THE DEVELOPMENT AND IMPLEMENTATION OF MOLECULAR TOOLS FOR SOYBEAN FUNCTIONAL GENOMICS

by

THOMAS BENJAMIN JACOBS

(Under the Direction of Wayne Parrott)

ABSTRACT

Soybean is one of the world's most important food crops. To ensure continued crop success and realize improved yields, it is important to understand the function of the soybean genes and how they interact with the environment and each other. The generation and implementation of molecular tools is useful for understanding gene action in soybean, and ultimately the improvement of soybean as a crop.

Functional-genomic studies rely on the manipulation of genes by either controlling their expression, or identifying null mutants. RNA interference is frequently used to silence the expression of genes. There are several methods to silence genes in plants, but the trans-acting small-interfering RNA (ta-siRNA) pathway is a particularly simple system. Chapter two details the design and implementation of ta-siRNA vectors to produce siRNAs and induce gene silencing in soybean.

In addition to RNAi, the ability to create targeted DNA modifications is powerful genomics tool. The CRISPR/Cas system has recently been used to create knock-out mutations in a wide range of organisms. Chapter three describes the development of a CRISPR/Cas system to efficiently modify specific DNA sequences in soybean. Targeted

DNA modifications were obtained in nearly 95% of the events evaluated. A high-throughput sequencing method was useful for identifying and quantifying modifications made at target and non-target loci.

In chapter four, the same sequencing approach was modified to identify transgene insertion sites in nine soybean events and transposition sites of *mPing* in soybean. Segregation of the insertions was evaluated in five transgenic lines. The methodology is simple, straightforward and hundreds of libraries can be generated within a week.

As a functional test of RNAi in soybean, chapter five describes transgenic events made with the goal of creating nematode-resistant soybean plants by engineering them with the ability to silence genes in the nematode necessary for parasitism. The work focusses on the characterization and selection of transgenic plants to identify those vectors and events able to produce small RNAs. It is clear that small RNA production is necessary but not sufficient, to obtain consistent nematode resistance. However, this work provides a foundation from which additional attempts can be made.

INDEX WORDS: Soybean, Plant Transformation, Hairy Roots, Gene Silencing, Trans-acting-siRNA, Genome Editing, CRISPR/Cas, Genome Walking, Plant-Parasitic Nematodes

THE DEVELOPMENT AND IMPLEMENTATION OF MOLECULAR TOOLS FOR SOYBEAN FUNCTIONAL GENOMICS

by

THOMAS BENJAMIN JACOBS

BS, University of Connecticut, 2008

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

THE DEVELOPMENT AND IMPLEMENTATION OF MOLECULAR TOOLS FOR SOYBEAN FUNCTIONAL GENOMICS

by

THOMAS BENJAMIN JACOBS

Major Professor:	Wayne Parrott
Committee:	Richard Hussey
	Scott Jackson
	Shavannor Smith
	Xiaoyu Zhang

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2014

ACKNOWLEDGEMENTS

I would just like to acknowledge a few of the many who have helped me along the way. First and foremost, Pete LaFayette for getting me started in molecular biology. Without the molecular expertise he was able to pass on to me, most of this would probably not have happened. Maria Ortega, my co-conspirator. Donna Tucker for answering my constant flow of questions. There were a number of people that were great to work and collaborate with; Noah Lawler, Lisa Kanizay, Nathan Hancock, CJ Tsai, and Bob Schmitz. I would also like to thank the boss, Wayne Parrott. I am sure I will feel the impact of his training for years to come. I very much appreciate the freedom he gave me to pursue my ideas.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Introduction

The world-wide soybean (*Glycine max* (L.) Merr.) crop makes up the largest portion of oilseed production and is second only to palm in vegetable oil production. In the United States, 31 million hectares of soybean were planted in 2013, and the crop value was approximately $42 billion. Current trends suggest that this value will continue to increase. In the United States, nearly 26.5 out of the 89.5 million metric tons produced in 2013 were used as feed by the livestock industry; primarily in swine and poultry production (soystats.com, 2014). It is widely believed that soybean was domesticated from *Glycine soja* Sieb. & Zucc., which is native to eastern Asia. *G. soja* has a viney habit, small black seeds, and pods that shatter, as compared to *G. max*, which is an upright plant, with yellow or black seeds, and pods that usually do not shatter. Both species produce small flowers that are cleistogamous, but out-crossing does occur with the assistance of insects [1].

The soybean genome consists of 20 chromosome pairs as a result of what is believed to be an ancient polyploidization event [2]. The soybean reference genome is published, and there are approximately 55,000 predicted genes [3]. In addition to sequence data, several groups have produced publicly available transcriptome datasets from several tissue types [4, 5].

RNAi mechanism

The RNAi pathways have been exploited and used by plant scientists for nearly a century, even if they were not aware of what they were doing. The first description of RNAi, though the mechanism was unknown at the time, was made in 1928 [6]. It was demonstrated that when a tobacco plant was inoculated with a non-lethal virus, the plant continued to grow, and with time, new plant tissues appeared that were normal and the plant was immune to additional viral infections. This acquired-immunity phenomenon was studied for many years, but not until the advent of transgenic plants, could it be studied in the necessary depth. It was thought the recovery response was a plant's version of the animal immune system, which is triggered by exogenous proteins. Transgenic experiments subsequently showed that when genes for viral coat proteins were transcribed, virus-resistance could be obtained in a process called coat-protein mediated resistance [7, 8].

Insight into the molecular mechanism first came from attempts to make a black petunia flower by over-expressing a chalcone synthase (CHS) gene. Napoli et al. [9] did not get darker flowers as expected; instead a number of petals were white, and the endogenous CHS mRNA level was reduced. The phenomenon was named co-suppression. Through genetic analyses, it was shown that the white petals segregated with the transgenic CHS gene, and those progeny that did not contain the suppressive transgene were phenotypically normal. While the molecular process was unknown, the authors speculated that some epigenetic effect may be involved [9].

Co-suppression and coat-protein mediated resistance were very useful, as some of the first transgenic crops used these RNAi pathways, e.g., the Flavr Savr tomato and

virus-resistant squash [10, 11]. In the late 1990's, work in *Caenorhabditis elegans* demonstrated that double-stranded RNA (dsRNA) was necessary and sufficient for the silencing of endogenous genes [12]. Observation in plants completed the story by showing the production of small, ~25-nt strands of RNA accumulated in plants after viral infection [13]. Since viruses are known to replicate via a dsRNA molecule, the model became clear. The RNAi mechanism begins with the recognition of dsRNA molecules that are degraded it into siRNAs. These siRNAs are able to recognize and cleave complementary sequences and thus induce their silencing, as there is no translatable mRNA left. In addition to viral sequences, transgenes can make dsRNA as a result of rearrangements during integration. Some of these rearrangements, e.g., inverted-tandem repeats, result in the production of dsRNAs, and ultimately silence complementary mRNA, which is consistent with what was observed in petunia [9] and the over-expression of viral coat proteins [7, 8].

RNA interference refers to two processes; transcriptional gene silencing (TGS) and post-transcriptional gene silencing (PTGS). Figure 1.1 presents a general overview of the RNAi pathways in plants. TGS involves the methylation of promoter sequences which prevents transcription. In PTGS, protein synthesis is inhibited by either a loss of target mRNA or, less frequently in plants, ribosomal stalling and an inhibition of translation [14]. However, the two pathways are not isolated from one another, and loci that undergo TGS can also have PTGS occurring at the same time [15]. Many RNAi mechanisms and pathways have been described in model organisms such as *C. elegans* and *Arabidopsis thaliana* and are well conserved between plants, animals, and fungi [14]. In the general RNAi pathway, siRNAs are incorporated into one of several Argonaute

(AGO) complexes to form an RNA-induced silencing complex (RISC) (Figure 1.1). Using the siRNA as a guide, the RISC complex specifically targets and cuts complementary RNA sequences, which are then degraded through normal cellular processes [16]. This process specifically degrades, or inhibits the translation of RNA complementary to the initial dsRNA.

In plants, dsRNA is cleaved by one of the RNase III-type DICER-like enzymes (DCL) to produce siRNAs 20-25 nt in size. Four DCL enzymes have been well characterized (DCL1-4) in arabidopsis. Each one produces siRNAs of a specific size that can enter the silencing pathway (Figure 1.1). DCL1 cleaves imperfect hairpins to produce primarily 21-nt microRNAs (miRNAs). DCL2 produces 22-nt siRNAs and DCL4 produces 21-nt siRNAs, and both appear to have overlapping roles in suppressing viral replication [17]. Furthermore, DCL2 is required for transitive silencing observed in transgenes [18], whereby additional siRNAs are produced along the entire length of the target mRNA, and not from just the initial dsRNA molecule. DCL4 is the principal DCL enzyme responsible for transgenic and viral PTGS [19, 20]. DCL3 generates 24-nt siRNAs that are responsible for *de novo* DNA methylation and TGS [21-23]. Studies have shown that siRNA produced from endogenous and transgenic dsRNA are processed by DCL2, DCL3, and DCL4, suggesting that these enzymes have some redundant functions [24, 25].

The number of siRNAs produced from a dsRNA can be amplified when an RNA-dependent RNA polymerase (RDR) synthesizes additional dsRNAs from cleaved RNAs (Figure 1.2). There are six known RDRs in arabidopsis. After initial slicing, target RNAs are aberrant in that they lack a 5’ cap or a polyA tail, and can then serve as a

substrate for an RDR to produce dsRNAs. It is thought that an mRNA cannot serve as a template for an RDR unless it has first been sliced by a RISC [26]. This reaction can be primer-independent, but often uses a siRNA as a primer. The newly formed dsRNAs are then substrates for one of the DCLs, leading to the production of more siRNAs. This positive-feedback loop, referred to as amplification, greatly enhances the silencing effect [26]. Amplification does not occur for every gene target, and appears to be associated with transgenes and highly expressed endogenous genes [27].

It is thought that the small RNA (sRNA) pathways in plants are derived from ancient and evolutionarily conserved mechanisms to protect genomes from invading viruses and/or transposons. Viruses are known to activate plant silencing pathways [17]. Double-stranded RNA intermediates produced during viral replication or by plant-derived RDRs are cleaved by DCL4 or, when DCL4 is not present, by DCL2. Amplification of the silencing signal is then initiated by RDR6 or RDR1. The sRNAs spread throughout the plant and inhibit the replication and spread of the virus. Many viruses encode proteins that suppress the RNAi pathways in plants by binding the siRNAs and preventing them from entering a RISC, suggesting an evolutionary arms race between plants and viruses [28, 29].

The sRNA pathways can be generally divided into two; the miRNA and the siRNA (Figure 1.1). The miRNA pathway starts with polymerase (Pol) II transcription of a *miRNA* (*MIR)* gene to produce an imperfect hairpin, called the primary-miRNA. DCL1, with other co-factors, cleaves the poly-A tail and 5' cap to generate the pre-miRNA, and then cleaves the rest of the strand to release a 21-nt dsRNA duplex, with 2-nt 3' overhangs. After duplexes are methylated and exported to the cytoplasm, miRNAs

interact with AGO1 to become RISCs, to either slice mRNA or inhibit translation [30]. The processing of miRNA is specific in that for most miRNAs, only one strand of the duplex is incorporated into AGO1 while the other is degraded. The result is an accumulation of a single RNA species and the relative absence of its complement. *MIR* genes are routinely identified in plants, and many are well conserved across the plant kingdom [30].

Several miRNAs have been implicated in controlling responses to abiotic and biotic stresses, as well as cell identity during development. For example, in arabidopsis, miR399 expression is triggered by phosphate starvation and leads to the accumulation of phosphate. When phosphate is again supplied, miR399 expression is suppressed. The over-expression of miR399 leads to the over-accumulation of phosphate [31]. Soybean has a number of miRNAs that are expressed, or suppressed in response to colonization by *Bradyrhizobium japonicum* and nodule formation [32]. In addition, the misexpression of some of these miRNAs leads to an increase in the number of nodules that are formed on transgenic roots, as the roots are unable to regulate nodulation [33].

Endogenous siRNAs can arise from several different types of loci in plants. Loci that have undergone inverted-tandem duplications may produce dsRNAs from read-through transcription. Double-stranded RNAs as long as 6.8 kb have been reported [24]. These inverted regions are far larger than those transformed into plants to induce silencing, which are typically 600 – 1600 bp in length [34]. In soybean, siRNAs produced from an inverted chalcone synthase locus reduce chalcone synthase mRNA in soybean seed coats, resulting in a buff-colored seed coat [35].

Plants also synthesize dsRNA from highly methylated, repetitive sequences, such as transposons and ribosomal sequences, in the RNA-dependent DNA methylation RdDM pathway, to create repeat-associated small interfering RNA (rasiRNA) [22, 23, 36, 37]. In the first step of the pathway, PolV synthesizes a single-stranded RNA molecule from a methylated sequence. The transcript is the template for RDR2 to produce a dsRNA, which is cleaved by DCL3 to into 24-nt siRNAs. AGO4 binds the 24-nt siRNAs, and with Pol IVb, directs the methylation of complementary DNA sequences in *cis* and *trans*. Transposons and other genomic regions are kept silenced through this pathway [16, 36, 38].

The trans-acting siRNA (ta-siRNA) pathway is a combination of the miRNA and the siRNA pathways [39]. Trans-acting siRNAs are produced from a transcript that is first cleaved by AGO1 directed by a miRNA, followed by RDR6 recruitment to generate a dsRNA molecule from the 3' miRNA-cleavage product (Figure 2.1). DCL4 cleaves the dsRNA every 21-nt from the initial miRNA cleavage point, resulting in a regular pattern of siRNAs, referred to as phased-siRNAs (phasiRNAs) [40]. There are many loci in plants that produced phasiRNA, but to be a true ta-siRNA, the siRNA must be demonstrated to target an RNA sequence in *trans*. Hence, this has led to the generation of yet another sRNA term, the phasiRNA [41].

Not all miRNAs recruit RDR6 to produce ta-siRNAs, and the miRNAs that can, appear to rely on one of two independent triggers, referred to as the one-hit and two-hit models. In the one-hit model, a single 22-nt miRNA is sufficient to recruit RDR6 and produce phasiRNAs. In arabidopsis, the 22-nt-miRNAs, miR173 and miR828, were sufficient for the production of phasiRNAs, whereas 21-nt versions were only able to

cleave the targeted site, and did not produce phasiRNAs. It was reasoned that the length of the miRNA somehow recruited RDR6. It has been demonstrated that an asymmetric bulge in the miRNA-miRNA* duplex (22-nt miRNA duplexes are typically asymmetric) is also sufficient for RDR6 recruitment, as 21-nt miRNAs with asymmetric bulges also produce phasiRNA [42]. Clearly, both the size and shape of the miRNA duplex influences the production of phasiRNAs.

In the ta-siRNA two-hit model, a ta-siRNA transcript has two miRNA targets, where one target is cleaved and the other is bound by the RISC. The intervening sequence is amplified by RDR6, and cleaved by DCL4 to produce ta-siRNAs [43]. The two-hit miR390 family is conserved in most plant species, and is likely ancestral to the one-hit model [44].

RNAi in research

With the knowledge that dsRNA can be used to produce sRNAs, transgenic vectors have been designed to generate artificial dsRNAs to efficiently silence target genes [45]. These vectors use portions of coding sequences arranged as inverted-repeat 'arms', separated by a 'loop', to form a dsRNA 'hairpin' upon transcription [45, 46]. Two types of loop sequences are used in hairpin vectors; one being intron sequences [45], and the second type are referred to as 'spacer' sequences (i.e. a non-functional portion of *GUS*) that do not code for anything [46].

Hairpin vectors can also target non-coding promoter sequences to induce silencing [47, 48]. It is thought that siRNAs generated from these hairpins do not direct the cleavage of transcripts; rather they target complementary promoter sequences for *de*

novo DNA methylation and initiate TGS. The methylation pattern is then maintained by DNA methyltransferases, and can be inherited without the original silencing signal [36, 37].

Virus-induced gene silencing (VIGS) is another popular method for inducing gene silencing in plants. Viral sequences are introduce into plant tissues either through *Agrobacterium*-infiltration, biolistic bombardment, or *in vitro* transcription and injection [49]. Once introduced, the virus replicates and moves throughout the plant. As the virus spreads, the RNAi machinery produces siRNAs from viral RNAs in an attempt to silence and limit the spread of the virus. Small-interfering RNAs are made from any sequence attached to the viral sequences, and the incorporation of target genes in VIGS vectors results in the silencing of complementary genes [49]. Virus-induced gene silencing has been used in a wide range of plant species [49-52], and the primary advantage over other RNAi systems is that transgenic plants do not have to be produced. However, depending on the pathogenicity of the virus used, infected plants exhibit a range of viral symptoms, which can make phenotyping difficult. In addition, the silencing effect may only be transient [49], and non-uniform across the plant [53], so proper timing, and sufficient numbers are required.

Other gene-silencing methods produce aberrant forms of RNA by including sequences in the transcript that induce the cleavage of the mRNA before a poly-A tail is added [54, 55]. Additionally, vectors can be made that simply lack a polyadenylation signal [56, 57]. It is thought that aberrant forms of RNA somehow bypass the normal RNA-degradation pathway and are instead amplified by RDR6 to enter into the siRNA-silencing pathway [16].

As the understanding of gene-silencing pathways increases, the next generation of gene-silencing vectors will offer more specific and efficient options to induce gene silencing. Artificial miRNAs (amiRNA) have been successfully used in a range of plant species [58-61]. These vectors have the advantage of producing a single miRNA that can be designed to specifically target individual, or multiple sequences, which is useful when working with highly similar gene families. Artificial miRNAs with an asymmetric duplex have also been used to induce the production of ta-siRNAs [62]. This method produces of a large variety of siRNAs from the target transcript, which may be able to increase the level of silencing as compared to amiRNA vectors.

MicroRNA target sites from the ta-siRNA pathway have also been used to generate silenced events. Insertion of the arabidopsis miR173 target site next to a portion of the phytoene desaturase (*PDS*) gene resulted in silencing of PDS, as evidenced by photobleaching and a reduction in PDS mRNA levels [63]. The ta-siRNAs were only produced 3' to the miR173 target site and were dependent on *RDR6* and *DCL4*. The arabidopsis miR173 has also been used to silence four other endogenous genes in arabidopsis in a system called miRNA-induced gene-silencing, or MIGS [64]. It has been suggested that the arabidopsis *miR173* gene could be introduced into other plant species to use the MIGS system [64]. However, a simpler alternative would be to use a plant's endogenous ta-siRNA system. If the plants are already expressing miRNAs that induce the ta-siRNA system, simply knowing the miRNA target sequence would be sufficient to create gene-silencing vectors. In soybean, phasiRNA loci have been computationally identified, but have yet to be experimentally verified [41, 65].

Soybean genes have been successfully silenced using various gene-silencing techniques. For example, hairpin vectors and aberrant-mRNA vectors have been used to silence fatty acid desaturase [54, 66] and VIGS has been used to silence phytoene desaturase [67], and various other pathways [50, 53]. Artificial miRNA vectors have also been successful in soybean hairy roots [68].

Targeted DNA mutations

The ability to selectively modify DNA sequences is incredibly useful for genetics research. Mutants of interest are typically obtained by random mutagenesis screens followed by mapping or TILLING. However, these methods can be tedious and time consuming. A simpler option is to directly modify the DNA sequence of interest. To this end, sequence-specific nucleases can be used to induce DNA double-stranded breaks (DSB), and ultimately the mutation of the target sequence. Non-homologous end-joining (NHEJ), and homologous recombination (HR), are two competing pathways that are used by eukaryotic cells to repair DNA DSBs [69]. NHEJ is the primary repair pathway, and generates short insertions and/or deletions (indels) at the break site, which typically result in frame-shift mutations. HR is less frequently used for repair, but can lead to the incorporation of foreign DNA at the break site. These pathways are frequently utilized to produce null mutations (NHEJ) or targeted insertions (HR).

Restriction enzymes are the most commonly used sequence-specific nuclease in molecular biology; however, they generally cut too frequently to be used in the large genomes of eukaryotes. Homing-endonucleases, zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and clustered regularly

interspaced short palindromic repeats (CRISPR) have been used to induce targeted DSB in plants and other organisms. Homing-endonucleases naturally target specific, 20 to 30-bp sequences, using a protein-DNA interaction, and are usually highly specific [70]. To use homing-endonucleases, as well as any other nuclease, the DNA-recognition motif needs to be modifiable to accommodate different DNA targets. However, the modification of homing-endonucleases is difficult, as changes to the recognition motif tend to eliminate nuclease activity [71]. Despite this limitation, trait stacking has been achieved in cotton using a modified I-*Cre*I homing-endonuclease [72].

Zinc-finger nucleases are chimeric proteins from the fusion of zinc-finger DNA-binding proteins, and the nuclease motif of the *Fok*I restriction enzyme [73]. A single zinc finger recognizes three DNA bases, and typically three zinc fingers are used in an array. Publically available reagents and software are available to design and create effective ZFN arrays [74]. The *Fok*I domain is functional only as a dimer, so two zinc-finger arrays are required to target any sequence. ZFNs have been used in a wide range of biological systems [75, 76], including plants [77-79], but they are expensive and time-consuming to build, and require extensive validation [80].

Transcription activator-like effectors (TALEs) are proteins that interact with DNA, and are used by the plant pathogen *Xanthamonas* to induce the expression of host genes during infection [81]. The design and use of TALENs is conceptually similar to that of ZFNs, in that the DNA-recognition motif, the TALE, is fused to a *Fok*I nuclease motif. Within a TALE is a 32 amino acid repeat, with two residues encoding for each DNA base. This simple protein-DNA code allows TALENs to theoretically target any DNA sequence [82]. Transcription activator-like effector nucleases are easier and less

expensive to construct than ZFNs, and, as with ZFNs, there are publically available reagents and design protocols available [83]. Both ZFNs and TALENs have been successfully used in soybean [74, 84, 85].

The development of CRISPRs is a great advance for targeted genome editing. The modification of DNA-recognition motifs is far simpler than for any other current nuclease technology. This ease of use has fueled its widespread and rapid adoption. Rather than a protein-DNA interaction, DNA-target recognition in CRISPRs uses RNA-DNA hybridization [86]. Since CRISPR targeting is directed by the 20-nt RNA molecule, additional targets can be made by simply changing 20 bp in a vector. Designing an oligonucleotide is a much easier process than the multi-step cloning modifications required for both ZFNs [74] and TALENs [83]. High-throughput methods have been developed that can make libraries of targeting RNAs for hundreds to thousands of gene targets [87]. With such a high throughput, CRISPRs can be used for forward genetic screens instead of the one-gene-at-a-time approach.

Naturally found in most bacteria and archaea, CRISPRs cleave the DNA of invading phages, providing anti-viral defense. When bacteria are infected with a virus, most of the cells die, but a few survive and the CRISPR system incorporates phage DNA into a CRISPR locus. The phage-derived sequences guide the Cas (CRISPR associated) 9 protein (a nuclease) to complementary phage sequences to cleave and destroy the invader [88]. Viral immunity was first shown in cultures of *Streptococcus thermophilus* when sequences complementary to phages were artificially incorporated into CRISPR loci, and resistance was conferred without prior exposure to the phages [89]. Work in

Streptococcus pyogenes demonstrated that only two components were required to cleave DNA sequences *in vitro* and *in vivo*; Cas9 and a guide RNA (gRNA) [90].

The most common target DNA sequence motif is the 23-nt $GN_{20}GG$; although other motif lengths have been used (Figure 1.3) [91]. The targeting range of this motif is limited by the 5' G and 3' GG, as compared to the theoretically unlimited targeting range of TALENs, but targets are still abundant in most DNA sequences. In wheat, the $GN_{19\text{-}21}GG$ motif is predicted to be present 21 times per cDNA, and in rice, the $AN_{19\text{-}21}GG$ motif is predicted 32.6 times per cDNA [92]. Thus, while the theoretical targeting range is lower than that of TALENs, the current state of CRISPR technology can target virtually any gene.

The 5' G is required by the Pol III U6 promoters used to drive the expression of the short gRNA molecules. However, there is some flexibility, as Pol III promoters with different first-base requirements have be used [92]. The 3' NGG portion of the target motif is called the proto-spacer adjacent motif (PAM), and is absent in the targeting gRNA. In bacteria, the PAM is present in the target phage genome, but absent in the bacterial genome. It is thought this is how bacteria can recognize self from non-self to avoid cleaving their own DNA. The PAM NGG is specific to the Cas9 protein from *S. pyogenes*, and other PAM sequences are found in different Cas9 proteins [86, 93] derived from different bacterial species. The identification and utilization of Cas9 proteins from different bacterial species should increase the targeting range of CRISPRs [86].

CRISPRs were first shown to induce DNA DSBs in bacteria [90] and human cells [93, 94] and were quickly applied to many other animals, plants, and fungi [91]. This

nuclease system can now be easily tailored to almost any model system, and a number of vectors are available from the non-profit vector repository, Addgene.

As with any DNA-targeting technology, there is the potential to inadvertently hit non-target sequences, resulting in off-target modifications. In bacteria, single mismatches between the gRNA and the target DNA on the 5' end of the gRNA still allowed DNA cleavage [90]. Mismatches closer to the 3' end reduced cleavage efficiency, and any mismatches beyond the 9th base completely inhibited DNA cleavage. It was suggested that there was a 'seed' region on the 3' end of the gRNA that must perfectly match for DNA cleavage to occur. Many of the first studies using CRISPRs detected few, if any, off-target modifications [92, 95-97]. However, a more systematic study identified a high frequency of off-target modifications in cultured human cells [98]. Single and multiple mismatches within the seed region still allowed DNA cleavage, though generally, mismatches on the 5'end were better tolerated than those on the 3' end. A follow-up study demonstrated that shorter gRNAs ($GN_{18}GG$) were just as effective at inducing DNA DSBs and had increased specificity [99]. While the potential for off-target modifications surely exists for CRISPRs, proper gRNA design can avoid many potential off targets. In addition, the off-target reports to date have primarily focused on identifying complementary off targets. A systematic study of whole genome re-sequencing experiments is needed to determine the extent of non-complementary, off-target modifications.

In plants, the CRISPR system has been most used in arabidopsis and rice [92, 100-105], but it has also been used in maize [106], wheat [107], citrus [108] and liverwort [109]. Modifications were readily made in each species, and so there does not

appear to be any limitation to the plant species that can be modified. However, the development and validation of CRISPR vector systems is needed for many other plant species.

Most attempts to use CRISPRs, and other nuclease systems, in plants have focused on the ability to induce short indels via NHEJ. However, a potentially more powerful and useful technique would be the specific replacement, or insertion, of DNA sequences via HR. Such methods are common in animal systems [94, 95], but are not routine in plants. The ability to selectively change individual DNA bases would be invaluable for validating causative SNPs, or modifying enzymatic pathways. Targeted insertions could also allow for the incorporation of multiple transgenes into a single segregating unit (stacking) which would be very beneficial in a breeding program.

In plants, targeted insertions are typically performed in protoplasts, whereby a large number of cells are transformed, and a strong selection pressure is used [103]. However, work in cotton has demonstrated that targeted insertions are possible with transgenic vectors containing homing endonucleases [72]. While only 1.8% of the recovered events contained the targeted insertion, having the stacked traits in a simple integration pattern is likely advantageous enough to warrant the generation of hundreds of events.

Soybean transformation

There are two widely used methods for obtaining transgenic soybean plants; particle bombardment and cotyledonary-node transformation with *Agrobacterium tumefaciens*. Particle bombardment uses small pieces of tungsten or gold coated with

DNA which are then shot into cells. The cells are selected and ultimately regenerated into whole plants [110]. Alternatively, the cotyledonary node can be inoculated with *A. tumefaciens* which transfers transfer-DNA (T-DNA) into target cells, which can then be regenerated to produce a transformed plant [110, 111].

While there are defined methods for obtaining transgenic soybean plants, the process requires at least one year to generate several independent, homozygous lines for testing [110]. Given this limitation, methods have been developed to rapidly generate transgenic tissues with *Agrobacterium rhizogenes*. *A. rhizogenes* transfers a T-DNA that contains genes that promote the production of roots from the transformed cells. The result is a transgenic root system referred to as 'hairy roots', owing to their highly branched structure. The transgenic roots can usually be maintained indefinitely in tissue culture. Given the ease and efficiency of transformation, hairy roots have been extensively used in soybean root studies [112-115]. However, since whole plants cannot be generated from soybean roots, hairy roots only provide a rapid testing system.

Transgene integration and mapping

The exact process of transgene integration is a very complex mechanism that is not entirely understood. There are different mechanisms for transgene integration for biolistic- or *Agrobacterium*-mediated transformation, and incorporation of transgenic DNA sequences may require microhomology between the transgene DNA and the insertion site [116]. It is widely believed that the integration of transgenes requires an initial DNA DSB. As the plant cell repairs the break, transgenic DNA can then be incorporated into the break site, resulting in a transgenic insertion event [116, 117].

During the integration process, transgenic DNA can rearrange to form concatemers, indels, local tandem-duplications, or associate with plant DNA to form large insertion arrays, interspersed with transgenic and genomic DNA [116, 118-122]. As previously mentioned, rearrangements of transgenes can result in the silencing of transgenes and complementary sequences [10]. Both biolistic- and *Agrobacterium*-mediated transformation of plant cells can result in the rearrangement of delivered DNA [117]. Being able to understand the process of transgene integration could help improve transformation efficiency and produce transgenic lines with cleaner arrangements of transgenes, which would be more desirable for consistent expression, downstream breeding efforts, and regulatory approval.

Many methods have been used to determine the structure of the insertion events, and the identification of DNA sequences flanking transgene insertion sites. Each method has inherent advantages and disadvantages. Southern blots are the standard method to determine copy number, but the presence of transgene concatemers can result in an incorrect interpretation of the Southern blot results [122]. Plasmid rescue has been used to thoroughly characterize large transgene arrays at the sequence level in oat [118]. But the plasmid rescue method requires the presence of a bacterial selectable marker, so such a method is not suitable when plant cells transformed with DNA lacking a vector backbone, or for *Agrobacterium* transformations for which usually only the T-DNA is incorporated, and therefore lacks a bacterial selectable marker. Fluorescent *in situ* hybridization on extended DNA fibers (fiber-FISH) is excellent for determining the structure of large transgenic arrays [120, 122], but this method does not have sequence-level resolution. Most recently, whole genome sequencing/resequencing (WGS) has been

used to characterize transgene loci in maize [123] and common bean [121], but the expense and bioinformatic resources required makes WGS cost-prohibitive for sequencing large numbers of samples. Furthermore, in WGS, assuming that most transgenic loci are 10 kb in length, many crop genomes are at least 1 Gb, and the sequencing reads are paired-end and 100 to 300 bp in length, only 0.2 to 0.6% of the sequencing reads will contain transgenic bases, which is an incredible waste of sequencing resources. Thermal-asymmetric interlaced (TAIL) PCR has also been a useful tool for identifying sequences flanking T-DNA insertions [124], but this technology requires knowing the ends of the T-DNA , which are usually the left- and right-borders. However, TAIL-PCR is not suitable for transgenes with complex rearrangements or unknown break points. Clearly there are many methods to interrogate transgene insertion sites. With the reduced cost of high-throughput sequencing, and increased read lengths, a method to specifically sequence transgenic sequences would be beneficial for characterizing a large number of transgenic events.

When transgene insertion sites are identified, unique markers for each transgene insertion can be created to determine the number of segregating loci and zygosity in segregating progeny. These markers can also aid the backcrossing of transgenes into an elite cultivar, since specific integration events could be selectively back-crossed in. While WGS has proven beneficial for identifying flanking sequences [123, 125], re-sequencing smaller, defined regions, would be a better use of resources. For example, when a sequence of interest is inserted randomly into a genome, it can be helpful to know where that sequence is located, without wasting sequencing and bioinformatic resources by re-sequencing unmodified regions. In the case of transgenic events, the junction

sequence of the transgenic DNA can aid in the identification of promoter/ enhancer traps, identify mutagenized genes (as in the case of T-DNA insertions or transposon tagging), and improve our general understanding of the transformation process.

In addition to whole-genome sequencing, several PCR-based methods exist to capture flanking sequences (genome walking); such as, TAIL-PCR [126], TOPO-vector ligation PCR [127], and inverse PCR [128]. Thermal asymmetric interlaced-PCR and similar methods require the use of random primers or adapters, high Tm-primers, and can lead to high background caused by non-specific amplification. Genome walking kits can be purchased from several companies, but are fairly expensive, and also have stringent requirements, thus are not amenable to high-throughput procedures. Whole-genome sequencing/ re-sequencing requires a reference genome, is cost-prohibitive for a large number of samples and requires bioinformatic expertise.

TOPO-vector ligation PCR is advantageous since the adapter is a cloning vector, and nested PCR primers can be designed to the vector 3' of the insert, which should reduce non-specific amplification [127]. In addition, the required materials are present in most molecular biology labs. The previously published TOPO-vector ligation protocol used cloning and Sanger sequencing to identify junction fragments [127]. However, such a method can only process a few samples at a time; the development of a high-throughput method would be beneficial.

Application of RNAi for nematode resistance

Of the several types of plant parasitic nematodes found in soybean fields, soybean-cyst (SCN, *Heterodera glycines*) and root-knot nematodes (RKN, *Meloidogyne*)

are the most damaging. In the years from 2006-2009, SCN in the U.S. was the most damaging pest or disease of soybean, reducing yields by 2.6 to 4.7 million metric tons, which puts soybean losses due to SCN at more than $1 billion [129]. Root-knot nematodes have less of an effect across the U.S., however they are an important pest in Georgia, as they are the number one pest or disease causing loss in 2003 and 2004, and second only to rust in 2005 [130]. In addition, RKNs have a wide host range and are an economically important pest around the world [131].

Soybean-cyst nematodes are obligate plant parasites. They have five life stages; four juvenile (J1-J4) and one adult stage. They undergo their first molt inside the egg and emerge in the J2 stage. At the J2 stage, SCN are free-living and search for roots to infest. As J2 juveniles, nematodes can penetrate and enter the roots of susceptible and resistant plants alike, but it is the formation of a feeding site near or within the vascular tissue that is critical for their development. The feeding site is a syncytium, a cluster of hundreds of cells, the formation of which is controlled by effector proteins secreted by the nematode. A single female can produce as many as 600 eggs, and this life cycle can be completed in as little as 22 days; several generations can occur within a growing season. Not all 600 eggs will hatch at once as some individuals are fast hatchers while others remain viable in the soil for years [132]. This staggered hatching rate results in a high load of nematode eggs in infested fields which are difficult to control.

Root-knot nematodes are unique in their ability to induce giant cells in plants which serve as feeding sites. As compared to syncytia, giant cells are derived from a few cells that undergo repeated mitosis without cell division [133]. Along with giant cell formation, there is an expansion of the root pericycle and cortical cells. Ultimately this

expansion produces the characteristic gall, called a root-knot (Figure 1.4). Root-knot nematode eggs are tough and resilient and may reside in the soil for years before hatching. The eggs are resistant to the intrusion of several compounds, making treatment with nematicides difficult. Like the SCN, RKN hatch at the J2 stage and are drawn to a host, likely using root exudates as a signal. Once the nematodes reach the root, several may accumulate at the zone of elongation. It is suspected that nematode exudates stimulate the root hairs of plants, including rapid ionic fluxes, cytoskeleton reorganization, and nuclear relocation. Once inside the plant, the RKN move intercellularly between the cortical and meristematic cells towards the root tip. Upon reaching the meristem at the root tip, they reverse direction and move upwards within the vascular cylinder. The RKNs stop and initiate a feeding site at the zone of differentiation. Multiple nematodes may follow the same path within a root resulting in large galls that may contain multiple individuals [134].

Both RKN and SCN use similar mechanisms to develop feeding sites within host plants. The nematodes secrete effectors (parasitism proteins) through their hollow stylet into cells from first penetration, movement, and the establishment and maintenance of a feeding site. The parasitism proteins are produced in the dorsal or subventral gland cells of the nematode. The morphology, contents, and activity of these cells changes throughout the parasitic process. As these genes are absent in non-parasitic nematodes, it is thought that these genes are what give the nematodes the ability to parasitize plants. If this is true, then the elimination or suppression of the parasitism proteins in feeding nematodes should eliminate the nematode's ability to properly infect and establish a

feeding site within a plant [135]. Many parasitism proteins have been identified and characterized for the different plant parasitic nematode species [136, 137]

Plant-parasitic nematode resistance induced by RNAi was first shown in arabidopsis plants expressing a hairpin vector generated from the RKN parasitism gene *16D10* [138]. Transgenic plants produced siRNAs to *16D10* and were able to reduce nematode infection. It was reasoned that when the nematodes fed upon the transgenic plants, they ingested the siRNAs, which were then able to enter the nematode RNAi pathway and silence 16D10, thus preventing the nematode from being an effective parasite. The *16D10* hairpin vectors have also been reported to be successful for obtaining RKN resistance in potato [139] and grape hairy roots [140].

In transgenic tobacco, *M. incognita* infection rates and target mRNA levels were reduced, but the production of siRNAs was not shown, making it difficult to establish the causal relationship of RNAi and resistance [141]. Also in tobacco, plants producing siRNAs we able to reduce *M. javanica* target mRNA levels, however there was no effect on nematode survival, suggesting that not all targets of RNAi lead to effective resistance. Arabidopsis plants expressing hairpin vectors to four putative parasitism genes were resistant to sugar-beet cyst nematode infection [142]. Corresponding with lower female and egg counts, there was a significant reduction in the parasitism gene transcripts in the nematodes. It is noteworthy that there was considerable variation in nematode counts between different experimental replications. Some homozygous transgenic lines reduced the number of nematodes by 40% in one experiment, but failed to reduce nematodes in others [142]. These data suggest that the resistance conferred by RNAi may be variable, which would be undesirable in a field setting.

RNAi-mediate resistance has not been limited to nematodes. Coleopteran resistance has also been reported in corn and cotton [143, 144]. In both cases, the dsRNA-producing events also produced siRNA to the target genes in the insect, and mRNA levels of the target genes were reduced. Resistance to the powdery mildew fungus has also been achieved in barley and wheat [145]. Thus, RNAi-mediated resistance appears to be a general process that can be used to defend against a variety of plant pests and pathogens.

Importantly, soybean expressing a dsRNA of an SCN sequence, homologous to sperm protein from *C. elegans*, led to a reduction in eggs in nematodes that fed upon the transgenic plants [146]. They also show a reduction in the target mRNA in the SCN and corresponding siRNA blots. However, the bioassays were conducted on T0 transgenic plants, which are the initial transgenic plants that come out of tissue culture, and tend to have abnormal phenotypes. There are several additional reports of nematode reduction in soybean hairy roots [147-150], but only one reports the production of the causative siRNAs [149]. Given these results, it seems likely that the expression of dsRNA in soybean of nematode genes could produce resistant soybean plants. However, there is a need to evaluate for nematode resistance in transgenic soybean plants over several generations, as only heritable resistance will be useful in a breeding program.

Based on Huang et al. [138], eight hairpin vectors targeting SCN and ten hairpin vectors targeting RKN parasitism genes were transformed into soybean. However, none of the homozygous events reduced the number of feeding nematodes or number of eggs [151]. Furthermore, using sRNA deep-sequencing analysis, very few, or no siRNAs were detected in most of the lines (Figure 5.2). The goal of this line of research will be to

explain some of the variables for the construction of high effective gene-silencing vectors, so that ultimately these vectors will be used for the creation of nematode silencing vectors, as well as testing nematode resistance in transgenic soybean plants.

References

1. Chung G, Singh RJ: **Broadening the genetic base of soybean: A multidisciplinary approach**. *Critical Reviews in Plant Sciences* 2008, **27**(5):295-341.

2. Shoemaker RC, Schlueter J, Doyle JJ: **Paleopolyploidy and gene duplication in soybean and other legumes**. *Current Opinion in Plant Biology* 2006, **9**(2):104-109.

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

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

5. Severin AJ, Woody JL, Bolon YT, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE *et al*: **RNA-Seq atlas of *Glycine max*: A guide to the soybean transcriptome**. *Bmc Plant Biology* 2010, **10**:16.

6. Wingard SA: **Hosts and symptoms of ring spot, a virus disease of plants.** 1928, **1928 Aug. 1 v. 37 no. 3**:p. 127-154.

7. Abel PP, Nelson RS, De B, Hoffmann N, Rogers SG, Fraley RT, Beachy RN: **Delay of disease development in transgenic plants that express the tobacco mosaic-virus coat protein gene**. *Science* 1986, **232**(4751):738-743.

8. Beachy RN, Loeschfries S, Tumer NE: **Coat protein-mediated resistance against virus-infection**. *Annual Review of Phytopathology* 1990, **28**:451-474.

9. Napoli C, Lemieux C, Jorgensen R: **Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans**. *Plant Cell* 1990, **2**(4):279-289.

10. Krieger EK, Allen E, Gilbertson LA, Roberts JK, Hiatt W, Sanders RA: **The Flavr Savr tomato, an early example of RNAi technology**. *Hortscience* 2008, **43**(3):962-964.

11. APHIS/USDA: **APHIS/USDA Petition 92-204-01 for Determination of Nonregulated Status for ZW-20 Squash**. In.; 1994.

12. Fire A, Xu SQ, Montgomery MK, Kostas SA, Driver SE, Mello CC: **Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans***. *Nature* 1998, **391**(6669):806-811.

13. Hamilton AJ, Baulcombe DC: **A species of small antisense RNA in posttranscriptional gene silencing in plants**. *Science* 1999, **286**(5441):950-952.

14. Carthew RW, Sontheimer EJ: **Origins and mechanisms of miRNAs and siRNAs**. *Cell* 2009, **136**(4):642-655.

15. Mourrain P, van Blokland R, Kooter JM, Vaucheret H: **A single transgene locus triggers both transcriptional and post-transcriptional silencing through double-stranded RNA production**. *Planta* 2007, **225**(2):365-379.

16. Brodersen P, Voinnet O: **The diversity of RNA silencing pathways in plants**. *Trends in Genetics* 2006, **22**(5):268-280.

17. Garcia-Ruiz H, Takeda A, Chapman EJ, Sullivan CM, Fahlgren N, Brempelis KJ, Carrington JC: ***Arabidopsis* RNA-dependent RNA polymerases and dicer-like**

proteins in antiviral defense and small interfering RNA biogenesis during turnip mosaic virus infection. *Plant Cell* 2010, **22**(2):481-496.

18. Mlotshwa S, Pruss GJ, Peragine A, Endres MW, Li JJ, Chen XM, Poethig RS, Bowman LH, Vance V: ***DICER-LIKE2* plays a primary role in transitive silencing of transgenes in *Arabidopsis***. *Plos One* 2008, **3**(3):11.

19. Dunoyer P, Himber C, Voinnet O: ***DICER-LIKE 4* is required for RNA interference and produces the 21-nucleotide small interfering RNA component of the plant cell-to-cell silencing signal**. *Nature Genetics* 2005, **37**(12):1356-1360.

20. Gasciolli V, Mallory AC, Bartel DP, Vaucheret H: **Partially redundant functions of *Arabidopsis* DICER-like enzymes and a role for DCL4 in producing trans-acting siRNAs**. *Current Biology* 2005, **15**(16):1494-1500.

21. Huettel B, Kanno T, Daxinger L, Bucher E, van der Winden J, Matzke AJM, Matzke M: **RNA-directed DNA methylation mediated by *DRD1* and *Pol IVb*: A versatile pathway for transcriptional gene silencing in plants**. *Biochimica Et Biophysica Acta-Gene Structure and Expression* 2007, **1769**(5-6):358-374.

22. Xie ZX, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Zilberman D, Jacobsen SE, Carrington JC: **Genetic and functional diversification of small RNA pathways in plants**. *Plos Biology* 2004, **2**(5):642-652.

23. Chan SWL, Zilberman D, Xie ZX, Johansen LK, Carrington JC, Jacobsen SE: **RNA silencing genes control *de novo* DNA methylation**. *Science* 2004, **303**(5662):1336-1336.

24. Dunoyer P, Brosnan CA, Schott G, Wang Y, Jay F, Alioua A, Himber C, Voinnet O: **An endogenous, systemic RNAi pathway in plants**. *Embo Journal* 2010, **29**(10):1699-1712.

25. Henderson IR, Zhang XY, Lu C, Johnson L, Meyers BC, Green PJ, Jacobsen SE: **Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning**. *Nature Genetics* 2006, **38**(6):721-725.

26. Voinnet O: **Use, tolerance and avoidance of amplified RNA silencing by plants**. *Trends in Plant Science* 2008, **13**(7):317-328.

27. Vaistij FE, Jones L, Baulcombe DC: **Spreading of RNA targeting and DNA methylation in RNA silencing requires transcription of the target gene and a putative RNA-dependent RNA polymerase**. *Plant Cell* 2002, **14**(4):857-867.

28. Eamens A, Curtin SJ, Waterhouse PM: **RNA Silencing in Plants**: Springer; 2010.

29. Sharma P, Ikegami M: **RNA-silencing suppressors of geminiviruses**. *Journal of General Plant Pathology* 2008, **74**(3):189-202.

30. Voinnet O: **Origin, biogenesis, and activity of plant microRNAs**. *Cell* 2009, **136**(4):669-687.

31. Bari R, Pant BD, Stitt M, Scheible WR: ***PHO2*, microRNA399, and *PHR1* define a phosphate-signaling pathway in plants**. *Plant Physiology* 2006, **141**(3):988-999.

32. Subramanian S, Fu Y, Sunkar R, Barbazuk WB, Zhu JK, Yu O: **Novel and nodulation-regulated microRNAs in soybean roots**. *Bmc Genomics* 2008, **9**:14.

33. Li H, Deng Y, Wu TL, Subramanian S, Yu O: **Misexpression of miR482, miR1512, and miR1515 increases soybean nodulation**. *Plant Physiology* 2010, **153**(4):1759-1770.

34. Wesley SV, Helliwell CA, Smith NA, Wang MB, Rouse DT, Liu Q, Gooding PS, Singh SP, Abbott D, Stoutjesdijk PA *et al*: **Construct design for efficient, effective and high-throughput gene silencing in plants**. *Plant Journal* 2001, **27**(6):581-590.

35. Tuteja JH, Zabala G, Varala K, Hudson M, Vodkin LO: **Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in *Glycine max* seed coats**. *Plant Cell* 2009, **21**(10):3063-3077.

36. Eamens A, Wang MB, Smith NA, Waterhouse PM: **RNA silencing in plants: Yesterday, today, and tomorrow**. *Plant Physiology* 2008, **147**(2):456-468.

37. Daxinger L, Kanno T, Bucher E, van der Winden J, Naumann U, Matzke AJM, Matzke M: **A stepwise pathway for biogenesis of 24-nt secondary siRNAs and spreading of DNA methylation**. *Embo Journal* 2009, **28**(1):48-57.

38. Rigal M, Mathieu O: **A "*mille-feuille*" of silencing: Epigenetic control of transposable elements**. *Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms* 2011, **1809**(8):452-458.

39. Fei QL, Xia R, Meyers BC: **Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks**. *Plant Cell* 2013, **25**(7):2400-2415.

40. Yoshikawa M, Peragine A, Park MY, Poethig RS: **A pathway for the biogenesis of trans-acting siRNAs in *Arabidopsis***. *Genes & Development* 2005, **19**(18):2164-2175.

41. Zhai JX, Jeong DH, De Paoli E, Park S, Rosen BD, Li YP, Gonzalez AJ, Yan Z, Kitto SL, Grusak MA *et al*: **MicroRNAs as master regulators of the plant NB-LRR**

defense gene family via the production of phased, trans-acting siRNAs. *Genes & Development* 2011, **25**(23):2540-2553.

42. Manavella PA, Koenig D, Weigel D: **Plant secondary siRNA production determined by microRNA-duplex structure**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**(7):2461-2466.

43. Axtell MJ, Jan C, Rajagopalan R, Bartel DP: **A two-hit trigger for siRNA biogenesis in plants**. *Cell* 2006, **127**(3):565-577.

44. Xia R, Meyers BC, Liu ZC, Beers EP, Ye SQ, Liu ZR: **MicroRNA superfamilies descended from miR390 and their roles in secondary small interfering RNA biogenesis in Eudicots**. *Plant Cell* 2013, **25**(5):1555-1572.

45. Smith NA, Singh SP, Wang MB, Stoutjesdijk PA, Green AG, Waterhouse PM: **Gene expression - Total silencing by intron-spliced hairpin RNAs**. *Nature* 2000, **407**(6802):319-320.

46. Miki D, Shimamoto K: **Simple RNAi vectors for stable and transient suppression of gene function in rice**. *Plant and Cell Physiology* 2004, **45**(4):490-495.

47. Mette MF, Aufsatz W, van der Winden J, Matzke MA, Matzke AJM: **Transcriptional silencing and promoter methylation triggered by double-stranded RNA**. *Embo Journal* 2000, **19**(19):5194-5201.

48. Dalakouras A, Moser M, Zwiebel M, Krczal G, Hell R, Wassenegger M: **A hairpin RNA construct residing in an intron efficiently triggered RNA-directed DNA methylation in tobacco**. *Plant Journal* 2009, **60**(5):840-851.

49. Senthil-Kumar M, Mysore KS: **New dimensions for VIGS in plant functional genomics**. *Trends in Plant Science* 2011, **16**(12):656-665.

50. Zhang C, Yang CL, Whitham SA, Hill JH: **Development and use of an efficient DNA-based viral gene silencing vector for soybean**. *Molecular Plant-Microbe Interactions* 2009, **22**(2):123-131.

51. Watson JM, Fusaro AF, Wang MB, Waterhouse PM: **RNA silencing platforms in plants**. *Febs Letters* 2005, **579**(26):5982-5987.

52. Burch-Smith TM, Anderson JC, Martin GB, Dinesh-Kumar SP: **Applications and advantages of virus-induced gene silencing for gene function studies in plants**. *Plant Journal* 2004, **39**(5):734-746.

53. Kandoth PK, Heinz R, Yeckel G, Gross NW, Juvale PS, Hill J, Whitham SA, Baum TJ, Mitchum MG: **A virus-induced gene silencing method to study soybean cyst nematode parasitism in *Glycine max***. *BMC research notes* 2013, **6**:255.

54. Buhr T, Sato S, Ebrahim F, Xing AQ, Zhou Y, Mathiesen M, Schweiger B, Kinney A, Staswick P, Clemente T: **Ribozyme termination of RNA transcripts down-regulate seed fatty acid genes in transgenic soybean**. *Plant Journal* 2002, **30**(2):155-163.

55. Wypijewski K, Hornyik C, Shaw JA, Stephens J, Goraczniak R, Gunderson SI, Lacomme C: **Ectopic 5 ' splice sites inhibit gene expression by engaging RNA surveillance and silencing pathways in plants**. *Plant Physiology* 2009, **151**(2):955-965.

