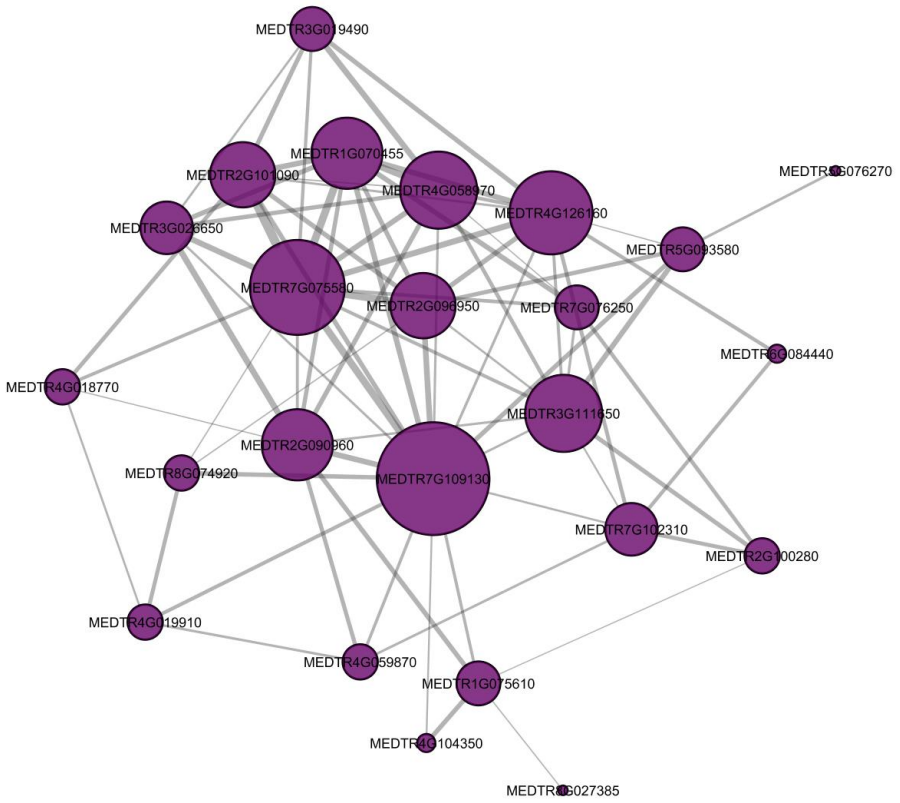


Figure 4. Overlap of Nod_A candidates in the Mt_JQL_Nodule network
Candidate genes for the JQL_nodule network for the Nod_A trait. Purple circles represent genes and grey lines represent co-expression between genes (minimum Z-Score of 2.5). The larger the circle, the more connections it has with other genes. The wider the line, the stronger the co-expression between genes.

Conclusions and future work

While significant progress has been made towards understanding the genomes and transcriptomes of legume species, the majority of genes still remain functionally uncharacterized. To meet the demands of the global food supply by the year 2050, significant progress must be made to achieve the desired agricultural output. One way researchers are trying to achieve this is through creating genetically modified organisms (GMO). While GMOs are widely used throughout the US, they are under tight regulation due to the concern with its safety.

Chapter two addressed an important issue that impacts the safety of GMO use in agriculture. Specifically, we investigated mutations rates within agrobacterium-mediated soybean lines by reanalyzing a publicly available dataset where the previous authors reported tens of thousands of mutations transcriptome-wide. Through reanalysis, we discovered that almost all mutations reported in their analysis were not due to transgenesis, but were due to improper genotype identity, heterogeneity, and non-optimal bioinformatic handling of the data. Their misinterpretation of the data highlights an important issue with using next-generation sequencing analysis for GMO regulation. Using two different analysis methods led to two different results, demonstrating that analysis pipelines for GMO regulations must be closely analyzed to reduce the number of false positives and/or false negatives. Further studies applying next-generation sequencing to study the GMO process would be beneficial, e.g. analyzing the variation induced though transgenesis, but our work highlights the importance of robust bioinformatics pipelines in drawing conclusions that can have broad impact on the GMO regulatory environment.

Chapter three explores the type of variation induced through CRISPR/Cas9-based mutagenesis by focusing on transgene integration sites in a series of CRISPR/Cas9 soybean lines. Through this analysis, we were able to see similar types of variation as reported in previous literature including deletions, insertions, duplications, as well as two independent instances where a transgene was inserted into a CRIPSR target site. Over half of the transgene insertions identified by this work occurred within different gene models, stressing the importance of knowing where transgenes are inserted in the genome when phenotyping lines that still contain transgene(s).

While Chapter three as well as other literature address the idea that agrobacterium mediated transformation induces variation at transgene integration sites, it remains to be seen if there is variation induced genome-wide. While our preliminary results appear to show evidence of off-targeting in CRISPR lines based on comparison to non-CRISPR lines, the number of samples needed to draw a generalizable conclusion is too low. More CRISPR lines will need to be sequenced in the future to follow up on our initial findings.

Chapter four explores the concept of integrating genome-wide association studies (GWAS) with co-expression networks. By combining two separate sources of data, we were able to prioritize candidate genes underlying GWAS peaks that we believe to be associated with nodulation. One of the genes identified, PEN-3-like, was previously validated by a different group using knockouts to demonstrate its functional role in nodulation. Similarly, other genes identified through this pipeline had annotations that could be directly associated with the phenotype.

Using these advances, we can now attempt to find new genes to target for breeding purposes. Instead of having to manually scan under each GWAS and trying to decide which gene(s) to target, researchers will be able to systematically prioritize candidate lists for future studies. This concept can be applied to a variety of different quantitative traits and a variety of different agricultural species to identify genes of interest with the intent of improving elite breeding lines to meet the future demands of the global food supply.

Bibliography

Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Methods 8: 61–65

Altpeter F, Baisakh N, Beachy R, Bock R, Capell T, Christou P, Daniell H, Datta K, Datta S, Dix PJ, et al (2005) Particle bombardment and the genetic enhancement of crops: myths and realities. Mol Breed 15: 305–327

Anderson JE, Michno JM, Kono TJY, Stec AO, Campbell BW, Curtin SJ, Stupar RM (2016) Genomic variation and DNA repair associated with soybean transgenesis: a comparison to cultivars and mutagenized plants. BMC Biotechnol 16: 41

Andrews S (2010) FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ http://www.bioinformatics.babraham.ac.uk/projects/

Aoki K, Ogata Y, Shibata D (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol 48: 381–390

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr. Protoc. Bioinforma. John Wiley & Sons, Inc., Hoboken, NJ, USA, p 11.10.1-11.10.33

Baltes NJ, Voytas DF (2015) Enabling plant synthetic biology through genome engineering. Trends Biotechnol 33: 120–131

Barsh GS, Copenhaver GP, Gibson G, Williams SM (2012) Guidelines for Genome-Wide Association Studies. PLoS Genet 8: e1002812

Belhaj K, Chaparro-Garcia A, Kamoun S, Nekrasov V (2013) Plant genome editing made easy: targeted mutagenesis in model and crop plants using the CRISPR/Cas system. Plant Methods 9: 39

Benedito VA, Torres-Jerez I, Murray JD, Andriankaja A, Allen S, Kakar K, Wandrey M, Verdier J, Zuber H, Ott T, et al (2008) A gene expression atlas of the model legume Medicago truncatula. Plant J 55: 504–513

Benfey PN, Chua N-H (1990) The Cauliflower Mosaic Virus 35S Promoter: Combinatorial Regulation of Transcription in Plants. Science (80-) 250: 959–966

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59

Bergelson J, Buckler ES, Ecker JR, Nordborg M, Weigel D (2016) A Proposal Regarding Best Practices for Validating the Identity of Genetic Stocks and the Effects of Genetic Variants: Table 1. Plant Cell 28: 606–609

Bhaya D, Davison M, Barrangou R (2011) CRISPR-Cas Systems in Bacteria and

Archaea: Versatile Small RNAs for Adaptive Defense and Regulation. Annu Rev Genet 45: 273–297

Bibikova M, Beumer K, Trautman JK, Carroll D (2003) Enhancing gene targeting with designed zinc finger nucleases. Science 300: 764

Bibikova M, Golic M, Golic KG, Carroll D (2002) Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. Genetics 161: 1169–1175

Bortesi L, Fischer R (2015) The CRISPR/Cas9 system for plant genome editing and beyond. Biotechnol Adv 33: 41–52

Brady SM, Orlando DA, Lee J-Y, Wang JY, Koch J, Dinneny JR, Mace D, Ohler U, Benfey PN (2007) A High-Resolution Root Spatiotemporal Map Reveals Dominant Expression Patterns. Science (80-) 318: 801–806

Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti AK, Woodward JE, May GD, Gentzbittel L, et al (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume Medicago truncatula. Proc Natl Acad Sci 108: E864–E870

Breseghello F, Coelho ASG (2013) Traditional and modern plant breeding methods with examples in rice (Oryza sativa L.). J Agric Food Chem 61: 8277–86

Butler NM, Atkins PA, Voytas DF, Douches DS (2015) Generation and Inheritance of Targeted Mutations in Potato (Solanum tuberosum L.) Using the CRISPR/Cas System. PLoS One 10: e0144591

Cai CQ, Doyon Y, Ainley WM, Miller JC, DeKelver RC, Moehle EA, Rock JM, Lee YL, Garrison R, Schulenberg L, et al (2009) Targeted transgene integration in plant cells using designed zinc finger nucleases. Plant Mol Biol 69: 699–709

Cai Y, Chen L, Liu X, Guo C, Sun S, Wu C, Jiang B, Han T, Hou W (2018) CRISPR/Cas9-mediated targeted mutagenesis of GmFT2a delays flowering time in soya bean. Plant Biotechnol J 16: 176–185

Cai Y, Chen L, Liu X, Sun S, Wu C, Jiang B, Han T, Hou W (2015) CRISPR/Cas9-Mediated Genome Editing in Soybean Hairy Roots. PLoS One 10: e0136064

Canver MC, Bauer DE, Dass A, Yien YY, Chung J, Masuda T, Maeda T, Paw BH, Orkin SH (2014) Characterization of genomic deletion efficiency mediated by clustered regularly interspaced palindromic repeats (CRISPR)/cas9 nuclease system in mammalian cells. J Biol Chem 289: 21312–21324

Cermak T, Curtin SJ, Gil-Humanes J, Čegan R, Kono TJY, Konečná E, Belanto JJ, Starker CG, Mathre JW, Greenstein RL, et al (2017) A multi-purpose toolkit to enable advanced genome engineering in plants. Plant Cell tpc.00922.2016

Cermak T, Doyle EL, Christian M, Wang L, Zhang Y, Schmidt C, Baller J a, Somia N V, Bogdanove AJ, Voytas DF (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res 39: e82

Chandrasekaran J, Brumin M, Wolf D, Leibman D, Klap C, Pearlsman M, Sherman A, Arazi T, Gal-On A (2016) Development of broad virus resistance in non-transgenic cucumber using CRISPR/Cas9 technology. Mol Plant Pathol 17: 1140–1153

Chee PP, Fober KA, Slightom JL (1989) Transformation of Soybean (Glycine max) by Infecting Germinating Seeds with Agrobacterium tumefaciens. Plant Physiol 91: 1212–8

Childs KL, Davidson RM, Buell CR (2011) Gene coexpression network analysis as a source of functional annotation for rice genes. PLoS One 6: e22196

Chilton M-DM (2003) Targeted Integration of T-DNA into the Tobacco Genome at Double-Stranded Breaks: New Insights on the Mechanism of T-DNA Integration. PLANT Physiol 133: 956–965

Cho KK, Cho KK, Sohn H, Ha IJ, Hong S, Lee H, Kim Y-M, Nam MH (2016) Network analysis of the metabolome and transcriptome reveals novel regulation of potato pigmentation. J Exp Bot 67: 1519–1533

Christian ML, Demorest ZL, Starker CG, Osborn MJ, Nyquist MD, Zhang Y, Carlson DF, Bradley P, Bogdanove AJ, Voytas DF (2012) Targeting G with TAL effectors: a comparison of activities of TALENs constructed with NN and NK repeat variable di-residues. PLoS One 7: e45383

Christou P (1992) Genetic transformation of crop plants using microprojectile bombardment. Plant J 2: 275–281

Clark KA, Krysan PJ (2010) Chromosomal translocations are a common phenomenon in Arabidopsis thaliana T-DNA insertion lines. Plant J 64: 990–1001

Collier R, Dasgupta K, Xing Y-P, Hernandez BT, Shao M, Rohozinski D, Kovak E, Lin J, de Oliveira MLP, Stover E, et al (2017) Accurate measurement of transgene copy number in crop plants using droplet digital PCR. Plant J 90: 1014–1025

Combier J-P, Frugier F, de Billy F, Boualem A, El-Yahyaoui F, Moreau S, Vernié T, Ott T, Gamas P, Crespi M, et al (2006) MtHAP2-1 is a key transcriptional regulator of symbiotic nodule development regulated by microRNA169 in Medicago truncatula. Genes Dev 20: 3084–8

Curtin SJ, Tiffin P, Guhlin J, Trujillo DI, Burghardt LT, Atkins P, Baltes NJ, Denny R, Voytas DF, Stupar RM, et al (2017) Validating Genome-Wide Association Candidates Controlling Quantitative Variation in Nodulation. Plant Physiol 173: 921–931

Curtin SJ, Voytas DF, Stupar RM (2012) Genome Engineering of Crops with Designer Nucleases. Plant Genome J 5: 42

