UNIVERSITY OF CALIFORNIA
RIVERSIDE

Gene Evolution in Solanaceae

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Plant Biology

by

Alex C. Rajewski

December 2020

Dissertation Committee:
Dr. Amy Litt, Chairperson
Dr. Patricia Springer
Dr. Jason Stajich

The Dissertation of Alex C. Rajewski is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgments

The text of this dissertation, in part, is a reprint of the material as it appears in Applications in Plant Sciences, February 7th, 2020. The co-author Amy Litt listed in that publication directed and supervised the research which forms the basis for this dissertation.

There are far too many people that have contributed to acknowledge in this small of a space, but I will try. My advisor, Amy Litt, has been indispensable in supporting and guiding both me and this project through some harrowing times. My committee members, Jason Stajich, Patty Springer, Norm Ellstrand, Dan Koenig, and John Heraty, have all provided sound advice, useful critiques, and necessary challenges. Jaimie Van Norman and Linda Walling have helped me immensely. I would also be remiss for not specifically recognizing Laura McGeehan who is the real-life Atlas holding up the graduate students.

I have leaned on numerous friends over the past six years. Liz Astrab reminded me that joy is everywhere. Michael Schwartz's sardonic comments made me laugh at my lowest points. Emily and Gabe provided stability and wine. Holly is twice the scientist I'll ever be, and I hope she never reads this. Liz Hann gassed me up but kept me in check. Without my Drake friends, Nate, Dan, Sam, Justine, Jenny and especially Danny, I honestly don't think I would have made it through 2020. Kevin DeBacker gave me a lot of love when I deserved it least.

Finally, I would like to thank my family, Vicki and Eva whose bottomless love, unwavering support, and selfless guidance quite literally mean the world to me.

Dedication

Dem Wahren, Schönen, Guten

ABSTRACT OF THE DISSERTATION

Gene Evolution in Solanaceae

by

Alex C. Rajewski

Doctor of Philosophy, Graduate Program in Plant Biology
University of California, Riverside, December 2020
Dr. Amy Litt, Chairperson

Among flowering plants there have been frequent evolutionary transitions from an ancestral dry fruit to a derived fleshy fruit. These transitions have dramatic consequences since fleshy fruits often attract animals, including humans, that disperse these fruits and their seeds over large distances. Rarely, however, these transitions occur in reverse, shifting from fleshy back to dry fruits. We set out to better understand these transitions and their genetic basis. To enable more detailed studies, we developed a transformation protocol for *Datura stramonium*, a species whose dry fruit exemplifies this evolutionary reversion. We then complemented this protocol with a draft genome assembly of this plant and used this to show an increased mutation rate following transformation but negligible impact on gene expression. We also found evidence in *D. stramonium* for lineage-specific gene duplications in a pathway that synthesizes medicinally important tropane alkaloids. Next, we analyzed gene expression patterns over time in the pericarps of five species with differing fruit types. This revealed a core set of 121 genes with conserved patterns of expression among all species. These core fruit development genes contained a number of known developmental regulators but also implicated unexpected developmental pathways potentially involving brassinosteroids or small RNAs.

Table of Contents

List of Figures

List of Tables

Chapter 1: General Introduction

Datura as a Model Organism

Historical Perspective

The genus *Datura* in the family Solanaceae contains several species of pharmacologically and medicinally important plants that have contributed notably to our understanding of species boundaries, chromosome complement, and secondary metabolism. Beginning in the early part of the 1900s, researchers began examining *Datura stramonium* more closely because of several conspicuous fruit and leaf phenotypes they had discovered (Blakeslee & Avery, 1919). Many of these were later determined to be caused by aneuploidy for one of the 12 chromosomes of the plant (Blakeslee, 1922). Because these mutants were easily identifiable by non-trained botanists and because the mutations were stably inherited, *D. stramonium*, often also called Jimson weed, began to be introduced in teaching laboratories as a model to study mendelian inheritance of traits (Blakeslee & Avery, 1917).

In addition to aneuploid mutations, where supernumerary copies of one or a few chromosomes are present in the nucleus, *D. stramonium* also gave rise to a number of haploid and polyploid mutants, with the complement of twelve chromosomes present either singly or in multiples higher than 3x, respectively. In fact, the first haploid flowering plants were discovered in *D. stramonium* and this was followed eventually by an entire series of polyploid series up to 8x (Blakeslee et al., 1922). These mutants provided the starting material for investigations into the interaction of ploidy, species, and self-

pollination that led to refined hypothesis about species boundaries within the genus (Belling & Blakeslee, 1922; Bergner et al., 1934; Buchholz et al., 1935; Sanders, 1948). As the researchers worked out the basis for the failures of many of these crosses, they began developing tissue culture based methods for embryo rescue, allowing the study of hybrids that would otherwise not be viable (Blakeslee & Satina, 1944). Many decades later, the application of new tissue culture techniques to *Datura* allowed the production of embryos directly from anthers and later the generation of somatic hybrids between species (Guha & Maheshwari, 1964, 1966; Schieder, 1978).

At that time, the long history of study and the multitude of techniques established in the genus made *Datura* an extremely practical system to study other aspects of plant biology. Like most plants in the family Solanaceae, *Datura spp.* produce a large number of secondary metabolites, chiefly tropane alkaloids. Although these alkaloids require elaborate biochemical pathways to synthesize, they deter pest insects and herbivores from attacking the plants (Kohnen-Johannsen & Kayser, 2019). In general alkaloids include many useful pharmacological compounds such as morphine, cocaine, heroin, nicotine, and caffeine. Tropane alkaloids of the sort produced by *Datura spp.* differ from these alkaloids by incorporating a tropane ring and chiefly include scopolamine and its precursor hyoscyamine (Parr et al., 1990). These are both extremely potent anticholinergics that produce hallucinations and delirium, but can also be used clinically to counteract a number of conditions including motion sickness, irritable bowel syndrome, eye inflammation and several other conditions (Lakstygal et al., 2019). In Native American cultures and in Ayurvedic medicine, *Datura spp.* are used to treat myriad conditions including asthma, ulcers, rheumatism, and many others (Gaire & Subedi, 2013). Because of the many uses for tropane alkaloids, *Datura* was next used as a system to study their

production and briefly as a platform for the production of the alkaloids themselves. This led to a number of studies that applied various tissue culture techniques, with and without transgenic transformation to *Datura* in order to synthesize tropane alkaloids (Hilton & Rhodes, 1990; Moyano et al., 2003; Parr et al., 1990; Payne et al., 1987; Rahman et al., 2008; Robins et al., 1991; Sangwan et al., 1991).

Beyond secondary metabolism, the genus is also notable for its fruit type from an evolutionary perspective. Within Solanaceae, early-diverging genera such as *Schizanthus*, *Petunia*, and *Nicotiana* possess a dry, capsular fruit, however, during the diversification of the subfamily Solanoideae, there was a shift in fruit type to a fleshy berry. Several notable genera are located in the subfamily Solanoideae including *Solanum* (tomato) and *Capsicum* (peppers), each of which produce fleshy berries. *Datura* is also located in this subfamily, but most species of the genus *Datura* have reverted to a dry, dehiscent capsule; however this capsule is developmentally distinct from the ancestral capsule as present outside of Solanoideae (Knapp, 2002; Pabon-Mora & Litt, 2011).

In the past, old questions have been successfully answered when enabled by new techniques. This has been the case in *Datura*, where advances in cytology and tissue culture allowed researchers to continue to propose new questions for over 100 years. Continued study of the basic questions regarding fruit-type evolution, tropane alkaloid evolution and synthesis, species boundaries, and ploidy will benefit from additional investment to develop new tools in *Datura*.

Stable Transformation and Genome Assembly

The first two chapters of this dissertation center on *D. stramonium* and provide two useful tools that will enable future genetic and genomic studies in this organism. The first

is a protocol for stable transgenic transformation of the plant. By adapting an *Agrobacterium tumefaciens*-based transformation protocol for tomato, I was able to introduce a transgene containing the *GREEN FLUORESCENT PROTEIN* gene and track the stable inheritance of this gene across several generations. The second resource developed here is the draft genome assembly and annotation of *D. stramonium*. Using a combination of Illumina short-read sequencing and Oxford Nanopore long reads, I assembled the 2 gigabase genome and subsequently annotated protein coding genes based on mRNAseq data. Combining these two resources, I characterized the impact of this stable transformation protocol on both mutation rate and leaf genes expression. These two resources provide the technical basis for future genomic and gene function studies and will streamline CRISPR-based mutagenesis experiments in the future.

Fruit Type Evolution

Background

Among flowering plants, there have been recurrent transitions in fruit evolution from an ancestral, dry fruit to a derived, fleshy fruit. This transition often has important ecological and economic consequences, as fleshy fruits provide nutrition for animals, allow for more effective seed dispersal, and are also cultivated and consumed by humans (Fleming & John Kress, 2011). One prominent example of such an evolutionary transition has occurred in the nightshade family (Solanaceae), which includes tomato, tobacco, bell pepper, eggplant and many other economically important crops. Within the nightshade family, early-diverging lineages, for example the horticulturally important plants Schizanthus, petunia, and tobacco, generally possess a dry, capsular fruit. However,

coinciding with the origin of the derived Solanoideae clade, containing tomato, pepper, and eggplant, there was a transition to a fleshy berry. This transition is not restricted to the nightshade family; many other families have seen evolutionary transitions from ancestral, dry fruits to derived, fleshy fruits (Bremer & Eriksson, 1992; Clausing et al., 2000; Cox, 1948; Givnish et al., 2005; Knapp, 2002; Plunkett et al., 1997; Spalik et al., 2001; Weber, 2004).

At lower taxonomic levels there can also be dramatic alterations in fruit morphology between species. One clear example of this is the domestication of tomato (*Solanum lycopersicum*) compared to its wild relative (*S. pimpinellifolium*). Domestication of tomato resulted in numerous changes to plant architecture, speed of development, and stress tolerance, but arguably the most conspicuous changes were to the fruits themselves (Bai & Lindhout, 2007). During the domestication of tomato a large amount of standing variation in wild species and populations was purged and many alleles that were rare among wild populations went to fixation in cultivated tomato (Blanca et al., 2015). These changes include an increase in fruit size and sugar content, but also changes in the architecture of the fruit itself. For example, the number of locules (chambers) in a wild tomato fruit is consistently two, however modern cultivated tomatoes often possess many more locules, and there are several other well studied quantitative trait loci (QTL) correlated with fruit shape, size, firmness, and color (Gonzalo & van der Knaap, 2008; Tanksley et al., 1996). These changes in fruit morphology have also been shown to coincide with a number of transcriptomic changes as well (Koenig et al., 2013).

Despite these dramatic changes in morphology between dry and fleshy fruits and between wild and cultivated fruits, there exist a large number of morphological and developmental parallels that make a comparative analysis fruitful. Fruit development can

be divided into a number of common stages regardless of the ultimate form of the fruit (Gillaspy et al., 1993). Between these fruit types, many morphological differences relate to the magnitude of processes like cell division and cell expansion rather than the presence or absence of these processes per se.