56. Akbudak MA, Nicholson SJ, Srivastava V: **Suppression of *Arabidopsis* genes by terminator-less transgene constructs**. *Plant Biotechnology Reports* 2013, **7**(4):415-424.

57. Luo ZH, Chen ZX: **Improperly terminated, unpolyadenylated mRNA of sense transgenes is targeted by RDR6-mediated RNA silencing in *Arabidopsis***. *Plant Cell* 2007, **19**(3):943-958.

58. Park W, Zhai JX, Lee JY: **Highly efficient gene silencing using perfect complementary artificial miRNA targeting AP1 or heteromeric artificial miRNA targeting *AP1* and *CAL* genes**. *Plant Cell Reports* 2009, **28**(3):469-480.

59. Ossowski S, Schwab R, Weigel D: **Gene silencing in plants using artificial microRNAs and other small RNAs**. *Plant Journal* 2008, **53**(4):674-690.

60. Warthmann N, Chen H, Ossowski S, Weigel D, Herve P: **Highly specific gene silencing by artificial miRNAs in rice**. *Plos One* 2008, **3**(3):10.

61. Alvarez JP, Pekker I, Goldshmidt A, Blum E, Amsellem Z, Eshed Y: **Endogenous and synthetic microRNAs stimulate simultaneous, efficient, and localized regulation of multiple targets in diverse species**. *Plant Cell* 2006, **18**(5):1134-1151.

62. McHale M, Eamens AL, Finnegan EJ, Waterhouse PM: **A 22-nt artificial microRNA mediates widespread RNA silencing in *Arabidopsis***. *Plant Journal* 2013, **76**(3):519-529.

63. Montgomery TA, Yoo SJ, Fahlgren N, Gilbert SD, Howell MD, Sullivan CM, Alexander A, Nguyen G, Allen E, Ahn JH *et al*: **AGO1-miR173 complex initiates phased siRNA formation in plants**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(51):20055-20062.

64. de Felippes FF, Wang JW, Weigel D: **MIGS: miRNA-induced gene silencing**. *Plant Journal* 2012, **70**(3):541-547.

65. Hu Z, Jiang QY, Ni ZY, Chen R, Xu S, Zhang H: **Analyses of a *Glycine max* degradome library identify microRNA targets and microRNAs that trigger secondary siRNA biogenesis**. *Journal of Integrative Plant Biology* 2013, **55**(2):160-176.

66. Flores T, Karpova O, Su XJ, Zeng PY, Bilyeu K, Sleper DA, Nguyen HT, Zhang ZJ: **Silencing of *GmFAD3* gene by siRNA leads to low alpha-linolenic acids (18 : 3) of fad3-mutant phenotype in soybean *Glycine max* (Merr.)**. *Transgenic Research* 2008, **17**(5):839-850.

67. Zhang CQ, Ghabrial SA: **Development of Bean pod mottle virus-based vectors for stable protein expression and sequence-specific virus-induced gene silencing in soybean**. *Virology* 2006, **344**(2):401-411.

68. Melito S, Heuberger AL, Cook D, Diers BW, MacGuidwin AE, Bent AF: **A nematode demographics assay in transgenic roots reveals no significant impacts of the *Rhg1* locus LRR-Kinase on soybean cyst nematode resistance**. *Bmc Plant Biology* 2010, **10**:14.

69. Mladenov E, Iliakis G: **Induction and repair of DNA double strand breaks: The increasing spectrum of non-homologous end joining pathways**. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* 2011, **711**(1-2):61-72.

70. Curtin SJ, Voytas DF, Stupar RM: **Genome engineering of crops with designer nucleases**. *Plant Genome* 2012, **5**(2):42-50.

71. Taylor GK, Petrucci LH, Lambert AR, Baxter SK, Jarjour J, Stoddard BL: **LAHEDES: the LAGLIDADG homing endonuclease database and engineering server**. *Nucleic Acids Research* 2012, **40**(W1):W110-W116.

72. D'Halluin K, Chantal V, Jolien VH, Joanna R, Ilse VDB, Anouk P, Kristel DH, Martine B, Derek J, Rene R *et al*: **Targeted molecular trait stacking in cotton through targeted double-strand break induction**. *Plant Biotechnology Journal* 2013:n/a--n/a.

73. Kim YG, Cha J, Chandrasegaran S: **Hybrid restriction enzymes: Zinc finger fusions to *Fok* I cleavage domain**. *Proceedings of the National Academy of Sciences of the United States of America* 1996, **93**(3):1156-1160.

74. Sander JD, Dahlborg EJ, Goodwin MJ, Cade L, Zhang F, Cifuentes D, Curtin SJ, Blackburn JS, Thibodeau-Beganny S, Qi YP *et al*: **Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA)**. *Nature Methods* 2011, **8**(1):67-U94.

75. Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD, Holmes MC: **Highly efficient endogenous human gene correction using designed zinc-finger nucleases**. *Nature* 2005, **435**(7042):646-651.

76. Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD: **Genome editing with engineered zinc finger nucleases**. *Nature Reviews Genetics* 2010, **11**(9):636-646.

77. Zhang F, Maeder ML, Unger-Wallace E, Hoshaw JP, Reyon D, Christian M, Li XH, Pierick CJ, Dobbs D, Peterson T *et al*: **High frequency targeted mutagenesis in *Arabidopsis thaliana* using zinc finger nucleases**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(26):12028-12033.

78. Shukla VK, Doyon Y, Miller JC, DeKelver RC, Moehle EA, Worden SE, Mitchell JC, Arnold NL, Gopalan S, Meng XD *et al*: **Precise genome modification in the crop species *Zea mays* using zinc-finger nucleases**. *Nature* 2009, **459**(7245):437-U156.

79. Townsend JA, Wright DA, Winfrey RJ, Fu FL, Maeder ML, Joung JK, Voytas DF: **High-frequency modification of plant genes using engineered zinc-finger nucleases**. *Nature* 2009, **459**(7245):442-U161.

80. DeFrancesco L: **Move over ZFNs (vol 29, pg 681, 2011)**. *Nature Biotechnology* 2012, **30**(1):1.

81. Moscou MJ, Bogdanove AJ: **A simple cipher governs DNA recognition by TAL effectors**. *Science* 2009, **326**(5959):1501-1501.

82. Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, Bogdanove AJ, Voytas DF: **Targeting DNA double-strand breaks with TAL effector nucleases**. *Genetics* 2010, **186**(2):757-U476.

83. Cermak T, Doyle EL, Christian M, Wang L, Zhang Y, Schmidt C, Baller JA, Somia NV, Bogdanove AJ, Voytas DF: **Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting**. *Nucleic Acids Research* 2011, **39**(12):11.

84. Curtin SJ, Zhang F, Sander JD, Haun WJ, Starker C, Baltes NJ, Reyon D, Dahlborg EJ, Goodwin MJ, Coffman AP *et al*: **Targeted mutagenesis of duplicated genes in soybean with zinc-finger nucleases**. *Plant Physiology* 2011, **156**(2):466-473.

85. Haun W, Coffman A, Clasen BM, Demorest ZL, Lowy A, Ray E, Retterath A, Stoddard T, Juillerat A, Cedrone F *et al*: **Improved soybean oil quality by targeted mutagenesis of the fatty acid desaturase 2 gene family**. *Plant biotechnology journal* 2014, **12**(7):934-940.

86. Hsu PD, Lander ES, Zhang F: **Development and applications of CRISPR-Cas9 for genome engineering**. *Cell* 2014, **157**(6):1262-1278.

87. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG *et al*: **Genome-scale CRISPR-Cas9 knockout screening in human cells**. *Science* 2014, **343**(6166):84-87.

88. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF *et al*: **Evolution and classification of the CRISPR-Cas systems**. *Nature Reviews Microbiology* 2011, **9**(6):467-477.

89. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes**. *Science* 2007, **315**(5819):1709-1712.

90. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E: **A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity**. *Science* 2012, **337**(6096):816-821.

91. Sander JD, Joung JK: **CRIRISPR-Cas systems for editing, regulating and targeting genomes**. *Nature Biotechnology* 2014, **32**(4):347-355.

92. Shan Q, Wang Y, Li J, Zhang Y, Chen K, Liang Z, Zhang K, Liu J, Xi JJ, Qiu J-L *et al*: **Targeted genome modification of crop plants using a CRISPR-Cas system**. *Nat Biotech* 2013, **31**(8):686-688.

93. Cong L, Ran FA, Cox D, Lin SL, Barretto R, Habib N, Hsu PD, Wu XB, Jiang WY, Marraffini LA *et al*: **Multiplex genome engineering using CRISPR/Cas systems**. *Science* 2013, **339**(6121):819-823.

94. Mali P, Yang LH, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM: **RNA-guided human genome engineering via Cas9**. *Science* 2013, **339**(6121):823-826.

95. Wang HY, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R: **One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering**. *Cell* 2013, **153**(4):910-918.

96. Gratz SJ, Cummings AM, Nguyen JN, Hamm DC, Donohue LK, Harrison MM, Wildonger J, O'Connor-Giles KM: **Genome engineering of *Drosophila* with the CRISPR RNA-Guided Cas9 Nuclease**. *Genetics* 2013, **194**(4):1029-+.

97. Nekrasov V, Staskawicz B, Weigel D, Jones JDG, Kamoun S: **Targeted mutagenesis in the model plant *Nicotiana benthamiana* using Cas9 RNA-guided endonuclease**. *Nat Biotech* 2013, **31**(8):691-693.

98. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, Sander JD: **High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells**. *Nat Biotech* 2013, **advance online publication**.

99. Fu YF, Sander JD, Reyon D, Cascio VM, Joung JK: **Improving CRISPR-Cas nuclease specificity using truncated guide RNAs**. *Nature Biotechnology* 2014, **32**(3):279-284.

100. Feng ZY, Mao YF, Xu NF, Zhang BT, Wei PL, Yang DL, Wang Z, Zhang ZJ, Zheng R, Yang L *et al*: **Multigeneration analysis reveals the inheritance, specificity, and patterns of CRISPR/Cas-induced gene modifications in *Arabidopsis***. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(12):4632-4637.

101. Xu R, Li H, Qin R, Wang L, Li L, Wei P, Yang J: **Gene targeting using the *Agrobacterium tumefaciens*-mediated CRISPR-Cas system in rice**. *Rice (New York, NY)* 2014, **7**(1):5.

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

103. Jiang WZ, Zhou HB, Bi HH, Fromm M, Yang B, Weeks DP: **Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in *Arabidopsis*, tobacco, sorghum and rice**. *Nucleic Acids Research* 2013, **41**(20):12.

104. Miao J, Guo DS, Zhang JZ, Huang QP, Qin GJ, Zhang X, Wan JM, Gu HY, Qu LJ: **Targeted mutagenesis in rice using CRISPR-Cas system**. *Cell Research* 2013, **23**(10):1233-1236.

105. Xie K, Yinong Y: **RNA-guided genome editing in plants using a CRISPR-Cas system**. *Molecular Plant* 2013.

106. Liang Z, Zhang K, Chen KL, Gao CX: **Targeted mutagenesis in *Zea mays* using TALENs and the CRISPR/Cas system**. *Journal of Genetics and Genomics* 2014, **41**(2):63-68.

107. Upadhyay SK, Kumar J, Alok A, Tuli R: **RNA-guided genome editing for target gene mutations in wheat**. *G3-Genes Genomes Genetics* 2013, **3**(12):2233-2238.

108. Jia HG, Wang N: **Targeted genome editing of sweet orange using Cas9/sgRNA**. *Plos One* 2014, **9**(4):6.

109. Sugano SS, Shirakawa M, Takagi J, Matsuda Y, Shimada T, Hara-Nishimura I, Kohchi T: **CRISPR/Cas9-mediated targeted mutagenesis in the liverwort *Marchantia polymorpha* L**. *Plant and Cell Physiology* 2014, **55**(3):475-481.

110. Trick HN, Dinkins RD, Santarém ER, Samoyolov R, Meurer C, Walker D, Parrott W, Finer J, Collins G: **Recent advances in soybean transformation**. *Plant Tissue Culture and Biotechnology* 1997, **3**:9-26.

111. Olhoft PM, Flagel LE, Donovan CM, Somers DA: **Efficient soybean transformation using hygromycin B selection in the cotyledonary-node method**. *Planta* 2003, **216**(5):723-735.

112. Collier R, Fuchs B, Walter N, Kevin Lutke W, Taylor CG: ***Ex vitro* composite plants: an inexpensive, rapid method for root biology**. *Plant Journal* 2005, **43**(3):449-457.

113. Mazarei M, Ying Z, Houtz RL: **Functional analysis of the Rubisco large subunit N-epsilon-methyltransferase promoter from tobacco and its regulation by light in soybean hairy roots**. *Plant Cell Reports* 1998, **17**(12):907-912.

114. Preiszner J, VanToai TT, Huynh L, Bolla RI, Yen HH: **Structure and activity of a soybean Adh promoter in transgenic hairy roots**. *Plant Cell Reports* 2001, **20**(8):763-769.

115. Subramanian S, Graham MY, Yu O, Graham TL: **RNA interference of soybean isoflavone synthase genes leads to silencing in tissues distal to the transformation site and to enhanced susceptibility to *Phytophthora sojae***. *Plant Physiology* 2005, **137**(4):1345-1353.

116. Somers DA, Makarevitch I: **Transgene intergration in plants: poking or patching holes in promiscuous genomes?** *Current Opinion in Biotechnology* 2004, **15**(2):126-131.

117. Altpeter F, Baisakh N, Beachy R, Bock R, Capell T, Christou P, Daniell H, Datta K, Datta S, Dix PJ *et al*: **Particle bombardment and the genetic enhancement of crops: myths and realities**. *Molecular Breeding* 2005, **15**(3):305-327.

118. Makarevitch I, Svitashev SK, Somers DA: **Complete sequence analysis of transgene loci from plants transformed via microprojectile bombardment**. *Plant Molecular Biology* 2003, **52**(2):421-432.

119. Svitashev SK, Pawlowski WP, Makarevitch I, Plank DW, Somers DA: **Complex transgene locus structures implicate multiple mechanisms for plant transgene rearrangement**. *Plant Journal* 2002, **32**(4):433-445.

120. Svitashev SK, Somers DA: **Genomic interspersions determine the size and complexity of transgene loci in transgenic plants produced by microprojectile bombardment**. *Genome* 2001, **44**(4):691-697.

121. Aragao FJL, Nogueira E, Tinoco MLP, Faria JC: **Molecular characterization of the first commercial transgenic common bean immune to the Bean golden mosaic virus**. *Journal of Biotechnology* 2013, **166**(1-2):42-50.

122. Jackson SA, Zhang P, Chen WP, Phillips RL, Friebe B, Muthukrishnan S, Gill BS: **High-resolution structural analysis of biolistic transgene integration into the genome of wheat**. *Theoretical and Applied Genetics* 2001, **103**(1):56-62.

123. Kovalic D, Garnaat C, Guo L, Yan YP, Groat J, Silvanovich A, Ralston L, Huang MY, Tian Q, Christian A *et al*: **The use of next generation sequencing and junction sequence analysis bioinformatics to achieve molecular characterization of crops improved through modern biotechnology**. *Plant Genome* 2012, **5**(3):149-163.

124. Liu YG, Mitsukawa N, Oosumi T, Whittier RF: **Efficient isolation and mapping of *Arabidopsis-thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR**. *Plant Journal* 1995, **8**(3):457-463.

125. Yang LT, Wang CM, Holst-Jensen A, Morisset D, Lin YJ, Zhang DB: **Characterization of GM events by insert knowledge adapted re-sequencing approaches**. *Scientific Reports* 2013, **3**:9.

126. Liu YG, Whittier RF: **Thermal asymmetric interlaced PCR - Automatable amplification and sequencing of insert end fragments from p1 and YAC clones for chromosome walking**. *Genomics* 1995, **25**(3):674-681.

127. Orcheski BB, Davis TM: **An enhanced method for sequence walking and paralog mining: TOPO(R) Vector-Ligation PCR**. *BMC research notes* 2010, **3**:61.

128. Ochman H, Gerber AS, Hartl DL: **Genetic applications of an inverse polymerase chain-reaction**. *Genetics* 1988, **120**(3):621-623.

129. Koenning SR, Wrather JA: **Suppression of soybean yield potential in the continental United States from plant diseases estimated from 2006 to 2009**. *Plant Health Prog* 2010.

130. Wrather JA, Koenning SR: **Estimates of disease effects on soybean yields in the United States 2003 to 2005**. *Journal of Nematology* 2006, **38**(2):173-180.

131. Jones JT, Haegeman A, Danchin EGJ, Gaur HS, Helder J, Jones MGK, Kikuchi T, Manzanilla-Lopez R, Palomares-Rius JE, Wesemael WML *et al*: **Top 10 plant-parasitic nematodes in molecular plant pathology**. *Molecular Plant Pathology* 2013, **14**(9):946-961.

132. Niblack TL, Lambert KN, Tylka GL: **A model plant pathogen from the kingdom animalia: *Heterodera glycines*, the soybean cyst nematode**. *Annual Review of Phytopathology* 2006, **44**:283-303.

133. Gheysen G, Mitchum MG: **How nematodes manipulate plant development pathways for infection**. *Current Opinion in Plant Biology* 2011, **14**(4):415-421.

134. Bird DM, Opperman CH, Williamson VM: **Plant infection by root-knot nematode**. In: *Plant Cell Monographs.* Edited by Berg RH, Taylor CG: Springer-Verlag Berlin; 2009: 1-13.

135. Davis EL, Mitchum MG: **Nematodes. Sophisticated parasites of legumes**. *Plant Physiology* 2005, **137**(4):1182-1188.

136. Elling AA, Mitreva M, Gai XW, Martin J, Recknor J, Davis EL, Hussey RS, Nettleton D, McCarter JP, Baum TJ: **Sequence mining and transcript profiling to explore cyst nematode parasitism**. *Bmc Genomics* 2009, **10**:17.

137. Davis EL, Hussey RS, Mitchum MG, Baum TJ: **Parasitism proteins in nematode-plant interactions**. *Current Opinion in Plant Biology* 2008, **11**(4):360-366.

138. Huang GZ, Allen R, Davis EL, Baum TJ, Hussey RS: **Engineering broad root-knot resistance in transgenic plants by RNAi silencing of a conserved and essential root-knot nematode parasitism gene**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(39):14302-14306.

139. Dinh PTY, Zhang LH, Brown CR, Elling AA: **Plant-mediated RNA interference of effector gene *Mc16D10L* confers resistance against *Meloidogyne chitwoodi* in diverse genetic backgrounds of potato and reduces pathogenicity of nematode offspring**. *Nematology* 2014, **16**:669-682.

140. Yang YZ, Jittayasothorn Y, Chronis D, Wang XH, Cousins P, Zhong GY: **Molecular characteristics and efficacy of *16D10* siRNAs in inhibiting root-knot nematode infection in transgenic grape hairy roots**. *Plos One* 2013, **8**(7):13.

141. Yadav BC, Veluthambi K, Subramaniam K: **Host-generated double stranded RNA induces RNAi in plant-parasitic nematodes and protects the host from infection**. *Molecular and Biochemical Parasitology* 2006, **148**(2):219-222.

142. Sindhu AS, Maier TR, Mitchum MG, Hussey RS, Davis EL, Baum TJ: **Effective and specific *in planta* RNAi in cyst nematodes: expression interference of four parasitism genes reduces parasitic success**. *Journal of Experimental Botany* 2009, **60**(1):315-324.

143. Mao YB, Cai WJ, Wang JW, Hong GJ, Tao XY, Wang LJ, Huang YP, Chen XY: **Silencing a cotton bollworm P450 monooxygenase gene by plant-mediated RNAi impairs larval tolerance of gossypol**. *Nature Biotechnology* 2007, **25**(11):1307-1313.

144. Baum JA, Bogaert T, Clinton W, Heck GR, Feldmann P, Ilagan O, Johnson S, Plaetinck G, Munyikwa T, Pleau M *et al*: **Control of coleopteran insect pests through RNA interference**. *Nature Biotechnology* 2007, **25**(11):1322-1326.

145. Nowara D, Gay A, Lacomme C, Shaw J, Ridout C, Douchkov D, Hensel G, Kumlehn J, Schweizer P: **HIGS: Host-induced gene silencing in the obligate biotrophic fungal pathogen *Blumeria graminis***. *Plant Cell* 2010, **22**(9):3130-3141.

146. Steeves RM, Todd TC, Essig JS, Trick HN: **Transgenic soybeans expressing siRNAs specific to a major sperm protein gene suppress *Heterodera glycines* reproduction**. *Functional Plant Biology* 2006, **33**(11):991-999.

147. Youssef RM, Kim KH, Haroon SA, Matthews BF: **Post-transcriptional gene silencing of the gene encoding aldolase from soybean cyst nematode by transformed soybean roots**. *Experimental Parasitology* 2013, **134**(2):266-274.

148. Ibrahim HMM, Alkharouf NW, Meyer SLF, Aly MAM, El-Din A, Hussein EHA, Matthews BF: **Post-transcriptional gene silencing of root-knot nematode in transformed soybean roots**. *Experimental Parasitology* 2011, **127**(1):90-99.

149. Li JR, Todd TC, Oakley TR, Lee J, Trick HN: **Host-derived suppression of nematode reproductive and fitness genes decreases fecundity of *Heterodera glycines* Ichinohe**. *Planta* 2010, **232**(3):775-785.

150. Klink VP, Kim KH, Martins V, MacDonald MH, Beard HS, Alkharouf NW, Lee SK, Park SC, Matthews BF: **A correlation between host-mediated expression of parasite genes as tandem inverted repeats and abrogation of development of female *Heterodera glycines* cyst formation during infection of *Glycine max***. *Planta* 2009, **230**(1):53-71.

151. Rambo-Martin B: **Assessing RNA interference gene targets of root-knot nematodes**. University of Georgia; 2009.

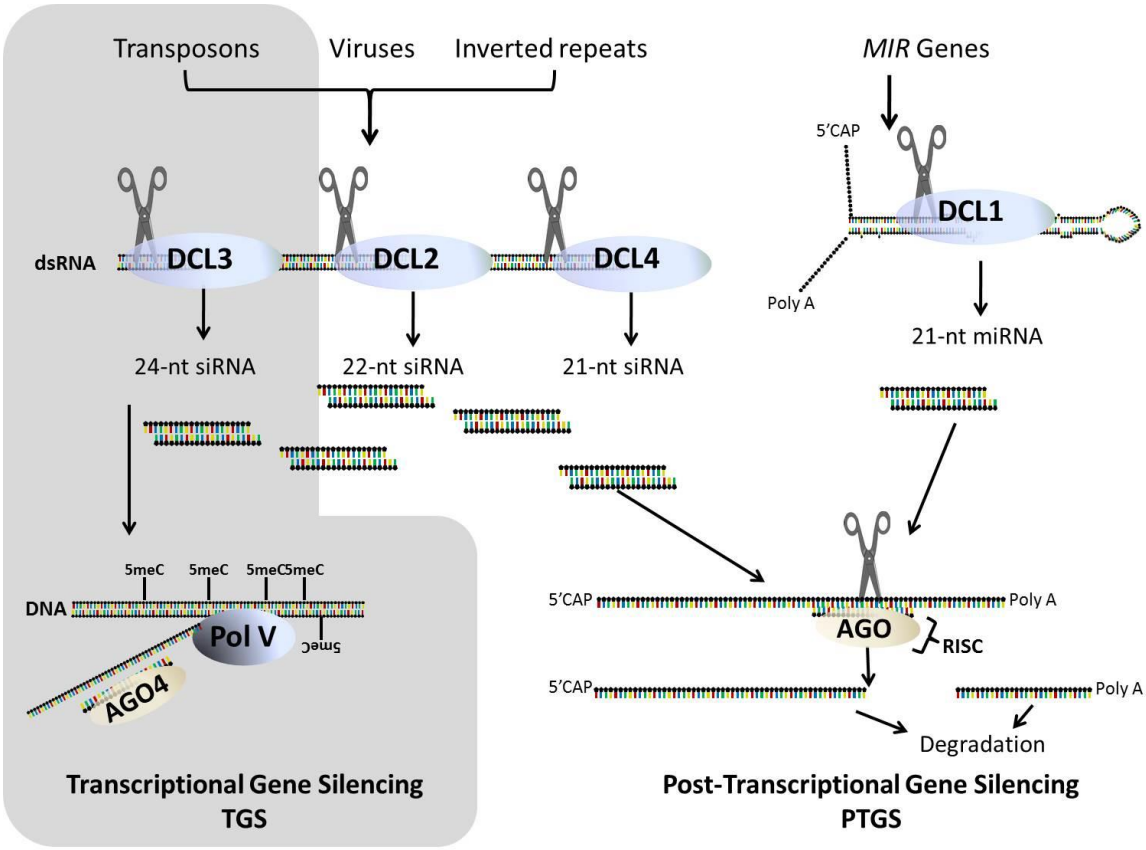


Figure 1.1. General overview of the small RNA pathways in plants

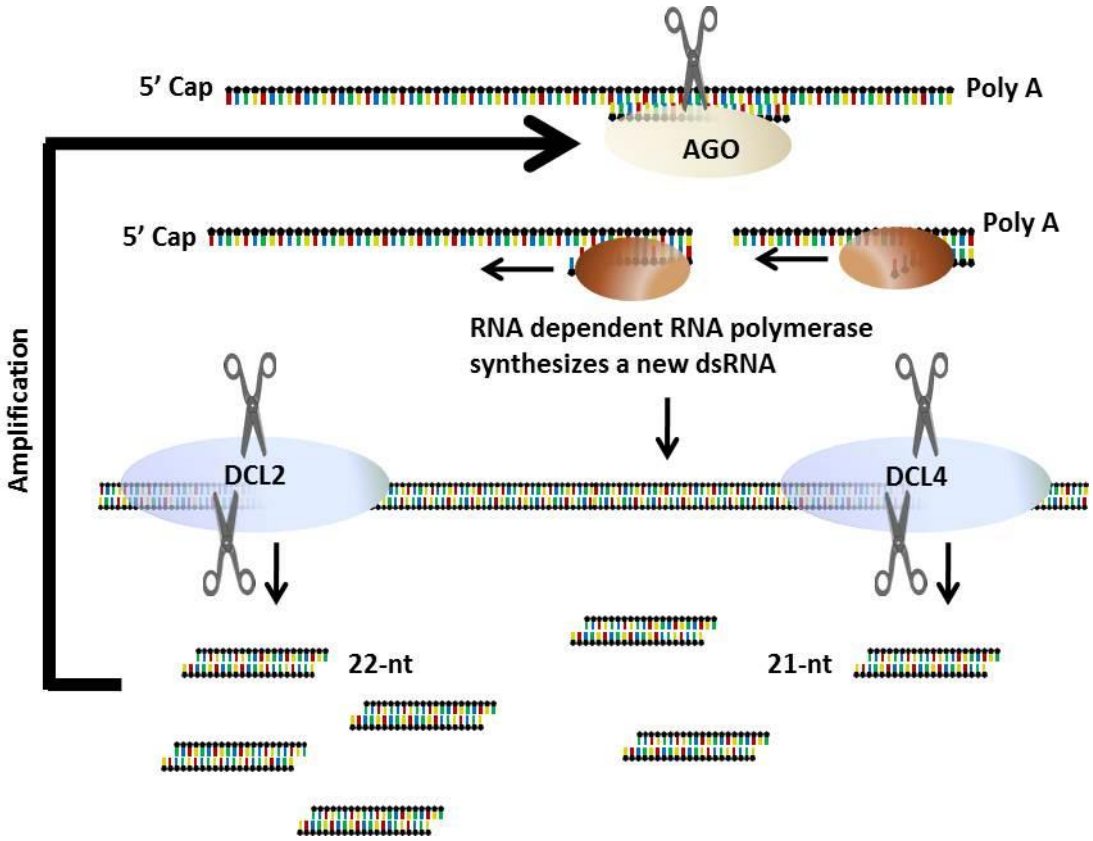


Figure 1.2. Amplification of small RNAs via RNA-dependent RNA polymerase.

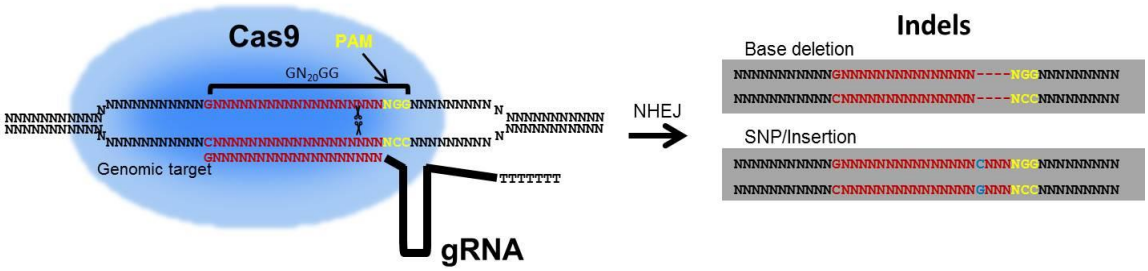


Figure 1.3. Diagram of CRISPR/Cas-mediated DNA double-strand break and repair via non-homologous end joining to produce indels.

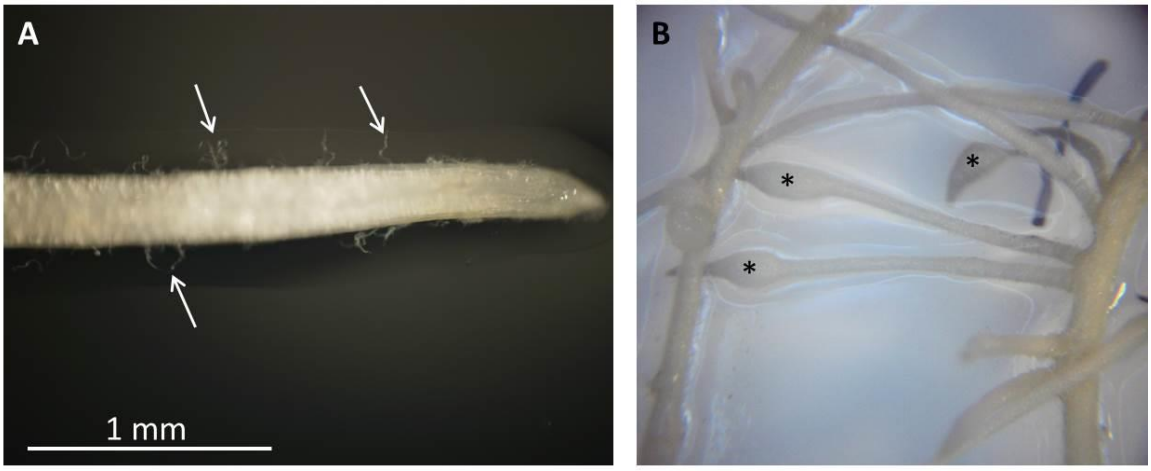


Figure 1.4. Root-knot nematodes infecting soybean hairy roots. A) Image of root-knot nematodes (arrows) surrounding and starting to penetrate a soybean hairy root in tissue culture. B) Characteristic root-knot galls (asterisks) forming on soybean hairy roots in tissue culture.

CHAPTER 2

SIMPLE GENE-SILENCING USING THE TRANS-ACTING SIRNA PATHWAY[1]

Thomas B. Jacobs, Noah J. Lawler, Peter R. LaFayette, Lila O. Vodkin and Wayne A. Parrott. Submitted to *Plant Biotechnology Journal*, 11/18/2014

Abstract

Background

In plants, some micro RNAs (miRNAs) induce the production of a class of small-interfering RNAs (siRNA) called trans-acting-siRNA (ta-siRNA), that lead to gene silencing. Ta-siRNA-inducing miRNAs initiate siRNA production by binding to, and cleaving, specific miRNA target sites upstream of a ta-siRNA template. Since a single miRNA target is sufficient for the production of ta-siRNAs, that target can be incorporated into a vector which induces the production of siRNAs, and ultimately gene-silencing. Ta-siRNAs have so far only been used to initiate gene silencing in arabidopsis. Several ta-siRNA loci have been identified in soybean, but prior to this work, few of the inducing miRNA have been identified, and experimentally validated, much less used to silence genes.

Results

Nine ta-siRNA loci and their respective miRNA targets were identified. The abundance of the ta-siRNA-inducing miRNAs varies dramatically in different tissues. The miRNA targets were experimentally verified by silencing a transgenic *GFP* gene and two endogenous genes in hairy roots and transgenic plants. Small RNAs were produced from both the positive and negative strands, indicating that dsRNA was produced, and processed through the ta-siRNA pathway. A side-by-side experiment showed that the ta-siRNA system is as effective at inducing gene silencing as traditional hairpin vectors. Transgenic plants produced siRNAs, and silencing was observed in the T1 generation.

Conclusions

Our results complement previous reports in arabidopsis by demonstrating that the ta-siRNA system is an efficient way to produce siRNAs and induce gene silencing in soybean. The miRNA targets identified here are simple to incorporate into silencing vectors, and offer an effective and efficient alternative to other gene silencing strategies.

Introduction

Induced gene-silencing in plants is an important tool for gene discovery and for the development of novel agronomic or quality traits, such as virus resistance [1, 2], insect resistance [3], and altered fatty acid compositions [4]. Co-suppresion was first used to describe the silencing of a native chalcone-synthase gene when a transgenic chalcone synthase was over-expressed in petunia plants [5]. It is now known that in plants gene silencing is initiated by a variety of small RNA (sRNA) molecules that are produced from double-stranded RNA (dsRNA) by the action of Dicer-like (DCL) enzymes. Small RNAs are incorporated into Argonaute (AGO) proteins to form RNA-induced silencing complexes (RISCs). The RISCs recognize complementary sequences and induce gene silencing by cleaving mRNA sequences, inhibiting translation, or initiating the methylation of DNA. In some gene-silencing pathways, RNA-dependent RNA polymerases (RDR) synthesize a dsRNA molecule from a cleaved transcript. The dsRNA can then be processed into additional sRNA molecules, thus greatly amplifying the original silencing signal [6].

Small RNAs are generally divided into two classes, miRNA and siRNA. The miRNA pathway starts with the transcription of *MIR* genes that produce primary-

miRNAs which form imperfect, stem-loop structures to create a dsRNA. Primary-miRNAs are processed by DCL1, ultimately resulting in a mature miRNA that is incorporated into AGO1 to form a RISC. The RISC directs the cleavage of mRNAs that are usually degraded by normal cellular processes [7].

Viruses, inverted-repeats, and methylated DNA can all result in long dsRNA molecules that are processed by DCL2, DCL3 and DCL4 to produce 22-nt, 24-nt, and 21-nt siRNAs, respectively. Small-interfering RNAs derived from DCL2 and DCL4 typically participate in post-transcription gene silencing (PTGS) by directing the cleavage of complementary transcripts [6]. DCL3-derived siRNAs direct the DNA-methylation machinery to complementary DNA sequences resulting in transcriptional gene silencing (TGS).

The use of dsRNA has been the cornerstone of current induced gene-silencing strategies reported to date, which rely on the use of hairpin vectors, whereby a portion of the target gene is cloned into a vector in an inverted-repeat orientation and is separated by a loop, which consists of either an intron or a spacer [8, 9]. More recently, artificial miRNAs have been used to induce gene-silencing [10, 11]. Nevertheless, more recent insights into the intricacies and pathways that plants use for gene-silencing makes it clear that the full potential of these pathways to be used for induced gene silencing has barely been tapped, and that it is possible to develop additional gene-silencing systems that are simpler or more efficient than is afforded by current technology.

The trans-acting-siRNA pathway is particularly amenable for use in gene-silencing. This pathway is a combination of the miRNA and siRNA pathways, and has two different mechanisms, referred to as the one-hit, or two-hit models [12]. In the one-

hit model, a single-stranded RNA transcript is first cleaved by AGO1 directed by a miRNA. Unlike the typical miRNA pathway in which the transcript is degraded, RDR6 is recruited to generate a dsRNA molecule from the 3' cleavage product. DCL4 cleaves the dsRNA every 21-nt from the initial miRNA cleavage point, resulting in a phased production of secondary siRNAs (Figure 2.1A)[13]. The term phasiRNA has been used to refer to secondary siRNAs that are produced in a phased pattern. However, the phasiRNA term is not interchangeable with ta-siRNA. Only when phasiRNAs are shown to target a transcript in *trans*, are they referred to as ta-siRNA [14]. PhasiRNAs have been identified in many crops such as tomato [15], tobacco [16], poplar [17], legumes [14] and rice [18]. In soybean, several phasiRNA [14, 19] and five miRNA targets [19] were identified, but not experimentally validated.

Only a handful of miRNAs induce the production of ta-siRNAs, and evidence from arabidopsis suggests that 22-nt miRNAs are sufficient to induce the production of ta-siRNAs, while 21-nt versions of the same miRNAs are not [20, 21]. An additional report has demonstrated that the asymmetry of the initial miRNA/miRNA* duplex, rather than the length of the mature miRNA, is responsible for the recruitment of RDR6 [22]. In the ta-siRNA two-hit model, a transcript with two miR390 target sequences is recognized by AGO7/miR390. The cleaved transcript serves as a template for RDR6 and leads to the production of ta-siRNAs [23]. The two-hit model is conserved in most plant species and is likely ancestral to the one-hit model [24].

The one-hit ta-siRNA pathway has been shown to induce the silencing of targeted genes in arabidopsis [25, 26]. When a vector containing the miRNA target, fused to a portion of the target gene is transformed into a plant cell, siRNAs are produced and

induce gene silencing of the target gene (Figure 2.1B). In the present study, six putative miRNA targets were identified using a sRNA sequencing dataset derived from soybean hairy roots. The putative miRNA targets were confirmed by the induction of gene-silencing of the *GFP* transgene and two endogenous soybean genes. The ta-siRNA system is a simple design given that a single 22-nt miRNA target is all that is required to induce the production of siRNAs from target sequences. The miRNA target sequences identified here extend the ta-siRNA system to soybean and potentially other legumes. These results further demonstrate the effectiveness of the ta-siRNA system for inducing gene-silencing in plants.

Results and discussion

Identification of miRNA targets

The goal of this work was to identify putative miRNA targets in soybean and determine if the sequences can be used to silence genes of interest. It was possible to identify nine putative *phasiRNA* (*PHAS*) loci corresponding to six unique miRNA targets (Table 2.1). Of the nine *PHAS* loci, six were previously identified [14, 19], which suggests the sRNA analysis is correct. All of the putative miRNAs identified here, and corresponding targets, are 22-nt in length, which is consistent with studies in arabidopsis [20, 21]. One putative miRNA, which is most similar to the peanut (*Arachis hypogaea* L.) miRNA, ahy-miR3514, had not been identified in soybean, and potentially targets three *PHAS* loci in soybean hairy roots. The two miRNAs, miR1509a and miR1509b.2, are predicted to target the same miRNA target sequence (Table 2.1). Hu *et al*. [19] also

identified the phasiRNA-inducing soybean miRNAs, miR1510 and miR1514, although they report miR1510 as 21- and 22-nt, and miR1514 as 21-nt.

Differential accumulation of triggering miRNAs across tissue types

Micro RNAs have complex expression profiles depending on developmental stage and environmental stresses [27]. To properly use the putative miRNAs for gene-silencing, it is important to understand their expression profile. To this end, the abundance of the putative miRNAs was evaluated in immature seed tissues (cotyledon, seed coat, and 12-14 day old seed), vegetative tissues (cotyledon from germinating seed, root, stem, shoot tip, and leaf) and hairy roots. The abundance of miR1510a.2 and miR1514a.2 is relatively low (0-20 reads per million sequenced (RPMS)), but consistent levels are present in almost all tissues (Figure 2.2). In contrast, miR1509b.2 is highly abundant (250-400 RPMS) in all tissues except shoot tips, leaves, and hairy roots, and peaks at 900 RPMS in root tissue. MiR5770.2 is only detected in soybean hairy roots at the relatively low level of 4 RPMS. These data clearly demonstrate the differential accumulation of the putative miRNAs across multiple tissue types.

The phasiRNA-inducing miRNAs are size variants of known miRNAs

Many of the putative miRNAs identified are size variants (isomiRs) of previously characterized soybean miRNAs. In miRBase (release 18), miR1514a, miR1509b, miR1510a, and miR5770 are all annotated as 21-nt miRNAs. Four of the putative 22-nt isomiRs identified here could be derived from the primary-miRNAs and are denoted with

the .2 notation. The miR1514, miR1509b, and miR1510a isomiRs have previously been observed in soybean [28].

The miR1510 identified here cannot be produced from the known *miR1510a* and *miR1510b* loci, as a 3' U is not present in the primary-miRNAs. The 3' uridylation of miRNAs is thought to be part of the miRNA degradation pathway in plants [29], but it has been suggested that such modifications could serve other unknown biological functions [30]. It is striking that the same isomiR was observed in such a diverse set of tissues, suggesting that 3' uridylation of miR1510 could be a common post-transcriptional modification.

Consistent and complete silencing of GFP with five miRNA targets

After identifying putative miRNA targets, their ability to induce silencing was tested first by targeting a *GFP* gene. All hairy-root events from the 1509:, 1510:, 1510a.2:, 1514a.2:, and 3514:GFP vectors, were strongly silenced, as indicated by a lack of fluorescence, while events from the GFP controls fluoresced (Figure 2.3). The abundance of GFP mRNA was reduced, on average, 92-96% as compared to the levels from the empty-vector control events (Figure 2.4). The fluorescence of GFP silenced events was comparable to the background autofluorescence in wild-type controls. While the 5770.2:GFP vector did not appear to induce silencing based on fluorescent imaging alone, expression actually was reduced 50%, on average, according to qRT-PCR and fluorescent-protein assays (Figure 2.4). Strong silencing, presumably caused by co-suppression, was observed in one 5770.2:GFP event and in two GFP target-only events, but absent in empty-vector controls.

Small RNA sequencing and mapping of reads to the GFP transgene produced coverage graphs that show the same pattern of sRNAs for each of the silencing vectors (Figure 2.5). In each of the events, tens to thousands of RPMS were produced from both the positive and negative strands. Reads from both strands indicates that dsRNA was produced from the transgenes, and subsequently cleaved by a DCL. There is an even distribution of reads between the positive and negative strands, with a slight bias towards the positive strand (Table 2.2). The 5770.2:GFP event only had 31 RPMS of GFP sRNAs, which is comparable to the background level of 25 RPMS in the empty-vector control. However, the 5770.2:GFP sRNA reads were distributed across the GFP target, similarly to the other silenced events.

The involvement of the ta-siRNA pathway is evident when looking at the reads that map around the miRNA target sequence. When DCL4 cleaves the dsRNA, the first sRNA molecule contains 10-nt from the miRNA target and 11-nt from the 3' sequence (Figure 2.1A). Therefore, the GFP-targeting vectors should produce chimeric reads with 10-nt of the miRNA target and 11-nt GFP. Chimeric reads were observed in all GFP silenced events, but not in the controls or the co-suppressed 5770.2 event (Figure 2.6). These data provide strong evidence that these vectors are using the ta-siRNA pathway, and that while the 5770.2 target is not inducing a strong silencing response, it too is likely participating in the ta-siRNA pathway.

The GFP silencing experiment confirmed that gene silencing and siRNA production could be induced with the identified miRNA targets. The fact that the siRNA patterns are the same for each of the vectors suggests that the same pathway is being used. The accumulation of siRNAs in a consistent pattern suggests that the siRNAs

produced from the transgenic vector can be predicted. In turn, the vector could be designed to avoid or encourage the silencing of closely related genes, depending on the goal. The level of silencing is also uniform and indicates that the ta-siRNA system is an efficient and effective way to induce gene silencing.

Silencing of two soybean genes with the ta-siRNA system

The ta-siRNA system was further tested by attempting to silence two endogenous genes with the miR1514a.2 target. *Glyma02g43860* encodes a nodulation-factor-receptor kinase 1α (NFR) that has been shown to control nodule number in soybean, and null mutants fail to produce nodules [31]. *Glyma07g14460* is a putative cytochrome P450 CYP51G1, and is highly expressed in all tissues based on RNA-seq data [32]. In arabidopsis, null mutants of CYP51G1 are embryo-lethal [33]. Hairy root events should be able to bypass embryo lethality.

All hairy-root events transformed with the 1514a.2:P450 vector had an average reduction of 88% for the P450 transcript, as compared to the empty vector control (Figure 2.7). In contrast, 1514a.2:NFR events had a wide range of expression, from 9-108% of wild type. While the 1514a.2:NFR events had, on average, a 60% reduction in expression compared to the expression of the empty-vector events, they were not significantly different than the NFR target-only events (Figure 2.7). The *NFR* gene is expressed at low levels in hairy root tissues, such that it may be difficult to observe a strong silencing effect for this target gene. NFR is most highly expressed in lateral roots, and is up-regulated by *Bradyrhizobium japonicum* inoculation [31]; therefore it may be necessary to inoculate to detect silencing. By comparison, the P450 gene is highly and

constitutively expressed; therefore, it may be a better reporter for detecting gene silencing.

Small RNA sequencing was performed on three events from each of the 1514a.2:target and control vectors (Table 2.3). Small RNA reads were detected in all of the 1514a.2:target events. Less than ten RPMS were observed in the target-only control events, and these are likely background levels. One NFR target-only event produced a large number of sRNAs (4606 RPMS), which is likely due to co-suppression.

The coverage graphs show that sRNA production mainly occurs 3' from the miR1514a.2 target, and ends in the rbcS terminator (Figures 2.8 and 2.9). Independent events from the same silencing vectors have similar patterns of sRNAs (Figures 2.8 and 2.9). These results demonstrate that the ta-siRNA system can also be used to silence endogenous genes in soybean in an efficient and predictable manner.

Ta-siRNA vectors are as effective as hairpin vectors at inducing gene silencing
A side-by-side experiment was conducted to compare silencing efficiency with hairpin and ta-siRNA vectors in soybean hairy roots. Genotyping for the vector components revealed that many of the hairpin-events were missing one hairpin arm or the other (Table 2.4). Incomplete hairpins should not produce dsRNA, and therefore, not be able to induce gene silencing. While the percentage of missing arms was not consistent between experimental replicates, on average, 18-43% of hairpin events were missing arms, whereas, on average, only 4-8% of the ta-siRNA events were missing the silencing vector (Table 2.4). Events with incomplete hairpins are routinely observed and the frequency can vary considerably for different hairpin vectors and the transformation method used

(data not shown). Yet, there is only one report in the literature describing the loss of hairpin-vector components [34].

Expression analysis from events containing complete vectors demonstrated that the ta-siRNA and hairpin vectors reduced expression levels by approximately 90% for both the GFP and the NFR targets (Figure 2.10). While the 1514a.2:P450 vector reduced expression by 95%, the hairpin-P450 vector reduced expression by 70%, and was not significantly different than the empty-vector control. These results demonstrate that the ta-siRNA vectors are equivalent to hairpin vectors at inducing gene silencing in soybean hairy roots. Given the multiple cloning steps required to produce hairpin vectors (see Methods) and the reduced number of events that contain complete hairpins, ta-siRNA vectors offer a simple and effective alternative.

Ta-siRNA silencing in transgenic plants

Hairy roots are a quick model system to study RNAi phenomena in soybean, but they cannot be used to generate whole plants. To determine if the ta-siRNA system is effective in whole plants, the vectors 1514a.2:NFR and 1514a.2:P450 were transformed into soybean. Small RNA sequencing of T0 leaf tissue from three 1514a.2:P450 events confirmed the production of siRNAs (Table 2.5). Small RNA reads mapped to the P450 mRNA show the same pattern of siRNAs that was observed in hairy roots (Figure 2.11).

In the T1 generation, a range of gene silencing was observed in lines containing the 1514a.2:P450 vector. Five lines have a consistent, although non-significant, reduction in P450 expression of 30-50% (Figure 2.12, lines 24, 11, 4, 47, and 18). Small RNA sequencing of the 1514a.2:P450 lines shows that siRNAs were produced in all of

the sequenced events, and the abundances ranged from 56 RPMS to 5,866 RPMS (Table 2.6). No correlation between siRNA abundance and level of silencing was observed. Line 11 had the second lowest P450 expression, while lines 7 and 33 had approximately 500-1800 more RPMS than line 11. A stronger level of silencing was observed in the soybean hairy roots than in the leaves of T1 plants. Since P450 null mutants are embryo lethal, the expectation is therefore that any plant recovered from somatic embryos or seed will need to have attenuated expression, as opposed to complete elimination of expression. Alternatively, the different levels of silencing could be due to differences in tissue type (leaf vs. hairy root).

Expression analysis and nodule counts were performed on three 1514a.2:NFR lines at the T1 generation. The qRT-PCR showed that line 9 consistently expressed the silencing construct, line 2 did not express at all, and only one plant from line 10 expressed the silencing construct (data not shown). Consistent with the hairy-root data, NFR expression was significantly reduced by 48% in line 9 for rep 1, but there was no significant difference in expression for rep 2 (Figure 2.13). No reduction in nodule numbers was observed (data not shown). The two NFR mutants, *rj1* and *nod49*, are recessive, loss-of-function mutants [31], so while line 9 had reduced levels of NFR, it may not be low enough to recapitulate the phenotype. SRNA sequencing was performed on the 1514a.2:NFR lines and, as expected, sRNAs targeting the NFR gene were detected in line 9 and the expressing line 10 plant, in patterns similar to those observed in hairy roots (Table 2.7). In hindsight, the selection of P450 and NFR as targets was not the best choice for a proof-of-concept experiment in transgenic plants. However, these data do

support that the ta-siRNA vectors can be used to make siRNAs that induce gene silencing.

Practical Considerations

The ta-siRNA silencing vectors are dependent on the phasiRNA-inducing miRNA. To use this system in a different plant species, the miRNAs and corresponding targets must be identified. A number of *PHAS* loci have been described in other plant species, so it should be simple to identify the inducing miRNAs and produce silencing vectors in a similar manner. Furthermore, the sRNA deep-sequencing results suggest that the putative soybean miRNAs have some level of tissue specificity. The tissue-specific profile for each of the different miRNAs will need to be characterized to ensure silencing in the tissue type of interest.

The lack of inverted repeats in ta-siRNA vectors makes them very easy to construct and amenable to high-throughput cloning. To facilitate vector construction, binary silencing vectors have been made that contain a multiple-cloning site next to each of the six reported miRNA targets. This vector series, called pGmute, is available from Addgene (plasmids 47025, 55768-55772).

Conclusions

The ta-siRNA silencing system is an effective and efficient gene-silencing tool for plant research. Here, six soybean miRNA targets were identified that can be used to induce ta-siRNA for a specific target in a transgenic system. Incorporating the miRNA targets into transgenic vectors led to consistent gene-silencing in hairy roots and

transgenic soybean plants. The simple vector design makes cloning and transformation straightforward. The predictable patterns of sRNAs that are produced and the specific expression profiles of the miRNA triggers may add a level of control not as easy to achieve with other gene-silencing systems.