Curtin SJ, Xiong Y, Michno JM, Campbell BW, Stec AO, Čermák T, Starker C, Voytas DF, Eamens AL, Stupar RM (2018) CRISPR/Cas9 and TALENs generate heritable mutations for genes involved in small RNA processing of Glycine max and Medicago truncatula. Plant Biotechnol J 16: 1125–1137

Curtin SJ, Zhang F, Sander JD, Haun WJ, Starker C, Baltes NJ, Reyon D, Dahlborg EJ,

Goodwin MJ, Coffman AP, et al (2011) Targeted mutagenesis of duplicated genes in soybean with zinc-finger nucleases. Plant Physiol 156: 466–73

Curtis MD (2003) A Gateway Cloning Vector Set for High-Throughput Functional Analysis of Genes in Planta. PLANT Physiol 133: 462–469

D'Halluin K, Vanderstraeten C, Van Hulle J, Rosolowska J, Van Den Brande I, Pennewaert A, D'Hont K, Bossut M, Jantz D, Ruiter R, et al (2013) Targeted molecular trait stacking in cotton through targeted double-strand break induction. Plant Biotechnol J 11: 933–941

Daber R, Sukhadia S, Morrissette JJD (2013) Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets. Cancer Genet 206: 441–448

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498

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

Dong W, Yang L, Shen K, Kim B, Kleter GA, Marvin HJ, Guo R, Liang W, Zhang D (2008) GMDD: a database of GMO detection methods. BMC Bioinformatics 9: 260

Downs GS, Bi Y-M, Colasanti J, Wu W, Chen XX, Zhu T, Rothstein SJ, Lukens LN (2013) A Developmental Transcriptional Network for Maize Defines Coexpression Modules. PLANT Physiol 161: 1830–1843

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. Science (80-) 323: 133–138

Elise S, Etienne-Pascal J, de Fernanda C-N, Gérard D, Julia F (2005) The Medicago truncatula SUNN Gene Encodes a CLV1-like Leucine-rich Repeat Receptor Kinase that Regulates Nodule Number and Root Length. Plant Mol Biol 58: 809–822

Endo M, Kumagai M, Motoyama R, Sasaki-Yamagata H, Mori-Hosokawa S, Hamada M, Kanamori H, Nagamura Y, Katayose Y, Itoh T, et al (2015) Whole-Genome Analysis of Herbicide-Tolerant Mutant Rice Generated by Agrobacterium-Mediated Gene Targeting. Plant Cell Physiol 56: 116–125

Estruch JJ, Carozzi NB, Desai N, Duck NB, Warren GW, Koziel MG (1997) Transgenic plants: An emerging approach to pest control. Nat Biotechnol 15: 137–141

Fadista J, Manning AK, Florez JC, Groop L (2016) The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur J Hum Genet 24: 1202–1205

Farno L, Keim KR, Edwards LH (2003) Registration of 'Washita' Soybean. Crop Sci 43: 1125

Fasoula VA, Boerma HR (2005) Divergent selection at ultra-low plant density for seed protein and oil content within soybean cultivars. F Crop Res 91: 217–229

Fasoula VA, Boerma HR (2007) Intra-Cultivar Variation for Seed Weight and Other Agronomic Traits within Three Elite Soybean Cultivars. Crop Sci 47: 367

Feng Z, Mao Y, Xu N, Zhang B, Wei P, Yang D-L, Wang Z, Zhang Z, Zheng R, Yang L, et al (2014) Multigeneration analysis reveals the inheritance, specificity, and patterns of CRISPR/Cas-induced gene modifications in Arabidopsis. Proc Natl Acad Sci 111: 4632–4637

Feng Z, Zhang B, Ding W, Liu X, Yang D-L, Wei P, Cao F, Zhu S, Zhang F, Mao Y, et al (2013) Efficient genome editing in plants using a CRISPR/Cas system. Cell Res 23: 1229–1232

Ficklin SP, Feltus F a (2011) Gene Coexpression Network Alignment and Conservation of Gene Modules between Two Grass Species: Maize and Rice. PLANT Physiol 156: 1244–1256

Ficklin SP, Luo F, Feltus FA (2010) The Association of Multiple Interacting Genes with Specific Phenotypes In Rice (Oryza sativa) Using Gene Co-Expression Networks. Plant Physiol 154: 13–24

Flint-Garcia SA, Thuillet A-CC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44: 1054–1064

Flores T, Karpova O, Su X, Zeng P, Bilyeu K, Sleper DA, Nguyen HT, Zhang ZJ (2008) Silencing of GmFAD3 gene by siRNA leads to low α-linolenic acids (18:3) of fad3-mutant phenotype in soybean [Glycine max (Merr.)]. Transgenic Res 17: 839–850

Fu F-F, Xue H-W (2010) Coexpression analysis identifies Rice Starch Regulator1, a rice AP2/EREBP family transcription factor, as a novel rice starch biosynthesis regulator. Plant Physiol 154: 927–938

Fukushima A, Nishizawa T, Hayakumo M, Hikosaka S, Saito K, Goto E, Kusano M (2012) Exploring Tomato Gene Functions Based on Coexpression Modules Using Graph Clustering and Differential Coexpression Approaches. PLANT Physiol 158: 1487–1502

Gao C, Nielsen KK (2013) Comparison Between Agrobacterium-Mediated and Direct Gene Transfer Using the Gene Gun. Biolistic DNA Deliv. Humana Press, Totowa, NJ, pp 3–16

Gavrin A, Kulikova O, Bisseling T, Fedorova EE (2017) Interface Symbiotic Membrane Formation in Root Nodules of Medicago truncatula: the Role of Synaptotagmins MtSyt1, MtSyt2 and MtSyt3. Front Plant Sci 8: 201

Glenn KC, Alsop B, Bell E, Goley M, Jenkinson J, Liu B, Martin C, Parrott W, Souder C, Sparks O, et al (2017) Bringing New Plant Varieties to Market: Plant Breeding and Selection Practices Advance Beneficial Characteristics while Minimizing

Unintended Changes. Crop Sci 57: 2906

Goda H, Sasaki E, Akiyama K, Maruyama-Nakashita A, Nakabayashi K, Li W, Ogawa M, Yamauchi Y, Preston J, Aoki K, et al (2008) The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. Plant J 55: 526–42

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17: 333–351

Guttikonda SK, Marri P, Mammadov J, Ye L, Soe K, Richey K, Cruse J, Zhuang M, Gao Z, Evans C, et al (2016) Molecular Characterization of Transgenic Events Using Next Generation Sequencing Approach. PLoS One 11: e0149515

Haun WJ, Hyten DL, Xu WW, Gerhardt DJ, Albert TJ, Richmond T, Jeddeloh J a, Jia G, Springer NM, Vance CP, et al (2011) The Composition and Origins of Genomic Variation among Individuals of the Soybean Reference Cultivar Williams 82. Plant Physiol 155: 645–55

Hernandez-Garcia CM, Martinelli AP, Bouchard R a., Finer JJ (2009) A soybean (Glycine max) polyubiquitin promoter gives strong constitutive expression in transgenic soybean. Plant Cell Rep 28: 837–849

Hinchee MAW, Connor-Ward D V., Newell CA, McDonnell RE, Sato SJ, Gasser CS, Fischhoff DA, Re DB, Fraley RT, Horsch RB (1988) Production of Transgenic Soybean Plants Using Agrobacterium-Mediated DNA Transfer. Nat Biotechnol 6: 915–922

Hirsch CD, Springer NM, Hirsch CN (2015) Genomic limitations to RNA sequencing expression profiling. Plant J 84: 491–503

Holst-Jensen A (2009) Testing for genetically modified organisms (GMOs): Past, present and future perspectives. Biotechnol Adv 27: 1071–1082

Homrich MS, Wiebke-Strohm B, Weber RLM, Bodanese-Zanettini MH (2012) Soybean genetic transformation: A valuable tool for the functional study of genes and the production of agronomically improved plants. Genet Mol Biol 35: 998–1010

Itkin M, Heinig U, Tzfadia O, Bhide AJ, Shinde B, Cardenas PD, Bocobza SE, Unger T, Malitsky S, Finkers R, et al (2013) Biosynthesis of Antinutritional Alkaloids in Solanaceous Crops Is Mediated by Clustered Genes. Science (80-) 341: 175–179

Jackson SA, Zhang P, Chen WP, Phillips RL, Friebe B, Muthukrishnan S, Gill BS (2001) High-resolution structural analysis of biolistic transgene integration into the genome of wheat. TAG Theor Appl Genet 103: 56–62

Jacobs TB, LaFayette PR, Schmitz RJ, Parrott W a (2015) Targeted genome modifications in soybean with CRISPR/Cas9. BMC Biotechnol 15: 1–10

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36: 338–345

James C, Krattiger A (1996) Global Review of the Field Testing and Commercialization of Transgenic Plants, 1986 to 1995: The First Decade of Crop Biotechnology.

Jardinaud M-F, Boivin S, Rodde N, Catrice O, Kisiala A, Lepage A, Moreau S, Roux B, Cottret L, Sallet E, et al (2016) A Laser Dissection-RNAseq Analysis Highlights the Activation of Cytokinin Pathways by Nod Factors in the Medicago truncatula Root Epidermis. Plant Physiol 171: 2256–2276

Jayaraman D, Richards AL, Westphall MS, Coon JJ, Ané J-M (2017) Identification of the phosphorylation targets of symbiotic receptor-like kinases using a high-throughput multiplexed assay for kinase specificity. Plant J 90: 1196–1207

Jiang C, Mithani A, Gan X, Belfield EJ, Klingler JP, Zhu J-K, Ragoussis J, Mott R, Harberd NP (2011) Regenerant Arabidopsis Lineages Display a Distinct Genome-Wide Spectrum of Mutations Conferring Variant Phenotypes. Curr Biol 21: 1385–1390

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C-S, et al (2017) Improved maize reference genome with single-molecule technologies. Nature 546: 524

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science (80-) 337: 816–821

Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, O'Brien SJ (2010) Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics 11: 724

Jupe F, Rivkin AC, Michael TP, Zander M, Motley TS, Sandoval JP, Slotkin KR, Chen H, Castagnon R, Nery JR, et al (2018) The complex architecture of plant transgene insertions. bioRxiv 282772

Kashima K, Mejima M, Kurokawa S, Kuroda M, Kiyono H, Yuki Y (2015) Comparative whole-genome analyses of selection marker–free rice-based cholera toxin B-subunit vaccine lines and wild-type lines. BMC Genomics 16: 48

Kawakatsu T, Kawahara Y, Itoh T, Takaiwa F (2013) A Whole-Genome Analysis of a Transgenic Rice Seed-Based Edible Vaccine Against Cedar Pollen Allergy. DNA Res 20: 623–631

Kessler DA, Taylor MR, Maryanski JH, Flamm EL, Kahl LS (1992) The safety of foods developed by biotechnology. Science 256: 1747–9, 1832

Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J 50: 347–63

Kim YG, Cha J, Chandrasegaran S (1996) Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. Proc Natl Acad Sci U S A 93: 1156–1160

Koren S, Phillippy AM (2015) One chromosome, one contig: complete microbial

genomes from long-read sequencing and assembly. Curr Opin Microbiol 23: 110–120

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9: 29

Kovalic D, Garnaat C, Guo L, Yan Y, Groat J, Silvanovich A, Ralston L, Huang M, Tian Q, Christian A, et al (2012) The Use of Next Generation Sequencing and Junction Sequence Analysis Bioinformatics to Achieve Molecular Characterization of Crops Improved Through Modern Biotechnology. Plant Genome J 0: 0

Labra M, Vannini C, Grassi F, Bracale M, Balsemin M, Basso B, Sala F (2004) Genomic stability in Arabidopsis thaliana transgenic plants obtained by floral dip. Theor Appl Genet 109: 1512–1518

Lambirth KC, Whaley AM, Blakley IC, Schlueter JA, Bost KL, Loraine AE, Piller KJ (2015a) A Comparison of transgenic and wild type soybean seeds : analysis of transcriptome profiles using RNA-Seq. BMC Biotechnol 15: 1–17

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

Lambirth KC, Whaley AM, Schlueter JA, Piller KJ, Bost KL (2016) Transcript Polymorphism Rates in Soybean Seed Tissue Are Increased in a Single Transformant of Glycine max. Int J Plant Genomics 2016: 1–12

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

Lareau CA, Clement K, Hsu JY, Pattanayak V, Joung JK, Aryee MJ, Pinello L (2018) Response to “Unexpected mutations after CRISPR–Cas9 editing in vivo.” Nat Methods 15: 238–239

Leal LG, López C, López-Kleine L (2014) Construction and comparison of gene co-expression networks shows complex plant immune responses. PeerJ 2: e610

Lee JM (2003) Genomic Gene Clustering Analysis of Pathways in Eukaryotes. Genome Res 13: 875–882

Levene MJ (2003) Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. Science (80-) 299: 682–686

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–9

Li J-F, Norville JE, Aach J, McCormack M, Zhang D, Bush J, Church GM, Sheen J (2013a) Multiplex and homologous recombination–mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. Nat Biotechnol

31: 688–691

Li T, Liu B, Spalding MH, Weeks DP, Yang B (2012) High-efficiency TALEN-based gene editing produces disease-resistant rice. Nat Biotechnol 30: 390–392

Li W, Teng F, Li T, Zhou Q (2013b) Simultaneous generation and germline transmission of multiple gene mutations in rat using CRISPR-Cas systems. Nat Biotechnol 31: 684–686

Li Z, Liu Z-B, Xing A, Moon BP, Koellhoffer JP, Huang L, Ward RT, Clifton E, Falco SC, Cigan AM (2015) Cas9-Guide RNA Directed Genome Editing in Soybean. Plant Physiol 169: 960–970

Lowe K, Wu E, Wang N, Hoerster G, Hastings C, Cho M-J, Scelonge C, Lenderts B, Chamberlin M, Cushatt J, et al (2016) Morphogenic Regulators Baby boom and Wuschel Improve Monocot Transformation. Plant Cell 28: 1998–2015

Makarevitch I, Svitashev SK, Somers DA (2003) Complete sequence analysis of transgene loci from plants transformed via microprojectile bombardment. Plant Mol Biol 52: 421–432

Mao L, Van Hemert JL, Dash S, Dickerson JA (2009) Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics 10: 346

Mao Y, Zhang H, Xu N, Zhang B, Gou F, Zhu J-K (2013) Application of the CRISPR–Cas System for Efficient Genome Engineering in Plants. Mol Plant 6: 2008–2011

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133–141

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10

Mauro Vigani AO (2015) Patterns and Determinants of GMO Regulations: An Overview of Recent Evidence.

