Searching Through the Weeds: Molecular and Bioinformatic Characterization of *Solanum sisymbriifolium*

A Dissertation

Presented in Partial Fulfillment of the Requirements for the

Degree of Doctorate of Philosophy

with a

Major in Plant Sciences

in the

College of Graduate Studies

University of Idaho

by

Alexander Q. Wixom

Major Professor: Allan B. Caplan, Ph.D.

Committee Members: Tanya Miura, Ph.D.; Zonglie Hong, Ph.D.; Joseph C. Kuhl, Ph.D.

Department Administrator: Robert R. Tripepi, Ph.D.

August 2018

Authorization to Submit Dissertation

This dissertation of Alexander Q. Wixom, submitted for the degree of Doctorate of Philosophy with a major in Plant Sciences and titled "Searching Through the Weeds: Molecular and Bioinformatic Characterization of *Solanum sisymbriifolium*," has been reviewed in final form. Permission, as indicated by the signatures and dates given below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: _____Date_____
Allan B. Caplan, Ph.D.

Committee Members: _____Date_____
Tanya Miura, Ph.D.

_____Date_____
Zonglie Hong, Ph.D.

_____Date_____
Joseph C. Kuhl, Ph.D.

Department Administrator: _____Date_____
Robert R. Tripepi, Ph.D.

Abstract

Solanum sisymbriifolium (SSI), also known as "Litchi Tomato" or "Sticky Nightshade," is an undomesticated and poorly researched plant related to potato and tomato. Unlike the latter species, SSI induces eggs of the cyst nematode, *Globodera pallida*, to hatch and migrate into its roots, but then arrests further nematode maturation. In order to provide researchers with a partial blueprint of its genetic make-up so that the mechanism of this response might be identified, we used single molecule real time (SMRT) sequencing to compile a high quality *de novo* transcriptome of 41,189 unigenes drawn from individually sequenced bud, root, stem, and leaf RNA populations. Functional annotation and BUSCO analysis showed that this transcriptome was surprisingly complete, even though it represented genes expressed at a single time point. By sequencing the 4 organ libraries separately, we found we could get a reliable snapshot of transcript distributions in each organ. A divergent site analysis of the merged transcriptome indicated that this species might have undergone a recent genome duplication and re-diploidization. Further analysis indicated that the plant then retained a disproportionate number of genes associated with photosynthesis and amino acid metabolism in comparison to genes with characteristics of R-proteins or involved in secondary metabolism. The former processes may have given SSI a bigger competitive advantage than the latter did. Further, SSI was found to be representative of some undomesticated plant species that appear to be recalcitrant to *Agrobacterium*-mediated infection. Transient expression assays in this weed using β-glucuronidase expression have proven highly variable and often failed. Systematic optimization experiments revealed that reducing wounding when bacteria were applied reduced variability. Transfection levels were also influenced by the number of leaves infected/plant, and by total leaves at the time of infection. This inverse relationship between leaf number and GUS expression resembled age-related resistance, but did not correlate with flowering, the accumulation of salicylic acid during bacterial infection, or with overt signs of senescence.

Acknowledgments

First, I would like to start by thanking my major professor, Allan Caplan. You saw something in me that other graduate schools did not. Thank you for pushing me to become a better researcher, writer, and scientist. Know that your example has been pivotal in my education. You will always have my respect and appreciation for everything that you have done for me over the years. I look forward to using what you taught me wherever I land next.

During my education, I have been blessed to have several vital mentors. Tanya Miura, thank you for giving me the push to switch to a Ph.D. and for allowing me to hone my computational skills and thoughts on a different system. Nate Schiele, thank you for trusting me to create a protocol in a system with which I was unfamiliar. Both of you have increased my confidence in my own abilities no matter where I'm working. Ami Wangeline, thank you for awakening my passion for research. The time in your lab changed my thoughts on scientific inquiry forever for the better.

Lastly, I would like to thank those also on my committee: Joe Kuhl, Zonglie Hong, and Tanya Miura for the comments and contributions that have led to the present document.

Dedication

Over the years, there has been an overwhelming number of people that have been instrumental in keeping me working towards my dreams and the completion of this degree. There is no way I can express my gratitude to every one of you; so please know that even if you are not specifically mentioned, if we ever shared a laugh, a smile, a drink, or a game, I would like to sincerely thank you for all the small things that make life enjoyable. These actions may not seem like much but can turn a long day in the lab (or at the computer) into a fulfilling one regardless of anything else.

Without my family, I would not be where I am today. My parents, Brad and Jana, have always pushed me to be the best that I could be; and my brothers, Andy and Nick, have always given me a mark to strive to be better than. Their love and support have been unwavering and have always been the cornerstone of my ability to push forward. My grandparents, aunts, uncles, and cousins have always been behind me as well pushing me onward. I will do my best to be there for each of you, pushing you all forward as you have done for me.

While many said (probably rightly so) that my working hours were insane. My friends were always there when I needed a break, no matter how long it had been since I had seen them. You all were the bright light at the end of a (sometimes very long) tunnel. Without you all, I probably would have lost my mind. Thank you from the bottom of my heart; Trevor, Keeley, Daniel, Chris, Kelsey, Jess, Amanda, Darcy, Jessie, Devyn, Carrie, Sierra, Andy, Sophia, and many others.

I would like to specially thank the Ultimate teams that accepted me as a coach and friend. Coaching and playing with frisbee with you all enforced me to take breaks when I needed them most but felt I didn't have time, so thank you.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Pre-sequencing

Historically, the pathway for characterizing the genome of a model species required a combination of techniques, many of which involved physical manipulation of the genetic material prior to sequence analysis. One of the earliest methods involved randomly fragmenting a genome into smaller chunks that were then placed into a vector to form circularized pieces of DNA called cosmids. These cosmids are plasmids containing a copy of the lambda phage *cos*-site. The addition of the *cos*-site allows this plasmid to be packaged *in vitro* using phage proteins to produce virus-like paricles that can transfer their DNA into cells where they then replicate like plasmids [Collins and Hohn 1978]. This method is able to transfer large pieces of DNA from a test tube into living cells more efficiently than any other bacterial transformation procedure currently in use [Collins and Hohn 1978]. In order to clone even bigger pieces of a genome, artificial chromosomes were created that have all the parts needed to be maintained and duplicated in yeast, including centromeres and telomeres [Murray and Szostak 1983]. These yeast artificial chromosomes (YACs) can have insert sizes up to 600-750 kilobases [Burke *et al.* 1987]. In a similar fashion, methods for constructing bacterial artificial chromosomes (BACs) have been developed, although they tend to have smaller inserts than YACs, on the order of 150-300 kilobase pairs. Another smaller insert size vector used for genome characterization is a fosmid [Hosoda *et al.* 1990]. Fosmids are similar to cosmids in that they both have *cos*-sites and can contain similar insert sizes. However, instead of replicating with a normal plasmid origin of replication, the fosmid carries sequences from an *E. coli* F factor, which stabilize large plasmids to ensure that they are maintained at 1-2 copies per cell. The sequences cloned in these libraries can be used to create a physical map of a genome with the help of end capture PCR, sequence tagged sites [Silverman *et al.* 1989], hybridization and digestion mapping.

1.2 First Generation of Sequencing

In the beginning of the genomic era, the DNA cloned in cosmids, fosmids, BACs, and YACs could only be sequenced in one of two ways. Gilbert-Maxam sequencing was the first widely used protocol capable of determining DNA content [Maxam and Gilbert 1977]. This method involved using chemical modifications and subsequent partial degradation of the DNA strands followed by gel electrophoresis to "read" the correct pattern of nucleotides in a sequence. Another strategy called "plus and minus" sequencing was developed by Sanger, involving the creation of a strand of DNA using a polymerase and then variably degrading it using an exonuclease [Sanger and Coulson 1975]. The latter approach was greatly improved by sequencing via dideoxy-chain termination, commonly known as Sanger sequencing [Sanger *et al.* 1977], where chain-terminating nucleotide analogues were introduced during the creation of the new strand of DNA removing the need to have exonuclease degradation of the sample. All of these methods run four different reactions of the same sample in parallel in the gel in order to identify the nucleotide at each successive position of a strand relative to its defined end. While both Gilber-Maxam and Sanger methods are able to sequence as many nucleotides as the resolution of the gel allows (an average of ∼1000 base pairs), neither method can be automated, and therefore costs are high and speed of analysis is slow.

1.3 First Multicellular Organism Genome Sequenced and Assembled

The first multicellular organism to have its genome fully sequenced was the nematode *C. elegans* [Consortium *et al.* 1998]. The process of sequencing *C. elegans* was the work of many groups carried out over many years. This genome was first physically mapped using 2527 cosmids but gaps were found that could not be resolved using only this preparation. In order to fill in these gaps, a YAC library was established and 257 YACs were discovered to cover 20% of the genome missed using the cosmids alone. 113 fosmids and 44 PCR

products [Cheng *et al.* 1994] were also used to link sequences unable to be verified otherwise. Sequencing on both strands was carried out using either dye termination, otherwise known as Sanger sequencing, or with dye tagged primer sequencing until each part of the sequence was covered 6 times. This multiplicity of sequencing is known as coverage, therefore *C. elegans* was sequenced to a coverage of 6. This was achieved by "shotgun" sequencing which involved taking each YAC and breaking it into even smaller sections, with ∼900 sections covering 40 kb of a YAC clone. Phred software [Ewing *et al.* 1998] was used to read DNA sequencing trace files to call bases with associated quality values which indicate when manual checking of the sequencing should be performed. To be able to build the full genome from all of the parts, Phrap [Green 1996] was implemented to help the previously manual process of aligning and assembling the now sequenced genome. Phrap uses the full read not just high quality bases, but it also allows for user-supplied input to improve assembly accuracy, especially in regions consisting of repeated bases or nucleotide motifs. Consed [Gordon *et al.* 1998] and GAP [Bonfield *et al.* 1995] were then used to visualize the alignments of the sequenced parts to help the manual analysis of the *in silico* assembly. Consed requires every base with a quality score less than 25% to be assessed by hand. GAP requires all sequence data to be assembled at a consensus level of >75% which requires further sequencing or other confirmation of that sequence to be accepted as a true assembly. These programs also suggest which PCR reactions could be performed in order to validate *in silico* assembly decisions in addition to helping predict digests to further assess the accuracy of Phrap assemblies. PCR reactions continued to be used for a large majority of the initial assembly until it was established that PCR experiment failure was more common than mis-assembly following the implemented protocol. The assemblies that couldn't be verified through PCR were verified though other methods i.e. isolating BACs, YACs, fosmids, or cosmids that spanned the assembled region. The main lesson to take away from this would be that despite manual verification of sequencing, additional checks were needed to be done at every step to ensure the validity of the sequencing.

1.4 New Sequencing Methods

In 2004, the completion of the first sequenced human genome [Consortium *et al.* 2004] prompted the founding of the National Human Genome Research Institute. This institute then began funding research into reducing the cost of sequencing the human genome to less than $1000 US dollars. The availability of this funding sped the development and production of many new sequencing technologies. These new methods became commonly known as next-generation sequencing (NGS). Many different NGS methods that were initially created, but two of the most popular were developed by Illumina and 454 Life Sciences (later acquired by Roche). Both of these NGS technologies, take advantage of "massively multiplicative parallel" sequencing. This is built off of PCR amplification and subsequent repeated reading of the same small piece of DNA. Refer to Figure 1.1A for details on an example of NGS library preparation.

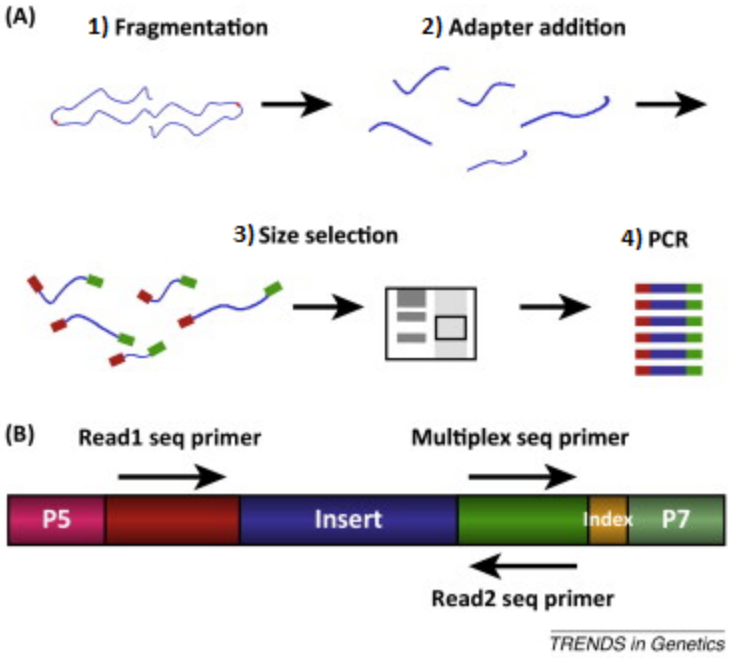


Figure 1.1: An example of a NGS library preparation adapted from van Dijk *et al.* [2014]. A) The library preparation process is very similar for all NGS sequencing. 1) DNA is broken into small fragments. 2) Adapters are ligated to the fragments. These are also include the Read1 and Read2 primers. 3) Similar size fragments are selected for uniformity of sequencing. 4) Non-specific PCR amplification is performed to allow for massive parallel sequencing. B) Different NGS methods will have differing P5 and P7 adapters (named as such only for this example) to allow for the differences in sequencing technologies.

1.4.1 Illumina sequencing

Illumina acquired Solexa in 2007 and the idea that using massive quantities of shorter sequencing reads to assemble full-length DNA sequences has remained the standard for many years (`https://bit.ly/2sIiHgY`). One of the greatest benefits of this approach is the massive coverage that one can obtain of the sequences being investigated. The prepared sequences (Figure 1.1B), which have been amplified by PCR, are annealed to a sequencing plate where each area is a specific sequence, and sequencing-by-synthesis is performed [Balasubramanian *et al.* 2004]. This uses a polymerase to synthesize a secondary strand of DNA and as each fluorescently-labeled nucleotide is added, an image is taken of the fluorescence produced. This allows for an extremely high number of strands to be sequenced at the same time, allowing for a high level of parallel processing and coverage. However, this increase in coverage came at a cost. First, the fracturing of the sequences created much smaller portions of the genetic material in a single read then were previously generated by Sanger sequencing. Second, the method of attaching the fragments to a plate increases the rates of errors near the ends of the fragment. Therefore, any assembly of these reads into full-length sequences would have to have the ability to error-correct for these tendencies. Unfortunately, due to the massive amount of sequencing data produced through these methods, running these corrections on assemblies has become costly, in both time and computational power.

1.4.2 454 sequencing

454 technology is similar in many ways to Illumina, but uses pyrosequencing [Ronaghi *et al.* 1996] segregated by beads [Shendure *et al.* 2005], rather than carried out at separated locations on a plate. Each bead is then placed into a well and one of the four nucleotides is flooded through at a time. This pyrosequencing relies on reading the amplitude of a signal produced in a well when nucleotides are added to the sequence during the flooding [Margulies *et al.* 2005]. This amplitude relates to the number of nucleotides being added

to the sequences at that point of time. Due to many copies of the same sequence on each bead, parallel sequencing is being performed on every sequence. This allows for slightly longer reads than Illumina as well as an ability to distinguish the bases in high GC regions where Illumina technology makes its most errors. In contrast, it is more likely than Illumina protocols to have errors with long mono-nucleotide repeats within a sequence. While, this method also produces fewer total reads than Illumina which gives lower coverage, it also provides less raw data to handle computationally due to fewer multiplicative reads.

1.4.3 SMRT sequencing

Pacific Biosciences developed a sequencing system called single molecule real-time (SMRT) sequencing [Eid *et al.* 2009] that skirts the line between Illumina/454 and Sanger. SMRT doesn't require fracturing the genetic material, thus allowing much longer nucleic acid molecules to be captured and characterized. In addition, this sequencing method can be applied to double-stranded genomic DNA or cDNA. The preparation of the library for SMRT sequencing can be seen in Figure 1.2. Briefly, double-stranded DNA has single-stranded hairpin adapters ligated to each end, creating a circular piece of single-stranded DNA. This is then put into the sequencer where a single piece of DNA will go into each well that has a modified DNA polymerase attached at the bottom. This polymerase will release a flash of light each time a nucleotide is attached to the newly synthesized strand which is visualized by the sequencer. The sequencing will continue around the circularized DNA until the polymerase fails or the sequencing cycle ends. Due to the circularized DNA molecules, SMRT sequencing allows for long read lengths while retaining a higher intra-molecular coverage than Sanger sequencing. This method can produce much longer reads (up to 100,000 bps currently) than Illumina or 454 but at a lower coverage than either. It is also currently more expensive than Illumina or 454 but costs are quickly decreasing due to improved sequencing chemistry allowing for higher throughput in a single sequencing run.

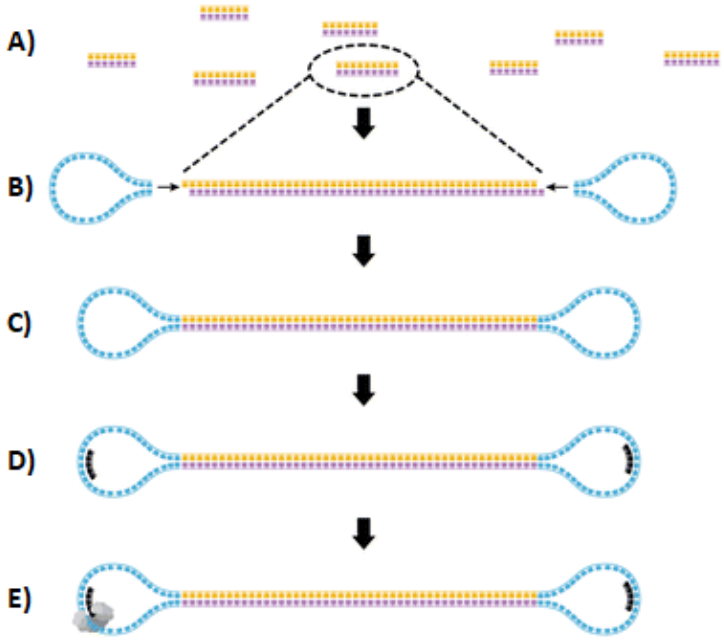


Figure 1.2: An example of a SMRT cDNA library preparation. This was adapted from Kong *et al.* [2017]. A) The library preparation process takes non-fragmented cDNA (yellow/red). B and C) Single-strand hairpin adapters (blue) are ligated to the cDNA to circularize it. D) Sequencing primers (black) are hybridized to the adapters. E) This creates a binding site for the polymerase (grey) to attach.

1.5 How has advances in sequencing changed research?

Next generation sequencing was designed to be performed robotically on hundreds to thousands of sequences in parallel. This has opened the door for unstudied species to be sequenced in a reasonable time and at a reasonable cost [Bentley 2006]. Whereas the yeast and human genome projects involved an international collaboration of hundreds of workers, NGS can sequence a genome with a handful of people. NGS has also made it feasible to sequence a transcriptome alone. A transcriptome contains only the genes being expressed at the time of isolation, in the tissue being isolated. While there are fewer sequences in transcriptomes, some of these sequences may outnumber others by hundreds of times. Traditional cloning-and-sequencing strategies might miss the rarer molecules, but because all current NGS platforms provide high coverage, NGS can, in principle, complete a transcriptome by capturing even low-copy cDNAs.

1.5.1 Genomes

The sequencing of a genome allows for the capture of all potential genes that make up an organism, and can be assembled from samples of any part of an organism at every life stage and every point in time. While this holistic approach to sequencing has the potential to reveal nearly everything that an organism can do, it can come at a cost. Firstly, the genome of an organism doesn't just contain genes it also contains noncoding or "junk DNA". This junk DNA has been stated to make up as much as 97% of the human genome [Nowak 1994], while *Rickettsia prowazekii* (the microbe that causes typhus) has just 24% "junk" [Andersson and Andersson 2001]. Oddly, there is no correlation with higher complexity organisms having higher amounts of junk DNA present in the genome [Pagel and Johnstone 1992]. With this junk DNA present, sequencing a multi-cellular complex organism becomes increasingly time-intensive and expensive due to the amount of sequencing needed to fully encapsulate the genome. Additionally, in complex multi-cellular organisms genes tend to be broken into exons and introns [Breathnach *et al.* 1978]. These exons have to be spliced together in the cell to form the complete and correct gene. Analogously, computational software is used to predict the regions of the sequence data that are exons. This prediction can be handled in several ways: identifying splice sites and predicting exon/introns from the boundaries [Mathé *et al.* 2002], searching for similar genes or synteny [Mathé *et al.* 2002], codon usage tendencies calculated from related species which look for individual exons [Solovyev *et al.* 1994], or some combination of all of these methods such as GenomeScan [Yeh *et al.* 2001]. The accuracy of these methods is further reduced by the existence of overlapping genes, frameshifts in sequencing, non-canonical splice sites, and many other potential issues. Finally, in a complex organism the order of the exons in the genome are not always the order of those exons in a gene [Blumenthal 1998]. This mis-assembling of exons into genes via predictive programs or through manual means can further introduce *in silico* artifacts into a biological assessment. To correct these possible errors, additional sequencing must occur

to validate the gene assignments within the genome.

1.5.2 Transcriptomes

The set of genes being expressed in an organism at any point in time is called a transcriptome [Yassour *et al.* 2009; Zhao *et al.* 2011; Grabherr *et al.* 2011]. To sequence the transcriptome, RNA is extracted from the cells instead of DNA. These transcripts contain all of the exons for their gene after the introns have been removed during the transcriptional process. In terms of sequencing, this reduces the complexity of the analysis on almost every front. First, the sheer amount of material needing to be sequenced is vastly lower [Yassour *et al.* 2009]. Second, the transcripts contain the exons in the correct combination and order needed for biological functionality. With these benefits, there is potential loss of informational robustness. This loss occurs due to the fact that the transcriptome changes from one life stage to the next. It will only ever contain the genes needed by the organism at the moment of extraction. Therefore, the completeness of the transcriptome will be correlated with the number of tissues or organs sequenced and with the age and life stage each sample was taken.

1.5.3 Which is better to sequence?

While sequencing either a genome or a transcriptome has its merits, sequencing both together is best. Doing so provides all of the information needed to accurately identify biologically relevant combinations and location of exons in the genome that provide genes of interest found expressed in the transcriptome. It is important to remember that either can be appropriate or inappropriate based on the questions being addressed.

1.6 How can you move from sequencing to biological testing?

After sequencing and annotating the genetic material, validating the annotations is paramount: many annotations today only consist of analogies and similarities, but not certainties of function. There are various methods that have been established to attempt this validation such as gene knock-outs, knock-ins, knock-downs, and knock-ups among others (Figure 1.3). A knock-out is when a gene that is normally expressed within an organism is either blocked from being expressed, or removed from the organism [Hooper *et al.* 1987; Kuehn *et al.* 1987]. A knock-in is where a gene that is normally not in an organism is added to that organisms gene pool that is expressed [Hanks *et al.* 1995]. Knock-downs and knock-ups are where the level of expression of the gene of interest is respectively changed [Bradley 1991]. Of these, knock-outs and knock-ins are the most common first attempts for annotation validation performed in plant species. Several techniques can be used to create knock-outs, but all of these rely on either selecting for homologous recombination events [Capecchi 1980], or some kind of sequence dependant targeting process like CRISPR [Shan *et al.* 2013]. Both protocols have proven to work inefficiently in plants. This makes gene knock-outs a less than ideal method when working with a novel plant species. In contrast, knock-ins generally involve transferring the gene of interest into a well-studied model species where the resulting phenotype can be assessed. However, this assessment relies on the gene of interesting operating without help from other genes only found in the original unstudied species. It is common practice for both knock-outs and knock-ins to be done in concert to assure the best assessment of the function of the gene of interest. When working with an unstudied species and gene of interest, knock-ins using a studied species tend to be the initial step. These knock-ins can be done in two different ways: stable transformation, or transient transformation.

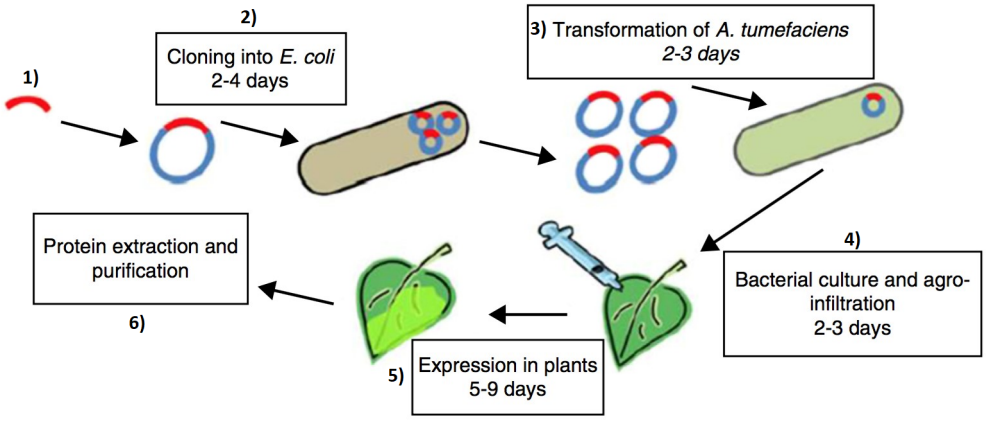


Figure 1.3: An example of a transient transfection protocol. This was adapted from `https://bit.ly/2Jovuzx`. 1) The gene of interest (red) is isolated and prepared for cloning. 2) It is then cloned into a vector, creating a plasmid, that can be maintained in both *E. coli* and *A. tumefaciens*. This plasmid is then transformed into *E. coli* to increase the concentration of the plasmid available. 3) The plasmid is isolated from *E. coli* and is used to transform *A. tumefaciens*. 4) The transformed *A. tumefaciens* is cultured and is pressed into leaf interstitial space using a syringe. 5) Several days of infection with the transformed *A. tumefaciens* allows for the transfer of the gene of interest into the leaf and the expression of the gene. 6) The product of the gene of interest can be assayed pre- or post-extraction from the leaf.