Multispecies Transcriptomes

The third chapter of this dissertation examines developmental series from the pericarp transcriptomes of five species. Between the two tomato species, I looked for shared and divergent patterns of gene expression to find genes that have potentially been affected by domestication and also profiles the expression patterns of several well-documented genes implicated in fruit ripening. Incorporating the transcriptome information from the dry-fruited desert tobacco (*Nicotiana obtusifolia*), I was able to generate a number of hypotheses about how the function of these ripening-related genes might have diverged between dry and fleshy fruits. Using wild and cultivated tomato, desert tobacco, melon (*Cucumis melo*), and the model plant *Arabidopsis thaliana*, I looked for gene expression patterns that are shared among all species in order to define a core set of genes that underlie fruit development.

References

Bai, Y., & Lindhout, P. (2007). Domestication and breeding of tomatoes: what have we gained and what can we gain in the future? *Annals of Botany*, *100*(5), 1085–1094.

Belling, J., & Blakeslee, A. F. (1922). The Assortment of Chromosomes in Triploid Daturas. *The American Naturalist*, *56*(645), 339–346.

Bergner, A. D., Cartledge, J. L., & Af, B. (1934). Chromosome behaviour due to a gene which prevents metaphase pairing in Datura. *Cytologia*, *6*(1), 19–37.

Blakeslee, A. F. (1922). Variations in Datura Due to Changes in Chromosome Number. *The American Naturalist*, *56*(642), 16–31.

Blakeslee, A. F., & Avery, B. T. (1917). ADZUKI BEANS AND JIMSON WEEDS: Favorable Class Material for Illustrating the Ratios of Mendel's Law—Actual Practice in Making Counts Is Necessary Before the Student Can Fully Grasp Modern Ideas of Heredity. *The Journal of Heredity*, *8*(3), 125–131.

Blakeslee, A. F., & Avery, B. T. (1919). Mutations in the Jimson Weed. *The Journal of Heredity*, *10*(3), 111–120.

Blakeslee, A. F., Belling, J., Farnham, M. E., & Bergner, A. D. (1922). A Haploid Mutant in the Jimson Weed, "Datura stramonium." *Science*, *55*(1433), 646–647.

Blakeslee, A. F., & Satina, S. (1944). New Hybrids from Incompatible Crosses in Datura through Culture of Excised Embryos on Malt Media. *Science*, *99*(2574), 331–334.

Blanca, J., Montero-Pau, J., Sauvage, C., Bauchet, G., Illa, E., Díez, M. J., Francis, D., Causse, M., van der Knaap, E., & Cañizares, J. (2015). Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genomics*, *16*, 257.

Bremer, B., & Eriksson, O. (1992). Evolution of fruit characters and dispersal modes in the tropical family Rubiaceae. *Biological Journal of the Linnean Society. Linnean Society of London*, *47*(1), 79–95.

Buchholz, J. T., Williams, L. F., & Blakeslee, A. F. (1935). Pollen-Tube Growth of Ten Species of Datura in Interspecific Pollinations. *Proceedings of the National Academy of Sciences of the United States of America*, *21*(12), 651–656.

Clausing, G., Meyer, K., & Renner, S. S. (2000). Correlations among fruit traits and evolution of different fruits within Melastomataceae. In *Botanical Journal of the Linnean Society* (Vol. 133, Issue 3, pp. 303–326). https://doi.org/10.1111/j.1095-8339.2000.tb01548.x

Cox, H. T. (1948). Studies in the Comparative Anatomy of the Ericales I. Ericaceae-Subfamily Rhododendroideae. In *American Midland Naturalist* (Vol. 39, Issue 1, p. 220). https://doi.org/10.2307/2421443

Fleming, T. H., & John Kress, W. (2011). A brief history of fruits and frugivores. *Acta Oecologica*, *37*(6), 521–530.

Gaire, B. P., & Subedi, L. (2013). A review on the pharmacological and toxicological aspects of Datura stramonium L. *Journal of Integrative Medicine*, *11*(2), 73–79.

Gillaspy, G., Ben-David, H., & Gruissem, W. (1993). Fruits: A Developmental Perspective. *The Plant Cell*, *5*(10), 1439–1451.

Givnish, T. J., Pires, J. C., Graham, S. W., McPherson, M. A., Prince, L. M., Patterson, T. B., Rai, H. S., Roalson, E. H., Evans, T. M., Hahn, W. J., Millam, K. C., Meerow, A. W., Molvray, M., Kores, P. J., O'Brien, H. E., Hall, J. C., Kress, W. J., & Sytsma,

K. J. (2005). Repeated evolution of net venation and fleshy fruits among monocots in shaded habitats confirms a priori predictions: evidence from an ndhF phylogeny. *Proceedings. Biological Sciences / The Royal Society*, *272*(1571), 1481–1490.
Gonzalo, M. J., & van der Knaap, E. (2008). A comparative analysis into the genetic bases of morphology in tomato varieties exhibiting elongated fruit shape. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *116*(5), 647–656.
Guha, S., & Maheshwari, S. C. (1964). In vitro Production of Embryos from Anthers of Datura. *Nature*, *204*, 497.
Guha, S., & Maheshwari, S. C. (1966). Cell Division and Differentiation of Embryos in the Pollen Grains of Datura in vitro. *Nature*, *212*, 97.
Hilton, M. G., & Rhodes, M. J. (1990). Growth and hyoscyamine production of “hairy root” cultures of Datura stramonium in a modified stirred tank reactor. *Applied Microbiology and Biotechnology*, *33*(2), 132–138.
Knapp, S. (2002). Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the Solanaceae. *Journal of Experimental Botany*, *53*(377), 2001–2022.
Koenig, D., Jiménez-Gómez, J. M., Kimura, S., Fulop, D., Chitwood, D. H., Headland, L. R., Kumar, R., Covington, M. F., Devisetty, U. K., Tat, A. V., Tohge, T., Bolger, A., Schneeberger, K., Ossowski, S., Lanz, C., Xiong, G., Taylor-Teeples, M., Brady, S. M., Pauly, M., … Maloof, J. N. (2013). Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(28), E2655–E2662.
Kohnen-Johannsen, K. L., & Kayser, O. (2019). Tropane Alkaloids: Chemistry, Pharmacology, Biosynthesis and Production. *Molecules* , *24*(4). https://doi.org/10.3390/molecules24040796
Lakstygal, A. M., Kolesnikova, T. O., Khatsko, S. L., Zabegalov, K. N., Volgin, A. D., Demin, K. A., Shevyrin, V. A., Wappler-Guzzetta, E. A., & Kalueff, A. V. (2019). DARK Classics in Chemical Neuroscience: Atropine, Scopolamine, and Other Anticholinergic Deliriant Hallucinogens. *ACS Chemical Neuroscience*, *10*(5), 2144–2159.
Moyano, E., Jouhikainen, K., Tammela, P., Palazón, J., Cusidó, R. M., Piñol, M. T., Teeri, T. H., & Oksman-Caldentey, K.-M. (2003). Effect of pmt gene overexpression on tropane alkaloid production in transformed root cultures of Datura metel and Hyoscyamus muticus. *Journal of Experimental Botany*, *54*(381), 203–211.
Pabon-Mora, N., & Litt, A. (2011). Comparative anatomical and developmental analysis of dry and fleshy fruits of Solanaceae. *American Journal of Botany*, *98*(9), 1415–1436.
Parr, A. J., Payne, J., Eagles, J., Chapman, B. T., Robins, R. J., & Rhodes, M. J. C. (1990). Variation in tropane alkaloid accumulation within the solanaceae and strategies for its exploitation. *Phytochemistry*, *29*(8), 2545–2550.
Payne, J., Hamill, J. D., Robins, R. J., & Rhodes, M. J. C. (1987). Production of hyoscyamine by “hairy root”cultures of Datura stramonium. *Planta Medica*, *53*(05), 474–478.
Plunkett, G., Soltis, D., & Soltis, P. (1997). Clarification of the relationship beteen Apiaceae and Araliaceae based on matK and rbcL sequence data. *American Journal of Botany*, *84*(4), 565.
Rahman, R. A., El-Din, E.-W., El-Said, A. G. A., & Khlifa, H. D. (2008). Agrobacterium-

mediated transformation of Datura metel (L.) and tropane alkaloids determination. *Cell Growth & Differentiation: The Molecular Biology Journal of the American Association for Cancer Research*, *2*(2), 62–66.

Robins, R. J., Parr, A. J., Bent, E. G., & Rhodes, M. J. (1991). Studies on the biosynthesis of tropane alkaloids in Datura stramonium L. transformed root cultures : 1. The kinetics of alkaloid production and the influence of feeding intermediate metabolites. *Planta*, *183*(2), 185–195.

Sanders, M. E. (1948). Embryo development in four Datura species following self and hybrid pollinations. *American Journal of Botany*, *35*(8), 525–532.

Sangwan, R. S., Ducrocq, C., & Sangwan-Norreel, B. S. (1991). Effect of culture conditions on Agrobacterium-mediated transformation in datura. *Plant Cell Reports*, *10*(2), 90–93.

Schieder, O. (1978). Somatic hybrids of Datura innoxia Mill.+Datura discolor Bernh. and of Datura innoxia Mill.+Datura stramonium L. var tatula L. *Molecular & General Genetics: MGG*, *162*(2), 113–119.

Spalik, K., Wojewódzka, A., & Downie, S. R. (2001). The evolution of fruit in Scandiceae subtribe Scandicinae (Apiaceae). *Canadian Journal of Botany. Journal Canadien de Botanique*, *79*(11), 1358–1374.

Tanksley, S. D., Grandillo, S., Fulton, T. M., Zamir, D., Eshed, Y., Petiard, V., Lopez, J., & Beck-Bunn, T. (1996). Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative L. pimpinellifolium. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *92*(2), 213–224.

Weber, A. (2004). Gesneriaceae. In *Flowering Plants · Dicotyledons* (pp. 63–158). https://doi.org/10.1007/978-3-642-18617-2_8

Chapter 2: *In Vitro* Plant Regeneration and *Agrobacterium tumefaciens*-Mediated Transformation of *Datura stramonium* (Solanaceae)

Abstract

Premise of the study: *Datura stramonium* is a pharmacologically and evolutionarily important plant species in the family Solanaceae. Stable transformation methodology of this species would be advantageous for future genetic studies.

Methods: *In vitro* plant regeneration and *Agrobacterium tumefaciens*-mediated transformation techniques were developed for *D. stramonium* based on methods reported for tomato. A binary vector containing p*At*UBQ10::erGFP was used for transformation.

Results: We recovered primary transformants harboring the GFP transgene that resulted in expression of fluorescence in all tissues analyzed. Transformants were allowed to self-pollinate, and two of five progeny contained the GFP transgene and displayed fluorescence identical to the primary transformants.

Discussion: We have demonstrated the first stable transformation in the genus *Datura*. This is a key first step to study the genetic basis of traits in this evolutionarily interesting species.

Introduction

Datura is a genus of pharmacologically important plants in the family Solanaceae. Like all members of the Solanaceae, *Datura* is notable for its production of toxic or psychoactive tropane alkaloids; however, the genus was also used extensively in the early 1900s as a model system to understand basic questions regarding hybridity,

intercrossability, and species boundaries (Blakeslee & Satina, 1944; Buchholz et al., 1935; Sanders, 1948). Early studies of polyploidy were also undertaken in *Datura*, and the first production of a haploid plant was reported in *Datura stramonium* (Blakeslee et al., 1922). From an evolutionary perspective, the genus is also notable for its fruit type. Within Solanaceae, early-diverging genera such as *Schizanthus*, *Petunia*, and *Nicotiana* possess a dry, capsular fruit. During the diversification of the subfamily Solanoideae, there was a shift in fruit type to a fleshy berry. *Datura* is located in the subfamily Solanoideae, along with fleshy-fruited genera such as *Solanum* and *Capsicum*; however, most species of the genus *Datura* have reverted to a dry, dehiscent capsule (Knapp, 2002).