McBlain BA, Fioritto RJ, St. Martin SK, Calip-Dubois AJ, Schmitthenner AF, Cooper RL, Martin RJ (1993) Registration of 'Thorne' Soybean. Crop Sci 33: 1406

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier A-S (2014) Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. PLoS One 9: e106689

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303

Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31–46

Michno JM, Wang X, Liu J, Curtin SJ, Kono TJ, Stupar RM (2015) CRISPR/Cas mutagenesis of soybean and Medicago truncatula using a new web-tool and a

modified Cas9 enzyme. GM Crops Food 6: 243–252

Miyao A, Nakagome M, Ohnuma T, Yamagata H, Kanamori H, Katayose Y, Takahashi A, Matsumoto T, Hirochika H (2012) Molecular Spectrum of Somaclonal Variation in Regenerated Rice Revealed by Whole-Genome Sequencing. Plant Cell Physiol 53: 256–264

Mochida K, Uehara-Yamaguchi Y, Yoshida T, Sakurai T, Shinozaki K (2011) Global landscape of a co-expressed gene network in barley and its application to gene discovery in Triticeae crops. Plant Cell Physiol 52: 785–803

Morisse P, Lecroq T, Lefebvre A (2017) Hybrid correction of highly noisy Oxford Nanopore long reads using a variable-order de Bruijn graph. bioRxiv 238808

Morisset D, Stebih D, Cankar K, Zel J, Gruden K (2008) Alternative DNA amplification methods to PCR and their application in GMO detection: a review. Eur Food Res Technol 227: 1287–1297

Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. Genomics 92: 255–264

Movahedi S, Van Bel M, Heyndrickx KS, Vandepoele K (2012) Comparative co-expression analysis in plant biology. Plant Cell Environ 35: 1787–98

Nacry P, Camilleri C, Courtial B, Caboche M, Bouchez D (1998) Major chromosomal rearrangements induced by T-DNA transformation in Arabidopsis. Genetics 149: 641–50

Nunes ACS, Vianna GR, Cuneo F, Amaya-Farfán J, de Capdeville G, Rech EL, Aragão FJL (2006) RNAi-mediated silencing of the myo-inositol-1-phosphate synthase gene (GmMIPS1) in transgenic soybean inhibited seed development and reduced phytate content. Planta 224: 125–132

Obayashi T, Okamura Y, Ito S, Tadaka S, Aoki Y, Shirota M, Kinoshita K (2014) ATTED-II in 2014: Evaluation of Gene Coexpression in Agriculturally Important Plants. Plant Cell Physiol 55: e6–e6

Obertello M, Shrivastava S, Katari MS, Coruzzi GM (2015) Cross-Species Network Analysis Uncovers Conserved Nitrogen-Regulated Network Modules in Rice. Plant Physiol. doi: 10.1104/pp.114.255877

Oldroyd GE, Engstrom EM, Long SR (2001) Ethylene inhibits the Nod factor signal transduction pathway of Medicago truncatula. Plant Cell 13: 1835–49

Osakabe Y, Watanabe T, Sugano SS, Ueta R, Ishihara R, Shinozaki K, Osakabe K (2016) Optimization of CRISPR/Cas9 genome editing to modify abiotic stress responses in plants. Sci Rep 6: 26685

Ozaki S, Ogata Y, Suda K, Kurabayashi A, Suzuki T, Yamamoto N, Iijima Y, Tsugane T, Fujii T, Konishi C, et al (2010) Coexpression analysis of tomato genes and experimental verification of coordinated expression of genes found in a functionally enriched coexpression module. DNA Res 17: 105–116

Padgette SR, Kolacz KH, Delannay X, Re DB, LaVallee BJ, Tinius CN, Rhodes WK, Otero YI, Barry GF, Eichholtz DA, et al (1995) Development, Identification, and Characterization of a Glyphosate-Tolerant Soybean Line. Crop Sci 35: 1451

Panagiotou OA, Ioannidis JPA (2012) What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. Int J Epidemiol 41: 273–286

Park J-H, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42: 570–575

Pauwels K, De Keersmaecker SCJ, De Schrijver A, du Jardin P, Roosens NHC, Herman P (2015) Next-generation sequencing as a tool for the molecular characterisation and risk assessment of genetically modified plants: Added value or not? Trends Food Sci Technol 45: 319–326

Paz MM, Martinez JC, Kalvig AB, Fonger TM, Wang K (2006) Improved cotyledonary node method using an alternative explant derived from mature seed for efficient Agrobacterium-mediated soybean transformation. Plant Cell Rep 25: 206–13

Potrykus I (2017) The GMO-crop potential for more, and more nutritious food is blocked by unjustified regulation. J Innov Knowl 2: 90–96

Puchta H, Fauser F (2013) Gene targeting in plants: 25 years later. Int J Dev Biol 57: 629–37

Pyott DE, Sheehan E, Molnar A (2016) Engineering of CRISPR/Cas9-mediated potyvirus resistance in transgene-free Arabidopsis plants. Mol Plant Pathol 17: 1276–1288

Qi Y, Zhang Y, Zhang F, Baller J a, Cleland SC, Ryu Y, Starker CG, Voytas DF (2013) Increasing frequencies of site-specific mutagenesis and gene targeting in Arabidopsis by manipulating DNA repair pathways. Genome Res 23: 547–54

Righetti K, Vu JL, Pelletier S, Vu BL, Glaab E, Lalanne D, Pasha A, Patel R V., Provart NJ, Verdier J, et al (2015) Inference of Longevity-Related Genes from a Robust Coexpression Network of Seed Maturation Identifies Regulators Linking Seed Storability to Biotic Defense-Related Pathways. Plant Cell 27: tpc.15.00632

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29: 24–26

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996) Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. Anal Biochem 242: 84–89

Rosa SF, Gatto F, Angers-Loustau A, Petrillo M, Kreysa J, Querci M (2016) Development and applicability of a ready-to-use PCR system for GMO screening. Food Chem 201: 110–119

Ruprecht C, Mendrinna A, Tohge T, Sampathkumar A, Klie S, Fernie AR, Nikoloski Z, Persson S, Mutwil M (2016) FamNet: A framework to identify multiplied modules driving pathway diversification in plants. Plant Physiol 170: pp.15.01281-

Ruprecht C, Mutwil M, Saxe F, Eder M, Nikoloski Z, Persson S (2011) Large-Scale Co-Expression Approach to Dissect Secondary Cell Wall Formation Across Plant Species. Front Plant Sci 2: 1–13

Sabot F, Picault N, El-Baidouri M, Llauro C, Chaparro C, Piegu B, Roulin A, Guiderdoni E, Delabastide M, McCombie R, et al (2011) Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. Plant J 66: 241–246

Sander JD, Cade L, Khayter C, Reyon D, Peterson RT, Joung JK, Yeh J-RJ (2011a) Targeted gene disruption in somatic zebrafish cells using engineered TALENs. Nat Biotechnol 29: 697–698

Sander JD, Dahlborg EJ, Goodwin MJ, Cade L, Zhang F, Cifuentes D, Curtin SJ, Blackburn JS, Thibodeau-Beganny S, Qi Y, et al (2011b) Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). Nat Methods 8: 67–9

Sanford J (1988) The biolistic process. Trends Biotechnol 6: 299–302

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94: 441–448

Sarkar NK, Kim Y-K, Grover A (2014) Coexpression network analysis associated with call of rice seedlings for encountering heat stress. Plant Mol Biol 84: 125–143

Sato S, Newell C, Kolacz K, Tredo L, Finer J, Hinchee M (1993) Stable transformation via particle bombardment in two different soybean regeneration systems. Plant Cell Rep 12–12: 408–413

Schaefer RJ, Briskine R, Springer NM, Myers CL (2014) Discovering functional modules across diverse maize transcriptomes using COB, the co-expression browser. PLoS One. doi: 10.1371/journal.pone.0099193

Schaefer RJ, Michno J-M, Jeffers J, Hoekenga O, Dilkes B, Baxter I, Myers C (2018) Integrating co-expression networks with GWAS to prioritize causal genes in maize. bioRxiv 221655

Schaefer RJ, Michno J-M, Myers CL (2017) Unraveling gene function in agricultural species using gene co-expression networks. Biochim Biophys Acta - Gene Regul Mech 1860: 53–63

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science (80-) 270: 467–470

Schiml S, Fauser F, Puchta H (2014) The CRISPR/Cas system can be used as nuclease for in planta gene targeting and as paired nickases for directed mutagenesis in Arabidopsis resulting in heritable progeny. Plant J 80: 1139–1150

Schmid-Burgk JL, Schmidt T, Kaiser V, Höning K, Hornung V (2013) A ligation-independent cloning technique for high-throughput assembly of transcription activator–like effector genes. Nat Biotechnol 31: 76–81

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU (2005) A gene expression map of Arabidopsis thaliana development. Nat Genet 37: 501–506

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463: 178–83

Schneider GF, Dekker C (2012) DNA sequencing with nanopores. Nat Biotechnol 30: 326–328

Scholtens IMJ, Molenaar B, van Hoof RA, Zaaijer S, Prins TW, Kok EJ (2017) Semiautomated TaqMan PCR screening of GMO labelled samples for (unauthorised) GMOs. Anal Bioanal Chem 409: 3877–3889

Schouten HJ, vande Geest H, Papadimitriou S, Bemer M, Schaart JG, Smulders MJM, Perez GS, Schijlen E (2017) Re-sequencing transgenic plants revealed rearrangements at T-DNA inserts, and integration of a short T-DNA fragment, but no increase of small mutations elsewhere. Plant Cell Rep 36: 493–504

Sekhon RS, Briskine R, Hirsch CN, Myers CL, Springer NM, Buell CR, de Leon N, Kaeppler SM (2013) Maize Gene Atlas Developed by RNA Sequencing and Comparative Evaluation of Transcriptomes Based on RNA Sequencing and Microarrays. PLoS One. doi: 10.1371/journal.pone.0061005

Sekhon RS, Lin H, Childs KL, Hansey CN, Buell CR, de Leon N, Kaeppler SM (2011) Genome-wide atlas of transcription during maize development. Plant J 66: 553–63

Shah DM, Horsch RB, Klee HJ, Kishore GM, Winter JA, Tumer NE, Hiornaka CM, Sanders PR, Gasser CS, Aykent S, et al (1986) Engineering Herbicide Tolerance in Transgenic Plants. Science (80-) 233: 478–481

Shah DM, Horsch RB, Klee HJ, Kishore GM, Winter JA, Tumer NE, Hironoka CM, Sanders PR, Gasser CS, Aylkent S, et al (1985) A Simple and General Method for Transferring Genes into Plants. Science (80-) 227: 1229–1231

Sham PC, Purcell SM (2014) Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet 15: 335–346

Shan Q, Wang Y, Li J, Zhang Y, Chen K, Liang Z, Zhang K, Liu J, Xi JJ, Qiu J-L, et al (2013a) Targeted genome modification of crop plants using a CRISPR-Cas system. Nat Biotechnol 31: 686–688

Shan Q, Wang Y, Li J, Zhang Y, Chen K, Liang Z, Zhang K, Liu J, Xi JJ, Qiu J-L, et al (2013b) Targeted genome modification of crop plants using a CRISPR-Cas system. Nat Biotechnol 31: 686–688

Simmonds DH, Donaldson PA (2000) Genotype screening for proliferative embryogenesis and biolistic transformation of short-season soybean genotypes. Plant Cell Rep 19: 485–490

Smil V (1999) Nitrogen in crop production: An account of global flows. Global Biogeochem Cycles 13: 647–662

Song Q, Hyten DL, Jia G, Quigley C V, Fickus EW, Nelson RL, Cregan PB (2013) Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. PLoS One 8: e54985

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

Stanton-Geddes J, Paape T, Epstein B, Briskine R, Yoder J, Mudge J, Bharti AK, Farmer AD, Zhou P, Denny R, et al (2013) Candidate Genes and Genetic Architecture of Symbiotic and Agronomic Traits Revealed by Whole-Genome, Sequence-Based Association Genetics in Medicago truncatula. PLoS One 8: e65688

Steeves RM, Todd TC, Essig JS, Trick HN (2006) Transgenic soybeans expressing siRNAs specific to a major sperm protein gene suppress Heterodera glycines reproduction. Funct Plant Biol 33: 991

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci 100: 9440–9445

Strauss SH, Sax JK (2016) Ending event-based regulation of GMO crops. Nat Biotechnol 34: 474–477

Stuart JM (2003) A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science (80-) 302: 249–255

Sun X, Hu Z, Chen R, Jiang Q, Song G, Zhang H, Xi Y (2015) Targeted mutagenesis in soybean using the CRISPR-Cas9 system. Sci Rep 5: 10342

Sun Z, Bhagwate A, Prodduturi N, Yang P, Kocher J-PA (2016) Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. Brief Bioinform 18: bbw069