1.6.1 Stable transformation

A stable transformation involves introducing the gene of interest into a (generally well-understood) species and having that gene become integrated into the genome [Hanks *et al.* 1995]. It needs to be integrated so that it can be passed along to the offspring. This also necessitates the need for that gene to be present in the entire organism. One of the easiest ways to ensure this occurs is to perform these transformations on reproductive organs. The passing of the gene of interest to progeny is considered a stable transformation. This gene transfer can be performed in several ways: using viral vectors, bacterial vectors, micro-injection, etc. Many of these techniques hijack nature's already established ability to genetically engineer the world around it. Over the years, particular genes have been removed from these vectors so they don't induce deleterious effects within the infected plants. To ensure the gene of interest has been transferred, a gene providing resistance to a selective

agent is included in the vector and the potentially transformed cells are grown in the presence of that antibiotic [Svab *et al.* 1990]. This will only allow the cells that have been transformed to continue to grow and proliferate.

1.6.2 Transient transformation

A transient transformation of a plant occurs in a very similar manner to a stable transformation, without the need to select for the transformed cells post-infection [Abel and Theologis 1994]. Once the genetic material has been transformed, there is a level of gene expression that occurs without the gene being integrated into the genome [Lee *et al.* 1989]. This allows for a localized expression of the gene of interest that will fade within 2 weeks post-infection [Weld *et al.* 2001]. This has the benefit of a much shorter time to assay the potential effect of the gene of interest within the known species. By not selecting for the cells that are transformed, there is no pressure for the cells to retain the gene that was introduced. Additionally, if the transformation wasn't performed on reproductive tissues, there is a negligible chance that the gene will be passed to the offspring. While the gene is being produced in the transformed tissue, assays can be performed allowing for the confirmation of the predicted annotation of the gene of interest.

1.7 Why does this matter?

With the combination of sequencing becoming cheaper and more reliable, the ability of a small lab to research a non-model organism is becoming more feasible. Our lab has taken steps to sequence and establish a base of knowledge for understanding one of these underutilized species, *Solanum sisymbriifolium*.

1.8 *Solanum sisymbriifolium* has the potential to be a new model organism for plant protective responses.

Solanum sisymbriifolium (SSI) is an undomesticated relative of potato and tomato and is part of the "spiny solanums" [Levin *et al.* 2006] (Figure 1.4). SSI is rightly considered a spiny solanum due to all aerial portions of the plant being covered in spines. This subgenus of *Solanum* has gone largely unstudied due to lack of available genetic resources [Yang *et al.* 2014]. A low level of non-genetic research has occurred for SSI. Specifically, SSI has been utilized as a trap-crop for plant parasitic nematodes for the past decade [Timmermans 2005; Dandurand and Knudsen 2016]. A trap-crop is a plant that induces the hatching and infection of a pathogen (in this case the nematode *Globodera pallida*) but restricts reproductive capabilities. Additionally, SSI has been investigated as a potentially useful source of anti-protozoan [Meyre-Silva *et al.* 2013] and anti-molluscan [Bagalwa *et al.* 2010] metabolites. These studies indicate that SSI could house a genetic wealth of defensive pathways towards pathogens. Chapter 3 of this body of work will offer a bioinformatic analysis of the transcriptome of hydroponically grown SSI. Chapter 4 details the first attempt to establish transient expression assays for SSI and to characterize its response to *Agrobacterium tumefaciens* and *Pseudomonas syringae*. The final chapter will present a very brief perspective on the planned uses of these tools to identify some of the components of the anti-nematodes defenses of SSI.

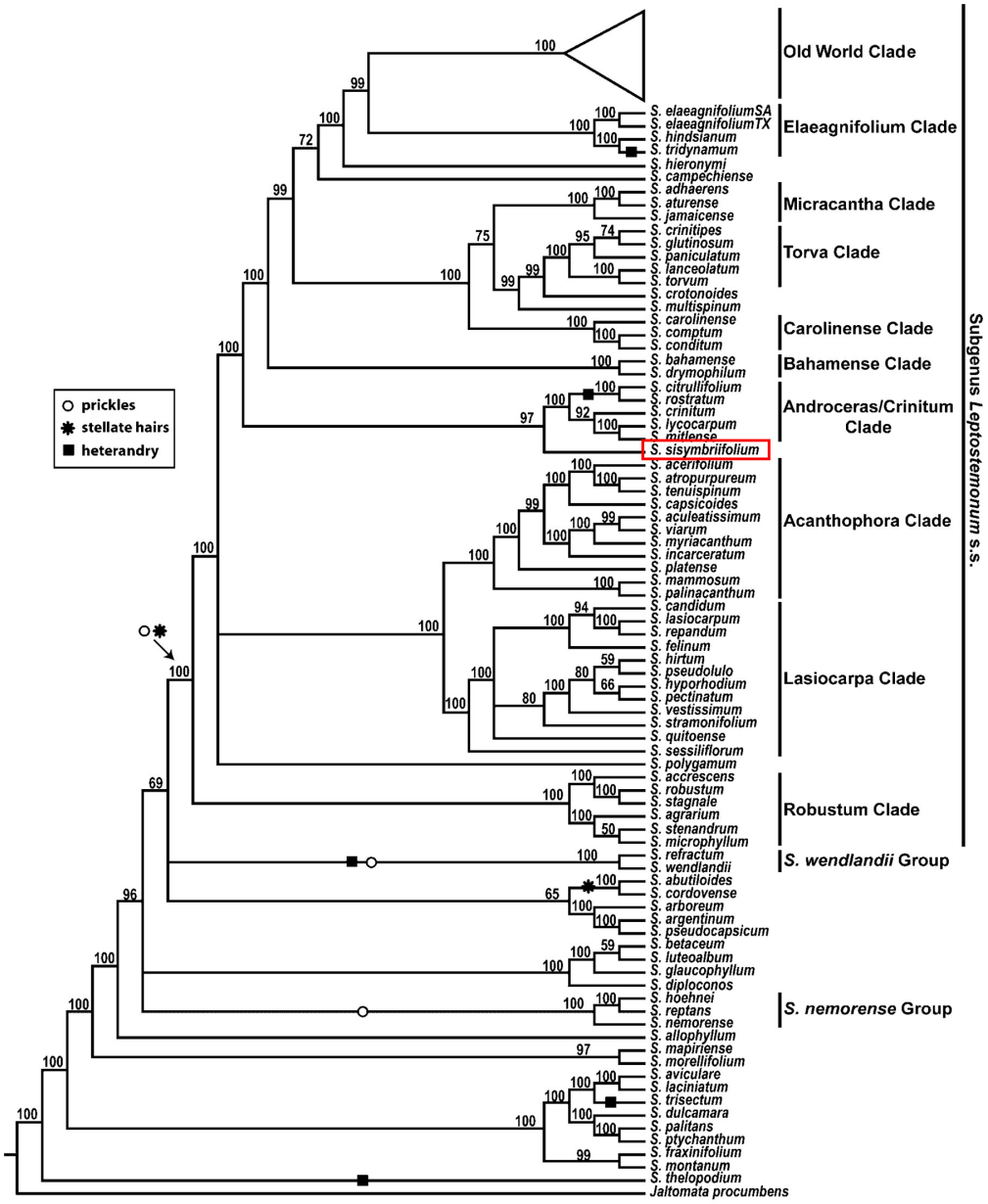


Figure 1.4: A phylogenetic assessment of *Solanums*. *Solanum sisymbriifolium* (red box) is indicated to be part of the subgenus *Leptostemonum* but has not been assigned to a clade. This has been adapted from Levin *et al.* [2006].

CHAPTER 2

Materials and Methods

2.1 Plant and culture conditions

S. sisymbriifolium (SSI) seeds obtained from C. Brown (USDA-ARS, Prosser WA) were germinated in soil. Nodes from a single plant were sterilized for 20 min using 10% NaClO with 0.05% Tween 20. Plant material was then washed 3x with sterile distilled H_2O and put into 120 mL baby food jars containing standard Murishige and Skoog salts, pH5.6, 3% sucrose, 0.7% agar, 100 µg mL^{-1} myo-inositol, 2.0 µg mL^{-1} glycine, 1 µg mL^{-1} thiamine, 0.5 µg mL^{-1} pyridoxine, and 0.5 µg mL^{-1} nicotinic acid. A single plant was chosen as the progenitor of all of the plants used in this study. All of its clones were maintained at 25°C in 16 h light, and subcultured vegetatively every 4 wk. Over the course of the project, rooted clones with at least 4-6 leaves were put into 2 L of hydroponic medium [Shouichi *et al.* 1976], referred to here as Fake Field. Each container was diffusely aerated through an aquarium stone, maintained at constant volume by the addition of distilled water, and emptied and refilled with fresh hydroponic medium every 7 d. Hydroponic containers were maintained at 22°C, 16 h light with an irradiance level of 0.0006 W m^{-2}. Illumination was provided by GE Lighting Fluorescent lamps (13781, F96T12/CW/1500). After a 2 wk lag-time, plants began producing 1-3 new leaves each wk, and flowered continuously afterwards. All experiments were performed on plants that had not been infected or wounded in any way previously.

2.2 RNA extraction

RNA was extracted from SSI bud, stem, leaf, and root and infected root organs adapted from the protocol in Casavant *et al.* [2017]. Adaptions included use of a coffee grinder to homogenize tissue with the addition of dry ice to maintain RNA integrity.

2.3 Genome size estimation by flow cytometry

Healthy green leaf tissues were collected from SSI plantlets growing *in vitro*. Roughly 1 cm^2 (0.01 g or less) of leaf was chopped in 1 mL ice cold LB01 buffer for 1.5-2 min [Doležel *et al.* 1989]. The LB01 buffer contained 50 µg mL^{-1} RNase stock and 50 µg mL^{-1} propidium iodide (25% PI stock in DMSO) per mL of LB01. Each sample was chopped with a fresh razor blade in a clean Pyrex petri plate. The finely chopped suspension was then filtered through a 50 µm nylon mesh filter (Partec 04-0042-2317). This filtered suspension was kept in the dark at 4° for between 15-90 minute before it was analyzed.

Genome size estimations were made using a BD FACSARIA Flow Cytometer (IBEST Imaging Core, University of Idaho, Moscow, ID, USA). A green laser at 488 nm was used to excite the propidium iodide stained cells and was then collected in the PE-A channel. Thresholds for PE-A were set at 1,000 and FSC at 500. The voltages were set so the major peak (2C) of the SSI samples were near 50,000 on the linear scale. Four suspensions were made from separate donor plants once a day for three consecutive days. Two replicates of two external standards were also used daily in addition to the 4 SSI samples. External standards included *Solanum lycopersicum* cv. Stupicke polni tyckove rane (2C = 1.96 pg DNA) and *Glycine max* cv. Polanka (2C = 2.50 pg DNA) [Doležel *et al.* 1992, 1994] which were chosen because their genome sizes were in the expected range of SSI. One repetition of internal standards was run using tomato and soybean. DNA content was estimated using the equation described by Doležel *et al.* [2007].

2.4 Illumina and 454 sequencing and *de novo* transcriptome assembly

Extracted total RNA from each previously stated organ was sent to Eurofins Genomics (Ebersberg, Germany) for library preparation and sequencing. Prior to library preparation, quality control (QC) was performed on individual tubes of RNA and equal aliquots of each

preparation were blended into one pool. The Illumina library cDNA was prepared using randomly-primed first and second strand synthesis, followed by gel sizing and PCR amplification. The library was then physically normalized and found to have insert sizes of 250-450 base pairs (bp).

Two libraries were prepared for 454 sequencing. The 3'-fragment cDNA library was prepared using polydT and synthesis, followed by gel sizing, PCR amplification, library purification and QC. The 5'-fragment library for 454 was prepared by dephosphorylating the mRNA, cleaving the 5'-cap structure and ligating a 5'-synthetic RNA adaptor. This was followed by randomly-primed first strand synthesis, library purification, and QC.

Illumina sequencing was performed on a MiSeq v3 2x300. The read sequences were clipped using Trimmomatic, version 0.32 [Bolger *et al.* 2014], and bases with a Phred score < 20 were removed. Trimmed reads shorter than 150 bp were removed; this step could remove none, one, or both mates of a read-pair. Prior to assembly, digital normalization was applied to the Illumina reads in order to reduce redundant information present in these large datasets. A coverage cutoff of 30 and a kmer size of 20 decreased the data to 28% of the initial 31,310,146 reads. BWA [Li and Durbin 2009] was used to map read pairs to the *Solanum tuberosum* cultivar Desiree chloroplast genome (GenBank accession DQ38616.2). Only unmapped reads were retained for assembly. The 454 3' and 5' fragment cDNA libraries were sequenced in 2x 1/2 segment of a full FLX++ run. These reads were converted to FASTQ format using the software convert project (`http://mira-assembler.sourceforge.net`).

These datasets were assembled using a Eurofins in-house pipeline which included Velvet (v1.2.10) and Oases (v0.2.08) [Zerbino and Birney 2008; Schulz *et al.* 2012]. A multi-kmer approach was applied using assemblies of 59, 69, 79, and 89 kmers that were finally merged into a single assembly using a kmer 29. This assembly was then clustered at an identity threshold of 0.99, corresponding to 1 mismatch in 100 bases, using the software CD-HIT-EST v4.6 [Li and Godzik 2006].

2.5 Library reparation for Iso-Seq

SMRT library preparation and sequencing were performed by the National Center for Genome Resources (Santa Fe, New Mexico). The Iso-Seq libraries for four organs, root, stem, leaf and bud, were prepared for Isoform Sequencing (Iso-Seq) using the Clontech SMARTer PCR cDNA Synthesis Kit and the BluePippin Size Selection System protocol as described by Pacific Biosciences (`https://goo.gl/ij71Hh`) with the following modifications. For cDNA conversion, 3 µg of total RNA was put into each Clonetech SMARTer reaction. From the PCR optimization procedure specified in the protocol, it was determined that 14 cycles of PCR would be sufficient for amplification of each organ's cDNA. Amplification was followed by size selection on each sample to obtain three size bins (0.5-2 kb, 1.5-3 kb and 2.5-6 kb) using the Blue Pippin (Sage Science, Beverly, Massachusetts) instrument. The amplified and size selected cDNA products were made into SMRTbell Template libraries per the Isoform Sequencing protocol referenced above. Libraries were prepared for sequencing by annealing a sequencing primer (component of the SMRTbell Template Prep Kit 1.0) and then binding polymerase to this primer-annealed template. The polymerase-bound template was bound to MagBeads (P/N 100-125-900) (`https://goo.gl/wdZErU`) and sequencing was performed on a PacBio RS II instrument. 12 v3 SMRTcells were run for the root tissues, 14 for the leaf tissues, 9 for the stem tissues, and 12 for the bud tissues for a total of 47 SMRTcells (Pacific Biosciences, P/N 100-171-800). The libraries from each organ were separately sequenced using P6C4 polymerase and chemistry and 240-minute movie times (Pacific Biosciences, P/N 100-372-700, P/N 100-356-200).

2.6 Single Molecule Real Time (SMRT) sequencing

All SMRTcells for a given organ were run through the Iso-Seq pipeline included in the SMRT Analysis software package. First, reads of insert (ROIs, previously known as circular consensus sequences, CCS) were generated using the minimum filtering requirement of

0 or greater passes of the insert and a minimum read quality of 75. This allowed for the high yields going into subsequent steps, while providing high accuracy consensus sequences where possible. The pipeline then classified the ROI in terms of full-length, nonchimeric and non-full length reads. This was done by identifying the 5' and 3' adapters used in the library preparation as well as the poly(A) tail. Only reads that contained all three in the expected arrangement and did not contain any additional copies of the adapter sequence within the DNA fragment were classified as full-length non-chimeric copies. Finally, all full-length non-chimeric reads were run through the Iterative Clustering for Error correction algorithm then further corrected by the Pacific Biosciences Quiver algorithm (`https://github.com/PacificBiosciences/cDNA\_primer/wiki/Understanding-PacBio-transcriptome-data`). Once the Iso-Seq pipeline result was available for each organ, the results were combined into a single data set and redundant sequences were removed using CD-HIT-EST [Li and Godzik 2006].

2.7 Sequence annotation

Mercator sequence annotation was performed using the TAIR, PPAP, KOG, CDD, IPR, BLAST CUTOFF of 80, and ANNOTATE options [Lohse *et al.* 2014].

2.8 Annotating protein domains of translated sequences

PfamScan [Finn *et al.* 2009] was run on the SSI transcriptomes following protocols set forth by Sarris *et al.* [2016].

2.9 Biological quality check of *in silico* sequences

45 clones from a cDNA library (Express Genomics, Average insert size=1 kb, Vector= pExpress 1) were randomly selected and sequenced via Sanger Dye-Deoxy DNA Sequencing (ABI 3730). These sequences were then aligned to the transcriptomes using Bowtie2 [Lang-

mead and Salzberg 2012] set for local alignment and best hit only. These aligned sequences were then manually compared for possible chimeric features.

2.10 Evolutionary comparison of *S. sisymbriifolium* to 13 other species

The evolutionary clustering and comparison protocols were adapted from those set out in Yang *et al.* [2014]. See Table 3.5 for species used and online download sources.

2.11 Divergent gene analysis to determine ploidy

Phasing of the SMRT transcriptome was completed using the unassembled Illumina sequences adapted from protocols established by Krasileva *et al.* [2013], with the addition of an in-house Python script to quantify single-nucleotide polymorphisms present per sequence (`https://github.com/AlexWixom/Transcriptome_scripts/freePloidy.py`).

2.12 Creation of expression snapshots using only SMRT sequences

In-house Python scripts were used to backtrack final transcriptome sequences to each organ using CD-HIT-EST cluster files (`https://github.com/AlexWixom/Transcriptome_scripts`).

2.13 Expression snapshot validation

Sequence specific oligonucleotides were designed for several genes that were then used to obtain semi-quantitative PCR expression snapshots on the same cDNA used to obtain our SSI transcriptome (referred to as the "Sequenced" sample), as well as on a second cDNA pool prepared from an independent RNA collection (referred to as the "Unsequenced" sample). PCR fragment bands were quantified with a local background subtracted and normalized to

actin (following the procedure established by Casavant *et al.* [2017]). The primers for these genes can be found in Table 4.2 with the proposed gene description.

2.14 Accession numbers

The SMRT sequenced transcriptome has been deposited at DDBJ/EMBL/GenBank under the accession GGFC00000000. The version described in this paper is the first version, GGFC01000000.

2.15 *Agrobaterium* culture

Single colonies of agrobacteria were picked from YEB [Vervliet *et al.* 1975] agar plates containing 50 $\mu g\,mL^{-1}$ kanamycin (Fisher Scientific (Fairlawn, NJ), cat. num. 25389-94-0), 100 $\mu g\,mL^{-1}$ penicillin (Life Technologies (Grand Island, NY), cat. num. 860-1830MJ) and inoculated into 5 mL YEB with kanamycin and penicillin to maintain selection on the vector and the nononcogenic Ti-plasmid, respectively. After shaking 18 h at 28°C, the saturated culture was diluted 10-fold and grown for an additional 5-8 h. Bacteria were then pelleted at 3k rpm for 20 min, re-suspended in 10 mL of induction medium (4.9 $mg\,mL^{-1}$ MES (Fisher Scientific (Fairlawn, NJ), cat. num. 145224-94-8), 2.5 $mg\,mL^{-1}$ glucose, 40 $mg\,mL^{-1}$ NH_4Cl, 12 $mg\,mL^{-1}$ $MgSO_4 \cdot 7\,H_2O$, 6.0 $mg\,mL^{-1}$ KCl, 4.0 $mg\,mL^{-1}$ $CaCl_2$, 1.0 $mg\,mL^{-1}$ $FeSO_4 \cdot 7\,H_2O$, and 0.13 $mg\,mL^{-1}$ NaH_2PO_4), and spun 3k rpm for 10 min. This final pellet was resuspended in 20 mL of induction medium supplemented with 130 $\mu g\,mL^{-1}$ acetosyringone, 50 $\mu g\,mL^{-1}$ kanamycin, and 100 $\mu g\,mL^{-1}$ penicillin. Cells were shaken at 28°C for 14 h, pelleted at 3 k, 20°C, and then resuspended to a final A600 of 0.9 in 10 mL MES buffer (4.9 $mg\,mL^{-1}$ MES) containing 40 $\mu g\,mL^{-1}$ acetosyringone.

2.16 Strains and plasmids

All transfections reported here were done with *Agrobacterium tumefasciens* strain GV3101::p*GV2260* containing the reporter plasmid, p*CAMBIA1301* that uses a CaMV 35S promoter to express a *uid*A coding region with catalase intron in the 5' untranslated region (Gene Bank : AF234297.1).

2.17 β-Glucuronidase analysis

β-Glucuronidase assays were carried out on the 4th or 5th leaf from the top at the time of harvest. Leaves were excised from their plants 6 d post infection unless otherwise indicated. In order to facilitate substrate infiltration, leaves were cut into approximately 1 cm diameter discs and submerged in beakers containing a solution of 150 mM K_2HPO_4 (pH 7.2), 5% Triton X-100, 0.25 mM $K_4Fe(CN)_6 \cdot 3\,H_2O$, 0.2 mM $K_3Fe(CN)_6$, and 1 mM X-Gluc (Gold Biotechnology (St. Louis, MO), G1281C2). While a lower concentration of X-Gluc reduced staining, doubling this concentration had no significant effect. Tissues were then submerged in this solution and subjected to a partial vacuum for 3 min. After breaking the vacuum, the beakers were swirled and vacuum-infiltrated once more. Once the vacuum was released, beakers were covered with aluminum foil to shield substrate from light and placed at 37°C for 24 h. Staining for only 4 h showed no indigo precipitant, while 48 h incubations were no better than 24 h ones. Tissues were then destained by replacing the substrate solution with 70% ethanol and incubating these pieces at 37°C for 24 h. The destaining solution was replaced with fresh 70% ethanol, and tissues were incubated once more for 24 h at 37°C. This step was followed by two, 24 h washes with 95% ethanol at 37°C. Additional incubations in 95% ethanol, 22°C were done if needed to clear the leaves.

2.18 Infections

Infections with induced bacteria were done either on clonally-propagated, hydroponically-grown SSI plants with 8-35 leaves (as indicated in the experiment) coming from the primary stem, or on seed-propagated, soil grown *N. benthamiana* with 7-11 leaves. Infections using pressure-infiltration were carried out as described previously [Schöb *et al.* 1997]. Infections with a wire "dog" brush were carried out by sterilizing the tines in 70% ethanol and, after the ethanol had evaporated, pushing them with minimal pressure into the underside of leaves as the leaves were cupped in the palm of a hand. The wounded leaves were then dipped into a solution of induced agrobacteria. Infections with a nylon brush were carried out by sterilizing the fibers in 70% ethanol, and after the ethanol had evaporated, painting induced agrobacteria onto the underside of leaves with a minimum of wounding. In order to assess the effects of exogenous salicylic acid (SA), the underside of leaves were painted with 0.1 mM SA no more than 1 h prior to application of agrobacteria.

2.19 Estimating the relative change in SA after bacterial infections

Endogenous SA levels were measured in cell-free extracts that had been prepared using established protocols [DeFraia *et al.* 2008] with a lux-based biosensor, *Acinetobacter* sp. ADPWH_lux [Huang *et al.* 2006; DeFraia *et al.* 2008]. Extracts were derived from 100 mg of tissue obtained from 3 positions (tip, middle, base) along the axis of the 4th or 5th leaf below the apical meristem of 10-13- ("young") or 26-32-leaf ("old") plants. Agrobacteria, cultured and induced as described above, or *Pseudomonas syringae* pv. Tomato DC3000 [Whalen *et al.* 1991] grown in King's broth [King *et al.* 1954], were painted onto the underside of 7 leaves 1 d prior to harvesting the 4th or 5th leaf. Fluorometric assays on each sample along with SA standards were conducted in 96-well, A/S white, untreated microwell plates (Nunc™) using a FLUOstar OPTIMA microplate reader (BMG Labtech, Ortenburg,

Germany) changing positions along an S-track. Treatments were separated from each other by empty wells to reduce light-contamination. Each value of SA ($\mu g\, g^{-1}$ fresh weight tissue) was an average of 3 samples (tip, middle, base)/leaf, 3 plants/treatment.

2.20 Estimating the relative change in gene expression after bacterial infection

Primer sets shown in supplementary files were used to amplify gene fragments from 3 populations of cDNA, each prepared using RNA from the 4th and 5th leaves from a different pair of plants with the following numbers of leaves: 12 and 30, 13 and 32, 15 and 33. Infections were carried out as indicated with either *Agrobacterium* or *Pseudomonas* for 24 h before harvesting. RNA was extracted, and following a DNase treatment with a TURBO DNA-free™Kit (Invitrogen, Lot 00389753), converted into cDNA using a SuperScript®III First-Strand Synthesis System for RT-PCR (Invitrogen, Lot 1777568) as described in [Casavant *et al.* 2017]. Approximately 1 µg cDNA was then amplified for 25 cycles with primers for the indicated genes using a SimpliAmp™Thermal Cycler (Applied Biosystems, SN 228006035) with a Phusion HF polymerase (New England BioLabs, Lot 0051509). The PCR reaction was performed following the Phusion protocol at 62°C (in accordance with primer design) with a 98°C melting and a 72°C elongation temperature. One half of each sample was run on a 1.5% agarose gel, and the resulting bands were quantified on an AlphaImager HP (ProteinSimple) and its corresponding software. The local background (consisting of the 10 lowest pixel values within the rectangle that surrounded the band) was subtracted from each. The final amounts of each PCR product were then normalized to the estimated amounts of actin, since levels of actin varied less from sample to sample than the other two potential standards included in each analysis, *ssEF1-alpha* and *ssSAM* (data not shown). In Table 4.3, "Young" is the average value of the samples taken from the plants with 12, 13, and 15 leaves mentioned above, while "Old" is the average of the samples taken from plants with 30, 32,

and 33 leaves.

2.21 Estimating bacterial titer on leaves

Agrobacteria containing p*CAMBIA1301* were brushed onto multiple 9- and 33-leaf plants. Within 1 h (T=0), and after 6 d, two 1 cm diameter discs of leaf were harvested from the 4th and the 5th leaves from the apical meristem of each plant. Different plants were used for the 0 and 6 d samples to avoid wound-induced effects on *in planta* bacterial survival. Each disc was homogenized in 0.5 mL M9 buffer [Sambrook *et al.* 1989] using a mortar and pestle and serially diluted from 10 and 10,000 fold. 0.1 mL of each dilution was spread on YEB agar containing 50 µg kanamycin. After 2-3 days of growth in a 30°C incubator, a photo was taken of each plate and colonies were counted on the photo using the Nikon NIS Elements v4.3 documentation package. Each survival assay was repeated 3 times from independently grown cultures.