Materials and Methods

Identification of putative ta-siRNA loci

Putative *PHAS* loci were identified by aligning small RNA sequences from a soybean cultivar 'Jack' hairy root library to the soybean genome (Glyma v1.0) using Geneious version 5.4 [35]. Regions with a high coverage of reads were evaluated by a number of criteria. First the sRNA reads had to be primarily 21 nt in length, consistent with DCL4 activity. Second, reads had to be from both positive and negative strands, which indicates the presence of dsRNA. Finally, a single species of sRNA (presumably a miRNA) had to align 5' to the majority of the aligned sequences, as such alignment can indicate the possibility of a miRNA inducing the production of the siRNAs. To get putative miRNAs to align to the reference sequence, the miRNA alignments allowed for mismatches. *PHAS* loci were also identified by aligning all miRNA from the Fabaceae deposited in miRBase release 18.0. Regions with sRNAs adjacent to possible miRNA cleavage sites were evaluated following the criteria outlined above. Once putative miRNA targets were identified, they were compared to the soybean genome via BLAST to identify other potential *PHAS* loci. Finally, sRNA reads from previously published datasets [36] were mapped to the 22-nt putative miRNAs using an in-house Bowtie script to test for tissue-specific expression

Vector Design

The binary vector p201N uses an *npt*II selectable cassette under the control of the *Solanum tuberosum* Ubi3 promoter and terminator [37], and was used for all hairy root experiments. The GmUbi promoter [38] drives the expression of the silencing and control cassettes, and is terminated by the *Pisum sativum* rbcS terminator [39]. Gene-silencing targets were created by amplifying the target regions from cDNA or DNA of the gene to be silenced with the miR1514a.2 target fused to the forward primer (Table 2.8). Control vectors were created with the same primers, except with the miRNA target sequence omitted. The amplicons were inserted into the p201N vector via its *Asc*I and *Avr*II restriction sites. Additional miRNA targets were fused to the GFP target using the same cloning scheme.

For hairpin vectors, a soybean *FAD3* intron [40] was inserted between the GmUbi and rbcs terminator. The *FAD3* intron is flanked by *Asc*I and *Swa*I restriction sites on the 5' end, and *Bam*HI and *Avr*II on the 3' end. Gene targets were amplified with forward primers with an *Asc*I and *Avr*II 5' tail, and reverse primers with a *Swa*I and *Bam*HI 5' tail. The hairpin arms were added in two successive cloning steps. First, the 3' arm of the hairpin was inserted between the rbcS terminator and FAD3 intron with the *Avr*II and *Bam*HI restriction sites. The 5' arm was then added with the *Asc*I and *Swa*I restriction sites. Due to an internal *Bam*HI site within the NFR target, the *Bgl*II restriction site was used in place of the *Bam*HI site in the NFR reverse primer since *Bgl*II and *Bam*HI produce identical DNA overhangs.

For biolistic transformation of soybean, a pSMART HC Kan (Lucigen Corporation, Middleton WI, accession number AF532107) cloning vector was modified

to contain a *hygromycin phosphotransferase* gene under the control of the Ubi3 promoter and terminator and the meganuclease I-*Ppo*I site and is referred to as pSPH2. The gene-silencing cassettes 1514a.2 P450 and 1514a.2:NFR were moved into pSPH2 as an I-*Ppo*I fragment. Prior to bombardment, the vectors were digested with *Pac*I to release the hygromycin resistance and gene-silencing cassettes from the vector backbone.

The pGmute series of vectors were created by inserting annealed oligos into the p201N vector using *Asc*I and *Bam*HI restriction sites. For each miRNA target, oligo pairs (Table 2.8) were designed such that after denaturing and re-annealing, inserts contain the miRNA target with the *Asc*I, *Bam*HI, and *Avr*II multiple cloning site on the 3' end. The vectors are available from Addgene (plasmids 47025, 55768-55772).

Hairy root transformation of soybean

Soybean Jack seeds were transformed with *A. rhizogenes* strain K599 [41] with slight modifications from the protocol previously described [41]. Briefly, soybean seeds were germinated for approximately one week under sterile conditions on a filter paper wetted with ½ MSO liquid medium (½ MS salts, B5 vitamins, 30 g L^{-1} sucrose) [42]. *Agrobacterium* from glycerol stocks was streaked out on YM medium [43] supplemented with 50 mg L^{-1} kanamycin and grown for two days at 28°C. The *Agrobacterium* was re-suspended in 600 µL of phosphate buffer (PB, 0.01 M Na2HPO4, 0.15M NaCl, pH 7.5) + 100 µM acetosyringone to an $O.D._{600}$ of 0.5-0.8. Soybean cotyledons were prepared similarly as for cot-node transformation [44]; the root and lower hypocotyl was removed from the cotyledons, leaving approximately 5 mm of hypocotyl. The apical shoot and hypocotyl were cut longitudinally to produce two cotyledons with a short hypocotyl

piece. The meristem was removed, and 1-mm deep cuts were made on the adaxial surface of the cotyledons, using a scalpel dipped in the solution of *Agrobacterium*. Cotyledons were co-cultivated with the *Agrobacterium* for three days on filter paper wetted with 2 mL of ½ MSO + 100 µM Acetosyringone. Cotyledons were then transferred according to Cho et al. [41] onto medium consisting of : ½ MS salts, 2 g L^{-1} Phytagel (Sigma-Aldrich Co.), and 500 mg L^{-1} timentin. Each root was treated as an individual event and transferred to ½ MSO solid medium with 10 mg L^{-1} of Geneticin (G418). No-vector control roots were grown on ½ MSO medium without G418. Those roots that grew on ½ MSO + G418 medium were considered events, and a 2-cm portion of a root tip was collected for CTAB DNA extraction [45]. PCR was performed to confirm the insertion of the gene of interest using primers in the promoter and terminator (Table 2.8). After selection and PCR verification, roots were grown on individual plates for two weeks. Root tissue was then harvested, frozen in liquid nitrogen, and stored at -80°C.

RNA Isolation

Tissues were harvested from culture plates or plants in the greenhouse, and immediately frozen in liquid nitrogen. Samples were ground with a mortar and pestle in liquid nitrogen or with a stainless-steel BB, with extraction buffer, in a 2010 Geno/Grinder® (Spex Sample Prep); 100 mg of tissue were used for each extraction. Total RNA was extracted with Tri-Reagent® (Life Technologies™) according to the manufacturer's instructions.

qRT-PCR

Total RNA was treated with Turbo DNase™ (Ambion®) to remove all contaminating DNA. Then, 100 ng of treated total RNA was used as the template in the Go Taq® 1-Step RT-qPCR system (Promega). The qRT-PCR reactions were performed in triplicate in a Light Cycler 480II (Roche Diagnostics) using : 37°C for 15 minutes; 95°C for 10 minutes; 40 cycles (95°C for 10 seconds; 60°C for 30 seconds; 72°C for 30 seconds); a melt-curve analysis from 60°C to 95° at a ramp rate of 0.11°C second^{-1}. The melt-curve analysis was used to confirm the specificity of the qRT-PCR reaction. Next, qRT-PCR amplicons were sequenced to confirm that it was the target gene that was amplified. The metalloprotease amplicon [46] was used to normalize expression of each of the target genes. A list of primers used and amplicon efficiencies can be found in Table 2.8. Finally, ΔCt values for each event were calculated by the LightCycler® 480 SW 1.5.1 program using the Advanced Relative Quantification analysis.

GFP protein quantification

At the time of harvest, 100 mg of root tissues were ground in a 2 mL microcentrifuge tube in PB. Samples were stored at -80. GFP was quantified using a Synergy 2 plate reader (BioTek Instruments, Inc.) according to [47]. Protein concentrations were measured by a Bradford assay [48]. Raw relative fluorescent units (RFU) were normalized by protein concentration.

GFP Imaging

After selection on ½ MSO + G418, root tips were imaged with an Olympus MVX10 microscope with a GFP filter cube and DP controller version 2.2.1.227 (Olympus America Inc.) imaging software. Blue-light images were taken with a 5 ms exposure.

Small RNA sequencing and Assembly

Small RNA libraries were prepared from 1µg of total RNA using Illumina's TruSeq™ small RNA library kit. The MiSeq was used for small RNA sequencing according to the manufacturer's instructions. Raw reads were separated by barcodes using the MiSeq Reporter software. Fastq files were imported into Geneious, which was used trim the adapters and select reads 18-25 nt in length for assembly. The reads were assembled to the respective silencing or control cassettes and to the target mRNA using the following parameters: Gaps not allowed, word length 18, index word length 13, ignore words repeated more than 5 times, maximum mismatches per read 0%, and maximum ambiguity 4.

Statistical Analysis

The ΔCt values were normalized to the mean ΔCt values of the empty-vector control events (ΔΔCt method). Protein quantification data were processed likewise. The normalized qRT-PCR and protein quantification values were analyzed with JMP® Pro 9.0.2 (SAS Institute Inc.), using the Fit Y by X function, with the vector as the X variable and the normalized values as the Y response variable(s). A one-way ANOVA was

performed to ensure a significant difference between the results from the different vectors. Means were separated using Student's T-test with $\alpha = 0.05$ for equal samples numbers or Tukey-Kramer HSD with $\alpha = 0.05$ for unequal sample numbers.

Biolistic Transformation of Soybean

Biolistic transformation of soybean was performed as previously described [49]. PCR was used to confirm the presence of transgenes in T0 plants. Small RNA analysis was performed on three independent 1514a.2:P450 events. RNA was extracted from immature leaves of T0 plants growing in the greenhouse.

Nodulation Assay

T1 seeds were germinated in 32-cell pack trays containing a mix of 3B (Fafard), field soil, and sand. After 2-3 weeks, seedlings were removed from the trays, and the roots were washed and nodules counted. Tissue from lateral roots was collected for RNA extraction. Leaf tips were taken for DNA extraction. Null segregants for the transgene were removed from the analysis. Two reps were performed approximately one month apart and each rep was processed on a single day.

1514a.2:P450 transgenic plants

Young, unexpanded leaves were taken from T0 1514a.2:P450 plants growing in the greenhouse and used for RNA extraction. T1 seeds from transgenic lines and 'Jack' controls were germinated in the lab under sterile conditions and transplanted to the greenhouse. After 3 weeks of growth, young, unexpanded leaves were taken for DNA

and RNA extractions. RNA was extracted from two T1 plants per line that tested positive for the transgene.

References

1. Tricoli DM, Carney KJ, Russell PF, McMaster JR, Groff DW, Hadden KC, Himmel PT, Hubbard JP, Boeshore ML, Quemada HD: **Field-evaluation of transgenic squash containing single or multiple virus coat protein gene constructs for resistance to cucumber mosaic-virus**. *Bio-Technology* 1995, **13**(13):1458-1465.

2. Tennant P, Fermin G, Fitch MM, Manshardt RM, Slightom JL, Gonsalves D: **Papaya ringspot virus resistance of transgenic rainbow and SunUp is affected by gene dosage, plant development, and coat protein homology**. *European Journal of Plant Pathology* 2001, **107**(6):645-653.

3. Baum JA, Bogaert T, Clinton W, Heck GR, Feldmann P, Ilagan O, Johnson S, Plaetinck G, Munyikwa T, Pleau M *et al*: **Control of coleopteran insect pests through RNA interference**. *Nature Biotechnology* 2007, **25**(11):1322-1326.

4. Flores T, Karpova O, Su XJ, Zeng PY, Bilyeu K, Sleper DA, Nguyen HT, Zhang ZJ: **Silencing of Gm*FAD3* gene by siRNA leads to low alpha-linolenic acids (18 : 3) of fad3-mutant phenotype in soybean *Glycine max* (Merr.)**. *Transgenic Research* 2008, **17**(5):839-850.

5. Napoli C, Lemieux C, Jorgensen R: **Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in *trans***. *Plant Cell* 1990, **2**(4):279-289.

6. Eamens A, Curtin SJ, Waterhouse PM: **RNA Silencing in Plants**: Springer; 2010.

7. Voinnet O: **Origin, biogenesis, and activity of plant microRNAs**. *Cell* 2009, **136**(4):669-687.

8. Wesley SV, Helliwell CA, Smith NA, Wang MB, Rouse DT, Liu Q, Gooding PS, Singh SP, Abbott D, Stoutjesdijk PA *et al*: **Construct design for efficient, effective and high-throughput gene silencing in plants**. *Plant Journal* 2001, **27**(6):581-590.

9. Miki D, Shimamoto K: **Simple RNAi vectors for stable and transient suppression of gene function in rice**. *Plant and Cell Physiology* 2004, **45**(4):490-495.

10. Ossowski S, Schwab R, Weigel D: **Gene silencing in plants using artificial microRNAs and other small RNAs**. *Plant Journal* 2008, **53**(4):674-690.

11. Warthmann N, Chen H, Ossowski S, Weigel D, Herve P: **Highly specific gene silencing by artificial miRNAs in rice**. *Plos One* 2008, **3**(3):10.

12. Fei QL, Xia R, Meyers BC: **Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks**. *Plant Cell* 2013, **25**(7):2400-2415.

13. Yoshikawa M, Peragine A, Park MY, Poethig RS: **A pathway for the biogenesis of trans-acting siRNAs in *Arabidopsis***. *Genes & Development* 2005, **19**(18):2164-2175.

14. Zhai JX, Jeong DH, De Paoli E, Park S, Rosen BD, Li YP, Gonzalez AJ, Yan Z, Kitto SL, Grusak MA *et al*: **MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs**. *Genes & Development* 2011, **25**(23):2540-2553.

15. Shivaprasad PV, Chen HM, Patel K, Bond DM, Santos B, Baulcombe DC: **A MicroRNA superfamily regulates nucleotide binding site-leucine-rich repeats and other mRNAs**. *Plant Cell* 2012, **24**(3):859-874.

16. Li F, Pignatta D, Bendix C, Brunkard JO, Cohn MM, Tung J, Sun HY, Kumar P, Baker B: **MicroRNA regulation of plant innate immune receptors**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**(5):1790-1795.

17. Klevebring D, Street NR, Fahlgren N, Kasschau KD, Carrington JC, Lundeberg J, Jansson S: **Genome-wide profiling of populus small RNAs**. *BMC Genomics* 2009, **10**:620.

18. Song XW, Li PC, Zhai JX, Zhou M, Ma LJ, Liu B, Jeong DH, Nakano M, Cao SY, Liu CY *et al*: **Roles of *DCL4* and *DCL3b* in rice phased small RNA biogenesis**. *Plant Journal* 2012, **69**(3):462-474.

19. Hu Z, Jiang QY, Ni ZY, Chen R, Xu S, Zhang H: **Analyses of a *Glycine max* degradome library identify microRNA targets and microRNAs that trigger secondary siRNA biogenesis**. *Journal of Integrative Plant Biology* 2013, **55**(2):160-176.

20. Chen HM, Chen LT, Patel K, Li YH, Baulcombe DC, Wu SH: **22-nucleotide RNAs trigger secondary siRNA biogenesis in plants**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(34):15269-15274.

21. Cuperus JT, Carbonell A, Fahlgren N, Garcia-Ruiz H, Burke RT, Takeda A, Sullivan CM, Gilbert SD, Montgomery TA, Carrington JC: **Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in *Arabidopsis***. *Nature Structural & Molecular Biology* 2010, **17**(8):997-U111.

22. Manavella PA, Koenig D, Weigel D: **Plant secondary siRNA production determined by microRNA-duplex structure**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**(7):2461-2466.

23. Axtell MJ, Jan C, Rajagopalan R, Bartel DP: **A two-hit trigger for siRNA biogenesis in plants**. *Cell* 2006, **127**(3):565-577.

24. Xia R, Meyers BC, Liu ZC, Beers EP, Ye SQ, Liu ZR: **MicroRNA superfamilies descended from miR390 and their roles in secondary small interfering RNA biogenesis in Eudicots**. *Plant Cell* 2013, **25**(5):1555-1572.

25. de Felippes FF, Wang JW, Weigel D: **MIGS: miRNA-induced gene silencing**. *Plant Journal* 2012, **70**(3):541-547.

26. de Felippes FF, Ott F, Weigel D: **Comparative analysis of non-autonomous effects of tasiRNAs and miRNAs in *Arabidopsis thaliana***. *Nucleic Acids Res* 2011, **39**(7):2880-2889.

27. Carthew RW, Sontheimer EJ: **Origins and mechanisms of miRNAs and siRNAs**. *Cell* 2009, **136**(4):642-655.

28. Kulcheski FR, de Oliveira LFV, Molina LG, Almerao MP, Rodrigues FA, Marcolino J, Barbosa JF, Stolf-Moreira R, Nepomuceno AL, Marcelino-Guimaraes FC *et al*: **Identification of novel soybean microRNAs involved in abiotic and biotic stresses**. *Bmc Genomics* 2011, **12**:17.

29. Li JJ, Yang ZY, Yu B, Liu J, Chen XM: **Methylation protects miRNAs and siRNAs from a 3 '-end uridylation activity in *Arabildopsis***. *Current Biology* 2005, **15**(16):1501-1507.

30. Wyman SK, Knouf EC, Parkin RK, Fritz BR, Lin DW, Dennis LM, Krouse MA, Webster PJ, Tewari M: **Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity**. *Genome Research* 2011, **21**(9):1450-1461.

31. Indrasumunar A, Searle I, Lin MH, Kereszt A, Men A, Carroll BJ, Gresshoff PM: ***Nodulation factor receptor kinase 1 alpha*** **controls nodule organ number in soybean (*Glycine max* L. Merr)**. *Plant Journal* 2011, **65**(1):39-50.

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

33. O'Brien M, Chantha SC, Rahier A, Matton DP: **Lipid signaling in plants. Cloning and expression analysis of the obtusifoliol 14alpha-demethylase from *Solanum chacoense* Bitt., a pollination- and fertilization-induced gene with both obtusifoliol and lanosterol demethylase activity**. *Plant Physiol* 2005, **139**(2):734-749.

34. Sunitha S, Shivaprasad PV, Sujata K, Veluthambi K: **High frequency of T-DNA deletions in transgenic plants transformed with intron-containing hairpin RNA genes**. *Plant Molecular Biology Reporter* 2012, **30**(1):158-167.

35. Drummond A, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S *et al*: **Geneious**. In., 5.4 edn; 2011.

36. Zabala G, Campos E, Varala KK, Bloomfield S, Jones SI, Win H, Tuteja JH, Calla B, Clough SJ, Hudson M *et al*: **Divergent patterns of endogenous small RNA populations from seed and vegetative tissues of *Glycine max***. *Bmc Plant Biology* 2012, **12**:26.

37. Joshi M, Niu C, Fleming G, Hazra S, Chu Y, Nairn CJ, Yang HY, Ozias-Akins P: **Use of green fluorescent protein as a non-destructive marker for peanut genetic transformation**. *In Vitro Cellular & Developmental Biology-Plant* 2005, **41**(4):437-445.

38. Hernandez-Garcia CM, Martinelli AP, Bouchard RA, Finer JJ: **A soybean (*Glycine max*) polyubiquitin promoter gives strong constitutive expression in transgenic soybean**. *Plant Cell Reports* 2009, **28**(5):837-849.

39. An G, Watson BD, Stachel S, Gordon MP, Nester EW: **New cloning vehicles for transformation of higher-plants**. *Embo Journal* 1985, **4**(2):277-284.

40. Siminszky B, Gavilano L, Bowen SW, Dewey RE: **Conversion of nicotine to nornicotine in *Nicotiana tabacum* is mediated by *CYP82E4,* a cytochrome P450 monooxygenase**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(41):14919-14924.

41. Cho HJ, Farrand SK, Noel GR, Widholm JM: **High-efficiency induction of soybean hairy roots and propagation of the soybean cyst nematode**. *Planta* 2000, **210**(2):195-204.

42. Murashige T, Skoog F: **A Revised medium for rapid growth and bio assays with tobacco tissue cultures**. *Physiologia Plantarum* 1962, **15**(3):473-497.

43. Lin JJ: **Optimization of the transformation efficiency of *Agrobacterium-tumefaciens* cells using electroporation.** *Plant Science* 1994, **101**(1):11-15.

44. Olhoft PM, Flagel LE, Donovan CM, Somers DA: **Efficient soybean transformation using hygromycin B selection in the cotyledonary-node method**. *Planta* 2003, **216**(5):723-735.

45. Murray MG, Thompson WF: **Rapid isolation of high molecular-weight plant DNA**. *Nucleic Acids Research* 1980, **8**(19):4321-4325.

46. Libault M, Thibivilliers S, Bilgin DD, Radwan O, Benitez M, Clough SJ, Stacey G: **Identification of four soybean reference genes for gene expression normalization**. *The Plant Genome* 2008, **1**(1):44-54.

47. Remans T, Schenk PM, Manners JM, Grof CPL, Elliott AR: **A protocol for the fluorometric quantification of mGFP5-ER and sGFP(S65T) in transgenic plants**. *Plant Molecular Biology Reporter* 1999, **17**(4):385-395.

48. Bradford MM: **Rapid and sensitive method for quantitation of microgram quantities of protein utilizing principle of protein-dye binding**. *Analytical Biochemistry* 1976, **72**(1-2):248-254.

49. Hancock CN, Zhang F, Floyd K, Richardson AO, LaFayette P, Tucker D, Wessler SR, Parrott WA: **The rice miniature inverted repeat transposable element *mPing* is an effective insertional mutagen in soybean**. *Plant Physiology* 2011, **157**(2):552-562.

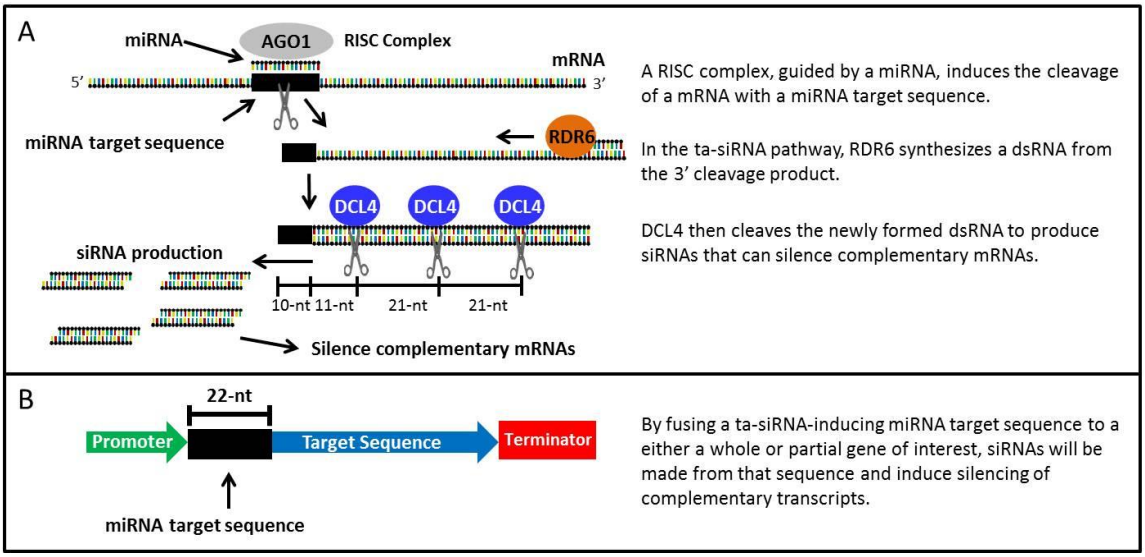


Figure 2.1. Outline of the ta-siRNA silencing pathway and how to adapt it to a transgenic system. (A) The general mechanism of the ta-siRNA pathway that leads to the production of siRNAs. (B) How the ta-siRNA mechanism can be co-opted for the production of siRNAs, and ultimately gene silencing.

miRNA Name	Related miRNA	Putative miRNA 5′ -> 3′	miRNA Target 5′ -> 3′	*PHAS* Locus Glyma model	*PHAS* Locus Characterization	*PHAS* Locus Coordinates
3514	Ahy-miR3514-3p	UCACCAUUAAGACAGAGACCUU	AAGGTCTCTGTCTTAATGGTGA	intergenic	uncharacterized, similar to PPR protein	Gm16:30,415,404..30,421,646
				Glyma18g03980	3' UTR, protein of unknown function	Gm18:2,747,299..2,748,817
				intergenic	*MIR4409* and *MIR5372*	Gm16:5,743,468..5,745,496
5770.2	Gma-miR5770b	UAGGACUAUGGUUUGGACAAGU	TCTTGTCCAAACCATAGTCCAA	Glyma17g02260	copper amino oxidase	Gm17:1,453,405..1,456,485
1510a.2	Gma-miR1510a-3p	UUGUUUUACCUAUUCCACCCAU	ATGGGTGGAATAGGGAAAACAA	Glyma04g39740	LRR protein	Gm04:45,906,669..45,909,231
1509a	Gma-miR1509a	UUAAUCAAGGAAAUCACGGUCG	CAACCTTGATTTCCTTGATTAA	intergenic	no information	Gm04:43,624,071..43,624,750
1509b.2	Gma-miR1509b	UUAAUCAAGGAAAUCACGGUUG				
1514a.2	Gma-miR1514a	UUCAUUUUUAAAAUAGGCAUUG	CAATGCCTATTTTAGAAATGAA	Glyma16g01940	No Apical Meristem protein, uncharacterized	Gm16:1,443,765..1,449,622
				Glyma07g05360	No Apical Meristem protein, uncharacterized	Gm07:3,999,218..4,001,905
1510	Gma-miR 1510a-3p	UGUUGUUUUACCUAUUCCACCU	AGGTGGAATAGGAAAAACAACT	Glyma12g27800	LRR protein	Gm12:31,264,569..31,268,587

Table 2.1. Summary of putative miRNAs and target sequences.

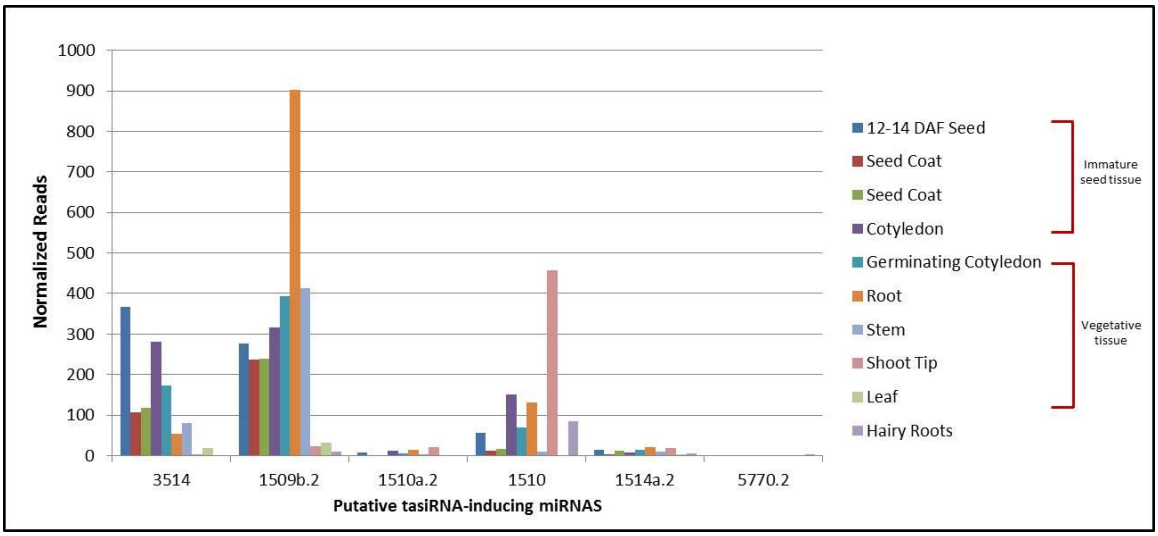


Figure 2.2. Relative abundance of the 22-nt miRNAs in different soybean tissues. The number of miRNA reads are normalized to the millions of sequenced reads sequenced (RPMS).

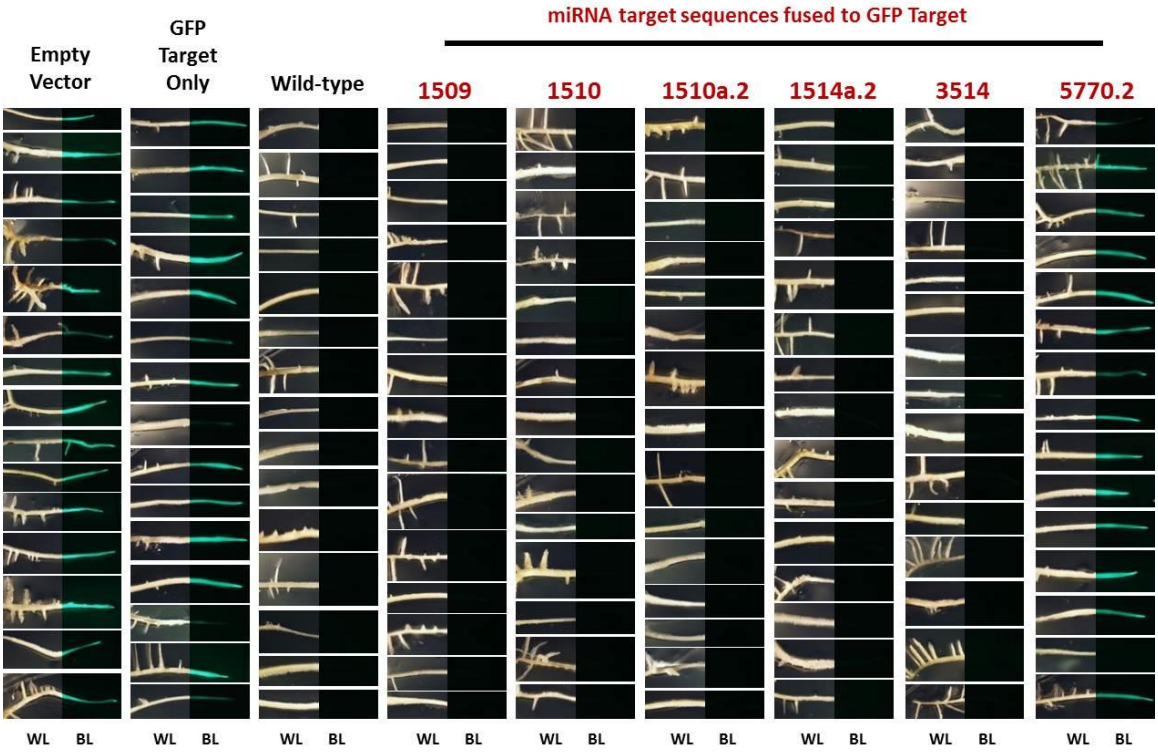


Figure2.3. Gene silencing of GFP in hairy roots. Silencing is evident by roots that do not fluoresce under blue light (BL). Each root is an individual transgenic event. Some silencing is observed in the control GFP-Target-Only treatment, but complete silencing is apparent in five of the six miRNA recognition sequences used.

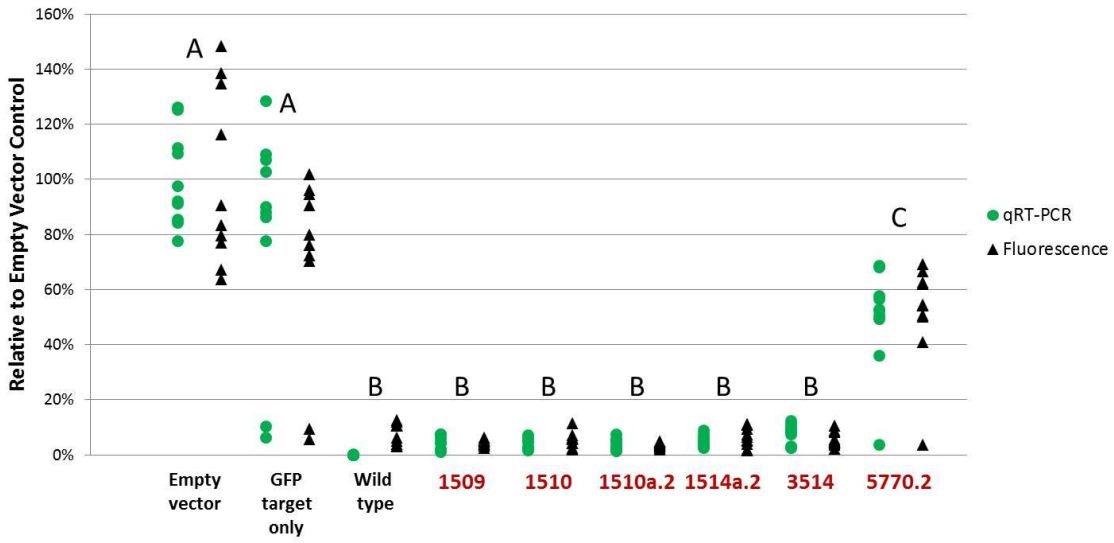


Figure 2.4. QRT-PCR and fluorescent-protein quantification of hairy roots. Values were normalized to the empty-vector control. Means were separated using Student's T-test $\alpha = 0.05$, n=10. Groups with the same letter are not significantly different. Results are the same for qRT-PCR and fluorescent-protein quantification.

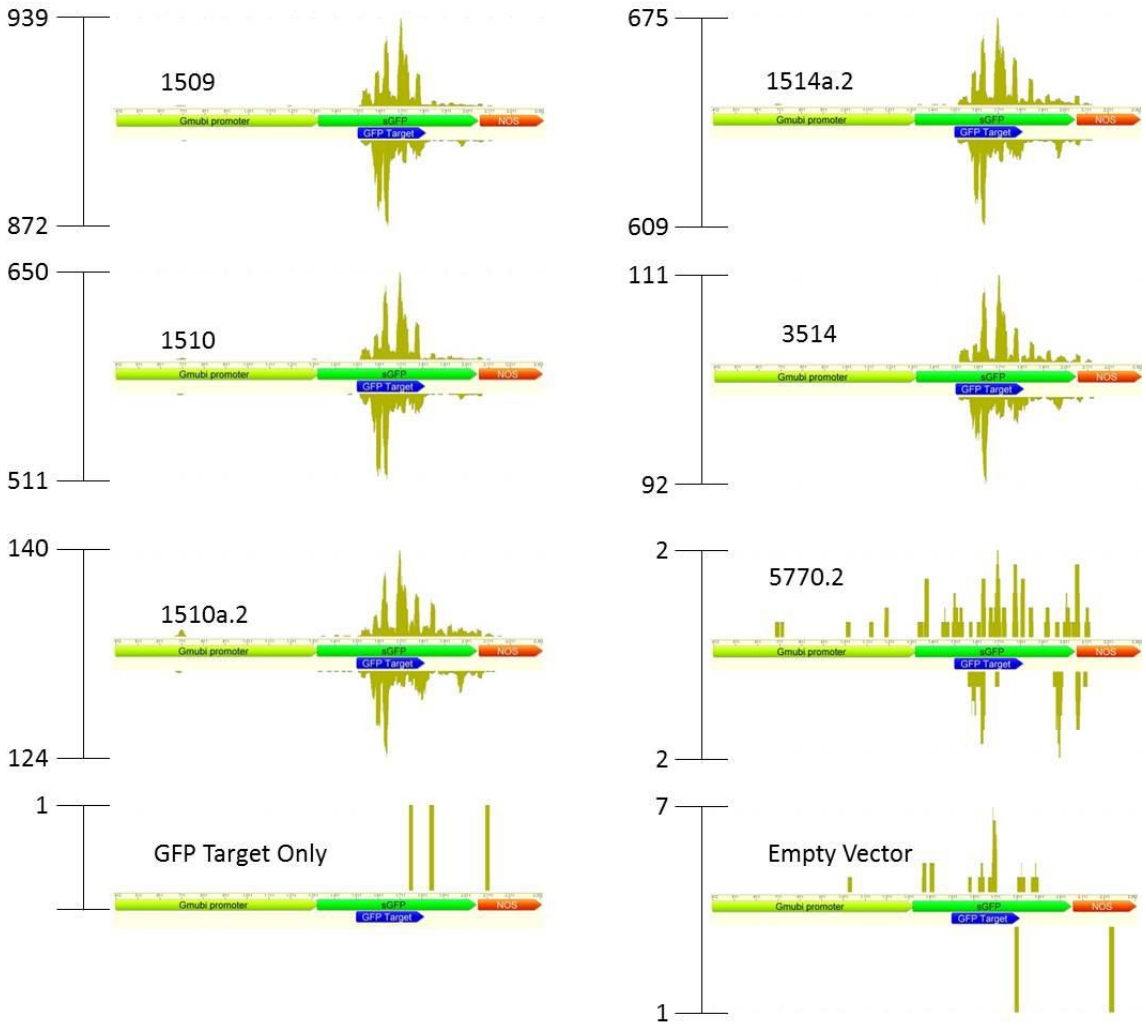


Figure 2.5. Mapping of sRNAs to the targeted GFP gene. The GFP sequence fused to the miRNA target is in blue. Scale bars are normalized to the millions of reads sequenced per library (reads per million).

Vector	Millions of reads	Positive strand	Negative strand	Total
1509	0.827	3898	3154	7052
1510	0.731	2492	1998	4490
1510a.2	1.450	692	516	1208
1514a.2	0.963	3272	2336	5608
5770	2.618	23	8	31
3514	1.008	481	332	813
GFP	0.756	4	0	4
K599	0.810	22	2	24

Table 2.2. Small RNA reads from miRNA target:GFP hairy-root events mapped to the GFP target gene. Reads are normalized to the millions of reads sequenced per library (reads per million).

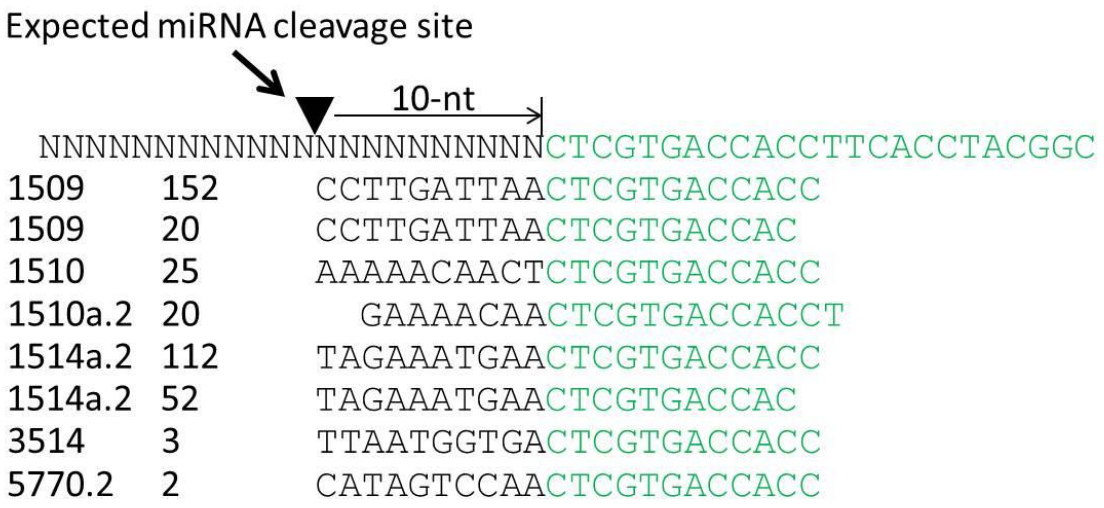


Figure 2.6. Chimeric siRNAs from target:GFP vectors. The triangle indicates the predicted miRNA cleavage site, between the 10th and 11th nucleotide of the miRNA. The total number of each respective chimeric reads are indicated in the second column. Only events containing the ta-siRNA silencing constructs produced chimeric reads, indicating that silencing is caused by the ta-siRNA pathway.

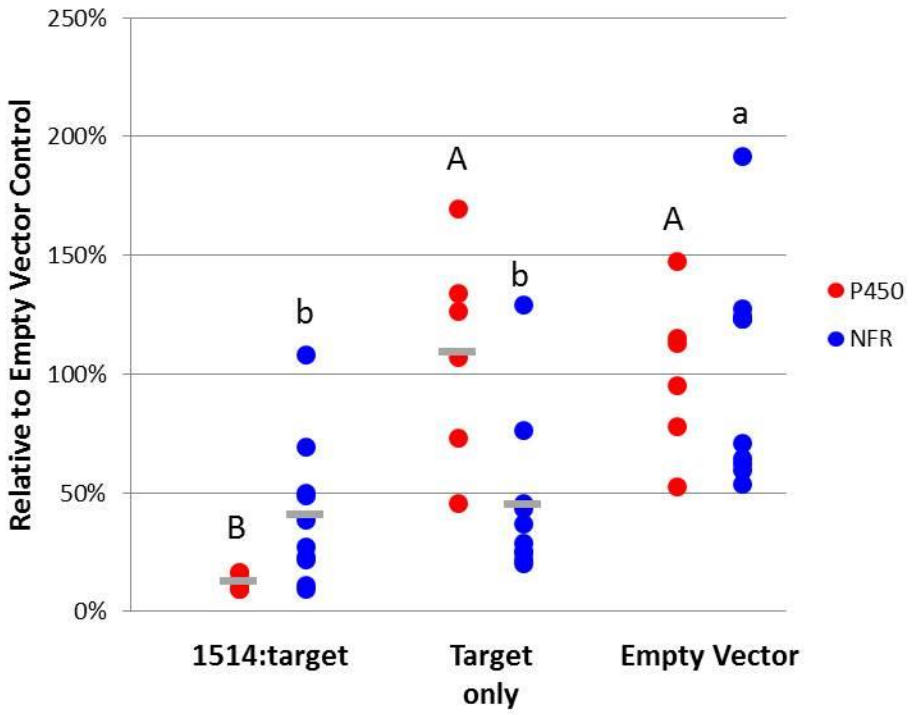


Figure 2.7. Silencing of the P450 and NFR genes in soybean hairy roots. Expression is relative to the empty vector control. Groups with the same letter are not significantly different. Means were separated using Student's T-test, $\alpha = 0.05$.

Vector	Millions of reads	positive strand	negative strand	Total
1514a.2:P450	0.365	728	1943	2670
	0.862	438	1546	1985
	1.023	217	741	958
P450 target only	1.130	2	1	3
	0.617	2	3	5
	0.644	5	0	5
1514a.2 NFR	0.531	2051	1075	3127
	0.803	1071	538	1610
	0.547	2160	882	3042
NFR target only	0.861	3919	687	4606
	1.097	2	1	3
	0.619	2	0	2

Table 2.3. Small RNAs mapping to P450 and NFR target genes. Reads are normalized to the millions of reads sequenced per library (reads per million). Strand bias is the number of negative strand reads divided by the number of positive strand reads

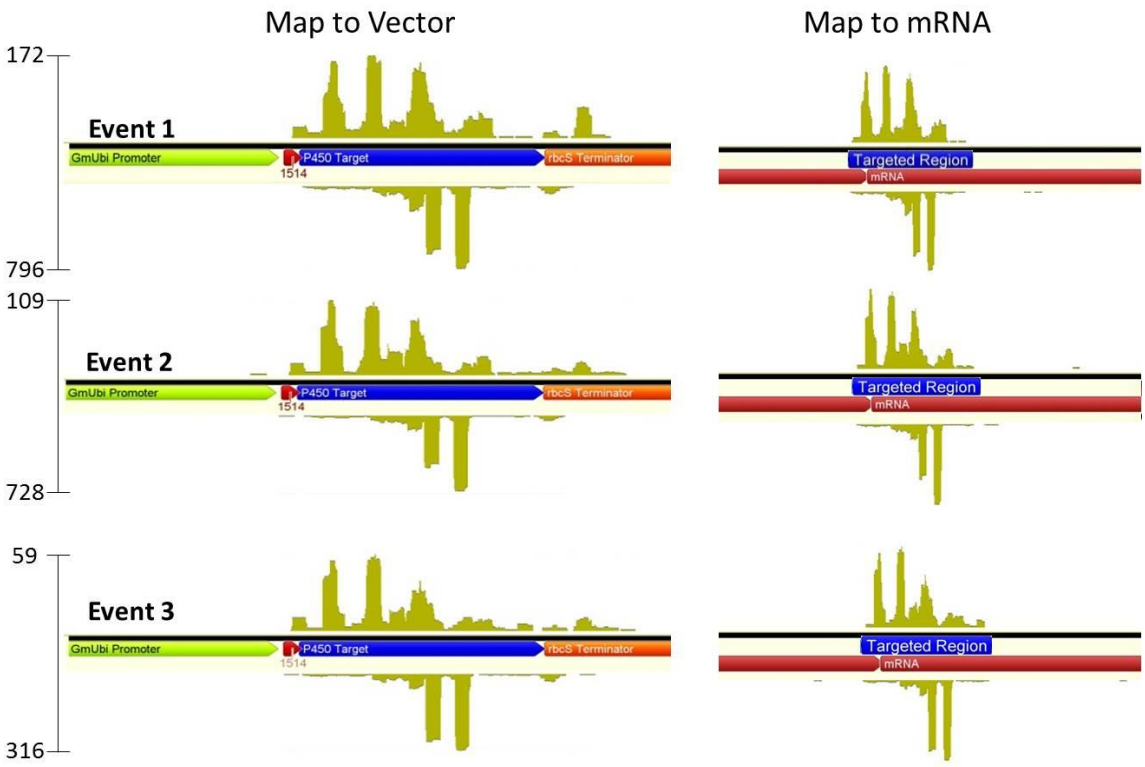


Figure 2.8. Mapping of sRNA reads to ta-siRNA vector and P450 mRNA target. Scale bars are normalized to the millions of reads sequenced per library (reads per million).

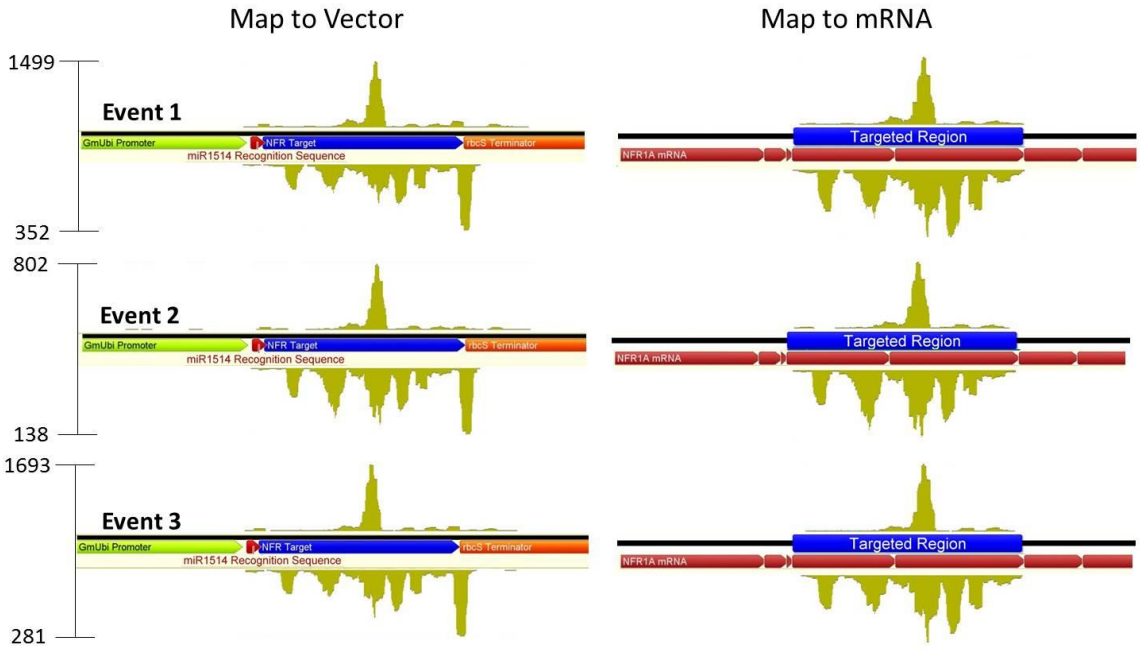


Figure 2.9. Mapping of sRNA reads to ta-siRNA vector and NFR mRNA target. Scale bars are normalized to the millions of reads sequenced per library (reads per million).

	# Events selected		Missing component		% Missing		
Vector	**Rep1**	**Rep2**	**Rep1**	**Rep2**	**Rep1**	**Rep2**	**Avg**
1514:GFP	22	18	0	3	0%	17%	8%
hpGFP	17	8	4	1	24%	13%	18%
1514:NFR	27	27	0	2	0%	7%	4%
hpNFR	14	22	2	10	14%	45%	30%
1514:P450	11	22	0	2	0%	9%	5%
hpP450	5	13	2	6	40%	46%	43%

Table 2.4. Presence and absence of vector components from hairy-root events transformed with 1514a.2:target and hairpin vectors.

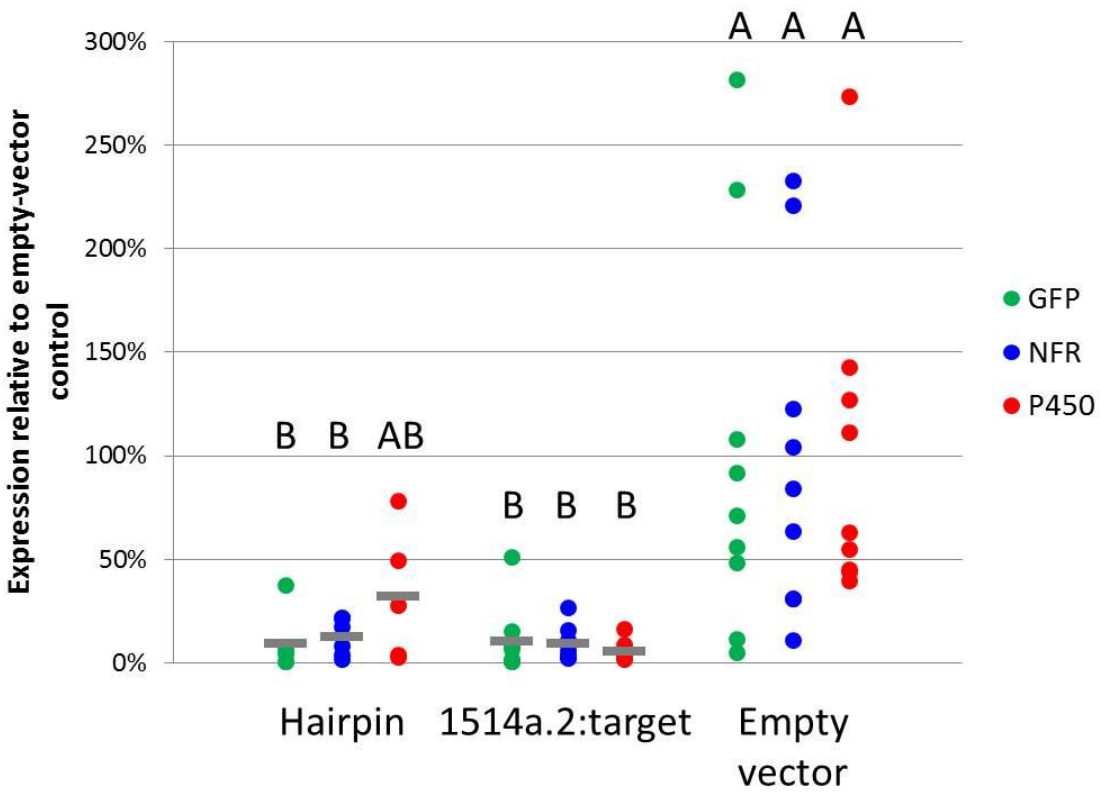


Figure 2.10. Relative expression of hairy-root events transformed with hairpin and 1514a.2:target vectors. Grey bars are the average of the expression values. Within each target, vectors labeled with the same letter are not significantly different. Means were separated with Tukey-Kramer HSD (α=0.05). For hairpin GFP n=6, 1514a.2:GFP n=9, Jack GFP n=9; hairpin NFR n=6, 1514a.2:NFR n=9, Jack NFR n=9; hairpin P450 n=5, 1514a.2:P450 n=10, Jack P450 n=9.