Svitashev SK, Somers DA (2001) Genomic interspersions determine the size and complexity of transgene loci in transgenic plants produced by microprojectile bombardment. Genome 44: 691–7

Swanson-Wagner R, Briskine R, Schaefer R, Hufford MB, Ross-Ibarra J, Myers CL, Tiffin P, Springer NM (2012) Reshaping of the maize transcriptome by domestication. Proc Natl Acad Sci 109: 11878–11883

Takagi K, Nishizawa K, Hirose A, Kita A, Ishimoto M (2011) Manipulation of saponin biosynthesis by RNA interference-mediated silencing of β-amyrin synthase gene expression in soybean. Plant Cell Rep 30: 1835–1846

Tang F, Yang S, Liu J, Zhu H (2016) Rj4 , a Gene Controlling Nodulation Specificity in Soybeans, Encodes a Thaumatin-Like Protein But Not the One Previously Reported. Plant Physiol 170: 26–32

Taylor A, Sajan S (2005) Testing for Genetically Modified Foods Using PCR. J Chem Educ 82: 597

Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13: 36–46

Turcatti G, Romieu A, Fedurco M, Tairi A-P (2008) A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis †. Nucleic Acids Res 36: e25–e25

Twyman RM, Christou P (2004) Plant Transformation Technology: Particle Bombardment. Handb Plant Biotechnol. doi: 10.1002/0470869143.kc015

Tzfira T (2003) Site-Specific Integration of Agrobacterium tumefaciens T-DNA via Double-Stranded Intermediates. Plant Physiol 133: 1011–1023

Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ 32: 1633–51

VandenBosch KA (2003) Summaries of Legume Genomics Projects from around the Globe. Community Resources for Crops and Models. PLANT Physiol 131: 840–865

Vaughn T, Cavato T, Brar G, Coombe T, DeGooyer T, Ford S, Groth M, Howe A, Johnson S, Kolacz K, et al (2005) A Method of Controlling Corn Rootworm Feeding Using a Protein Expressed in Transgenic Maize. Crop Sci 45: 931

Veena V, Taylor CG (2007) Agrobacterium rhizogenes: recent developments and promising applications. Vitr Cell Dev Biol - Plant 43: 383–403

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five Years of GWAS Discovery. Am J Hum Genet 90: 7–24

Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ, et al (2013) The genome sequence of the colonial chordate, Botryllus schlosseri. Elife 2: e00569

Wagner N, Mroczka A, Roberts PD, Schreckengost W, Voelker T (2011) RNAi trigger fragment truncation attenuates soybean FAD2-1 transcript suppression and yields intermediate oil phenotypes. Plant Biotechnol J 9: 723–728

Wahler D, Schauser L, Bendiek J, Grohmann L (2013) Next-Generation Sequencing as a Tool for Detailed Molecular Characterisation of Genomic Insertions and Flanking Regions in Genetically Modified Plants: a Pilot Study Using a Rice Event Unauthorised in the EU. Food Anal Methods 6: 1718–1727

Wang G, Xu Y (2008) Hypocotyl-based Agrobacterium-mediated transformation of soybean (Glycine max) and application for RNA interference. Plant Cell Rep 27: 1177–1184

Wang Y, Cheng X, Shan Q, Zhang Y, Liu J, Gao C, Qiu J-L (2014) Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. Nat Biotechnol 32: 947–951

Wang Y, Yang Q, Wang Z (2015) The evolution of nanopore sequencing. Front Genet 5:

449

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

Wasson AP (2006) Silencing the Flavonoid Pathway in Medicago truncatula Inhibits Root Nodule Formation and Prevents Auxin Transport Regulation by Rhizobia. PLANT CELL ONLINE 18: 1617–1629

Weber N, Halpin C, Hannah LC, Jez JM, Kough J, Parrott W (2012) Editor's Choice: Crop Genome Plasticity and Its Relevance to Food and Feed Safety of Genetically Engineered Breeding Stacks. PLANT Physiol 160: 1842–1853

Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang X-J, Buck D, Au KF (2017) Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research 6: 100

Wickham H (2011) ggplot2. Wiley Interdiscip Rev Comput Stat 3: 180–185

Wickham H (2006) An introduction to ggplot : An implementation of the grammar of graphics in R. 1–8

Willems S, Fraiture M-A, Deforce D, De Keersmaecker SCJ, De Loose M, Ruttink T, Herman P, Van Nieuwerburgh F, Roosens N (2016) Statistical framework for detection of genetically modified organisms based on Next Generation Sequencing. Food Chem 192: 788–798

Wolfe D, Dudek S, Ritchie MD, Pendergrass SA (2013) Visualizing genomic information across chromosomes with PhenoGram. BioData Min 6: 18

Xing H-L, Dong L, Wang Z-P, Zhang H-Y, Han C-Y, Liu B, Wang X-C, Chen Q-J (2014) A CRISPR/Cas9 toolkit for multiplex genome editing in plants. BMC Plant Biol 14: 327

Yan W, Chen D, Kaufmann K (2016) Efficient multiplex mutagenesis by RNA-guided Cas9 and its use in the characterization of regulatory elements in the AGAMOUS gene. Plant Methods 12: 23

Yin X, Biswal AK, Dionora J, Perdigon KM, Balahadia CP, Mazumdar S, Chater C, Lin H-C, Coe RA, Kretzschmar T, et al (2017) CRISPR-Cas9 and CRISPR-Cpf1 mediated targeting of a stomatal developmental gene EPFL9 in rice. Plant Cell Rep 36: 745–757

Young ND, Udvardi M (2009) Translating Medicago truncatula genomics to crop legumes. Curr Opin Plant Biol 12: 193–201

Zhang D, Wang Z, Wang N, Gao Y, Liu Y, Wu Y, Bai Y, Zhang Z, Lin X, Dong Y, et al (2014) Tissue Culture-Induced Heritable Genomic Variation in Rice, and Their Phenotypic Implications. PLoS One 9: e96879

Zhang X, Sato S, Ye X, Dorrance AE, Morris TJ, Clemente TE, Qu F (2011) Robust RNAi-Based Resistance to Mixed Infection of Three Viruses in Soybean Plants

Expressing Separate Short Hairpins from a Single Transgene. Phytopathology 101: 1264–1269

Zhang Y, Liang Z, Zong Y, Wang Y, Liu J, Chen K, Qiu J-L, Gao C (2016a) Efficient and transgene-free genome editing in wheat through transient expression of CRISPR/Cas9 DNA or RNA. Nat Commun 7: 12617

Zhang Z, Mao Y, Ha S, Liu W, Botella JR, Zhu J-K (2016b) A multiplex CRISPR/Cas9 platform for fast and efficient editing of multiple genes in Arabidopsis. Plant Cell Rep 35: 1519–1533

Zheng Z-L, Zhao Y (2013) Transcriptome comparison and gene coexpression network analysis provide a systems view of citrus response to “Candidatus Liberibacter asiaticus” infection. BMC Genomics 14: 27

Zhu T, Budworth P, Chen W, Provart N, Chang H-S, Guimil S, Su W, Estes B, Zou G, Wang X (2002) Transcriptional control of nutrient partitioning during rice grain filling. Plant Biotechnol J 1: 59–70

Appendix

Appendix 1:

Chapter 2 supplemental figures

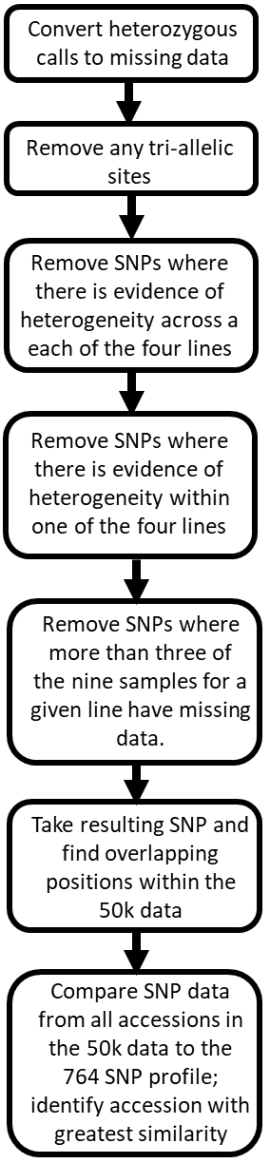


Figure S1. Pipeline to identify the background genotype of 764. The steps used to narrow down consensus calls between each of the four lines where only one of the four exhibited an alternate allele compared to the 'Williams 82' reference genome. The intersect between the resulting SNP positions and positions within the 50k data were then used to calculate the percentage of SNPs that match the 764 series.

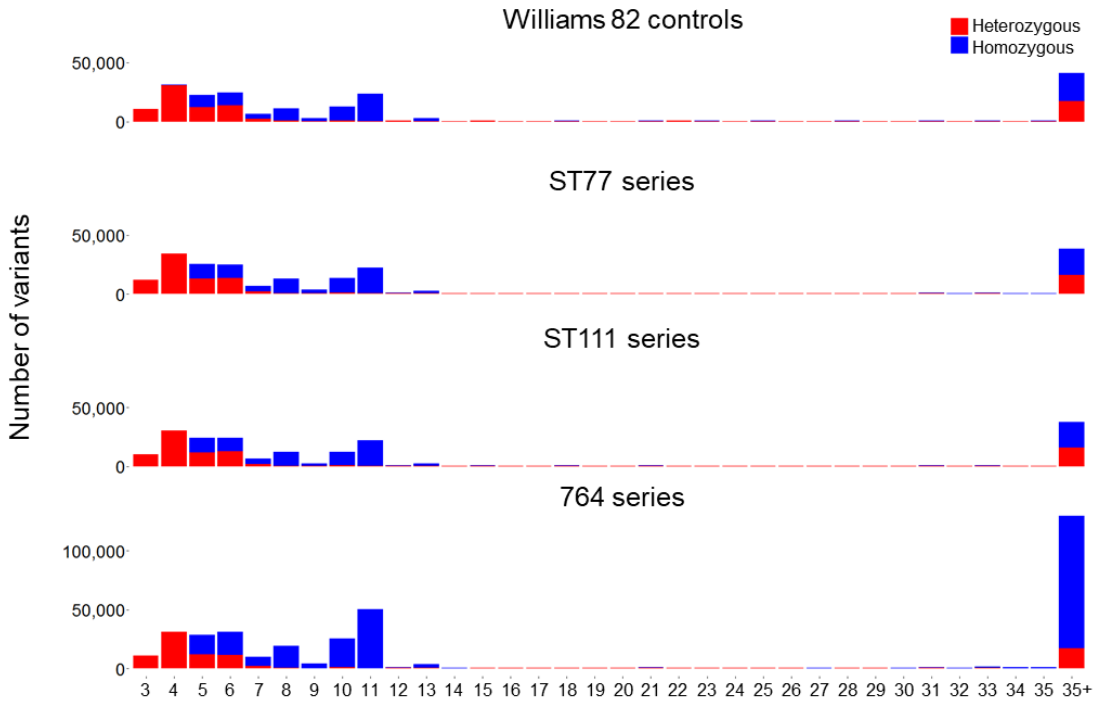


Figure S2. Quality scores for all polymorphic variants (SNPs and indels) called in the Lambirth et al. (2016) study. The polymorphic calls shown here were made between each sample and the reference genome 'Williams 82', without consideration for the uniqueness of the call among series or reproducibility among different plants within the series. Homozygous calls are shown in blue and heterozygous calls are shown in red. Each bar sums the number of polymorphisms across the nine plants that were called at each read depth (e.g., we are showing the ~211,448 total variants called in series ST77 across the nine plants; ST77 averaged 23,494 variants per plant). Note the larger peak in the 35+ category for the 764 series; many of these (mostly homozygous) calls likely represent standing variants between lines 'Thorne' and 'Williams 82'. The 35+ peaks in the other three groups (ST77, ST111, and 'Williams 82' controls) may derive from various factors, most obviously the clusters of variants that are found within heterogeneous regions of different sub-lines of 'Williams 82'.