2.22 SPAD-502 measurement of leaf chlorophyll

The SPAD-502 meter was calibrated using the internal standard of the instrument. Four readings were then taken on day of inoculation and on the 6th d (prior to harvesting samples for titering bacteria) from each of the 4th and 5th leaves of the plants. The SPAD unit measurements were averaged by the machine.

2.23 Microscopy and image analysis

Photographs were taken with a Leica S6D microscope using a Leica EC3 camera and Leica application suite 3.0.0 software. The background surrounding the leaves in each image was removed using Microsoft PowerPoint. The area stained as a percent of total leaf area was determined using *CompuEye* following program instructions established by Bakr [2005].

2.24 Statistical analyses

SAS 9.4 (SAS Institute Inc., Cary, NC) was used to perform a statistical analysis of our data. In order to determine the significance of the investigated treatments, the data were split into different subsets consisting of multiple leaf infections, single leaf infections, SA treated infections, and untreated control infections. All of the data in the subsets were highly skewed when assessed using the PROC UNIVARIATE procedure. Normalization was achieved by transforming each data point into the natural log of each "count". Each count came from the average area stained per disc based on 4-12 discs. After transformation, the PROC REG (regression) procedure was performed on each data subset. A dummy variable replacement (DVR) was then performed by running PROC GLM (generalized linear model) to compare the regression lines formed by the two variables present in each subset. This DVR incorporated both regressions into one equation so that the difference in the lines and the slopes of the regressions could be calculated and the statistical significance of that difference determined. The probability that the lines could have arisen by chance alone will be represented as P_{lines} and the probability that the slopes could have arisen by chance alone will be represented by P_{slopes}.

CHAPTER 3

Assessment of an organ-specific *de novo* transcriptome of the nematode trap-crop, *Solanum sisymbriifolium*.

"Assessment of an organ-specific *de novo* transcriptome of the nematode trap-crop, *Solanum sisymbriifolium*." *G3: Genes|Genomes|Genetics*, vol. 8, no. 7, 2018, pp. 2135–2143.

3.1 Overview of Bioinformatics on a Novel Species

Solanum sisymbriifolium (SSI), otherwise known as "litchi tomato", "*morelle de Balbis*", or "sticky nightshade", is an undomesticated relative of potato and tomato. For more than a decade, SSI has been investigated as a trap-crop (a plant that attracts nematodes but kills them before they can reproduce) for nematodes such as *Globodera pallida* that normally parasitize potatoes and tomatoes [Timmermans 2005; Dandurand and Knudsen 2016]. It has an effective antibacterial defense against non-oncogenic *Agrobacterium tumefaciens* [Wixom *et al.* 2018], and is also a potential source of anti-protozoan [Meyre-Silva *et al.* 2013] and anti-molluscan [Bagalwa *et al.* 2010] metabolites. If the genetic basis for these protective processes could be identified, it might be possible to transfer these traits, either through cross-breeding or through modern transgenic technologies, from this weed to its domesticated relatives. However, while the genomes of potato and tomato have been studied extensively, spiny solanums, like SSI, have not [Yang *et al.* 2014]. Only 54 SSI nucleotide sequences have been submitted to NCBI as of 2016. This ignorance about the biology and genetics of the spiny solanums could be masking a wealth of genetic resources that could be used to protect agriculturally important crops.

Most bioinformatic analyses of a species begin with the assembly and annotation of a complete genome. Once assembled, these data can be searched for genes encoding a particular protein or RNA sequence. For those working on a species that has not been studied exten-

sively in the past, and which is only being studied now in order to conduct a limited number of experiments, whole genome sequencing can be more expensive and time consuming than can be justified. In these circumstances, alternative methods using sequencing technologies that are generally referred to as next-generation sequencing (NGS), have allowed researchers to by-pass whole genome sequencing in favor of generating a smaller database, one depleted of the silent regions of the genome and of genes that are not contributing to the phenotypes of interest. Most commonly, this is done using Illumina or 454 platforms that generate 10's and 100's of millions of short reads from cDNA copies of all of the mRNAs expressed during a given moment of time. Once obtained, these sequences can then be merged *in silico* into full length protein coding sequences. However, this de novo transcriptome can sometimes prove problematic. Short reads derived from highly conserved coding domains and repetitively organized genes can potentially be aligned and joined into chimeric assemblies that cannot be verified or removed because there is no independently sequenced genome available to serve as an extended template or scaffold to ensure that the merged sequences are indeed co-linear [Yang and Smith 2013]. A recent technical improvement, Pacific Biosciences' single-molecule real-time (SMRT) "sequencing by synthesis" strategy, has become sufficiently accurate and attainably priced to be utilized by small research groups. The benefit of using SMRT sequencing is that it produces vastly longer reads than previous methodologies, although with lower coverage [Eid *et al.* 2009]. The longer reads allow researchers to establish a transcriptome consisting of nearly complete open reading frames free of the kinds of errors possible when sequences must be assembled *in silico* from short reads [Ocwieja *et al.* 2012; Zhang *et al.* 2014].

The specific goal of the current project was to establish a four organ (bud, leaf, stem and root) *de novo* transcriptome of SSI. In doing so, we wanted to ensure that the final sequences were high-quality and consisted of genes that were biologically relevant and not artifacts of some *in silico* process. This transcriptome will provide a reference library to be used in future RNA-seq experiments to identify genes for nematode and other pathogen resistances

in SSI.

3.2 Results

3.2.1 Establishing an Illumina and 454 sequenced transcriptome

Before any sequencing was attempted, the genome size of SSI was estimated using flow cytometry (Figure 3.1). This showed that the genome mass of SSI was approximately 4.73 pg per 2C (1C = amount of DNA in a haploid nucleus), or 4.63 mega-base pairs. By comparison, Arumuganathan and Earle [1991] using the same technology estimated that the tomato genome massed between 1.88 to 2.07 pg per 2C while tetraploid potato massed between 3.31 to 3.86 pg per 2C. Thus, these initial measurements gave SSI a genome size greater than tetraploid potato. Despite their unusual length [Paul and Banerjee 2015], SSI has 24 chromosomes like diploid potatoes and many other Solanaceae [Acosta *et al.* 2012; Yang *et al.* 2014]. Due to the size of this genome, and our interest in generating a database of protein-coding genes, we elected to sequence the SSI transcriptome rather than its genome.

The first attempt at generating an SSI transcriptome was done using a combination of Illumina and Roche 454 technologies. Each was used to compensate for the different types of mistakes inherent in the technology of the other, such as errors in sequence interpretation of homopolymer regions with 454, and in GGC motifs with Illumina [Loman *et al.* 2012; Luo *et al.* 2012]. Plant material was harvested from leaves, roots, stems, and buds of uninfected plants growing in hydroponic solution, and RNA was extracted following protocols established in Casavant *et al.* [2017]. RNA from each organ was converted to cDNA and normalized. The normalized pools were then blended in equal amounts and sequenced in separate batches using Illumina MiSeq 2x300 and Roche 454 platforms. The 454 3' library consisted of 632,108 reads while the 454 5' library provided 326,451 reads. The Illumina library sequenced 31,310,146 reads (15,655,073 read pairs) which was digitally normalized to reduce redundant reads that could hinder efficient assembly. This normalization used a

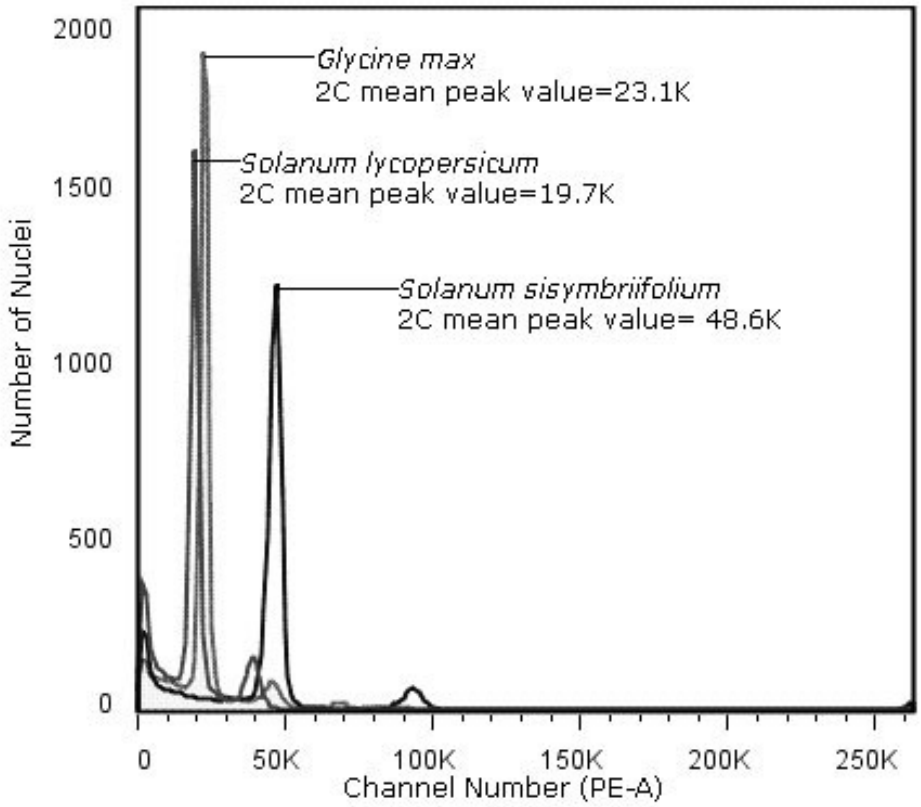


Figure 3.1: An overlay of three nuclear DNA content histograms of species interrogated using a BDFACSAria Flow Cytometer. Major peaks represent the estimated 2C DNA content for three species; *Solanum lycopersicum*, *Glycine max*, and *Solanum sisymbriifolium*. Based on these standards, the SSI genome is estimated to mass 4.73 pg/ 2C value (2,315 Mb/ 1C). The y axis indicates the number of propidium iodide stained nuclei counted out of 10,000 events. The number of stained nuclei per sample varied among species.

coverage cutoff of 30 and a kmer size of 20 to reduce the number of reads to 8,891,416 reads (4,446,708 read pairs). All remaining read pairs were then mapped to the chloroplast genome of *Solanum tuberosum*, and the unmapped reads that remained were used for assembly.

The assembly was performed using a Eurofins in-house script which included Velvet (v1.2.10) and Oases (v0.2.08). Further details of the assembly can be found in Materials and Methods (Section 2.4). This created a set of 351,982 contigs containing 148,324 transcripts with a GC content of 39.1% which was then reduced at a 99% identity using CD-HIT-EST to produce a final set of 102,226 unigenes that collectively had a GC content of 39.1% (Table 3.1). This GC content was slightly lower than expected when compared with other Solanum species, such as potato (40.1%) and tomato (40.7%). This, added to the fact that the number of unigene transcripts exceeded the expected number of genes for plants of this family which has been estimated to range between 35,000 - 40,000 genes [Mueller *et al.* 2009; Gálvez *et al.* 2016] created a need to check the quality of the sequencing and assembly processes.

Table 3.1: Summary of Illumina and 454 assembly.

Illumina/454 Assembly		SSI
Total raw reads		32,268,705
Read lengths		2x300, 2x400
Total raw reads size (bp)		9,466,315,095
GC content		40.95
Contigs	Number	351,982
	Total length	87,840,250
	N50	446
	Max length	14,128
	GC content	38.30
Transcripts	Number	148,324
	Total length	276,561,793
	N50	2,641
	Max length	24,516
	GC content	39.11
Unigenes	Number	102,226
	Total length	207,933,177
	N50	2,773
	Max length	24,516
	GC content	39.12

3.2.2 Bioinformatic assessment of the SSI transcriptome

Our transcriptome represented a "snap-shot" of the metabolic pathways operating in a single healthy plant growing in hydroponic medium. In order to judge whether the data we obtained contained representatives of the sequences that have been found in other species, we performed an annotation of it using Mercator [Lohse *et al.* 2014]. Mercator weights the annotations in its database by the reliability of the source material. One of this program's additional advantages is that it also bins the sequences based on the molecular pathway in which they are most likely to participate (Figure 3.2). In general, the percentage of the SSI sequences binning to each functional class was similar to the percentages binned using the much more intensively studied tomato and potato transcriptomes despite the fact that we had sampled RNA from only 4 organs during a single phase of growth. Not only were

the percentages of "classifiable" genes in each of these species similar, but so too was the percentage of "Not assigned" sequences as seen in Figure 3.3. Finally, we carried out one further commonly employed quality check using PfamScan [Finn *et al.* 2008]. PfamScan is a set of scripts that uses HMMer [Eddy 1998] which is a domain annotation program. This program in combination with protocols established by Sarris *et al.* [2016], allowed for the annotation of domains found in the amino acid sequences translated from the assembled transcriptome. Here, too, all three species showed very similar percentages of sequences without an annotated domain, as can be seen in Table 3.2.

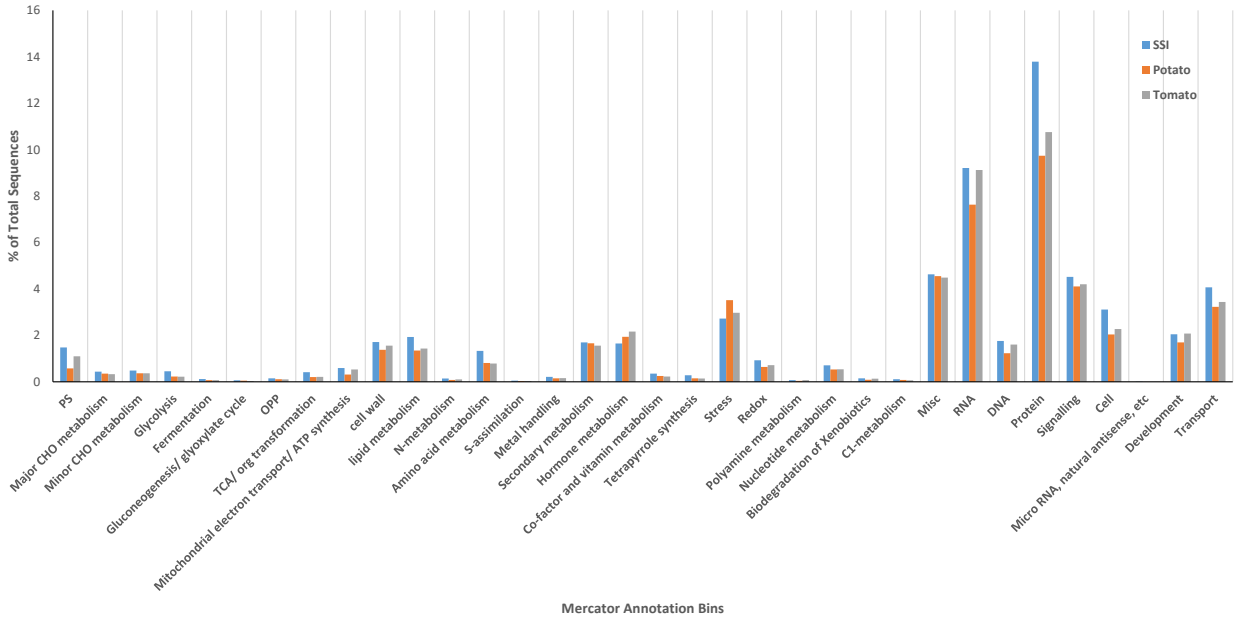


Figure 3.2: Profile of gene ontology (GO) bins extracted from the transcriptomes of three species: *Solanum sisymbriifolium*, SSI; *Solanum tuberosum*, STU; *Solanum lycopersicum*, SLY. The current SSI transcriptome proved comparable to the extensively analyzed SLY and STU genomes. PS, Photosynthesis; CHO, carbohydrates; OPP, oxidative pentose phosphate pathway; TCA, tricarboxylic acid cycle.

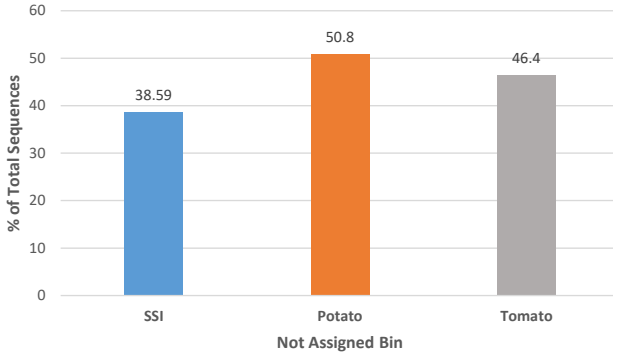


Figure 3.3: Comparison of Mercator "Not assigned" bin of SSI, SLY, and STU transcriptomes. The SSI dataset had fewer "Not assigned" unigenes than found in either potato or tomato genomes, suggesting 10% or more of the SSI genome was not being expressed at the time of sampling.

Table 3.2: HMMer annotation of protein domains recognized in the transcriptomes of tomato (SLY), potato (STU), and SSI translated transcriptomes using PfamScan. The SSI database has a decreased percentage of unannotated sequences compared to the established STU and SLY databases.

	Domain annotated sequences	Total amino acid sequences	Unannotated sequences (%)
SSI (SMRT)	34,912	38,938[a]	10.3
SSI (Illumina/454)	95,022	117,340[b]	19.0
STU	42,785	56,218	23.9
SLY	27,754	34,725	20.1

[a]The decreased number of amino acid sequences compared to the nucleotide sequences is due to lack of open reading frame within the sequence.

[b]The increased number of amino acid sequences based on multiple open reading frames present in a single transcript.

3.2.3 Evidence-based Quality Control of Illumina/454 Transcriptome

The conventional sequence assembly analyses that we performed left nearly 50% of the SSI sequences unannotated (Figure 3.3). While this was not inconsistent with previous studies [Yang *et al.* 2014], it was high, and could obstruct future searches for SSI genes. Thus, in addition to comparing our sequences with those of other species, we performed an internal quality check by sequencing 45 randomly chosen clones from a cDNA library using Sanger Dye Deoxy technology (ABI 3730, Applied Biosystems). Bowtie2 [Langmead and Salzberg 2012] was then used to find the most likely equivalent of each clone in the Illumina/454 assembly. A manual comparison of the Sanger-sequenced clone and the assembled transcript was done using DNA Strider [Marck 1988]. This revealed that out of 45 randomly selected clones, only 35 were found in the Illumina/454 assembled transcripts (Figure 3.4A). Of these 35, only 19 of the clones were identical to at least 1 sequence in the transcriptome database, and more worryingly, 16 appeared to be chimeric (Figure 3.4B). After consideration of alternative explanations, we concluded these chimeras were generated through computational misassembly of short-read data.

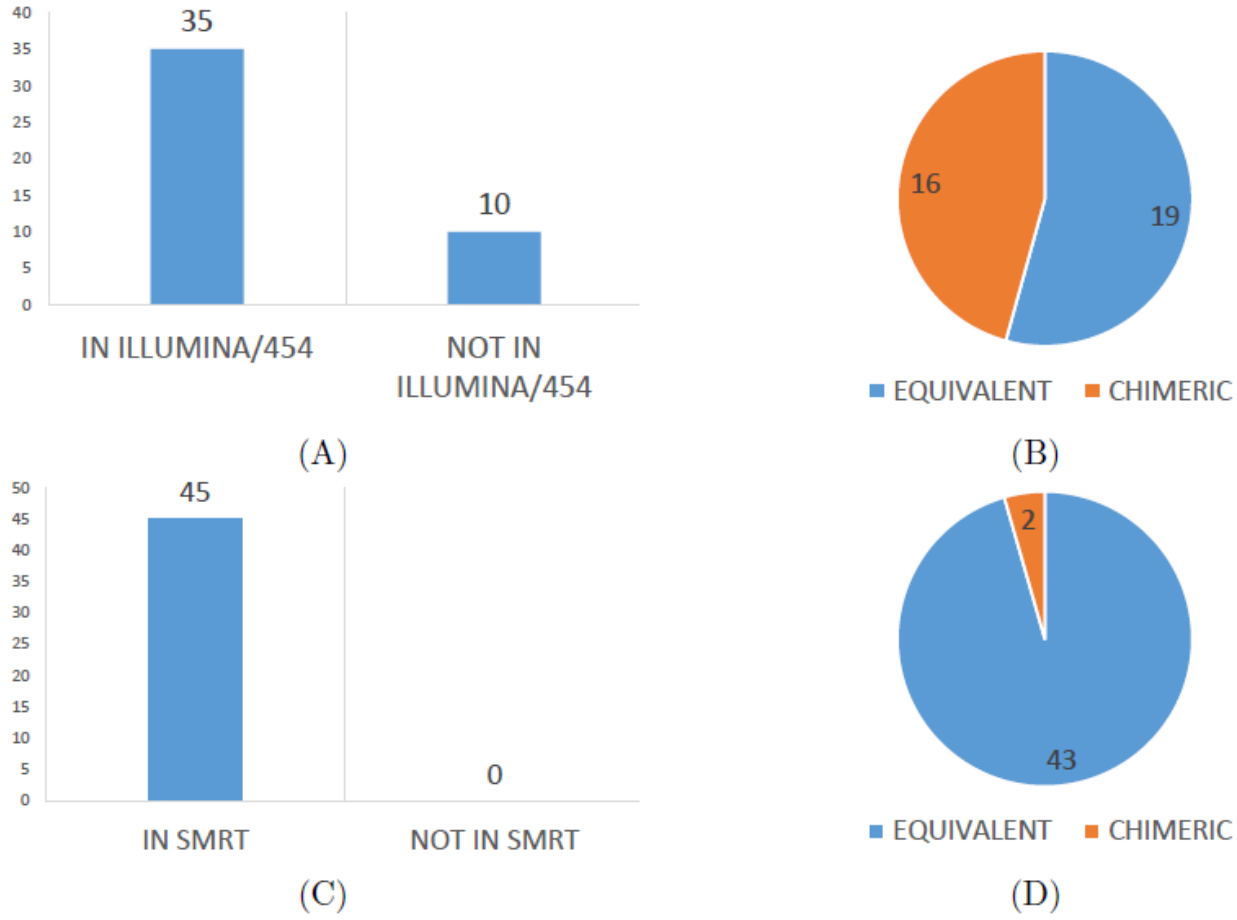


Figure 3.4: Correspondence between 45 randomly selected Sanger-sequenced SSI cDNAs with Illumina/ 454 and SMRT transcriptomes. A) Bowtie2 was used to determine the presence of 45 cDNA clones in the Illumina/454 transcriptome. B) Alignment of matched sequences to Illumina/454 SSI transcriptome was performed using DNA Strider and manual evaluation of alignments yielded equivalent or chimeric assignments. C) Bowtie2 was used to determine the presence of 45 cDNA clones in the SMRT SSI transcriptome. D) Alignment of matched sequences to SMRT SSI transcriptome was performed using DNA Strider and manually evaluated as either equivalent or chimeric. More Sanger sequenced clones were found in the SMRT dataset and a lower percentage were found to be chimeric than in the Illumina/454 assembled dataset.

3.2.4 Establishing a SMRT sequenced transcriptome

The chimericism found in many of the NGS transcriptome sequences would unquestionably hinder future attempts to use this database as a reference library for gene identification. For this reason, we turned to a newer generation of sequencing, called Single-Molecule Real Time (SMRT) sequencing by PacBio Sciences [Eid *et al.* 2009]. Rather than producing only short reads, SMRT can provide reads up to 60,000 bp along a single molecule of DNA. This method can capture entire genes with one read rather than chunking it into many small bits that have to be assembled later. This sequencing strategy gives higher coverage than Sanger-based reactions like those performed on Applied Biosystems™gene analysis instruments, and longer reads than Illumina or Roche 454.

The Iso-Seq pipeline classifies the sequences as either full-length non-chimeric (FLNC), or non-full length reads. Full length reads are those containing both 5' and 3' adapters, in addition to the poly(A) tail. Non-chimeric reads contained each of these parts in the expected order, i.e. 5' adapter–poly(A)–3' adapter, with no additional copies. The non-full length reads were then used to correct the FLNC reads using Iterative Cluster for Error correction and the Pacific Biosciences Quiver algorithm (`https://github.com/PacificBiosciences/SMRT-Analysis/wiki/ConsensusTools-v2.3.0-Documentation`).

In an attempt to improve our ability to detect differences in the suites of genes expressed in different parts of the plant, we generated cDNA from 4 organs; leaves, stems, roots, and unopened flower buds. We then independently carried out SMRT sequencing of all 4 samples. Finally, all corrected FLNC reads were merged in silico and redundancy was removed using CD-HIT-EST [Li and Godzik 2006].

This SMRT sequencing strategy created 231,712 total corrected FLNC sequences (Table 3.3) using the aforementioned pipeline. These sequences had a GC content of 41.2%. CD-HIT-EST was then used to reduce the redundant sequences to sets with 100% identity. This lowered the number of sequences to 139,611 with a GC content of 41.0%. The GC

content continued to decrease further as the identity was reduced using CD-HIT-EST. At 80% identity, there were 32,315 sequences, with an estimated GC content of 39.7%. The decrease in GC content could be due to the methodology of CD-HIT-EST that retains the longest sequence during the reduction process. Because of this, reads that spanned untranslated regions of a transcript were expected to be favored over those only consisting of coding regions.

Table 3.3: Summary of the SSI transcriptome derived using SMRT technology. Organ sub-transcriptomes were sequenced and combined from 33,170 root, 99,924 bud, 50,825 leaf, and 47,793 stem reads. See Figure 3.2 and Figure 3.3 for gene ontology bins of this transcriptome.

SMRT Assembly		SSI
Total raw reads		231,712
Read lengths		300–7883
Total raw reads size (bp)		362,086,346
GC content		41.17
Transcripts	Number	139,611
	Total length	237,865,670
	N50	2,050
	Max length	7,883
	GC content	40.97
Unigenes	Number	41,189
	Total length	74,642,518
	N50	2,158
	Max length	7,883
	GC content	39.90

In the end, we chose to work with a final SMRT dataset that had been reduced to 90% identity and consisted of a set of 41,189 sequences with a GC content of 39.9%. We judged that this estimate of the number of transcripts present in the 4 organs would most likely err on the high side, yet still retain most splice variants within a gene, as well as many paralogs and single nucleotide polymorphisms between alleles of this obligate outbreeder.