Additional detailed studies on the genetic basis of tropane alkaloid production, fruit-type evolution, species boundaries, and other topics would benefit from the ability to stably genetically modify *Datura*. Several groups have reported and optimized various plant regeneration protocols for *Datura spp.* (Amiri et al., 2011; Amiri & Kazemitabar, 2011; Guha & Maheshwari, 1964, 1966; Sharma et al., 1993). Transient hairy-root transformation has been reported (Hilton & Rhodes, 1990; Payne et al., 1987), and other groups have transformed *D. metel*, but did not demonstrate stable inheritance of the transgene (Rahman et al., 2008). To our knowledge no one has demonstrated stable inheritance of transgenes in any species in the genus *Datura*.

Here we report the adaptation of a straightforward transformation protocol developed in cultivated tomato (*Solanum lycopersicum*) for use with *Datura stramonium* (J. Van Eck et al., 2018; Joyce Van Eck et al., 2006). This adaptation was successfully used to integrate a green fluorescent protein-encoding transgene into *D. stramonium*, and the transgene was stably inherited by the progeny of these primary transformants.

Methods

Germination and Callus Induction

Datura stramonium seeds were obtained in 2013 from JL Hudson Seedsman (La Honda, CA) and were grown under greenhouse conditions at the University of California, Riverside through several generations. To aid germination, the outer seed coat of 15 seeds was removed under a stereoscope. Seeds were surface sterilized for 3 hours with chlorine gas according to Lindsey (2017) and transferred to medium designated 1/2MS0 containing 2.15 g/L Murashige and Skoog (MS) salts, 100 mg/L myo-inositol, 2 mg/L thiamine, 0.5 mg/L pyridoxine, 0.5 mg/L nicotinic acid, 10 g/L sucrose, and 8 g/L agar. Petri dishes (100 mm x 20 mm) containing 1/2MS0 were used for germination. After 12 days, cotyledons had fully emerged and expanded, but the first true leaves had not yet appeared. The cotyledons were excised under sterile conditions, cut into ~1 cm segments and placed adaxial side down on KCMS medium [4.3 g/L MS salts, 100 mg/L myo-inositol, 1.3 mg/L thiamine, 0.2 mg/L 2,4-dichlorophenoxy acetic acid, 200 mg/l KH_2PO_4, 0.1 mg/L kinetin, 30 g/L sucrose, 5.2 g/L Agargel (Sigma-Aldrich), pH 6.0]. The cultures were maintained at 22ºC for 24 hours under 100 µmol m^{-2} s^{-1} light conditions at a 16-hour photoperiod.

Transformation and Co-cultivation

Agrobacterium tumefaciens GV3101 containing an *AtUBQ10:erGFP* binary vector (Fig. 1) was kindly provided by Dr. Jaimie Van Norman (Van Norman et al., 2014) and grown in 25 mL of liquid LB medium supplemented with gentamicin and spectinomycin to an OD_{600} of 0.6 (approximately 48 hours). The culture was pelleted by centrifugation at

4000 RPM for 10 minutes and resuspended in 25 mL of liquid 2% MSO medium [4.3 g/L MS Salts, 100 mg/L myo-inositol, 0.4 mg/L thiamine, 0.5 mg/L pyridoxine, 0.5 mg/L nicotinic acid, 2 mg/L glycine, 20 g/L sucrose, pH 5.6].

Cotyledon segments were incubated in the *Agrobacterium* suspension for approximately 5 minutes, then placed adaxial side down on a new plate of KCMS medium for cocultivation in the dark for 48 hours.

Shoot Regeneration

After cocultivation, 70 cotyledon segments were moved to 2ZBT medium containing 4.3 g/L MS salts with Nitsch Vitamins (Caisson Labs, Smithfield, UT), 100 mg/L myo-inositol, 20 g/L sucrose, 2 mg/L zeatin, 300 mg/L timentin, 9 mg/L phosphinothricin, 5.2 g/L Agargel, pH 6.0. Filter-sterilized zeatin, timentin, and phosphinothricin were added after autoclaving once the medium reached ~55ºC. The cotyledon segments were incubated under the same light conditions used for seed germination. Over the next 2 weeks, the cotyledon segments were transferred to new 2ZBT plates 3 times.

During this period, 23 cotyledon segments became necrotic and were discarded. The 47 surviving segments displayed callus growth and were transferred to 16-ounce polypropylene deli containers (Fabri-Kal, Kalamazoo, MI) containing 1ZBT medium, identical to 2ZBT except for the addition of 1 mg/L of zeatin instead of 2 mg/L. Six weeks after co-cultivation, the calli began to produce leaves. Over the next several weeks, the calli produced approximately 24 shoots.

Rooting and Greenhouse Transfer

The survival of these plants for six weeks on antibiotic-containing media indicated that they were antibiotic resistance and therefore transformed. To speed rooting, shoots that were 1-2 cm tall were excised and placed on non-selective rooting medium [4.3 g/L MS salts with Nitsch vitamins, 30 g/L sucrose, 1 mg/L indole-3-acetic-acid (IAA), 8 g/L Difco Bacto agar, pH 6.0]. After one week, bacterial contamination was evident in the containers and we therefore selected nine robust shoots for direct rooting in soil in order to avoid the loss of explants to the bacterial contamination. These shoots were excised, the cut stems dipped in Rootone (Bayer CropScience, USA), and placed directly in soil under a plastic dome to maintain humidity until root growth was evident.

For 4 weeks, the nine primary transformants did not elongate or display vigorous leaf growth as they developed roots. Three primary transformants survived this acclimatization period, while the other six were lost likely due to stressful conditions in the growth room. Approximately 4.5 months after cocultivation with *Agrobacterium*, the three remaining primary transformants were transferred into a greenhouse, where one further plant was lost to pest damage. The surviving primary transformants (T_0-1 and T_0-2) had vigorous growth and produced typical-sized leaves, fruits, and seeds after their transfer to the greenhouse. These two surviving T_0 plants were selected for further phenotyping to confirm GFP fluorescence and the presence of the transgene.

T_1 Plants

Before the transfer to the greenhouse, the 4-5 cm tall T_0 plants began to flower despite, at this stage, having very few leaves (usually fewer than three). Flowers from two

plants self-pollinated and set fruit. The T_1 seeds collected from fruits before the transfer to the greenhouse were small compared to wild-type (~1 mm vs 4 mm for wild-type), and, upon dissection, most were determined to be empty seed coats. Five viable T_1 seeds were produced by the primary transformants and pooled. The seed coats of these were removed, and the seeds were surface sterilized, germinated on 1/2MSO medium, and transferred to soil. These T_1 plants grew normally compared to wild-type plants, and displayed typical flowering time and seed set.

DNA Extraction and PCR conditions

Young leaf tissue (~3 cm^2) from T_0, T_1, and wild-type plants was harvested in 2 mL collection tubes and snap frozen in liquid nitrogen. Genomic DNA was extracted according to King et al. (2014). See https://dx.doi.org/10.17504/protocols.io.sgpebvn for a step-by-step protocol.

Primers to amplify 982bp of the GFP coding sequence were designed and checked for dimerization and deleterious secondary structure using the IDT OligoAnalyzer 3.1 (https://www.idtdna.com/calc/analyzer). The primer sequences were forward 5'-CTGTCAGTGGAGAGGGTGAAGG-3' and reverse 5'-TAAAGTTGCTCGAGGTACCCGG-3'. Approximately 50 ng of genomic DNA from each plant was used to amplify a region of the GFP coding sequence using EconoTaq Plus Green 2x Master Mix (Lucigen, Middleton, WI). Cycling conditions were an initial denaturation at 94ºC for 2m; followed by 25 cycles of 94ºC for 20s, 56ºC for 20s, and 72ºC for 60s; and a final extension step at 72ºC for 5m. PCR amplification of approximately 650bp of *ACTIN* was used as a positive control. Primers for *ACTIN* were forward 5'-GATGGATCCTCCAATCCAGACACTGTA-3' and reverse 5'-

GTATTGTGTTGGACTCTGGTGATGGTGT-3'. Cycling conditions consisted of an initial denaturation at 95ºC for 3m; followed by 20 cycles of 95ºC for 30s, 55ºC for 30s, and 72ºC for 30s; and a final extension step at 72ºC for 10m. These amplicons were visualized on a 2% agarose gel stained with GelRed (Biotium, Fremont, CA).

GFP Visualization

Vegetative and reproductive organs of wild-type plants, primary transformants (T_0), and T_1 progeny were imaged on a Leica M165FC stereoscope (Leica Microsystems, Switzerland) using white light or, for GFP, using a 40 nm-bandwidth excitation filter centered at 470 nm with a 50 nm-bandwidth barrier filter centered at 525 nm to block chlorophyll fluorescence. All white light images were taken with an exposure time of 75-100 milliseconds, and all images for GFP fluorescence were taken with a 3 second exposure time using a Leica D450 C digital microscope camera.

Results

GFP Transgene Amplification

Two primary transformants showed strong amplification for the expected 982 bp PCR product (Fig. 2). Of the five T_1 progeny assayed, three (T_1-1, T_1-2, and T_1-3) failed to show amplification for the GFP PCR product, however two others (T_1-4 and T_1-5) did produce a band of the expected size (Fig. 2). Genomic DNA from one wild-type plant, two primary transformants, and all T_1 plants was amplified for the presence of *ACTIN* as a control for DNA quality, and all showed the expected band (Fig. 2).

Fluorescence

The abaxial leaf surface from two of the primary transformants was imaged for GFP fluorescence and both individuals showed consistent and uniform fluorescence across the leaf epidermis; however, fluorescence was greater in the vasculature than in the epidermal tissue (Fig. 3). Adaxial leaf tissue also displayed uniform fluorescence. Tissue from all 4 floral whorls, immature fruits, and stem cross sections were also imaged. Stamens and pistils showed very strong fluorescence, as did nectaries and pollen. Fluorescence was very weak but detectable in the sepals and petals (data not shown). No visual evidence of mosaicism was observed. Although the erGFP reporter construct was designed in part for its even expression in Arabidopsis root tissues, it is expressed in all aerial tissues of the plant. Because we grew many of our plants in soil and not on agar plates, we chose the easier, above-ground tissue for screening and did not image below-ground-tissue for fluorescence.

The GFP transgene was not detected in three T_1 progeny (T_1-1, T_1-2, and T_1-3), and these also failed to show fluorescence above background levels. However, the two T_1 plants that did show PCR amplification of the GFP transgene also showed fluorescence similar to the primary transformants. As observed in the primary transformants, GFP fluorescence was very strong in the stamens, pistil, pollen, and nectaries, and moderate fluorescence was consistently observed in the leaf tissue.

Wild-type plants did not show fluorescence in leaf, stem and most reproductive tissues. Background fluorescence was elevated in anthers and stigmatic tissue, identical to that seen in the anthers and stigmas of non-transgenic T_1 plants.