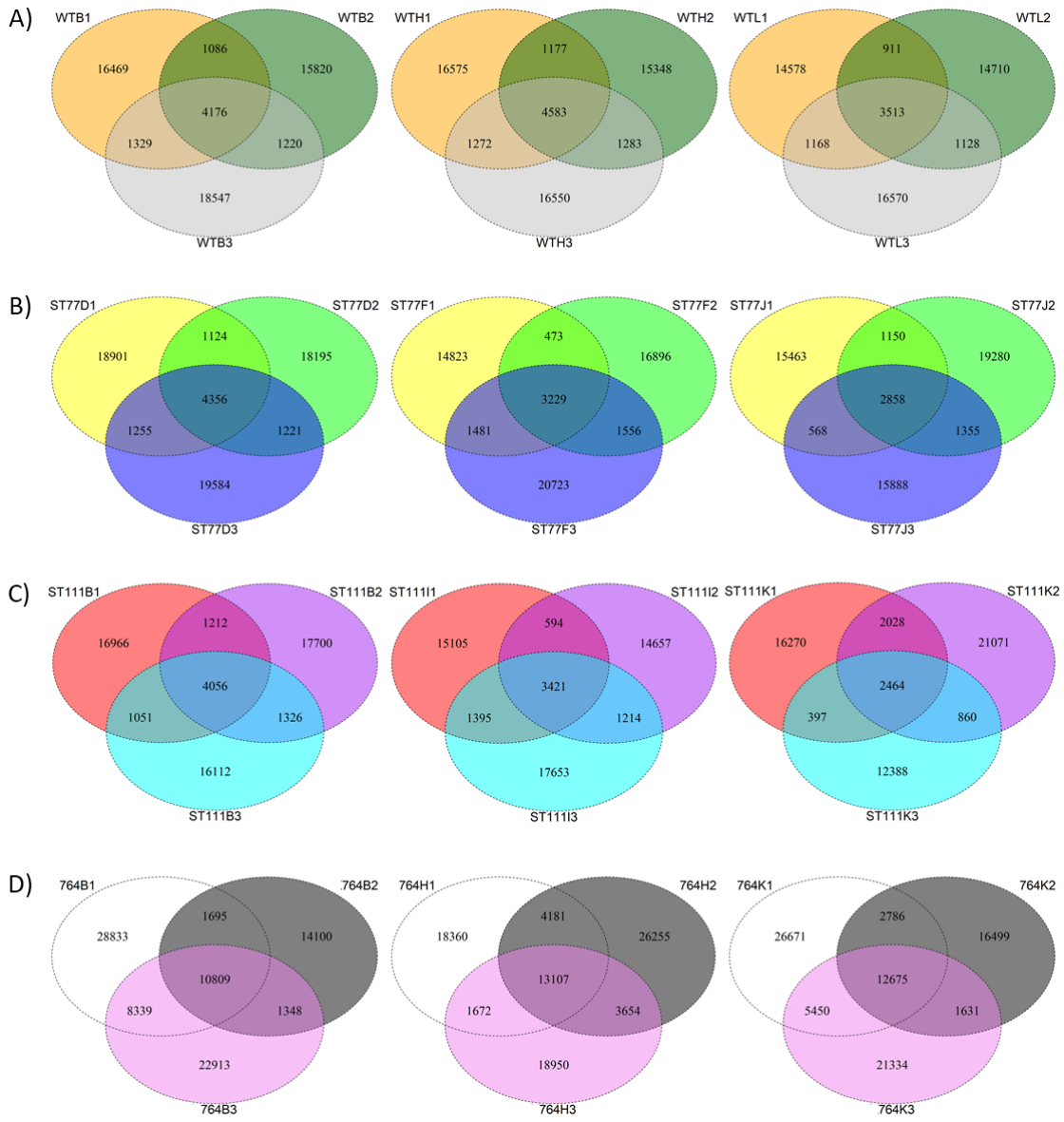


Figure S3. Number of overlapping polymorphisms in the Lambirth et al. (2016) study within each of the 12 sibling families studied. A) Venn diagram showing of the number of sequence variants alternate to the reference genome that overlapped between three different groups of three siblings each of ‘Williams 82’, the wild type (WT) control in this study. Heterozygous and homozygous alternate calls are not differentiated in this analysis. B-D) Similar Venn diagrams of the number of polymorphism that overlapped between the F7:8 siblings in each of the ST77D (B), ST111 (C), and 764 (D) transgenic families.

Appendix 2:

Chapter 3 supplemental tables

WPT	Type	Target	Target #2	gRNA	gRNA2	Backbone	Promoter	Plant selectable marker	gRNA Type	Cas9	Agro Strain	Transgene was heritable	Targeted mutation was heritable
WPT536-2	Cas9	Glyma16g12160	NA	GACATTCAAC GCAAAGTACT GACATTCAAC	NA	MDC123	gmubiXL	BAR	Arabidopsis U6	Gmax codon optimized	18r12	+	+
WPT536-2-13-15	Cas9	Glyma16g12160	NA	GCAAAGTACT GACATTCAAC	NA	MDC123	gmubiXL	BAR	Arabidopsis U6	Gmax codon optimized	18r12	-	-
WPT536-2-13-16	Cas9	Glyma16g12160	NA	GCAAAGTACT	NA	MDC123	gmubiXL	BAR	Arabidopsis U6	Gmax codon optimized	18r12	-	+
WPT553-6	Cas9	Glyma.18g041100	NA	GAAGTTGGGG AAGAGACACA	NA	MDC32	35S	Hygromycin	Arabidopsis U6	Gmax codon optimized	18r12	-	+
WPT553-6-8	Cas9	Glyma.18g041100	NA	GAAGTTGGGG AAGAGACACA	NA	MDC32	35S	Hygromycin	Arabidopsis U6	Gmax codon optimized	18r12	-	+
WPT553-6-11	Cas9	Glyma.18g041100	NA	GAAGTTGGGG AAGAGACACA	NA CCAGTT GGTGAG	MDC32	35S	Hygromycin	Arabidopsis U6	Gmax codon optimized	18r12	-	+
WPT608-1	Cas9	Glyma.16G209100, Glyma.09G159900	Glyma.09 G159900	TGCCAAGAAT AAGAAACAAG	TGAAAC CA CCAGTT GGTGAG	218 gmubi		BAR	Arabidopsis U6/ At7	Jen Sheen Arabidopdis optimized Cas9	18r12	-	-
WPT608-3	Cas9	Glyma.16G209100	Glyma.09 G159900	TGCCAAGAAT AAGAAACAAG	TGAAAC CA	218 gmubi		BAR	Arabidopsis U6/ At7	Jen Sheen Arabidopdis optimized Cas9	18r12	+	+

Table S1. Whole-plant transformation construct metadata
Chapter 3 supplemental Figures

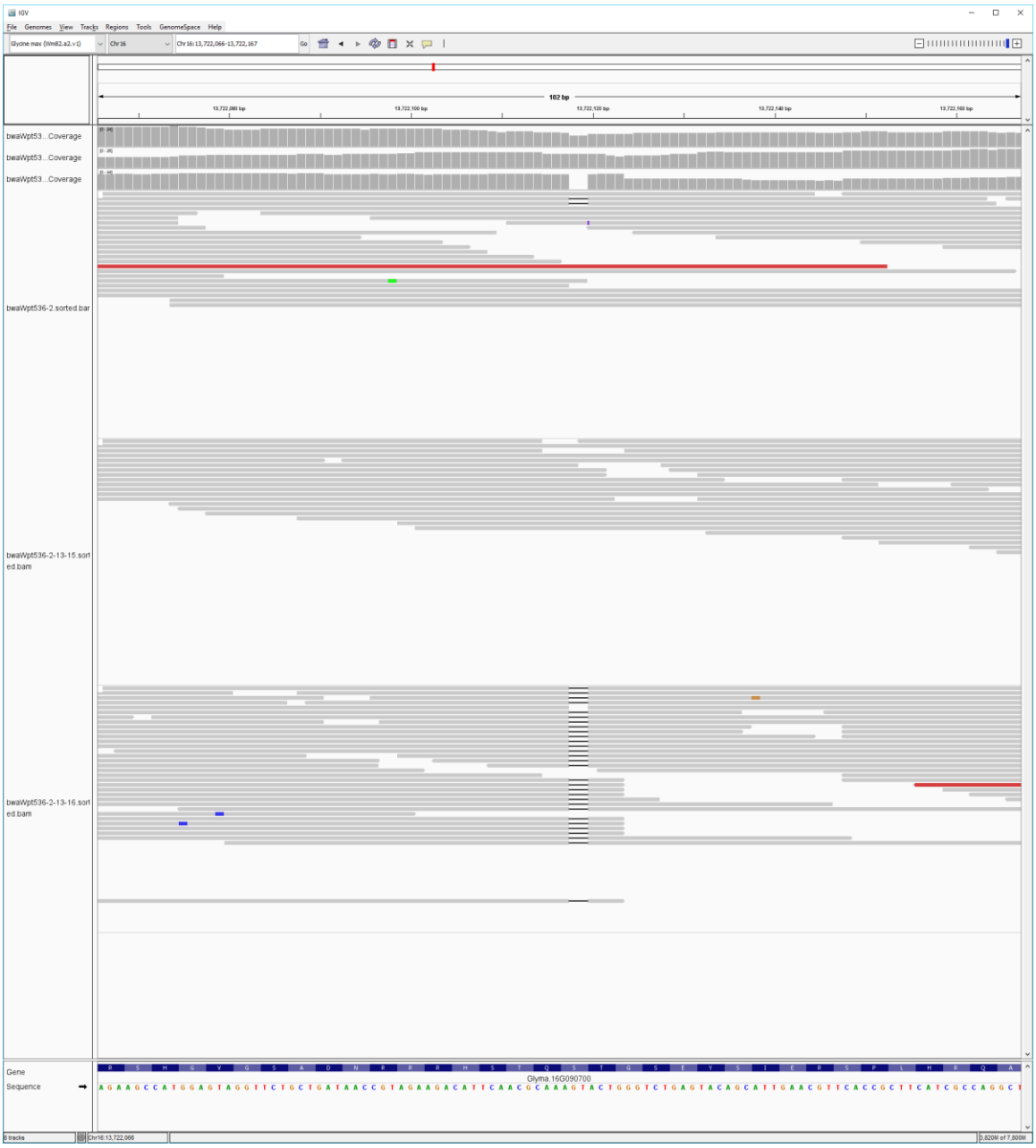


Figure S1. Rin4b gRNA target site. IGV screenshot of WGS at the gRNA target site for the Rin4b T0 and two of its T2 progeny.

Figure S2. Rin4b transgene insertion event. IGV screenshot of the transgene insertion event using WGS. Red bar at the top of the figure represents where the transgene inserted itself into the genome.

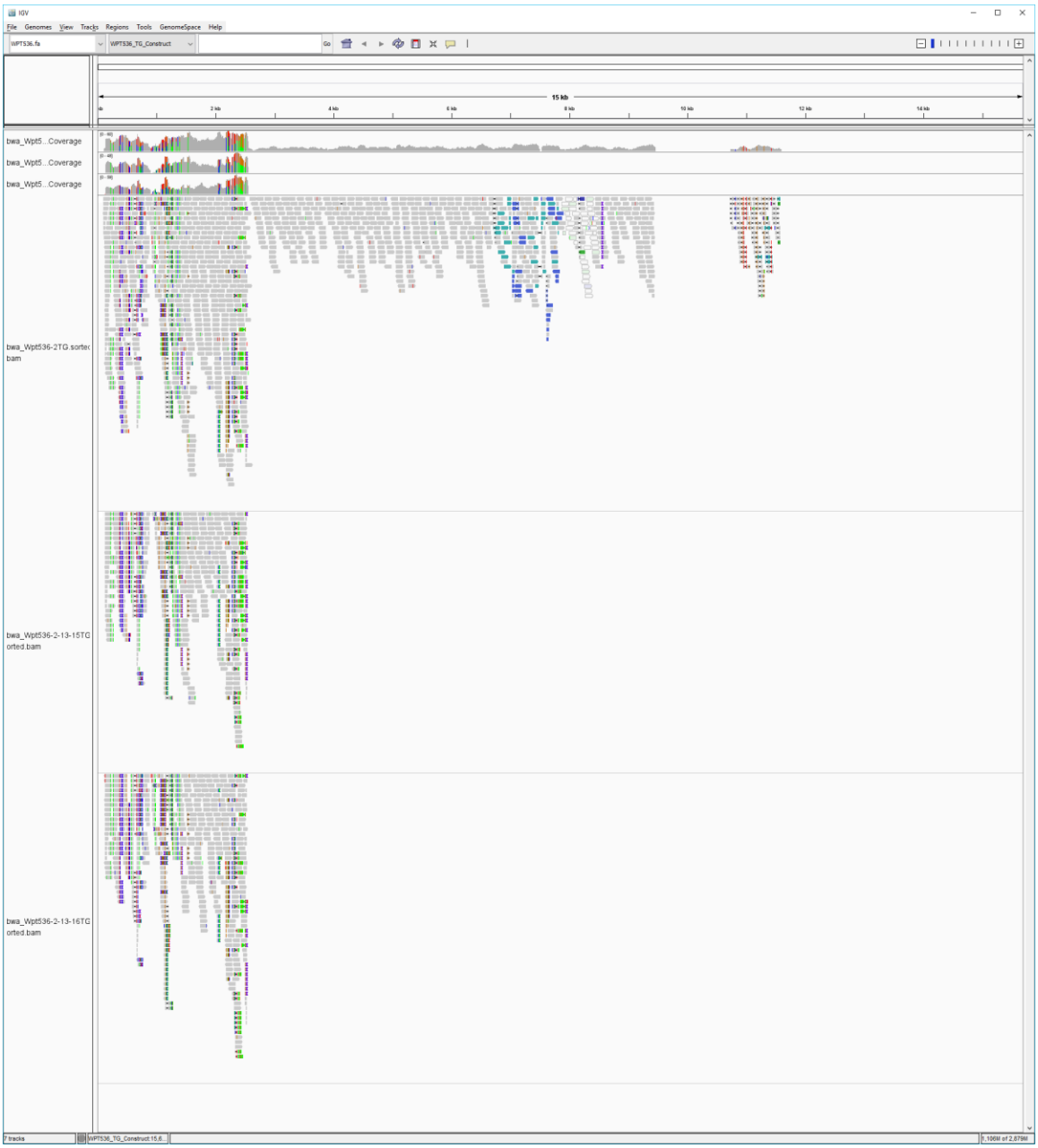


Figure S3. Rin4b transgene mapping coverage. IGV screenshot of WGS read mapped directly to the transgene sequence.