3.2.5 Evidence based Quality Control of the SMRT Transcriptome

We performed an internal quality check by sequencing 45 randomly chosen clones from a cDNA library using Sanger Dye Deoxy technology (ABI 3730, Applied Biosystems). Bowtie2 [Langmead and Salzberg 2012] was then used to find the most likely equivalent of each clone in our SSI transcriptome. A manual comparison of the Sanger-sequenced clone and the assembled transcript was done using DNA Strider [Marck 1988]. Firstly, all 45 cDNAs were found in the SSI transcriptome. Secondly, only two of these SMRT-derived sequences appeared to be chimeric, and based on the length of non-homologous stretches, could have been transcribed from different members of the same gene family rather than been created by misassembly. During our analysis of the SSI transcriptome, we did find entries that consisted of inverted repeats of entire gene sequences. These inverted repeats likely occurred during the preparation of the cDNA library prior to sequencing rather than during sequencing or subsequent computational processing as can be found in Illumina or 454 assemblies [Loman *et al.* 2012; Luo *et al.* 2012].

When the SSI transcriptome was analyzed using Mercator [Lohse *et al.* 2014], it contained 38.6% unannotated sequences. This was markedly fewer than the percent unannotated sequences of either potato (50.8%) or tomato (46.4%) transcriptomes processed in the same way via Mercator (Figure 3.3). Other than that, the binned profile of SSI was very similar to the published transcriptomes (Figure 3.2) of these plants. This led us to believe that our transcriptome was at least of comparable quality with the working transcriptomes of these two better studied species.

PfamScan [Finn *et al.* 2008] was also used to annotate the domains of the transcriptome. This program uses HMMer [Eddy 1998] domain annotations, and used in combination with protocols established by Sarris *et al.* [2016], allowed for the annotation of domains found in the amino acid sequences translated from the assembled transcriptome. This annotated 84.7% of the transcriptome with at least one recognizable domain. There were fewer unannotated sequences in the SSI transcriptome than in the STU and SLY transcriptomes (Table 3.2). The reduced number of unannotated sequences found in the SMRT transcriptome

might reflect the fact that this set had undergone a conservative reduction to 90% identity. Alternatively, the reduced number of unannotated SSI sequences could have resulted from the fact that we had only sampled the four most frequently studied organs of a "normally" growing plant, that is, plants manifesting a physiological state which has been extensively studied in numerous species, while the STU and SLY transcriptomes were compiled from plants sampled over a much broader range of life-history stages and growth conditions ranging from fruit and tuber development to exposure to biotic and abiotic stresses where the functions of many genes are still under investigation.

To further test the quality and completeness of our transcriptome, BUSCO benchmarking [Simão *et al.* 2015] was performed using the CyVerse Discovery Environment [Goff *et al.* 2011]. The BUSCO database was established to allow researchers to assess the completeness of newly completed genomes or transcriptomes based on the detection of a set of universal, single-copy orthologs. We found 93% intact BUSCO archetypes (889 genes) in the SSI SMRT transcriptome, 30.2% (289) of these were found in multiple copies, while an additional 2.2% (21) of the BUSCO archetypes were present in fragments, and 4.8% (46) were missing entirely (Table 3.4). These numbers representing genes expressed during a single growth condition of SSI, were only 4.7 percentages different from the numbers of BUSCO archetypes found in the entire SMRT sequenced genome of *A. thaliana*.

Table 3.4: BUSCO assessment for completeness of 3 transcriptomes and one genome. The SSI transcriptome appears to be nearly complete, but contains a disproportion number of duplicated sequences. See Table 3.5 for sources of datasets. ATH, *Arabidopsis thaliana*.

	SSI	Tomato	Potato	ATH (Genome)
Complete BUSCOs	93%	96.2%	86.7%	97.7%
Complete Single-copy BUSCOs	62.8%	94.2%	64.1%	N/A
Complete Duplicated BUSCOs	30.2%	2.0%	22.6%	N/A
Fragmented BUSCOs	2.0%	0.8%	4.7%	0.6%
Missing BUSCOs	4.1%	3.0%	8.6%	1.7%

Although the average SSI chromosome is larger than most other Solanaceae [Paul and Banerjee 2015], independent studies have indicated that it is a normal diploid with n=24

Table 3.5: Datasets for orthologous group assessments in Figure 3.5 were downloaded from online sources.

Species	Download source
Solanum lycopersicum	ftp://ftp.solgenomics.net/tomato_genome/annotation/ITAG2.4_release/
Solanum tuberosum	https://solgenomics.net/organism/Solanum_tuberosum/genome
Arabidopsis thaliana	ftp://ftp.arabidopsis.org/home/tair/Proteins/TAIR10_protein_lists/
Carica papaya	ftp://ftp.plantgdb.org/pub/Genomes/CpGDB/
Vitis vinifera	ftp://ftp.ensemblgenomes.org/pub/plants/release-31/fasta/vitis_vinifera/pep/
Populus trichocarpa	ftp://ftp.plantgdb.org/pub/Genomes/PtGDB/
Prunus persica	ftp://ftp.ensemblgenomes.org/pub/plants/release-31/fasta/prunus_persica/pep/
Citrus sinensis	http://citrus.hzau.edu.cn/orange/download/data.php
Medicago truncatula	ftp://ftp.jcvi.org/pub/data/m_truncatula/Mt4.0/Annotation/Mt4.0v1/
Zea mays	ftp://ftp.ensemblgenomes.org/pub/plants/release-31/fasta/zea_mays/pep/
Oryza sativa	ftp://ftp.plantgdb.org/pub/Genomes/OsGDB/
Solanum melongena	ftp://ftp.solgenomics.net/genomes/Solanum_melongena/

[Acosta *et al.* 2012; Paul and Banerjee 2015]. Chemically-induced tetraploids have been found to have twice this number, while crosses between these lines and presumed diploids produced sterile triploid progeny (Kuhl and Tripepi, data not shown). Despite appearing to be diploid, the BUSCO analysis showed that SSI had more duplicate copy archetypes than diploid and tetraploid potato. This high number of similar sequences could point to the fact that our transcriptome has not been reduced far enough, or could be one line of evidence that SSI has undergone extensive genome duplication or hybridization in the past. This latter hypothesis was evaluated by divergent gene analysis as has been done with plants such as wheat [Krasileva *et al.* 2013]. When the program Freebayes [Garrison and Marth 2012] was run using a defined diploid setting, it output information stating there were genes that had more than 2 alleles or paralogues. We redid the analysis using defined triploid and tetraploid settings and found that even after merging sequences with more than 90% identity using CD-HIT-EST, the SSI transcriptome contained 1,348 genes with 3 distinguishable alleles or paralogues and furthermore, 44 genes with 4 distinguishable copies (Table 3.6). It was noteworthy that no gene had more than 4 alleles or paralogues. A simple explanation for these multiple gene variants, that would be consistent with the BUSCO analysis, was that SSI underwent a genome duplication followed by diploidization in the past and that over time, some of the duplicated loci acquired additional mutations while other loci were lost. To determine if this proposed duplication was restricted to one chromosome, or one chromosomal arm, the 44 genes with 4 alleles were mapped onto SLY chromosomes (Table 3.7). There were 4–allele genes found on 11 of 12 SLY chromosomes which indicated, assuming that genes dispersed in tomato were not linked when the two species diverged, that SSI has undergone a full genome duplication rather than a segmental duplication within one chromosome.

Since SSI is not as well-known as other Solaneacae, we employed OrthoMCL v2.0.9 [Li *et al.* 2003] to illustrate some of the common features its gene complement showed with those of other plants. Protein sequences from our SSI transcriptome (translated using the program

Table 3.6: Divergent gene assessment of allele and/or paralog number in the SSI transcriptome. 4-allele genes were mapped to tomato chromosomes, see Table 3.2.

Allele or Paralog	# of genes
Homozygous	17773
2	22098
3	1358
4	44

Table 3.7: 4-allele genes of SSI have homologs on most SLY chromosomes. Assuming conserved synteny, the locations of these genes could indicate a full genome expansion within SSI.

Tomato chromosome	4 allele SSI genes
1	10
2	4
3	2
4	2
5	5
6	2
7	2
8	0
9	1
10	1
11	5
12	2

ESTScan [Iseli *et al.* 1999]), and protein sequences from tomato (SLY), potato (STU), eggplant (SME), *Arabidopsis thaliana* (ATH), papaya (CPA), grapes (VVI), peaches (PPE), black cottonwood (PTR), oranges (CSI), alfalfa (MTR), maize (ZMA) and rice (OSA) were merged into 45,234 orthologous groups (gene families). In this set, 6097 orthologous groups were shared by all 13 species (Figure 3.5), an overlap well within the range of previous studies [Yang *et al.* 2014]. Each species had many additional groups that were not shown in this diagram because they were not shared with all members of this set of plants. Interestingly, even closely related species like SSI, STU, SLY, and SME had hundreds of groups not found in each other. When the annotations of the SSI unique set were compared to the full transcriptome, several functional groups showed a disproportionate increase. It is possible that these disproportionately expanded sets, that included photosynthetic genes, and genes for amino acid and vitamin metabolism, diverged so much more than groups such as those for cell wall composition, and hormone and secondary metabolism, because expansion of the former traits gave SSI a competitive edge over other species in their habitat (Figure 3.6). Overall, though, there were fewer groups of genes unique to SSI than unique to STU and SME. As noted previously, this could merely reflect the fact that our data came from a single-point snapshot of only 4 organs and so would have lacked those transcripts specifically expressed during fruit and seed set, germination, senescence, abiotic stress, pathogen attacks, and numerous other stages of a plant's lifecycle.

Using highly conserved orthologous genes, i.e. subunits of Rubisco, provisional phylogenies were created for nuclear-encoded and chloroplast-encoded genes using the aforementioned species (data not shown). In doing so, we concluded that nuclear SSI was most closely related to eggplant, which has been noted previously [Särkinen *et al.* 2015], while chloroplast SSI was more closely related to tomato. This dichotomy has also been seen by others [Miz *et al.* 2008] and interpreted to indicate that SSI had undergone an ancient hybridization and afterwards retained the chloroplast genome from one parent, and much of the nuclear genome from another, however, many more SSI genes will have to be compared with the genes of

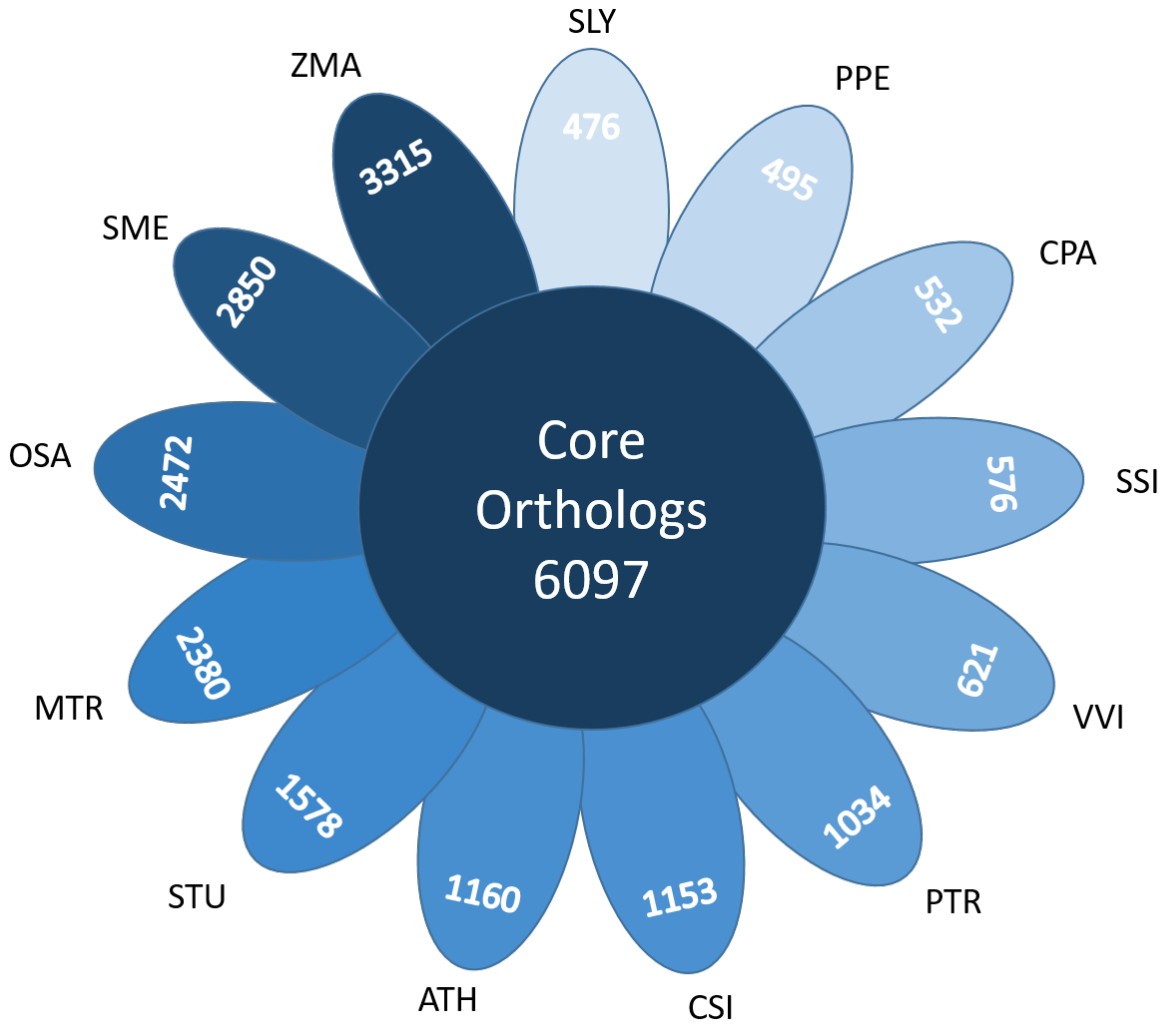


Figure 3.5: Shared and restricted orthologous genes among 13 species. All species shown here shared 6067 core orthologs. Each petal shows the number of gene groups unique to each species. For visualization purposes, each group was differentially shaded according to the number of genes in the set, ranging from SLY (least) to ZMA (most). Not shown are groups shared by only 2–12 species. *Solanum sisymbriifolium*, SSI; *Solanum tuberosum*, STU; *Solanum lycopersicum*, SLY; *Solanum melongena*, SME; *Arabidopsis thaliana*, ATH; *Carica papaya*, CPA; *Vitis vinifera*, VVI; *Prunus persica*, PPE; *Populus trichocarpa*, PTR; *Citrus sinensis*, CSI; *Medicago truncatula*, MTR; *Zea mays*, ZMA; and *Oryza sativa*, OSA. See Figure 3.6 for gene ontology bins for the SSI unique groups, and Table 3.5 for sources of datasets.

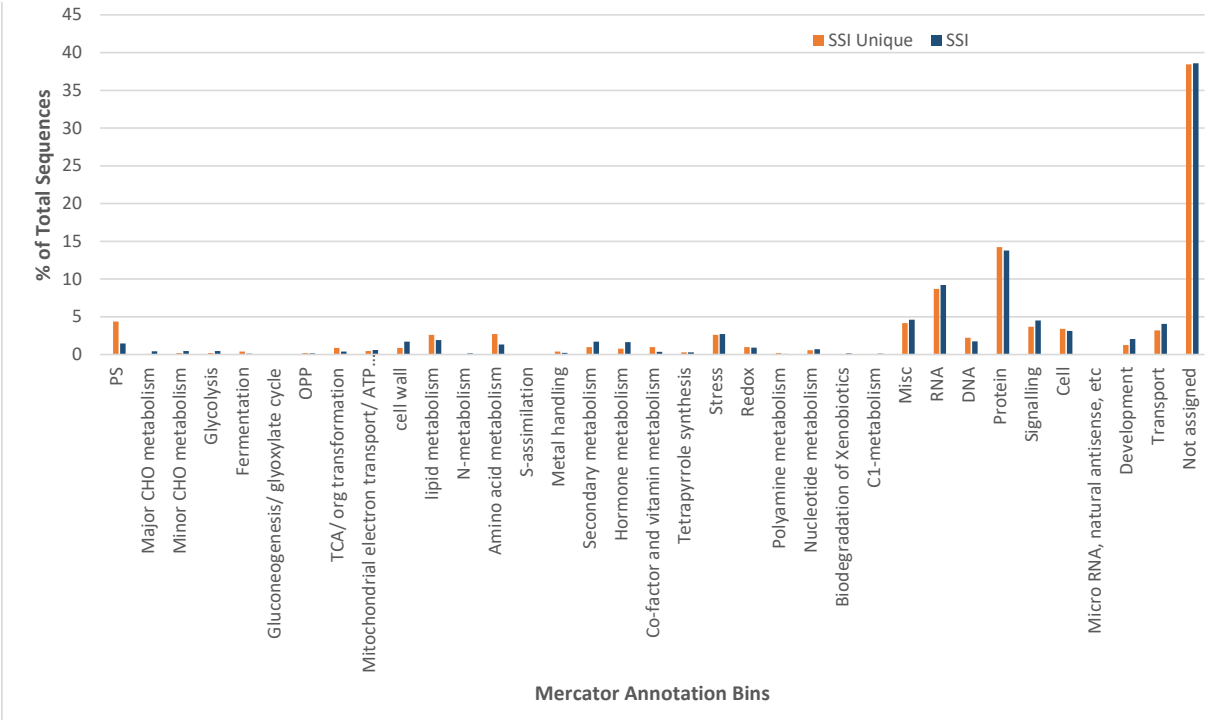


Figure 3.6: Comparison of Mercator bin annotations between the SSI transcriptome and the grouped sequences unique to SSI as found in Figure 3.5. While SSI-unique sequences were found in almost every bin, several were over-represented, including Photosynthesis, lipid synthesis, and amino acid synthesis. PS, Photosynthesis; CHO, carbohydrates; OPP, oxidative pentose phosphate pathway; TCA, tricarboxylic acid cycle.

many more South American *Solanum* species to confirm that this hybridization occurred.

3.2.6 Building a snapshot of organ-associated gene expression

Since we had maintained separate cDNA pools from individual organs, it was possible to backtrack each sequence within the final transcriptome to obtain a provisional profile of gene expression throughout the plant (Figure 3.7A). This analysis showed that there were 8019 sequences expressed solely in buds, 4957 solely in roots, 5349 solely in leaves, 4198 solely in stems, and 7212 sequences expressed in all tissues. That left 11,538 sequences that were expressed in more than one organ but not in all 4.

This backtracking allowed us to construct an expression snapshot that showed how different genes were being expressed at the time the organs were harvested. Using several in-house Python scripts, we recorded the number of reads for genes that had common annotations for several different physiological processes.

A set of light-harvesting complex genes (LHC-I) were predictably found in aerial organs with few exceptions (Figure 3.7B), demonstrating that the backtracking program could extract biologically useful information about sequences with specified characteristics from the merged transcriptome. In order to determine if this kind of analysis of SMRT sequences could categorize the expression of very different sets of genes, we constructed an inter-organ expression profile of genes that encoded both a leucine-rich repeat (LRR) domain and a nucleotide binding (NB-ARC) domain (Figure 3.7C), a pairing frequently found in pathogen resistance genes (R-genes). A profile of R-gene prevalence in SSI, potato, and tomato indicates there is a reduction of these genes in the SMRT transcriptome compared to the other two species (Table 3.8). Three of these potential R-genes were then assayed by semi-quantitative PCR (primers found in Table 4.2) and quantified using a sample of cDNA from the same pool that had been sequenced, and a sample from an independently-prepared, unsequenced cDNA pool (Figure 3.8). In order to assess whether a SMRT data set could be a reliable indicator of gene expression, both the *in silico* and PCR measurements of gene expression were normalized in kind to an actin sequence (Ssi032526). The physical measure-

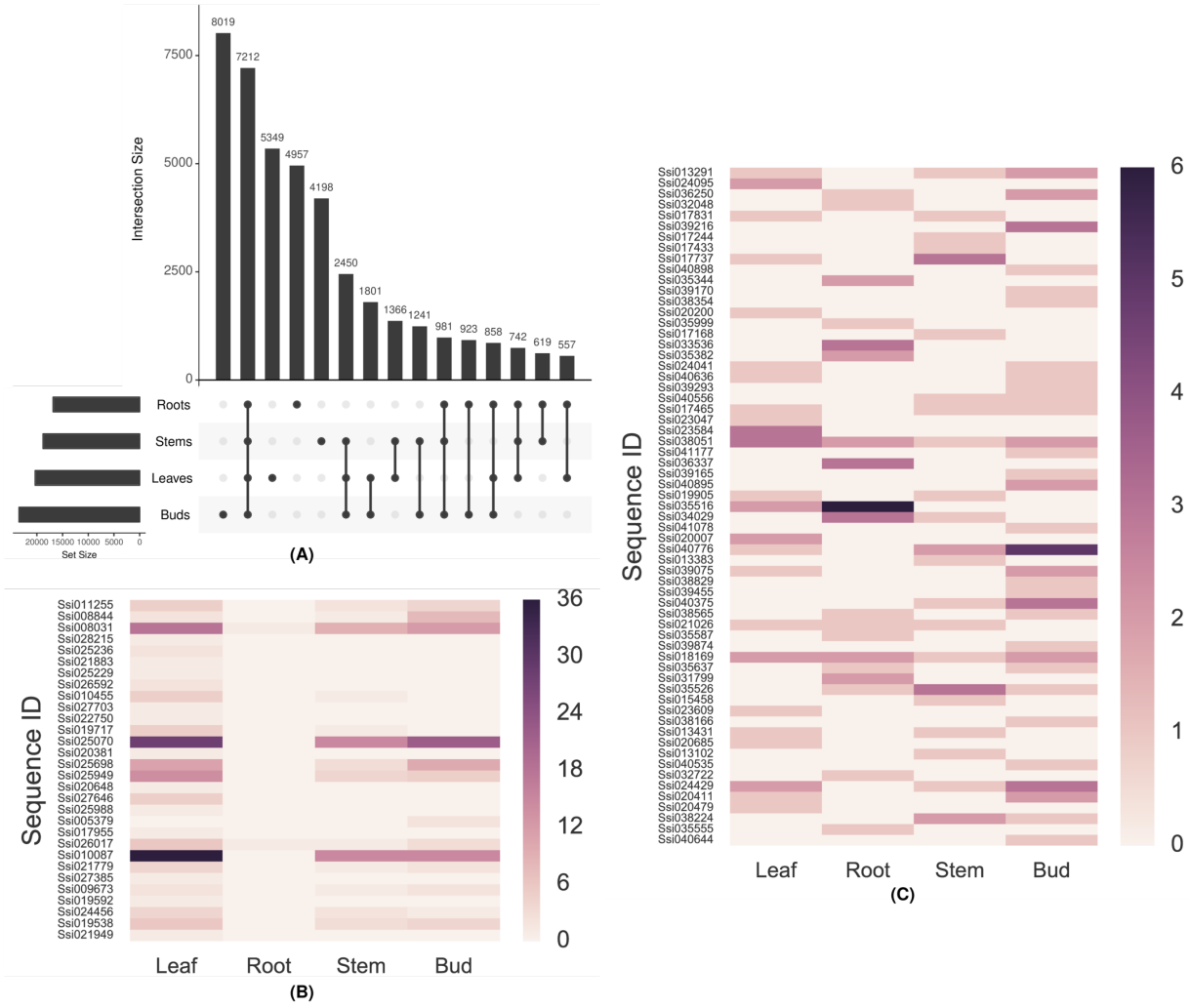


Figure 3.7: Final SMRT transcriptome sequences were backtracked through the de-redundification process to the organ sub-transcriptomes. A) Upset plot [Lex *et al.* 2014] of genes expressed in one or more organs. Each vertical bar shows the number of genes expressed in the organ(s) indicated by the intersection matrix below it (where a single dot in the matrix is a single organ, 2 dots = 2 organs, etc.). The number of genes found to have a homolog in at least one of the organs is indicated by the horizontal bar graph extending to the left. B) Green-tissue specificity of sequences annotated as genes involved in the light harvesting complex-I pathway via Mercator. C) Sequences annotated as putative resistance genes because they contained a nucleotide binding domain (NB-ARC) and leucine rich repeat (LRR) domains show varied expression patterns. As shown on the scales on the right of (B) and (C), the darker the color, the more times the sequence was found in that organ.

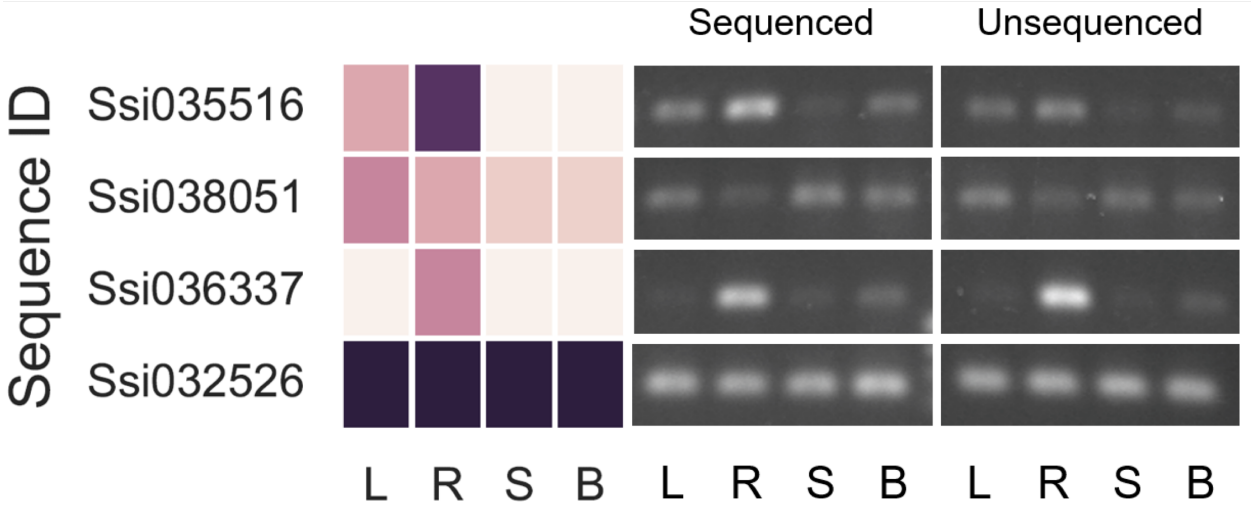
ments of expression of two of the three genes matched the expression snapshot extremely well, but the third gene (Ssi038051) was more abundant in stems and buds than expected based on its SMRT expression snapshot. This confirms that whole transcriptome snapshots can provide a provisional picture of organ differences in gene expression, but further shows that the expression of each gene of interest needs to be verified by independent tests.