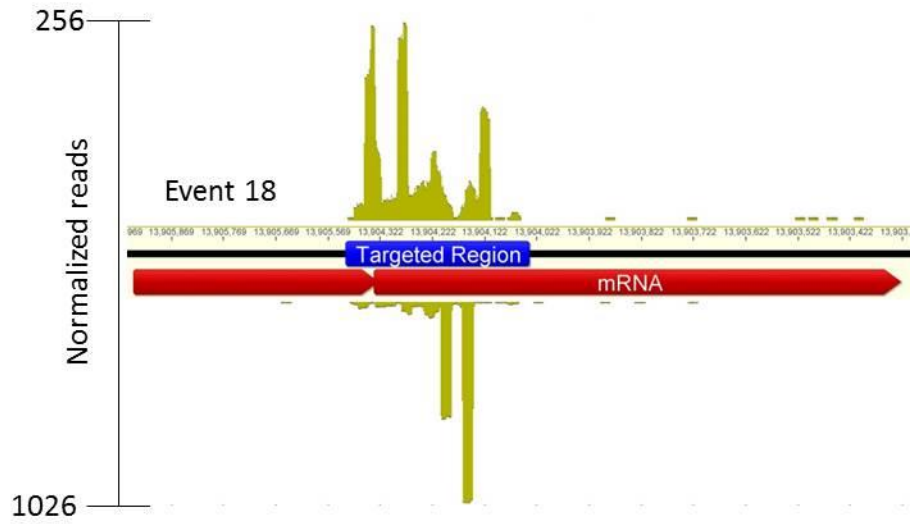


Figure 2.11. Assembly of sRNAs to the transgenic vector and target mRNA from T0 leaf tissue. Scale bar is RPM.

Line	Millions of reads	Positive strand	Negative strand	Total
12	0.984	37	300	337
18	1.145	255	1872	2127
20	0.407	170	1274	1444

Table 2.5. Small RNA reads from 1514a.2:P450 T0 events mapped to the P450 target gene. Reads are normalized to the millions of reads sequenced per library (reads per million).

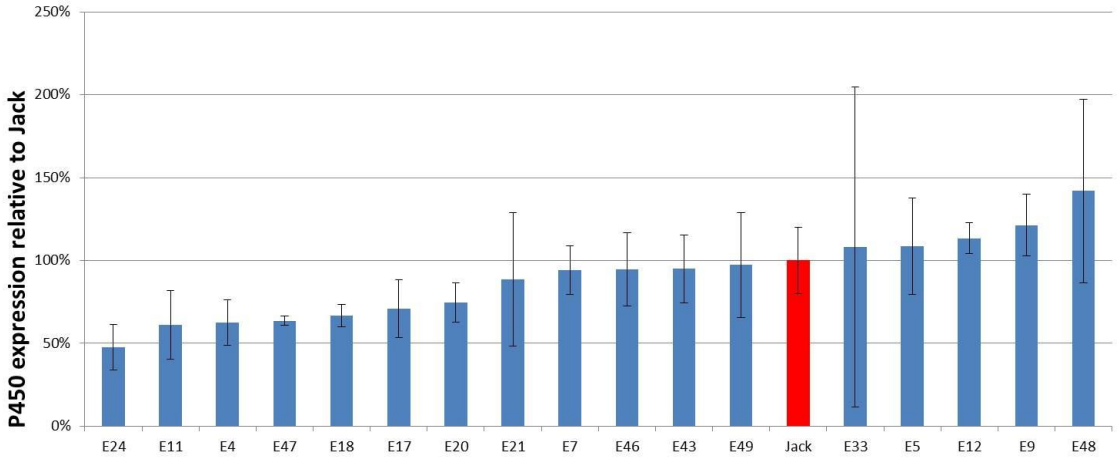


Figure 2.12. Silencing of the P450 gene in T1 plants. Results are an average of two T1 plants (blue bars) and are normalized to the expression of two Jack plants (red bar). Error bars are standard deviation.

Line_T1 Plant	Millions of reads	Positive strand	Negative strand	Total
7_1	1.226	116	1119	1235
7_2	1.368	173	1519	1692
11_1	1.179	171	842	1013
11_2	0.828	132	627	759
12_1	0.825	5	90	95
12_2	1.269	6	50	56
24_1	1.183	934	4932	5866
24_2	0.978	689	3151	3840
33_1	1.089	522	2515	3037
33_2	0.997	366	2312	2678
Jack	0.837	0	0	0

Table 2.6. Small RNA reads from 1514a.2:P450 T1 events mapped to the P450 target gene. Reads are normalized to the millions of reads sequenced per library (reads per million).

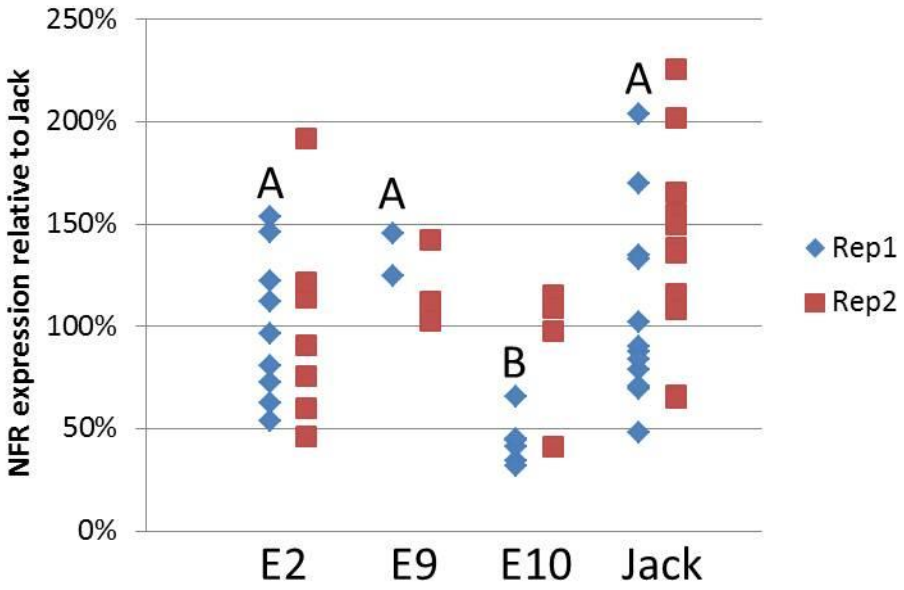


Figure 2.13. Relative expression of the NFR gene in T1 1514a.2:NFR events. A significant difference was only detected in rep 1 (p<0.0073). Groups labeled with the same letter are not significantly different. Means were separated with Tukey-Kramer HSD (α=0.05). Rep1, E2 n=2, E9 n=6, E10 n=12, Jack n=9; Rep2, E2 n=3, E9 n=4, E10 n=11, Jack n=7.

Line_T1 Plant	Millions of reads	Positive strand	Negative strand	Total	Strand Bias (-/+)
2_1	2.642	0	0	0	N/A
2_2	2.213	0	0	0	N/A
9_1	2.737	214	303	517	1.4
9_2	1.931	312	395	707	1.3
10_1	1.079	7	9	16	N/A
10_2	2.241	164	192	356	1.2
Jack	2.764	1	0	1	N/A

Table 2.7. Small RNA reads from 1514a.2:NFR T1 events mapped to the NFR target gene. The Line_1 plants are from rep1 and the line_2 plants are from rep2. Reads are normalized to the millions of reads sequenced per library (reads per million).

Primers for qRT		
	F	PCR efficiency
NFR4F	GTCTTGATGCACAGTTGCTGCAGC	1.84
NFR4R	CTCGGGTACAATGGAACATAGTCTCC	
GFP1F	GCCGACAAGCAGAAGAACGGCA	1.942
GFP1R	AGGTAGTGGTTGTCGGGCAGCA	
P4501F	ACAGAAGCTGAAGTGACTGGGCTT	1.944
P4501R	CACAGCTGAAAGGTACTGGTTGTT	
rbcsTF	GTTCGAGTATTATGGCATTGGGAAAACTG	1.962
rbcsTR	CACAGTTCGATAGCGAAAACCGAAT	

Cloning Primers	
AscI 1514:GmNFR1a	ATATA**GGCGCGCC**CAATGCCTATTTTAGAAATGAACAGGTCTTGCTAGGGGTGC
AvrII GmNFRR1a	TATAT**CCTAGG**CTCGCCTCTCAGTTCTGCATAA
AscI 1514:GmP450	ATATA**GGCGCGCC**CAATGCCTATTTTAGAAATGAACTTAGGGCCAACAAGCTCAAGG G
AvrII GmP450	TATAT**CCTAGG**CTGATTTGCTTGCACTTTTGCG
AscI GmNFR1a	ATATA**GGCGCGCC**CAGGTCTTGCTAGGGGTGC
AscI GmP450	ATATA**GGCGCGCC**CTTAGGGCCAACAAGCTCAAGGG
AscI 1509:GFP	ATATA**GGCGCGCC**CAACCTTGATTTCCTTGATTAACTCGTGACCACCTTCA
AscI 1510:GFP	ATATA**GGCGCGCC**AGGTGGAATAGGAAAAACAACTCTCGTGACCACCTTCA
AscI 1510a.2:GFP	ATATA**GGCGCGCC**ATGGGTGGAATAGGGAAAACAACTCGTGACCACCTTCA
AscI 1514:GFP	ATATA**GGCGCGCC**CAATGCCTATTTTAGAAATGAACTCGTGACCACCTTCA
AscI 3514:GFP	ATATA**GGCGCGCC**AAGGTCTCTGTCTTAATGGTGACTCGTGACCACCTTCA
AscI 5770:GFP	ATATA**GGCGCGCC**TCTTGTCCAAACCATAGTCCAACTCGTGACCACCTTCA
AscI GFP	ATATA**GGCGCGCC**CTCGTGACCACCTTCA
AvrII GFP R	TATAT**CCTAGG**GCCGTTCTTCTGCTTGTC

AscI AvrII NFR F	TATAT**GGCGCGCCCTAGG**CAGGTCTTGCTAGGGGTGC
BglII SwaI NFR R	TATAT**ATTTAAATAGATC**TCTCGCCTCTCAGTTCTGCATAA
AscI AvrII P450F	TATAT**GGCGCGCCCTAGG**CTTAGGGCCAACAAGCTCAAGGG
BamHI SwaI P450R	ATATA**GGATCCATTTAAAT**CTTAGGGCCAACAAGCTCAAGGG
AscI AvrII GFPF	TATAT**GGCGCGCCCTAGG**CTCGTGACCACCTTCA
BamHI SwaI GFPR	ATATA**GGATCCATTTAAAT**GCCGTTCTTCTGCTTGTC

Restriction sites in bold
miRNA recognition sequences underlined

Genotyping Primers	
GmUbi842 F	CGAGATTGCTTCAGATCCGTA
rbcsT272 R	CGGTCATTAGAGGCCACGAT

Table 2.8. Primers used in this study. Restriction sites are in bold. The miRNA recognition sequences are underlined.

CHAPTER 3

TARGETED GENOME MODIFICATIONS IN SOYBEAN WITH CRISPR/CAS9[1]

Thomas B Jacobs, Peter R LaFayette, Robert J Schmitz, Wayne A Parrott. Submitted to *BMC Biotechnology*, 09/26/14

Abstract

Background

The ability to selectively alter genomic DNA sequences *in vivo* is a powerful tool for basic and applied research. The CRISPR/Cas9 system precisely mutates DNA sequences in a number of organisms. Here, the CRISPR/Cas system is shown to be effective in soybean by knocking-out a green fluorescent protein (GFP) transgene and modifying nine endogenous loci.

Results

Targeted DNA mutations were detected in 95% of 88 hairy-root transgenic events analyzed. Bi-allelic mutation were involved in all cases. Small deletions were the most common type of mutation produced, although SNPs and short insertions were also observed. Homoeologous genes were successfully targeted singly and together, demonstrating that CRISPR/Cas9 can both selectively, and generally, target members of gene families. Somatic embryo cultures were also modified to enable the production of plants with heritable mutations, with the frequency of DNA modifications increasing with culture time. A novel cloning strategy and vector system based on In-Fusion® cloning was developed to simplify the production of CRISPR/Cas9 targeting vectors, which should be applicable for targeting any gene in any organism.

Conclusions

The CRISPR/Cas9 is a simple, efficient, and highly specific genome editing tool in soybean. Although some vectors are more efficient than others, it is possible to edit

duplicated genes relatively easily. The vectors and methods developed here will be useful for the application of CRISPR/Cas9 to soybean and other plant species.

Introduction

Methods to specifically target and modify DNA sequences are indispensable for basic and applied research. Recently, the type II bacterial clustered, regularly interspaced, short palindromic repeats (CRISPR) system emerged as a simple and efficient tool to target and modify DNA sequences of interest in a variety of organisms, including; cultured human cells [1, 2], zebrafish embryos [3], yeast [4], mice [5], and plants such as rice [6-9], *Arabidopsis thaliana* [10], maize [11] and liverwort [12].

There are two components to the CRISPR system: a nuclear-localized CRISPR-associated (Cas) 9 protein and a guide RNA (gRNA). Cas9 is a large protein containing two nuclease domains, and the most commonly used one is derived from *Streptococcus pyogenes*. The gRNA is a synthetic 100 nucleotide (nt) RNA molecule, of which the first approximately 20 nt are the targeting site, and the 3' end forms a hairpin structure that interacts with the Cas9 protein [13]. Cas9 and the gRNA interact to identify DNA sequences complementary to the gRNA and generate a DNA double-strand break (DSB).

When a DNA DSB occurs in eukaryotic cells, the imprecise repair mechanism, non-homologous end joining (NHEJ), can result in the insertion and/or deletion of sequences at the breakage site, typically resulting in frame-shift mutations [14]. In plants, such targeted DSBs can be used to knock-out genes [15, 16], modify gene expression by disrupting promoter sequences [17], or insert transgenes at a specific location via homologous recombination [18-22].

This work characterizes and further extends the use of CRISPRs for the genetic modification of soybean genes. CRISPR vectors targeting 11 loci were introduced into soybean via *Agrobacterium rhizogenes* to generate transgenic hairy roots. Custom-amplicon sequencing of DNA from these roots show that genetic modifications were made in 95% of the tested events. Modifications were also detected in somatic embryo cultures, and these should result in soybean lines with germinal modifications. Differences between *Agrobacterium*- and particle bombardment-mediated transformation were observed and may be important considerations for transformation experiments. To facilitate CRISPR mutagenesis efforts, a series of CRISPR vectors and a novel gRNA cloning method were produced.

Results and Discussion

Knock-out of a GFP transgene

The first test of the CRISPR system in soybean was with a GFP (*Green Fluorescent Protein*)-expressing soybean line, as GFP knock-outs are easily observed by a loss of fluorescence. Two GFP-targeting gRNA vectors were designed; one gRNA was designed to target the 5' end of *GFP* (5'-target) and a second was designed to target the 3' end (3'-target) (Figure 3.1A). The vectors were introduced into the GFP line via *A. rhizogenes* to produce hairy roots. 15 out of 17 5'-target events and four of the 22 3'-target events were knock-outs as evident by a loss of fluorescence under blue-light (Figure 3.2). Controls containing either Cas9 or the gRNAs alone, all fluoresced (Figure 3.2). Since the GFP soybean line used is homozygous for GFP, these results show that

the CRISPR system is able modify both GFP alleles, which is the only way to get loss of fluorescence.

Custom-amplicon sequencing was used to determine the genetic modifications at the *GFP* transgene. The most abundant mutations at the 5'-target were short (1-21-nt) deletions (Table 3.1). For event 10, a wild-type sequence was observed in 16% of the reads, which is consistent with fluorescent imaging (Figures 3.1 and 3.2). The 3'-target is less efficient; wild-type sequences were observed in seven of the events, with one event being completely unmodified (Table 3.1). Events with wild-type and modified sequences may be due to a single GFP allele being modified, or to the presence of chimeric tissues. Four of the 3'-target events contained SNPs and one event contained a T insertion, whereas the 5'-target events did not contain any SNPs or insertions. A single SNP at the 3'-target was routinely observed in the modified events and Cas9 control and may be due to errors during library preparation or sequencing.

Modifying a soybean gene

Given the successful modifications of the GFP target, the next attempt was to modify the single-copy soybean gene, Glyma07g14530, which is a putative glucosyl-transferase. Glyma07g14530 custom amplicons from ten independent events were sequenced, and these showed a variety of mutations, including deletions, SNPs, insertions, and replacements (Table 3.1). Replacements are defined as two or more bases that were incorporated after a deletion event. Three events contained only modified sequences, six events had both wild-type and modified sequences, and one event had no

modifications. These results indicate that both mono- and biallelic modifications were made and/or chimeric tissues were present.

Targeting homoeologous genes

Soybean is a paleopolyploid [23] and thus most genes have a homoeolog. For functional genomic studies, it would be beneficial if the CRISPR system could be used to target a homoeologous gene-pair singly and at the same time. To test this, the soybean genes Glyma01g38150 and Glyma11g07220 (orthologs of the *Arabidopsis thaliana DDM1* gene) were targeted. Three gRNAs were designed; one to target Glyma01g38150 (01gDDM1), one to target Glyma11g07220 (11gDDM1), and a third to target both (01g+11gDDM1). Both single-targeting gRNAs resulted in average indel frequencies greater than 70% (Figure 3.3). For 01gDDM1, eight events had indel frequencies between 87-97%. Two events only had indel frequencies of 1-2%, but these were still higher than the Cas9 control (0.14%). All but one of the 11gDDM1 events had indel frequencies greater than 95% (Figure 3.3). The 01gDDM1 gRNA was specific for the intended chr1 target, but the 11gDDM1 gRNA led to a small but detectable level (2-13%) of off-target modifications at the chr1 sequence (Figure 3.4).

Genetic modifications at both DDM1 genes were detected in events containing the 01g+11gDDM1 gRNA, but the average indel frequency was only 21% for chr1 and 8.9% for chr11 (Figure 3.3). Average indel frequencies greater than 97% were observed in events targeting a different homoeologous gene pair Glyma04g36150 and Glyma06g18790 (*Arabidopsis thaliana MET1* orthologs), suggesting that the lower indel

frequency of the 01g+11gDDM1 vector is due to the gRNA itself and not a result of targeting multiple genes at once.

It is noteworthy that unique insertions of the *A. rhizogenes* root-inducing (Ri) plasmid [GenBank: AJ271050] were present in two 11gDDM1 events. The Ri insertions were identified in 4.8% of the reads from event 3 and 79.2% of the reads from event 4. Both insertions are from the left-border end of the Ri plasmid, approximately 1 kb apart from each other. Cloning and sequencing of event 4 showed a 252-bp insertion from the Ri plasmid (Figure 3.5). These results are particularly interesting since it should be possible to increase the chances of obtaining targeted insertions, as has been shown with other nuclease systems [24].

Targeting MIR *genes*

MicroRNAs (miRNAs) are small RNA molecules responsible for regulating a wide range of processes in plants [25]. MicroRNAs are encoded by *MIR* genes that are typically short (~500 bp), non-coding sequences. These features, coupled with the genetic redundancy of *MIR* families, may decrease the likelihood of isolating *MIR* mutants in mutagenesis screens [26]. Thus, the specific targeting of Cas9, and the large number of targets for any given gene, may make the Cas9 system well suited for generating *MIR* mutants. Two soybean miRNAs, miR1514 and miR1509 were targeted with Cas9. The short length of the *MIR* genes limited the number of possible Cas9 targets. Finding a *MIR1514* target near the mature miRNA was particularly difficult. Since mismatches are tolerated on the 5' end of the gRNA [13], a C to G mismatch between the target and gRNA was made on the 5' base (Figure 3.3) to get a target close to the mature miRNA.

Indel frequencies greater than 95% were observed in all four miR1509-, and three out of four miR1514-targeted events. None of the short deletions (1-16 bp) were within the mature miRNA sequences, thus, none the mutations are expected to alter the production of the miRNAs. However, these results demonstrate that short, non-coding sequences, such as *MIRs*, can be readily targeted by the CRISPR/Cas system.

Genetic modification of somatic embryos

Hairy roots are an excellent transgenic model system for soybean, however, they cannot generate whole plants, and therefore heritable mutations cannot be made. To evaluate CRISPR mutagenesis in whole plants, somatic embryo cultures of soybean were biolistically transformed with Cas9 constructs. Eight Glyma07g14530 and 24 01g+11gDDM1 hygromycin-resistant events were recovered. Although each event contained portions of the gRNA and Cas9 genes (data not shown), only two Glyma07g14530 and three 01g+11gDDM1 events contained a complete Cas9 gene as determined by long-distance PCR (Figure 3.6). When hairy-root events (*Agrobacterium* transformation) were screened, a full Cas9 product was observed in all ten events (Figure 3.6). These results suggest that the Cas9 gene fragmented during biolistic-mediated transformation, but not upon *Agrobacterium*-mediated transformation.

As with other Cas9 systems [10], the continued activity of Cas9 in the somatic embryos resulted in additional genetic modifications. DNA samples were taken from all events once there was enough tissue, approximately 2-4 weeks after selection, and used for amplicon sequencing. At this first sequencing time-point, event 24 had approximately 2.5 % modified sequences on chr1 and chr11, whereas events 10 and 21 had none.

Although individual modified sequences made up fewer than 1% of the reads in event 24 (Table 3.1), such deletions were not observed in any of the other 23 events sequenced, indicating that these deletions were not due to sequencing errors. When DNA was collected approximately two weeks after the first sequencing experiment, the indel frequency increased to 4.3% in event 24. Events 10 and 21 had 20% and 4-5% modified sequences, respectively, for both targets (Figure 3.7).

The two Glyma07g14530 events did not survive tissue culture and no modifications were detected in DNA from somatic embryos (data not shown). Individual embryos from event 24 range in indel frequency from 0-14%, with most of the events at 4% (Figure 3.7). Therefore continued expression of Cas9 leads to additional mutations during the development of these embryos.

Mutation Efficiency

Of the nine targeting vectors used in this study, seven resulted in average indel frequencies greater than 70% (GFP 5', 01gDDM1, 11gDDM1, Glyma04g36150, Glyma06g18790, MIR1509, and MIR1514). In hairy roots, the 01g+11gDDM1 vector had the lowest average, with 21% and 8.9% for the chr1 and chr11 targets, respectively. A similar frequency was observed in the somatic embryos (Figure 3.7). It should be noted that the 01g+11gDDM1 gRNA is one base shorter than the rest of the gRNAs in this study ($GN_{19}GG$). However, this target length has been used in plants [27], and shorter gRNAs ($GN_{18}GG$) have been shown to be as effective as commonly used gRNA ($GN_{20}GG$) in cultured human cells [28]. It seems unlikely that a shorter gRNA led to a decrease it indel frequency, but a thorough testing of gRNA lengths in plants has not been

reported. Although each of the vectors had a range of indel frequencies, only four out of 88 (5%) hairy-roots were unmodified, demonstrating that CRISPR mutagenesis in soybean is a robust system.

The three 01g+11gDDM1 somatic-embryo events with the complete Cas9 gene contained targeted genetic modifications. These were three out of 24 hygromycin-resistant lines. These data demonstrates that when the complete Cas9 is incorporated, genetic modifications are made, although the complete Cas9 gene is only incorporated in 12.5% biolistically-transformed events. Of the recent reports of CRISPRs being used in plants, several have shown the recovery of whole-plants. One publication reported the biolistic transformation of rice, in which 9.4% and 7.1% of the T0 rice plants recovered contained mutations at their respective targets [27]. In this report, the Cas9 and gRNA cassettes were located on separate plasmids, and it is unclear if the complete Cas9 and gRNA cassettes were incorporated in all events. In contrast, transgenic *Arabidopsis thaliana* and rice plants transformed with *Agrobacterium tumefaciens* had efficiencies of 20-90% for several targets [6, 7, 9, 29]. Our data suggest that the disparity between biolistic and *Agrobacterium*-mediated transformation could be due to incomplete incorporation of the complete Cas9 gene upon biolistic-mediated transformation.

Types of mutations

The types of mutations obtained here are similar to those observed in other plants obtained with zinc finger nucleases (ZFN) [15, 21]**,** transcription activator-like effector nucleases (TALENS) [17, 18] and CRISPRs [27, 29-31]; small deletions were the most frequent mutations; SNPs were less common.

The different targeting sequences tested led to a distinctive gamut of mutations. The seven most effective vectors almost exclusively generated short deletions, whereas the lower efficiency vectors contained more insertions/SNPs (Table 3.1). Of the ten 07g14530 events, seven had insertions of one or more bases. These results suggest that the differences were determined by either the target sequence or the gRNA. Therefore, multiple targeting vectors may be needed for any potential target sequence, depending on the frequencies/types of mutations desired. Obtaining a greater variety of mutations may be desirable when the intent is to produce an allelic series.

The types of mutations between the hairy-root events and somatic embryos are consistent between chromosomal targets and between transformation methods. Within the ten 01g+11gDDM1 hairy-root events, six contained an A insertion on chr1 at the same position. From those same ten events, five contained an A insertion on the homoeologous target on chr11 (Table 3.1). Each of the somatic-embryo events has the same A insertion for both chr1 and chr11, and in many cases, it is the most abundant read (Table 3.1). Given the consistent insertion pattern, it is tempting to speculate that there may be rules governing the types of mutations that are possible for a given target.

Evaluation of off-target modifications

One limitation of the CRISPR system is the potential for off-target modifications, i.e., the modification of sequences similar to the intended target sequence [13, 32]. To determine the extent to which there may be off-target modifications, putative off-target sites were identified for the Glyma07g14530, DDM1, MET1, and miR1514 vectors. Each putative off-target site has two to six mismatches relative to the gRNA (Figure 3.4).

Two gRNAs created off-target mutations. The 11gDDM1 chr1 off target was modified in 2-13% of the sequenced reads, which is considerably lower than the indel frequency at the intended chr11 target (95-100%). When off-targeting occurred at miR1514 18g, there was a range of frequencies; 100%, 25%, and 5%. The 07g14530-15g and -17g off-target loci had indel frequencies of 2.8% and 2.2%, respectively. However, the increased indel frequencies were also observed in the Cas9 control, showing that they were due to sequencing errors caused by long stretches of T's in the amplicons. These results indicate that while off-targeting does occur, at least for the tested gRNAs, it is not common, and was generally at a much lower frequency than at the intended target.

gRNA vector construction

In this work, a rapid cloning method (Figure 3.8) was developed to create new gRNAs. It consists of a single PCR reaction with two 41-bp primers and an In-Fusion® reaction and can be used to clone any gRNA target sequence. The pUC gRNA shuttle vector makes the construction of gRNAs simple and inexpensive. The use of the In-Fusion® cloning system has the benefit of reducing handling steps, to the point where it should be simple to automate the entire cloning process. Binary Cas9 vectors with four different selectable makers (*nptII*, GFP, hygromycin, *bar*) were also created to facilitate plant transformation experiments.

Conclusions

This work shows that the Cas9 system is functional in two stably transformed plant system, hairy roots and somatic embryos. It was possible to efficiently mutate all

11 loci chosen for testing; only two of the targeting vectors resulted in off-target mutations at predicted off-target loci. The different gRNA targets produced different types of mutations. Combined with a vector system developed to efficiently assemble the necessary gRNAs, these results confirm that the CRISPR system will be a simple and inexpensive method for genome editing in soybean, thus facilitating the use of genome editing to confirm candidate genes, develop novel alleles/phenotypes, and engineer plants with important agronomic or quality traits.

Experimental procedures

Vector Construction

The human codon-optimized Cas9 gene [2] was obtained from Addgene (plasmid 41815). Two flanking primers with added *Nhe*I and *Sac*II sites were used to amplify the coding sequence, including the SV40 nuclear localization signal, with the KAPA HiFi polymerase (KAPA BioSystems). The amplicon was digested with the two restriction enzymes and ligated to the vector, pM35S, between the double 35S promoter and nopaline synthase (nos) terminator (Figure 3.9). The entire cassette is flanked with I-*Sce*I restriction sites, which were used to move the Cas9 cassette into p201N to create p201N:Cas9 (Addgene plasmid 59175). The p201N vector is a p201BK [33] vector modified to include an *nptII* selectable marker cassette and I-*Sce*I and I-*Ppo*I restriction sites (Figure 3.9).

For biolistic transformation of soybean, a pSMART HC Kan (Lucigen Corporation, [GenBankAF532107]) cloning vector was modified to contain a *hygromycin phosphotransferase* (*hph*) gene under the control of the Ubi3 promoter and terminator

[34] and the meganuclease I-*Ppo*I site, and is referred to as pSPH2. The vector pSPH2 was digested with I-*Ppo*I and DNA overhangs were removed with T4 DNA polymerase. To prepare the Cas9 insert, p201N:Cas9:gRNA-Glyma07g14530 was digested with *Spe*I and *Pme*I and DNA overhangs were removed with T4 DNA polymerase. The vector and insert were ligated to create the plasmid pSPH2:Cas9:gRNA-Glyma07g14530. The Glyma07g14530 gRNA was then replaced with the 01g+11gDDM1 (Glyma01g38150 and Glyma11g07220) gRNA via I-*Ppo*I to produce pSPH2:Cas9:gRNA-01g+11gDDM1.

Additional binary Cas9 vectors were produced by replacing *npt*II from p201NCas9, with *hph*, *bar* (phosphinothricin resistance), or GFP. The *hph* cassette was moved from pSPH2 into the p201N Cas9 vector with the *Pac*I and *Spe*I restriction sites to produce p201H:Cas9 (Addgene plasmid 59176). The bar and GFP cassettes (double 35S promoter, nos terminator) were amplified with the *Spe*I 35SF and *Pac*I nosR primers (Table 3.2), and moved into the p201N Cas9 vector with the *Pac*I and *Spe*I restriction sites to produce p201B:Cas9 (Addgene plasmid 59177) and p201G:Cas9 (Addgene plasmid 59178).

The gRNA targets were designed as previously described [2], with the exception of the U6 promoter, which was replaced with the *Medicago truncatula* U6.6 polymerase III promoter [35] for efficient transcription in soybean. For the gRNA targets, 22-23-bp targets were chosen that had the $GN_{19-20}GG$ motif as previously described [2]. The GN_{18-19} portion of the genomic target motif was incorporated into the gRNA target molecule. The GFP, Glyma07g14530, and DDM1 gRNA target sequences were synthesized by IDT using gBlocks. The gBlocks were amplified by PCR with flanking primers containing I-*Ppo*I restriction sites. All primer sequences can be found in Table 3.2. The products

were then digested with I-*Ppo*I and inserted into the p201N vector. The MET1 (Glyma04g36150 and Glyma06g18790), miR1514, and miR1509 gRNA target sequences were produced with the pUC gRNA shuttle vector system described below. Plasmids were electroporated into *Agrobacterium rhizogenes* strain K599 and used for hairy-root transformation. Vectors containing both the Cas9 and gRNA target cassettes were combined by inserting the gRNA target cassette into the p201N Cas9 I-*Ppo*I site.

Hairy-root transformation of soybean

Soybean 'Jack-GFP [36]' and 'Jack' germinating seeds were used for transformation with slight modifications from the protocol previously described [37]. Briefly, soybean seeds were germinated for approximately one week under sterile conditions on a filter paper wetted with a ½ MSO liquid germination medium [38] supplemented with B5 vitamins [39]. *A. rhizogenes* (strain K599) containing the vectors-of-interest were streaked from glycerol stocks onto YM medium [40] supplemented with 50 mg L^{-1} kanamycin. Soybean cotyledons were prepared in a manner similar to that described for cotyledonary node transformation [41]; the root and lower hypocotyl were removed from the cotyledons, leaving approximately 5 mm of hypocotyl. The apical shoot and hypocotyl were cut longitudinally to produce two symmetrical cotyledons with a short hypocotyl piece. The apical meristem was removed and 1-mm-deep cuts were made in the cotyledons on the adaxial surface with a scalpel dipped in a solution of *A. rhizogenes* (PB Buffer (0.01 M Na2HPO4, 0.15M NaCl, pH 7.5) + 100 µM acetosyringone). Cotyledons were co-cultivated with *A. rhizogenes* for 3 days on filter paper wetted with 2 mL of liquid germination medium + 100 µM acetosyringone.

Cotyledons were transferred to a hairy-root growth (HRG) medium according to Cho et al. [37] with the following modifications: ½ MS salts, 2 g L^{-1} Phytagel, and 500 mg L^{-1} timentin to inhibit *A. rhizogenes*. Each root was treated as an individual event and transferred to HRG medium with 10 mg L^{-1} of Geneticin (G418). Those roots that grew on HRG + G418 were considered events, and a 2-cm portion of a root tip was collected for CTAB DNA extraction [42]. PCR was performed to confirm the presence of the Cas9 and gRNA genes with the primers listed in Table 3.2. Long-distance PCR was performed with a Promega long-distance PCR master mix according to manufacturer's instructions.

GFP Imaging

After selection on HRG + G418, root tips were imaged with an Olympus MVX10 microscope with a GFP filter cube and the imaging software DP controller version 2.2.1.227 (Olympus America Inc.). Blue-light images were taken with a 5 ms exposure.

Custom-Amplicon Sequencing and Analysis

Genomic DNA was amplified with the KAPA HiFi polymerase (KAPA Biosystems) with tailed primers under the conditions 95˚C for 3 min; 30 cycles (98˚C for 15 sec, 60˚C for 15 sec, 72˚C for 30-45 sec); and 72˚C for 5 min. PCR products were run on a 1% agarose, 1X TBE gel and visualized on a UV transilluminator to verify amplification. PCR products were pooled across amplicons, diluted 1:100, and used as a template for a second PCR with the conditions 95˚C for 3 min; 10 cycles (98˚C for 15 sec, 60˚C for 15 sec, 72˚C for 30-45 sec); and 72˚C for 10 min. The second PCR was used to add the final Illumina adapters and indexes. PCR products were again visualized

to ensure amplification. All products were pooled and concentrated with DNA clean and concentrator columns (Zymo Research). The pooled samples were run on a 1.5% agarose, 1X TAE + cytidine gel and the proper fragments were gel extracted with the Zymoclean Gel DNA Recovery Kit. Purified libraries were quantified with the KAPA Library Quantification Kit (KAPA Biosystems) and run on an Illumina MiSeq (Illumina Inc.). Reads were de-multiplexed with the MiSeq reporter software version 2.3.32.

Reads were imported into the software Geneious (Biomatters Ltd.) version 7. Reads were trimmed for quality and separated by amplicon using the separate-reads-by-barcode function using the forward sequencing primer + five bases downstream as the barcode. The five downstream bases were essential to remove primer-dimers from the analysis. After quality and barcode trimming, only reads within five bases of the expected length were extracted for analysis. Reads were trimmed to regions approximately 20-bp upstream and downstream of the gRNA target site. Sequences that were the length of wild-type sequences were extracted. Indel frequency was then calculated by subtracting the number of wild-type sequences from the total number of extracted reads.

For each of the targeted loci, unique sequences were extracted from the trimmed total extracted reads using the find-duplicates function. The most abundant, unique reads are reported in Table 3.1.

Off-target Sequence Identification

Potential off-target sites were identified by comparing the 23-bp gRNA target sequences using BLAST to the soybean reference genome (Glyma v1.1), on Phytozome,

setting the e-value threshold to 5 since the query sequence is only 23 nt. Only loci that had the required PAM NGG motif at the 3' end of the sequence were considered for analysis. Primers used for amplifying the off-target loci are in Table 3.2.

Biolistic Transformation of Somatic Embryos

Biolistic transformation of soybean was performed as previously described [43]. DNA was isolated from somatic embryo cultures for PCR and custom-amplicon sequencing.

gRNA Shuttle Plasmid

To facilitate the construction of gRNA targets, a shuttle plasmid was created that makes construction quick and inexpensive. The *Medicago truncatula* U6.6 promoter was fused to the gRNA scaffold [2], and the entire gRNA is flanked by I-*Ppo*I restriction sites. To produce a novel gRNA target, forward and reverse primers were designed with tails that encode the new target sequence (Figure 3.8). Fifteen bp of homology on the primer tails allowed for In-Fusion® cloning (Clonetech Laboratories Inc., Mountain View, CA). After transformation, the new gRNA target molecule was inserted between the promoter and gRNA scaffold. Sanger sequencing was performed with the commonly used M13-reverse primer to confirm the sequence of the gRNA. I-*Ppo*I was then used to move the functional gRNA target cassette into a vector of choice. The pUC gRNA Shuttle plasmid can be obtained from Addgene (plasmid 47024).

References

1. Cong L, Ran FA, Cox D, Lin SL, Barretto R, Habib N, Hsu PD, Wu XB, Jiang WY, Marraffini LA *et al*: **Multiplex genome engineering using CRISPR/Cas systems**. *Science* 2013, **339**(6121):819-823.

2. Mali P, Yang LH, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM: **RNA-guided human genome engineering via Cas9**. *Science* 2013, **339**(6121):823-826.

3. Hwang WY, Fu YF, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh JRJ, Joung JK: **Efficient genome editing in zebrafish using a CRISPR-Cas system**. *Nature Biotechnology* 2013, **31**(3):227-229.

4. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM: **Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems**. *Nucleic Acids Research* 2013, **41**(7):4336-4343.

5. Wang HY, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R: **One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering**. *Cell* 2013, **153**(4):910-918.

6. Xu R, Li H, Qin R, Wang L, Li L, Wei P, Yang J: **Gene targeting using the *Agrobacterium tumefaciens*-mediated CRISPR-Cas system in rice**. *Rice (New York, NY)* 2014, **7**(1):5.

7. Zhang H, Zhang J, Wei P, Zhang B, Gou F, Feng Z, Mao Y, Yang L, Zhang H, Xu N *et al*: **The CRISPR/Cas9 system produces specific and homozygous targeted gene editing in rice in one generation**. *Plant biotechnology journal* 2014, **12**(6):797-807.

8. Jiang WZ, Zhou HB, Bi HH, Fromm M, Yang B, Weeks DP: **Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in *Arabidopsis*, tobacco, sorghum and rice**. *Nucleic Acids Research* 2013, **41**(20):12.

9. Miao J, Guo DS, Zhang JZ, Huang QP, Qin GJ, Zhang X, Wan JM, Gu HY, Qu LJ: **Targeted mutagenesis in rice using CRISPR-Cas system**. *Cell Research* 2013, **23**(10):1233-1236.

10. Feng ZY, Mao YF, Xu NF, Zhang BT, Wei PL, Yang DL, Wang Z, Zhang ZJ, Zheng R, Yang L *et al*: **Multigeneration analysis reveals the inheritance, specificity, and patterns of CRISPR/Cas-induced gene modifications in *Arabidopsis***. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(12):4632-4637.

11. Liang Z, Zhang K, Chen KL, Gao CX: **Targeted mutagenesis in *Zea mays* using TALENs and the CRISPR/Cas system**. *Journal of Genetics and Genomics* 2014, **41**(2):63-68.

12. Sugano SS, Shirakawa M, Takagi J, Matsuda Y, Shimada T, Hara-Nishimura I, Kohchi T: **CRISPR/Cas9-mediated targeted mutagenesis in the liverwort *Marchantia polymorpha* L**. *Plant and Cell Physiology* 2014, **55**(3):475-481.

13. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E: **A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity**. *Science* 2012, **337**(6096):816-821.

14. Mladenov E, Iliakis G: **Induction and repair of DNA double strand breaks: The increasing spectrum of non-homologous end joining pathways**. *Mutation*

Research-Fundamental and Molecular Mechanisms of Mutagenesis 2011, **711**(1-2):61-72.

15. Curtin SJ, Zhang F, Sander JD, Haun WJ, Starker C, Baltes NJ, Reyon D, Dahlborg EJ, Goodwin MJ, Coffman AP *et al*: **Targeted mutagenesis of duplicated genes in soybean with zinc-finger nucleases**. *Plant Physiology* 2011, **156**(2):466-473.

16. Osakabe K, Osakabe Y, Toki S: **Site-directed mutagenesis in *Arabidopsis* using custom-designed zinc finger nucleases**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(26):12034-12039.

17. Li T, Liu B, Spalding MH, Weeks DP, Yang B: **High-efficiency TALEN-based gene editing produces disease-resistant rice**. *Nature Biotechnology* 2012, **30**(5):390-392.

18. Zhang Y, Zhang F, Li XH, Baller JA, Qi YP, Starker CG, Bogdanove AJ, Voytas DF: **Transcription activator-like effector nucleases enable efficient plant genome engineering**. *Plant Physiology* 2013, **161**(1):20-27.

19. Fauser F, Roth N, Pacher M, Ilg G, Sanchez-Fernandez R, Biesgen C, Puchta H: ***In planta* gene targeting**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**(19):7535-7540.

20. Shukla VK, Doyon Y, Miller JC, DeKelver RC, Moehle EA, Worden SE, Mitchell JC, Arnold NL, Gopalan S, Meng XD *et al*: **Precise genome modification in the crop species *Zea mays* using zinc-finger nucleases**. *Nature* 2009, **459**(7245):437-U156.

21. Townsend JA, Wright DA, Winfrey RJ, Fu FL, Maeder ML, Joung JK, Voytas DF: **High-frequency modification of plant genes using engineered zinc-finger nucleases**. *Nature* 2009, **459**(7245):442-U161.

22. D'Halluin K, Chantal V, Jolien VH, Joanna R, Ilse VDB, Anouk P, Kristel DH, Martine B, Derek J, Rene R *et al*: **Targeted molecular trait stacking in cotton through targeted double-strand break induction**. *Plant Biotechnology Journal* 2013:n/a--n/a.

23. Schlueter JA, Lin JY, Schlueter SD, Vasylenko-Sanders IF, Deshpande S, Yi J, O'Bleness M, Roe BA, Nelson RT, Scheffler BE *et al*: **Gene duplication and paleopolyploidy in soybean and the implications for whole genome sequencing**. *Bmc Genomics* 2007, **8**:16.

24. Qi Y, Zhang Y, Zhang F, Baller JA, Cleland SC, Ryu Y, Starker CG, Voytas DF: **Increasing frequencies of site-specific mutagenesis and gene targeting in *Arabidopsis* by manipulating DNA repair pathways**. *Genome research* 2013, **23**(3):547-554.

25. Bartel DP: **MicroRNAs: Target recognition and regulatory functions**. *Cell* 2009, **136**(2):215-233.

26. Jones-Rhoades MW, Bartel DP, Bartel B: **MicroRNAs and their regulatory roles in plants**. In: *Annual Review of Plant Biology.* vol. 57. Palo Alto: Annual Reviews; 2006: 19-53.

27. Shan Q, Wang Y, Li J, Zhang Y, Chen K, Liang Z, Zhang K, Liu J, Xi JJ, Qiu J-L *et al*: **Targeted genome modification of crop plants using a CRISPR-Cas system**. *Nat Biotech* 2013, **31**(8):686-688.

28. Fu YF, Sander JD, Reyon D, Cascio VM, Joung JK: **Improving CRISPR-Cas nuclease specificity using truncated guide RNAs**. *Nature Biotechnology* 2014, **32**(3):279-284.

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

30. Li J-F, Norville JE, Aach J, McCormack M, Zhang D, Bush J, Church GM, Sheen J: **Multiplex and homologous recombination-mediated genome editing in *Arabidopsis* and *Nicotiana benthamiana* using guide RNA and Cas9**. *Nat Biotech* 2013, **31**(8):688-691.

31. Nekrasov V, Staskawicz B, Weigel D, Jones JDG, Kamoun S: **Targeted mutagenesis in the model plant *Nicotiana benthamiana* using Cas9 RNA-guided endonuclease**. *Nat Biotech* 2013, **31**(8):691-693.

32. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, Sander JD: **High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells**. *Nat Biotech* 2013, **advance online publication**.

33. Covert SF, Kapoor P, Lee MH, Briley A, Nairn CJ: ***Agrobacterium tumefaciens*-mediated transformation of *Fusarium circinatum***. *Mycological Research* 2001, **105**:259-264.

34. Joshi M, Niu C, Fleming G, Hazra S, Chu Y, Nairn CJ, Yang HY, Ozias-Akins P: **Use of green fluorescent protein as a non-destructive marker for peanut genetic transformation**. *In Vitro Cellular & Developmental Biology-Plant* 2005, **41**(4):437-445.

35. Kim GB, Nam YW: **Isolation and characterization of *Medicago truncatula* U6 promoters for the construction of small hairpin RNA-mediated gene silencing vectors**. *Plant Molecular Biology Reporter* 2013, **31**(3):581-593.

36. Hernandez-Garcia CM, Martinelli AP, Bouchard RA, Finer JJ: **A soybean (*Glycine max*) polyubiquitin promoter gives strong constitutive expression in transgenic soybean**. *Plant Cell Reports* 2009, **28**(5):837-849.

37. Cho HJ, Farrand SK, Noel GR, Widholm JM: **High-efficiency induction of soybean hairy roots and propagation of the soybean cyst nematode**. *Planta* 2000, **210**(2):195-204.

38. Murashige T, Skoog F: **A revised medium for rapid growth and bio assays with tobacco tissue cultures**. *Physiologia Plantarum* 1962, **15**(3):473-497.

39. Gamborg OL, Miller RA, Ojima K: **Nutrient requirements of suspension cultures of soybean root cells**. *Experimental Cell Research* 1968, **50**(1):151-&.

40. Lin JJ: **Optimization of the transformation efficiency of *Agrobacterium-tumefaciens* cells using electroporation**. *Plant Science* 1994, **101**(1):11-15.

41. Olhoft PM, Flagel LE, Donovan CM, Somers DA: **Efficient soybean transformation using hygromycin B selection in the cotyledonary-node method**. *Planta* 2003, **216**(5):723-735.

42. Murray MG, Thompson WF: **Rapid isolation of high molecular-weight plant DNA**. *Nucleic Acids Research* 1980, **8**(19):4321-4325.

43. Hancock CN, Zhang F, Floyd K, Richardson AO, LaFayette P, Tucker D, Wessler SR, Parrott WA: **The rice miniature inverted repeat transposable element *mPing* is an effective insertional mutagen in soybean**. *Plant Physiology* 2011, **157**(2):552-562.

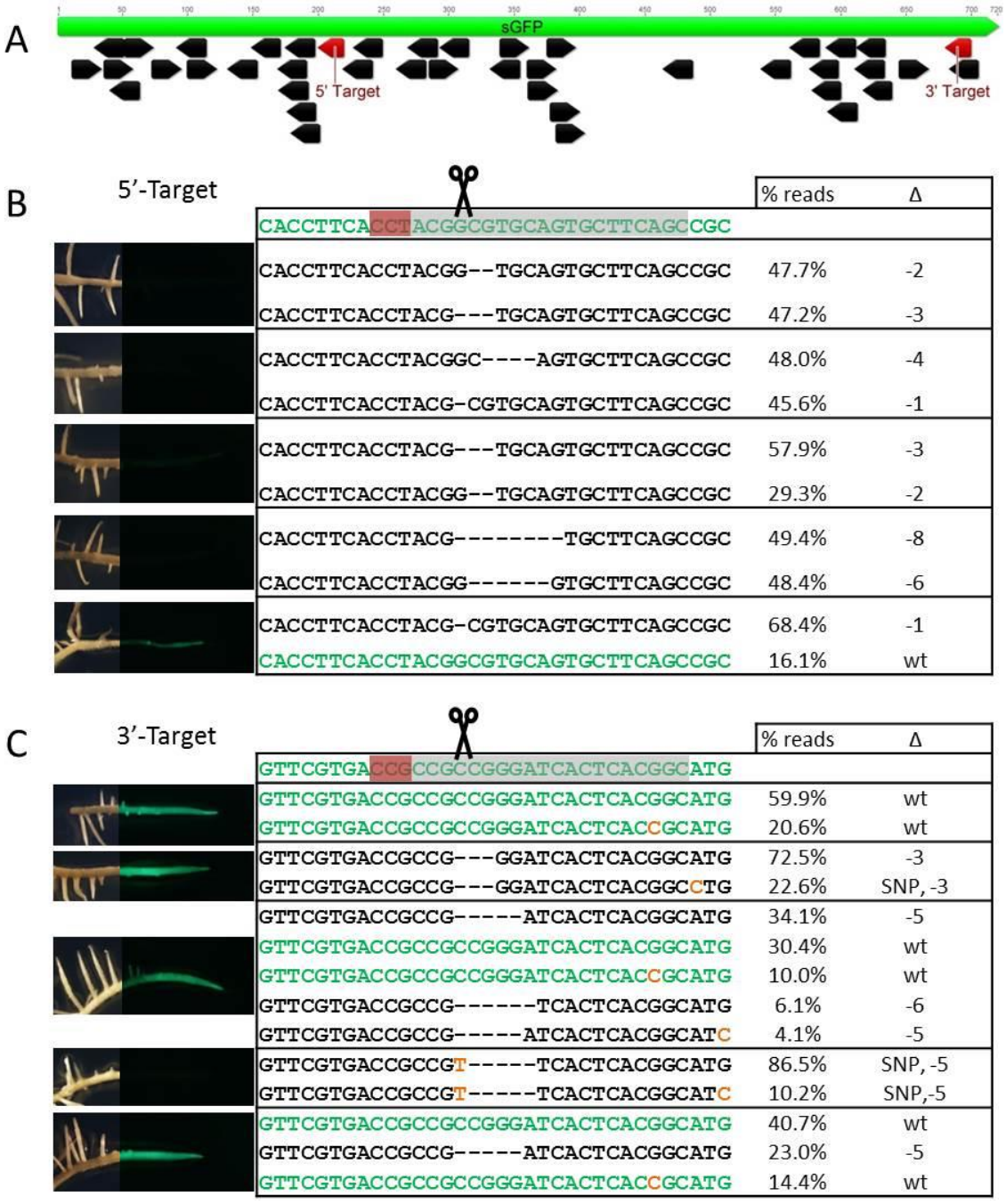


Figure 3.1. Cas9 targeting of a GFP gene in soybean hairy roots. (a) Schematic showing the targeted GFP sequences. The targets were designed to the negative strand of GFP. Black arrows are all possible $GN_{20}GG$ target motifs. GFP imaging and amplicon

sequencing of representative (b) C9 + GFP 5’ target events and (c) C9 + GFP 3’ target events. Each panel is an independent event and blue-light images were overlaid onto white-light images of roots. Wild-type sequences are in green, deletions are shown as dashes, and SNPs are shown in orange. The targeted sequences are highlighted in grey and the PAM is highlighted in red. Percentages next to sequences indicate the number of reads with sequence over the number of total reads sequenced. The same magnification was used for all images.