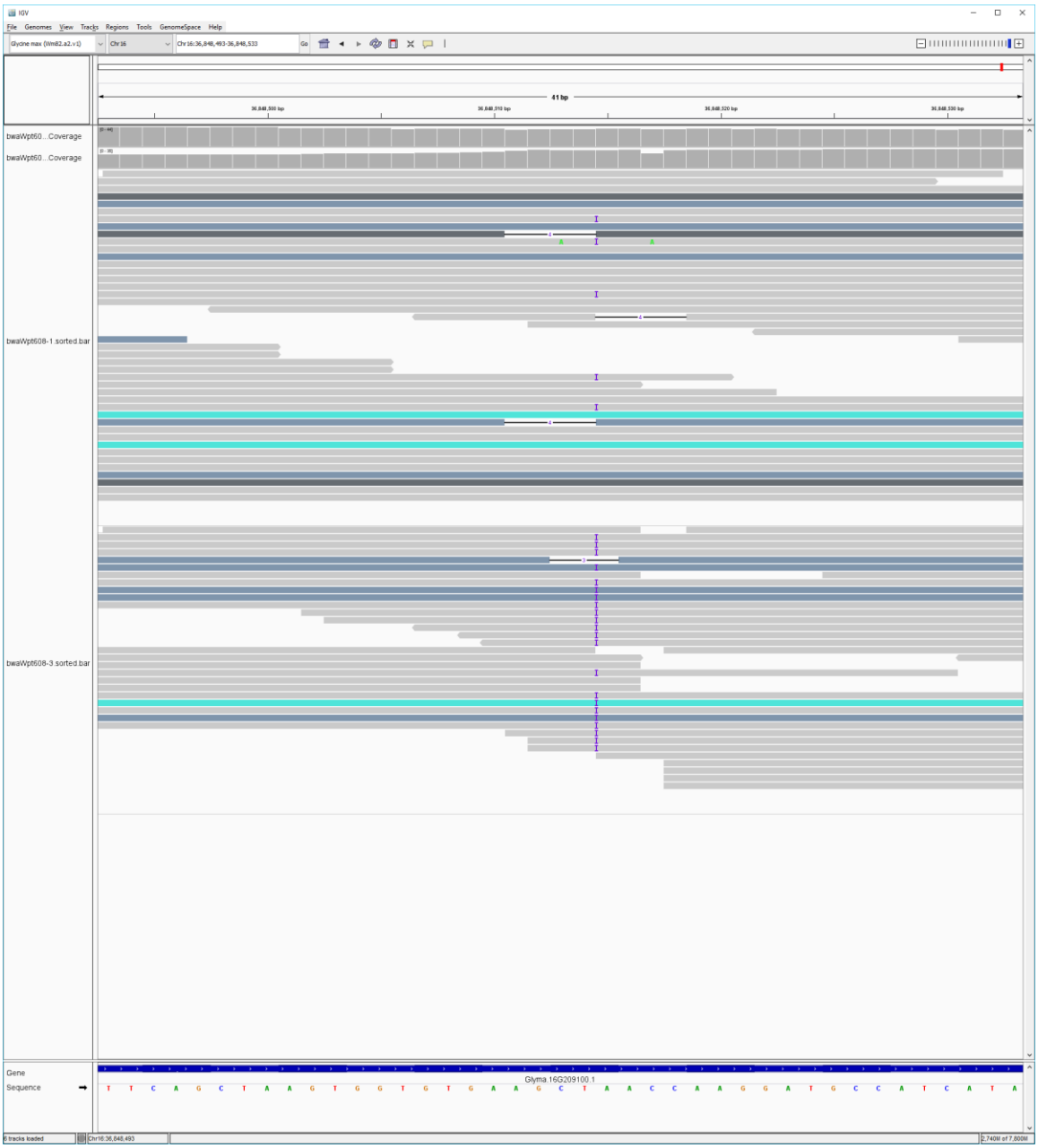


Figure S4. IGV screenshot of Glyma.16G209100 gRNA target site and transgene insertion on chromosome 16.

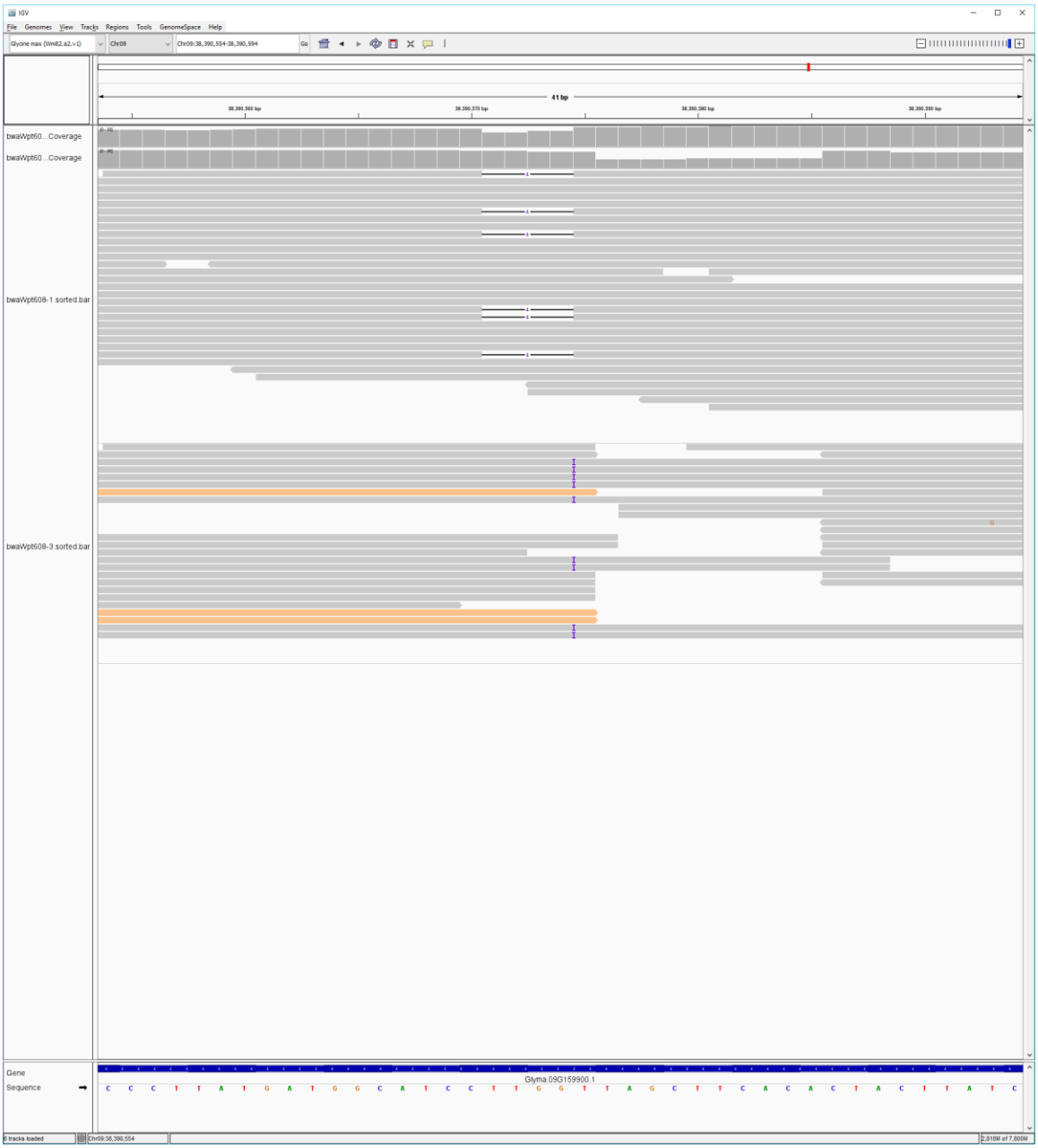


Figure S5. IGV screenshot of the Glyma.16G209100 paralog gRNA target site and transgene insertion on chromosome 9.

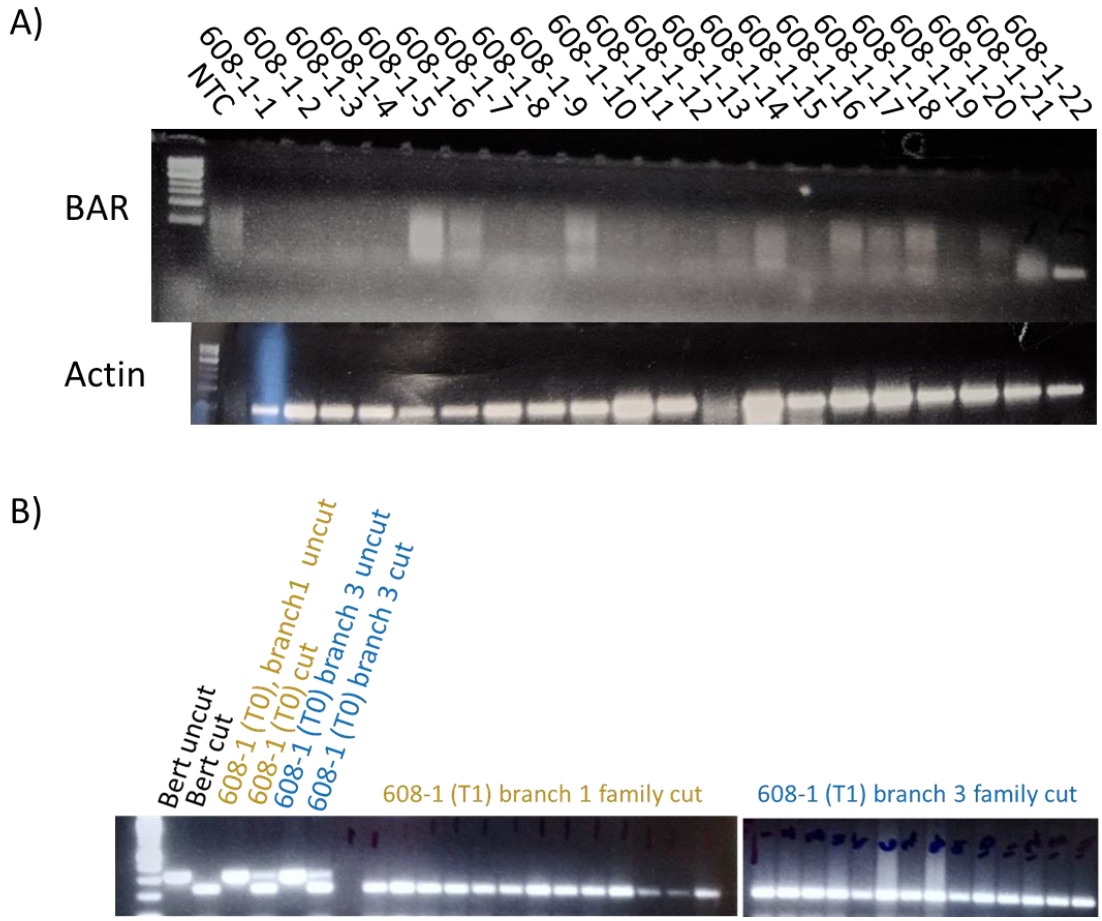
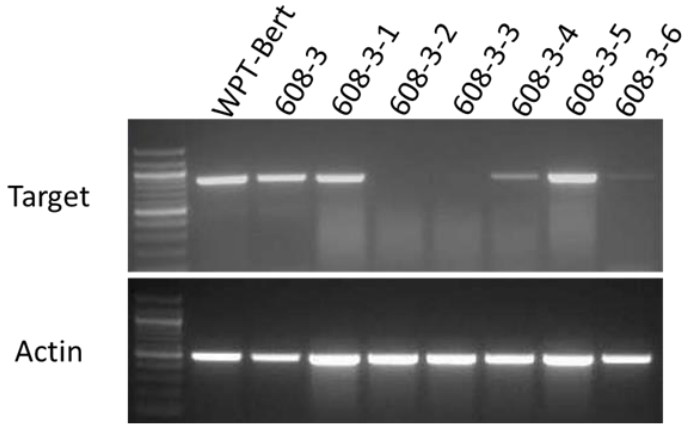


Figure S6. PCR assay for transgene presence in 608-1 offspring. A) PCR assay to amplify a portion of the BAR gene belonging to the transgenic cassette as well as an Actin control. B) 608-1 T0 and T1 family CAPS assay using Sty1 to test for the presence of mutations. The T0 parent and two different branches from each T1 progeny plant were tested.

A)



B)

	Subclones sequenced	allele
608-3 (T0)	12	+A
608-3-1 (T1)	11	+A
608-3-5(T1)	13	+A

C)

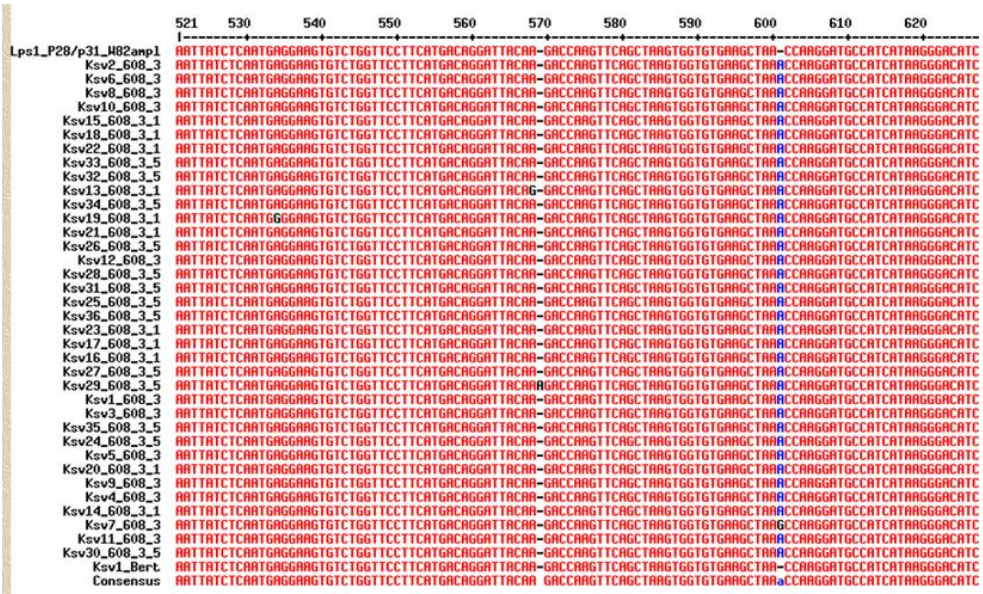


Figure S7. WPT608-1 amplification of gRNA target site on chromosome 16. A) PCR of markers flanking the guide RNA target site for gene target Glyma.16G209100. B) The proportions of subclones sequenced from PCR amplification as well as the detected mutations. C) Sanger sequencing results of the subclones for WPT608-3, WPT608-3-1 and WPT608-5.