Table 3.8: R-gene profile of potato (STU), tomato (SLY), and the SSI transcriptome. The SSI database had fewer assigned R genes (based on the presence of nucleotide-binding domains and leucine-rich repeats within the same open reading frame) than either SLY or STU genomes. Refer to Table 3.2 for full domain annotation statistics.

Transcriptome	R-genes
SSI (SMRT)	67
STU	309
SLY	137

Table 3.9: PCR primers for expression snapshot validation.

SSI sequence	*A. thaliana* ortholog	Forward 5'-3'	Reverse 5'-3'
Ssi036337	RPM1 (at3g07040)	GCAAAGGATGGAGCTGGAGA	CAGCTGAGAACGAACACACACA
Ssi035516	ADR1-L1 (at4g33300)	TCTCCCTGAGGCAATTGGTTGCT	AGTCCTCAGCAACCTGCAGGGAAATAAC
Ssi038051	NB-ARC domain-containing disease resistance protein (at1g50180)	TGCGGCGAAATGAAGTCATCCTAAGTCC	TGGGGTCTGGGGAGGGTAGAG
Ssi032526	Actin7 (at5g09810)	TGCTGGTATGGAGAAGTTTGG	TTCGACAAGGGATGGTGTAAC



(A)

SSI sequence ID	*A. thaliana* homologue	Leaf	Root	Stem	Bud
Ssi035516	ADR1-L1 (at4g33300)	0.64 ± 0.22	0.99 ± 0.13	0.20 ± 0.02	0.39 ± 0.00
Ssi038051	NB-ARC domain-containing disease resistance protein (at1g50180)	0.88 ± 0.40	0.43 ± 0.24	0.69 ± 0.06	0.64 ± 0.13
Ssi036337	RPM1 (at3g07040)	0.22 ± 0.02	1.56 ± 0.56	0.19 ± 0.05	0.36 ± 0.10

(B)

Figure 3.8: Comparison of expression of 3 putative R-gene sequences in the SMRT database to semi-quantitative PCR from 2 cDNA preparations. A) The expression of three genes with LRR and NB-ARC domains characteristic of the R-genes and an actin isoform is shown in the heat map at the left and compared on the right to semi-quantitative PCR of those same genes in two independently prepared cDNA pools, one from the pool used to generate the transcriptome (Sequenced) and one prepared independently, and not used to make the transcriptome (Unsequenced). B) The expression of each PCR product from each pool was quantified and then normalized to the expression of an actin isoform (Ssi032526). Data (biological replicates, n=2) are represented as mean ± STD. See Table 4.2 for primers used.

3.3 Discussion

The creation of a *de novo* transcriptome necessitates massive amounts of follow-up analyses, both *in silico* and biologically, to estimate its reliability.

We initially employed both Illumina and 454 sequencing in order to compensate for the different kinds of errors to which each method was prone [Luo *et al.* 2012]. Screening this assembly with genes randomly selected from an SSI cDNA library revealed that 20% of these genes failed to match any of the assembled sequences in this database, and of those that matched, 40% appeared to be chimeric (Figure 3.4A and B). In contrast, all of these cDNAs were found in our SMRT sequenced transcriptome and few were patently chimeric (Figure 3.4C and D).

A number of factors are known to exasperate misassembly including the presence of large gene families and of repeatedly occurring kmers in the dataset [Moreton *et al.* 2015]. Even though we did not sequence the SSI genome, we found 4 lines of evidence indicating that it might be complex enough to pre-dispose our transcriptome to these kinds of assembly mistakes. First, the nuclear DNA content of SSI was larger than most diploid Solanaceae, roughly the same size as a tetraploid potato (Figure 3.1). Second, divergent gene analysis indicated that the SSI transcriptome was unusually complex and contained 3 and 4 distinguishable alleles for many genes (Table 3.6). Third, there were only 67 putative R-genes, that is, genes containing a nucleotide binding domain (NB-ARC) and a leucine-rich repeat domain (LRR), in the SMRT sequenced dataset compared to the 309 in STU, and 137 in STU (Table 3.8). Finally, an unusually high percentage of the BUSCO gene set were present in multiple copies in SSI even though our transcriptome could only consist of a portion of all the genes that are likely to be encoded in its DNA (Table 3.4). One model consistent with these 4 facts was that SSI had, sometime in the past, undergone a partial or complete genome duplication. Over time, as diploidy was re-established, some of the duplicated alleles or paralogues diverged, while others were lost. Nevertheless, enough of the expanded gene

families remained to confound the alignment programs that tried to differentiate between their members. While these kinds of errors might be correctable with the use of other assembly programs, we chose, instead to create an assembly-independent transcriptome using SMRT technology.

At the moment, SMRT technology does not provide the sequence coverage or depth that can be obtained with Illumina or 454 sequencing. In order to increase our chances of sampling uncommon organ-specific transcripts, we prepared independent cDNA pools from 4 organs of the plant. Using an in-house script (`https://github.com/AlexWixom/Transcriptome_scripts`), we were able to increase the value of the final library by generating expression snapshots for genes of interest in each organ. These expression snapshots are no substitute for a more thorough RNA-seq study, but they do provide a preliminary assessment of a plant's biology at the time of harvest. Using these snapshots, we recognized different patterns of expression of individual LHC-1 genes (Figure 3.7B) within the photosynthetic parts of the plant. We also saw that 2 of the 3 R-genes re-examined by PCR showed the same expression pattern in two independent RNA and cDNA preparations as found in the transcriptome itself (Figure 3.8). Thus, in the absence of RNA-seq studies or experimental evidence for the role of a specific locus, this kind of library assembly could be used to direct researchers to the subset of R-genes most likely responsible for the resistance in a given organ.

R-genes coding for recognition proteins are commonly perceived as sentinels that would be awaiting activation by molecules introduced during infection [Jones and Dangl 2006]. Therefore, the reduced number of R-genes found in SSI (67), compared to both potato (309), or tomato (137) was unexpected based on the seemingly enriched protective responses in SSI. This discrepancy could be explained in any one of several ways arising either from the way our data was collected, or from the biology of this species. First, the seeming paucity of R-genes could simply reflect the fact that we only sequenced RNA from 4 tissues and that this material was harvested only once. However, Yang *et al.* [2014] established two different Solanaceae *de novo* transcriptomes (turkeyberry and eggplant) based on 3 tissues

and obtained much higher R-gene numbers (281 and 172 respectively), close to those we found in the coding DNA sequences of tomato and potato (Table 3.8). A second explanation would be that the sequencing depth in the present study might simply have been inadequate to capture all R-genes that were actually being expressed at low levels. A third possibility would be that SSI could be relying on rapidly inducing transcription of many of its R-genes after an infection has occurred. Finally, SSI might be using proteins with novel domain structures in place of classic R-genes. Any one of these hypotheses is worthy of continuing analysis.

With this transcriptome as an example, we have established a protocol that opens the door to further genetic mining of previously uncharacterized species. The completeness of our database indicates that *de novo* transcriptomes not only provide an economical and time-saving way to study a new species, but can also provide expression data that could not be gleaned from a genomic sequence.

CHAPTER 4

Solanum sisymbriifolium plants become more recalcitrant to *Agrobacterium* transfection as they age.

"*Solanum sisymbriifolium* plants become more recalcitrant to *Agrobacterium* transfection as they age." *Physiological and Molecular Plant Pathology*, vol. 102, 2018, pp. 209–218.

4.1 Overview of Transient Expression.

Agrobacterium tumefaciens is the best known and most widely used representative of a very small set of bacteria that are able to transfer a portion of their genomes into eukaryotic cells [Lacroix and Citovsky 2016]. When *Agrobacterium* successfully infects plants, this transferred DNA (T-DNA) brings in genes that promote affected cells to proliferate extensively, and to redirect their metabolic pathways into synthesizing compounds that can only be catabolized by the tumor-inciting bacteria. Unlike the narrow host ranges of most bacterial pathogens, early studies involving 1193 greenhouse- and garden-grown plants showed that approximately 60% of the tested dicot and gymnosperm species developed tumor-like growths when infected with a naturally occurring *A. tumefaciens* strain called B6 [De Cleene and De Ley 1976]. A subsequent screen, this time assaying for the product of a T-DNA encoded reporter gene rather than for tumor formation, detected the diagnostic enzymatic activity in 50 out of 248 cell cultures representing a spectrum of different plant species [Sindarovska *et al.* 2014]. Even though there is still no definitive explanation for the recalcitrance that the remaining plants showed to infection [Pitzschke 2007], this extraordinarily eclectic capacity to infect has made *Agrobacterium* virtually indispensable for genetically modifying most common dicotyledonous crops, and, if grown in optimized culture conditions, is also capable of transferring T-DNA sequences to monocots [Chan *et al.* 1993], fungi [Bundock *et al.* 1995], and mammalian cells [Kunik *et al.* 2001].

Although most work with *Agrobacterium* has centered on isolating stable transgenic cell

lines and regenerated plants weeks after the initial application of bacteria, the first signs of transgene activity can be detected within days of infection. Without further selection for sustained expression, this persists until the transgenes have either been lost or silenced [Weld *et al.* 2001]. This "transient expression" was first demonstrated by monitoring nopaline synthase activity [Horsch and Klee 1986], but today, virtually all researchers measure gene transfer efficiency using either a fluorescent protein like GFP [Chalfie 1994] or the more sensitive enzymatic assay for β-glucuronidase (GUS) [Jefferson *et al.* 1987; De Ruijter *et al.* 2003]. Frequently, the development of a protocol for transient expression has been the first step towards development of a method to stably introduce genes to assess whether they accentuate [Chang *et al.* 2002] or suppress [Saedler and Baldwin 2004] plant phenotypes [Komarova *et al.* 2011; McQualter *et al.* 2015].

The vast majority of transient expression assays have been done using one species, *Nicotiana benthamiana* [Goodin *et al.* 2008]. Its leaves are thin and easily infiltrated with bacteria as well as with the substrates needed to visualize GUS activity *in situ*. As an added advantage, this relative of tobacco has a defective RNA-dependent RNA polymerase gene that could otherwise silence the transgenes under investigation [Yang *et al.* 2004]. While agronomically relevant species such as potato have also been used for transient expression analyses, the outcomes have proven highly variable and affected by leaf position, plant age, and the particular cultivar being tested [Bhaskar *et al.* 2009]. Because of this, more transient assays of potato and tomato transgenes have been carried out in *N. benthamiana* than in their original hosts [Dan *et al.* 2006; Bhaskar *et al.* 2009].

This paper describes an attempt to apply the lessons learned from transient expression studies in other plants to an undomesticated relative of potato and tomato, *Solanum sisymbriifolium*, also known as "litchi tomato", "*morelle de Balbis*", or "sticky nightshade". This plant has been investigated as a potentially useful source of anti-protozoan [Meyre-Silva *et al.* 2013] and anti-molluscan [Bagalwa *et al.* 2010] metabolites, and as a trap-crop for plant parasitic nematodes [Timmermans 2005; Dandurand and Knudsen 2016]. It has

been viewed as an example of the kind of reservoir of genetic diversity that may need to be tapped in order to provide future plant breeding and genetic engineering programs with traits missing from more familiar species. Our studies have shown that in comparison to plants like *Nicotiana benthamiana* (Fig. 4.2), *Solanum sisymbriifolium* is somewhat recalcitrant to *Agrobacterium*-mediated delivery of DNA by typical methods including wounding [Joos *et al.* 1983], pressure-infusion [Schöb *et al.* 1997], and vacuum infiltration [Kapila *et al.* 1997] applications of *Agrobacterium*. However, we found that minimizing epidermal damage when applying bacteria led to β-glucuronidase expression in about 80-90% of the treated plants. The percentage of leaf surface showing this measure of bacterial infection was negatively correlated with the chronological age of the plant as measured by its number of leaves. This decline in susceptibility to an invading bacterium thus resembled the salicylic acid (SA)-dependent, *npr1*-independent process [Kus *et al.* 2002] called "age-related resistance" (ARR) that *Arabidopsis* plants develop against pathogens [Carella *et al.* 2015]. When ARR begins, however, has not been resolved. In *Arabidopsis*, several studies have shown that ARR is correlated with the onset of flowering, or with stressful conditions that advance the time of flowering [Kus *et al.* 2002; Wilson *et al.* 2013; Carella *et al.* 2015], but other studies have concluded that disease resistance increases after plants produce a critical number of leaves [Wilson *et al.* 2013]. Furthermore, ARR-like responses are triggered in other species by distinctly different developmental cues [Develey-Rivière and Galiana 2007]. For example, tomato develops an ARR response against the fungus *Cladosporium fulvum* as late as the beginning of fruit set [Panter *et al.* 2002]. In the present study, during which the age of *S. sisymbriifolium* plants was measured not from germination, but by the number of leaves present after vegetative propagation, age-related resistance manifested itself just after more than 17 leaves had emerged and expanded, at which time the oldest leaf began showing signs of senescence. It is possible that transgene expression studies in other recalcitrant species could be improved by first determining when ARR begins.

4.2 Results

4.2.1 *S. sisymbriifolium* responds adversely to *Agrobacterium.*

One of the many reasons that *N. benthamiana* has become the standard host for transient expression studies is that a simple handheld syringe can force a solution of agrobacteria throughout a broad area of a leaf's mesophyll [Schöb *et al.* 1997]. However, neither this method nor vacuum infiltration of bacteria into isolated leaf discs produced GUS-expressing *S. sisymbriifolium* (SSI) explants (data not shown). When we compared the infiltration zones of the two plants, we noted that little liquid penetrated into the densely packed cell layers of SSI leaves when compared to the amount that was infused into *N. benthamiana* (data not shown). Moreover, we noticed that SSI leaves had a much more pronounced, potentially anti-bacterial, response to *Agrobacterum* than did leaves of *N. benthamiana*. This was particularly noticeable when bacteria were pressed into the tissues using a wire dog brush (Fig. 4.1A). Figure 4.1C, E, and G show a representative grid of puncture marks created on *N. benthamiana* and SSI with this brush. As expected, tissues around the circumference of each *N. benthamiana* wound showed substantial GUS activity when assayed 6 days after infection with the *Agrobacterium* strain, GV3101 (p*GV2260*; p*CAMBIA1301*) (Fig. 4.1C). On the other hand, the damaged surfaces of an SSI leaf infected with the same strain, in concentrations ranging from an OD_{600} of 0.05 to 0.9, only developed a dark brown pigment, possibly a type of melanin like that produced in wounded chickpea leaves [Yadav *et al.* 2017]. No such pigment formed on *N. benthamiana* leaves (Fig. 4.1C), and very little pigment formed when SSI leaves were wounded, but not infected with bacteria (Fig. 4.2). This pigment production may have only been the most visible component of a multi-factorial defense response that inhibited bacterial infection, or killed affected plant cells before detectable levels of β-glucuronidase could accumulate.

In order to reduce wounding, we explored a less abrasive way to apply the bacteria that involved painting induced bacteria onto the underside of leaves with a soft brush (Fig. 4.1B)

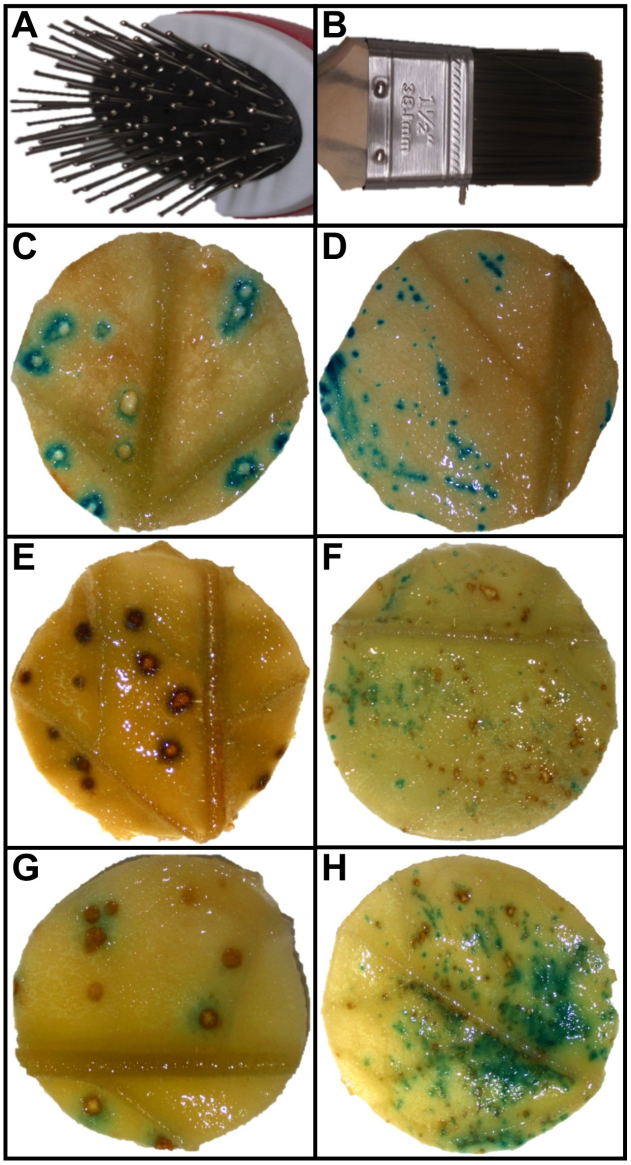


Figure 4.1: Minimizing wounding enhanced the detection of transfection. A) Wire dog brush. B) Paint brush. C) and D) β-glucuronidase activity was consistently detected in *N. benthamiana* when GV3101 containing p*CAMBIA1301* was applied to leaves wounded, respectively, with a dog brush or a paint brush. All *N. benthamiana* plants tested for this experiment had between 8 and 11 leaves, and were assayed after 6 d. Note the absence of melanic pigment at the wound sites. (E–H) Two randomly chosen discs cut from SSI leaves with either 25-30 leaves (E, F) or 12-15 leaves (G, H) after they had been wounded with a dog brush (E, G) or a paint brush (F, H) and infected for 6 d with Agrobacteria. The horizontal bar is 1 mm in length. Note in panels (F) and (G) that the majority of GUS^+ spots developed where there was very little wound reaction.

that only produced very shallow breaks in the epidermis. We continued to use the highest concentration of bacteria that we had tested previously (OD_{600}= 0.9) because it did not,

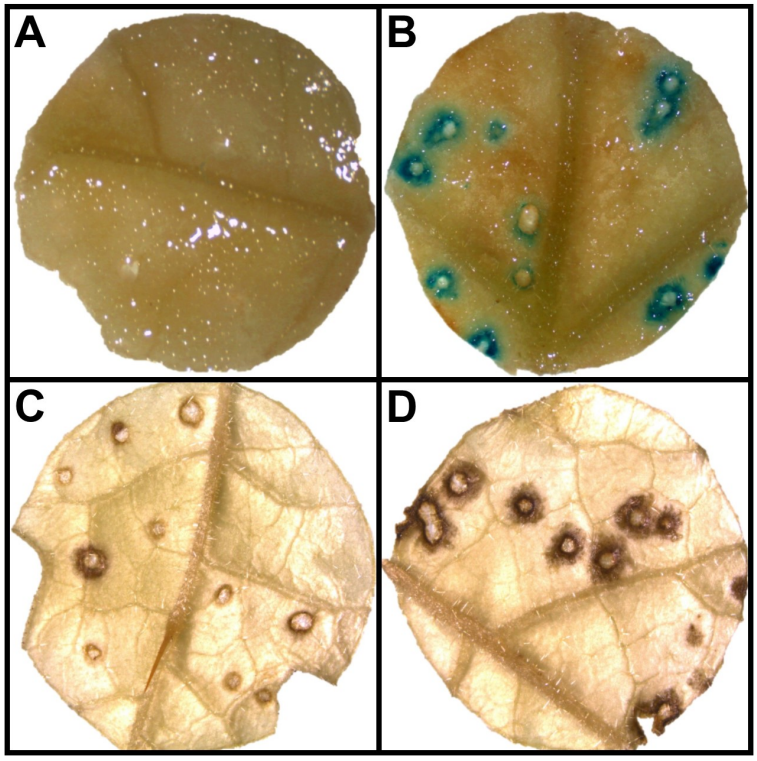


Figure 4.2: *Agrobacterium* induced pigment accumulation on wounded SSI. *N. benthamiana* (A, B) and *S. sisymbriifolium* (C, D) leaves were wounded, dipped into induced GV3101 (p*CAMBIA1301*) (B, D) or not inoculated (A, C), and after 6 d, cut into discs, stained for GUS activity, and then destained for chlorophyll. The horizontal bar is 1 mm in length. While wounding alone induced some pigment accumulation, note that application of *Agrobacterium* triggered more extensive melanin accumulation surrounding wounds on SSI but not on *N. benthamiana.*

on its own, appear to damage SSI leaves. Bacterial concentrations between 0.05 and 0.50 rarely succeeded at all (data not shown). The effectiveness of this approach is shown in Figure 4.1 E–H. *N. bentamiana* leaves expressed GUS activity regardless of how GV3101 (p*GV2260*; p*CAMBIA1301*) was applied. By comparison, GUS activity characteristic of successful gene transfer and sustained transgene expression was consistently detected only when bacteria were painted onto SSI (Fig. 4.1F, H), potato (Fig. 4.3), and tomato leaves (Fig. 4.4). Note that brown spots still arose on painted leaves, but the majority of these were GUS^-. Inclusion of the anti-oxidant N-acetylcysteine in the bacterial "paint" did not consistently enhance GUS expression, nor did inclusion of the other chemicals reported to enhance transformation such as glucose (to prolong bacteria growth *in planta*), autoclaved

glutamine solution [Sandal *et al.* 2011], Silwet detergent [Kim *et al.* 2009], silver [Singh and Prasad 2016], and the anti-oxidants, ascorbic or lipoic acids. Since none gave consistent improvements at concentrations used in these publications, none of these additives were pursued further.

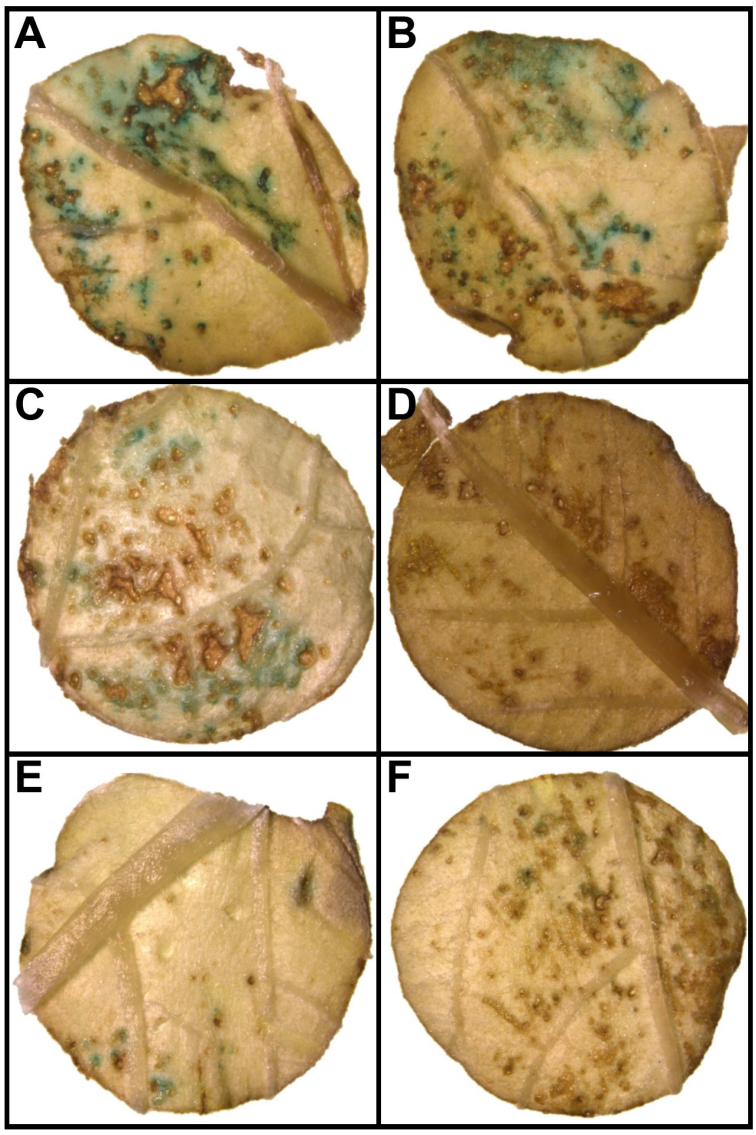


Figure 4.3: Potato leaves could be infected with a paint brush. Leaves of *Solanum tuberosum* were painted with induced GV3101 (p*CAMBIA1301*) as illustrated in Figs. 4.1F and H, and after 6 d, cut into discs, stained for GUS activity, and then destained for chlorophyll. Panels A and B represent the highest levels of infection, while C-F represent randomly chosen examples shown here to illustrate variability within the tests. The horizontal bar is 1 mm in length.