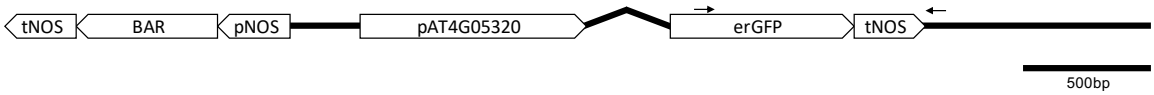
Discussion

Although GFP signal was clearly visible, the relatively low GFP fluorescence observed, especially in leaf tissues, could be due to a number of factors. The GFP transgene used in this study is endoplasmic reticulum-localized and driven by the *Arabidopsis UBIQUITIN 10* (At4g05320) promoter. Because of the comparatively large vacuoles in many plant cells, the endoplasmic reticulum is often pressed against the cell membrane, making the GFP signal in a single cell dense; however, across a given tissue, the signal will potentially appear more diffuse. Additionally, it has been reported that, when present in the oxidizing environment of the ER lumen, GFP folding can be disrupted and promote the formation of disulfide bonds between GFP molecules, potentially reducing fluorescent intensity (Aronson et al., 2011; Jain et al., 2001).

We have successfully regenerated transgenic plants from callus tissue of *Datura stramonium* and demonstrated stable inheritance of the GFP transgene. To our knowledge this is the first report of stable transformation and transgene inheritance of any species in the genus *Datura* and represents an important tool for genetic studies in this evolutionarily important genus. Availability of methodology for recovery of stable transgenic lines is a critical first step for *Datura* gene function studies through approaches such as overexpression and gene-editing by CRISPR/*Cas9* or other editing technology.

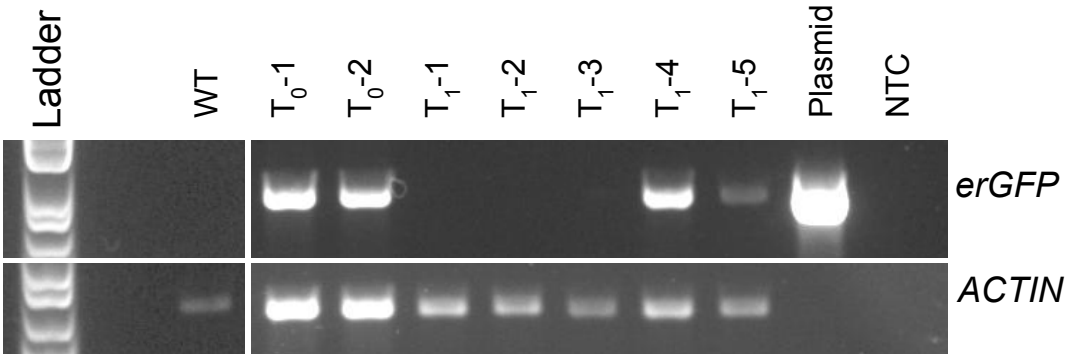
Figures

Figure 2.1 – Transgene Schematic



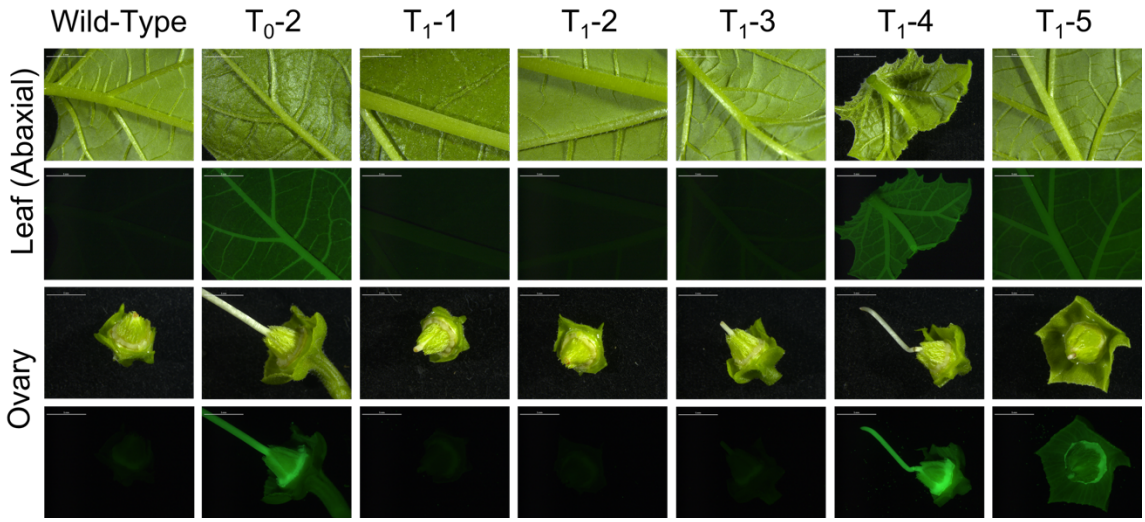
Schematic representation of the T-DNA region of the binary vector used for transformation. This vectors encodes a Basta herbicide selective marker (BAR) driven by the nopaline synthase promoter (pNOS) and terminated by the nopaline synthase terminator (tNOS). The ER-localized GFP transgene (erGFP) is driven by the Arabidopsis UBIQUITIN10 putative promoter including the 5' UTR (pAT4G05320) and is flanked upstream by the first intron of AtUBQ10. Transcription of the transgene is terminated by the nopaline synthase terminator (tNOS). Arrows above the schematic represent the locations of PCR primers used to amplify the GFP transgene. Left border and right border sequences of the binary vector (not shown) are located on the left and right sides of the schematic.

Figure 2.2 – Transgene Amplification Across Generations



PCR amplification of a 982bp region of the erGFP transgene (top row) and a ca. 650bp region of the ACTIN control (bottom row) in a wild-type plant (WT), two primary transformants (T0-1 and T0-2), five progeny of the primary transformants (T1-1 through T1-5), the vector used for transformation (Plasmid), and a negative control (NTC). All lanes with Datura DNA amplify for ACTIN, with the band falling between the 650bp and 850bp points on the ladder. Only the primary transformants, two progeny (T1-4 and T1-5), and the transformation vector amplify for the erGFP region with a band falling between the 850bp and 1000bp points on the ladder.

Figure 2.3 – GFP Fluorescence Across Generations



White light (first and third rows) and GFP-fluorescent (second and fourth rows) images of abaxial leaf surfaces (first and second rows) and ovaries (third and fourth rows) from a wild-type plant (WT), a primary transformant (T0-2), and five progeny of the primary transformants (T1-1 through T1-5). Fluorescence can been seen in leaf and ovary tissues of the primary transformant and two progeny plants (T1-4 and T1-5). All scale bars 5 mm.

References

Amiri, S., & Kazemitabar, S. K. (2011). Enhancement of callus induction and regeneration efficiency from embryo cultures of Datura stramonium by adjusting carbon sources and concentrations. *African Journal of Biotechnology*, *10*(50), 10101–10107.

Amiri, S., Kazemitabar, S. K., Ranjbar, G. A., & Azadbakht, M. (2011). In vitro propagation and whole plant regeneration from callus in Datura (Datura stramonium. L). *African Journal of Biotechnology*, *10*(3), 442–448.

Aronson, D. E., Costantini, L. M., & Snapp, E. L. (2011). Superfolder GFP is fluorescent in oxidizing environments when targeted via the Sec translocon. *Traffic* , *12*(5), 543–548.

Blakeslee, A. F., Belling, J., Farnham, M. E., & Bergner, A. D. (1922). A Haploid Mutant in the Jimson Weed, “Datura stramonium.” *Science*, *55*(1433), 646–647.

Blakeslee, A. F., & Satina, S. (1944). New Hybrids from Incompatible Crosses in Datura through Culture of Excised Embryos on Malt Media. *Science*, *99*(2574), 331–334.

Buchholz, J. T., Williams, L. F., & Blakeslee, A. F. (1935). Pollen-Tube Growth of Ten Species of Datura in Interspecific Pollinations. *Proceedings of the National Academy of Sciences of the United States of America*, *21*(12), 651–656.

Guha, S., & Maheshwari, S. C. (1964). In vitro Production of Embryos from Anthers of Datura. *Nature*, *204*, 497.

Guha, S., & Maheshwari, S. C. (1966). Cell Division and Differentiation of Embryos in the Pollen Grains of Datura in vitro. *Nature*, *212*, 97.

Hilton, M. G., & Rhodes, M. J. (1990). Growth and hyoscyamine production of “hairy root” cultures of Datura stramonium in a modified stirred tank reactor. *Applied Microbiology and Biotechnology*, *33*(2), 132–138.

Jain, R. K., Joyce, P. B., Molinete, M., Halban, P. A., & Gorr, S. U. (2001). Oligomerization of green fluorescent protein in the secretory pathway of endocrine cells. *Biochemical Journal*, *360*(Pt 3), 645–649.

King, Z., Serrano, J., Roger Boerma, H., & Li, Z. (2014). Non-toxic and efficient DNA extractions for soybean leaf and seed chips for high-throughput and large-scale genotyping. *Biotechnology Letters*, *36*(9), 1875–1879.

Knapp, S. (2002). Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the Solanaceae. *Journal of Experimental Botany*, *53*(377), 2001–2022.

Lindsey, B. E., 3rd, Rivero, L., Calhoun, C. S., Grotewold, E., & Brkljacic, J. (2017). Standardized Method for High-throughput Sterilization of Arabidopsis Seeds. *Journal of Visualized Experiments: JoVE*, *128*. https://doi.org/10.3791/56587

Payne, J., Hamill, J. D., Robins, R. J., & Rhodes, M. J. C. (1987). Production of hyoscyamine by “hairy root”cultures of Datura stramonium. *Planta Medica*, *53*(05), 474–478.

Rahman, R. A., El-Din, E.-W., El-Said, A. G. A., & Khlifa, H. D. (2008). Agrobacterium-mediated transformation of Datura metel (L.) and tropane alkaloids determination. *Cell Growth & Differentiation: The Molecular Biology Journal of the American Association for Cancer Research*, *2*(2), 62–66.

Sanders, M. E. (1948). Embryo development in four Datura species following self and hybrid pollinations. *American Journal of Botany*, *35*(8), 525–532.

Sharma, V. K., Jethwani, V., & Kothari, S. L. (1993). Embryogenesis in suspension cultures of Datura innoxia Mill. *Plant Cell Reports*, *12*(10), 581–584.

Van Eck, J., Kirk, D. D., & Walmsley, A. M. (2006). Tomato (Lycopersicum esculentum). *Methods in Molecular Biology* , *343*, 459–473.

Van Eck, J., Tjahjadi, M., & and Keen, P. (2018). Agrobacterium tumefaciens-mediated transformation of tomato. In S. Kumar, P. Barone, & M. Smith (Eds.), *Transgenic Plants: Methods and Protocols*. Springer Science+Business Media.

Van Norman, J. M., Zhang, J., Cazzonelli, C. I., Pogson, B. J., Harrison, P. J., Bugg, T. D. H., Chan, K. X., Thompson, A. J., & Benfey, P. N. (2014). Periodic root branching in Arabidopsis requires synthesis of an uncharacterized carotenoid derivative. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(13), E1300–E1309.