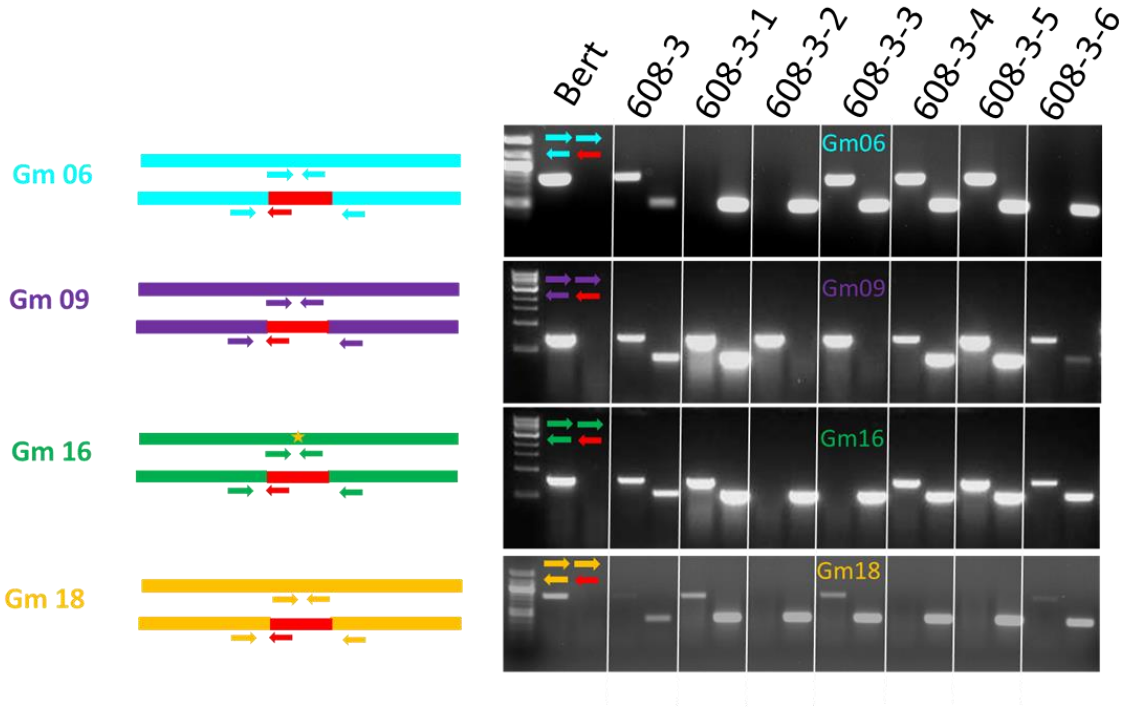


Figure S8. PCR assay to for detection of the transgene in all offspring for 608-3. PCR assay to amplify transgene insertion events identified from next-generation sequencing. Two set of primers were used for each event; one pair flanking either side of the transgene insertion event, and the other pair using one primer flanking the transgene insertion event and the other within the transgene insertion event (red).

Figure S9. GS1 mutations at the gRNA target site mutations induced by CRISPR/Cas9. IGV screenshot of WGS at the gRNA target site for GS1.

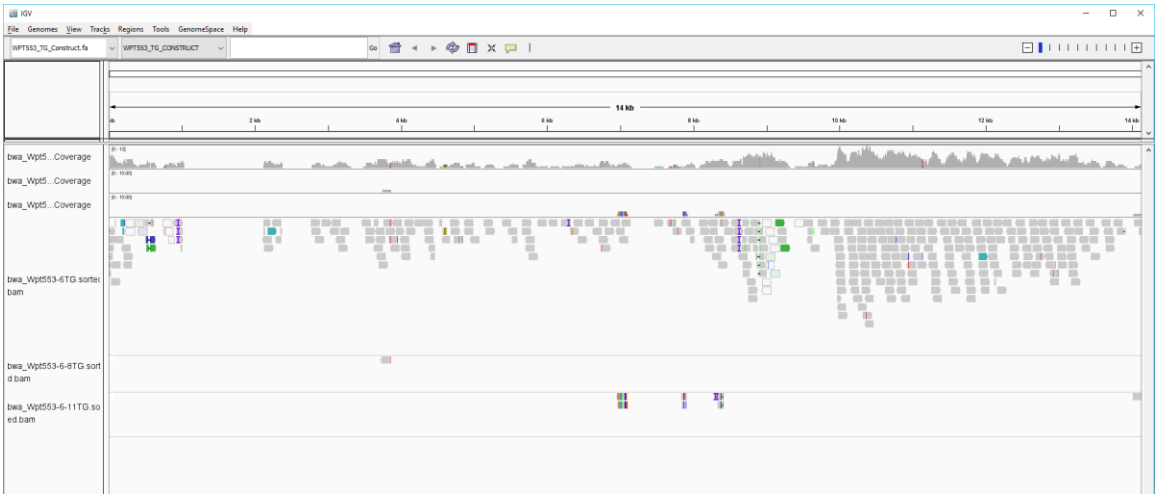


Figure S10. GS1 transgene mapping coverage. IGV screenshot of WGS read mapped directly to the transgene sequence.

Appendix 3:

Chapter 4 supplemental tables

GWAS Trait	Description	Number of SNP's	SNP's included after collapse (window size) 10kb	20kb	50kb	Min p-value
Height	Plant height	197	139	133	127	3.00E-05
Total_Nod	Total number of nodules	163	124	122	119	3.00E-05
Nod_A	Total number of nodules in the top 5 cm of roots	523	294	275	255	9.96E-06
Nod_B	Total number of nodules below the top 5 cm of roots	232	185	178	165	3.00E-05
Flowering Date	Flowering date	550	150	120	100	6.94E-06
OccupancyA	Strain occupancy in the top 5 cm of roots	292	230	226	209	3.00E-05
OccupancyB	Strain occupancy below the top 5 cm of roots	27	17	17	14	9.61E-05

Table S1. GWAS trait information and the number of SNP's used for analysis. "Collapse" refers to SNP's removed due to overlapping windows between sets of SNP's

Sample ID	Tissue	M. truncatula accession	Sinorhizobium species and strain	Nitrogen	D7 Index	Barcode	D5 Index	Barcode
N128	Nodule	HM056	S. meliloti (KH46c)	0	D701	ATTACTCG	D501	TATAGCCT
N86	Nodule	HM056	S. meliloti (KH46c)	0	D702	TCCGGAGA	D501	TATAGCCT
N73	Nodule	HM056	S. medicae (WSM419)	0	D704	GAGATTCC	D501	TATAGCCT
N137	Nodule	HM056	S. medicae (WSM419)	0	D705	ATTCAGAA	D501	TATAGCCT
N48	Nodule	HM056	S. medicae (WSM419)	0	D706	GAATTCGT	D501	TATAGCCT
N88	Nodule	HM101	Both	0	D707	CTGAAGCT	D501	TATAGCCT
N103	Nodule	HM101	Both	0	D708	TAATGCGC	D501	TATAGCCT
N25	Nodule	HM101	Both	0	D709	CGGCTATG	D501	TATAGCCT
N9	Nodule	HM101	S. meliloti (KH46c)	0	D710	TCCGCGAA	D501	TATAGCCT
N121	Nodule	HM101	S. meliloti (KH46c)	0	D711	TCTCGCGC	D501	TATAGCCT
N39	Nodule	HM101	S. meliloti (KH46c)	1	D712	AGCGATAG	D501	TATAGCCT
N75	Nodule	HM101	S. meliloti (KH46c)	1	D701	ATTACTCG	D502	ATAGAGGC
N146	Nodule	HM101	S. meliloti (KH46c)	1	D702	TCCGGAGA	D502	ATAGAGGC
N83	Nodule	HM101	S. medicae (WSM419)	0	D704	GAGATTCC	D502	ATAGAGGC
N56	Nodule	HM101	S. medicae (WSM419)	0	D705	ATTCAGAA	D502	ATAGAGGC
N14	Nodule	HM101	S. medicae (WSM419)	0	D706	GAATTCGT	D502	ATAGAGGC
N64	Nodule	HM101	S. medicae (WSM419)	0	D707	CTGAAGCT	D502	ATAGAGGC
N122	Nodule	HM101	S. medicae (WSM419)	1	D708	TAATGCGC	D502	ATAGAGGC
N46	Nodule	HM101	S. medicae (WSM419)	1	D709	CGGCTATG	D502	ATAGAGGC
N41	Nodule	HM101	S. medicae (WSM419)	1	D710	TCCGCGAA	D502	ATAGAGGC
N107	Nodule	HM101	S. medicae (WSM419)	1	D711	TCTCGCGC	D502	ATAGAGGC
N62	Nodule	HM340	Both	0	D712	AGCGATAG	D502	ATAGAGGC
N21	Nodule	HM340	Both	0	D701	ATTACTCG	D503	CCTATCCT
N160	Nodule	HM340	S. meliloti (KH46c)	0	D702	TCCGGAGA	D503	CCTATCCT
N115	Nodule	HM340	S. meliloti (KH46c)	1	D704	GAGATTCC	D503	CCTATCCT
N131	Nodule	HM340	S. meliloti (KH46c)	1	D705	ATTCAGAA	D503	CCTATCCT
N143	Nodule	HM340	S. meliloti (KH46c)	1	D706	GAATTCGT	D503	CCTATCCT
N80	Nodule	HM340	S. medicae (WSM419)	0	D707	CTGAAGCT	D503	CCTATCCT
N92	Nodule	HM340	S. medicae (WSM419)	0	D708	TAATGCGC	D503	CCTATCCT
N26	Nodule	HM340	S. medicae (WSM419)	0	D709	CGGCTATG	D503	CCTATCCT
N8	Nodule	HM340	S. medicae (WSM419)	0	D710	TCCGCGAA	D503	CCTATCCT

N42	Nodule	HM340	S. medicae (WSM419)	1	D711	TCTCGCGC	D503	CCTATCCT
N47	Nodule	HM340	S. medicae (WSM419)	1	D712	AGCGATAG	D503	CCTATCCT
N111	Nodule	HM340	S. medicae (WSM419)	1	D701	ATTACTCG	D504	GGCTCTGA
N120	Nodule	HM056	S. medicae (WSM419)	0	D704	GAGATTCC	D502	ATAGAGGC
N40	Nodule	HM101	S. meliloti (KH46c)	1	D705	ATTCAGAA	D502	ATAGAGGC
N11	Nodule	HM340	S. meliloti (KH46c)	0	D706	GAATTCGT	D502	ATAGAGGC
R51	Root	HM101	S. meliloti (KH46c)	0	D702	TCCGGAGA	D504	GGCTCTGA
R5	Root	HM101	S. meliloti (KH46c)	0	D705	ATTCAGAA	D504	GGCTCTGA
R125	Root	HM101	None	1	D706	GAATTCGT	D504	GGCTCTGA
R171	Root	HM101	None	1	D707	CTGAAGCT	D504	GGCTCTGA
R142	Root	HM101	None	1	D708	TAATGCGC	D504	GGCTCTGA
R83	Root	HM101	S. medicae (WSM419)	0	D709	CGGCTATG	D504	GGCTCTGA
R56	Root	HM101	S. medicae (WSM419)	0	D710	TCCGCGAA	D504	GGCTCTGA
R14	Root	HM101	S. medicae (WSM419)	0	D711	TCTCGCGC	D504	GGCTCTGA
R64	Root	HM101	S. medicae (WSM419)	0	D712	AGCGATAG	D504	GGCTCTGA
R160	Root	HM340	S. meliloti (KH46c)	0	D701	ATTACTCG	D505	AGGCGAAG
R11	Root	HM340	S. meliloti (KH46c)	0	D702	TCCGGAGA	D505	AGGCGAAG
R44	Root	HM340	None	1	D704	GAGATTCC	D505	AGGCGAAG
R13	Root	HM340	None	1	D705	ATTCAGAA	D505	AGGCGAAG
R33	Root	HM340	None	1	D706	GAATTCGT	D505	AGGCGAAG
R80	Root	HM340	S. medicae (WSM419)	0	D707	CTGAAGCT	D505	AGGCGAAG
R92	Root	HM340	S. medicae (WSM419)	0	D708	TAATGCGC	D505	AGGCGAAG
R26	Root	HM340	S. medicae (WSM419)	0	D709	CGGCTATG	D505	AGGCGAAG
R8	Root	HM340	S. medicae (WSM419)	0	D710	TCCGCGAA	D505	AGGCGAAG
R9	Root	HM101	S. meliloti (KH46c)	0	D707	CTGAAGCT	D502	ATAGAGGC
R34	Root	HM340	S. meliloti (KH46c)	0	D708	TAATGCGC	D502	ATAGAGGC
L70	Leaf	HM056	S. meliloti (KH46c)	0	D712	AGCGATAG	D505	AGGCGAAG
L128	Leaf	HM056	S. meliloti (KH46c)	0	D701	ATTACTCG	D506	TAATCTTA
L152	Leaf	HM056	S. meliloti (KH46c)	0	D702	TCCGGAGA	D506	TAATCTTA
L86	Leaf	HM056	S. meliloti (KH46c)	0	D709	CGGCTATG	D502	ATAGAGGC
L20	Leaf	HM056	None	1	D704	GAGATTCC	D506	TAATCTTA
L59	Leaf	HM056	None	1	D705	ATTCAGAA	D506	TAATCTTA
L60	Leaf	HM056	None	1	D706	GAATTCGT	D506	TAATCTTA
L61	Leaf	HM056	None	1	D707	CTGAAGCT	D506	TAATCTTA
L120	Leaf	HM056	S. medicae (WSM419)	0	D708	TAATGCGC	D506	TAATCTTA