We proceeded to score the reliability and uniformity of each bacterial application by

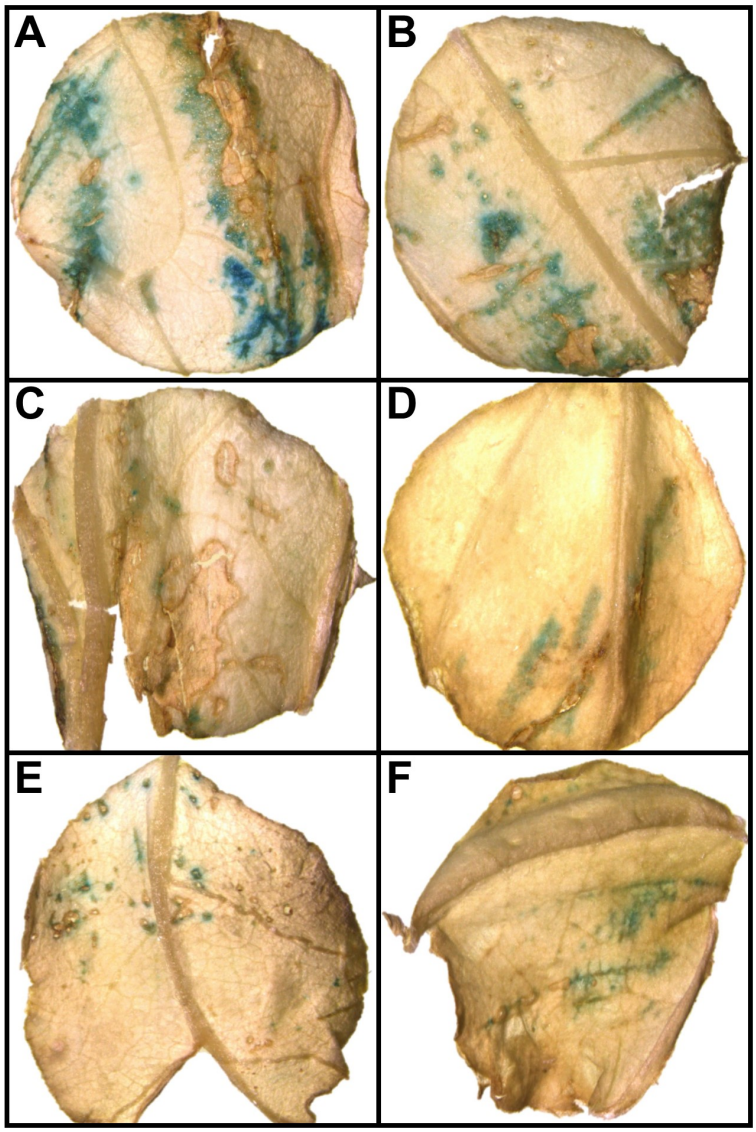


Figure 4.4: Tomato leaves could be infected with a paint brush. Leaves of *Solanum lycopersicum* were painted with induced GV3101 (p*CAMBIA1301*)and after 6 d, cut into discs, stained for GUS activity, and then destained for chlorophyll. The horizontal bar is 1 mm in length. Panels A and B represent the highest levels of infection, while C-F represent randomly chosen examples to illustrate variability in the experiment.

measuring the percentage of the leaf surface staining positively for GUS activity. By assaying leaves of the highly susceptible *N. benthamiana* plants at 0, 2, 4, 6, and 8 d, we determined that the percent of detected infections peaked at 6 d. From that point on, all infections of SSI and other plants were assayed at 6 d as well. We then endeavored to reduce experiment-to-experiment variability not only by adhering to published protocols for pressure infiltration of

bacteria [Bashandy *et al.* 2015], but also by using plants that had been vegetatively derived from a single progenitor and grown in a defined hydroponic solution. Despite these efforts, the percentages of infected leaf surface areas of SSI appeared to fluctuate greatly. Co-infection with Agrobacteria harboring a copy of the *Turnip Mosaic Virus* Hc-Pro [Chapman *et al.* 2004] to suppress potential silencing did not alter this result (data not shown). However, when the results were plotted as a function of the number of partially or fully expanded leaves on each plant at the time of infection (Fig. 4.5), two independent factors emerged that seemed to be contributing to the scatter.

The first variable we isolated was the number of infections carried out on each plant. Researchers using *N. benthamiana* or tomato routinely perform multiple infections on multiple leaves of single plants [Wu *et al.* 2004; Wroblewski *et al.* 2005]. However, we noticed that when we did the same, the median GUS-positive surface area on leaf 4 or 5 from the apex of each plant decreased from 2.3% when a single leaf was infected to 0.31% when plants were infected on 7 leaves (Fig. 4.6A). This indicated that inoculating multiple leaves induced a system-wide reaction against *Agrobacterium* that phenotypically resembled the systemic immunity other plant species show against their pathogens [Mou *et al.* 2003]. This defense either inhibited the bacteria before they could transfer their T-DNAs to the plant, or reduced expression of the T-DNAs after they were transferred, or a combination of the two.

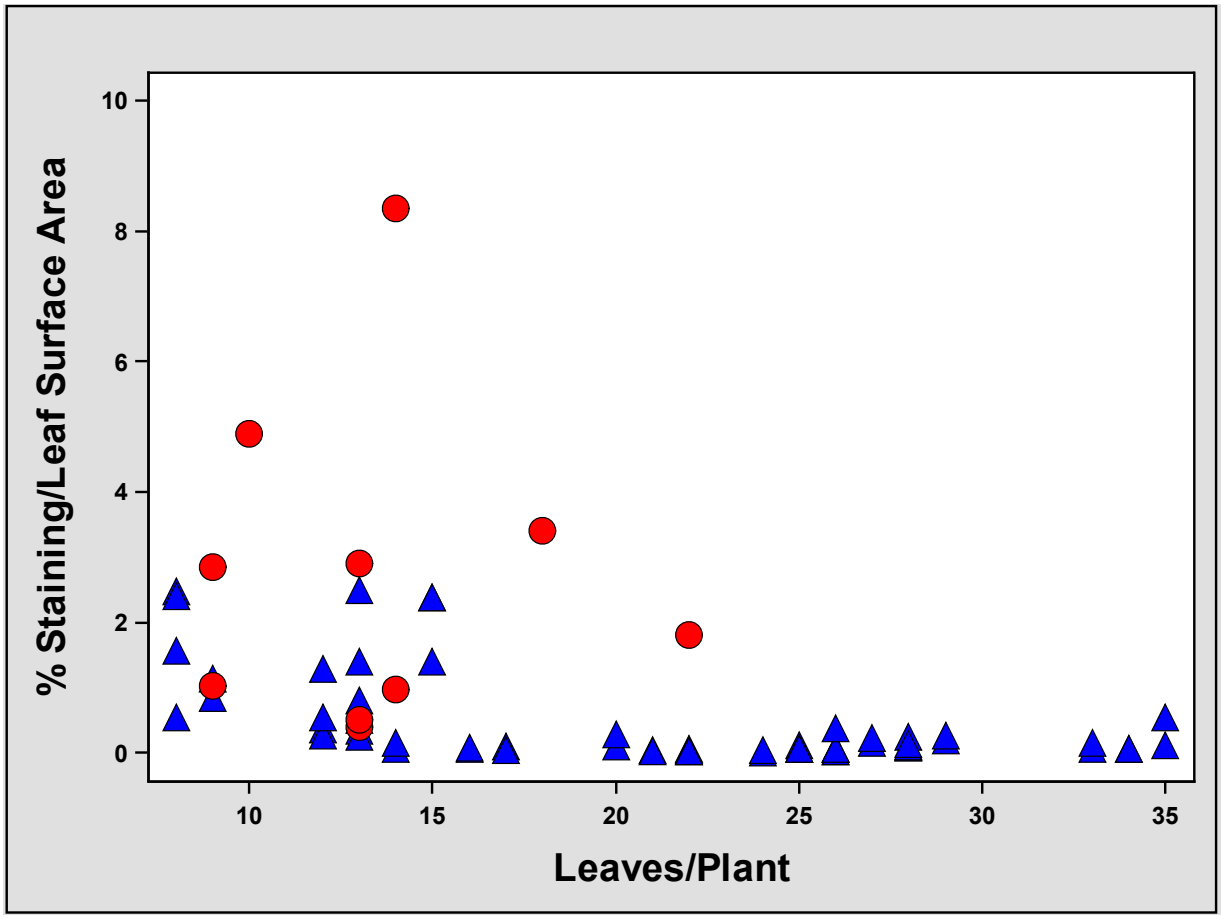


Figure 4.5: Variation in the transfection efficiency of SSI. Plants with different numbers of leaves were painted with induced GV3101 (p*CAMBIA1301*) either on a single leaf (●), or on 7 leaves (▲). These plants were then cultured 6 d after which the 4th or 5th leaf of each plant was cut into 4–12 discs. These discs were assayed for GUS expression, de-stained for chlorophyll, and the percentage of the disc stained was determined as described in Materials and Methods (Section 2.23). One naïve plant was used for each data point. Each data point, in turn, was the average percent GUS^+ area from all of the discs excised from the 4th or 5th leaf of each plant. These experiments were performed on 72 vegetatively propagated plants derived from a single individual and subcultured over the course of a yr.

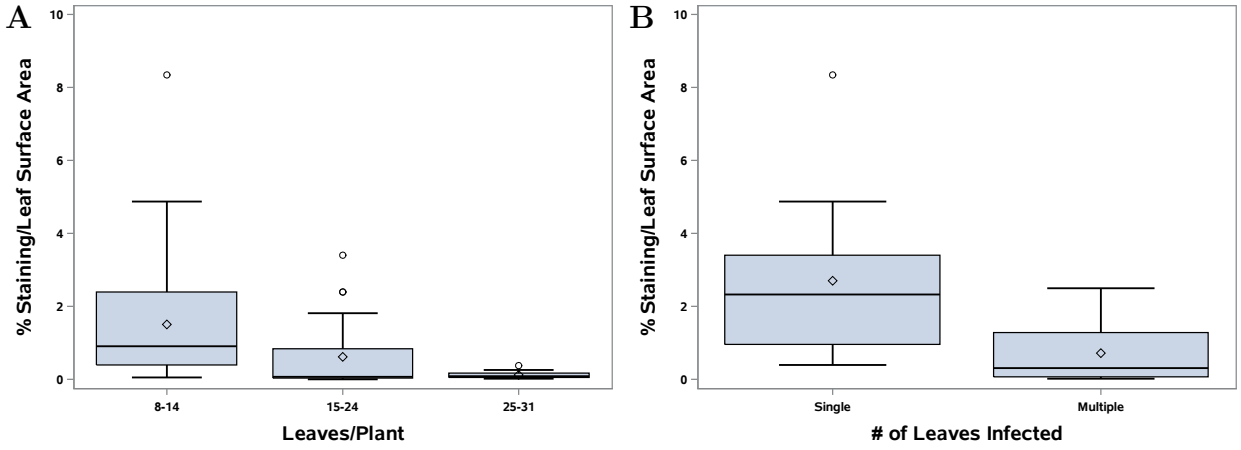


Figure 4.6: Transfection levels in SSI plants were inversely correlated with the number of leaves infected and with plant age. A). Plants with between 9 and 23 leaves were matched by age with each other. One set was painted with *Agrobacterium* on a single leaf (either 4th or 5th from apex), and the other set on 7 leaves. After 6 d, the 4th or 5th leaf from each plant was removed, assayed for GUS expression, destained for chlorophyll, and quantified as in Fig. 4.5. Each panel shows (ascending from the x-axis) the minimum observation, Q1 (the 25th percentile, indicated by the lower boundary of the grey box), the median (black line spanning grey box), the mean (◇), Q3 (the 75th percentile, indicated by the upper boundary of the grey box), the maximum observation in each set, and any excluded outlying points (○). Based on this analysis, plants infected on 1 leaf tended to show higher transgene expression than plants infected on multiple leaves. B). Similar statistical analyses were performed on all singly and multiply infected plants in this study grouped by leaf number (age) into 3 sets. The youngest plants showed the broadest range of GUS^+ staining/leaf, while the oldest plants showed consistently less staining.

4.2.2 The percentage of the leaf surface infected was affected by the plant's chronological age.

The second factor affecting expression appeared to be the chronological age of the plant as measured by the number of expanded leaves: plants with fewer leaves tended to express more β-glucuronidase activity (compare Figs. 4.1E and F with 4.1G and H). While our assay did not reveal a sharp transition point indicating when the success of T-DNA delivery and expression changed, Fig. 4.6B shows that the median percent of GUS-expressing cells declined from 0.90% for plants with 8-14 leaves, to 0.09% for plants with 25-31 leaves. The median expression value (0.07%) for the intermediate set of plants closely matched the older plants, but varied over a far wider range as if the physiologies of these leaves were at different stages of a transition. This reduction in transgene expression resembled ARR as it has been defined by the developmental response of *Arabidopsis* against some of its pathogens [Kus *et al.* 2002; Carviel *et al.* 2009; Wilson *et al.* 2013; Carella *et al.* 2015].

There is still a great deal of uncertainty concerning what triggers the transition from pathogen-sensitivity to pathogen-resistance over the course of a plant's life. In tobacco, the onset of ARR coincides with its ability to flower [Hugot *et al.* 1999], however in *Arabidopsis*, ARR has most recently been linked to leaf number [Wilson *et al.* 2013]. Where it has been studied, ARR is induced throughout a plant and not limited to the position of a leaf, the site of infection on the leaf, or to the onset of senescence [Kus *et al.* 2002]. SSI seems to behave differently. In this plant, the onset of the ARR-like phenomenon did not correlate with developmental maturity since plants set flowers continuously from the time they had 5-8 leaves, regardless of whether they were growing *in vitro*, or in hydroponic conditions. Moreover, since this species is an obligate out-crosser, and all SSI used were cloned from one individual, fruit set was extremely low, and none of the plants in the assayed age classes produced seed-bearing fruit. We also saw no evidence that ARR was brought on by the more common forms of environmental stress that might occur in nature or in a greenhouse.

In fact, the hydroponic culture conditions used here were developed so that it would be possible to minimize nutrient depletion, or fluctuations in light and temperature throughout the lifespans of the plants. At no point did plants show overt signs of wilting, disease, or delays in leaf production. Other than leaf number, the only trait that roughly correlated to the onset of ARR was chlorosis at the tips of the largest leaves on 15-leaf plants (Fig. 4.7). These leaves were the first ones that emerged after plants were transferred to hydroponic conditions. However, all infections were carried out on much younger leaves that showed no signs of yellowing either before or after they were treated with bacteria.

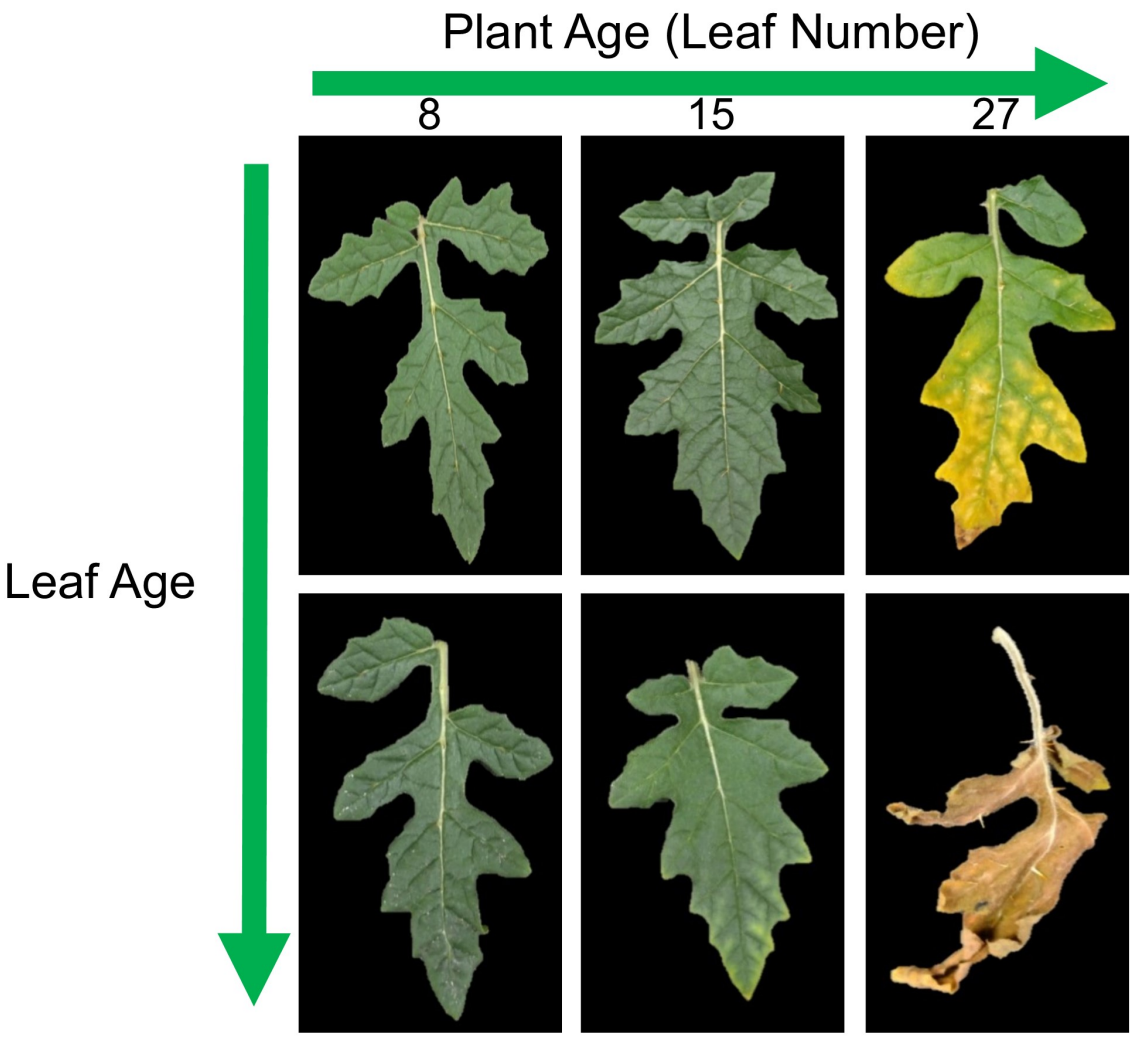


Figure 4.7: The first signs of aging-associated senescence might coincide with ARR. Plant ages were assigned based on the number of leaves attached to the main stem at the beginning of the experiment; leaf ages were assigned based on their rank position from top to bottom along the main stem. Leaves in these photographs varied in length from 10-15 cm. The lower panels show the oldest leaves of 8, 15, and 27 leaf plants, while the upper panels show the penultimate leaves from the same plants. Fifteen-leaf SSI began showing signs of ARR at the age when they also began to yellow at the tips of their oldest leaves. Only the 4th and 5th leaves from the tip were used in the experiments presented throughout this paper.

4.2.3 The age-related defense of SSI could be mimicked by salicylic acid treatments.

Before additional analyses of the data in Fig. 4.5 could be performed statistically, it was necessary to limit the analyses of each treatment to plants grouped according to their

age-classes. We therefore extracted truncated data sets made up of plants with 8-22 or 8-25 leaves, as data pairs allowed. These data were then normalized by transforming each point into the natural log of the percent surface area stained relative to the total surface area that was assayed. Fig. 4.8 shows the original values that were then analyzed superimposed, for the purpose of presentation, with the lines generated from the transformed data. As shown in Fig. 4.8A, this revealed that infecting single leaves gave significantly more GUS^+ infections than infecting multiple leaves did ($P_{lines} < 0.0001; P_{slopes} < 0.0044$).

Since many antibacterial defenses, including age-related antibacterial resistance, are brought on by rises in SA levels [Kus *et al.* 2002; Carviel *et al.* 2009; Carella *et al.* 2015], we investigated whether treating SSI with 0.1 mM SA by painting the chemical onto the underside of leaves at the time of infection duplicated the decline in GUS^+ transfections seen with older plants. Fig. 4.8B shows that this low dose of SA had no significant effect on GUS^+ transfections when multiple leaves were infected (P lines ¡0.2901; P slopes ¡0.4207). However, when these treatments were repeated on singly infected plants (Fig. 4.8C), transfection levels declined relative to untreated controls ($P_{lines} < 0.084; P_{slopes} < 0.8754$), and in fact fell to levels indistinguishable (Fig. 4.8D) from those seen during multiple infections of older plants ($P_{lines} < 0.2317; P_{slopes} < 0.2185$). Since age, multiple infection, and exogenous SA did not produce additive responses, it is possible they each induced the same anti-transfection defense.

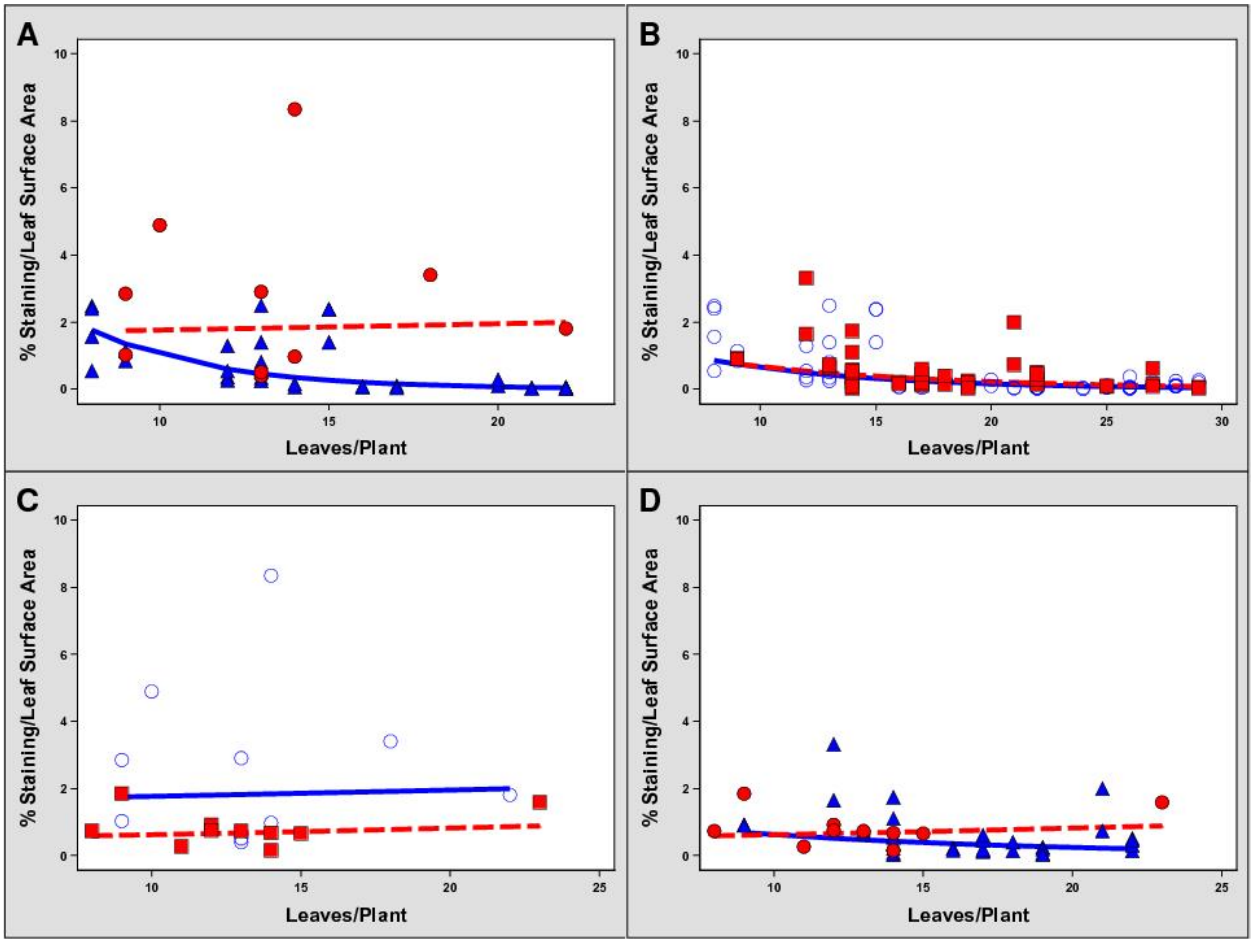


Figure 4.8: Statistical analysis of pair-wise comparisons of the effect of supplementary salicylic acid on transfection success. The set of data in Fig. 4.5 was truncated to those points that could be paired (±1 leaf) with each other. A). Singly infected plants (●) showed higher transfection frequencies than multiply-infected plants (▲). $P_{lines} < 0.0001; P_{slopes} < 0.0044$. B). In multiply-infected plants, SA treatment (■) produced similar transfection frequencies to infections without SA (○). $P_{lines} < 0.2901; P_{slopes} < 0.4207$. C). In singly-infected plants, SA treatment (■) decreased transfection frequencies compared to infections without SA (○). $P_{lines} < 0.084; P_{slopes} < 0.8754$. D). SA treatments caused singly-infected plants (●) to have the same transfection frequencies as multiply-infected plants (▲) with (or without) SA. $P_{lines} < 0.2317; P_{slopes} < 0.2185$.

4.2.4 *Agrobacterium* infections did not promote SA accumulation.

Prior research [Yuan *et al.* 2007] has shown that plants overproducing 4.5-fold more SA than wild-type plants [Bowling *et al.* 1994] had 2-fold or fewer *Agrobacterium* infections. Based on this observation, it was possible that older SSI, or SSI plants infected on mul-

tiple leaves could become less transformable than younger plants if they accumulated free SA. In order to determine whether SSI increased SA levels when treated with a recognized pathogen, 7 leaves were painted with *Pseudomonas syringae* pv. Tomato DC3000 [Whalen *et al.* 1991] and assayed [Huang *et al.* 2006; DeFraia *et al.* 2008]. Table 4.1 shows that this pseudomonad evoked a dramatic rise in SA levels in both old and young plants within 24 h. In order to test whether *Agrobacterium* evoked a similar reaction, 10-13- or 26-32-leaf plants were treated with bacteria on 7 leaves and assayed after 1 d. Unlike the changes provoked by *Pseudomonas* infection, SA levels in both uninfected and *Agrobacterium*-treated plants were statistically indistinguishable from background. This demonstrates that the response to *Agrobacterium* did not depend on large-scale changes in the concentration of this particular signaling molecule, either prior to, or after, infection. However, because we do not have mutants in SA production, we cannot rule out that low-level, or highly localized accumulations of SA might still be important.

Table 4.1: Estimates of SA levels ($\mu g\, g^{-1}$ fresh weight) in young and old plants before and after bacterial infections. Each value of SA ($\mu g\, g^{-1}$ fresh weight tissue) was an average of 3 samples (tip, middle, base)/leaf, 3 plants/treatment. The basal levels of SA in *Agrobacterium*-infected and uninfected plants are not statistically different from background.