Chapter 3: Datura Genome Reveals Duplications of Psychoactive Alkaloid Biosynthetic Genes and High Mutation Rate Following Tissue Culture

Abstract

Background

Datura stramonium (Jimsonweed) is a medicinally and pharmaceutically important plant in the nightshade family (Solanaceae) known for its production of various toxic, hallucinogenic, and therapeutic tropane alkaloids. Recently, we published a tissue-culture based transformation protocol for *D. stramonium* that enables more thorough functional genomics studies of this plant. However, the tissue culture process can lead to undesirable phenotypic and genomic consequences independent of the transgene used. Here, we have assembled and annotated a draft genome of *D. stramonium* with a focus on tropane alkaloid biosynthetic genes. We then use mRNA sequencing and genome resequencing of transformants to characterize changes following tissue culture.

Results

Our draft assembly conforms to the expected 2 gigabasepair haploid genome size of this plant and achieved a BUSCO score of 94.7% complete, single-copy genes. The repetitive content of the genome is 61%, with *Gypsy*-type retrotransposons accounting for half of this. Our gene annotation estimates the number of protein-coding genes at 52,149 and shows evidence of duplications in two key alkaloid biosynthetic genes, tropinone

reductase I and hyoscyamine 6 β-hydroxylase. Following tissue culture, we detected only 186 differentially expressed genes, but were unable to correlate these changes in expression with either polymorphisms from resequencing or positional effects of transposons.

Conclusions

We have assembled, annotated, and characterized the first draft genome for this important model plant species. Using this resource, we show duplications of genes leading to the synthesis of the medicinally important alkaloid, scopolamine. Our results also demonstrate that following tissue culture, mutation rates of transformed plants are quite high ($1.16x10^{-3}$ mutations per site), but do not have a drastic impact on gene expression.

Background

Datura stramonium (Jimsonweed) is an important medicinal plant in the nightshade family (Solanaceae) and is known for its production of various tropane alkaloids. These alkaloids primarily consist of hyoscyamine and scopolamine, which are extremely potent anticholinergics that produce hallucinations and delirium, however, they can also be used clinically to counteract motion sickness, irritable bowel syndrome, eye inflammation and several other conditions (Lakstygal et al., 2019). Beyond Western medicine, *D. stramonium* is also used extensively in Native American cultures and in Ayurvedic medicine to treat myriad conditions including asthma, ulcers, rheumatism, and many others (Gaire & Subedi, 2013). While total synthesis of scopolamine and related precursor alkaloids is possible, extraction from plants is currently the most feasible production method (Grynkiewicz & Gadzikowska, 2008; Nocquet & Opatz, 2016). There

has been significant interest in genetic engineering or breeding for increased alkaloid content in *D. stramonium*, but like many species, we lack the genetic or genomic tools to enable this (Georgiev et al., 2013; Xia et al., 2016).

Like many plants, stable genetic engineering of *D. stramonium* requires a complex process of tissue culture. During tissue culture, the researcher uses a combination of phytohormones, typically auxins and cytokinins, to de-differentiate and maintain the starting tissue in an unorganized, totipotent mass of cells called a callus. Researchers can then infect this callus with genetically modified *Agrobacterium tumefaciens*, which integrates foreign DNA sequences randomly into the nuclear genome of the callus cells (Gelvin, 2017). Alternatively, foreign DNA sequences can be conjugated to tungsten particles (or other dense metals) and directly fired into cells using a gene gun (Sanford et al., 1987). To select for transformed cells, the introduced DNA sequences typically contain a selectable marker, often an antibiotic resistance gene. The transformed cells of the callus are then regenerated into whole plants using a combination of phytohormones to induce shoot and later root growth. This allows the production of a stable transgenic plant.

Unfortunately, in addition to being very time consuming, this process can have several unwanted genotypic and phenotypic outcomes (Filipecki & Malepszy, 2006). Many early studies documented aberrant phenotypes of plants emerging from tissue culture (Heinz & Mee, 1971; Larkin & Scowcroft, 1981). In the case of tissue culture with transformation, these aberrant phenotypes can be a result of the inserted transgene itself. T-DNA from *Agrobacterium* have been shown to preferentially integrate into transcriptionally active regions of the genome, such as the promoters of protein-coding genes thereby disrupting transcription (Alonso & Stepanova, 2003; Koncz et al., 1992; Sha et al., 2004). T-DNA constructs used for transgenic transformation also often contain

one or more strong enhancer and promoter elements such as the Cauliflower Mosaic Virus 35S promoter, which can alter transcriptional levels of genes or generate antisense transcripts (Benfey & Chua, 1985; Fitch et al., 1992; Ichikawa et al., 2003; Ko et al., 2018; Rang et al., 2005). Insertion of T-DNA sequences has also been shown to disrupt genome structure both on small and large scales, causing deletions, duplications, translocations, and transversion (Forsbach et al., 2003; Herman et al., 1990; Wenck et al., 1997).

Apart from the direct effects of the transgene insertion, tissue culture is an extremely physiologically stressful process for plant tissue. This includes exposure to exogenous and highly concentrated phytohormones and to antibiotics. In the case of *Agrobacterium* transformation, this stress is compounded by exposure to a modified (formerly) pathogenic bacterium. Each of these stressors has been independently documented to cause changes in development and to alter the genome of the plant (Bardini et al., 2003; Karp, 1991; Lucht et al., 2002; Schmitt et al., 1997; Veilleux & Johnson, 1998).

Other compelling explanations for the phenotypic and genetic changes observed following tissue culture include DNA hypomethylation, (less frequently) DNA hypermethylation, generally elevated mutation rates, and bursts of transposon activity (Hirochika, Otsuki, et al., 1996; Hirochika, Sugimoto, et al., 1996; Kaeppler et al., 2000; Kaeppler & Phillips, 1993; Kikuchi et al., 2003; Larkin & Scowcroft, 1981; LoSchiavo et al., 1989). These bursts of transposon activity were first detected by Barbara McClintock and form the basis of the genomic shock hypothesis, which posits that the up-regulation of transposable elements and their movement can lead to large scale changes in genome structure and gene expression (McClintock, 1984; Naito et al., 2009).

These genomic, genetic, and epigenetic changes are heritable in future generations after tissue culture (Kaeppler & Phillips, 1993; Marcotrigiano & Jagannathan, 1988; Oono, 1985; Stroud et al., 2013). This presents a problem for many subsequent studies as phenotypes caused by the transgene introduced with tissue culture can be confounded with unintended phenotypes resulting from the tissue culture process itself.

Importantly, the drivers of unintended but heritable changes following tissue culture are not uniform across species. For instance, although transposon bursts have been widely documented in many plant species emerging from tissue culture, this phenomenon was not detected in *Arabidopsis thaliana* plants after tissue culture (Labra et al., 2004). Because the nature and magnitude of unintended but heritable changes vary across species, tissue culture is expected to have greater or lesser impacts on the interpretations of transgenic phenotypes in each species (Johnson et al., 1987; Lee & Phillips, 1987). In contrast to *A. thaliana,* stable transgenic transformation of solanaceous plants, such as the horticulturally important species tomato (*Solanum lycopersicum*), potato (*S. tuberosum*), bell pepper (*Capsicum annuum*), petunia (*Petunia spp.*), tobacco (*Nicotiana spp.*), and *Datura stramonium* requires tissue culture, despite unreproducible claims of other transformation methods (Zhao et al., 2017). This makes characterizing the genomic impacts of tissue culture on these plants important in order to contextualize subsequent genetic and genomic studies in these species.

Previously, we published a tissue-culture based transformation protocol for *D. stramonium* and demonstrated stable inheritance and expression of a green fluorescent protein (GFP) transgene (Rajewski et al., 2019). To examine the impacts of this passage through tissue culture on genomic structure, we sequenced and assembled a reference genome of an untransformed progenitor plant. We then resequenced the genomes of

three third-generation (T3) transformants and combined this with mRNA-seq of leaf tissue to determine the impact these potential genomic changes have on gene expression.

Results

D. stramonium has a Moderately Repetitive, Average-Sized Genome for Solanaceae

Historically, *Datura stramonium* was used as a model plant to study the consequences of polyploidy and aneuploidy (Belling & Blakeslee, 1922; Blakeslee, 1921, 1922; Blakeslee et al., 1922). For this reason, we assessed the ploidy of our reference-genome prior to assembly using Smudgeplot (Ranallo-Benavidez et al., 2020). Encouragingly, raw sequencing reads supported this plant as having a diploid genome (Supplemental Fig. 3.1), and supported our strategy using general genome assembly tools.

The short-read only assembly with ABySS produced a highly fragmented 1.9Gbp assembly with over three million contigs, 85% of which were shorter than 500bp (Table 3.1). This draft assembly size corresponded well to the haploid genome size estimate of 2.0Gbp determined by flow cytometry (Kubešová et al., 2010). The completeness of this draft is also supported by the BUSCO value for complete, single-copy genes of 67.7% with an additional 14.2% fragmented genes. Due to the large number of very small contigs, the contig N50 was very low, 1.89kbp (Table 3.1). These values were the best among the various ABySS assemblies across all kmer values, and we proceeded with this assembly for further analysis.

We next used error-corrected, low-coverage long reads to scaffold this assembly and improve contiguity. In contrast to contigs, scaffolds routinely contain ambiguous bases (Ns) representing estimated gaps between contigs. Thirteen rounds of long-read scaffolding, long-read gap filling, and short-read polishing drastically improved most metrics of assembly quality. The largest scaffold increased from approximately 53kbp to nearly 1.5Mbp, while the largest contig increased to 232kbp. Similarly, the scaffold and contig N50 values increased from 1.89kbp to 103.5kbp and 5.4kbp, respectively. Because the long-read coverage was low (~7x), gap filling was only partially successful and the scaffolded assembly contained approximately 19% ambiguous bases, representing gaps between contigs in the scaffolds.

The extremely large number of small contigs (≤500bp) presented a number of computational and biological problems, in addition to negatively skewing assembly quality metrics. Working with a large number of small contigs drastically increased computing time for assembly, scaffolding, polishing, and annotation. Furthermore, their inclusion did not affect BUSCO scores, which remained at 94% after their removal, indicating that their contribution to the genic portion of the genome is negligible. Finally, genome assembly submissions to many online repositories such as NCBI limit contig size to 500bp or larger. For these reasons, after long-read scaffolding, we filtered the assembly to include only contigs and scaffolds larger than 500bp.

Polishing and gap filling using kmers derived from short-read data resulted in a 2.1Gbp assembly containing approximately 24% gaps. This increased the BUSCO score of the final assembly to 94.7%. The contig and scaffold N50 values are 13kbp and 164kbp, respectively. The largest contig and scaffold are 235kbp and 1.48Mbp, respectively (Table 3.1).

Preliminary repeat masking with RepeatModeler and RepeatMasker classified approximately 56.6% of the genome as repetitive elements (Smit et al., 2013; Smit & Hubley, 2008). Taking the 24% of the assembly represented by ambiguous nucleotides in gaps into account, this resulted in approximately 19% of the genome available for gene annotation. In relation to other sequenced Solanaceae genomes, this estimate of repetitive content for the assembled genome is comparable to that of *Nicotiana benthamiana* (61%) and *Petunia spp.* (60-65%), but much less than *Capsicum annuum* (76%), *S. lycopersicum* (72%), *N. tomentosiformis*, and *N. sylvestris* (75% and 72%, respectively) (Bombarely et al., 2016; Hosmani et al., 2019; Kim et al., 2017; Schiavinato et al., 2019; Sierro et al., 2013).