Event	GFP 5' Sequence	# Seqs.	Freq	Δ
	CACCCTCGTGACCACCTTCACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAA			
1	CACCCTCGTGACCACCTTCACCTACGG--TGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	3165	47.7%	-2
	CACCCTCGTGACCACCTTCACCTACG---TGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	3132	47.2%	-3
2	CACCCTCGTGACCACCTTCACCTACGGC----AGTGCTTCAGCCGCTACCCCGACCACATGAA	3115	48.0%	-4
	CACCCTCGTGACCACCTTCACCTACG-CGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	2959	45.6%	-1
3	CACCCTCGTGACCACCTTCACCTACG---TGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	2985	57.9%	-3
	CACCCTCGTGACCACCTTCACCTACGG--TGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	1507	29.3%	-2
4	CACCCTCGTGACCACCTTCACCTACG--------TGCTTCAGCCGCTACCCCGACCACATGAA	2853	49.4%	-8
	CACCCTCGTGACCACCTTCACCTACGG------GTGCTTCAGCCGCTACCCCGACCACATGAA	2797	48.4%	-6
5	CACCCTCGTGACCACCTTCACCTACGG--TGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	3005	95.3%	-2
6	CACCCTCGTGACCACCTTCACCTACGG--TGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	1655	48.9%	-2
	CACCCTCGTGACCACCTTCACCTACG---TGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	1592	47.0%	-3
7	CACCCTCGTGACCACCTTCACCTACGGC----AGTGCTTCAGCCGCTACCCCGACCACATGAA	1677	50.1%	-4
	CACCCTCGTGACCACCTTCACC----GCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	1527	45.6%	-4
8	CACCCGCGTGCAGTGCTTCA--------------------GCCGCTACCCCGACCACATGAA	1607	47.1%	-21
	CACCCTCGTGACCACCTTCACCTACGGC----AGTGCTTCAGCCGCTACCCCGACCACATGAA	851	24.9%	-4
	CACCCTCGTGACCACCTTCACCTACG---TGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	829	24.3%	-3
9	CACCCTCGTGACCACCTTCACCTACG---TGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	2595	48.4%	-3
	CACCCTCGTGACCACCTTCACCTACGG--TGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	2544	47.5%	-2
10	CACCCTCGTGACCACCTTCACCTACG-CGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	318	68.4%	-1
	CACCCTCGTGACCACCTTCACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	75	16.1%	wt
Cas9	CACCCTCGTGACCACCTTCACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAA	3795	96.7%	wt

Event	GFP 3' Sequence	# Seqs.	Freq	Δ
	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACGGCATG			
1	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACGGCATG	3691	59.9%	wt
	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACCGCATG	1267	20.6%	wt
	GGTCCTGCTGGAGTTCGTGACCGCCG-----ATCACTCACGGCATG	249	4.04%	-5
2	GGTCCTGCTGGAGTTCGTGACCGCCG---GGATCACTCACGGCATG	7438	72.5%	-3
	GGTCCTGCTGGAGTTCGTGACCGCCG---GGATCACTCACGGCCTG	2319	22.6%	SNP, -3
3	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACGGCATG	8738	72.0%	wt
	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACCGCATG	2795	23.0%	wt
4	GGTCCTGCTGGAGTTCGTGACCGCCG-----ATCACTCACGGCATG	8809	34.1%	-5
	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACGGCATG	7869	30.4%	wt
	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACCGCATG	2597	10.0%	wt
	GGTCCTGCTGGAGTTCGTGACCGCCG------TCACTCACGGCATG	1565	6.1%	-6
	GGTCCTGCTGGAGTTCGTGACCGCCG-----ATCACTCACGGCATC	1070	4.1%	-5, SNP
5	GGTCCTGCTGGAGTTCGTGACCGCCGT-----TCACTCACGGCATG	2307	86.5%	-6, +1
	GGTCCTGCTGGAGTTCGTGACCGCCGT-----TCACTCACGGCATC	222	10.2%	SNP
6	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACGGCATG	2179	40.7%	wt
	GGTCCTGCTGGAGTTCGTGACCGCCG-----ATCACTCACGGCATG	1232	23.0%	-5
	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACCGCATG	773	14.4%	wt
7	GGTCCTGCTGGAGTTCGTGACCGCCCCCGGGATCACTCACGGCATG	1694	26.8%	SNP
	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACGGCATG	1338	21.2%	wt
	GGTCCTGCTGGAGTTCGTGACCGCCG--GGGATCACTCACGGCATG	780	12.4%	-2
	GGTCCTGCTGGAGTTCGTGACCGCCCCCGGGATCACTCACCGCATG	625	9.9%	SNP
	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACCGCATG	474	7.5%	wt
8	GGTCCTGCTGGAGTTCGTGACCGCC--CGGGATCACTCACGGCATG	2877	41.9%	-2
	GGTCCTGCTGGAGTTCGTGACCGCCG--------ACTCACGGCATG	1640	23.9%	-8
	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACGGCATG	1276	18.6%	wt
9	GGTCCTGCTGGAGTTCGTGACCGCCG------TCACTCACGGCATG	1581	85.6%	-6

	GGTCCTGCTGGAGTTCGTGACCGCCG------TCACTCACGGCATG	135	7.3%	-6
10	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACGGCATG	4103	49.3%	wt
	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACCGCATG	1614	19.4%	wt
	GGTCCTGCTGGAGTTCGTGACCGCCG------TCACTCACGGCATG	1480	17.8%	-6
Cas9	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACGGCATG	4545	67.6%	wt
	GGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACCGCATG	1843	27.4%	wt

Event	07g 14530Sequence	# Seqs.	Freq	Δ
	CAAAAAGTGGGTGTGAATGTTTATTGT..GGTTG...GATAAACAACCACCTAATTCA			
1	CAAAAAGTGGGTGTGAATGTT------..GGTTG...GATAAACAACCACCTAATTCA	8050	48.5%	-6
	CAAAAAGTGGGTGTGAATGTTTATTGT..GGTTG...GATAAACAACCACCTAATTCA	6315	38.1%	wt
2	CAAAAAGTGGGTGTGAATGTTTAT---..GGTTG...GATAAACAACCACCTAATTCA	25419	50.6%	-3
	CAAAAAGTGGGTGTGAATGTTTATTGT..GGTTG...GATAAACAACCACCTAATTCA	14950	29.8%	wt
	CAAAAAGTGGGTGTGAATGTTTATTGT.TGGTTG...GATAAACAACCACCTAATTCA	4368	8.7%	+1
3	CAAAAAGTGGGTGTGAATGTTTATT--..GGTTG...GATAAACAACCACCTAATTCA	7111	49.1%	-2
	CAAAAAGTGGGTGTGAATGTTTATTGT..-GTTG...GATAAACAACCACCTAATTCA	4652	32.1%	-1
	CAAAAAGTGGGTGTGAATGTTTATTGT.TGGTTG...GATAAACAACCACCTAATTCA	1794	12.4%	+1
4	AGATCACTCTTTGAT----------..-----...--------ACCACCTAATTCA	5946	39.0%	-25, insertion
	CAAAAAGTGGGTGTGAATGTTTATTGT..GGTTG...GATAAACAACCACCTAATTCA	4787	31.4%	wt
	CAAAAAGTGGGTGTGAATGTTTATT--..GGTTG...GATAAACAACCACCTAATTCA	1337	8.8%	-2
	CAAAAAGTGGGTGTGAATGTTTATTGT..GGTGG...----------	1118	7.3%	SNP,-34
	CAAAAAGTGGGTGTGAATGTTTATT-T..GGTTG...GATAAACAACCACCTAATTCA	604	4.0%	-1
5	CAAAAAGTGGGTGTGAATGTTTATTGT..GGTTG...GATAAACAACCACCTAATTCA	10860	95.9%	wt
6	CAAAAAGTGGGTGTGAATGTTT-----..GGTTG...GATAAACAACCACCTAATTCA	39937	50.0%	-5
	CAAAAAGTGGGTGTGAATGTTTATTGT.TGGTTG...GATAAACAACCACCTAATTCA	9864	12.3%	+1
	CAAAAAGTGGGTGTGAATGTTTATTGT..GGTTG...GATAAACAACCACCTAATTCA	9444	11.8%	wt
	CAAAAAGTGGGTGTGAATGTTTATTGT..GG---...-ATAAACAACCACCTAATTCA	6596	8.3%	-4

	CAAAAAGTGGGTGTGAATGTTTATTGTAAGGTTG...GATAAACAACCACCTAATTCA	4070	5.1%	+2
7	CAAAAAGTGGGTGTGAATGTTTATTGT..GGTTG...GATAAACAACCACCTAATTCA	10233	26.3%	wt
	CAAAAAGTGGGTGTGAATGTTTAT-GT..GTCAACATGATAAACAACCACCTAATTCA	6818	17.5%	+2
	CAAAAAGTGGGTGTGAATGTTTAT---..GGTTG...GATAAACAACCACCTAATTCA	6715	17.3%	-3
	CAAAAAGTGGGTGTGAATGTTTATTGT.TGGTTG...GATAAACAACCACCTAATTCA	5623	14.5%	+1
	CAAAAAGTGGGTGTGAATGTTTATT--..GGTTG...GATAAACAACCACCTAATTCA	3509	9.0%	-2
8	CAAAAAGTGGGTGTGAATGTTTATTGT.AGGTTG...GATAAACAACCACCTAATTCA	5085	42.1%	+1
	CAAAAAGTGGGTGTGAATGTTTATT--..GGTTG...GATAAACAACCACCTAATTCA	5009	41.5%	-2
	CAAAAAGTGGGTGTGAATGTTTATTGT..GGTTG...GATAAACAACCACCTAATTCA	1542	12.8%	wt
9	CAAAAAGTGGGTGTGAATGT-------..-----...--------------------	21219	47.2%	-46
	CAAAAAGTGGGTGTGAATGTTTATTGT.TGGTTG...GATAAACAACCACCTAATTCA	19064	42.4%	+1
10	CAAAAAGTGGGTGTGAATGTTTATT--..GGTTG...GATAAACAACCACCTAATTCA	10659	37.8%	-2
	CAAAAAGTGGGTGTGAATGTTTATTG-..AGTTG...GATAAACAACCACCTAATTCA	5808	20.6%	-1
	CAAAAAGTGGGTGTGAATGTTTA----..GGTTG...GATAAACAACCACCTAATTCA	5575	19.8%	-4
	CAAAAAGTGGGTGTGAATGTTTATTGT.TGGTTG...GATAAACAACCACCTAATTCA	5002	17.8%	+1
Cas9	CAAAAAGTGGGTGTGAATGTTTATTGT..GGTTG...GATAAACAACCACCTAATTCA	29923	96.2%	wt

Event	DDM chr1 sequence	# Seqs.	Freq	Δ
	AGAGAAGCTACTTGAAGCTAGGAT-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT			
1	AGAGAAGCTACTT----------AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	820	46.0%	-11
	AGAGAAGCTACTTGAAG------T-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	770	43.2%	-6
	AGAGAAGCTACTTGAAGCTAGGAT-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	94	5.3%	wt
2	AGAGAAGCTACTTGAAG------T-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	1008	86.0%	-6
	AGAGAAGCTACTTGAAGCTAGGAT-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	97	8.3%	wt
3	AGAGAAGCTACTTGAAGCTA---T-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	770	83.4%	-3
	AGAGAAGCTACTTGAAGCTAGGAT-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	107	11.6%	wt
4	AGAGAAGCTACTTGAAGCT----T-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	1185	46.4%	-4
	AGAGAAGCTACTTGAAGCTA------AAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	1101	43.1%	-5

	AGAGAAGCTACTTGAAGCTAGGAT-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	170	6.7%	wt
5	AGAGAAGCTACTTGAAGCTAGGAT-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	4321	96.1%	wt
6	AGAGAAGCTACTTGAAGCTA------AAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	1367	86.6%	-5
	AGAGAAGCTACTTGAAGCTAGGAT-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	116	7.3%	wt
7	AGAGAAGCTACTTGAA-------T-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	2749	47.4%	-7
	AGAGAAGCTACTTGAAGCTAGGA----AGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	2698	46.5%	
8	AGAGAAGCTACTTGAAGCTAGGAT-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	2140	95.5%	wt
9	AGAGAAGCTACTTGAAGCTA---T-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	3368	59.1%	-3
	AGAGAAGCTACTTGAAGCTAGGAT-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	695	12.2%	wt
	AGAGAAGCTACTTGAAGCTAGGATTAAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	358	6.3%	+1
10	AGAGAAGCTACTTGAAGCTAGGA--AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	1839	89.6%	-1
	AGAGAAGCTACTTGAAGCTAGGAT-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	97	4.7%	wt
Cas9	AGAGAAGCTACTTGAAGCTAGGAT-AAAGGAAGAGGAGGTGCAGTATGAGGAGGCAGT	16066	97.9%	wt

Event	DDM chr11 sequence	# Seqs.	Freq	Δ
	AATTGAAGCTAGGATGAAGGAAGAGGAGGTACAGTGTGAGGAGGTACCTGACCTTAAT			
1	AATTGAAGCTAGGATGAAGGAAGAGGAGGTA---TGTGAGGAGGTACCTGACCTTAAT	21199	48.4%	-3
	AATTGAAGCTAGGATGAAGGAAGAGGA-------TGTGAGGAGGTACCTGACCTTAAT	19516	44.5%	-7
	----------GTGTGAGGAGGTACCTGACCTTAAT	1926	4.4%	-44
2	AATTGAAGCTAGGATGAAGGAAGAGGAGGTA---TGTGAGGAGGTACCTGACCTTAAT	20757	50.3%	-3
	AATTGAAGCTAGGATGAAGGAAGAGGAGGTA----GTGAGGAGGTACCTGACCTTAAT	19505	47.2%	-4
3	AATTGAAGCTAGGATGAAGGAAGAGGAGGTAC--TGTGAGGAGGTACCTGACCTTAAT	16111	91.5%	-2
	AATTGAAGCTAGGATGAAGGAAGAGGAGGTACAGTG-GAGGAGGAAGAAGTGACT..	841	4.8%	insert
4	AATTGAAGCTAGGATGAAGGAAGAGGAGGTACAACCAACGTCTTCGCCATACCGA..	5337	79.2%	insert
	AATTGAAGCTAGGATGAAGGAAGAGGAGGT-----GTGAGGAGGTACCTGACCTTAAT	412	6.1%	-5
	AATTGAAGCTAGGATGAAGGAAGAGGAGGT----TGTGAGGAGGTACCTGACCTTAAT	272	4.0%	-4
5	AATTGAAGCTAGGATGAAGGAAGAGGAGGTAC--------------CTGACCTTAAT	21126	48.8%	-15
	AATTGAAGCTAGGATGAAGGAAGAGGAGGTA---TGTGAGGAGGTACCTGACCTTAAT	19548	45.2%	-3

6	AATTGAAGCTAGGATGAAGGAAG----------GTACCTGACCTTAAT	23260	51.2%	-20
	AATTGAAGCTAGGATGAAGGAAGAGGAGGTA---TGTGAGGAGGTACCTGACCTTAAT	21190	46.7%	-3
7	AATTGAAGCTAGGATGAAGGAAGAGGAGGTACAGTGTGAGGAGGTACCTGACCTTAAT	83999	97.8%	wt
8	AATTGAAGCTAGGATGAAGGAAGAGGAGGTA----------CCTGACCTTAAT	44717	64.7%	-15
	AATTGAAGCTAGGATGAAGGAAGAGGAGGTAC--TGTGAGGAGGTACCTGACCTTAAT	22916	33.2%	-2
9	AATTGAAGCTAGGATGAAGGAAAAC----------AAATGTACTCTGACCTTAAT	27888	53.1%	-22, +9
	AATTGAAGCTAGGATGAAGGAAGAGGAGGTACA-TGTGAGGAGGTACCTGACCTTAAT	23555	44.8%	-1
10	AATTGAAGCTAGGATGAAGGAAGAGGAGGTA----------CCTGACCTTAAT	30806	50.9%	-15
	AATTGAAGCTAGGATGAAGGAAGAGGAGGTA---TGTGAGGAGGTACCTGACCTTAAT	28429	47.0%	-3
11	AATTGAAGCTAGGATGAAGGAAGAGGAGGTACAGTGTGAGGAGGTACCTGACCTTAAT	25721	98.4%	wt

pink is insertion of A. rhizogenes plasmid

Event	both DDM, chr1 Sequence	# Seqs.	Freq	Δ
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA			
1	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	6070	63.7%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAAATGGGTCTTGGGAAGACAATCCAA	2434	25.5%	+1
2	TGTGGCAAAATGGATTGAATGGGATTCTTGC-----.---.--GGTCTTGGGAAGACAATCCAA	2139	34.2%	-10
	TGTGGCAAAATGGATTGAATGGGATTCTTGC----C.AAA.TGGGTCTTGGGAAGACAATCCAA	2100	33.6%	-4
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	1804	28.8%	wt
3	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	14415	90.3%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAAATGGGTCTTGGGAAGACAATCCAA	450	2.8%	+1
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGA--.--A.-GGTCTTGGGAAGACAATCCAA	193	1.2%	-5
4	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	1993	73.0%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCT----.--A.TGGGTCTTGGGAAGACAATCCAA	668	24.5%	-6
5	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	13553	93.1%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAAATGGGTCTTGGGAAGACAATCCAA	414	2.8%	+1
6	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	14145	88.2%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGA--.AAA.TGGGTCTTGGGAAGACAATCCAA	1553	9.7%	-2

7	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	7341	96.0%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAAATGGGTCTTGGGAAGACAATCCAA	76	1.0%	+1
8	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	4673	91.6%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAAATGGGTCTTGGGAAGACAATCCAA	159	3.1%	+1
9	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	6703	92.8%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAAATGGGTCTTGGGAAGACAATCCAA	268	3.7%	+1
10	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	5084	58.9%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATCTAAA.TGGGTCTTGGGAAGACAATCCAA	1985	23.0%	+1
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AA-.TGGGTCTTGGGAAGACAATCCAA	1042	12.1%	-1
Cas9	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	11996	98.3%	wt

Event	both DDM, chr11 Sequence	# Seqs.	Freq	Δ
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAATGGGTCTTGGGAAGACAATCCAA			
1	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAATGGGTCTTGGGAAGACAATCCAA	461	68.4%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGA--.AAATGGGTCTTGGGAAGACAATCCAA	89	13.2%	-2
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATCAAAATGGGTCTTGGGAAGACAATCCAA	56	8.3%	+1
2	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAATGGGTCTTGGGAAGACAATCCAA	199	92.6%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATCAAAATGGGTCTTGGGAAGACAATCCAA	6	2.8%	+1
3	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAATGGGTCTTGGGAAGACAATCCAA	792	86.4%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCT----.-AATGGGTCTTGGGAAGACAATCCAA	35	3.8%	-5
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATCAAAATGGGTCTTGGGAAGACAATCCAA	34	3.7%	+1
4	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAATGGGTCTTGGGAAGACAATCCAA	107	96.4%	wt
5	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAATGGGTCTTGGGAAGACAATCCAA	869	90.5%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATCAAAATGGGTCTTGGGAAGACAATCCAA	46	4.8%	+1
6	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAATGGGTCTTGGGAAGACAATCCAA	1171	94.2%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGAT-.--ATGGGTCTTGGGAAGACAATCCAA	38	3.1%	-3
7	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAATGGGTCTTGGGAAGACAATCCAA	424	96.6%	wt
8	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAATGGGTCTTGGGAAGACAATCCAA	245	85.1%	wt

	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGAT-.AAATGGGTCTTGGGAAGACAATCCAA	18	6.3%	-1
9	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAATGGGTCTTGGGAAGACAATCCAA	341	92.7%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATCAAAATGGGTCTTGGGAAGACAATCCAA	15	4.1%	+1
10	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAATGGGTCTTGGGAAGACAATCCAA	509	93.6%	wt
	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.-AATGGGTCTTGGGAAGACAATCCAA	11	2.0%	-1
Cas9	TGTGGCAAAATGGATTGAATGGGATTCTTGCTGATC.AAATGGGTCTTGGGAAGACAATCCAA	2243	98.6%	wt

Event	04gMet1 Sequence	# Seqs.	Freq	Δ
	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGTCTATGAGTGGTAGCAAATGCTTTTCTGGAG			
1	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGT----GAGTGGTAGCAAATGCTTTTCTGGAG	135576	96.7%	-4
2	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGT----GAGTGGTAGCAAATGCTTTTCTGGAG	45743	96.5%	-4
3	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGT----GAGTGGTAGCAAATGCTTTTCTGGAG	21	58.3%	-4
	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGTCT--GAGTGGTAGCAAATGCTTTTCTGGAG	6	16.7%	-2
	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGTCTATGAGTGGTAGCAAATGCTTTTCTGGAG	5	13.9%	wt
4	GTCTTGATGAGTTACTTGCTGGCATGGTGCGG------AGTGGTAGCAAATGCTTTTCTGGAG	37770	48.3%	-5
	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGTCT--GAGTGGTAGCAAATGCTTTTCTGGAG	37226	47.6%	-2
5	GTCCTCA------------------------------AGTGGTAGCAAATGCTTTTCTGGAG	54793	51.0%	-31
	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGTCT--GAGTGGTAGCAAATGCTTTTCTGGAG	48309	45.0%	-2
Cas9	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGTCTATGAGTGGTAGCAAATGCTTTTCTGGAG	75850	96.8%	wt

Event	06gMet1 Sequence	# Seqs.	Freq	Δ
	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGTCTAT.GAGTGGTAGCAAATGCTTTTCTGGAA			
1	GTCTTGATGAGTTACTTGCTGGCATGGTGCGG-----.-AGTGGTAGCAAATGCTTTTCTGGAA	78477	48.7%	-6
	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGT----.GAGTGGTAGCAAATGCTTTTCTGGAA	77433	48.0%	-4
2	GTCTTGATGAGTTACTTGCTGGCATGGTGCGG-----.----GGTAGCAAATGCTTTTCTGGAA	46617	74.8%	-9
	GCGGAGAGACAGAGACATAAATAACAAAATCCTTTATAGATGCAGTTCCAGAAAAGCATTTGCT	12366	19.9%	inversion
3	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGT---T.GAGTGGTAGCAAATGCTTTTCTGGAA	8217	55.8%	-3
	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGTCTATTGAGTGGTAGCAAATGCTTTTCTGGAA	5986	40.7%	+1

4	GTCTTGATGAGTTACTTGCTGGCATGGTGCGG-----.-AGTGGTAGCAAATGCTTTTCTGGAA	61583	48.7%	-6
	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGTC---.GAGTGGTAGCAAATGCTTTTCTGGAA	60313	47.7%	-3
5	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGT---T.GAGTGGTAGCAAATGCTTTTCTGGAA	80835	97.1%	-3
Cas9	GTCTTGATGAGTTACTTGCTGGCATGGTGCGGTCTAT.GAGTGGTAGCAAATGCTTTTCTGGAA	110961	97.1%	wt

underlined is part of inverted repeat

Event	Gma-miR1509 Sequence	# Seqs.	Freq	Δ
	ACTGCATCTTTTTAATCAAGGAAATCACGGTTGAGTGTGAAGGAGAGAAAGTGGCTTCAGATT			
1	ACTGCATCTTTTTAATCAAGGAAATCACGGTTGAGTG----------GCTTCAGATT	9491	50.5%	-16
	ACTGCATCTTTTTAATCAAGGAAATCACGGTT-----TGAAGGAGAGAAAGTGGCTTCAGATT	8980	47.8%	-5
2	ACTGCATCTTTTTAATCAAGGAAATCACGGTT-----TGAAGGAGAGAAAGTGGCTTCAGATT	21187	50.7%	-5
	ACTGCATCTTTTTAATCAAGGAAATCACGGTTGAGT-TGAAGGAGAGAAAGTGGCTTCAGATT	19818	47.4%	-1
3	ACTGCATCTTTTTAATCAAGGAAATCACGGTTGAGT-TGAAGGAGAGAAAGTGGCTTCAGATT	10705	97.4%	-1
4	ACTGCATCTTTTTAATCAAGGAAATCACGGTTGAG----------AAAGTGGCTTCAGATT	5159	97.7%	-12
Cas9	ACTGCATCTTTTTAATCAAGGAAATCACGGTTGAGTGTGAAGGAGAGAAAGTGGCTTCAGATT	18479	98.0%	wt

Event	Gma-miR 1514 Sequence	# Seqs.	Freq	Δ
	TGTTTTCATTTTAAAATAGGCATTGGGATAGGAAAGGAAAAGGAGGAACAAGAAGGGACCCCA			
1	TGTTTTCATTTTAAAATAGGCATTGGGATAGGAAAGGAAAAGGAGGAACAAGAAGGGACCCCA	41873	93.9%	wt
	TGTTTTCATTTTAAAATAGGCATTGGGATAGGAAA--AAAAGGAGGAACAAGAAGGGACCCCA	1575	3.5%	-2
2	TGTTTTCATTTTAAAATAGGCATTGGGATAGGA----------GGAACAAGAAGGGACCCCA	20285	50.0%	-11
	TGTTTTCATTTTAAAATAGGCATTGGGATAGGAAA---AAAGGAGGAACAAGAAGGGACCCCA	19498	48.1%	-3
3	TGTTTTCATTTTAAAATAGGCATTGGGATAGGAAAG-AAAAGGAGGAACAAGAAGGGACCCCA	5893	48.8%	-1
	TGTTTTCATTTTAAAATAGGCATTGGGATAGGAAA--AAAAGGAGGAACAAGAAGGGACCCCA	5696	47.2%	-2
4	TGTTTTCATTTTAAAATAGGCATTGGGATAGGA----------GGAACAAGAAGGGACCCCA	16572	25.6%	-11
	TGTTTTCATTTTAAAATAGGCATTGGGATAG------AAAAGGAGGAACAAGAAGGGACCCCA	16322	25.2%	-6
	TGTTTTCATTTTAAAATAGGCATTGGGATAGGAAA--AAAAGGAGGAACAAGAAGGGACCCCA	15342	23.7%	-2
	TGTTTTCATTTTAAAATAGGCATTGGGATAGGAAA----AAGGAGGAACAAGAAGGGACCCCA	7947	12.3%	-4

	TGTTTTCATTTTAAAATAGGCATTGGGATAGGAAA-----AGGAGGAACAAGAAGGGACCCCA	4643	7.2%	-5
	TGTTTTCATTTTAAAATAGGCATTGGGATAGGAAAGGAAAAGGAGGAACAAGAAGGGACCCCA	2502	3.9%	wt
Cas9	TGTTTTCATTTTAAAATAGGCATTGGGATAGGAAAGGAAAAGGAGGAACAAGAAGGGACCCCA	35599	97.8%	wt

Shot event 24, bothDDM 1g, timepoint 1	# Seqs.	Freq	Δ
TGGATTGAATGGGATTCTTGCTGATCAAA.TGGGTCTTGGGAAGACAATCCAA	45726	95.43%	wt
TGGATTGAATGGGATTCTTGCTGATCAAAATGGGTCTTGGGAAGACAATCCAA	353	0.74%	+1
TGGATTGAATGGGATTCTTGCTGATCAAA.TGTGTCTTGGGAAGACAATCCAA	201	0.42%	wt
TGGATTGAATGGGATTCTTGCTGATCAA-.TGGGTCTTGGGAAGACAATCCAA	117	0.24%	-1
TGGATTGAATGGGATTCTTGC-----AAA.TGGGTCTTGGGAAGACAATCCAA	74	0.15%	-5
TGGATTGAATGGGATTCTTGCTGATCAAA.TGGGTCTTGGGAAGACAATCCAA	68	0.14%	wt
TGGATTGAATGGGATTCTTGCT---CAAA.TGGGTCTTGGGAAGACAATCCAA	68	0.14%	-3
TGGATTGAATGGGATTCTTGCT----AAA.TGGGTCTTGGGAAGACAATCCAA	67	0.14%	-4
TGGATTGAATGGGATTCTTGCTGATCAAA.TGCGTCTTGGGAAGACAATCCAA	57	0.12%	wt
TGGATTGAATGGGATTCTTGC----CAAA.TGGGTCTTGGGAAGACAATCCAA	54	0.11%	-4

Shot event 24, bothDDM 11g, timepoint 1	# Seqs.	Freq	Δ
TGGATTGAATGGGATTCTTGCTGAT.C.AAATGGGTCTTGGGAAGACAATCCAA	14863	94.83%	wt
TGGATTGAATGGGATTCTTGCTGAT.CAAAATGGGTCTTGGGAAGACAATCCAA	116	0.74%	+1
TGGATTGAATGGGATTCTTGCTGAT.C.AAATGTGTCTTGGGAAGACAATCCAA	65	0.41%	wt
TGGATTGAATGGGATTCTTG-----.C.AAATGGGTCTTGGGAAGACAATCCAA	36	0.23%	-5
TGGATTGAATGGGATTCTTGCTGAT.C.--ATGGGTCTTGGGAAGACAATCCAA	34	0.22%	-2
TGGATTGAATGGGATTCTTGCTG--.-.-AATGGGTCTTGGGAAGACAATCCAA	32	0.20%	-4
TGGATTGAATGGGATTCTTGCTGAT.C.-AATGGGTCTTGGGAAGACAATCCAA	25	0.16%	-1
TGGATTGAATGGGATTCTTGCTGAT.C.AAATGGGTCTTGGGAAGACAATCCAA	20	0.13%	wt
TGGATTGAATGGGATTCTTGCTGATCC.AAATGGGTCTTGGGAAGACAATCCAA	20	0.13%	+1
TGGATTGAATGGGATTCTTGCTG--.-.--ATGGGTCTTGGGAAGACAATCCAA	17	0.11%	-5

Shot event 10, bothDDM 1g, timepoint 2			
TGGATTGAATGGGATTCTTGCTGATCAAA.TGGGTCTTGGGAAGACAATCCAA	13997	77.65%	wt
TGGATTGAATGGGATTCTTGC-----AAA.TGGGTCTTGGGAAGACAATCCAA	3436	19.06%	-5
TGGATTGAATGGGATTCTTGCTGATCAAAATGGGTCTTGGGAAGACAATCCAA	185	1.03%	+1
TGGATTGAATGGGATTCTTGCTGATCAA-.TGGGTCTTGGGAAGACAATCCAA	19	0.11%	-1
TGGATTGAATGGGATTCTTGCTGAT----.-GGGTCTTGGGAAGACAATCCAA	16	0.09%	-5
TGGATTGAATGGGATTC---------AAA.TGGGTCTTGGGAAGACAATCCAA	16	0.09%	-9
TGGATTGAATGGGATTCTTGCTGATCAAA.TGGGTCTTGGGAAGACAATCCAA	14	0.08%	wt
TGGATTGAATGGGATTCTTGCT----AAA.TGGGTCTTGGGAAGACAATCCAA	12	0.07%	-4
TGGATTGAATGGGATTCTTGCTGATCAAA.TGGGTCTTGGGAAGACAATCCAA	10	0.06%	wt
TGGATTGAATGGGATTCTTGCT-----AA.TGGGTCTTGGGAAGACAATCCAA	10	0.06%	-5

shot event 21, bothDDM 1g, timepoint 2			
TGGATTGAATGGGATTCTTGCTGATCAAA.TGGGTCTTGGGAAGACAATCCAA	25034	95.82%	wt
TGGATTGAATGGGATTCTTGCTGATCAAAATGGGTCTTGGGAAGACAATCCAA	220	0.84%	+1
TGGATTGAATGGGATTCTTGCTGA--AAA.TGGGTCTTGGGAAGACAATCCAA	110	0.42%	-2
TGGATTGAATGGGATTCTTGCTGATCAA-.TGGGTCTTGGGAAGACAATCCAA	44	0.17%	-1
TGGATTGAATGGGAT----------.-GGGTCTTGGGAAGACAATCCAA	43	0.16%	-15
TGGATTGAATGGGATTCTTGC-----AAA.TGGGTCTTGGGAAGACAATCCAA	33	0.13%	-5
TGGATTGAATGGGATTCTTGCTGATCAAA.TGGGTCTTGGGAAGACAATCCAA	28	0.11%	wt
TGGATTGAATGGGATTCTTGCTGA----A.TGGGTCTTGGGAAGACAATCCAA	27	0.10%	-4
TGGATTGAATGGGATTCTTGCT----AAA.TGGGTCTTGGGAAGACAATCCAA	22	0.08%	-4
TGGATTGAATGGGATTCTTGCTGAT----.-GGGTCTTGGGAAGACAATCCAA	20	0.08%	-5

shot event 24, bothDDM 1g, timepoint 2			
TGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	22659	94.35%	wt
TGGATTGAATGGGATTCTTGCTGATC.AAAATGGGTCTTGGGAAGACAATCCAA	360	1.50%	+1
TGGATTGAATGGGATTCTTGCTGA--.-AA.TGGGTCTTGGGAAGACAATCCAA	95	0.40%	-3

TGGATTGAATGGGATTCTTGCTGATC.AA-.TGGGTCTTGGGAAGACAATCCAA	67	0.28%	-1
TGGATTGAATGGGATTCTTGCT---C.AAA.TGGGTCTTGGGAAGACAATCCAA	63	0.26%	-3
TGGATTGAATGGGATTCTTGCT----.AAA.TGGGTCTTGGGAAGACAATCCAA	54	0.22%	-4
TGGATTGAATGGGATTCTTGC-----.AAA.TGGGTCTTGGGAAGACAATCCAA	48	0.20%	-5
TGGATTGAATGGGATTCTTGCTGATC.AAA.TGGGTCTTGGGAAGACAATCCAA	33	0.14%	wt
TGGATTGAATGGGATTCTTGC-----.-AA.TGGGTCTTGGGAAGACAATCCAA	32	0.13%	-6
TGGATTGAATGGGATTCTTGCTGATCCAAA.TGGGTCTTGGGAAGACAATCCAA	31	0.13%	+1

shot event 10, bothDDM 1g, timepoint 2

TGGATTGAATGGGATTCTTGCTGATCAAA.TGGGTCTTGGGAAGACAATCCAA	16111	80.31%	wt
TGGATTGAATGGGATTCTTGCTGA--AAA.TGGGTCTTGGGAAGACAATCCAA	2033	10.13%	-2
TGGATTGAATGGGATTCTTGCTGA---AA.TGGGTCTTGGGAAGACAATCCAA	952	4.75%	-3
TGGATTGAATGGGATTCTTGCTGATCAAAATGGGTCTTGGGAAGACAATCCAA	226	1.13%	+1
TGGATTGAATGGGATTCTTGC-----AAA.TGGGTCTTGGGAAGACAATCCAA	65	0.32%	-5
TGGATTGAATGGGATTCTTGCTGATCAA-.TGGGTCTTGGGAAGACAATCCAA	42	0.21%	-1
TGGATTGAATGGGATTCTTGCT---CAAA.TGGGTCTTGGGAAGACAATCCAA	29	0.14%	-3
TGGATTGAATGGGATTCTTGCTGAT-AAA.TGGGTCTTGGGAAGACAATCCAA	26	0.13%	-1
TGGATTGAATGGGATTCTTG---------.----------	24	0.12%	-33
TGGATTGAATGGGATTCTTGCT----AAA.TGGGTCTTGGGAAGACAATCCAA	21	0.10%	-4

shot event 21, bothDDM 1g, timepoint 2

TGGATTGAATGGGATTCTTGCTGATCAAA.TGGGTCTTGGGAAGACAATCCAA	13824	93.82%	wt
TGGATTGAATGGGATTCTTGCTGATCAAAATGGGTCTTGGGAAGACAATCCAA	298	2.02%	+1
TGGATTGAATGGGATTCTTGCTGATCAA-.TGGGTCTTGGGAAGACAATCCAA	38	0.26%	-1
TGGATTGAATGGGATTCTTGC-----AAA.TGGGTCTTGGGAAGACAATCCAA	35	0.24%	-5
TGGATTGAATGGGATTCTTGCTGA---AA.TGGGTCTTGGGAAGACAATCCAA	34	0.23%	-3
TGGATTGAATGGGATTCTTGCT----AAA.TGGGTCTTGGGAAGACAATCCAA	27	0.18%	-4

TGGATTGAATGGGATTCTTGCTGAT-AAA.TGGGTCTTGGGAAGACAATCCAA	25	0.17%	-1
TGGATTGAATGGGATTCTTGCTGAT----.-GGGTCTTGGGAAGACAATCCAA	22	0.15%	-5
TGGATTGAATGGGATTCTTGCTGAT--AA.TGGGTCTTGGGAAGACAATCCAA	21	0.14%	-2
TGGATTGAATGGGATTCTTGCTGATCAAA.TGGGTCTTGGGAAGACAATCCAA	18	0.12%	wt

shot event 24, bothDDM 1g, timepoint 2			
TGGATTGAATGGGATTCTTGCTGATCAAA.TGGGTCTTGGGAAGACAATCCAA	15641	94.32%	wt
TGGATTGAATGGGATTCTTGCTGATCAAAATGGGTCTTGGGAAGACAATCCAA	157	0.95%	+1
TGGATTGAATGGGATTCTTGCTGAT--AA.TGGGTCTTGGGAAGACAATCCAA	97	0.58%	-2
TGGATTGAATGGGATTCTTGCTGA----A.TGGGTCTTGGGAAGACAATCCAA	62	0.37%	-4
TGGATTGAATGGGATTCTTGCTGAT-AAA.TGGGTCTTGGGAAGACAATCCAA	58	0.35%	-1
TGGATTGAATGGGATTCTTGC-----AAA.TGGGTCTTGGGAAGACAATCCAA	58	0.35%	-5
TGGATTGAATGGGATTCTTGCTGATCAA-.TGGGTCTTGGGAAGACAATCCAA	47	0.28%	-1
TGGATTGAATGGGATTCTTGCTGA--AAA.TGGGTCTTGGGAAGACAATCCAA	26	0.16%	-2
TGGATTGAATGGGATTCTTGCTGATC---.-GGGTCTTGGGAAGACAATCCAA	23	0.14%	-4
TGGATTGAATGGGATTCTTGCT----AAA.TGGGTCTTGGGAAGACAATCCAA	15	0.09%	-4

Table 3.1. Unique sequences from all events in this study. The most abundant reads for each event are reported. The number of reads, the respective percentages, and the type of modification (Δ) is listed for each event. Wild-type sequences are in green, dashes are deletions, SNPs are orange, insertions are red, replacements are orange, an insertion of Ri plasmid is pink, and an inversion is in purple.

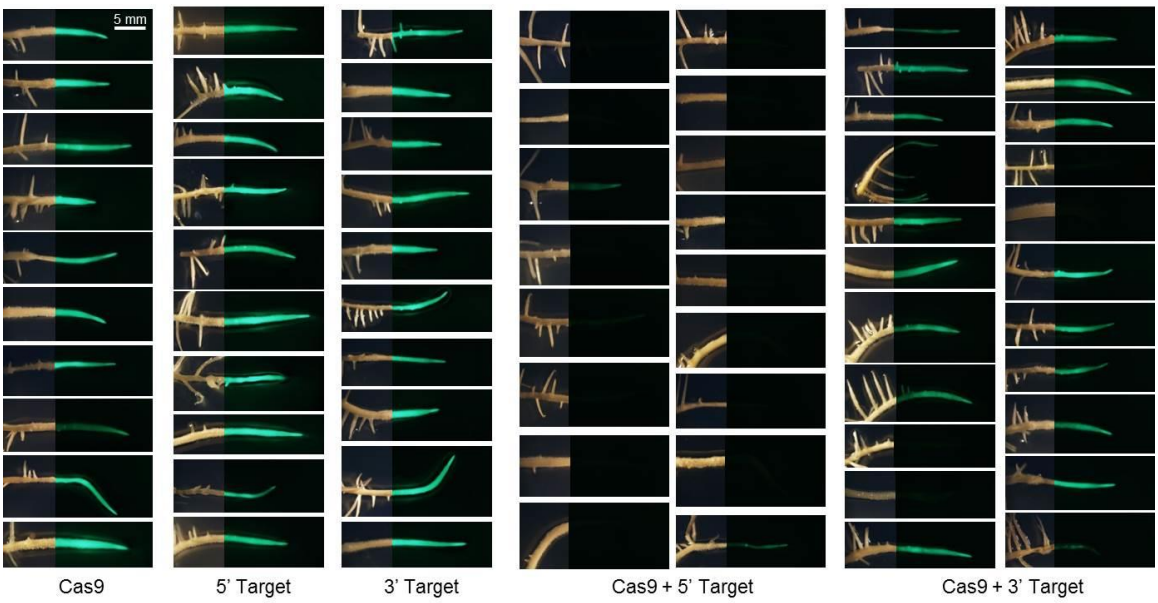


Figure 3.2. GFP imaging of modified GFP events and controls. Each panel is an independent event and blue-light images were overlaid onto white-light images of roots. Scale bar is shown as 5mm and all images are taken with the same magnification.

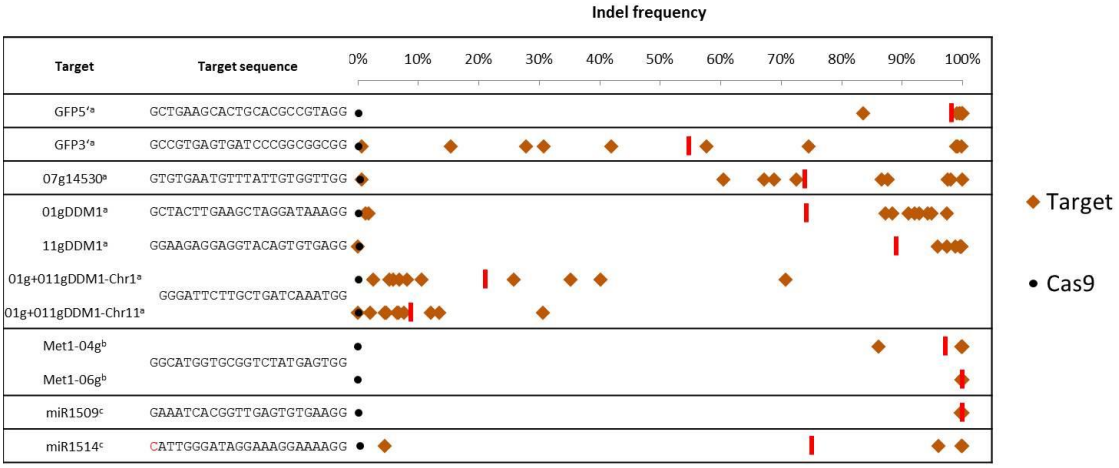


Figure 3.3. Modification efficiency for hairy root events. Custom-amplicon sequencing was used to measure indel frequency for each of the targeting constructs. Individual events are in orange triangles, the Cas9-tranformed control is in black circles, and average indel frequencies are vertical red bars. The MiR1514 target sequence has a single mismatch to the gRNA in red. [a] n=10, [b] n=5, [c] n=4.

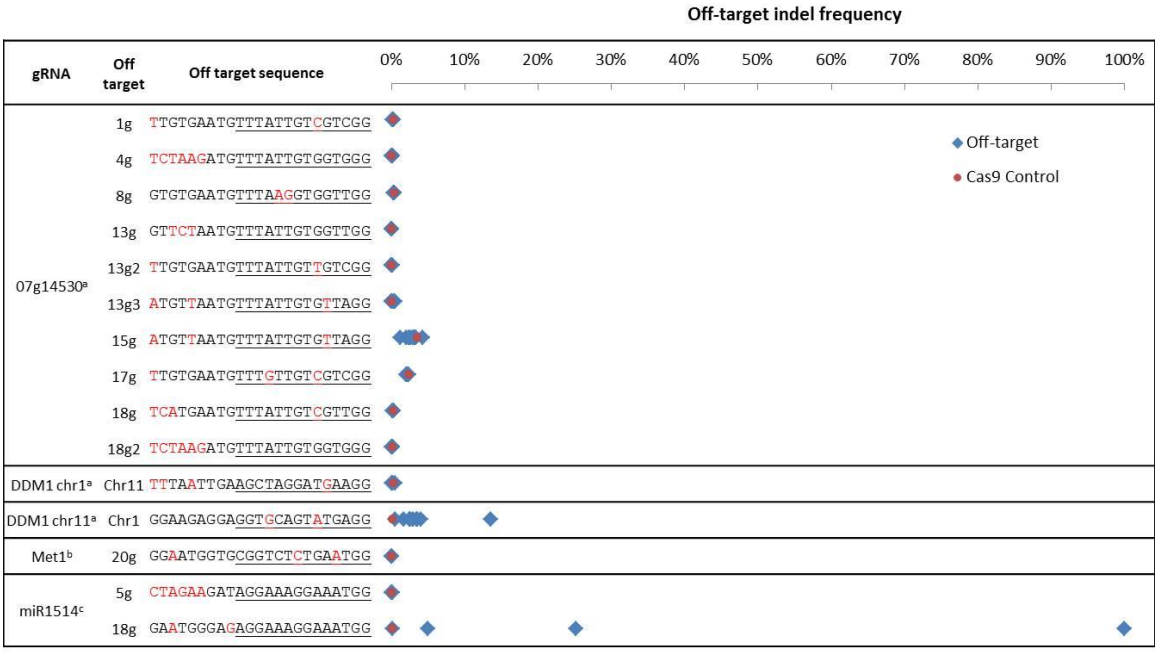


Figure 3.4. Off-target indel frequency for hairy-root events. The measured indel frequency is represented by a blue diamond for each event and a red dot for the Cas9 control. Mismatches between the gRNA and the off-target sequence are in red. The critical 'seed' region is underlined. [a] n=10, [b] n=5, [c] n=4.

NNNNNNNNNNTTNNNGTTCNATTTTTTCAATGNNNGGCTGTCTGTGCCCCAAAAGAAGAAGTTAAGTTGGAGGAAGAAGT

GACTGCAGACATCAAAGATGATGGGACCTCTCTTATATCGAAAACAATGGTGGAGGAGGAAGAGAATTTAATTGAAGCTA

GGATGAAGGAAGAGGAGGTACAACCAACGTCTTCGCCATACCGAATGAGGATGCTCTCTCTGCAATTCTTGGCGATCTCG

GCGTGTTGTTGCCCGATCTGAGCAAGTGATCGCTCCGATCTGCCGCTGTTCTATTACATCCATAGTCACCGTAGGAAGGG

AAAGAAGTCATAATTGCTTCAGTTGTTACTTAGTTTTTTCTATTTTAGTGTCTATGATCCGCCACTCCGGCGACTATTTC

AATACACAACGCCAGAAATAATATAATATCTCCATGTGAGGAGGTACCTGACCTTAATGACACACAGTTTAACAAATTGG

ATGAGCTTTTGACTCAAACCAAACTGTACTCTGAGTTTCTGCTGGAGAAAATGGACGACATCACACTTGTAATTTTTAAC

CCTCTTTGCTTCTTTTTTTTTTCTGTGATAACTTTTTACACGTGTTGATTTGTTGTTCTTGTTGCATCATTCTTATTCTT

TGAATATTTGTTGCTAGGCTGTGGGTGAACAAGAGAATAGGGAAAAGCAAGAGAGCAATCCTTCTGCAAAGAANAAGGGC

TGTGGATCAAAAAGAA

Figure 3.5. Cloned sequence from modified 11gDDM1 event that contains a 252-bp insertion of the Ri plasmid. Red is gRNA target, underline is insertion.

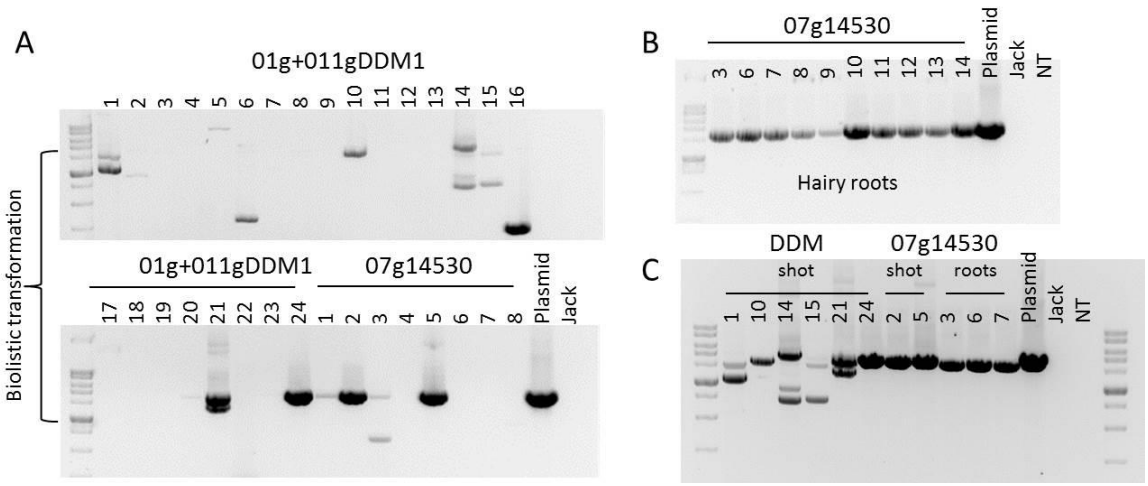


Figure 3.6. Long-distance PCR for the Cas9 gene in somatic embryos (A) and hairy-root events (B). Events positive for Cas9 were re-run together to get appropriate sizing (C). Three 01g+011gDDM1 and two 07g14530 biolistic-events have the correct 4.3kb band. All hairy-root events are positive.