	Young	Old
Uninfected	-7.5 ± 36	16 ± 29
Infected with *A. tumefaciens*	5.1 ± 2.2	8.7 ± 14
Infected with *P. syringae*	330 ± 43	205 ± 60

We still do not know how ARR is imposed in different species, or even whether ARR operates the same way in all plants [Develey-Rivière and Galiana 2007]. With few precedents to base our choices on, we investigated the expression of a very small set of genes believed to be involved in the endogenous production of, or response to, SA, wounding, or pathogen recognition (see Table 4.2 for the names of the genes). When normalized to the expression of an actin sequence, the majority of these genes appeared to be expressed at similar levels in 9-13- vs. 25-35-leaf plants, regardless of whether they were infected with *Agrobacterium* or not (Table 4.3). However, we did note that two genes with potential roles in increasing resistance

to bacteria, the SSI homologs to the *Arabidopsis* genes, *NPR1* [Cao *et al.* 1997], and *PAO* [Fincato *et al.* 2010], were expressed 1.4-1.5 fold higher in older plants than young ones. Following *Agrobacterium* infection of older plants, SSI *PEN3*, a protein like the one that participates in an *Arabidopsis* defense against pathogens [Xin *et al.* 2013], also increased 1.4 fold relative to its level in uninfected plants. All of these changes in expression were considerably less pronounced than the changes that occurred when plants were infected with a recognized pathogen, *P. syringae* pv. Tomato DC3000. Nevertheless, this sampling suggests that a more complete survey of the SSI transcriptome could be justified.

Table 4.2: PCR primers based on SSI sequences used to characterize gene expression. The SSI sequenced transcriptome has been deposited at DDBJ/EMBL/GenBank under the accession GGFC00000000.

Sequence	*A. thaliana* ortholog	LT Forward 5'-3'	LT Reverse 5'-3'
Ssi017026	*ssEps1* (at1g20760)	ATTCTTGAACTGGAGACTGCC	GCCTTCCCTATATCGTTCCATC
Ssi009799	*ssMes11* (at3g29770)	TTGGAAGAGATTCGTGGGTG	AGTTCAGAAACTGAGTTAGCCTC
Ssi006031	*ssNpr1* (at3g29770)	AATGCAGGAGATGGAAGAGC	CAGACATTGGGTTTTGGACG
Ssi021114	*ssPal2* (at3g53260)	GACAAGTCTTGGTGGATCATGC	TTGGTGCCTTTCCATTCTCC
Ssi023809	*ssPao1* (at2g43020)	TCAGCGCGCATGTGTATAGTC	ACCAGTCTGACCCATCACCA
Ssi007964	*ssPen1* (at3g11820)	CGAGGGTGTTGCCTTTCCTA	GAGGGTACTCCCTCAGCTCCA
Ssi035343	*ssPen3* (at1g59870)	TGCATGGACTGTTTATGGTTGC	CAGCAACTGGTGCCATGAAG
Ssi028673	*ssTsa1* (at3g54640)	GGAGCTCGTGCATCTGTGA	ACCAACAATGACGCCATCAG
Ssi032526	*ssAct7* (at5g09810)	TGCTGGTATGGAGAAGTTTGG	TTCGACAAGGGATGGTGTAAC

Table 4.3: Representative RT-PCR analysis of SA associated genes in young and old plants. All measurements are relative to actin expression. The S. sisymbriifolium sequenced transcriptome has been deposited at DDBJ/EMBL/GenBank under the accession GGFC00000000.

	Uninfected		Infected with *A. tumefaciens*		Infected with *P. syringae*	
SSI homologs	Young	Old	Young	Old	Young	Old
ssEps1	0.45 ± 0.07	0.53 ± 0.08	0.39 ± 0.13	0.47 ± 0.09	1.03 ± 0.28	0.76 ± 0.14
ssMes11	0.29 ± 0.01	0.33 ± 0.04	0.35 ± 0.07	0.34 ± 0.02	0.53 ± 0.07	0.37 ± 0.02
ssNpr1	0.48 ± 0.02	0.73 ± 0.04	0.57 ± 0.07	0.53 ± 0.02	1.28 ± 0.33	0.91 ± 0.13
ssPal2	0.30 ± 0.04	0.30 ± 0.06	0.28 ± 0.02	0.31 ± 0.01	0.68 ± 0.19	0.78 ± 0.10
ssPao1	0.57 ± 0.06	0.80 ± 0.02	0.66 ± 0.06	0.88 ± 0.12	0.85 ± 0.14	1.11 ± 0.21
ssPen1	0.37 ± 0.19	0.40 ± 0.02	0.32 ± 0.03	0.40 ± 0.07	0.75 ± 0.05	1.19 ± 0.36
ssPen3	0.33 ± 0.07	0.29 ± 0.02	0.33 ± 0.05	0.41 ± 0.08	0.99 ± 0.08	1.45 ± 0.33
ssTsa1	0.63 ± 0.03	0.63 ± 0.08	0.60 ± 0.07	0.53 ± 0.03	0.80 ± 0.01	0.74 ± 0.11

4.2.5 Mutually antagonistic interactions between plants and bacteria.

The onset of age-related resistance in plants like *Arabidopsis* correlates with a decline in the proliferation of bacterial pathogens *in planta* [Kus *et al.* 2002]. However, as noted previously, we were unable to correlate the onset of ARR with any gross change in the appearance of the plants other than with first signs of chlorosis in the oldest leaf of 15-leaf plants. Nevertheless, based on this phenotype, we assessed whether bacterial survival on 9-leaf "pre-senescence" plants (those showing no sign of chlorosis) differed from survival on 33-leaf "post-senescence" plants (those undergoing severe chlorosis on their oldest leaf). In each case, induced cultures of *Agrobacterium* were brushed onto 7 leaves and then titered using discs cut only from the plants' 4th and 5th leaves in order to reduce leaf position variation (Table 4.4). After 6 d, both sets of plants had fewer bacteria than were present initially, but the older plants had a 4-fold more dramatic reduction than did younger plants. This 4-fold difference in bacterial titer was accompanied by an 8-fold reduction in GUS-expression on plants of the same ages (Table 4.4). For purposes of comparison, bacteria were also painted onto leaves of *N. benthamiana*. Here too, Agrobacteria titers dropped by 6 d, but to a much lesser degree than seen with either 9-leaf or 33-leaf SSI: whereas populations on *N. benthamiana* fell 8-fold during the treatment period, populations on 9-leaf SSI plants fell 22-fold, and those on 33-leaf SSI fell 94-fold. As could be expected, the reduced bacterial loss led to an order of magnitude more GUS expression on *N. benthamiana* than on SSI.

As previously noted, each of the plants tested in Table 4.4 had been infected on 7 leaves. In agreement with previous statistical analyses (Fig. 4.8B), when these "multiply-infected" SSI plants were treated with SA, there was no change in either bacterial survival or GUS-expression.

ARR operates to fight pathogen infections; however, other than sporadic pigment accumulation along some of the brush strokes (Fig. 4.1F, H), SSI plants did not appear to be

responding to *Agrobacterium*. In order to objectively assess the health of infected plants, we assayed them with a Minolta SPAD-502 Chlorophyll Meter (Ramsey, NJ) that measures relative chlorophyll loss [Yang *et al.* 2004]. While this parameter did not change very much in the leaves of older plants, younger plants lost about 10% of their SPAD units (Table 4.4) when infected by Agrobacteria. Despite higher transfection levels (based on GUS activity), *N. benthamiana* plants showed increased SPAD units over the 6 d period.

Table 4.4: Bacterial survival correlated with better transfection. GV3101 (p*CAMBIA1301*) Agrobacteria were painted onto 7 leaves of 9-leaf (pre-senescent) and 33-leaf (post-senescent) SSI, or onto leaves of 7-11-leaf *Nicotiana benthamiana* plants. At T=0 d, SPAD 502 measurements were made on the 4th and 5th leaves. These leaves were removed, broken up with a mortar and pestle, diluted, and titered on YEB kanamycin agar plates. Six days later, SPAD measurements were made on the 4th and 5th leaf of a second pair of plants. Afterwards, these leaves were removed, crushed, and titered as before. Colonies were counted after 2 d growth at 30°C. This shows the average of 3 independent bacterial infections and titers, each performed on a single, previously untreated plant. The percent transfection (measured as GUS^+ surface area/area of each disc) was measured on T=6 d using separate plants. ND=No data.

	9-Leaf Plants		33-Leaf Plants		*N. benthamiana*	
Supplemental SA	No	Yes	No	Yes	No	Yes
Bacterial titer at 0 d	$5.1 \pm 0.83\,(10^6)$	$5.1 \pm 0.83\,(10^6)$	$5.1 \pm 0.83\,(10^6)$	$5.1 \pm 0.83\,(10^6)$	$3.9 \pm 2.2\,(10^6)$	$3.9 \pm 2.2\,(10^6)$
Bacterial titer after 6 d	$0.23 \pm 0.10\,(10^6)$	$0.21 \pm 0.02\,(10^6)$	$0.05 \pm 0.03\,(10^6)$	$0.07 \pm 0.02\,(10^6)$	$0.49 \pm 0.17\,(10^6)$	$0.51 \pm 0.15\,(10^6)$
Percent leaf surface stained	0.97 ± 0.81	0.87 ± 0.84	0.12 ± 0.10	0.16 ± 0.21	8.4 ± 2.1	ND
SPAD units at 0 d	62 ± 1.0	60 ± 1.0	67 ± 1.0	71 ± 1.0	47 ± 2.4	47 ± 3.5
SPAD units at 6 d	57 ± 1.0	58 ± 1.0	66 ± 1.0	70 ± 1.0	50 ± 2.1	53 ± 4.1

4.3 Discussion

Agrobacterium, especially non-oncogenic strains, are virtually undetectable by many host plant species [Lee *et al.* 2009; Gohlke and Deeken 2014]. As a result, *Agrobacterium* has been successfully used to transform at least 20% of the plant species that have been tested [Sindarovska *et al.* 2014]. However, the efficiencies of transformation, as measured by GUS transgene activity as well as by other assays, have varied by nearly 3 orders of magnitude [Sindarovska *et al.* 2014]. Even within the genus *Nicotiana*, which includes species that are routinely used for transgene research, efficiencies differ by 8-fold [Sheludko *et al.* 2007]. There appears to be no single cause for these differences. Some ecotypes of *Arabidopsis* appear to be recalcitrant because they fail to bind the bacteria efficiently, while others fail to incorporate the DNA in the genome [Mysore *et al.* 2000]. Still other species may be recalcitrant because they are able to induce defense measures quickly enough to kill the bacteria before they can transfer their T-DNAs [Tie *et al.* 2012].

The ability of plants to protect themselves changes over time. We failed to appreciate the contribution that this made to experimental variability until we plotted transfection efficiency against the number of leaves on each plant (Fig. 4.5). That graph showed that assay results that we thought were random and unreproducible, in fact correlated to changes in the life history of each plant, declining as plants grew larger and produced more leaves. This graph also showed that infecting multiple leaves, as is often done, produced a system-wide reaction that reduced the efficiency of transfection/plant. Even when we took all of these lessons into account, we still could not detect GFP transfections of *S. sisymbriifolium* using constructs and application conditions that worked quite well on *N. benthamiana* (data not shown). This suggests that the number of SSI cells infected, the number of T-DNAs introduced per cell, or the amount of expression from each T-DNA was lower than that obtained using model species and as a result, can only be measured by reporters like GUS that produce an accumulating product.

The present study indicated that SSI blocked *Agrobacterium*-mediated transfections in at least two different ways. First, unlike *N. benthamiana*, which showed no wound response, wounds on SSI leaves turned much darker brown when they were inoculated with Agrobacteria (Fig. 4.1). This kind of reaction to infection has been associated with the activities of defense-related polyphenol oxidases [Sullivan 2015]. Close examination of the treated leaves showed that very few of the indigo spots indicative of GUS expression co-localized with the brown spots indicative of cell death (Fig. 4.1F, H). This suggested that the transfection events that we measured occurred in those cells, or clusters of cells, where the antibacterial defenses had failed.

Second, the initial bacterial population fell 10-20-fold within 6 d of application (Table 4.4) but this die-off was far greater than the percentage of surface area showing either a wound response, or successful transfection (Fig. 4.5). Since the major portion of the leaf remained green and symptom-free, the population loss was probably not solely due to the plant's hypersensitive response at wounds. It seems more likely that bacterial death was caused either by bactericidal proteins or metabolites produced during normal growth, or by unmet bacterial nutritional needs on the leaf.

We could not determine whether the defense occurring at wounds, and the defense acting elsewhere on the leaf, were coordinated by a common signal. One of the most frequently used signaling molecules in plants is SA which mediates a number of systemic anti-pathogen defenses in potato [Yu *et al.* 1997] and tomato [Block *et al.* 2005]. While free SA accumulated during *P. syringae* infection, no amounts above background were detected when *Agrobacterium* was applied to multiple leaves, that is, when a system-wide reaction seemed to be occurring to block transfection (Fig. 4.8A). On the other hand, when less severe infections were applied, as when *Agrobacterium* was applied to single leaves, SSI did respond to exogenous SA (Fig. 4.8C). The fact that severe infections and exogenous SA affected transfection efficiency at all made it less likely that recalcitrance was due solely to the lack of an essential surface protein needed for bacterial attachment [Wagner and Matthysse 1992],

a protein responsible for importing the T-DNA into the nucleus [Tzfira *et al.* 2001], or any of the proteins needed to activate the 35S-promoter [Katagiri *et al.* 1989].

Very little is known about how ARR is coupled to plant age or size. Our system of vegetatively propagating SSI produced plants that flowered continuously both before and after the onset of ARR indicating that this anti-bacterial defense was most likely not determined solely by the time elapsed since germination, nor triggered solely by the onset of a reproductive phase. It remains possible that ARR, at least in SSI, is coordinated with the beginnings of senescence in the oldest leaves of the plant, but studies in *Arabidopsis* where senescence-associated genes have been monitored, have not supported such a link there [Kus *et al.* 2002]. On the other hand, it is fair to point out that this correlation may be coincidental. Instead ARR might have been triggered when plants reached a critical size that allowed their metabolic output to exceed their metabolic requirements for maintenance or reproduction. In other words, the onset of ARR might not have required the expression of new gene pathways, but instead the gradual accumulation of constitutively produced antimicrobial metabolites or proteins. Certainly, at the time of infection, SPAD measurements, and by inference, chlorophyll content, for 33-leaf plants were higher than they were for 9-leaf ones (Table 4.4). Finally, in order to assess whether a deeper exploration of the transcriptome might be justified, we surveyed the expression of several potential wound- and/or pathogen-responsive genes. As might be expected, *P. syringae* pv. Tomato DC3000 caused extensive lesions across the surface of the leaf (data not shown). *Agrobacterium*, by comparison, did not. Perhaps tellingly, *Pseudomonas* infection increased expression in both old and young plants of characteristically induced genes including *PEN1*, *PEN3*, and *PAL2*. By comparison, *Agrobacterium* infections decreased *NPR1* expression, and increased *PEN3* expression, in older plants while the expression of these genes in younger plants was unaffected. Although the changes measured here were small, it is possible that they summed together with equally slight changes in the expression of other genes to produce the 4-fold differences in transfection that we measured as ARR began. It is possible that more can

be learned once we better understand how gene expression changes as plants increase in mass. To this end, we have constructed a reference transcriptome (DDBJ/EMBL/GenBank under the accession GGFC00000000) that can be queried using RNAseq methodologies for changes in gene expression that correlate with leaf age. Regardless of its cause, this study indicates that some recalcitrant species may have brief periods of heightened susceptibility to *Agrobacterium* during their growth cycle when cultured under conditions like we have used here.

CHAPTER 5

Summary and Conclusions

Prior to the research in these chapters, very few genetic studies chose *S. sisymbriifolium* (SSI) as their focus other than those dealing with phylogeny [Levin *et al.* 2006; Miz *et al.* 2008; Paul and Banerjee 2015; Särkinen *et al.* 2015] or with the use of the plant as a trap crop for plant parasitic nematodes [Timmermans 2005; Dandurand and Knudsen 2016]. While I greatly doubt this plant will ever serve as a model for plant biologists in the same way that tobacco or tomatoes have, these few prior investigations have indicated that SSI has some useful things to teach us about anti-pathogen defenses. In order to assist future researchers determine how these defenses operate, I sought to provide them with the minimum set of tools they might need. I have established a high quality *de novo* transcriptome (Chapter 3) for SSI and demonstrated that the recalcitrance associated with transiently expressing transfected genes can be mitigated through the optimization of *Agrobacterium* infection timing (Chapter 4). I showed that SSI develops an "age-related resistance" (ARR) to *Agrobacterium* infection, but one that doesn't follow the time course operating in other plants. Based on this foundation, it should now be possible to delve into the protective responses SSI has towards *Globodera pallida*, *Agrobacterium tumefaciens*, and other pathogens.

It is important to note that despite the combination of bioinformatic and wet-lab correlative assays, the identification of SSI's protective capability may still prove elusive. No mutants have ever been identified that affect any relevant phenotype. Without this tool that has been so useful in studying developmental and physiological processes in plants such as *Arabidopsis*, maize, rice, and tomato, we can only speculate on what a given protein is doing through comparisons between it and a similar protein from other, generally unrelated, species. In order to verify that function, we can only introduce it into a heterologous species. Assaying the transgenics made using potatoes or other species related to SSI assumes that the relative has the rest of the genes needed for the introduced one to establish a new or altered pathway. Conversely, it is possible that SSI is missing a particular gene that produces

a protein necessary for infection by these pathogens. In this case, we would need to look not for genes induced by the pathogen, but rather genes missing in an SSI expression profile or genome but present in sensitive plants. I would then show that knocking them out in that sensitive plant creates a resistant line.

Infected plant cells produce diffusible signaling molecules such as jasmonic acid (JA; [Penninckx *et al.* 1998]) and salicylic acid (SA; [Métraux *et al.* 1990]) in response to pathogen infection. However, these chemicals are produced in response to many unrelated pathogens as well. In general, JA is triggered when "necrotrophic" pathogens attack, while SA is produced during biotrophic interactions [Glazebrook 2005]. Once one of these pathways has been induced, a range of anti-pathogen proteins are produced, although only some of them are likely to have any effect on any given invader. In order to identify which of the induced genes is most likely to be targeting nematodes, I believe we need to distinguish nematode-specific responses from non-specific responses. In order to do this, the current work described in this dissertation should now be followed up by a new series of studies. Since SSI has not been studied before, the first experiment should be determining whether either JA or SA is induced during *G. pallida* infection. For example, if there was an increase in SA concentration at that time, then it would be logical to focus future work on determining which genes are induced both by SA and by nematode invasion. If we find no change in SA but instead an increase in JA levels during infection, we might tentatively conclude that *G. pallida* provokes a necrotrophic response in SSI and so should look accordingly for genes associated with that kind of response in other species.

To investigate these responses, we will need to perform an RNA-seq analysis. This set of RNA-seq experiments could encompass both the pathogen infections as well as exogenous treatment of the plants with hormones. In addition, the plants could be wounded and RNA could be isolated from that as well since it is commonly believed that nematodes wound plants as they move through the roots. Using this combination of treatments, genes can be classified as pathogen-induced, hormone-induced, wounding-induced, or sets responding

to a combination of these treatments. Using these gene sets with the previously discussed correlative hormone changes, research can focus on a smaller subset of the RNAseq data to establish the genes or pathways potentially responsible for protection. The identification of these potential pathways and genes is predicated on the assumption that there is a change in gene expression that is causing the protective response. If the gene or pathway that is defending SSI from these pathogens is constitutively expressed, it could still respond post-translationally without showing up in a RNA-seq analysis.

Another way to determine the pathways that are responsible for the defense of SSI to nematodes would be to pre-treat the plants with a hormone and then attempt infections with the pathogens. This can be done with any of the hormones found to change in the chemical screen. This pre-treatment could change the infectivity of *G. pallida* and lead to a correlation of pathogenesis to a particular signalling pathway. I tried this approach in my investigations of ARR in Chapter 4, and found evidence that SA reduces *Agrobacterium* infection. In these proposed studies using *Globodera*, I would include treatments of potato to provide a positive infection control. We know SSI is inherently less susceptible to *Globodera* than is potato, therefore it might not be possible to see this susceptibility decrease. On the other hand, we should be able to see such a decrease in potato. The addition of potato infections would also allow for correlation with the protective gene(s).

There are several ways to interpret how a hormone treatment of potato could decrease the number of successful infections by *G. pallida*. First, the gene(s) needed for protection from *G. pallida* might be present in potato, but not expressed during normal nematode infection in high enough concentrations to be effective. Second, the gene for recognition might not be present in potato, but the downstream pathway for defense could be present and could be stimulated by the same hormone that induced protection in SSI. Third, neither gene nor pathway is present but the hormone turns on a different defense pathway that provides protection.

Once these potential genes or pathways have been identified using the combination of

the methods described above, further evidence for their protective potential needs to be gathered. To do this, gene knock-outs, knock-ins, and changing the level of gene expression in SSI will need to be evaluated for a change in protective capability. Gene knock-ins would need to be performed in a species that allows for pathogen infection, i.e. potatoes in the case of the pathogen *G. pallida*. Knock-outs in SSI could be accomplished using homologous recombination (which is problematic in all plants), transposon screens, or possibly the use of CRISPR/Cas9 technology. In each case, knocking out a single allele would probably have no effect and so we would have to go through a lengthy screening program to identify the loss of a single locus, and then an equally lengthy breeding program to create a homozygote that might show a phenotypic change in nematode sensitivity.

I feel we must weigh the cost-benefit balance of this work before embarking on such extensive studies.

Abel, S. and Theologis, A. (1994). Transient transformation of arabidopsis leaf protoplasts: a versatile experimental system to study gene expression. *The Plant Journal*, **5**(3):421–427.

Acosta, M. C., Guerra, M., and Moscone, E. A. (2012). Karyological relationships among some south american species of solanum (solanaceae) based on fluorochrome banding and nuclear dna amount. *Plant systematics and evolution*, **298**(8):1547–1556.

Andersson, J. O. and Andersson, S. G. (2001). Pseudogenes, junk dna, and the dynamics of rickettsia genomes. *Molecular biology and evolution*, **18**(5):829–839.

Arumuganathan, K. and Earle, E. (1991). Nuclear dna content of some important plant species. *Plant molecular biology reporter*, **9**(3):208–218.

Bagalwa, J.-J. M., Voutquenne-Nazabadioko, L., Sayagh, C., and Bashwira, A. S. (2010). Evaluation of the biological activity of the molluscicidal fraction of solanum sisymbriifolium against non target organisms. *Fitoterapia*, **81**(7):767–771.

Bakr, E. (2005). A new software for measuring leaf area, and area damaged by tetranychus urticae koch. *Journal of applied Entomology*, **129**(3):173–175.

Balasubramanian, S., Klenerman, D., and Bentley, D. (2004). Arrayed biomolecules and their use in sequencing. US Patent 6,787,308.

Bashandy, H., Jalkanen, S., and Teeri, T. H. (2015). Within leaf variation is the largest source of variation in agroinfiltration of nicotiana benthamiana. *Plant methods*, **11**(1):47.

Bentley, D. R. (2006). Whole-genome re-sequencing. *Current opinion in genetics & development*, **16**(6):545–552.

Bhaskar, P. B., Venkateshwaran, M., Wu, L., Ané, J.-M., and Jiang, J. (2009). Agrobacterium-mediated transient gene expression and silencing: a rapid tool for functional gene assay in potato. *PLoS One*, **4**(6):e5812.

Block, A., Schmelz, E., O'Donnell, P. J., Jones, J. B., and Klee, H. J. (2005). Systemic acquired tolerance to virulent bacterial pathogens in tomato. *Plant physiology*, **138**(3):1481–1490.

Blumenthal, T. (1998). Gene clusters and polycistronic transcription in eukaryotes. *Bioessays*, **20**(6):480–487.

Bolger, A., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics (Oxford, England)*, **30**(15):2114–2120.

Bonfield, J. K., Smith, K. F., and Staden, R. (1995). A new dna sequence assembly program. *Nucleic acids research*, **23**(24):4992–4999.

Bowling, S. A., Guo, A., Cao, H., Gordon, A. S., Klessig, D. F., and Dong, X. (1994). A mutation in arabidopsis that leads to constitutive expression of systemic acquired resistance. *The Plant Cell*, **6**(12):1845–1857.

Bradley, A. (1991). Modifying the mammalian genome by gene targeting. *Current Opinion in Biotechnology*, **2**(6):823–829.

Breathnach, R., Benoist, C., O'hare, K., Gannon, F., and Chambon, P. (1978). Ovalbumin gene: evidence for a leader sequence in mrna and dna sequences at the exon-intron boundaries. *Proceedings of the National Academy of Sciences*, **75**(10):4853–4857.

Bundock, P., den Dulk-Ras, A., Beijersbergen, A., and Hooykaas, P. (1995). Trans-kingdom t-dna transfer from agrobacterium tumefaciens to saccharomyces cerevisiae. *The EMBO journal*, **14**(13):3206.

Burke, D. T., Carle, G. F., and Olson, M. V. (1987). Cloning of large segments of exogenous dna into yeast by means of artificial chromosome vectors. *Science*, **236**:806–813.

Cao, H., Glazebrook, J., Clarke, J. D., Volko, S., and Dong, X. (1997). The arabidopsis npr1 gene that controls systemic acquired resistance encodes a novel protein containing ankyrin repeats. *Cell*, **88**(1):57–63.

Capecchi, M. R. (1980). High efficiency transformation by direct microinjection of dna into cultured mammalian cells. *Cell*, **22**(2):479–488.

Carella, P., Wilson, D. C., and Cameron, R. K. (2015). Some things get better with age: differences in salicylic acid accumulation and defense signaling in young and mature arabidopsis. *Frontiers in plant science*, **5**:775.

Carviel, J. L., AL-DAOUD, F., Neumann, M., Mohammad, A., Provart, N. J., Moeder, W., Yoshioka, K., and Cameron, R. K. (2009). Forward and reverse genetics to identify genes involved in the age-related resistance response in arabidopsis thaliana. *Molecular plant pathology*, **10**(5):621–634.

Casavant, N. C., Kuhl, J. C., Xiao, F., Caplan, A., and Dandurand, L.-M. (In press 2017). Assessment of *Globodera pallida* rna extracted from j2 larvae and nematode-infected solanum roots. In press.

Chalfie, M. (1994). Green fluorescent protein as a marker for gene expression. *Trends in Genetics*, **10**(5):151.

Chan, M.-T., Chang, H.-H., Ho, S.-L., Tong, W.-F., and Yu, S.-M. (1993). Agrobacterium-mediated production of transgenic rice plants expressing a chimeric α-amylase promoter/β-glucuronidase gene. *Plant molecular biology*, **22**(3):491–506.

Chang, J. H., Tai, Y.-S., Bernal, A. J., Lavelle, D. T., Staskawicz, B. J., and Michelmore, R. W. (2002). Functional analyses of the pto resistance gene family in tomato and the identification of a minor resistance determinant in a susceptible haplotype. *Molecular plant-microbe interactions*, **15**(3):281–291.