Our final nuclear genome annotation suggested 52,149 potentially protein-coding genes and an additional 1,392 tRNA loci. Most of the identified genes have few exons, with a median exon number of 2 (mean 3.8), but a midasin protein homolog with 66 exons was annotated as well (Garbarino & Gibbons, 2002). Across the genome, the median size of exons was 131bp (mean 208bp), while introns tended to be much larger with a median size of 271bp (mean 668bp) and a range between 20bp and over 14kb (Fig. 3.1A). These numbers largely agree with those from *S. lycopersicum* (Fig. 3.1B).

This estimate of gene number is based on multiple sources of evidence including mRNA-seq transcript alignments, protein sequence alignments, and several *ab initio* gene

prediction softwares. Despite this support, the total number of gene models is higher than closely related species such as tomato (34,075) and pepper (34,899) (Table 3.2) (Hosmani et al., 2019; Kim et al., 2014).

Heteroplasmy of Chloroplast Genome

We recovered sufficient reads to reconstruct the complete chloroplast genomes from our reference plant. The program GetOrganelle produced two distinct chloroplast genome assemblies, both of 155,895bp. This corresponds well to the 155,871bp size of the first published chloroplast genome of *D. stramonium* and to the 155,884bp size from a pair of more recently published *D. stramonium* chloroplast assemblies (De-la-Cruz & Núñez-Farfán, 2020; Yang et al., 2014). Following annotation with GeSeq, we noticed that our two assemblies differed from one another only in the orientation of their small single-copy region, but otherwise displayed the typical quadripartite structure of most angiosperm plastid genomes (Fig. 3.2) (Tillich et al., 2017). Inversion polymorphism within an individual is quite common among plants and has been documented many times since its discovery nearly 40 years ago (J. D. Palmer, 1983). Independent pairwise alignments of the small single-copy region and of the large single-copy region with both flanking inverted-region regions from our two genomes show no further polymorphisms between our chloroplast genomes apart from this inversion of the small single-copy region. Because the assemblies from De la Cruz et al have not been released, we aligned the complete sequence of the original assembly from Yang et al to the corresponding assembly from our study and observed a 99.97% similarity between the two, suggesting very little divergence.

Lineage-Specific Duplications Cannot Explain High Gene Number

Given the high apparent gene number in our genome annotation, we undertook a number of analyses to ascertain if this represented bona fide gene family expansions, whole genome duplications, or if it was an artifact of our annotation methods. We first looked for genes with evidence of transcription. Our mRNA-seq data from leaf tissue could provide support for 62.8% of annotated genes, leaving approximately 19,900 genes with only theoretical evidence.

To examine the possibility of lineage-specific increases in gene number, we first used OrthoFinder2 to cluster protein sequences from *D. stramonium* and 12 other angiosperm species into orthologous groups and identify gene duplication events (Emms & Kelly, 2019). Overall, approximately 12% of the over 400,000 protein sequences were assigned to lineage-specific orthogroups, suggesting a non-trivial amount of lineage-specific expansion events, but only 482 orthogroups contained genes present in a single copy across all 13 species. Together the low number of single-copy orthogroups and the moderate number of lineage-specific orthogroups suggests that there is some amount of misestimation associated with the orthology analysis, potentially caused by the large evolutionary distance between the 13 species and the polyploid history of several taxa. However, the majority of protein sequences were successfully grouped, and the inferred species tree from this analysis largely matched the previously established phylogeny of these angiosperm species (Fig. 3.3) (The Angiosperm Phylogeny Group et al., 2016). One exception to this is the Solanoideae subfamily where *C. annuum* is placed sister to a clade containing both *D. stramonium* and *S. lycopersicum*. The only other topological anomaly in the species tree is that *A. thaliana* and *V. vinifera* form a grade sister the asterid species

instead of forming their own rosid clade. The relationship of the aforementioned taxa in the species tree is inconsistent with several other phylogenetic studies, which used wider taxon sampling, and should thus be interpreted with some caution (Dupin & Smith, 2018; Särkinen et al., 2013; The Angiosperm Phylogeny Group et al., 2016).

When examining duplication events mapped onto the species tree, *D. stramonium* stands out among Solanaceae for having 14,057 lineage-specific duplication events. This number is much larger than the lineage-specific events for other Solanaceae, which ranged from 4,830 (*S. lycopersicum*) to 8,750 (*C. annuum*) (Table 3.2). Across the entire species tree, *Helianthus annuus* has more lineage-specific duplications, with 18,131. However, this species has evidence of polyploidy events after its divergence from Solanaceae (Badouin et al., 2017; Schmutz et al., 2010). Although Solanaceae is hypothesized to have undergone a whole genome triplication event, the expansion events inferred in *D. stramonium* by OrthoFinder2 were not shared with the other members of Solanaceae (Bombarely et al., 2016; Tomato Genome Consortium, 2012). If the expansion events in *D. stramonium* represent a burst of recent lineage-specific expansions, then these paralogous genes should share higher sequence similarity with each other than with orthologous genes in other Solanaceae species. To examine this possibility, we constructed Ks plots showing the frequency of synonymous substitutions (Ks) for 1) paralogous genes within *D. stramonium*, 2) paralogous genes within *S. lycopersicum*, and 3) orthologous genes between *D. stramonium* and *S. lycopersicum*. If many of these genes in *D. stramonium* are truly lineage-specific expansions since the divergence of *D. stramonium* and *S. lycopersicum*, and not the result of a misannotation, then they would be represented as a peak in the Ks plot of *D. stramonium* with a lower Ks value (more similar) than the peak for orthologous genes between *D. stramonium* and

S. lycopersicum (Cui et al., 2006). The peak Ks value for orthologous genes was approximately 0.19, and this matched individual Ks peak estimates for both *D. stramonium* and *S. lycopersicum* (Fig. 3.2). We did not detect well-supported Ks peaks for paralogous genes in either species with lower Ks values than this. Thus our analysis fails to support a burst of lineage-specific gene family expansions since the divergence of *Solanum* and *Datura*. Taken together, the large number of genes without mRNA-seq support, without obvious orthologs in 13 other angiosperms, and without evidence of evolutionarily recent lineage-specific expansions suggest that the higher number of genes in *D. stramonium* compared to other Solanaceae is likely due to overestimates of gene number rather than a bona fide increase in gene number.

We performed a GO term enrichment analysis on all of the genes from lineage-specific duplications in *D. stramonium* and *S. lycopersicum* to look for trends among these lineage-specific genes Figure 2E-F). Between these species, many of the GO terms were very broad. For example, translation (GO:0006412), oxidation-reduction processes (GO:0055114), and response to auxin (GO:0009733) were enriched in both species' datasets. Other categories of lineage-specific duplications were related to defense such as gene silencing by RNA (GO:0031047), chitin catabolic processes (GO:0006032), and response to wounding (GO:0009611).

Lineage-Specific Duplications of Alkaloid Biosynthetic Genes

Because of the medicinal and pharmaceutical importance of *D. stramonium* tropane alkaloids, we wanted to examine our genome assembly and annotation for evidence of changes in copy number of tropane alkaloid biosynthetic genes. The tropane alkaloid biosynthesis pathway is fairly well characterized and most of the enzymes

responsible for the creation of *Datura spp.* predominant tropane alkaloids have already been elucidated (Kohnen-Johannsen & Kayser, 2019).

Interestingly, in the lineage-specific duplication events for *D. stramonium*, we detected significant enrichment for the polyamine biosynthetic processes GO term (Fig. 3.2E, GO:0006596, $p=1.9x10^{-4}$). Polyamines, such as putrescine, are precursor molecules for the production of tropane alkaloids, which accumulate to high levels in *D. stramonium* (De-la-Cruz et al., 2020; Kohnen-Johannsen & Kayser, 2019). Looking into the gene trees inferred by OrthoFinder2, we also detected lineage-specific duplications in *D. stramonium* of the genes encoding the enzyme tropinone reductase I (TRI) (Fig. 3.3B). Tropinone reductases function on tropinone to shunt the biosynthetic pathway toward pseudotropine and eventually calystegines in the case of tropinone reductase II (TRII) or toward tropine and the eventual production of the pharmacologically important alkaloids atropine and scopolamine in the case of tropinone reductase I (TRI) (Kohnen-Johannsen & Kayser, 2019). These duplications were not observed in *S. lycopersicum* or *C. annuum*.

One further lineage-specific duplication appears to have occurred in *D. stramonium* for the biosynthetic enzyme hyoscyamine 6 β-hydroxylase (H6H, Figure 3C). This enzyme converts hyoscyamine into scopolamine, a more potent and fast-acting hypnotic (Alizadeh et al., 2014). The two paralogous H6H loci in *D. stramonium* are arranged in a tandem array approximately 2kb apart and share nearly 80% amino acid sequence identity. Our OrthoFinder search resolved two *P. axillaris* genes in the same orthogroup as the tandem *D. stramonium* H6H genes, but failed to find loci from any of the other 11 species. One of the petunia genes appears to encode a fusion protein of two tandemly-arrayed H6H genes transcribed in-frame, which we split apart for our phylogenetic analysis to examine their evolutionary relationships independently.

Interestingly, the two regions of the fused petunia genes do not appear most closely related to each other in our phylogenetic analysis, as could be expected if they were the result of a recent tandem duplication followed by a fusion. Instead, the C-terminal region of the fusion appears quite similar to the unfused gene. Other solanaceous genes identified via a BLAST search group separately from the petunia and *D. stramonium* genes recovered by OrthoFinder, suggesting that these might not be true orthologs. Broader taxon sampling from genera that are known to produce scopolamine such as *Atropa, Scopolia*, or *Hyoscyamus* could clarify this evolutionary relationship. Taken together, the duplications of two structural enzymes in the scopolamine biosynthetic pathway of *D. stramonium* confirm the importance of tropane alkaloid production in this *D. stramonium*.

Transposable Element Inventory

We applied the Extensive *de novo* TE Annotator (EDTA) pipeline to achieve a more comprehensive and detailed inventory of transposable elements across this genome (Ou et al., 2019). This pipeline annotated an additional ~4% of the ungapped genome as transposable elements or repeats compared to the RepeatModeler pipeline alone. A summary of repetitive elements delineated by superfamilies as defined by Wicker et al. is presented in Table 3.2 (Wicker et al., 2007). Over half of the annotated repetitive elements belong to the *Gypsy* superfamily of Long Terminal Repeat (LTR) retrotransposons, with unclassified LTRs and the *Mutator* superfamily of Terminal Inverted Repeat (TIR) DNA transposons making up the next two most numerous classes of repetitive elements. *Gypsy*-type LTRs also make up roughly a third of the genomes of several sequenced *Solanum* species, and the repetitive content of the genomes of *Capsicum annuum* and

C. chinense are also approximately half *Gypsy*-type LTRs (Bolger et al., 2014; Hosmani et al., 2019; Kim et al., 2014; Razali et al., 2018).

Impacts of Tissue Culture-based Transformation

Previously we developed a tissue culture regeneration protocol for *D. stramonium* and used this to demonstrate the first stable transgenic transformants in the genus by introducing a green fluorescent protein (GFP) transgene (Rajewski et al., 2019). Because all transgenic transformation protocols for solanaceous plants developed thus far require a tissue culture phase, we sought to characterize the potential genomic and transcriptome impacts of this process.