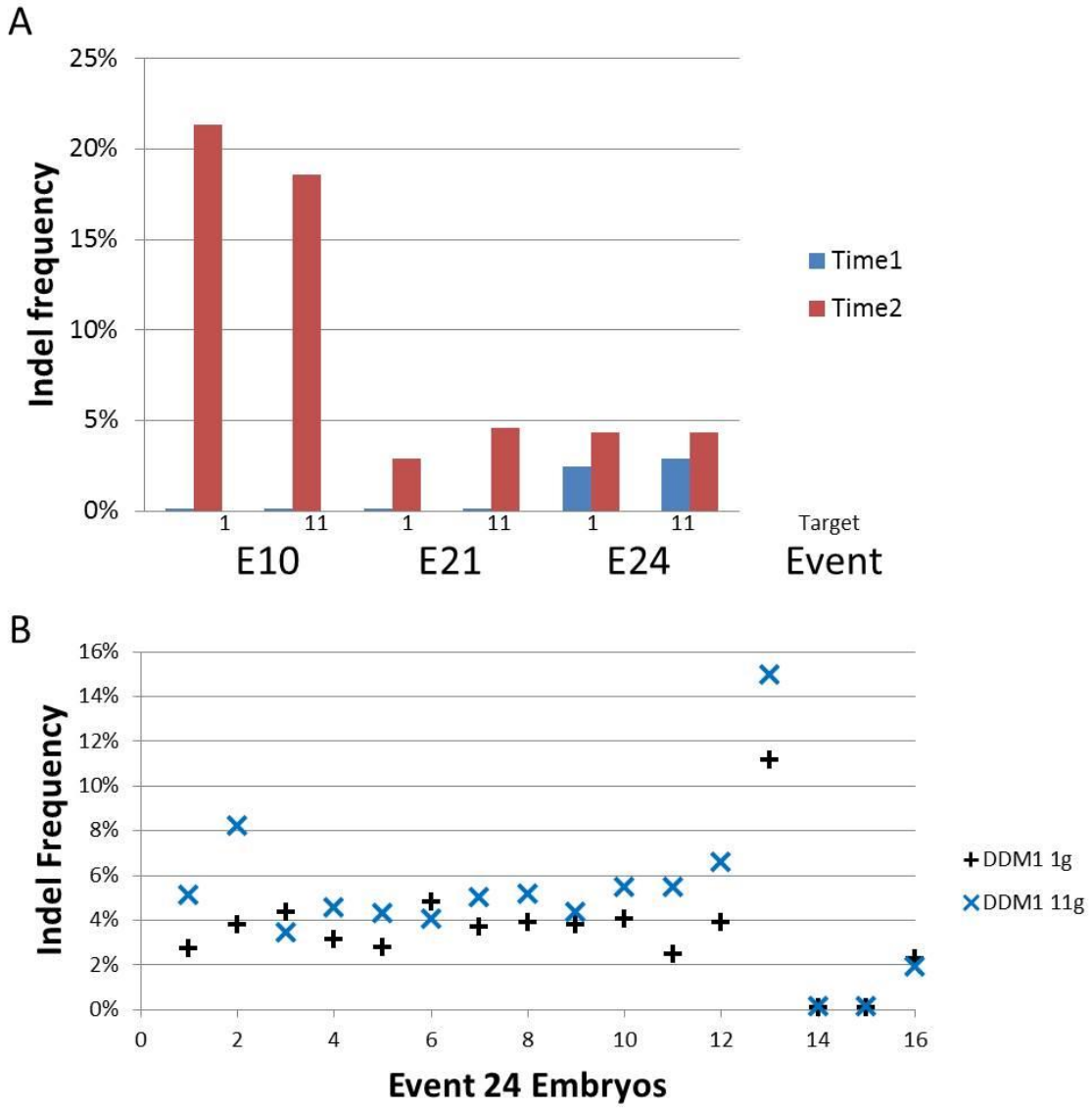


Figure 3.7. Modification of both DDM genes in somatic embryos. (a) Modifications were detected in three events transformed with the bothDDM vector. At the initial time-point, modifications were only detected in event 24. When samples were taken approximately 2 weeks later, modifications were detected in all three events. (b) Modifications were detected in 14 out of 16 individual regenerating embryos from event 24.

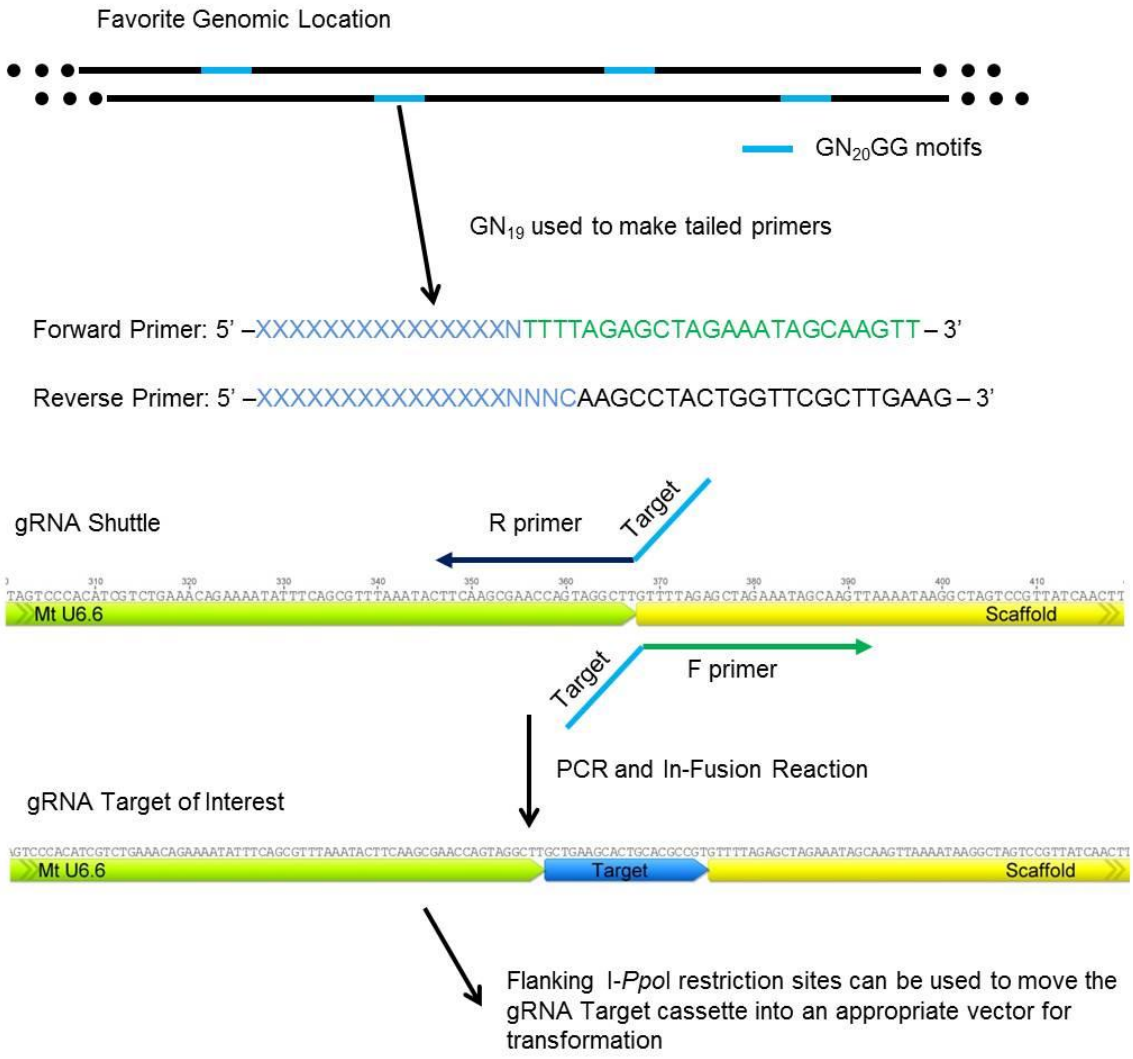


Figure 3.8. Cas9/gRNA targeting and cloning scheme to produce gRNAs. $GN_{20}GG$ motifs are identified in a genomic region of interest. Tailed forward and reverse primers are designed to amplify the entire 3 kb gRNA Shuttle Plasmid. The primer tails contain sequences for the target (blue) and share 15 bp of homology (X's) for the In-Fusion® protocol. PCR products can then undergo In-Fusion® cloning, resulting in the creation of the gRNA Target Plasmid. The gRNA cassette is in the middle of a multiple-cloning site for easy transfer to a final vector. This pUC gRNA Shuttle plasmid can be used for plant modifications, but the cloning scheme will work for any gRNA target.

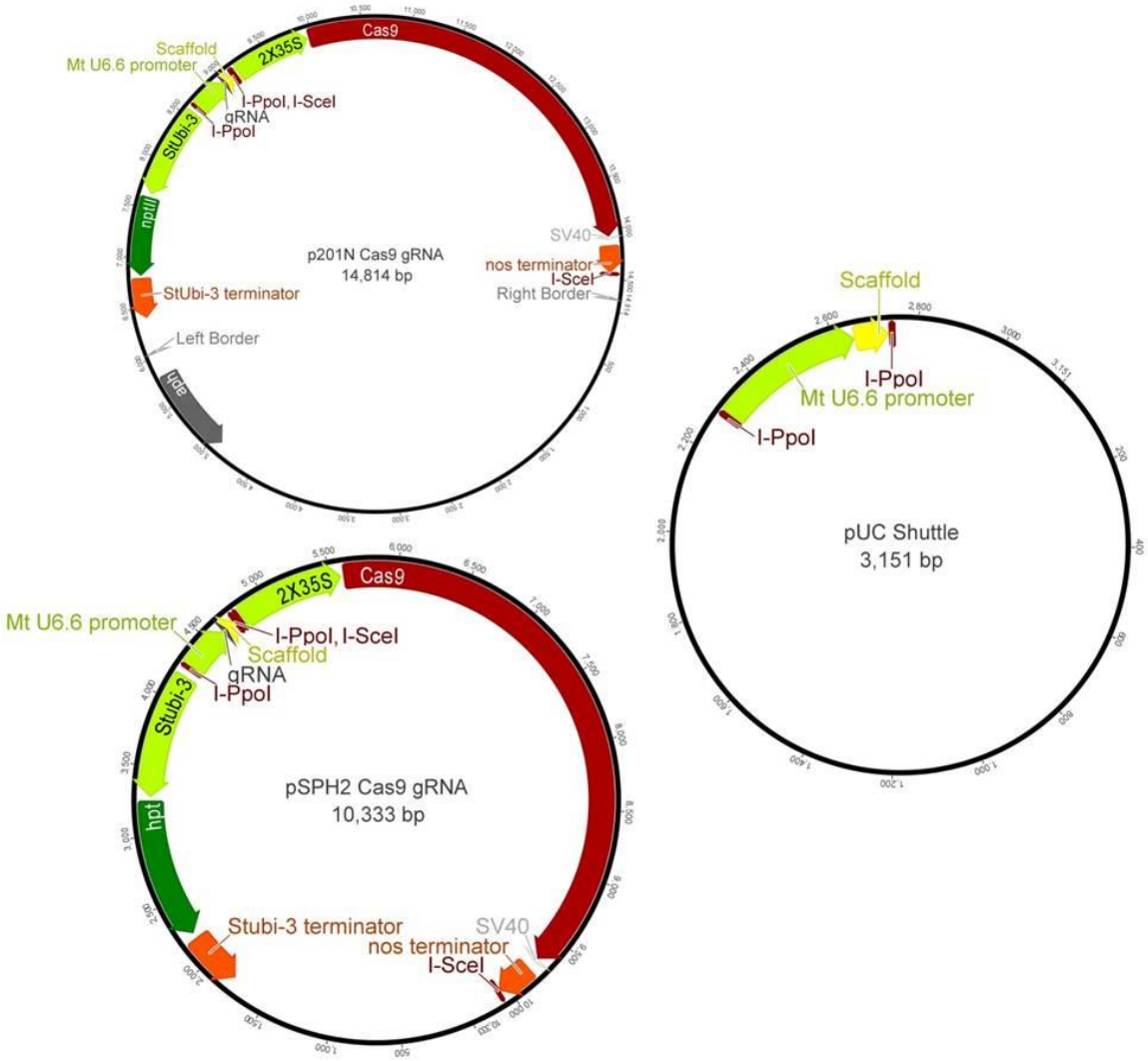


Figure 3.9. Vectors used in this study.

Cas9 Primers	5' -> 3'
Underlines denote restriction sites	
*Nhe*I_Cas9F	ATATAGCTAGCATGGACAAGAAGTACTCCATTGGGCT
*Sac*II_Cas9R	ATATACCGCGGTCACACCTTCCTCTTCTTCTTGGGG
35S2xF	TCATTTCATTTGGAGAGGACAC
Cas9R1	GCTGTGGCGATCGGTATTGC
Cas9F1	CATATGATCAAATTTCGGGGACACTTC
Cas9F2	CGATCAGTCTAAAAATGGCTACGCC
Cas9F3	CGTGGACCTCCTCTTCAAGACG
Cas9F4	CGTCAAAGTAATGGGAAGGCATAAGC
Cas9F5	GACAAACTGATTCGAGAGGTGAAAG
Cas9F6	GTACTGGTTGTGGCCAAAGTGG
*IPpo*I gRNA F	ATATACTCTCTTAAGGTAGCATGCCTATCTTATATGATCAATGAGG
*IPpo*I gRNA R	TGCTACTCTCTTAAGGTAGCAAAAAAAGCACCGACTCGGTG
MtU6 183 R	TTTGTAGCAGTGCACCTTGTCTTT
MtU6 63F	CAGCTCCTGGCTTGGGAATGAT
Ubi3p 218R	ACATGCACCTAATTTCACTAGATGT
nosT-Rev	TGATAATCATCGCAAGACC
*Pac*I nosR	ATATATTAATTAAATCAATTCCCGATCTAGTAACATAGATG
*Spe*I 35SF	ATATAACTAGTCCCAACATGGTGGAGCACGAC

puC gRNA Cloning primers	
gRNA Shuttle F	nxxxxxxxxxxxxxxxGTTTTAGAGCTAGAAATAGCAAGTT
gRNA Shuttle R	nnnnxxxxxxxxxxxxxxxxAAGCCTACTGGTTCGCTTGAAG

Blue nucleotides are part of the gRNA Target. X's denote 15bp of homology between primers required for In-Fusion® Cloning

gRNA Met1F	TGGTGCGGTCTATGAGGTTTTAGAGCTAGAAATAGCAAGTT
gRNA Met1R	TCATAGACCGCACCATGCCAAGCCTACTGGTTCGCTTGAAG
gRNA miR1509 F	TCACGGTTGAGTGTGAGTTTTAGAGCTAGAAATAGCAAGTT
gRNA miR1509 R	CACACTCAACCGTGATTTCAAGCCTACTGGTTCGCTTGAAG
gRNA miR1514 F	GGGATAGGAAAGGAAAGTTTTAGAGCTAGAAATAGCAAGTT
gRNA miR1514 R	TTCCTTTCCTATCCCAATCAAGCCTACTGGTTCGCTTGAAG

	Custom-amplicon sequencing
F Tail	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
R Tail	CTGGAGTTCAGACGTGTGCTCTTCCGATC
Final F Primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC
Final R with index	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGT
index in red	
Amplicon	
GFP5'	CGGCCACAAGTTCAGCGTG
	GCCGTTCTTCTGCTTGTC

GFP3'	GCCGACAAGCAGAAGAACGGCA
	TGATAATCATCGCAAGACC
07g14530	CGTTGCCAAAGATACCGTAGTGC
	CCAAACCCTGCCCTTTGGTTCTCT
01DDM	GGTTTGTTGTTTGCTTTTACGTTGCTAAC
	TGCAACAAGAACAAAAAATCAACACGCT
11DDM	GGAGTGAATTGGTTTTAATGTGTGTTG
	CTTTTTGATCCACAGCCCTTCTTCTTT
bothDDM1g	GAGGGTTGATAAAGGAGCAATTATCATGTAC
	CTGAGGAAGAAAGAGTTGAGAAAGAGCAG
bothDDM11g	GAGGGTTGATAGAGGAGCAATTATCATGTAT
	GAGCAAGTTGGTAAAATGTTATTTGACCAT
Met16G	GCCTTGGATGTCATGTCCAAACCAATT
	CCAAATAGGCGCATCACGGAGTTCATT
Met14G	GCCTTGGATGTCATATCCAAACCAACA
	ATGAGAAGTCCTGCCTCATTGAAACAG
miR1509	CTCCCAAGCGCAGGAGAGAGC
	GATCGGAATACAGAACAACGAAAAGGGAG
miR1514	GCCGTAATAGAGTCAGAGACTGTGG
	GAAGAAGATGAGGCGCCCTTGTCC
Off targets	
07g14530 1g	CGCACCTCTTCTTCTACGACGATG
	CACAACCAGTAAGGCTGTCCTTATAACCC
07g14530 4g	TTTGTTGCGGTTAACACGGG
	TGTGTTTATGCTCCCTCAGGT
07g14530 8g	GTATTCGTTCTTTCTTCAACGAAGC
	CCTTGGTCCTCTCAATGAACCG
07g14530 13g	CGGATAACTGTAACCTCGGTTTAAGCG
	GCTTTATCCTACCAACTTGCCCCGAAG
07g14530 13g2	GGGAACTACTTGGGCGTACTTGG
	CACACGGATTGGTGCCAATCTATATG
07g14530 13g3	TCTCACTTTCCAGGCAGACA
	AGCTCAACATGCCTGATTCA
07g14530 15g	CCTGGATACACGATGAAGATTGG
	CCACTAGTCAATACCCTGGAGG
07g14530 17g	ACACCTCTGTCAAAGAATCTCTCC
	GGTGACGAAAAATCACTAACTTCACCG
07g14530 18g	GGCAACGAAAAACCACCAAACTTCACC
	GGGTCAATGTATTTTTGCAGGTACACCC
07g14530 18g2	TCGCTCCTCTTGTCAACCTC
	ACTCAAAGCAAACGAGCACA
miR1514 5g	TCGAATATAGATTTCAGCTCTCACACT
	TCCAAGAAGCTACGAAGAGGC

miR1514 18g	AAATAATGTGAGAGAGGGAGGGA
	TTTGTTGTACCGTTTATTCAACCT
Met1 20g	GTCTAGCGCAAGTTGAAGCA
	TGAACTTGTTGTTGACTTGTGGG

Table 3.2 Primers used in this study. Underline denotes restriction site. X's are 15 bp of homology between primers required for In-Fusion® Cloning. Blue nucleotides are part of the gRNA target.

CHAPTER 4

HTSTUF: HIGH-THROUGHPUT SEQUENCING TO LOCATE UNMAPPED DNA FRAGMENTS[1]

*Jacobs, Thomas B.; *Kanizay, Lisa B.; Gillespie, Kevin; Newsome, Jade A.; Spaid, Brittany N.; Parrott, Wayne A. Submitted to *The Plant Genome*, 10/22/2014

*These authors contributed equally to this work.

Abstract

Advances in high-throughput sequencing have led to many new technologies for assessing genomes and population diversity. In spite of this, it remains difficult to pinpoint the location of transgenes and other specific sequences in large, complex genomes. Here we report the use of a modified T/A cloning and Illumina sequencing method called high-throughput sequencing to locate unmapped DNA fragments (Htstuf). Transgenic insertion sites were identified and confirmed in nine out of ten transgenic lines. Additionally this method was used to map *mPing* transposition events in four T6 lines derived from a single event. Fifteen of these insertion sites have been validated with PCR. Together, these data demonstrate the simplicity and effectiveness of this sequencing method.

Introduction

The reduced cost and increased power of high-throughput sequencing has allowed many genomes to be re-sequenced as a means of assessing genome diversity, identifying novel quantitative trait loci (QTL) [1, 2], and determining transgene integration sites [3]. While whole-genome re-sequencing (WGS) has proven beneficial for these purposes, it requires a reference genome and is cost-prohibitive for a large number of samples. Additionally, WGS analysis requires bioinformatic expertise. Some questions are better addressed with re-sequencing smaller, defined regions with techniques such as genotyping by sequencing (GBS) [4]. Furthermore, when a sequence of interest is inserted randomly into a genome, it can be helpful to know where that sequence is located, without wasting sequencing and bioinformatic resources by re-sequencing

unmodified regions. In the case of transgenic events, the junction sequence of the transgenic DNA can aid in the identification of promoter/ enhancer traps, mutagenized genes, T-DNA insertions, or transposon tags, and can improve the general understanding of the transformation process.

In addition to WGS, several PCR-based methods exist to capture junction fragments (genome walking); such as, TAIL-PCR [5], TOPO-vector ligation PCR [6], transposon display [7], and inverse PCR [8]. TAIL-PCR and similar methods require the use of random adapters, high Tm-primers, and can lead to false-positives caused by non-specific amplification, making it technically difficult to target a wide range of templates. Genome walking kits such as APAgene™ GOLD (BioS&T) and Universal GenomeWalker ™ 2.0 (Clontech) are fairly expensive, have stringent requirements and are not amenable to high-throughput procedures.

TOPO-vector ligation PCR is advantageous since the adapter is a cloning vector, and nested PCR primers can be designed to flank the insert, which should reduce non-specific amplification [6]. In addition, nested universal primer sites exist in the vector for sequencing purposes, and the required materials are present in most molecular biology labs. The previously published TOPO-vector ligation protocol used cloning and Sanger sequencing to identify junction fragments [6]. It should be possible to take advantage of the 'cleaner' amplification strategy and use the amplification products as templates for high-throughput Illumina sequencing.

PCR-based addition of Illumina adapters and barcodes has been used to generate transposon specific sequencing libraries in maize [9]. The maize UniformMu project uses adapter ligation followed by PCR addition of barcodes and multiplexing IDs for

Illumina sequencing. This method is effective at identifying the location of the endogenous *Mu* transposons in maize. The generation and analysis of the sequence data generated in the UniformMu project is inhibited by the abundance of *Mu* elements in maize but aided by the ability to use phenotypic markers for detecting transposon activity. Since not all species have the same capacity for genetics as maize or a plethora of well characterized endogenous transposons, we aimed to develop a simplified sequencing strategy for the identification and mapping of transposons and transgenes.

The goal of the method presented here was to specifically sequence portions of the soybean genome directly flanking transgene or transposon insertion sites. To increase throughput, the method utilizes Illumina sequencing and T/A cloning. The modified TOPO-vector ligation PCR yields site-specific amplicons that can be directly sequenced with Illumina technology, but with the correct primer sequences, any sequencing technology could be used.

Materials and Methods

Plant materials

Somatic embryos of soybean cultivar 'Williams82' were biolistically transformed as previously described [10]. The DNA used for bombardment was a gel-extracted 5081 bp *Pac*I (NEB) linear fragment, containing a *hygromycin phosphotransferase* gene under the control of a *Solanum tuberosum* promoter and terminator [11] and gene-of-interest cassette driven by the soybean GmUbi promoter [12] and *Pisum sativum* rbcS terminator [13] (Figure 4.3). DNA for library preparation was extracted from young leaves of T0 plants in the greenhouse.

To determine the zygosity of T1 plants from event 16, the Invader assay (Hologic Corp.) was run on a Synergy 2 plate reader (BioTek Instruments, Inc.) according to manufacturer instructions, except using a single reaction for each plant. The assay contains a probe that produces fluorescence when bound to the *hygromycin phosphotransferase* gene that is used to quantify the relative abundance of the target sequence in a genomic sample.

The transgenic soybean cultivar 'Jack' lines used for *mPing* sequence analysis are T6 lines derived from a previously reported event [10].

Library preparation and primer design

For an overview of the library preparation method see Figure 4.1. DNA was collected from young soybean leaves and extracted using a modified CTAB protocol [14]. Approximately 100 – 1000 ng of genomic DNA were fragmented to 1 – 5 kb with Fragmentase® (NEB) in 10-µl reactions, according to manufacturer instructions. Digestion times varied depending on the size of the initial DNA sample. Thirty minutes of digestion were used to obtain a 1 kb mean fragment size for most samples. After fragmentation, DNA samples were cleaned using Zymo Clean & Concentrator™ kit (Zymo Research), eluted with 11 µl of 10 mM Tris-HCl. Eluted samples were run on a bioanalyzer with a high-sensitivity DNA chip (Agilent Technologies) or on a Fragment Analyzer (Advanced Analytical) to confirm proper fragmentation.

Fragmented samples contain overhangs that need to be removed prior to A-tailing and ligation. The overhangs were removed using a T4 Polymerase reaction (per reaction: 2 µl 10X NEB Buffer2, 2 µl 10X BSA, 1 µl 2mM dNTPs, 0.2 µl T4 DNA Polymerase, 3

µl DNA, 11.8 µl water, and incubated at 12°C for 15 minutes), which was stopped by immediately cleaning with Zymo columns and eluted with 8 µl 10 mM Tris-HCl. Samples were A-tailed and ligated to a pGEM-T vector (Promega) according to manufacturer instructions. A 4°C overnight ligation was used. The ligation was diluted 1:10, and 1 µl was used as a template for primary PCR. A gene specific primer (GSP) 1 was used with the universal M13 reverse primer for primary PCR with KAPA 2X HiFi HotStart ReadyMix (KAPA Biosystems) (per reaction: 5 µl 2X ReadyMix, 0.3 µl of each 10 µM primer, 1 µl diluted ligation, water to 10 µl) with the following conditions: 95°C 3 minutes; 30 cycles (98°C 20 seconds, 60°C 15 seconds, 72°C 1 minute); 72°C 5 minutes; hold at 12°C. PCR products were visualized on a 1% agarose, TBE gel. A smear with a large product at approximately 3 kb was often observed (Figure 4.2). The primer sequences used in library preparation are in supplemental Table 4.1.

The primary PCR products were then diluted 1:100, and 1 µl was used in a secondary touchdown-PCR reaction with the conditions 95°C 3 minutes; 10 cycles (98°C 20 seconds, 70°C to 60°C 15 seconds (-1C per cycle), 72°C 1 minute); 20 cycles (98°C 20 seconds, 60°C 15 seconds, 72°C 1 minute); 72°C 5 minutes; hold at 12°C) with nested primers GSP2 and pGEM reverse. The secondary primers contain 5' tails that are used to start adding on the Illumina adapter sequences. A touchdown PCR method was used to ensure specific amplification. PCR products were visualized on a 1% agarose, TBE gel. Multiple banding patterns and smears were typically observed. The PCR product should be obvious at this point and greater than 250 bp in length; shorter fragments are likely from primer-dimers.

Secondary PCR products were diluted 1:100. At this point, if there were multiple amplicons per sample, all secondary reactions were pooled within a sample. The tertiary PCR primers bind the tails that were added during the secondary PCR and amplify any secondary product (including primer-dimers). The tertiary adds 6-nt indexing barcodes [15] to the samples and produces a final PCR product with complete Illumina TruSeq-style adapters. The tertiary cycle conditions are as follows: 95°C 3 minutes; 10-13 cycles (98°C 20 seconds, 60°C 15 seconds, 72°C 1 minute); 72°C 5 minutes; hold at 12°C. Tertiary PCR products were visualized on 1% TBE agarose gels to ensure amplification.

Tertiary PCR products were then pooled, run on a 1% agarose, TAE + cytidine gel, and 500-1000 bp molecules were gel-extracted. The gel extraction was performed with a Zymo gel extraction kit, and the final libraries were eluted in 10 µl of 10 mM Tris-HCl. Libraries were quantified with the qPCR KAPA library quantification kit (KAPA Biosystems) according to manufacturer instructions, and reactions were run on a LightCycler480II (Roche). Libraries were also run on a Bioanalyzer with a high-sensitivity chip to ensure the correct size of library fragments. Libraries were then prepared and run on a MiSeq (Illumina Inc.) according to manufacturer instructions. Raw reads were de-multiplexed with the 6-nt indexes using the MiSeq Reporter software, version 2.3.32 (Illumina Inc.).

Sequence Analysis and Flanking Sequence Confirmation

Transgene insertion mapping

Fastq files were imported into the commercial software Geneious (Biomatters Ltd.) versions 6 or 7. Low-quality reads were trimmed from read1 and read2 using the

trim-ends function with an error-probability limit of 1%. The pGEM_R primer was removed from the 3' end of read1 and the Illumina adapter primer was removed from the 3' end of read2. Read1 reads were then sorted by amplicons, using the GSP2 primers as barcodes in the separate-reads-by-barcode function, with one mismatch allowed. Individual amplicon reads from read1 were paired with reads from read2, resulting in a set of paired reads for each amplicon. The paired reads were *de novo* assembled into contigs with the following settings: Don't merge contigs when there is a variant with coverage over approximately 6, merge homopolymer variants, do not allow gaps, minimum overlap 25, no minimum overlap identity, word length 24, index word length 14, ignore words repeated more than 100 times, reanalyze threshold 16, maximum mismatches per read 5%, and maximum ambiguity 4.

For each amplicon, the *de novo* contigs were mapped to the linear 5,081 bp transgene as the reference (Figure 4.3), with the default setting of medium sensitivity. Contigs that did not map to the reference were also saved. Of the contigs that mapped to the fragment, only those with at least 100 raw reads were considered for additional analysis. In the assembly viewer, contigs with large stretches of mismatches to the reference ($\geq$ 20 bp) were run through BLAST on the NCBI database and the soybean genome in Phytozome, along with unused reads. BLAST hits are summarized in Table 4.2. Soybean genomic sequences within the contigs with were considered putative flanking sequences.

Reverse primers were designed to the putative flanking sequences and were used in PCR with the respective GSP1 primers. Four templates were used for each primer, T0 DNA, non-transformed DNA, sequencing library, and no-template control. The

sequencing library was used as a positive control. An example can be seen in Figure 4.4. The PCR reaction was as follows: 5 µl 2xApex Master Mix (Genesee), 0.3 µl 10 µM primers, 1 µl template, 3.4 µl water, using the following conditions: 95°C 3 minutes; 32 cycles (95°C 15 seconds, 60°C 15 seconds, 72°C 30 seconds); 72°C 5 minutes; hold at 12°C. PCR products were visualized on a 1% agarose, TBE gel. Only primers sets with amplification in the T0 and sequencing library were considered positive (Figure 4.4). PCR products were purified and sequenced with Sanger sequencing to ensure PCR products were from the expected DNA sequence.

Transposon insertion mapping

Individuals from four lines were sequenced twice, once as individuals and once in pools of 4-6 siblings so that more samples could be processed at once. These sequences were analyzed in two ways. First, to get a visual assessment of the method's efficacy, sequences were processed using Geneious 7. The 3' Illumina adapter was removed from all sequences using the trim primer function. Next, to separate *mPing*-containing sequences from background, reads containing the last 55 bp from the 3' and/ or 5' end of *mPing* were filtered using the separate reads by barcode function, with one mismatch allowed. These *mPing*-only reads were mapped back to the reference soybean genome with the following custom sensitivity settings; no fine tuning, maximum gap size of 50 with no more than 15% of the read having gaps, a word length of 20 nt and index word length of 12 nt, a maximum of 30% mismatches per read, maximum ambiguity of 4, and allowing read mapping to repeat regions. The sequences were mapped to a concatenation of the 20 soybean chromosomes. The reads were also mapped to each chromosome

individually so that the specific location of each *mPing* insertion could be determined (Table 4.3). General stats for the number of reads generated and mapped can be found in Table 4.4.

To rapidly process the large number of samples, fastq files from the individuals analyzed in Geneious, as well as all other sequenced *mPing* individuals, were uploaded to the University of Georgia's computing cluster (Zcluster). Reads were filtered for quality using FastQC, and set to keep only reads for which 75% of the read length has a quality score of 37 or higher. Illumina adapters and *mPing* were trimmed from the reads using Cutadapt [16]. Reads containing *mPing* were placed into a new file while the remaining sequences were discarded. To reduce read redundancy, reads greater than 50 bp after *mPing* trimming were assembled with CAP3 [17], then contigs were mapped to the soybean reference genome using Bowtie2 [18] with the 'very sensitive' option. These programs were joined together into a shell script for efficient processing.

Primers pairs were designed between 100-300 bp upstream and downstream of each putative *mPing* insertion site. Primers were tested first with the samples in which the insertion was sequenced and with a no-template control. If these results were positive, the primers were used in a broader range of DNA samples, from siblings of the sequenced lines. PCR reaction was as follows: 5 µl 2xApex Master Mix (Genesee), 0.2 µl 10 µM primers, 1 µl template, 3.6 µl water, using the following conditions: 95°C 3 minutes; 34 cycles (95°C 30 seconds, 55°C-60°C 20 seconds, 72°C 1 minute); 72°C 5 minutes; hold at 12°C. PCR products were visualized on a 1% agarose, TBE gel. Primer sequences for validated insertion sites are publicly available from SoyBase.org.

Results and discussion

Genome Fragmenting

To generate sequence-specific PCR libraries compatible with Illumina sequencing, genomic DNA was first fragmented to a usable size. A Covaris sonicator (Covaris Inc.) was used initially for mechanical sheering, but this method proved to be too costly and required too much input DNA (results not shown). Therefore, DNA Fragmentase®, an enzyme mix that randomly cuts DNA was evaluated. Several digestion times were tried using different concentrations of DNA. A digestion time of 30 minutes proved to be effective for a wide range of DNA concentrations.

Library Construction

T/A ligation was used to generate templates for sequencing libraries, since the use of a vector with T overhangs, as opposed to linear adapters or blunt ligation, should limit the formation of DNA concatemers [6]. The conditions used for primary and secondary amplification were a compromise between specificity and amplification. When a touchdown protocol was used for primary amplification, few samples produced secondary products. When a non-touchdown protocol was used for secondary amplification, amplification could be observed in the genomic control sample. Both of these are undesirable for the production of amplicons. For this reason, a low-stringency primary amplification followed by a high-stringency, touchdown, nested secondary PCR strategy was adopted. After primary amplification, a large product at approximately 3 kb (presumably the pGEM vector), and a smear of smaller products, were observed in all DNA samples (Figure 4.2). After the secondary amplification, samples with the

amplicons-of-interest had the most PCR products, typically as a smear with some banding (Figure 4.2).

Secondary amplification products within an event can be pooled into a single tertiary reaction, since the individual amplicons can be de-multiplexed after sequencing. The tertiary reaction is a low-cycle PCR to add on the necessary Illumina adapters and to index the samples. This amplification results in a smear for all samples (Figure 4.2). The samples can then be pooled and loaded into a single well for gel extraction. When many indexes are used (>20), it is useful to pool samples, column purify, and load a concentrated sample in a single well. Gel-purified samples are then ready for sequencing. No attempt was made to normalize the amount of DNA used per sample or per amplicon.

Transgene sequencing and mapping

The initial attempt to identify genomic flanking sequences used the walking protocol to sequence from the rbcS and StUbi3 terminators, which are on the end of the linear vector (Figure 4.5). However, only vector rearrangements were detected (data not shown), indicating that there was a complex integration of the linear fragments. Therefore, it was reasoned that the linear vectors were rearranged during the transformation process and a flanking sequence could be found anywhere along the length of the vector. Primers were designed to capture amplicons approximately every 500-800 bp along the vector (Figure 4.5). With 14 amplicons per event, it was necessary to use a high-throughput sequencing method.

Sequencing ten transgenic events resulted in 2,364,016 paired reads (read1 + read2). Although the input DNA was not normalized, there was an even distribution of reads (8.4-10.6%) across the ten events (Table 4.5). Within each event, the reads were also evenly distributed across the 14 amplicons. Twelve amplicons had, on average, 6-9% of the total reads (Table 4.6) which is expected if the reads were evenly distributed. The two amplicons Stubi262F and Stubi389R were underrepresented, with only 2.4 and 5% of the sequenced reads, respectively. These amplicons likely did not amplify as efficiently as the others; however, this did not prevent the detection of flanking sequences (Table 4.6). Fifty-six percent of the paired-reads were assembled into contigs and were considered usable (Table 4.5). These data suggest that the efficiency of the library preparation method could be improved, but the quantity and quality of the data were still sufficient for identifying flanking sequences.

Ninety putative flanking sequences were identified, with four to 16 flanking sequences per event. Twenty flanking sequences were confirmed in T0 DNA in nine out of ten events (Table 4.5). The confirmed flanking sequences were evenly distributed across the 14 amplicons, with only four amplicons not having a genomic flanking sequence (Table 4.6). This last point underscores the fact that the vector DNA randomly rearranged during integration. These results, and the complex arrangements of the transgenes, demonstrate that a flanking sequence can be found at any position along the vector DNA. Similar vector rearrangements have been observed in transgenic oat [19]. Such integration patterns would be difficult if not impossible to dissect with the more traditional Southern blot technology.

Three events have pairs of confirmed flanking sequences that mapped to the same chromosome, several Kb to one Mb apart. For example, event 8 has four flanking sequences; two are mapped to chromosome 2, eight Kb apart, and two to chromosome 14, 406 kb apart (Table 4.2). Similar results can be seen for events 9, and 13. Interestingly, event 31 has five confirmed flanking sequences that map to the chloroplast or chromosome 9, as well as chromosome 15. Three flanking sequences are located in repetitive regions of the genome, so it is difficult to precisely map their locations. These data are reminiscent of the complex arrangements of transgenes, interspersed with chloroplast [20] or genomic [19, 21, 22] sequences in other transgene mapping reports.

Segregation analysis of flanking sequences in transgenic events

Segregation analysis was performed on five lines to determine the linkage of the transgenes and the confirmed flanking sequences. There are three segregating units for event 8 and, surprisingly, the two flanking sequences that mapped to chromosome 14, only 406 kb apart (<1cM), segregate independently (Table 4.7). This may be due to chromosome 14 sequences being incorporated during the transgene integration process at other unidentified loci. The five confirmed flanking sequences in event 31 all form one segregating unit, and there is also a second, un-identified insertion, as four of the T1 plants are positive for the gene-of-interest (GOI) and negative for the five flanking sequences (Table 4.8). A single segregating unit was observed in the three events 13, 16, and 36.

Determining the zygosity of segregating progeny, and identifying homozygous lines, is an important step in transgene analysis. To this end, the Invader Assay is

routinely used to identify homozygous plants. Initially, it was reasoned that if the location of the transgene insertion was known, it would be possible to design primers in the flanking genomic regions that could test for the presence or absence of the insertion. However, despite repeated attempts, only one zygosity primer set for event 16 worked. Two primers were designed to flank the chromosome 20 insertion (Figure 4.6B). In this case, only wild-type sequences are able to amplify. The transgenic insertion is too large or complex for the PCR conditions, thus a negative result would indicate a homozygous line. Upon PCR amplification, three T1 plants (6, 21, 27) tested negative (Figure 4.6A). These three plants were positive for the flanking sequence, and the Invader assay also identified these three individuals as homozygous (Figure 4.6C). The two methods had perfect correlation, calling heterozygotes and null segregants. These data illustrate that even though it was difficult to generate zygosity primer sets for all events, knowledge of the genomic insertion site can be used to determine the zygosity of transgenic plants.

While the development of a zygosity-PCR test proved challenging, the mapping approach still provided unique markers for tracking insertions across generations. With the flanking markers, it is possible to identify plants with multiple segregating loci with a simple PCR test, as was the case for events 8 and 31. Such information could be obtained with a Southern blot, but these are generally more technically challenging and have a lower throughput. With these markers, it is possible to segregate the independent insertions away from each other, or to keep them together for a stacked event.

mPing sequencing and mapping

The DNA transposon *mPing* is a small, 430-bp element originally identified in rice [23]. It was previously transformed into soybean with the goal of generating mutations in soybean genes [10]. This sequencing method was tested on lines derived from a single event as a way to rapidly map *mPing* insertions in a large population and distinguish germinal insertions from somatic insertions. Sequencing of 15 individuals produced over 600,000 reads. These reads were processed using Geneious 7, and approximately 220,000 *mPing*-containing reads were identified (~37% of total reads). Of these, 170,030 were mapped to the soybean reference genome. Based on shared insertions and coverage of mapped loci, seven loci stood out above background levels (Figure 4.7). To expand the analysis, 24 pools containing four to six DNA samples were sequenced. These sequences contained the same 15 initially sequenced individuals, as well as 84 additional ones. More than 620,000 *mPing*-containing reads were mapped to 22 unique locations in the genome, and 15 were PCR-validated (Table 4.3).

The PCR-validated insertions were shared between different individuals, indicating that they are germinal insertions, i.e., occurring in the previous generation's germ line (Table 4.3), rather than in somatic tissues that does not contribute to gamete formation. In fact, many of these insertions were shared between individuals from different lines, indicating they occurred at least two generations prior. Five insertions that were previously validated by cloning of transposon-display (TD) products ([10]and personal communication) were identified here as well, providing further support of the validity and robustness of this method.

Utility of HtStuf method for mapping transposon and transgenes

Transposon display is a modified AFLP method [7] that requires running a polyacrylamide gel, extracting bands and/or cloning, and sequencing the bands or clones (Figure 4.8) [10]. To identify the five insertions validated here, a TD reaction was cloned and 96 colonies were sequenced (personal communication). The use of TD to identify and validate all insertions in a large population of *mPing* lines is costly and time prohibitive (for a cost comparison, see Figure 4.8). Moreover, the restriction enzyme used for the initial production of genomic DNA fragments can result in sampling biases.

The bioinformatics analysis is also very straightforward and can be done in user-friendly software such as Geneious. Unlike the work with mapping Mu transposon insertions in maize, the samples here were not tagged with multiplexing IDs. This largely simplifies the separation of individual samples in our analysis, which can be done in MiSeq Reporter. Constructing libraries with known priming sites allows the trimming and isolation of *mPing*-containing sequences with pre-existing informatics programs.

This sequencing method is robust and accurate; flanking sequences were confirmed in nine out of ten transgenic events, and 15 individuals from four *mPing* lines. The power of this approach comes from the combination of random fragmentation, T/A cloning, and the use of next-generation sequencing technology, which allows pooling many amplicons and samples for simultaneous sequencing. Typical gene-walking experiments clone and then Sanger-sequence PCR products to identify flanking sequences [24]. With the complex integration pattern observed with the transgenic events, or with the large numbers of individuals in a transposon mutagenesis screen, such an approach would be tedious, time-consuming, and expensive. Instead, this method can

go from DNA isolation to sequence analysis and primer design within one week. Furthermore, the molecular techniques used are straight-forward enough that student workers have been trained to independently generate sequencing libraries.

One limitation with this method is the amount of background PCR and ligation artifacts that are sequenced. Only 22% (20/90) of the putative transgene flanking sequences were shown to be real insertions. While this may seem like a low frequency, the PCR screening process to verify putative flanking sequences can be accomplished in relatively little time. Additionally, because such high numbers of reads are generated with Illumina sequencing, more than enough useful sequence data are produced. One possible way to improve the overall specificity is to sequence multiple individuals from the same line, as was shown with the transposon mapping. All PCR-confirmed *mPing* insertions were mapped in multiple individuals. Insertions in single individuals are either artifacts or, more likely, new, somatic insertions. Since this method yields a large number of sequences, and the sequencing libraries are relatively simple to prepare, making two libraries per transgenic event to improve specificity would not be a burden.

For transgenic mapping, one advantage of this method over WGS is the enrichment of the transgenic sequences. Here, 56% of the sequenced reads could be assembled into contigs and used for mapping to the transgene. If WGS were used, assuming that most transgenes are 10 kb in size, most crop genomes are at least 1 Gb, and the sequencing reads are paired-end and 100 to 300 bp in length, then only 0.2 to 0.6% of the sequencing reads will contain transgenic bases. This is an incredible waste of sequencing resources. Using WGS for the identification of transgenic insertions may be

better suited to evaluating high-value transgenic lines to ensure other DNA modifications were not made during the transformation process.

While we used this method for the identification of transgene and transposon flanking sequences, there are additional applications for this type of technology. In non-sequenced or poorly sequenced genomes, this walking method can be used to close gaps in contigs for genome assembly. Degenerate primers could also be designed to amplify sequences in related organisms to capture sequences from large gene families. In cases of gene-amplification such as EPSPS [25], this technique could be used to identify unique insertion sites. With the increased sequencing lengths offered by new technologies such as PacBio, it should be possible to modify the primer sequences to work with any system.

Conclusions

The ability to sequence specific unknown DNA loci with a high-throughput technology is useful for the characterization of transgenic plants. The data presented demonstrates that HtStuf is a quick and reliable method for determining the flanking sequences of transgenes and transposons in the soybean genome. The effectiveness of the technology is based on the combined use of a modified TOPO vector-ligation PCR method with the power of Illumina sequencing. Sequence data can be generated and analyzed quickly, and flanking sequences are identified in nearly all sequenced individuals. This sequencing technology is not limited to soybean and should be applicable in any other species where flanking sequences need to be known.

References

1. Hufford MB, Xu X, van Heerwaarden J, Pyhajarvi T, Chia JM, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM *et al*: **Comparative population genomics of maize domestication and improvement**. *Nature Genetics* 2012, **44**(7):808-U118.

2. Huang XH, Feng Q, Qian Q, Zhao Q, Wang L, Wang AH, Guan JP, Fan DL, Weng QJ, Huang T *et al*: **High-throughput genotyping by whole-genome resequencing**. *Genome Research* 2009, **19**(6):1068-1076.

3. Kovalic D, Garnaat C, Guo L, Yan YP, Groat J, Silvanovich A, Ralston L, Huang MY, Tian Q, Christian A *et al*: **The use of next generation sequencing and junction sequence analysis bioinformatics to achieve molecular characterization of crops improved through modern biotechnology**. *Plant Genome* 2012, **5**(3):149-163.

4. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: **A Robust, simple genotyping-by-sequencing (GBS) approach for high diversity species**. *Plos One* 2011, **6**(5):10.

5. Liu YG, Whittier RF: **Thermal asymmetric interlaced PCR - automatable amplification and sequencing of insert end fragments from p1 and YAC clones for chromosome walking**. *Genomics* 1995, **25**(3):674-681.

6. Orcheski BB, Davis TM: **An enhanced method for sequence walking and paralog mining: TOPO(R) Vector-Ligation PCR**. *BMC research notes* 2010, **3**:61.

7. Van den Broeck D, Maes T, Sauer M, Zethof J, De Keukeleire P, D'Hauw M, Van Montagu M, Gerats T: **Transposon display identifies individual transposable elements in high copy number lines**. *Plant Journal* 1998, **13**(1):121-129.

8. Ochman H, Gerber AS, Hartl DL: **Genetic applications of an inverse polymerase chain-reaction**. *Genetics* 1988, **120**(3):621-623.

9. McCarty DR, Settles AM, Suzuki M, Tan BC, Latshaw S, Porch T, Robin K, Baier J, Avigne W, Lai JS *et al*: **Steady-state transposon mutagenesis in inbred maize**. *Plant Journal* 2005, **44**(1):52-61.

10. Hancock CN, Zhang F, Floyd K, Richardson AO, LaFayette P, Tucker D, Wessler SR, Parrott WA: **The rice miniature inverted repeat transposable element *mPing* is an effective insertional mutagen in soybean**. *Plant Physiology* 2011, **157**(2):552-562.

11. Garbarino JE, Belknap WR: **Isolation of a ubiquitin-ribosomal protein gene (*UBI3*) from potato and expression of its promoter in transgenic plants**. *Plant Molecular Biology* 1994, **24**(1):119-127.

12. Hernandez-Garcia CM, Martinelli AP, Bouchard RA, Finer JJ: **A soybean (*Glycine max*) polyubiquitin promoter gives strong constitutive expression in transgenic soybean**. *Plant Cell Reports* 2009, **28**(5):837-849.

13. An G, Watson BD, Stachel S, Gordon MP, Nester EW: **New cloning vehicles for transformation of higher-plants**. *Embo Journal* 1985, **4**(2):277-284.

14. Murray MG, Thompson WF: **Rapid isolation of high molecular-weight plant DNA**. *Nucleic Acids Research* 1980, **8**(19):4321-4325.

15. Faircloth BC, Glenn TC: **Not all sequence tags are created equal: Designing and validating sequence identification tags robust to indels**. *Plos One* 2012, **7**(8):11.

16. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads**. *2011* 2011, **17**(1).

17. Huang XQ, Madan A: **CAP3: A DNA sequence assembly program**. *Genome Research* 1999, **9**(9):868-877.

18. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature Methods* 2012, **9**(4):357-U354.

19. Svitashev SK, Pawlowski WP, Makarevitch I, Plank DW, Somers DA: **Complex transgene locus structures implicate multiple mechanisms for plant transgene rearrangement**. *Plant Journal* 2002, **32**(4):433-445.

20. Aragao FJL, Nogueira E, Tinoco MLP, Faria JC: **Molecular characterization of the first commercial transgenic common bean immune to the Bean golden mosaic virus**. *Journal of Biotechnology* 2013, **166**(1-2):42-50.

21. Svitashev SK, Somers DA: **Genomic interspersions determine the size and complexity of transgene loci in transgenic plants produced by microprojectile bombardment**. *Genome* 2001, **44**(4):691-697.

22. Jackson SA, Zhang P, Chen WP, Phillips RL, Friebe B, Muthukrishnan S, Gill BS: **High-resolution structural analysis of biolistic transgene integration into the genome of wheat**. *Theoretical and Applied Genetics* 2001, **103**(1):56-62.

23. Naito K, Cho E, Yang GJ, Campbell MA, Yano K, Okumoto Y, Tanisaka T, Wessler SR: **Dramatic amplification of a rice transposable element during recent domestication**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(47):17620-17625.

24. Leoni C, Volpicella M, De Leo F, Gallerani R, Ceci LR: **Genome walking in eukaryotes**. *Febs Journal* 2011, **278**(21):3953-3977.

25. Gaines TA, Zhang WL, Wang DF, Bukun B, Chisholm ST, Shaner DL, Nissen SJ, Patzoldt WL, Tranel PJ, Culpepper AS *et al*: **Gene amplification confers glyphosate resistance in *Amaranthus palmeri***. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(3):1029-1034.

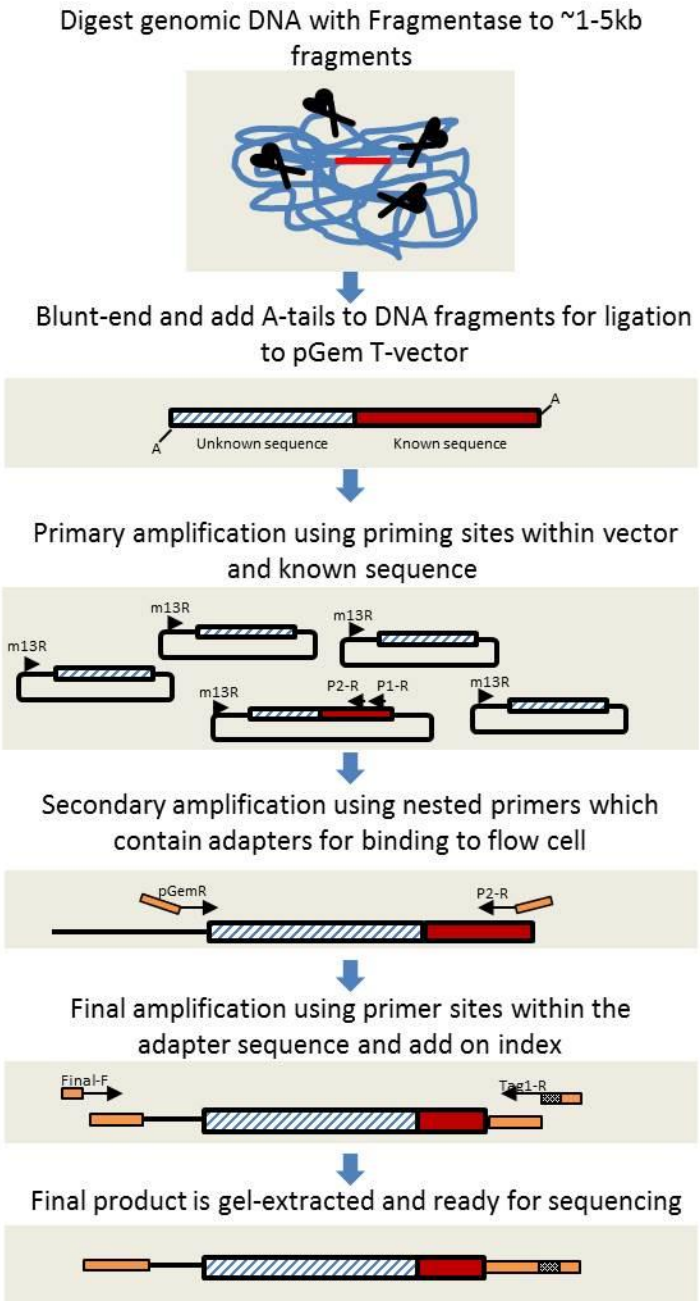


Figure 4.1. Library generation overview.