Chapman, E. J., Prokhnevsky, A. I., Gopinath, K., Dolja, V. V., and Carrington, J. C. (2004). Viral rna silencing suppressors inhibit the microrna pathway at an intermediate step. *Genes & development*, **18**(10):1179–1186.

Cheng, S., Fockler, C., Barnes, W. M., and Higuchi, R. (1994). Effective amplification of long targets from cloned inserts and human genomic dna. *Proceedings of the National Academy of Sciences*, **91**(12):5695–5699.

Collins, J. and Hohn, B. (1978). Cosmids: a type of plasmid gene-cloning vector that is packageable *in vitro* in bacteriophage lambda heads. *Proceedings of the National Academy of Sciences*, **75**(9):4242–4246.

Consortium, I. H. G. S. *et al.* (2004). Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011):931.

Consortium, S. *et al.* (1998). Genome sequence of the nematode *c. elegans*: A platform for investigating biology. *Science*, **282**:2012–2018.

Dan, Y., Yan, H., Munyikwa, T., Dong, J., Zhang, Y., and Armstrong, C. L. (2006). Microtom-a high-throughput model transformation system for functional genomics. *Plant cell reports*, **25**(5):432–441.

Dandurand, L.-M. and Knudsen, G. (2016). Effect of the trap crop solanum sisymbriifolium and two biocontrol fungi on reproduction of the potato cyst nematode, globodera pallida. *Annals of Applied Biology*, **169**(2):180–189.

De Cleene, M. and De Ley, J. (1976). The host range of crown gall. *The Botanical Review*, **42**(4):389–466.

De Ruijter, N., Verhees, J., Van Leeuwen, W., and Van der Krol, A. (2003). Evaluation and comparison of the gus, luc and gfp reporter system for gene expression studies in plants. *Plant Biology*, **5**(02):103–115.

DeFraia, C. T., Schmelz, E. A., and Mou, Z. (2008). A rapid biosensor-based method for quantification of free and glucose-conjugated salicylic acid. *Plant Methods*, **4**(1):28.

Develey-Rivière, M.-P. and Galiana, E. (2007). Resistance to pathogens and host developmental stage: a multifaceted relationship within the plant kingdom. *New Phytologist*, **175**(3):405–416.

Doležel, J., Binarová, P., and Lcretti, S. (1989). Analysis of nuclear dna content in plant cells by flow cytometry. *Biologia plantarum*, **31**(2):113–120.

Doležel, J., Doleželová, M., and Novák, F. (1994). Flow cytometric estimation of nuclear dna amount in diploid bananas (musa acuminata andm. balbisiana). *Biologia plantarum*, **36**(3):351–357.

Doležel, J., Greilhuber, J., and Suda, J. (2007). Estimation of nuclear dna content in plants using flow cytometry. *Nature protocols*, **2**(9):2233–2244.

Doležel, J., Sgorbati, S., and Lucretti, S. (1992). Comparison of three dna fluorochromes for flow cytometric estimation of nuclear dna content in plants. *Physiologia plantarum*, **85**(4):625–631.

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics*, **14**(9):755.

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

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome research*, **8**(3):175–185.

Fincato, P., Moschou, P. N., Spedaletti, V., Tavazza, R., Angelini, R., Federico, R., Roubelakis-Angelakis, K. A., and Tavladoraki, P. (2010). Functional diversity inside the arabidopsis polyamine oxidase gene family. *Journal of Experimental Botany*, **62**(3):1155–1168.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., *et al.* (2009). The pfam protein families database. *Nucleic acids research*, **38**(suppl_1):D211–D222.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L., *et al.* (2008). The pfam protein families database. *Nucleic acids research*, **36**(suppl 1):D281–D288.

Gálvez, J. H., Tai, H. H., Barkley, N. A., Gardner, K., Ellis, D., and Strömvik, M. V. (2016). Understanding potato with the help of genomics. *AIMS Agriculture and Food*, **2**(agrfood-02-00016):16.

Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. https://arxiv.org/abs/1207.3907v2.

Glazebrook, J. (2005). Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annu. Rev. Phytopathol.*, **43**:205–227.

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., *et al.* (2011). The iplant collaborative: cyberinfrastructure for plant biology. *Frontiers in plant science*, **2**:34.

Gohlke, J. and Deeken, R. (2014). Plant responses to agrobacterium tumefaciens and crown gall development. *Frontiers in plant science*, **5**:155.

Goodin, M. M., Zaitlin, D., Naidu, R. A., and Lommel, S. A. (2008). Nicotiana benthamiana: its history and future as a model for plant–pathogen interactions. *Molecular Plant-Microbe Interactions*, **21**(8):1015–1026.

Gordon, D., Abajian, C., and Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome research*, **8**(3):195–202.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, **29**(7):644.

Green, P. (1996). Phrap-sequence-assembly program. *Version 0.96073*, **1**.

Hanks, M., Wurst, W., Anson-Cartwright, L., Auerbach, A. B., and Joyner, A. L. (1995). Rescue of the en-1 mutant phenotype by replacement of en-1 with en-2. *Science*, **269**(5224):679–682.

Hooper, M., Hardy, K., Handyside, A., Hunter, S., and Monk, M. (1987). Hprt-deficient (lesch–nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature*, **326**(6110):292.

Horsch, R. and Klee, H. (1986). Rapid assay of foreign gene expression in leaf discs transformed by agrobacterium tumefaciens: Role of t-dna borders in the transfer process. *Proceedings of the National Academy of Sciences*, **83**(12):4428–4432.

Hosoda, F., Nishimura, S., Uchida, H., and Ohki, M. (1990). An f factor based cloning system for large dna fragments. *Nucleic acids research*, **18**(13):3863–3869.

Huang, W. E., Huang, L., Preston, G. M., Naylor, M., Carr, J. P., Li, Y., Singer, A. C., Whiteley, A. S., and Wang, H. (2006). Quantitative in situ assay of salicylic acid in tobacco leaves using a genetically modified biosensor strain of acinetobacter sp. adp1. *The Plant Journal*, **46**(6):1073–1083.

Hugot, K., Aimé, S., Conrod, S., Poupet, A., and Galiana, E. (1999). Developmental regulated mechanisms affect the ability of a fungal pathogen to infect and colonize tobacco leaves. *The Plant Journal*, **20**(2):163–170.

Iseli, C., Jongeneel, C. V., and Bucher, P. (1999). Estscan: a program for detecting, evaluating, and reconstructing potential coding regions in est sequences. In *ISMB*, volume 99, pages 138–148.

Jefferson, R. A., Klass, M., Wolf, N., and Hirsh, D. (1987). Expression of chimeric genes in caenorhabditis elegans. *Journal of molecular biology*, **193**(1):41–46.

Jones, J. D. and Dangl, J. L. (2006). The plant immune system. *Nature*, **444**(7117):323–329.

Joos, H., Timmerman, B., Van Montagu, M., and Schell, J. (1983). Genetic analysis of transfer and stabilization of agrobacterium dna in plant cells. *The EMBO journal*, **2**(12):2151.

Kapila, J., De Rycke, R., Van Montagu, M., and Angenon, G. (1997). An agrobacterium-mediated transient gene expression system for intact leaves. *Plant science*, **122**(1):101–108.

Katagiri, F., Lam, E., and Chua, N.-H. (1989). Two tobacco dna-binding proteins with homology to the nuclear factor creb. *Nature*, **340**(6236):727.

Kim, M. J., Baek, K., and Park, C.-M. (2009). Optimization of conditions for transient agrobacterium-mediated gene expression assays in arabidopsis. *Plant cell reports*, **28**(8):1159–1167.

King, E. O., Ward, M. K., and Raney, D. E. (1954). Two simple media for the demonstration of pyocyanin and fluorescin. *Translational Research*, **44**(2):301–307.

Komarova, T. V., Kosorukov, V. S., Frolova, O. Y., Petrunia, I. V., Skrypnik, K. A., Gleba, Y. Y., and Dorokhov, Y. L. (2011). Plant-made trastuzumab (herceptin) inhibits her2/neu+ cell proliferation and retards tumor growth. *PLoS One*, **6**(3):e17541.

Kong, N., Ng, W., Thao, K., Agulto, R., Weis, A., Kim, K. S., Korlach, J., Hickey, L., Kelly, L., Lappin, S., *et al.* (2017). Automation of pacbio smrtbell ngs library preparation for bacterial genome sequencing. *Standards in genomic sciences*, **12**(1):27.

Krasileva, K. V., Buffalo, V., Bailey, P., Pearce, S., Ayling, S., Tabbita, F., Soria, M., Wang, S., Akhunov, E., Uauy, C., *et al.* (2013). Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biology*, **14**(6):R66.

Kuehn, M. R., Bradley, A., Robertson, E. J., and Evans, M. J. (1987). A potential animal model for lesch–nyhan syndrome through introduction of hprt mutations into mice. *Nature*, **326**(6110):295.

Kunik, T., Tzfira, T., Kapulnik, Y., Gafni, Y., Dingwall, C., and Citovsky, V. (2001). Genetic transformation of hela cells by agrobacterium. *Proceedings of the National Academy of Sciences*, **98**(4):1871–1876.

Kus, J. V., Zaton, K., Sarkar, R., and Cameron, R. K. (2002). Age-related resistance in arabidopsis is a developmentally regulated defense response to pseudomonas syringae. *The Plant Cell*, **14**(2):479–490.

Lacroix, B. and Citovsky, V. (2016). Transfer of dna from bacteria to eukaryotes. *MBio*, **7**(4):e00863–16.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, **9**(4):357–359.

Lee, B., Murdoch, K., Topping, J., Kreis, M., and Jones, M. G. (1989). Transient gene expression in aleurone protoplasts isolated from developing caryopses of barley and wheat. *Plant molecular biology*, **13**(1):21–29.

Lee, C.-W., Efetova, M., Engelmann, J. C., Kramell, R., Wasternack, C., Ludwig-Müller, J., Hedrich, R., and Deeken, R. (2009). Agrobacterium tumefaciens promotes tumor induction by modulating pathogen defense in arabidopsis thaliana. *The Plant Cell*, **21**(9):2948–2962.

Levin, R. A., Myers, N. R., and Bohs, L. (2006). Phylogenetic relationships among the "spiny solanums"(solanum subgenus leptostemonum, solanaceae). *American Journal of Botany*, **93**(1):157–169.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). Upset: visualization of intersecting sets. *IEEE transactions on visualization and computer graphics*, **20**(12):1983–1992.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, **25**(14):1754–1760.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, **13**(9):2178–2189.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13):1658–1659.

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., Tohge, T., Fernie, A. R., Stitt, M., and Usadel, B. (2014). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, cell & environment*, **37**(5):1250–1258.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., and Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, **30**(5):434–439.

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., and Konstantinidis, K. T. (2012). Direct comparisons of illumina vs. roche 454 sequencing technologies on the same microbial community dna sample. *PloS one*, **7**(2):e30087.

Marck, C. (1988). 'dna strider': a 'c'program for the fast analysis of dna and protein sequences on the apple macintosh family of computers. *Nucleic acids research*, **16**(5):1829–1836.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057):376.

Mathé, C., Sagot, M.-F., Schiex, T., and Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic acids research*, **30**(19):4103–4117.

Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing dna. *Proceedings of the National Academy of Sciences*, **74**(2):560–564.

McQualter, R. B., Petrasovits, L. A., Gebbie, L. K., Schweitzer, D., Blackman, D. M., Chrysanthopoulos, P., Hodson, M. P., Plan, M. R., Riches, J. D., Snell, K. D., *et al.* (2015). The use of an acetoacetyl-coa synthase in place of a β-ketothiolase enhances poly-3-hydroxybutyrate production in sugarcane mesophyll cells. *Plant biotechnology journal*, **13**(5):700–707.

Métraux, J., Signer, H., Ryals, J., Ward, E., Wyss-Benz, M., Gaudin, J., Raschdorf, K., Schmid, E., Blum, W., and Inverardi, B. (1990). Increase in salicylic acid at the onset of systemic acquired resistance in cucumber. *Science*, **250**(4983):1004–1006.

Meyre-Silva, C., Niero, R., Bolda Mariano, L. N., Gomes do Nascimento, F., Vicente Farias, I., Gazoni, V. F., dos Santos Silva, B., Giménez, A., Gutierrez-Yapu, D., Salamanca, E., *et al.* (2013). Evaluation of antileishmanial activity of selected brazilian plants and identification of the active principles. *Evidence-Based Complementary and Alternative Medicine*, **2013**.

Miz, R. B., Mentz, L. A., and Souza-Chies, T. T. (2008). Overview of the phylogenetic relationships of some southern brazilian species from section torva and related sections of "spiny solanum"(solanum subgenus leptostemonum, solanaceae). *Genetica*, **132**(2):143–158.

Moreton, J., Izquierdo, A., and Emes, R. D. (2015). Assembly, assessment, and availability of de novo generated eukaryotic transcriptomes. *Frontiers in genetics*, **6**.

Mou, Z., Fan, W., and Dong, X. (2003). Inducers of plant systemic acquired resistance regulate npr1 function through redox changes. *Cell*, **113**(7):935–944.

Mueller, L. A., Lankhorst, R. K., Tanksley, S. D., Giovannoni, J. J., White, R., Vrebalov, J., Fei, Z., van Eck, J., Buels, R., Mills, A. A., *et al.* (2009). A snapshot of the emerging tomato genome sequence. *The Plant Genome*, **2**(1):78–92.

Murray, A. W. and Szostak, J. W. (1983). Construction of artificial chromosomes in yeast. *Nature*, **305**(189):l93.

Mysore, K. S., Kumar, C., and Gelvin, S. B. (2000). Arabidopsis ecotypes and mutants that are recalcitrant to agrobacterium root transformation are susceptible to germ-line transformation. *The Plant Journal*, **21**(1):9–16.

Nowak, R. (1994). Mining treasures from'junk dna.'(includes related glossary). *Science*, **263**(5147):608–611.

Ocwieja, K. E., Sherrill-Mix, S., Mukherjee, R., Custers-Allen, R., David, P., Brown, M., Wang, S., Link, D. R., Olson, J., Travers, K., *et al.* (2012). Dynamic regulation of hiv-1 mrna populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic acids research*, **40**(20):10345–10355.

Pagel, M. and Johnstone, R. A. (1992). Variation across species in the size of the nuclear genome supports the junk-dna explanation for the c-value paradox. *Proc. R. Soc. Lond. B*, **249**(1325):119–124.

Panter, S. N., Hammond-Kosack, K. E., Harrison, K., Jones, J. D., and Jones, D. A. (2002). Developmental control of promoter activity is not responsible for mature onset of cf-9b-mediated resistance to leaf mold in tomato. *Molecular plant-microbe interactions*, **15**(11):1099–1107.

Paul, A. and Banerjee, N. (2015). Phylogenetic relationship of some species of *SOLANUM* based on morphological, biochemical and cytological parameters. *Indian Journal of Fundamental and Applied Life Sciences*, **5**(3):51–56.

Penninckx, I. A., Thomma, B. P., Buchala, A., Métraux, J.-P., and Broekaert, W. F. (1998). Concomitant activation of jasmonate and ethylene response pathways is required for induction of a plant defensin gene in arabidopsis. *The Plant Cell*, **10**(12):2103–2113.

Pitzschke, A. (2007). Agrobacterium infection and plant defense-transformation success hangs by a thread. *Agrobacterium biology and its application to transgenic plant production*, **4**:115.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. (1996). Real-time dna sequencing using detection of pyrophosphate release. *Analytical biochemistry*, **242**(1):84–89.

Saedler, R. and Baldwin, I. T. (2004). Virus-induced gene silencing of jasmonate-induced direct defences, nicotine and trypsin proteinase-inhibitors in nicotiana attenuata. *Journal of Experimental Botany*, **55**(395):151–157.

Sambrook, J., Fritsch, E. F., Maniatis, T., *et al.* (1989). *Molecular cloning: a laboratory manual*, volume Ed. 2. Cold spring harbor laboratory press.

Sandal, I., Bhattacharya, A., Saini, U., Kaur, D., Sharma, S., Gulati, A., Kumar, J. K., Kumar, N., Dayma, J., Das, P., *et al.* (2011). Chemical modification of l-glutamine to alpha-amino glutarimide on autoclaving facilitates agrobacterium infection of host and non-host plants: A new use of a known compound. *BMC chemical biology*, **11**(1):1.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of molecular biology*, **94**(3):441–448.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, **74**(12):5463–5467.

Särkinen, T., Barboza, G. E., and Knapp, S. (2015). True black nightshades: Phylogeny and delimitation of the morelloid clade of solanum. *Taxon*, **64**(5):945–958.

Sarris, P. F., Cevik, V., Dagdas, G., Jones, J. D., and Krasileva, K. V. (2016). Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC biology*, **14**(1):8.

Schöb, H., Kunz, C., and Meins Jr, F. (1997). Silencing of transgenes introduced into leaves by agroinfiltration: a simple, rapid method for investigating sequence requirements for gene silencing. *Molecular and General Genetics MGG*, **256**(5):581–585.

Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: Robust *de novo* rna-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**(8):1086–1092.

Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., Zhang, K., Liu, J., Xi, J. J., Qiu, J.-L., *et al.* (2013). Targeted genome modification of crop plants using a crispr-cas system. *Nature biotechnology*, **31**(8):686.

Sheludko, Y., Sindarovska, Y., Gerasymenko, I., Bannikova, M., and Kuchuk, N. (2007). Comparison of several nicotiana species as hosts for high-scale agrobacterium-mediated transient expression. *Biotechnology and bioengineering*, **96**(3):608–614.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**(5741):1728–1732.

Shouichi, Y., Douglas, A. F., James, H. C., and Kwanchai, A. G. (1976). Laboratory manual for physiological studies of rice. *Manila: IRRI*, **56**:69–77.

Silverman, G. A., Ye, R. D., Pollock, K. M., Sadler, J. E., and Korsmeyer, S. J. (1989). Use of yeast artificial chromosome clones for mapping and walking within human chromosome segment 18q21. 3. *Proceedings of the National Academy of Sciences*, **86**(19):7485–7489.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**(19):3210.

Sindarovska, Y., Golovach, I., Belokurova, V., Gerasymenko, I., Sheludko, Y., and Kuchuk, N. (2014). Screening of plant cell culture collection for efficient host species for agrobacterium-mediated transient expression. *Cytology and genetics*, **48**(4):208–217.

Singh, R. K. and Prasad, M. (2016). Advances in agrobacterium tumefaciens-mediated genetic transformation of graminaceous crops. *Protoplasma*, **253**(3):691–707.

Solovyev, V. V., Salamov, A. A., and Lawrence, C. B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Research*, **22**(24):5156–5163.

Sullivan, M. L. (2015). Beyond brown: polyphenol oxidases as enzymes of plant specialized metabolism. *Frontiers in plant science*, **5**:783.

Svab, Z., Hajdukiewicz, P., and Maliga, P. (1990). Stable transformation of plastids in higher plants. *Proceedings of the National Academy of Sciences*, **87**(21):8526–8530.

Tie, W., Zhou, F., Wang, L., Xie, W., Chen, H., Li, X., and Lin, Y. (2012). Reasons for lower transformation efficiency in indica rice using agrobacterium tumefaciens-mediated transformation: lessons from transformation assays and genome-wide expression profiling. *Plant molecular biology*, **78**(1-2):1–18.

Timmermans, B. G. (2005). *Solanum sisymbriifolium (Lam.): a trap crop for potato cyst nematodes*. PhD dissertation, Wageningen University.

Tzfira, T., Vaidya, M., and Citovsky, V. (2001). Vip1, an arabidopsis protein that interacts with agrobacterium vire2, is involved in vire2 nuclear import and agrobacterium infectivity. *The EMBO Journal*, **20**(13):3596–3607.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, **30**(9):418 – 426.

Vervliet, G., Holsters, M., Teuchy, H., Van Montagu, M., and Schell, J. (1975). Characterization of different plaque-forming and defective temperate phages in agrobacterium strains. *Journal of General Virology*, **26**(1):33–48.

Wagner, V. and Matthysse, A. (1992). Involvement of a vitronectin-like protein in attachment of agrobacterium tumefaciens to carrot suspension culture cells. *Journal of bacteriology*, **174**(18):5999–6003.

Weld, R., Heinemann, J., and Eady, C. (2001). Transient gfp expression in nicotiana plumbaginifolia suspension cells: the role of gene silencing, cell death and t-dna loss. *Plant molecular biology*, **45**(4):377–385.

Whalen, M. C., Innes, R. W., Bent, A. F., and Staskawicz, B. J. (1991). Identification of pseudomonas syringae pathogens of arabidopsis and a bacterial locus determining avirulence on both arabidopsis and soybean. *The Plant Cell*, **3**(1):49–59.

Wilson, D. C., Carella, P., Isaacs, M., and Cameron, R. K. (2013). The floral transition is not the developmental switch that confers competence for the arabidopsis age-related resistance response to pseudomonas syringae pv. tomato. *Plant molecular biology*, **83**(3):235–246.

Wixom, A. Q., Casavant, N. C., Kuhl, J. C., Xiao, F., Dandurand, L.-M., and Caplan, A. B. (2018). Solanum sisymbriifolium plants become more recalcitrant to agrobacterium transfection as they age. *Physiological and Molecular Plant Pathology*, **102**:209 – 218.

Wroblewski, T., Tomczak, A., and Michelmore, R. (2005). Optimization of agrobacterium-mediated transient assays of gene expression in lettuce, tomato and arabidopsis. *Plant Biotechnology Journal*, **3**(2):259–273.

Wu, A.-J., Andriotis, V. M., Durrant, M. C., and Rathjen, J. P. (2004). A patch of surface-exposed residues mediates negative regulation of immune signaling by tomato pto kinase. *The Plant Cell*, **16**(10):2809–2821.

Xin, X.-F., Nomura, K., Underwood, W., and He, S. Y. (2013). Induction and suppression of pen3 focal accumulation during pseudomonas syringae pv. tomato dc3000 infection of arabidopsis. *Molecular plant-microbe interactions*, **26**(8):861–867.

Yadav, R., Mehrotra, M., Singh, A. K., Niranjan, A., Singh, R., Sanyal, I., Lehri, A., Pande, V., and Amla, D. (2017). Improvement in agrobacterium-mediated transformation of chickpea (cicer arietinum l.) by the inhibition of polyphenolics released during wounding of cotyledonary node explants. *Protoplasma*, **254**(1):253–269.

Yang, S.-J., Carter, S. A., Cole, A. B., Cheng, N.-H., and Nelson, R. S. (2004). A natural variant of a host rna-dependent rna polymerase is associated with increased susceptibility to viruses by nicotiana benthamiana. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(16):6297–6302.

Yang, X., Cheng, Y.-F., Deng, C., Ma, Y., Wang, Z.-W., Chen, X.-H., and Xue, L.-B. (2014). Comparative transcriptome analysis of eggplant (solanum melongena l.) and turkey berry (solanum torvum sw.): phylogenomics and disease resistance analysis. *BMC genomics*, **15**(1):412.

Yang, Y. and Smith, S. A. (2013). Optimizing de novo assembly of short-read rna-seq data for phylogenomics. *BMC genomics*, **14**(1):328.

Yassour, M., Kaplan, T., Fraser, H. B., Levin, J. Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., Gnirke, A., *et al.* (2009). Ab initio construction of a eukaryotic transcriptome by massively parallel mrna sequencing. *Proceedings of the National Academy of Sciences*, **106**(9):3264–3269.

Yeh, R.-F., Lim, L. P., and Burge, C. B. (2001). Computational inference of homologous gene structures in the human genome. *Genome research*, **11**(5):803–816.

Yu, D., Liu, Y., Fan, B., Klessig, D. F., and Chen, Z. (1997). Is the high basal level of salicylic acid important for disease resistance in potato? *Plant Physiology*, **115**(2):343–349.

Yuan, Z.-C., Edlind, M. P., Liu, P., Saenkham, P., Banta, L. M., Wise, A. A., Ronzone, E., Binns, A. N., Kerr, K., and Nester, E. W. (2007). The plant signal salicylic acid shuts down expression of the vir regulon and activates quormone-quenching genes in agrobacterium. *Proceedings of the National Academy of Sciences*, **104**(28):11790–11795.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de bruijn graphs. *Genome research*, **18**(5):821–829.

Zhang, W., Ciclitira, P., and Messing, J. (2014). Pacbio sequencing of gene families-a case study with wheat gluten genes. *Gene*, **533**(2):541–546.

Zhao, Q.-Y., Wang, Y., Kong, Y.-M., Luo, D., Li, X., and Hao, P. (2011). Optimizing de novo transcriptome assembly from short-read rna-seq data: a comparative study. In *BMC bioinformatics*, volume 12, page S2. BioMed Central.

Appendix A.1: G3 Permission

From: Ruth Isaacson <ruth.isaacson@thegsajournals.org>
Date: Saturday, July 28, 2018 at 7:01 AM
To: GENETICS Editorial Office <genetics-gsa@thegsajournals.org>, "Caplan, Allan (acaplan@uidaho.edu)" <acaplan@uidaho.edu>
Subject: Re: Remarq Feedback

Dear Dr. Caplan:

Thank you for your message. All G3 articles are an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Please consider this email as confirmation that it is permissible to publish the paper in your PhD thesis. If you need anything further, please let me know.

Regards,

Ruth Isaacson

--
Ruth Isaacson
Managing Editor
GENETICS & G3
ruth.isaacson@thegsajournals.org
412-226-5933
genetics.msubmit.net | g3.msubmit.net
genetics.org | g3journal.org

Appendix A.2: PMPP Permission

ELSEVIER

SEARCH CART MENU

Personal use

Authors can use their articles, in full or in part, for a wide range of scholarly, non-commercial purposes as outlined below:

- Use by an author in the author's classroom teaching (including distribution of copies, paper or electronic)
- Distribution of copies (including through e-mail) to known research colleagues for their personal use (but not for Commercial Use)
- Inclusion in a thesis or dissertation (provided that this is not to be published commercially)
- Use in a subsequent compilation of the author's works
- Extending the Article to book-length form
- Preparation of other derivative works (but not for Commercial Use)
- Otherwise using or re-using portions or excerpts in other works

These rights apply for all Elsevier authors who publish their article as either a subscription article or an open access article. In all cases we require that all Elsevier authors always include a full acknowledgement and, if appropriate, a link to the final published version hosted on Science Direct.