We resequenced the genomes of three plants derived from GFP-transformants in the 2019 study. All three individuals were derived from the same transgenic event and were propagated through single-seed descent of selfed plants for three generations after tissue culture. The estimated genome coverage for resequencing varied from 2-5x among the three plants. Overall, we detected over two million variants among the three transformants, with over half of the variants being SNPs. Indels ranged in size from 28bp deletions to 22bp insertions, but over 66% of indels were only ±1bp. The vast majority of these polymorphisms were intergenic (74.3%, Table 3.3) with an additional 21.8% appearing proximally (±5kb) upstream or downstream of coding regions. Only 1% of polymorphisms were present within exons and 2.8% were present in introns or at splice junctions. Of the exonic variants, about one third produced silent mutations while 64% created missense mutations. Nonsense mutations only accounted for 2.2% of variants.

Although this analysis did not reveal strong evidence of duplicated genomic regions, we wanted to confirm that the transformants were still euploid diploids following

tissue culture (Hang & Bregitzer, 1993). We used Smudgeplot to estimate the ploidy of each resequenced transformant from kmer frequencies (Ranallo-Benavidez et al., 2020). Because sequencing depth varied across the three samples, we created smudgeplots with four kmer lengths (13, 15, 17, and 19bp). While longer kmers are more specific and lead to a more robust analysis, their per-kmer coverage is lower and diminishes the power of the analysis at low sequencing coverage. Transformant #1 had the highest resequencing coverage and was determined to be a diploid regardless of the kmer length used. The two other resequenced transformants were assigned as diploids based on three of the four kmer lengths. Transformant #2 was determined to be a triploid with k=15, while Transformant #3 was determined to be a triploid with k=13. (Fig. 3.2)

Using all three transformants as replicates, we then conducted an mRNAseq experiment to look for potential differential expression of genes between the wild-type and transformed plants. With a FDR threshold of 0.01 and a $\log_2$ fold change threshold of 2, we were only able to detect 186 differentially expressed genes. Of these, 81 had lower expression in the GFP transformants compared to wild-type, and 105 had higher expression in the GFP transformants. We performed a GO term enrichment to determine if and to what extent the differentially expressed genes fell into distinct functional groups. The genes downregulated in the GFP transformants were slightly but significantly enriched for transport-related GO terms, specifically anion and organic acid transport (GO:0098656, GO:1903825, GO:1905039, GO:0006820, and GO:0009611). However only 1-2 genes fell into each of these partially overlapping categories. In contrast, the upregulated genes were generally enriched for regulatory GO terms, but spanned several regulatory terms from regulation of gene expression (GO:0010468) to regulation of nitrogenous compound

metabolic processes (GO:0051171). These regulatory GO terms each represented between 8 and 10 genes.

To explore the effects of mutations following tissue culture on gene expression, we attempted to correlate the polymorphisms present in the resequencing data with our mRNA-seq data. We reasoned that changes in gene expression following tissue culture could be due to tissue-culture-induced mutations in regulatory regions or gene body regions important for transcript stability or transcription efficiency. We used the program snpEff to describe the impact proximal polymorphisms might have on genes (Cingolani et al., 2012). The program assigns polymorphisms into 26 categories describing their magnitude, effect, and location. Three categories summarize the potential magnitude of the polymorphism as having a high, moderate, or low impact on the coding sequence, while a fourth summary category is used for polymorphisms in potential regulatory (non-coding) regions. The remaining 22 categories further define the polymorphisms' likely consequence, for example frameshifts, splice donor variants, synonymous variants, etc. Using a hypergeometric test, we asked if the differentially expressed gene set was enriched for any of the snpEff polymorphism categories compared to the rest of the genes in the genome. Two of the summary categories (impact_LOW and impact_MODERATE) and two more detailed categories (effect_conservative_inframe_deletion and effect_synonymous_variant) showed enrichment with p-values less than 0.05. For each of the four categories, we performed a linear regression, regressing the $\log_2$ fold change of expression on either the number of polymorphisms in each gene or simply the presence/absence of polymorphisms in each gene. In no case were any of these categories sufficient to explain changes in gene expression between the wild-type plants and GFP transformants ($p >> 0.1$).

In a separate attempt to explain the differentially expressed genes following tissue culture, we leveraged the transposable element inventory of the sequenced genome to look for correlations between differentially expressed genes and nearby transposable elements. Most genes contained both proximal (<5kb up- or downstream) as well as internal transposable elements. Here again, we performed a series of linear regressions, regressing $\log_2$ fold change of expression on distance to the nearest transposable element.

We partitioned the data into several subsets for regression analyses. This included removing or including transposable elements located between the start and stop codons; only considering upstream transposable elements; using absolute distance from the gene body or signed distance from the gene body; and including all differentially expressed genes (DEGs), only upregulated DEGs, or only downregulated DEGs. All regressions were run both breaking the dataset apart by transposon superfamily and considering all transposable elements together. In all cases, distance to the nearest transposable element was capped at 5kb. There were very few elements located further than this distance, and these rare data points had very high leverage in the regression analyses potentially inflating trends and p-values.

When combining all transposable elements regardless of superfamily, we failed to see a statistically significant ($p<0.05$) dependence between any of the distance metrics and $\log_2$ fold change of expression. We also saw no statistically significant dependence when examining helitrons (DHH) and all types of retrotransposons (RIX, RLC, RLG, and RLX). Additionally, looking only at upstream transposable elements likewise failed to show any statistical significance associations between $\log_2$ fold change of expression and distance. However, for several superfamilies of DNA transposons as well as for

uncategorized transposable elements, we detected a statistically significant association between absolute distance to the nearest transposable element and $\log_2$ fold change. This association was present when partitioning the data into up- or downregulated DEGs such that, as distance to the nearest transposable element in the given superfamily increases, the magnitude of differential expression also increases (Fig. 3.1).

Discussion

For well over a century, geneticists, biochemists, and evolutionary biologists have studied *Datura stramonium* for its interesting fruit and leaf phenotypes, startling alterations in ploidy, and useful production of various alkaloids (Blakeslee et al., 1922; Blakeslee & Avery, 1917, 1919; De-la-Cruz et al., 2020). Here we continue this advance by providing a draft reference genome, demonstrating lineage specific duplications of alkaloid biosynthetic pathways, and characterizing the impacts of a recently developed transformation technique for this model plant.

Our reference genome assembly corresponds well to the previously estimated 2Gbp haploid genome size of *D. stramonium* based on flow cytometry and contains a very high percentage of BUSCO complete and single copy genes (Table 3.1) (Kubešová et al., 2010). Combined with the associated draft annotation of protein coding genes, this resource will better enable future genetic and genomic studies in this species and perhaps allow us to revisit unanswered morphological and evolutionary questions from classical studies.

Our annotation suggests a higher than expected number of protein-coding genes in the genome of *D. stramonium*. This estimate could represent an accurate count of true protein coding loci in the genome, however, protein-coding gene number in other diploid

members of Solanaceae is roughly 35,000 (Barchi et al., 2019; Bombarely et al., 2016; Hosmani et al., 2019; Kim et al., 2017; Xu et al., 2017). Because our kmer- and Ks-based analyses are consistent with *D. stramonium* being a eudiploid plant lacking evidence of substantial lineage-specific gene family expansions since its split from *S. lycopersicum*, we expect a similar number of genes in *D. stramonium*. Our mRNA-seq data from leaf tissue can provide support for 33,629 genes, and including the publically available mRNA-seq data used to train our gene prediction software only provided support for an additional 1,841 genes. We expect that future studies of other tissues and conditions will incrementally increase evidence for this smaller set of well-supported genes. Additionally, improved contiguity of the assembly in the future is likely to lead to more thorough masking of the repetitive DNA content from the gene annotation. Indeed, a similar pattern was seen with the eggplant (*Solanum melongena*) genome, where the draft assembly estimate of 85,446 genes was revised downward significantly to 34,916 as later assemblies improved contiguity and mRNA-seq sampling (Barchi et al., 2019; Hirakawa et al., 2014).

In terms of repetitive DNA content, our assembly suggests that *D. stramonium* is unremarkable amongst other Solanaceae with its 61% repetitive DNA, comparable to Petunia and *N. benthamiana*, but lower than tomato, pepper, and several other tobacco species. Over half of the annotated repetitive elements belong to the *Gypsy* superfamily of Long Terminal Repeat (LTR) retrotransposons (Table 3.2).This results is in keeping with our knowledge of other closely related plants. Indeed *Gypsy*-type LTRs similarly make up about a third of the genomes of several sequenced *Solanum* species (Bolger et al., 2014; Razali et al., 2018; Tomato Genome Consortium, 2012). The repetitive portion of the *Capsicum annuum* and *C. chinense* genomes are also approximately half *Gypsy*-type LTRs; however, these genomes contain more repetitive DNA overall (Kim et al., 2014).

Within the Solanaceae family but outside the Solanoideae subfamily, which contains *Datura stramonium*, *Capsicum spp.* and *Solanum spp.*, *Gypsy* superfamily LTRs also make up much of the repetitive DNA. This superfamily alone comprises between one third and one half of the genomes of several *Nicotiana* species (Xu et al., 2017). *Gypsy*-type LTRs are the most abundant superfamily of repetitive elements in the *Petunia axillaris* genome as well; however, *Copia*-type LTRs make up a nearly equal share of the genome, unlike other solanaceous species (Bombarely et al., 2016).

These results should be interpreted with two caveats in mind. First, our assembly contains approximately 24% ambiguous bases, representing gaps of known size but unknown sequence between contigs. Precisely because our sequencing methods could not resolve these gaps, it is very likely that they correspond to highly repetitive regions of the genome such as centromeres, rDNA loci, or intergenic regions with nested/tandem transposable element insertions. Therefore resolving these gaps with additional long-read sequencing technologies in the future is likely to alter the inventory of transposable elements and refine our protein-coding gene annotation. Second, our scaffolds are not yet assigned to chromosome-scale linkage groups or pseudomolecules. Our current assembly comprises over 200,000 scaffolds; however, based on previous karyotyping studies and the assumed conservation of base chromosome number in the subfamily Solanoideae, we expect *D. stramonium* to have 12 haploid chromosomes (x=12) (Blakeslee et al., 1922; Särkinen et al., 2013). In the future, additional long-read sequencing, optical mapping, proximity ligation sequencing, or other techniques could achieve a more contiguous, chromosome-scale assembly. Such a chromosome-scale assembly would also provide better evidence for genome size and would necessarily affect repetitive DNA content, transposable element annotation, and gene annotation. The

resolution of full-scale chromosomes would also enable a more precise characterization of structural variation following tissue culture. Overall however, our kmer-based Smudgeplot analysis, BUSCO duplicate genes score, and paralog Ks plots are all consistent with our reference genome deriving from a typical eudiploid plant and support its use in future genetic and genomic studies.