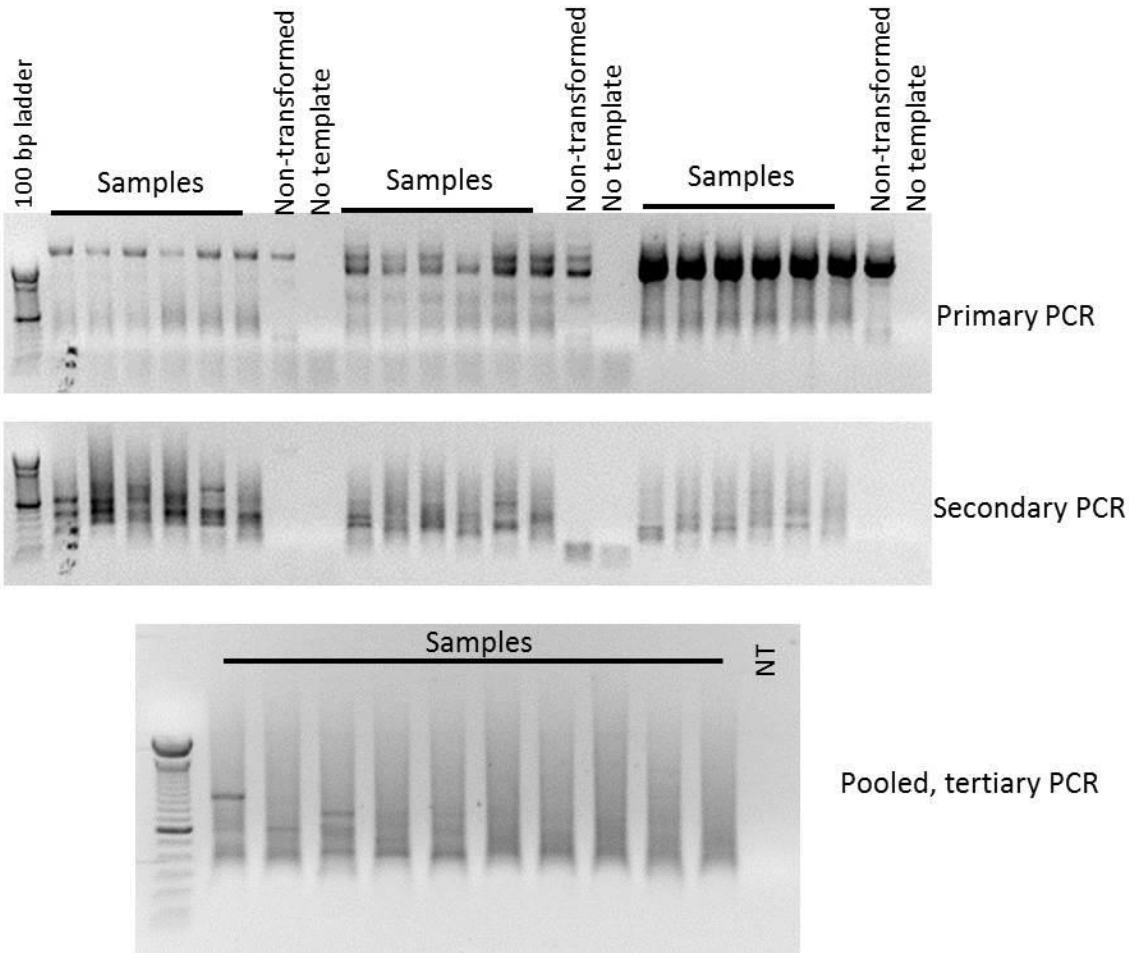


Figure 4.2. An example of the amplification process. The low-stringency primary PCR typically results in a smear for all samples, including non-transformed controls. The secondary PCR is a touch-down to provide increased stringency. Only samples containing the amplicons-of-interest produce a visible product. Amplicons within an event can then be pooled and run on a tertiary PCR that adds Illumina barcodes and adapters. Barcoded tertiary products are pooled together and ready for sequencing.

Figure 4.3. An example of mapping of contigs (grey and black bars) to the transgenic insert. The contigs were derived from a single amplicon from a single vector. Grey portions match the transgene sequence and black portions do not match the reference. The amplicon is derived from the StUbi3 terminator. Note the reads mapping to the rbcS Terminator, indicating rearrangement of the transgenic vector. The gene-of-interest (GOI) and a portion of chloroplast sequence is also fused in one contig. Scale bar in base pairs is across top of figure.

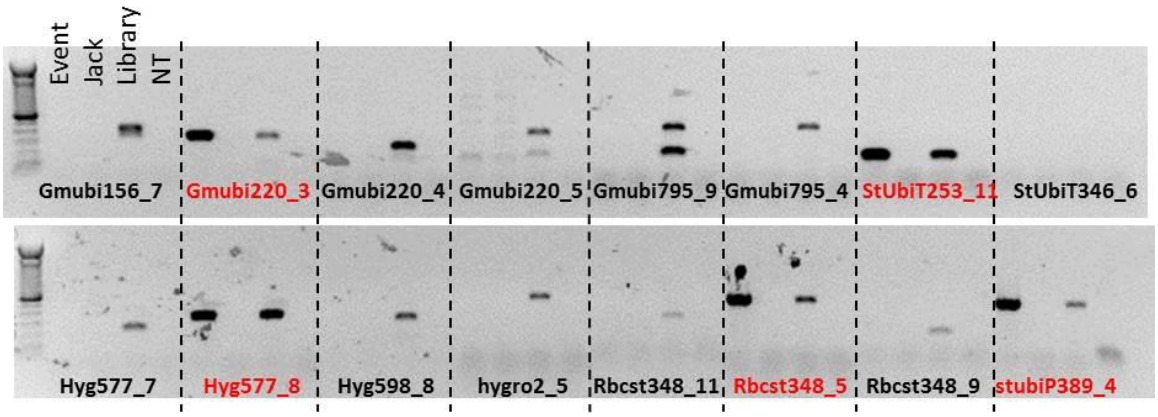


Figure 4.4. An example of PCR validation of the putative flanking sequences. Sequences that only amplify with the Event and Library DNA are considered confirmed insertions (red text). Most flanking sequences are just positive for the Library indicating these are ligation/PCR artifacts. Jack is non-transformed genomic DNA control. NT is no-template PCR control.

Primer Name	Transgene ID	Sequence
GSP1_1	gmubi100F	GCAATTCTCAATTTCCTAGAAGGACTCTC
GSP2_1	GmUbi766F	CGAGAGATATTGCTCAGATCTGTTAGC
GSP3_1	hyg 620R	CCAACCACGGCCTCCAGAAG
GSP4_1	hyg167R	CCATCGGCGCAGCTATTTACCC
GSP5_1	rbcst295F	CTTGTAGTTGTACCATTATGCTTATTCACTAGGC
GSP6_1	StUbiP343R	GAGATTCTTTTCTCAATTTTGTATGTCTAGAG
GSP7_1	StUbiT402 R	GTCTCGACAGACACATAGCACCTAAC
GSP8_1	GmUbi270R	GAAGGAAGGAGGAGGGGTG
GSP9_1	gmubi923R	CCGCTGTCGAGTCAACAATCACAG
GSP10_1	GSPa	CAACACACTAGCAATTTGGCTGCAGCGTATGG
GSP11_1	hyg 540 F	TGATGCTTTGGGCCGAGGAC
GSP12_1	rbcs451R	GAATCTGACAAGGATTCTGGAAAATTACATA
GSP13_1	StUbiP218F	ACATCTAGTGAAATTAGGTGCATGT
GSP14_1	StUbiP840F	GGCAGAGCTTACACTCTCATTCC
GSP1_2	GmUbi156F	CGTGTCATAGGCACCAAGTGA
GSP2_2	GmUbi795F	CTGCCTTGTTTGTTGATTCTATTGCCG
GSP3_2	GSPcR	TACTCAACCCAATGAGCATAAAGACTGTA
GSP4_2	hyg 577R	GAATCCCCGAACATCGCCTC
GSP5_2	hyg114R	GCTGAAAGCACGAGATTCTTCGC
GSP6_2	rbcst348F	CCTAGAAAAGCTGCAAATGTTACTGAATAC
GSP7_2	StUbiP389R	TTCACTCTAGTTGGTTGTTGCTTTG
GSP8_2	GmUbi220R	CCCTTCTCCAGTCATATTGTGACGC
GSP9_2	GmUbi842R	GTACGGATCTGAAGCAATCTCGTG
GSP10_2	GSPb	GGATTATGGAACTATCAAGTCTGTGGGATCG
GSP11_2	hyg 598F	CTCCAACAATGTCCTGACGGAC
GSP12_2	hygro2F	TGAAAAAGCCTGAACTCACCGCG
GSP13_2	rbcs418R	GGAAAGTTCATAAATGTCTAAAACACAAGAGG
GSP14_2	StUbiP262F	GGGCATATATCTTGATCTAGATAATTAACG

GSP5'_1	mping	CCTTCTCTCTCATCCCCATTTCATGCAA
GSP5'_2	mping	CTTACCCCTATTAAATGTGCATGACACACC
GSP3'_1	mping	GGGGATTGTTTCATAAAAGATTTCATTTGAGAGAAG
GSP3'_2	mping	GGTATAATATTTTGGGTAGCCGTGCAATGAC
		CTGGAGTTCAGACGTGTGCTCTTCCGATCCGGCCGCGAATT
IllR_pGEM	Illumina Primer	CACTAGTG
	Illumina Primer	
IllF_GSP_2	tail	ACACTCTTTCCCTACACGACGCTCTTCCGATCT…GSP_2
Final R with	Illumina Primer,	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGA
index	red is barcode	GTTCAGACGT
		AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA
Final F	Illumina Primer	CGAC

Table 4.1. Primers used in library construction. Underline is pGEMR.

Event	Primer	Contig ID	#reads	Glyma1.1 position
8	GmUbi156F	2	1266	Gm02:26,768,025..26,768,100
	GmUbi842R	4	901	Gm02:26,776,749..26,776,849; repeat
	rbcst348F	1	3329	Gm14:5,793,219..5,793,372
	StUbiP262F	1	643	Gm14:5,387,136..5,387,383
9	GmUbi842R	4	275	Gm03:41,393,632..41,393,874
	StUbiT253F	2	543	Gm03:36,156,081..36,156,374
13	rbcst348F	6	162	Gm01:10,803,728..10,803,847; repeat
	StUbiP389R	3	611	Gm01:10,877,434..10,877,780
16	StUbiP262F	3	244	Gm20:45,226,343..45,226,503
20	GmUbi156F	3	814	Gm13:5,725,568..5,725,733
	StUbiT253F	several		Gm01:50,969,475..50,969,635
	hyg114R	3	427	Gm16:1,355,529..1,355,745
22	GmUbi795F	9	310	Gm08:9,325,670..9,325,750
31	GmUbi220R	3	476	Gm15:39,172,663..39,172,790
	StUbiT253F	11	135	Gm09:22,925,954..22,926,254 or chloropast
	hyg 577R	8	143	Gm09:12,126,093..12,126,332 or chloroplast
	rbcst348F	5	317	Chloroplast
	StUbiP389R	4	757	Chloroplast
36	GmUbi220R	3	1377	Gm12:16,905,590..16,905,722; repeat
38	GmUbi156F	2	1222	Gm06:45,240,174..45,240,532

Table4.2. Confirmed flanking loci for transgenic events.

			Event 2-9 B2													
			3-47-2-3				19-6-16-2				16-9-9-5			32-13-A-11		
PCR validated	Gmax v1.1 location	Gmax v2.0 location	4	5	6	7	13	14	15	18	4	5	6	8	10	11
not tested	01:1429049	01:1424207	X	X	X	X	X	X	X	X				X	X	X
yes *	01:2606076	01:2622091	X		X	X	X	X			X		X		X	X
not tested	01:4222382	01:4244346								X		X				
no	02:10936581	02:11029144													X	
no	03:39420017	03:37402214	X	X	X	X										
yes	4:5309363	04:5375178		X			X	X	X		X	X	X		X	X
yes	4:49160246	04:52305882		X	X	X	X				X	X	X	X	X	X
no	05:8103550	05:1020942													X	
no	05:4822311	05:6537162	X										X			
yes *	05:36268911	05:36551688			X	X	X	X	X	X	X	X		X		X
yes	05:38666410	05:41682411											X			
yes	06:1589315	06:1607130	X		X	X										
yes	08:5836631	08:5843843	X			X					X					X
yes	08:14637964	08:14568353	X		X	X	X						X			
yes *	08:44669303	08:45103227			X		X					X				
yes	10:6392637	10:6420224	X			X					X	X				
no	10:38503885	10:39052083												X	X	X
yes *	10:40875417	10:41422677	X	X	X	X	X			X	X		X	X		X
yes	11:1057718	11:1066579	X	X	X	X	X	X	X	X						
yes	11:36804300	11:32347400									X			X	X	X
yes *	12:35682591	12:35656855	X	X	X	X	X				X	X	X		X	X
yes	15:43604507	15:44363652										X	X			

Table 4.3. Mapped mPing insertions. * indicates previously identified insertions.

sample name	#reads	3'	5'	total mPing reads	%mPing	#mapped reads	%mapped
3-47-2-3-4	59967	9380	9684	19064	31.8	13,832	72.5
3-47-2-3-5	58653	16469	11060	27529	46.9	18891	68.6
3-47-2-3-6	34781	12	11482	11494	33.0	11214	97.5
3-47-2-3-7	37150	15	16193	16208	43.6	13691	84.5
19-6-16-2-13	27369	23	11974	11997	43.8	7961	66.3
19-6-16-2-14	47062	8561	5451	14012	29.7	10465	74.6
19-6-16-2-15	59441	7742	6180	13922	23.4	11397	81.8
19-6-16-2-18	45733	3433	4583	8016	17.5	5632	70.2
16-9-9-5-4	32413	17	14229	14246	43.9	13436	94.3
16-9-9-5-5	17684	15	5546	5561	31.4	4241	76.2
16-9-9-5-6	30921	7	14056	14063	45.4	9808	69.7
32-13-A-11-8	48555	15	20461	20476	42.1	14288	69.8
32-13-A-11-9	40512	10049	4419	14468	35.7	8051	55.6
32-13-A-11-10	36374	9	11414	11423	31.4	11136	97.5
32-13-A-11-11	54132	11775	6463	18238	33.7	15987	87.6
Total	630747	67522	153195	220717	34.9	170030	77.0

Table 4.4. Read analysis summary for *mPing* samples analyzed in Geneious.

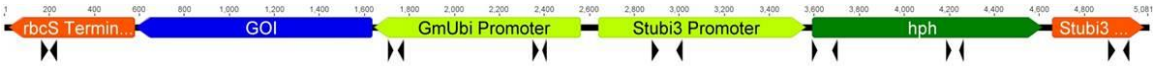


Figure 4.5. Linear vector used for biolistic transformation of transgene events analyzed. Arrow head indicate the 14 GSP_2 primer locations.

Event	# of reads	% of library	# reads assembled	% used	# flanking sequences ID	# PCR confirmed
8	197,950	8.37%	113,354	57.26%	8	4
9	240,386	10.17%	140,479	58.44%	5	2
13	248,334	10.50%	154,330	62.15%	4	2
16	236,280	9.99%	119,384	50.53%	5	1
20	247,962	10.49%	141,466	57.05%	7	3
22	236,434	10.00%	129,379	54.72%	5	1
28	248,910	10.53%	136,395	54.80%	12	0
31	210,538	8.91%	123,894	58.85%	16	5
36	246,550	10.43%	127,299	51.63%	12	1
38	250,672	10.60%	139,387	55.61%	16	1
Total	2,364,016		1,325,367	56.06%	90	20

Table 4.5. Distribution of reads, flanking sequences, and PCR-confirmed flanking sequences across transgenic events.

Amplicon	Avg	StDev	# confirmed
GmUbi156F	7.50%	1.00%	3
GmUbi220R	7.53%	2.04%	2
GmUbi795F	9.16%	1.31%	1
GmUbi842R	6.04%	0.58%	2
StUbiT253F	6.94%	2.33%	3
StUbiT346R	7.94%	2.14%	0
hyg114R	8.15%	1.04%	1
hyg577R	6.55%	1.60%	1
hyg598F	9.27%	1.47%	0
hygro2F	6.76%	1.42%	0
rbcs418R	8.70%	2.40%	0
rbcst348F	8.05%	4.14%	3
Stubi262F	2.43%	1.43%	2
Stubi389R	4.99%	1.77%	2

Table 4.6. Distribution of reads and PCR-confirmed flanking sequences across amplicons.

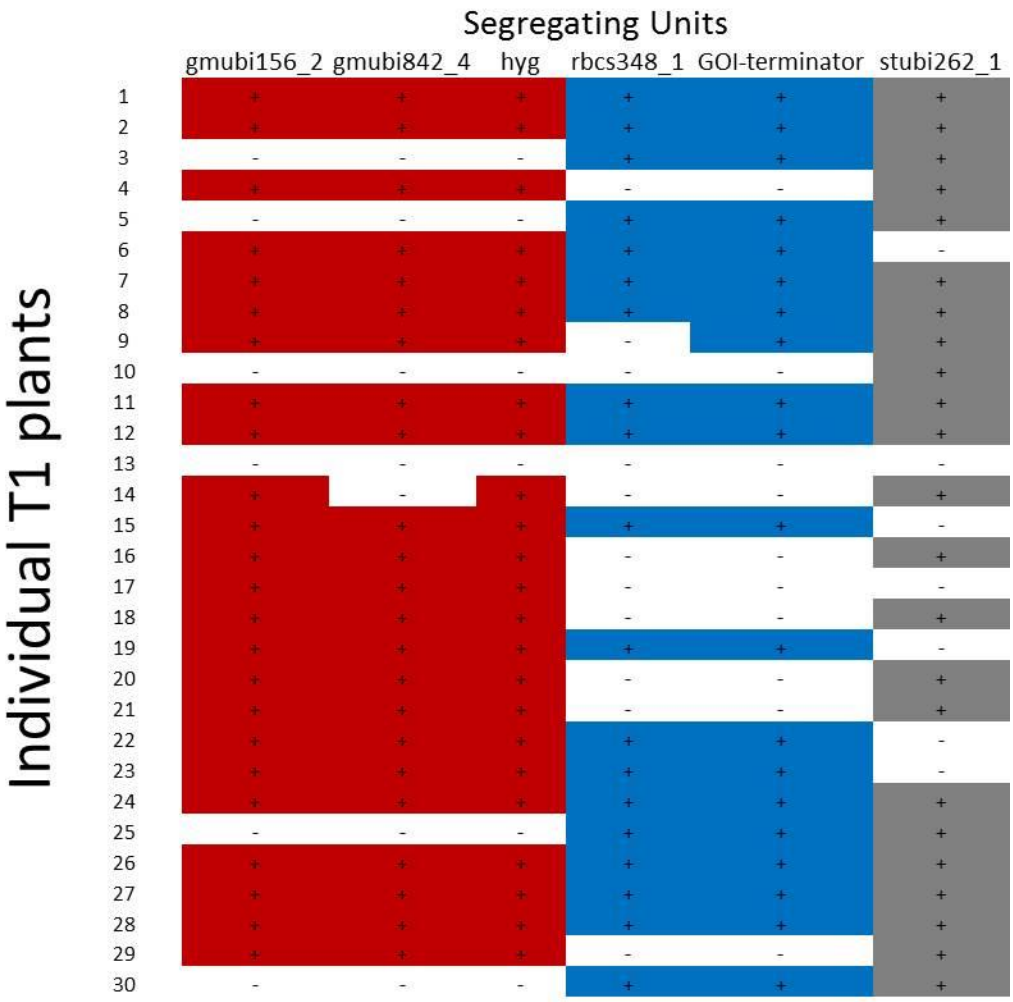


Table 4.7. Segregation analysis of segregating units in event 8. Individual T1 plants positive for the respective amplicons are indicated with a + and a colored box. Three segregating units are observed for this event.

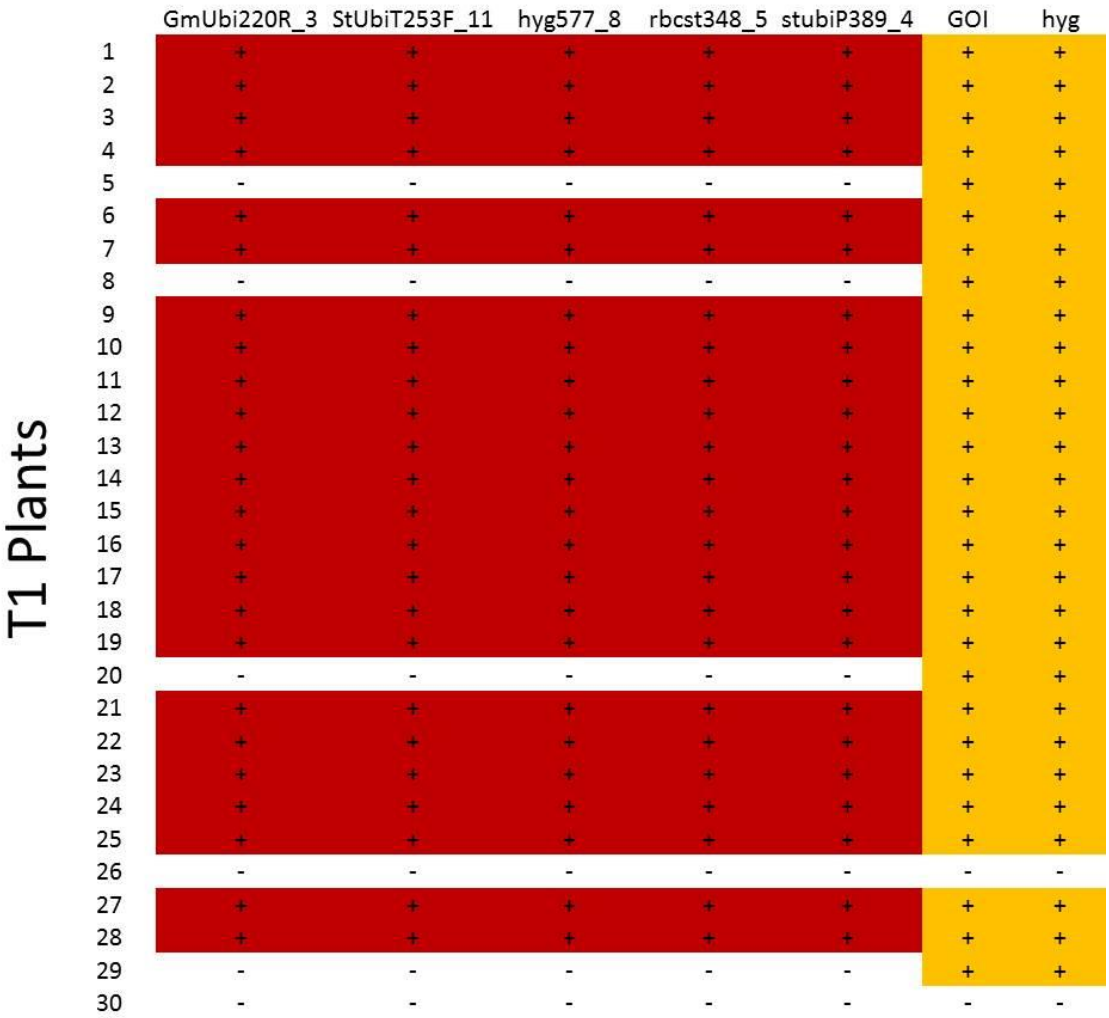


Table 4.8. Segregation analysis of segregating units in event 31. Individual T1 plants positive for the respective amplicons are indicated with a + and a colored box. Two segregating units are observed for this event, one known and one unknown. The five flanking sequences are genetically linked, suggesting that the transgenic insertion event is interspersed with genomic DNA.

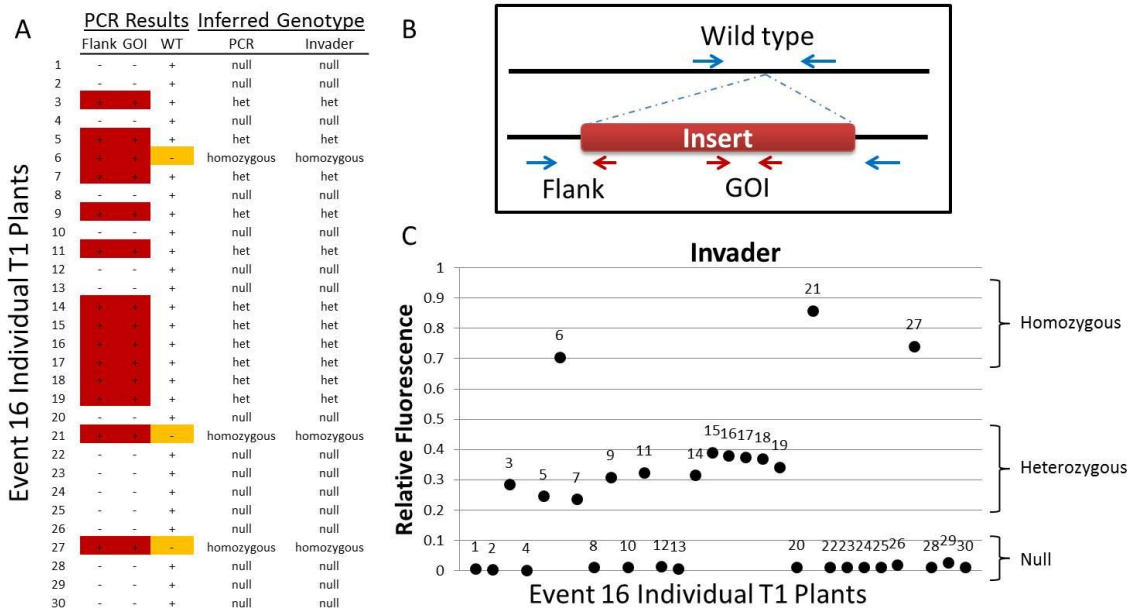


Figure 4.6. Segregation and zygosity check for event 16 individuals. PCR (A) and Invader assays (C) were performed and genotype was inferred for 30 individuals from event 16 (A). The primers used were designed within and to flank the transgene insert (B). Individual plants positive for the flank and gene-of-interest (GOI) amplicons are indicated with a + and colored red. Individual plants negative for the wild-type (WT) amplicon (homozygotes) are indicated with a – and colored gold. Invader assay results perfectly correlate with the PCR data (A, C).

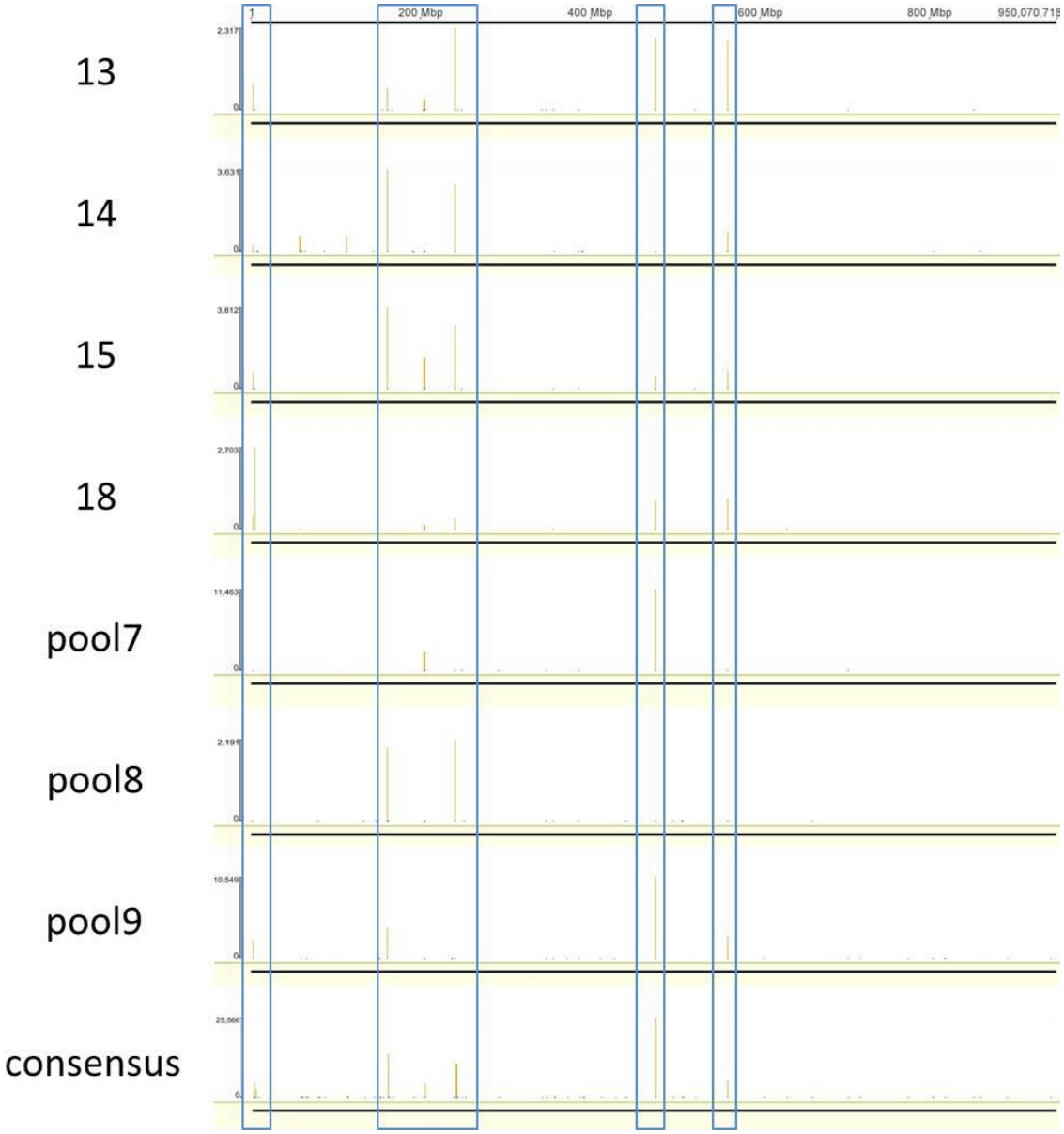


Figure 4.7. Coverage graphs of *mPing*-containing reads mapped to the soybean genome. Individuals, 13, 14, 15, and 18 are siblings from the same line. Pools 7, 8, and 9 are also from this line and contain 5-6 individuals. Insertions that are germinal (shared between individuals) clearly show up in the consensus graph (blue rectangles). Somatic insertions show up in individual graphs (e.g. shorter orange bars in 14).

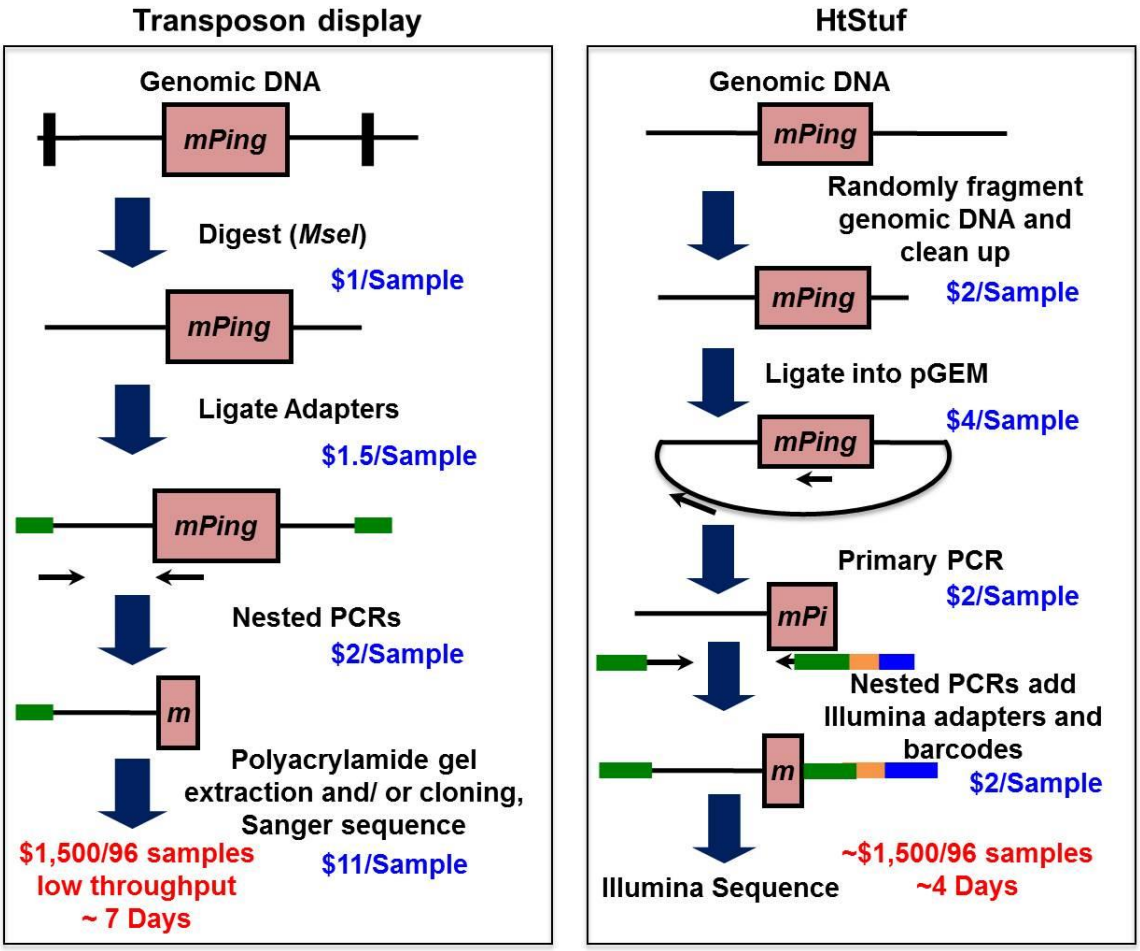


Figure 4.8. Cost comparison of major steps in transposon display (TD) and HtStuf methods. The cost for either method is roughly the same, however the man hours for HtStuf are much less than for TD. Moreover, the cost for doubling the number of samples would not change much for the HtStuf method since many more samples can be run using a higher-capacity Illumina sequencing kit.

CHAPTER 5

AN RNAI MEDIATED APPROACH TO NEMATODE RESISTANCE IN SOYBEAN: FACTORS THAT AFFECT SIRNA PRODUCTION[1]

Jacobs, Thomas; Hussey, Richard; Finer, John; LaFayette, Peter; Mitchum, Melissa; Rambo-Martin, Benjamin; Vodkin, Lila; Zernova, Olga; Parrott, Wayne. To be submitted to *Nematology*

Abstract

Soybean is one of the world's most important food crops, and in the U.S., the largest loss yield of is due to plant-parasitic nematodes. Parasitic nematodes secrete effector proteins into plants to establish feeding sites that are essential for their survival. RNA interference (RNAi) provides a potentially unique avenue by which resistance can be engineered in plants. RNAi produces small-interfering RNA (siRNA) molecules that are presumably ingested by the feeding nematodes, and result in the silencing of nematode genes. Previous reports have shown that *Arabidopsis* plants engineered to induce the silencing of nematode effector genes are resistant to nematode infection. We attempted to use the same approach to generate nematode-resistant soybean lines. Small-interfering RNAs were readily detected in transgenic soybean lines, and moderate, though non-significant, resistance was observed in one nematode bioassay. We were unable to reproduce these results in an additional bioassay. These results suggest that either RNAi is not an effective strategy for nematode resistance in soybean, or silencing the target genes is not sufficient for resistance.

Introduction

Several species of plant parasitic nematodes are found in soybean fields, and the most damaging are the soybean-cyst (SCN, *Heterodera glycines*) and the root-knot nematodes (RKN, *Meloidogyne spp*). U.S. soybean losses due to SCN are estimated to be more than one billion dollars annually [1]. RKNs are less problematic in the U.S.; however they are an important pest in the Southeast USA and in semitropical regions around the world [2].

Many quantitative trait loci have been associated with SCN resistance in soybean [3], with the *Rhg1* locus being the most widely used in commercial cultivars. As these resistance genes have been incorporated into soybean cultivars, virulent populations of nematodes have emerged in the field [4]. Continued selection for virulent nematodes will likely require the discovery and use of additional resistance genes, both endogenous and transgenic. Transgenic approaches are therefore being designed to take advantage of the nematode's feeding habits, as these provide a weak point that can be altered as a method of control.

SCN and RKN are obligate plant parasites. They enter the roots of susceptible and resistant plants alike, but it is the formation of a feeding site near or within the vascular tissue that is essential for their development and survival [5, 6]. Both RKN and SCN secrete effectors (parasitism proteins) through their hollow stylet into plant cells during the entire infection process. Parasitism proteins are produced in the dorsal or subventral gland cells in the nematode, and more than a hundred putative effector genes have been identified [7-10]. These genes are largely absent in non-parasitic nematodes and are likely what give the nematodes the ability to parasitize plants. If true, then the elimination of parasitism proteins should eliminate the nematode’s ability to properly infect, and establish a feeding site within a plant [11].

Nematode resistance in plants by RNAi was first demonstrated in *Arabidopsis* expressing a construct targeting the RKN parasitism gene, *16D10* [12]. The authors suggested that when the nematodes fed upon the transgenic plants, they ingested small-interfering RNAs (siRNAs), which were then able to enter the nematode RNAi pathway and down-regulate the *16D10* gene. Vectors targeting *16D10* have since been used to

reduce the number of *M. chitwoodi* egg masses and eggs in transgenic potato lines, and a corresponding reduction in the expression of the *16D10* gene in J2 offspring was observed [13, 14]. However, in both reports it was unclear if small RNAs were being produced in the transgenic plants. In transgenic grape roots, *M. incognita* egg production was reduced; however, there was no effect on numbers of egg masses and the expression level of the nematode target gene was not quantified [15].

In *Arabidopsis*, plants expressing RNAi constructs to four putative parasitism genes were resistant to *H. schachtii* infection [16]. Corresponding with the lower *H. schachtii* female and egg counts, there was a significant reduction in the parasitism gene transcripts in the nematodes. *H. schachtii* resistance has also been obtained in *Arabidopsis* by targeting two other parasitism proteins [17, 18].

In plants, RNAi can be initiated from transgenic vectors containing an inverted repeat of the coding sequence from a gene target. The inverted repeat is separated by a loop sequence that is usually an intron [19]. Upon transcription, the inverted repeats hybridize with one another to form a double-stranded RNA (dsRNA) stem-loop, referred to as the hairpin. The inverted repeats make up the arms of the hairpin, and the intron is the loop. The dsRNA is then processed by a DICER-like enzyme to produce siRNAs. The siRNAs are bound by an Argonaute (AGO) protein to form an RNA-induced silencing complex (RISC). The RISC then cleaves mRNAs that are complementary to the siRNA, thus resulting in gene silencing [20]. The general RNAi mechanism is well conserved between animals, plants, and fungi [20].

RNA interference has also been used to generate transgenic plants resistant to other plant pests. Reports in corn [21] and cotton [22] have shown that transgenic events

that accumulate dsRNA are resistant to two coleopteran species, western corn rootworm and cotton bollworm, respectively. In both cases, the dsRNA-producing events also produced siRNAs to the target genes, and target mRNA levels were reduced in the target species. Powdery mildew resistance has also been obtained in wheat and barley with RNAi [23]. However, not all RNAi reports have been successful. Fairbairn et al. [24] showed a reduction in the mRNA levels of the zinc-finger transcription factor *MjTisII* in feeding *M. javanica* and the production siRNAs *in planta*; however, there was no effect on nematode survival. This report suggests that not all targets of RNAi will lead to effective resistance.

In soybean, the expression of an SCN dsRNA, homologous to a sperm protein essential for reproduction in *C. elegans*, led to a reduction in egg number from nematodes that fed upon the transgenic plants [25]. While the expression level of the target mRNA was not evaluated, siRNAs to the target gene were detected. There are several reports where composite plants (wild-type shoots, transgenic roots) engineered with RNAi vectors are nematode resistant [26-29]. The hairpin vectors used in these reports target putatively essential genes in SCN that are preferentially expressed in infecting juveniles. Although siRNA analyses were performed in only one report [28], the reduction in SCN females was attributed to an RNAi effect in all of them.

This report details several attempts to obtain nematode resistance in soybean by silencing nematode parasitism genes in SCN and RKN. Using small RNA deep sequencing, we identify a number of possible factors that contribute to the production of siRNAs from hairpin vectors. Despite being able to produce soybean lines with high numbers of siRNAs, none of the transgenic plants were resistant to nematode infection.

Results and Discussion

Hairpin vectors targeting eight SCN [9] and ten RKN [10] parasitism genes were transformed into soybean (Table 5.1). However, none of the 65 SCN-targeting and 114 RKN-targeting events had reduced numbers of feeding nematodes or eggs in nematode bioassays (data not shown). Using sRNA deep-sequencing analysis, appreciable levels of siRNAs were only detected in one line expressing 18H08 dsRNA, and it primarily produced siRNAs from the positive strand, which would be unable to cleave the target mRNA (Figure 5.1). The data suggest that the lack of nematode resistance was because the hairpin vectors were not producing siRNAs, or when they were, the siRNAs were to the wrong strand. In contrast, two of these nematode genes, *16D10* and *30C02*, elicited siRNA production and nematode resistance when placed in pHANNIBAL hairpin vectors and transformed into *Arabidopsis* [18, 30].

Troubleshooting began by re-validating the hairpin vectors by targeting the transgene *GUSPlus* for silencing. Multiple configurations of the hairpins were made to determine if the orientation of the hairpin arms, or the type of loop used, might influence the ability to silence and produce siRNAs. Hairpin vectors contained convergent or divergent arms, and an intron or spacer sequence (portion of *gusA*) for the loops (Figure 5.2A). The convergent:*FAD3* orientation was the configuration used in the initial SCN and RKN hairpin vectors. The different hairpin and control vectors were transformed into a soybean line expressing the *GUSPlus* gene. In hairy roots, silencing was observed in events derived from each vector. GUSPlus activity was at, or below, 20% of the empty-vector control in 31/36 roots (Figure 5.2). Two empty-vector control events were spontaneously silenced for GUSPlus, indicating that a background level of GUSPlus

silencing was occurring in the hairy roots. To confirm that silencing was an RNAi-mediated response, small RNA sequencing was performed on the silenced lines. Large numbers of siRNAs were detected from each of the hairpin configurations (Figure 5.3). These results clearly demonstrate that the hairpin vectors used were able to produce siRNAs and silence transgenes very effectively, even if they were unable to do likewise for the nematode-derived targets.

The robust silencing of *GUSPlus*, and production of siRNAs, may be due to the presence of the target template (GUSPlus mRNA) in the plant, which would permit a siRNA amplification step. In plants, small RNA amplification occurs when siRNAs prime dsRNA synthesis by RNA-dependent RNA polymerases, using mRNA as a template [31], leading to large numbers of secondary siRNAs. Given that the nematode-targeting hairpin vectors do not have a complementary mRNA sequence in the plant from which to amplify from, the siRNAs produced would be limited to the primary siRNA pool, which is much smaller than the amplified pool.

To determine if amplification could increase the production of siRNAs from the nematode-derived genes, the respective nematode target sequences were transformed into the hairpin-containing lines. In case the length or stability of the mRNA templates affected amplification, three target templates were made; one arm from the original hairpin (300 bp), the full-length nematode target transcript (292-1138 bp), or the hairpin arm fused to a *bar* coding sequence, figuring the bar sequence could stabilize the mRNA. These vectors were tested in hairy roots from genotypes previously transformed with the corresponding hairpin vectors. If amplification occurred, then there should be an increase in the siRNA production. The target templates were able to increase the abundance of

siRNAs in two lines, 18H08 and 9H10, with 9H10 being the most responsive (Table 5.2). The siRNAs produced in the 33A09 and 5C03B lines were not affected. Generally, the bar-fusion vector was most effective, as the 18H08 and 9H10 bar-fusion roots had over 1,000 total normalized small RNAs. The possibility that the target templates were being co-suppressed by the plants cannot be ruled out, which would result in the production of siRNAs without the hairpins.

The hairpin cassettes used for soybean transformation have several differences from the pHANNIBAL vectors used in previous nematode-silencing reports in *Arabidopsis* [16-18]. The hairpin vectors are driven by the GmUbi promoter [32], have a *FAD3* intron for the loop [33], and are terminated by the *rbcS* terminator [34], whereas the pHANNIBAL vector uses a 35S promoter, a *PDK* intron, and the *nos* terminator [35]. To rule out any possible vector differences to generate siRNA, a factorial experiment was performed to determine if any of the vector components (promoter, loop, or terminator) could explain the low siRNA yields. The *30C02* nematode gene was used for the hairpin arms, and since *30C02* is not similar to any other gene in the soybean genome, only primary siRNAs should be detected. Three hairy root events per vector were pooled for small RNA sequencing, and sequencing results confirmed the production of small RNAs in events from each of the vectors (Table 5.3). The promoter used was the only significant effect of the different components ($p = 0.0362$). On average, GmUbi-promoter events had 10-fold fewer siRNAs than events with the 35S promoter. These data suggest that the 35S promoter is superior to the GmUbi promoter for the production of primary siRNAs in soybean roots. This result was not expected; the GmUbi promoter is a strong constitutive promoter that is about five times stronger than the 35S promoter

when driving the expression of the *GFP* gene [36]. There may be an optimal level of expression to induce optimal siRNA production.

While the data show that the hairpin vectors were able to produce siRNAs and silence transgenes in hairy roots, a pHANNIBAL-based hairpin vector was used for further work since it produced siRNAs from nematode-derived targets when the original hairpin could not, and had been used to produce nematode-resistant *Arabidopsis* in previous reports [16-18]. For the next round of soybean transformations, efforts centered on five SCN targets that had shown a measure of effectiveness in *Arabidopsis*; 10A06, 4G06, 8H07 [16], 4F01 [17] and 30C02 [18]. The *10A06*, *4G06*, and *8H07* genes are similar to genes in the SCF complex responsible for ubiquitination [16], *4F01* is similar to a plant annexin [17], and 30C02 is novel protein that interacts with a β-1, 3-endoglucanase in plant cells [18].

Several quality-control checks were used to screen the recovered transgenic soybean lines. Across all transformation labs, events missing one hairpin arm or the other were observed, so events were screened for the presence of both hairpin arms by PCR. Second, transgenic events had to express the silencing construct as determined by end-point RT-PCR. Through qRT-PCR and small RNA sequencing analyses, events expressing the hairpin vector were also produced siRNAs, while the lines not expressing the vector did not produce any siRNAs (Figure 5.4). Within the expressing lines, no correlation between expression of the hairpin construct and siRNA production was observed. Therefore, end-point RT-PCR is sufficient to identify lines that produce siRNAs. Only lines that contained both hairpin arms and were expressing the hairpin vectors were used for nematode bioassays.

Twelve lines expressing the 10A06 (1 line), 4F01 (4 lines), 30C02 (3 lines), 8H07 (3 lines) and 4G06 (1 line) pHANNIBAL hairpins were used in a nematode bioassay. All lines had, on average, a lower nematode count as compared to the wild-type controls (Figure 5.5). Average nematode counts in the transgenic lines ranged from 68-97% of the wild-type controls. However, none of the differences were significant ($p = 0.1134$). PCR genotyping indicated that some of the lines were segregating for the hairpin vectors. Removing these data points from the analysis did not alter the results. The null segregants had normalized nematode counts that covered the same range as the events containing the hairpin vectors (Figure 5.5).

The siRNAs produced by silencing vectors are thought to be the causative molecules that induce RNAi; therefore, it should be possible to correlate siRNA abundance with nematode counts. Such a correlation was tested by sequencing small RNAs on select plants within lines that had either a high or low nematode count. Tens, to tens of thousands of normalized small RNAs were detected in the ten lines sequenced (Table 5.4). The number of siRNAs was consistent between individual plants within a line. While some of the individual plants were accumulating siRNAs to a very high level, there was no correlation between siRNA abundance and nematode counts. For instance, the 30C02-targeting events behaved similarly in the bioassay (83-97%), yet lines 25 and 28 barely produced any siRNAs, whereas line 43 had, on average, 13,000 normalized siRNAs (Table 5.4).

Five of the most promising lines were used in a second nematode bioassay; however, no reduction in nematode counts was observed (Figure 5.6). The inconsistent bioassay results between rep 1 and 2 are reminiscent of a previous report in *Arabidopsis*

silencing 10A06, 4G06, and 8H07 [16]. In this report, there was high variance within transgenic lines and between experiments. Some transgenic lines had significant reductions in nematodes in one replication, but would not be significantly different in an additional replication, and more likely reflect escapes than true resistance.

Our bioassay results are in contrast with previous reports using RNAi to obtain nematode resistance in soybean. In this study, we attempted to silence nematode genes that are suspected of being required for parasitism, as opposed to the other reports that target putatively essential genes previously identified in *C. elegans* [25-29]. Parasitism genes are interesting gene targets as they should be specific to plant-parasitic nematodes and are less likely to have off-target effects in beneficial nematodes or other animals. However, this potential advantage may be countered by the inability to obtain strong nematode resistance in soybean.

Plant-parasitic nematodes may use hundreds of effectors to infect their host plants [7]. It has been suggested that some of these nematode effectors have redundant functions [37, 38], such that if one effector fails to work, others can be used to accomplish the same task. Silencing a single parasitism gene may not be sufficient to observe a phenotype; therefore, if parasitism genes are to be targeted, it may be necessary to use another approach such as targeting multiple effectors at a time. This can be accomplished by creating silencing vectors with multiple target sequences fused together.