L73	Leaf	HM056	S. medicae (WSM419)	0	D709	CGGCTATG	D506	TAATCTTA
L137	Leaf	HM056	S. medicae (WSM419)	0	D710	TCCGCGAA	D506	TAATCTTA
L48	Leaf	HM056	S. medicae (WSM419)	0	D711	TCTCGCGC	D506	TAATCTTA
L88	Leaf	HM101	Both	0	D712	AGCGATAG	D506	TAATCTTA
L103	Leaf	HM101	Both	0	D701	ATTACTCG	D507	CAGGACGT
L25	Leaf	HM101	Both	0	D702	TCCGGAGA	D507	CAGGACGT
L158	Leaf	HM101	Both	0	D710	TCCGCGAA	D502	ATAGAGGC
L51	Leaf	HM101	S. meliloti (KH46c)	0	D704	GAGATTCC	D507	CAGGACGT
L9	Leaf	HM101	S. meliloti (KH46c)	0	D705	ATTCAGAA	D507	CAGGACGT
L5	Leaf	HM101	S. meliloti (KH46c)	0	D707	CTGAAGCT	D507	CAGGACGT
L39	Leaf	HM101	S. meliloti (KH46c)	1	D708	TAATGCGC	D507	CAGGACGT
L75	Leaf	HM101	S. meliloti (KH46c)	1	D709	CGGCTATG	D507	CAGGACGT
L146	Leaf	HM101	S. meliloti (KH46c)	1	D710	TCCGCGAA	D507	CAGGACGT
L40	Leaf	HM101	S. meliloti (KH46c)	1	D711	TCTCGCGC	D507	CAGGACGT
L125	Leaf	HM101	None	1	D712	AGCGATAG	D507	CAGGACGT
L171	Leaf	HM101	None	1	D701	ATTACTCG	D508	GTACTGAC
L142	Leaf	HM101	None	1	D702	TCCGGAGA	D508	GTACTGAC
L83	Leaf	HM101	S. medicae (WSM419)	0	D711	TCTCGCGC	D502	ATAGAGGC
L56	Leaf	HM101	S. medicae (WSM419)	0	D704	GAGATTCC	D508	GTACTGAC
L14	Leaf	HM101	S. medicae (WSM419)	0	D705	ATTCAGAA	D508	GTACTGAC
L64	Leaf	HM101	S. medicae (WSM419)	0	D706	GAATTCGT	D508	GTACTGAC
L62	Leaf	HM340	Both	0	D707	CTGAAGCT	D508	GTACTGAC
L21	Leaf	HM340	Both	0	D708	TAATGCGC	D508	GTACTGAC
L118	Leaf	HM340	Both	0	D709	CGGCTATG	D508	GTACTGAC
L49	Leaf	HM340	Both	0	D710	TCCGCGAA	D508	GTACTGAC
L160	Leaf	HM340	S. meliloti (KH46c)	0	D711	TCTCGCGC	D508	GTACTGAC
L11	Leaf	HM340	S. meliloti (KH46c)	0	D701	ATTACTCG	D501	TATAGCCT
L34	Leaf	HM340	S. meliloti (KH46c)	0	D702	TCCGGAGA	D501	TATAGCCT
L44	Leaf	HM340	None	1	D711	TCTCGCGC	D501	TATAGCCT
L13	Leaf	HM340	None	1	D704	GAGATTCC	D501	TATAGCCT
L33	Leaf	HM340	None	1	D705	ATTCAGAA	D501	TATAGCCT
L80	Leaf	HM340	S. medicae (WSM419)	0	D706	GAATTCGT	D501	TATAGCCT
L92	Leaf	HM340	S. medicae (WSM419)	0	D707	CTGAAGCT	D501	TATAGCCT
L26	Leaf	HM340	S. medicae (WSM419)	0	D708	TAATGCGC	D501	TATAGCCT
L8	Leaf	HM340	S. medicae (WSM419)	0	D709	CGGCTATG	D501	TATAGCCT

L121	Leaf	HM101	S. meliloti (KH46c)	0	D706	GAATTCGT	D507	CAGGACGT
JQL01	Nodule	HM101	S. meliloti (KH46c)	0	NA	NA	NA	NA
JQL02	Nodule	HM101	S. meliloti (KH46c)	0	NA	NA	NA	NA
JQL03	Nodule	HM101	S. meliloti (KH46c)	0	NA	NA	NA	NA
JQL04	Nodule	HM101	S. medicae (WSM419)	0	NA	NA	NA	NA
JQL05	Nodule	HM101	S. medicae (WSM419)	0	NA	NA	NA	NA
JQL06	Nodule	HM101	S. medicae (WSM419)	0	NA	NA	NA	NA
JQL07	Root	HM101	None	0	NA	NA	NA	NA
JQL08	Root	HM101	None	0	NA	NA	NA	NA
JQL09	Root	HM101	None	0	NA	NA	NA	NA
JQL10	Nodule	HM056	S. meliloti (KH46c)	0	NA	NA	NA	NA
JQL11	Nodule	HM056	S. meliloti (KH46c)	0	NA	NA	NA	NA
JQL12	Nodule	HM056	S. meliloti (KH46c)	0	NA	NA	NA	NA
JQL13	Nodule	HM056	S. medicae (WSM419)	0	NA	NA	NA	NA
JQL14	Nodule	HM056	S. medicae (WSM419)	0	NA	NA	NA	NA
JQL15	Nodule	HM056	S. medicae (WSM419)	0	NA	NA	NA	NA
JQL16	Root	HM056	None	0	NA	NA	NA	NA
JQL17	Root	HM056	None	0	NA	NA	NA	NA
JQL18	Root	HM056	None	0	NA	NA	NA	NA
JQL19	Nodule	HM340	S. meliloti (KH46c)	0	NA	NA	NA	NA
JQL20	Nodule	HM340	S. meliloti (KH46c)	0	NA	NA	NA	NA
JQL21	Nodule	HM340	S. meliloti (KH46c)	0	NA	NA	NA	NA
JQL22	Nodule	HM340	S. medicae (WSM419)	0	NA	NA	NA	NA
JQL23	Nodule	HM340	S. medicae (WSM419)	0	NA	NA	NA	NA
JQL24	Nodule	HM340	S. medicae (WSM419)	0	NA	NA	NA	NA
JQL25	Root	HM340	None	0	NA	NA	NA	NA
JQL26	Root	HM340	None	0	NA	NA	NA	NA
JQL27	Root	HM340	None	0	NA	NA	NA	NA
JQL28	Nodule	HM034	S. meliloti (KH46c)	0	NA	NA	NA	NA
JQL29	Nodule	HM034	S. meliloti (KH46c)	0	NA	NA	NA	NA
JQL30	Nodule	HM034	S. meliloti (KH46c)	0	NA	NA	NA	NA
JQL31	Nodule	HM034	S. medicae (WSM419)	0	NA	NA	NA	NA
JQL32	Nodule	HM034	S. medicae (WSM419)	0	NA	NA	NA	NA
JQL33	Nodule	HM034	S. medicae (WSM419)	0	NA	NA	NA	NA
JQL34	Root	HM034	None	0	NA	NA	NA	NA

JQL35	Root	HM034	None	0	NA	NA	NA	NA
JQL36	Root	HM034	None	0	NA	NA	NA	NA

Table S2. Metadata regarding the 138 samples used for analysis

Network Name	Mt_General	Mt_Leaf	Mt_Nodule	Mt_Root	Mt_JQL	Mt_JQL_Nodule
Tissue type(s)	Leaf, Root Nodule	Leaf	Nodule	Root	Root and Nodule	Nodule
Samples	102	45	37	20	36	24
Genes included	24,067	21,822	21,054	23,773	23,131	22,123
Edges	289,598,211	238,088,931	221,624,931	282,565,878	267,510,015	244,702,503

Table S3: Statistics associated with co-expression networks built from different tissue types.

Chapter 4 supplemental figures

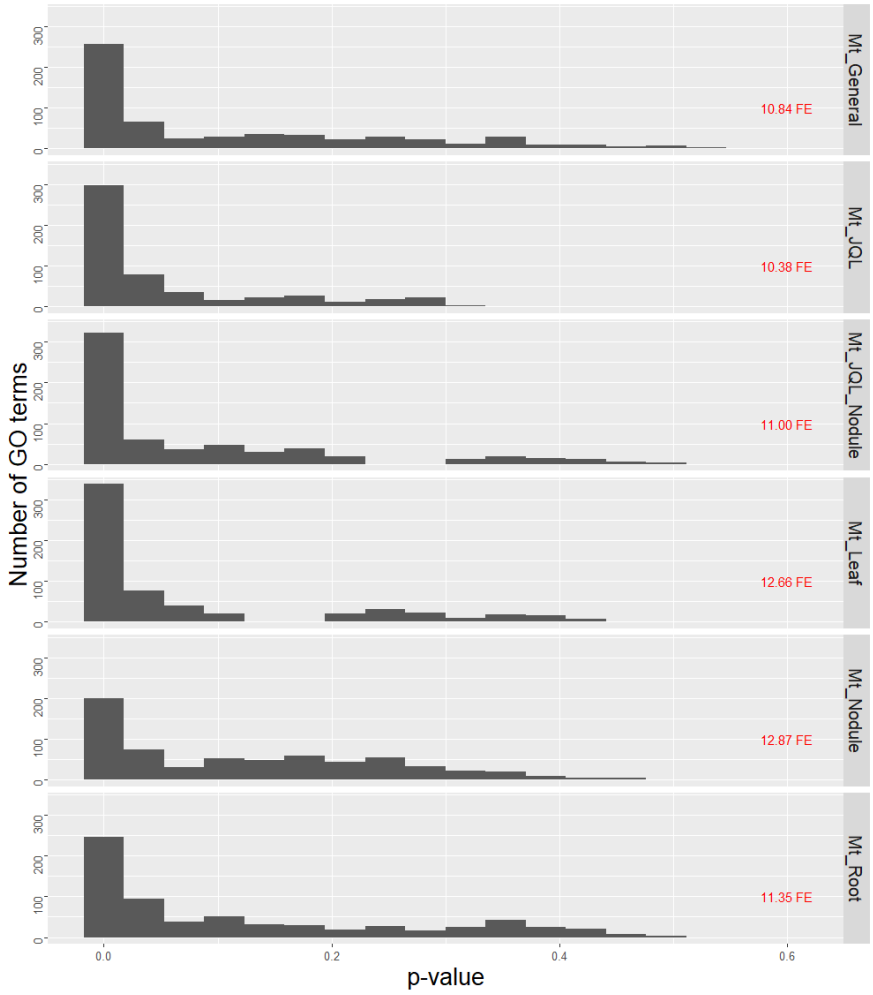
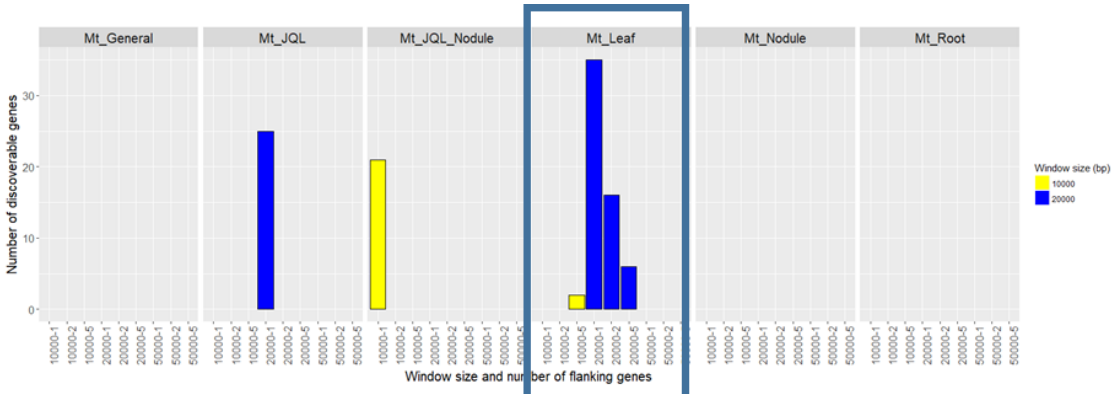


Figure S1. Network GO term enrichment
Distribution of p-values from density-based GO-term enrichment. A histogram of p-values for each density-based GO-term enrichment test based on its density, relative to the distribution of density values from random gene sets similar in size.

A)



B)

Mt_Leaf/Height GWAS									
Window size (bp)	10000			20000			50000		
Number of flanking genes	1	2	5	1	2	5	1	2	5
Starting number of SNP's	197			197			197		
SNPs after collapse	139			133			127		
Number of genes tested	138	182	198	165	257	326	196	363	663
Number of discoverable genes	0	0	2	35	16	6	0	0	0

Figure S2. Co-expression/ Height GWAS discoverable gene summary
Flow chart of candidate gene identification in the height GWAS trait. A) Number of discoverable genes (FDR < 0.35) using the height GWAS with each co-expression network. Colors represent the window size parameter used with Camoco. B) The number of SNP's and genes that were included in each analysis.