This is the first report evaluating RNAi to obtain nematode resistance in the progeny of transgenic soybean plants. Previous reports in soybean were either in hairy-root assays [26-29] or T0 transgenic plants [25]. Since plants cannot be generated from hairy roots, the effect of transgenes cannot be evaluated across generations. T0's are the

initial transgenic plants that emerge from tissue culture and tend to have abnormal phenotypes. Furthermore, the results from the T0 plants were from a single bioassay where only 14 plants, derived from four independent events, were evaluated. Given the high variability of nematode bioassays, several individual plants per line are required to determine true resistance from escapes. Many of the null segregants identified here had nematode counts that were 50% that of the negative controls (Figure 5.5). This apparent high variability also underscores the importance of testing for siRNAs to show the causal link between hairpin vectors and observed resistance.

Conclusions

The use of RNAi to control nematodes remains a viable goal, particularly since there are so few alternatives for effective and sustainable control. Nevertheless, it remains a challenging goal, and while there have been many promising reports, definitive proof in crop plants is still lacking. Initial attempts to silence nematode parasitism genes to obtain nematode resistance were unsuccessful, but did give insight into the technological challenges associated with hairpin vectors that need to be overcome. It is clear from this work that not all gene-vector combinations are able to produce siRNAs. While the hairpin vectors were able to generate siRNAs from *GUSPlus* and the nematode gene *30C02*, no siRNAs were detected in the initial transgenic lines. The use of target templates for the generation of secondary siRNAs via amplification might be useful in some cases. Promoter strength is important in determining siRNA production, but the basis for this phenomenon remains to be explored. Even when transgenic soybean plants transformed with pHANNIBAL hairpins produced a large number of siRNAs, a

significant nematode reduction was not observed. Therefore, the last outstanding question is whether or not the target genes are being silenced in the nematode. Once this issue is resolved, future experiments may need to target multiple nematode parasitism genes or resort to different types of target genes. Results from model systems such as *Arabidopsis* and hairy roots do not extrapolate to whole soybean plants.

Materials and Methods

Vector construction

The binary vector p201N is a modified p201BK [39] vector that contains an nptII selectable cassette (StUbi-3P:nptII:StUbi-3T) controlled by the Solanum tuberosum Ubi-3 promoter and terminator [40]. The GmUbi promoter [32] drives the expression of the hairpin cassette (GmUbi:hairpin:rbcST) and is terminated by the Pisum sativum rbcS terminator [41]. The soybean FAD3 intron [33] or gusA fragment [42], was inserted into the p201N vector in-between the GmUbi promoter and rbcS terminator, and was flanked by the restriction sites AscI/SwaI on the 5' end, and BamHI/AvrII on the 3' end. The convergent GUSPlus hairpin arms were amplified by a forward primer containing AscI/AvrII restriction sites on the 5' tail, and a reverse primer with SwaI/BamHI restriction sites on the 5' tail (Table 5.5). Divergent arms were made by swapping the restriction sites between the forward and reverse primers. The 3' hairpin arm was first then inserted into the vector using the AvrII/BamHI restriction sites. Then the 5' hairpin arm was inserted into the p201N vector by the AscI/SwaI restriction sites.

The empty target-template cassette (CaMV35S:MCS:nosT) was inserted into a modified p201BK vector [39], p201G2 (CsVMV-P:GFP:nosT) with the I-SceI restriction

site to make p201G2-35. The bar gene was amplified and inserted into the MCS between AvrII and XhoI make p201G2-35Bar. The 5C03B target was amplified and inserted into p201G2-35 with AvrII/XhoI restriction sites, and into p201G2-35Bar with SwaI/XhoI restriction sites to maintain the reading frame of bar. For all other targets, In-Fusion® cloning (Clonetech Laboratories Inc.) was used to insert the target template or target template + bar into p201G2-35 between the AvrII and XhoI restriction sites.

For the full-factorial hairpin vectors, within the p201N vector, GmUbi, FAD3, and rbcS were replaced with their respective components from the pHANNIBAL vector. The same hairpin arm cloning strategy used for the GUSPlus hairpin vectors was used to introduce the 30C02 hairpin arms. All binary vectors were transformed into Agrobacterium rhizogenes strain K599 [43] via electroporation.

The pHANNIBAL hairpin vectors were created by placing a hygromycin phosphotransferase gene under the control of the Ubi3 promoter and terminator and inserted into a pSMART HC Kan (Lucigen Corporation, Middleton WI, accession number AF532107) vector to produce pSPH2. The pHANNIBAL hairpin cassettes were moved from the pART27 binary vectors [16-18] with NotI and into pSPH2 for the selection of transformed soybean cells.

Hairy-root transformation of soybean and MUG analysis

Hairy-transformations were performed as previously described [43] with slight modifications. Soybean seeds were germinated under sterile conditions in a Petri dish wetted with a ½ MSO liquid germination medium [44] supplemented with B5 vitamins

[45]. A. rhizogenes containing the hairpin vectors were grown for two days at 28°C on YM medium [46] supplemented with 50 mg L^{-1} kanamycin. Soybean cotyledons were inoculated as previously described [47]; the root, lower hypocotyl, and primary shoot were excised from the cotyledons, leaving an approximately 5 mm tail of hypocotyl attached to the cotyledons. The hypocotyl tails were cut longitudinally to produce two half cotyledons, each with a short hypocotyl piece. A scalpel dipped in a solution of A. rhizogenes (PB Buffer (0.01 M Na2HPO4, 0.15M NaCl, pH 7.5) + 100 µM acetosyringone) was used to make 1mm-deep cuts on the adaxial side of the cotyledons and hypocotyl piece. The explants were co-cultivated for three days on filter paper wetted with 2 mL of ½ MSO liquid germination medium + 100 µM acetosyringone. Explants were transferred to a solid ½ MSO medium according to Cho et al. [43] with the following modifications: ½ MS salts, 2 g L^{-1} Phytagel, and 500 mg L^{-1} timentin. Each emerging root was considered an individual event and transferred to solid ½ MSO medium with 10 mg L^{-1} of the selective agent Geneticin (G418). Roots growing on the selective medium were considered events.

GUSPlus-targeting, hairpin vectors were introduced into a soybean line homozygous for pCAMBIA1305.2 [Genbank AF354046], which contains the GUSPlus gene under the expression of the double 35S promoter. Twelve independent events from each vector were randomly selected, and 100 mg of hairy-root tissue was ground in 1X LB Buffer [48] for MUG analysis [49]. Fluorescence was measured with a Synergy 2 plate reader (BioTek Instruments, Inc.). Total protein was measured with a Bradford assay [50], and was used to normalize MUG values.

For the full-factorial experiment, roots were prepared as described above. Three roots per vector, containing both hairpin arms, as determined by PCR, were used for small RNA sequencing. Primer sequences can be found in table 5.5.

RNA extraction and Small RNA sequencing

Plant materials used for small RNA sequencing were immediately frozen in liquid nitrogen and stored at -80°C until processing. Samples were then freeze-dried and total RNA was extracted using phenol-chloroform and lithium-chloride precipitation as previously described [51].

Small RNA libraries for Illumina sequencing were prepared according to manufacturer's instructions and sequenced using the Illumina GAIIx at the University of Illinois Keck Center. After trimming the 3' adapter sequence, reads 18-25 nt in length were mapped to the respective gene targets using Bowtie2 [52].

Small RNA sequencing and qRT-PCR of 10A06 samples

Total RNA was extracted from the roots of homozygous 10A06 events using Tri-Reagent (Ambion®) according to manufacturer's instructions. One microgram of total RNA was used to generate small RNA sequencing libraries using Illumina's TruSeq™ small RNA library kit. The MiSeq was used for small RNA sequencing according to the manufacturer's instructions. Raw reads were separated by barcodes using the MiSeq Reporter software. Fastq files were imported into Geneious, which was used trim the adapters and select reads 18-25 nt in length for assembly. The reads were mapped to the 10A06 target mRNA using the following conditions: Gaps not allowed, word length 18,

index word length 13, ignore words repeated more than 5 times, maximum mismatches per read 0%, and maximum ambiguity 4.

For qRT-PCR, total RNA was treated with Turbo DNase™ (Ambion®). One-hundred nanograms of DNase treated total RNA were used as the template in the Go Taq® 1-Step RT-qPCR system (Promega). The qRT-PCR reaction was performed in a Light Cycler 480II (Roche Diagnostics GmbH) at 37°C for 15 minutes; 95°C for 10 minutes; 40 cycles (95°C for 10 seconds; 60°C for 30 seconds; 72°C for 30 seconds); a melt-curve analysis from 60°C to 95° at a ramp rate of 0.11°C/second. The melt-curve analysis was used to confirm the specificity of the qRT-PCR reaction. OCS amplicons were Sanger sequenced to ensure amplification of the correct gene. The metalloprotease amplicon [53], was used to normalize expression. The ΔCt values for each event were calculated by the LightCycler® 480 SW 1.5.1 program using the Advanced Relative Quantification analysis.

End-point RT-PCR

End-point RT-PCR was used to confirm the expression of the pHANNIBAL hairpin vectors in all transgenic lines. Five-hundred nanograms of DNase-treated, total RNA was used for cDNA synthesis with an oligo(dT) primer using GoScript™ reverse transcriptase (Promega) according to manufacturer's instructions. The OCSF2/R2 primer set was used under standard PCR conditions to amplify the 5' end of the ocs terminator. PCR products were visualized on a 2% agarose, TBE gel.

DNA Extraction

DNA was extracted from leaf and root tissues using a modified CTAB protocol [54]. Genotyping was performed with primers in table 5.5 under standard PCR conditions. PCR products were visualized on a 1% agarose, TBE gel.

Nematode bioassays

Nematode bioassays were performed as previously described [55]. The transgenic lines used in the bioassay were in the background Jack cross. Eight seedlings were used for each transgenic line and wild-type controls. As the bioassays were disassembled, leaf samples were immediately frozen for DNA and RNA extraction. The first bioassay used transgenic lines 10A06-1, 4F01-18, 30C02-43, 8H07-82, 8H07-84, and 4G06-89. An independent bioassay was performed on transgenic lines 4F01-22, 4F01-17, 4F01-24, 30C02-25, 30C02-28, and 8H07-71. Nematode counts were normalized to the respective mean nematode count of the non-transformed controls, and were combined for the results in figure 5.5. Based on the results from the first set of bioassays, lines 10A06-1, 30C02-43, 30C02-25, 8H07-82 and 4F01-22 were selected for a second bioassay.

References:

1. Koenning SR, Wrather JA: **Suppression of soybean yield potential in the continental United States from plant diseases estimated from 2006 to 2009**. *Plant Health Prog* 2010.

2. Wrather JA, Koenning SR: **Estimates of disease effects on soybean yields in the United States 2003 to 2005**. *Journal of Nematology* 2006, **38**(2):173-180.

3. Concibido VC, Diers BW, Arelli PR: **A decade of QTL mapping for cyst nematode resistance in soybean**. *Crop Science* 2004, **44**(4):1121-1131.

4. Niblack TL, Colgrove KB, Colgrove AC: **Soybean cyst nematode in Illinois from 1990 to 2006: Shift in virulence phenotype of field populations**. *Journal of Nematology* 2006, **38**(2):285-285.

5. Niblack TL, Lambert KN, Tylka GL: **A model plant pathogen from the kingdom animalia: *Heterodera glycines*, the soybean cyst nematode**. *Annual Review of Phytopathology* 2006, **44**:283-303.

6. Gheysen G, Mitchum MG: **How nematodes manipulate plant development pathways for infection**. *Current Opinion in Plant Biology* 2011, **14**(4):415-421.

7. Elling AA, Mitreva M, Gai XW, Martin J, Recknor J, Davis EL, Hussey RS, Nettleton D, McCarter JP, Baum TJ: **Sequence mining and transcript profiling to explore cyst nematode parasitism**. *Bmc Genomics* 2009, **10**:17.

8. Davis EL, Hussey RS, Mitchum MG, Baum TJ: **Parasitism proteins in nematode-plant interactions**. *Current Opinion in Plant Biology* 2008, **11**(4):360-366.

9. Gao BL, Allen R, Maier T, Davis EL, Baum TJ, Hussey RS: **The parasitome of the phytonematode *Heterodera glycines***. *Molecular Plant-Microbe Interactions* 2003, **16**(8):720-726.

10. Huang GZ, Gao BL, Maier T, Allen R, Davis EL, Baum TJ, Hussey RS: **A profile of putative parasitism genes expressed in the esophageal gland cells of the root-knot nematode *Meloidogyne incognita***. *Molecular Plant-Microbe Interactions* 2003, **16**(5):376-381.

11. Davis EL, Mitchum MG: **Nematodes. Sophisticated parasites of legumes**. *Plant Physiology* 2005, **137**(4):1182-1188.

12. Huang GZ, Allen R, Davis EL, Baum TJ, Hussey RS: **Engineering broad root-knot resistance in transgenic plants by RNAi silencing of a conserved and essential root-knot nematode parasitism gene**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(39):14302-14306.

13. Dinh PTY, Zhang LH, Brown CR, Elling AA: **Plant-mediated RNA interference of effector gene *Mc16D10L* confers resistance against *Meloidogyne chitwoodi* in diverse genetic backgrounds of potato and reduces pathogenicity of nematode offspring**. *Nematology* 2014, **16**:669-682.

14. Dinh PTY, Brown CR, Elling AA: **RNA interference of effector gene *Mc16D10L* confers resistance against *Meloidogyne chitwoodi* in *Arabidopsis* and Potato**. *Phytopathology* 2014, **104**(10):1098-1106.

15. Yang YZ, Jittayasothorn Y, Chronis D, Wang XH, Cousins P, Zhong GY: **Molecular characteristics and efficacy of *16D10* siRNAs in inhibiting root-knot nematode infection in transgenic grape hairy roots**. *Plos One* 2013, **8**(7):13.

16. Sindhu AS, Maier TR, Mitchum MG, Hussey RS, Davis EL, Baum TJ: **Effective and specific in planta RNAi in cyst nematodes: expression interference of four parasitism genes reduces parasitic success**. *Journal of Experimental Botany* 2009, **60**(1):315-324.

17. Patel N, Hamamouch N, Li CY, Hewezi T, Hussey RS, Baum TJ, Mitchum MG, Davis EL: **A nematode effector protein similar to annexins in host plants**. *Journal of Experimental Botany* 2010, **61**(1):235-248.

18. Hamamouch N, Li CY, Hewezi T, Baum TJ, Mitchum MG, Hussey RS, Vodkin LO, Davis EL: **The interaction of the novel 30C02 cyst nematode effector protein with a plant beta-1,3-endoglucanase may suppress host defense to promote parasitism**. *Journal of Experimental Botany* 2012, **63**(10):3683-3695.

19. Smith NA, Singh SP, Wang MB, Stoutjesdijk PA, Green AG, Waterhouse PM: **Gene expression - Total silencing by intron-spliced hairpin RNAs**. *Nature* 2000, **407**(6802):319-320.

20. Carthew RW, Sontheimer EJ: **Origins and mechanisms of miRNAs and siRNAs**. *Cell* 2009, **136**(4):642-655.

21. Baum JA, Bogaert T, Clinton W, Heck GR, Feldmann P, Ilagan O, Johnson S, Plaetinck G, Munyikwa T, Pleau M *et al*: **Control of coleopteran insect pests through RNA interference**. *Nature Biotechnology* 2007, **25**(11):1322-1326.

22. Mao YB, Cai WJ, Wang JW, Hong GJ, Tao XY, Wang LJ, Huang YP, Chen XY: **Silencing a cotton bollworm P450 monooxygenase gene by plant-mediated RNAi impairs larval tolerance of gossypol**. *Nature Biotechnology* 2007, **25**(11):1307-1313.

23. Nowara D, Gay A, Lacomme C, Shaw J, Ridout C, Douchkov D, Hensel G, Kumlehn J, Schweizer P: **HIGS: Host-induced gene silencing in the obligate biotrophic fungal pathogen *Blumeria graminis***. *Plant Cell* 2010, **22**(9):3130-3141.

24. Fairbairn DJ, Cavallaro AS, Bernard M, Mahalinga-Iyer J, Graham MW, Botella JR: **Host-delivered RNAi: an effective strategy to silence genes in plant parasitic nematodes**. *Planta* 2007, **226**(6):1525-1533.

25. Steeves RM, Todd TC, Essig JS, Trick HN: **Transgenic soybeans expressing siRNAs specific to a major sperm protein gene suppress *Heterodera glycines* reproduction**. *Functional Plant Biology* 2006, **33**(11):991-999.

26. Youssef RM, Kim KH, Haroon SA, Matthews BF: **Post-transcriptional gene silencing of the gene encoding aldolase from soybean cyst nematode by transformed soybean roots**. *Experimental Parasitology* 2013, **134**(2):266-274.

27. Ibrahim HMM, Alkharouf NW, Meyer SLF, Aly MAM, El-Din A, Hussein EHA, Matthews BF: **Post-transcriptional gene silencing of root-knot nematode in transformed soybean roots**. *Experimental Parasitology* 2011, **127**(1):90-99.

28. Li JR, Todd TC, Oakley TR, Lee J, Trick HN: **Host-derived suppression of nematode reproductive and fitness genes decreases fecundity of *Heterodera glycines* Ichinohe**. *Planta* 2010, **232**(3):775-785.

29. Klink VP, Kim KH, Martins V, MacDonald MH, Beard HS, Alkharouf NW, Lee SK, Park SC, Matthews BF: **A correlation between host-mediated expression of parasite genes as tandem inverted repeats and abrogation of development of female *Heterodera glycines* cyst formation during infection of *Glycine max***. *Planta* 2009, **230**(1):53-71.

30. Huang G, Allen R, Davis EL, Baum TJ, Hussey RS: **Engineering broad root-knot resistance in transgenic plants by RNAi silencing of a conserved and essential root-knot nematode parasitism gene**. *Proc Natl Acad Sci U S A* 2006, **103**(39):14302-14306.

31. Voinnet O: **Use, tolerance and avoidance of amplified RNA silencing by plants**. *Trends in Plant Science* 2008, **13**(7):317-328.

32. Hernandez-Garcia CM, Martinelli AP, Bouchard RA, Finer JJ: **A soybean (*Glycine max*) polyubiquitin promoter gives strong constitutive expression in transgenic soybean**. *Plant Cell Reports* 2009, **28**(5):837-849.

33. Siminszky B, Gavilano L, Bowen SW, Dewey RE: **Conversion of nicotine to nornicotine in *Nicotiana tabacum* is mediated by *CYP82E4*, a cytochrome P450 monooxygenase**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(41):14919-14924.

34. Watson JM, Fusaro AF, Wang MB, Waterhouse PM: **RNA silencing platforms in plants**. *Febs Letters* 2005, **579**(26):5982-5987.

35. Wesley SV, Helliwell CA, Smith NA, Wang MB, Rouse DT, Liu Q, Gooding PS, Singh SP, Abbott D, Stoutjesdijk PA *et al*: **Construct design for efficient, effective and high-throughput gene silencing in plants**. *Plant Journal* 2001, **27**(6):581-590.

36. Chiera JM, Bouchard RA, Dorsey SL, Park E, Buenrostro-Nava MT, Ling PP, Finer JJ: **Isolation of two highly active soybean (*Glycine max* (L.) Merr.) promoters and their characterization using a new automated image collection and analysis system**. *Plant Cell Reports* 2007, **26**(9):1501-1509.

37. Rosso MN, Jones JT, Abad P: **RNAi and functional genomics in plant parasitic nematodes**. In: *Annual Review of Phytopathology*. vol. 47. Palo Alto: Annual Reviews; 2009: 207-232.

38. Mitchum MG, Hussey RS, Baum TJ, Wang XH, Elling AA, Wubben M, Davis EL: **Nematode effector proteins: an emerging paradigm of parasitism**. *New Phytologist* 2013, **199**(4):879-894.

39. Covert SF, Kapoor P, Lee MH, Briley A, Nairn CJ: ***Agrobacterium tumefaciens*-mediated transformation of *Fusarium circinatum***. *Mycological Research* 2001, **105**:259-264.

40. Joshi M, Niu C, Fleming G, Hazra S, Chu Y, Nairn CJ, Yang HY, Ozias-Akins P: **Use of green fluorescent protein as a non-destructive marker for peanut genetic transformation**. *In Vitro Cellular & Developmental Biology-Plant* 2005, **41**(4):437-445.

41. An G, Watson BD, Stachel S, Gordon MP, Nester EW: **New cloning vehicles for transformation of higher-plants**. *Embo Journal* 1985, **4**(2):277-284.

42. Mann DGJ, LaFayette PR, Abercrombie LL, King ZR, Mazarei M, Halter MC, Poovaiah CR, Baxter H, Shen H, Dixon RA *et al*: **Gateway-compatible vectors for high-throughput gene functional analysis in switchgrass (*Panicum virgatum* L.) and other monocot species**. *Plant Biotechnology Journal* 2012, **10**(2):226-236.

43. Cho HJ, Farrand SK, Noel GR, Widholm JM: **High-efficiency induction of soybean hairy roots and propagation of the soybean cyst nematode**. *Planta* 2000, **210**(2):195-204.

44. Murashige T, Skoog F: **A revised medium for rapid growth and bio assays with tobacco tissue cultures**. *Physiologia Plantarum* 1962, **15**(3):473-497.

45. Gamborg OL, Miller RA, Ojima K: **Nutrient requirements of suspension cultures of soybean root cells**. *Experimental Cell Research* 1968, **50**(1):151-&.

46. Lin JJ: **Optimization of the transformation efficiency of *Agrobacterium-tumefaciens* cells using electroporation.** *Plant Science* 1994, **101**(1):11-15.

47. Olhoft PM, Flagel LE, Donovan CM, Somers DA: **Efficient soybean transformation using hygromycin B selection in the cotyledonary-node method**. *Planta* 2003, **216**(5):723-735.

48. Sivamani E, Qu R: **Expression enhancement of a rice polyubiquitin gene promoter**. *Plant Molecular Biology* 2006, **60**(2):225-239.

49. Leckie F, Devoto A, Delorenzo G: **Normalization of GUS by luciferase activity from the same cell extract reduces transformation variability**. *Biotechniques* 1994, **17**(1):52-&.

50. Bradford MM: **Rapid and sensitive method for quantitation of microgram quantities of protein utilizing principles of protein-dye binding.** *Analytical Biochemistry* 1976, **72**(1-2):248-254.

51. Todd JJ, Vodkin LO: **Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family**. *Plant Cell* 1996, **8**(4):687-699.

52. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature Methods* 2012, **9**(4):357-U354.

53. Libault M, Thibivilliers S, Bilgin DD, Radwan O, Benitez M, Clough SJ, Stacey G: **Identification of four soybean reference genes for gene expression normalization**. *The Plant Genome* 2008, **1**(1):44-54.

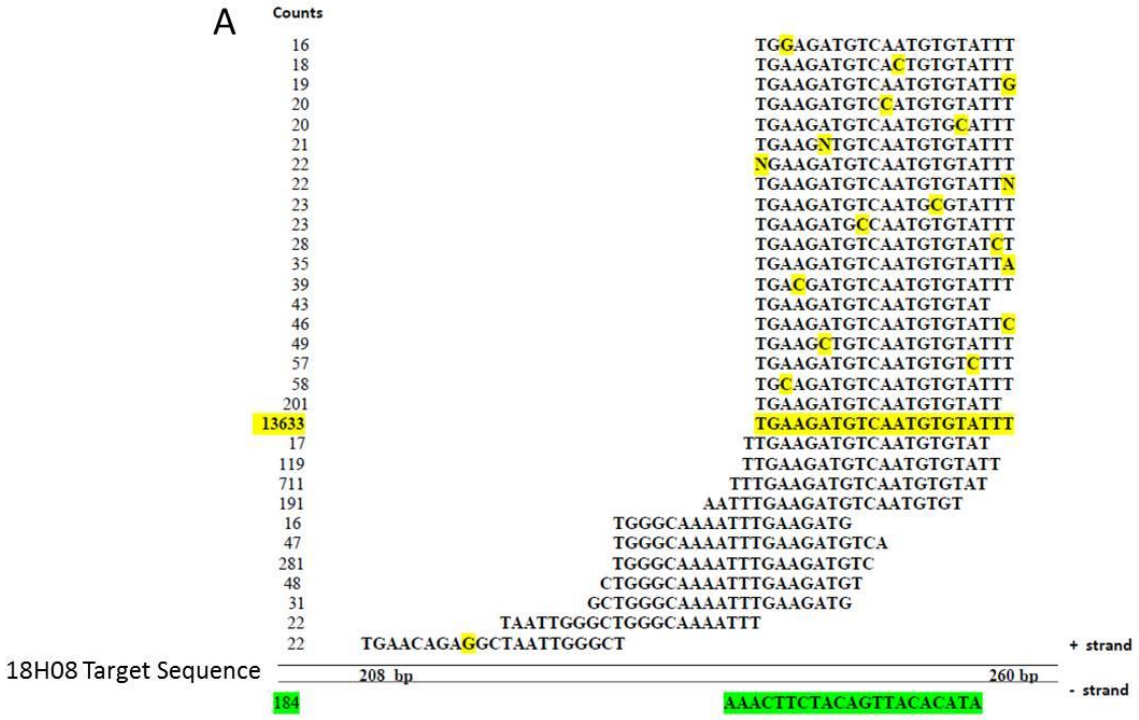
54. Murray MG, Thompson WF: **Rapid isolation of high molecular-weight plant DNA**. *Nucleic Acids Research* 1980, **8**(19):4321-4325.

55. Liu XH, Liu SM, Jamai A, Bendahmane A, Lightfoot DA, Mitchum MG, Meksem K: **Soybean cyst nematode resistance in soybean is independent of the *Rhg4* locus LRR-RLK gene**. *Functional & Integrative Genomics* 2011, **11**(4):539-549.

	Parasatism Gene	Lab	Lines Tested
SCN	16B09	GA	8
	18H08	IL	12
	20E03	OH	6
	23G12	OH	10
	28B03	IL	7
	30C02	GA	10
	30D08	KS	2
	33A09	OH	10
	Total		65
RKN	2E07	GA, IL	25
	7A01	IL	14
	7E12	OH	12
	7H08	OH, GA	24
	5C03B	OH	6
	30H07	GA	12
	16D10	GA	9
	Total		96

Table 5.1. Parasitism genes targeted by hairpin vectors. The parasitism genes are those from previous publications [9, 10]. The lab column indicates which group produced the specific events: GA, University of Georgia; IL, University of Illinois; OH, The Ohio State University; and KS, Kansas State University. None of the events had a significant reduction in nematode counts.

A



B

Target	Positive Strand	Negative Strand	Total
18H08	3975	46	4021
28B03	0	0	0
16B09	6	4	10
30C02	10	1	11
20E03	0.5	0.7	1.2
23G12	0	0	0
33A09	0	0	0

Figure 5.1. One hairpin line produced siRNAs to any appreciable amount. A) For the 18H08 hairpin, a single sRNA species dominates (yellow), but it is to the positive strand and will not bind the target mRNA. A single negative strand siRNA was detected (green), but it has relatively few counts. B) Table of hairpin vectors evaluated by small RNA sequencing.

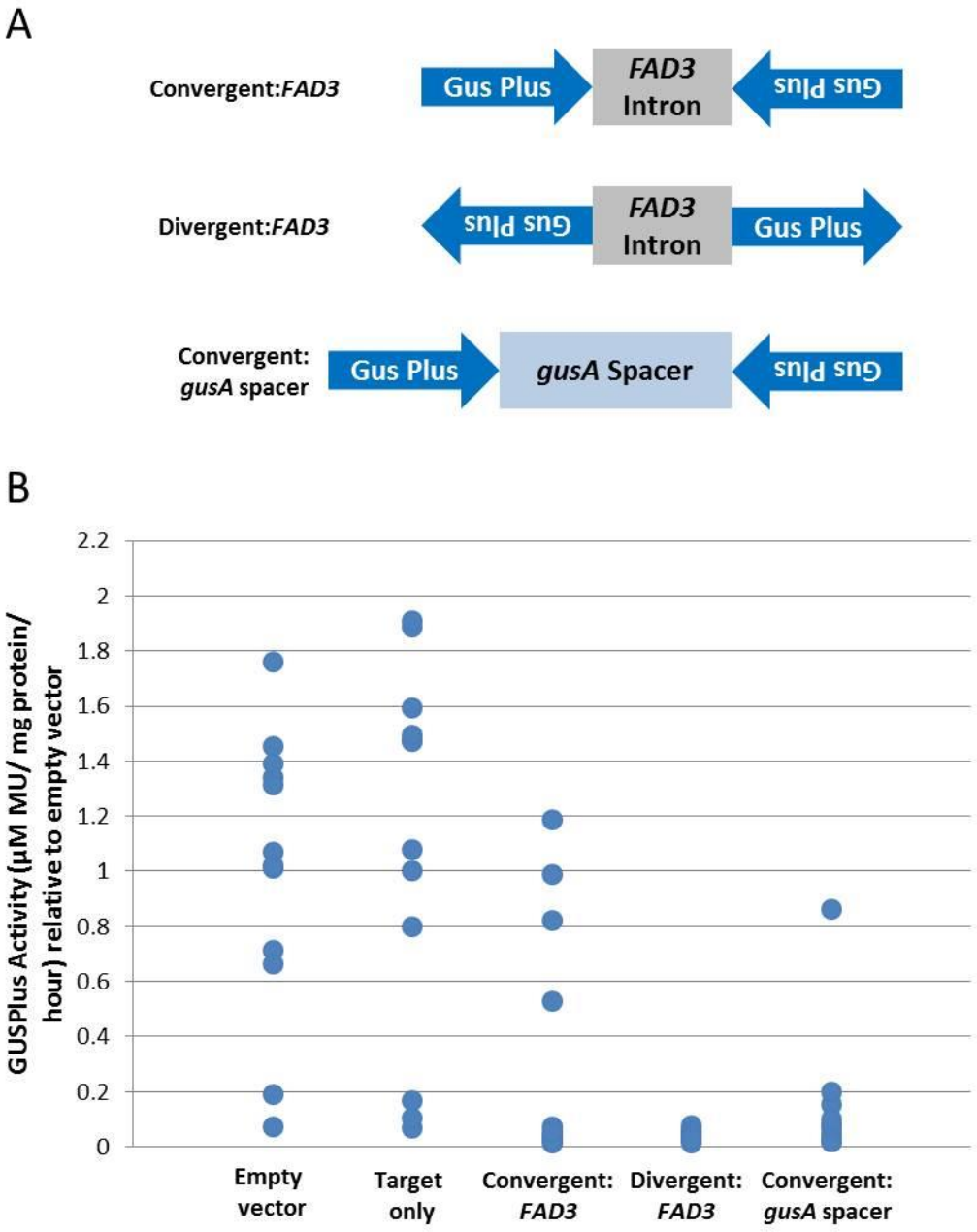


Figure 5.2. Validation of in-house hairpin vectors by silencing GUSPlus. Hairy roots were generated with three hairpin vectors, with different hairpin arms and loops. Convergent hairpin arms have the coding sequence target directed towards the loop, and divergent is the opposite. The loops used were either the soybean *FAD3* intron or a *gusA* spacer. n=12 for each vector.

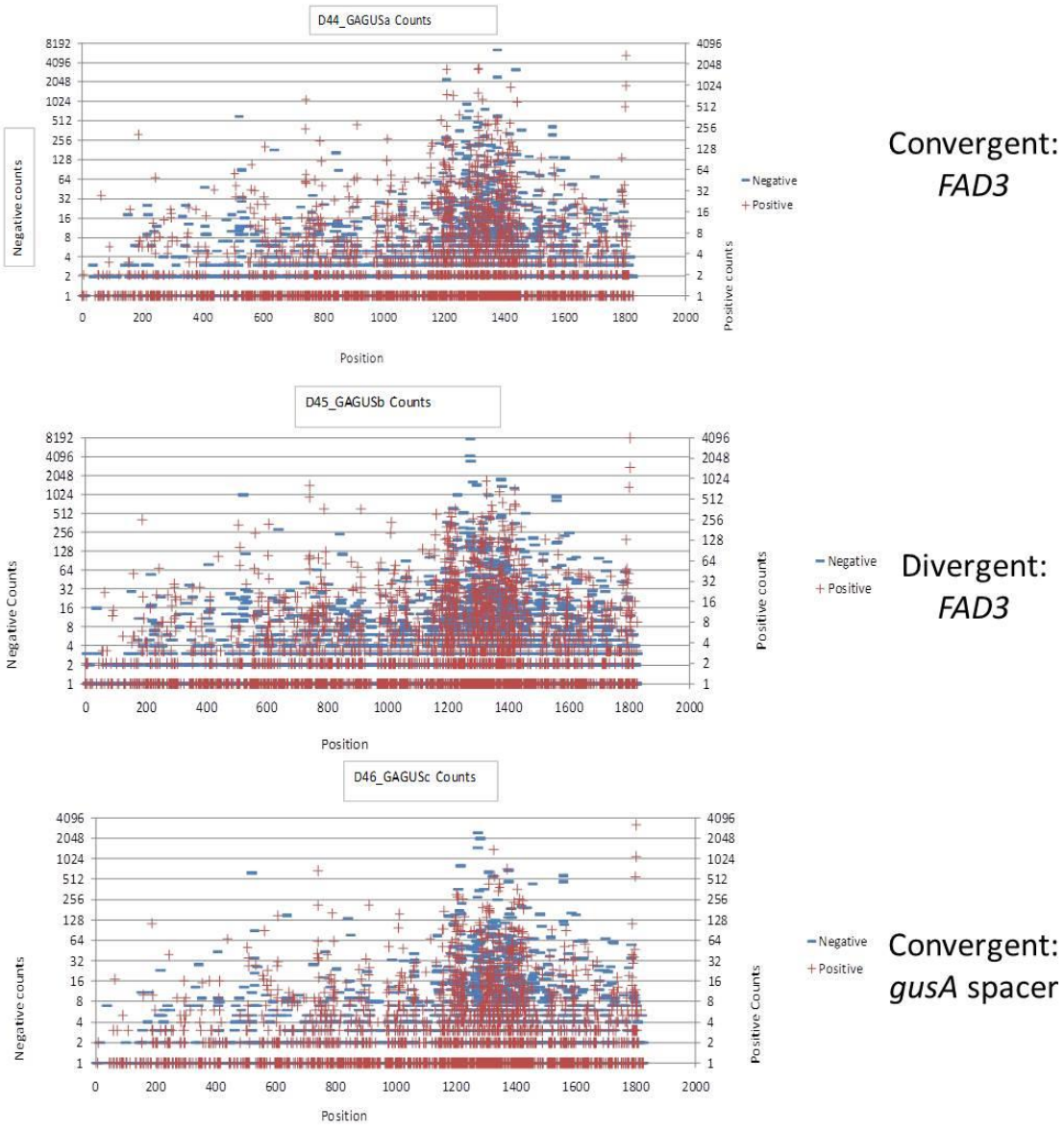


Figure 5.3. Small RNA sequencing of three pooled GUSPlus silenced events per vector. Small RNAs were mapped to the entire GUSPlus coding sequence.

Target	Template	Positive Strand	Negative Strand	Total
33A09	Target	23	5	28
	Full-length	0.8	0	0.8
	Bar fusion	13	0	13
18H08	Target	75	27	102
	Full-length	95	17	112
	Bar fusion	854	370	1224
9H10	Target	41	81	122
	Full-length	152	156	308
	Bar fusion	2287	2082	4369
5C03B	Target	7	6	13
	Full-length	5	5	10
	Bar fusion	8	1	9

Table 5.2. Small RNA reads from the target amplification experiment. Values are reported as normalized small RNA reads (no. of perfectly matching reads per millions of reads sequenced). Each treatment is a pool of three individual events.

Promoter	Loop	Terminator	Positive strand	Negative strand	Total
35S	PDK	OCS	4561	2866	7427
GmUbi	PDK	OCS	65	50	115
35S	FAD3	OCS	791	768	1559
GmUbi	FAD3	OCS	81	82	162
35S	PDK	rbcS	3763	2763	6526
GmUbi	PDK	rbcS	71	43	114
35S	FAD3	rbcS	1968	1775	3743
GmUbi	FAD3	rbcS	912	616	1528

Table 5.3. Small RNA reads from the hairpin component full-factorial experiment. Values are reported as normalized small RNA reads (no. of perfectly matching reads per millions of reads sequenced). Each treatment is a pool of three individual events.

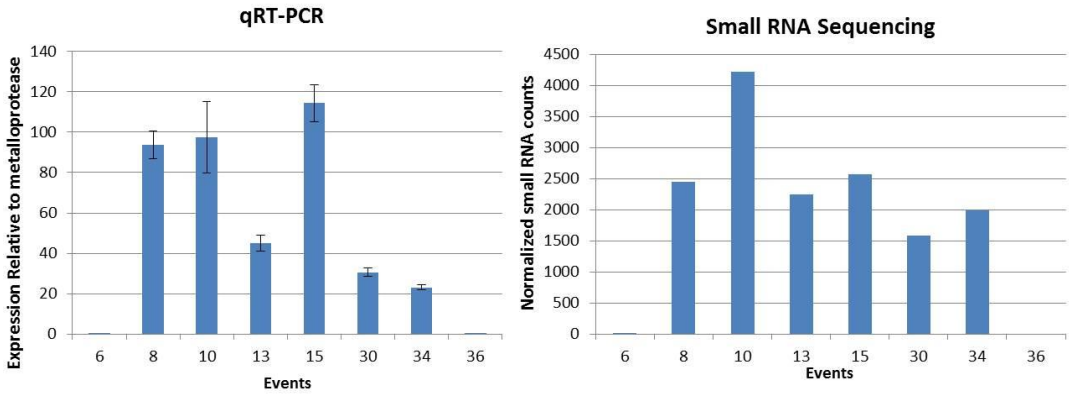


Figure 5.4. qRT-PCR and small RNA sequencing of six 10A06 hairpin lines. Lines 6 and 36 do not express the hairpin construct as determined by qRT-PCR, and do not produce small RNAs. qRT-PCR results are from three technical replicates and error bars are standard deviation.

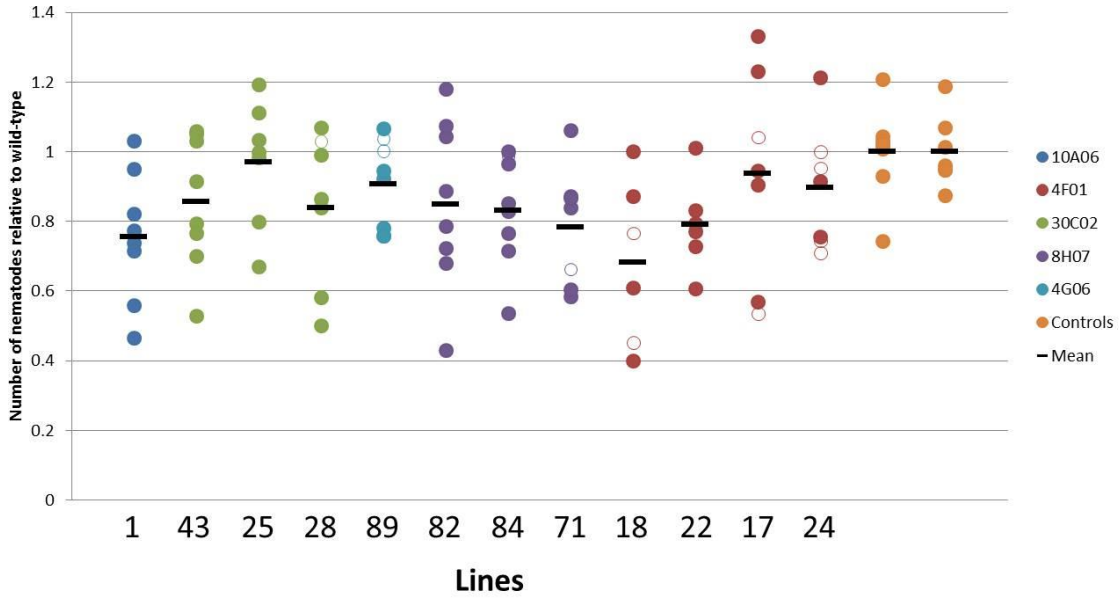


Figure 5.5. Nematode bioassay results. Individual data points represent individual plants. Data is normalized to mean value of respective wild-type controls. Open circle indicate null segregant as determined by PCR.

Event	Line	Plant	Normalized nematode counts	Small RNAs
10A06	1	1	1.03	283
		8	0.56	135
		5	0.46	321
30C02	25	3	1.19	12
		7	0.67	17
	28	2	1.07	8
		7	0.58	7
		6	0.50	10
	43	4	1.06	12174
		2	0.53	14831
8H07	71	7	1.06	9063
		8	0.60	9354
		4	0.58	6957
	82	2	1.18	5610
		5	0.43	6063
	84	4	1.00	3374
		5	0.54	5049
4F01	17	2	1.33	10590
		6	0.57	19282
	18	2	1.00	7336
		7	0.40	9082
	22	1	0.61	35291

Table 5.4. Small RNA sequencing of select plants from nematode bioassay.

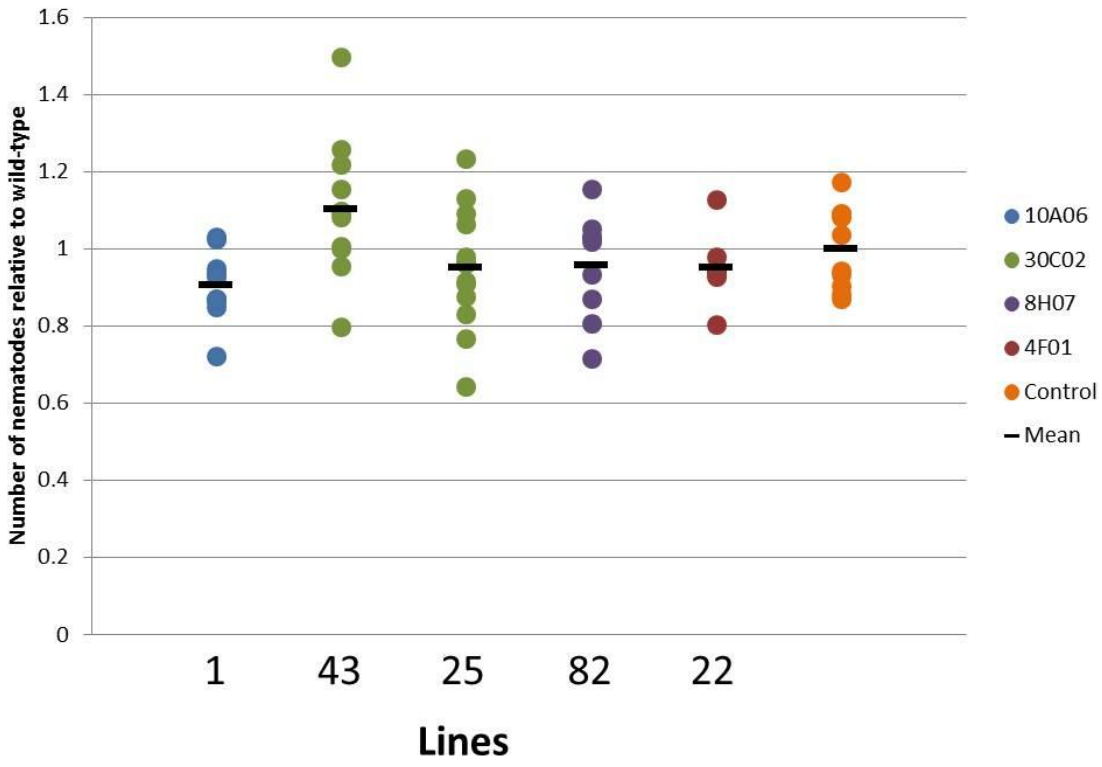


Figure 5.6. Second nematode bioassay results. Individual data points represent individual plants. Data is normalized to mean value of wild-type control.

Primers	
Genotyping	
35SFor	CTATCCTTCGCAAGACCCTT
PDK-L	TCTTCGTCTTACACATCACTTG
PDK-R	ATAACAAAGCGCAAGATCTATC
OCS-R	CAACGTGCACAACAGAATTGA
rbcST-R	CCGCGGATTGATGCATGTTGTCA
FAD3IN_R	GATGTATAAAGAAAGTGTGAGAGTGAGATTACCATT
FAD3IN_F	CTTTATACATCGCACACCAGTGTG
GmUbiF	GCGGGCCCAATATAACAACGAC
RT-PCR	
OCSF2	GACGCCTATGATCGCATGATATTTGCTTTC
OCSR2	CATTAGAATGAACCGAAACCGGCGG
Cloning primers	
gusp-1750R_BS	TTA GGA TCC ATT TAA ATA CTT TGG CCG CTT CGA GA
gusp-1460F_AA	ATA CCT AGG CGC GCC ATC AAG ACG TTC TCC G
gusp-1460F_BS2	TTA GGA TCC ATT TAA ATC AAG ACG TTC TCC G
gusp-1705R_AA	ATA CCT AGG CGC GCC ACT TTG GCC GCT TCG AGA
30C02_A_AA	TAT ATC CTA GGC GCG CCT ACT AGT GAT TAA GCA GTG GTA ACA AC
30C02_B_BS	ATA TAG GAT CCA TTT AAA TGA TTA AGC AGT GGT AAC AAC GC
30C02_C_BS	ATA TAG GAT CCA TTT AAA TGA GGA AAC TTC CAT TCT TGC
30C02_D_AA	TAT ATC CTA GGC GCG CCG GTG AAA TGC GTT TTT CCA
BarF	CGA CCG CTA GCC TAG GAT GAG CCC AGA ACG ACG C
BarR_S	CTC GGC CGG CCT CGA GAT TTA AAT CTC GGT GAC GGG CAG
18H08_TA_F	CGA CCG CTA GCC TAG GGC GCC TAC AAT TTG TCA CTT C
18H08_FLA_F	CGA CCG CTA GCC TAG GAT GGC ATT TCT CCT GTT GTC A
18H08_TA_R	CTC GGC CGG CCT CGA GTA AAT ACA CAT TGA CAT CTT CAA ATT T
BAR:18H08_F	CCC GTC ACC GAG ATT GCG CCT ACA ATT TGT CAC TTC
BAR:18H08_R	ACA AAT TGT AGG CGC AAT CTC GGT GAC GGG CA
33A09_TA_R	CTC GGC CGG CCT CGA GTA AAT ACA CAT TGA CAT CTT CAA ATT T
33A09_TA_F	CGA CCG CTA GCC TAG GGT TAA AAA TCA GTG AAG CAA TGT CT

33A09_R2*	CTC GGC CGG CCT CGA GTC AAT CGG A
33A09_FLA_F	CGA CCG CTA GCC TAG GAA ATG GGC GAG TGC TGC
33A09_FLA_R	CTC GGC CGG CCT CGA GTC ATC AAT CAC AAT ATC ATG CAA G
BAR:33A09_F*	CCC GTC ACC GAG ATT GAG TTA AAA AT
BAR:33A09_R*	ACT GAT TTT TAA CTC AAT CTC GGT GAC
9H10 R2	CTC GGC CGG CCT CGA GCT AAA TTT CCA ACT CTG CGG C
9H10_TA_R	CTC GGC CGG CCT CGA GTA AAT TTC CAA CTC TGC GGC
9H10_TA_F	CGA CCG CTA GCC TAG GCA GAT CAA GGT GAT TCC ATA GTC
9H10_FLA_F	CGA CCG CTA GCC TAG GAT GTC AAA CAA TTT TAA AAC TTG CC
9H107_FLA_R	CTC GGC CGG CCT CGA GTT AAA CCA TAT TAA TAG CCC TTT TAC A
BAR:9H10 F	CCC GTC ACC GAG ATT GCA GAT CAA GGT GAT TCC ATA GTC
BAR:9H10R	ATC ACC TTG ATC TGC AAT CTC GGT GAC GGG CA
BAR_S:9H10 F	CCG TCA CCG AGA TTT CTG CAG ATC AAG GTG ATT CCA
BaRS:5C03B F	CCG TCA CCG AGA TTT CCT TGG CTT TCG GAA GTC
5C03B_TA_F	CGA CCG CTA GCC TAG GTC CTT GGC TTT CGG AAG TC
5C03B_TA_R	CTC GGC CGG CCT CGA GAA ATC GTT ACA ATT AAA TTA GTC GTT TA
5C03B_FLA_F	CGA CCG CTA GCC TAG GAA TTT ATT CTC CAT GAT GAA ATT AAT

Table 5.5. Primers used in this study.

CHAPTER 6

CONCLUSIONS

This dissertation describes the development and application of three molecular techniques for the study and improvement of soybean, as well as testing the functional application of one, RNAi, for plant-parasitic nematode resistance in soybean. The ta-siRNA pathway in soybean was utilized to simply and efficiently induce gene silencing. Compared to traditional hairpin vectors, it is easier to build gene-silencing vectors, and recover silenced plants with the ta-siRNA system. Gene silencing was observed in the roots and leaves of transgenic plants and is likely occurring in other tissues as well. Based on the ta-siRNAs identified in soybean, a family of gene-silencing vectors suitable for high-throughput cloning applications is available to the scientific community and will be useful for the silencing soybean genes.

The CRISPR/Cas system is a powerful genome-editing tool. The vectors developed are very efficient at inducing biallelic DNA modifications at all targeted loci. The data demonstrated that homoeologous genes can be targeted specifically and at the same time. This is particularly useful for the highly duplicated soybean genome. A novel cloning technique was developed for the construction of new CRISPR/Cas targeting vectors. Using this system, new targets can be made quickly and for little expense. A series of CRISPR/Cas vectors have also been made available to the scientific community and will be useful for the modification of specific DNA sequences in soybean and other plant species.

The selective sequencing of DNA sequences is an efficient use of high-throughput sequencing technologies. A novel sequencing method to enrich for genomic DNA sequences flanking transgene insertion sites was developed for mapping transgenes and a transposon in soybean. The high-throughput nature of this method enables the rapid identification and confirmation of flanking sequences. This technology will be useful for the development of insert-specific transgene markers that can be used for the characterization of transgenic events. Furthermore, transposon insertion sites can now be mapped in a large mutagenic population of soybean.

The promise of obtaining resistance to plant-parasitic nematodes via RNAi has been an area of intense research for nearly a decade; however, the data presented here suggests that resistance is not possible with the current strategy. Transgenic vectors and plants were rigorously evaluated for the production of siRNAs, but consistent nematode resistance was not observed. However, a number of factors were evaluated for the production of siRNAs and will be important considerations for future applications of RNAi. This work is important for understanding the biology of the plant-nematode interaction and for the development of practical nematode control strategies.

This body of work produced a number of functional genomics tools that will be useful for studying gene function in soybean and potentially other species. These molecular tools were made with the hope that they would be used to improve soybean as a crop and be adopted by other researchers to facilitate their own work.