One of our key findings was the lineage-specific duplications in two tropane alkaloid biosynthetic genes. Early in this pathway, the enzyme tropinone reductase I (TRI) acts to shunt the production toward tropine and the derivative tropane alkaloids by competing with tropinone reductase II (TRII), which produces pseudotropine leading eventually to calystegine alkaloids (Hashimoto et al., 1992; Kohnen-Johannsen & Kayser, 2019). We show evidence for a lineage-specific duplication of TRI in *D. stramonium* that is not shared with the other members of the Solanoideae subfamily of Solanaceae that we examined (Fig. 3.3B). Following the formation of tropine, several other biochemical reactions can eventually lead to the production of hyoscyamine, a pharmaceutically important tropane alkaloid in its own right. However, many *Datura spp.* are known to accumulate the hyoscamine derivative, scopolamine, as the primary tropane alkaloid instead (Parr et al., 1990). At this step in the biosynthesis pathway, we discovered the second lineage-specific gene duplication in *D. stramonium*, with a tandem duplication of hyoscyamine 6β-hydroxylase (H6H, Figure 3C). This gene was successfully targeted in a previous effort to increase tropane alkaloid content in *Atropa belladonna (Xia et al., 2016).* Our initial search for orthologs of these genes with OrthoFinder2 found two genes in *P. axillaris* and none in the other 11 species included in this study. Unlike the tandem duplicates in *D. stramonium*, the two petunia genes are located on different scaffolds. Remarkably however, the ortholog *Peaxi162Scf00075g01545* does appear to encode a

fusion protein of two tandemly transcribed in-frame H6H genes, suggesting that perhaps the evolutionary history of these H6H genes in Solanaceae is more complex. Our BLAST search did recover similar proteins among other solanaceous species, but these grouped distinctly in our phylogenetic analysis. This arrangement could be an artifact of narrow taxonomic sampling or possibly independent derivations from an ancestral protein of unknown function. Importantly, our dataset does not include other solanaceous species with notable production of tropane alkaloids, as we were unable to find assembled genomes for any of these species. Including sequences from genera such as *Atropa*, *Scopolia*, or *Hyoscyamus* could shed more light on the evolution of this enzyme and clarify this unlikely grouping of *Petunia* and *Datura* protein sequences. Broader sampling is likely to clarify the history of gene duplication and loss that could have led to the phylogenetic arrangement we observed.

Our previously published protocol for transformation of *D. stramonium* enabled more thorough functional genetic studies, but also carried with it the possibility of genomic changes induced by tissue-culture itself (Rajewski et al., 2019). To better characterize these potential changes, we resequenced the genomes of three plants descended from a transformant in the original study. We detected several million polymorphisms (SNPs and indels) among the resequenced plants compared to the reference genome. This amounts to 1.16×10^{-3} mutations per site, which is much higher than the estimated mutation rates following tissue culture in either *A. thaliana* (between 4.2×10^{-7} and 24.2×10^{-7} mutations per site) or *O. sativa* (5.0×10^{-5} mutations per site) (Jiang et al., 2011; Zhang et al., 2014). Our analysis pipeline took PCR duplicates from library preparation and potential sequencing errors into account, so we expect that our analysis is detecting bona fide polymorphisms

between the transformants and the reference genome. Our plants were allowed to self-pollinate prior to reference genome sequence and between tissue culture and genome resequencing, however our methods cannot rule out that some of these polymorphisms are due to standing heterozygosity in the resequenced plants not captured by the reference genome. Importantly, this mutation rate from tissue culture is not contextualized with a background mutation rate from untransformated *D. stramonium*. We expect the mutation rate following tissue culture to be higher than normal, but we also know that mutation rates are not uniform across species or even cultivar boundaries (Jiang et al., 2011; Miyao et al., 2012; Zhang et al., 2014). Our estimate could be further refined with long-term mutation-accumulation studies (Weng et al., 2019).

Although our transformants accumulated a large number of mutations compared to the reference genome, their impact appears low. The mutations following tissue culture were overwhelmingly found in intergenic regions of the genome. Only 27,000 exonic mutations are present across over two million mutations in the three individuals. However, nearly two thirds of these exonic mutations are not silent and could potentially affect protein function, secondary structure, etc. Notably, we did not examine changes to the epigenome, which is frequently connected with aberrant phenotypes of transformants (Filipecki & Malepszy, 2006; Joyce et al., 2003; Ko et al., 2018). It is also possible that mobilization of transposable elements is responsible for some alteration in the transformants (Huang et al., 2009; Kobayashi et al., 2004). This movement along with other large scale structural changes to the genome have been observed following tissue culture; however, we were unable to successfully apply the computational tools to detect this given the fragmentation of our assembly (Lu et al., 2017; Naito et al., 2006, 2009).

Despite these unknowns, it is encouraging that when we examined the transcriptomic impacts of tissue culture on our transformants, the results were negligible. Using our thresholds of differential expression (FDR<0.01 and $\log_2$ fold change>2), we were only able to call 186 genes as differentially expressed between the transformed and untransformed plants. We did detect significant GO term enrichment for certain classes of genes among the 186 differentially expressed genes, including regulatory terms and transmembrane transport. Our attempts to explain this small number of differentially expressed genes through correlation with polymorphisms or transposons did not produce robust results though some weak association between magnitude of differential expression and distance to certain DNA transposons superfamilies was present and has been remarked on by other studies as well (Eichten et al., 2012; Hollister & Gaut, 2009). Overall it seems that other factors not captured by our study could be behind the differential expression of this subset of genes.

Conclusions

Our assembled and annotated 2 gigabasepair draft genome of this plant, is the first in the genus and will be an excellent resource for others working on functional genomic studies in this system. Future work involving long-read sequencing technologies should improve the contiguity and annotation of this draft. Using this new resource along with mRNAseq and genome resequencing, we show that following tissue culture, mutation rates of transformed plants are quite high, but do not have a drastic impact on gene expression.

Abbreviations

bp: Basepair, BUSCO: Benchmarking University Single-Copy Orthologs, DEG: Differentially Expressed Gene, DHH: Helitron, DMM: *Maverick*-type DNA Transposon, DTA: *hAT*-type DNA Transposon, DTC: *CACTA*-type DNA Transposon, DTH: *PIF-Harbinger*-type DNA Transposon, DTM: *Mutator*-type DNA Transposon, DTT: *Tc1-Mariner*-type DNA Transposon, FDR: False Discovery Rate, GFP: Green Fluorescent Protein, GO: Gene Ontology, H6H: Hyoscyamine 6 β-hydroxylase, indel: Insertion-Deletion, Ks: Synonymous Substitution, LTR: Long Terminal Repeat, mRNA: Messenger RNA, rDNA: Ribosomal DNA, RIX: Unidentified LINE Retrotransposon, RLC: *Copia*-type LTR Retrotransposon, RLG: *Gypsy*-type LTR Retrotransposon, RLR: *Retrovirus*-type LTR Retrotransposon, RLX: Unidentified LTR Retrotransposon, SNP: Single Nucleotide Polymorphism, T-DNA: Transfer DNA, TE: Transposable Element, TIR: Terminal Inverted Repeat, TRI: Tropinone Reductase I, TRII: Tropinone Reductase II, XXX: Unknown Transposon

Methods

Plant Material

Growth Conditions

For genome sequencing, wild-type *Datura stramonium* seeds were obtained in 2013 from J. L. Hudson Seedsman (La Honda, California, USA), sown directly on soil, and grown under greenhouse conditions at the University of California, Riverside for three generations with self pollination to increase homozygosity prior to genome sequencing.

For genome resequencing and gene expression analyses of transgenic plants, we used GFP-transgene harboring seeds previously described in Rajewski et al. (Rajewski et al., 2019). These seeds correspond to the second generation seed from individual T_1-4, making these seeds three generations removed from tissue culture. We selected progeny

of T_1-4 based on its brighter GFP fluorescence than that of its siblings in order to aid screening. To increase germination efficiency, we dissected away the outer seed coat of these seeds. All plants for gene expression analyses and genome resequencing were maintained at 22°C for 24 h under 100 μmol m^{-2} s^{-1} light conditions at a 16h light and 8h dark photoperiod.

For a wild-type gene expression analysis, we selected sibling seed of the genome sequenced individual, dissected the seed coat away, and germinated them under the same conditions as the GFP-transgenic seeds.

Nucleic Acid Isolation

For short read sequencing, we isolated DNA from a single developing leaf of one wild-type, greenhouse-grown *Datura stramonium* plant described above using the E.Z.N.A Plant DNA Kit (Omega Bio-tek, Norcross, GA) according to the manufacturer's instructions, and quantified its purity and concentration using a biospectrometer (Eppendorf AG, Hamburg, Germany). In order to isolate high molecular weight DNA for Oxford Nanopore sequencing, we used a CTAB DNA extraction with several modifications to reduce shearing of genomic DNA (Doyle & Doyle, 1987; Harkess, 2017). The DNA was stored at -70 until needed for library construction.

For gene expression analyses, we collected one immature leaf (~3cm in length) each from three wild-type and three plants harboring the GFP transgene. We snap froze this tissue in liquid nitrogen, ground each sample using steel BBs in a Retsch MM400 mixer mill (Haan, Germany), and isolated RNA with the RNeasy Plant Mini Kit (QIAGEN, Hilden, Germany). RNA isolation proceeded according to the manufacturer's protocol

except that the lysis step of this protocol was modified to use buffer RLC instead of RLT and supplemented with 2.5% (w/v) polyvinylpyrrolidone (PVP). We removed DNA contamination with an on-column RNAse-Free DNAse kit (QIAGEN, Hilden, Germany) according to the manufacturer's protocol. The UCR Genomics Core assessed the integrity of the isolated RNA using an Agilent 2100 Bioanalyzer. We stored the material at -70ºC.

Reference Genome Sequencing

We used the SeqOnce Rapid DNA-seq preparation Kit (Beta Version 4.0d, SeqOnce Biosciences, Pasadena, CA) to prepare a DNA sequencing library. This library was sequenced across two partial Illumina NovaSeq 2x150bp runs at the University of California San Francisco Functional Genomics Core Facility, and produced 165Gbp of sequencing data, corresponding to ~100x haploid genome coverage. For long-read Oxford Nanopore sequencing, we used the high molecular weight DNA (greater than 28kb) and the Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore, UK) to create a 1D sequencing library. We sequenced this on a MinION flow cell R9.4 to generate approximately 13Gbp of data (~9x haploid genome coverage). Read sizes ranged from 330kb to 500b with a mean of 9.4kb.

Reference Genome Assembly and Annotation

All scripts used to assemble and annotate this reference genome are available in a public Github repository (https://github.com/rajewski/Datura-Genome).

We first created several short-read only assemblies using ABySS (v2.0.2) with odd kmer sizes from 33-121bp, but ultimately selected k=101 as the optimal kmer size based

on the assembly's BUSCO score using the embryophyta version 9 lineage dataset (Jackman et al., 2017; Simão et al., 2015).

Following base calling by Guppy, we error-corrected the Nanopore reads using LoRDEC (v0.9) (Salmela & Rivals, 2014). We then used the optimal ABySS assembly for several iterations of scaffolding, gap-filling, and polishing using LINKS (v1.8.4), RAILS (v1.5.1), and ntEdit (v1.3.0), respectively (Warren, 2016; Warren et al., 2015, 2019). For LINKS scaffolding, we selected a relatively high kmer size of 19bp because we were using error-corrected Nanopore reads. We scaffolded with insert sizes of 750bp, 1kb, 5kb, 10kb, 15kb, 20kb, 30kb, 40kb, 60kb, 70kb, 80kb, 90kp, and 100kb. Gap filling with RAILS also used the error-corrected LoRDEC reads. Polishing with ntEdit was run several times after each scaffolding or gap-filling step until the number of edits stabilized. The kmer size for ntEdit was 50bp.

Prior to gene annotation, we used RepeatModeler (v1.0.11) and RepeatMasker (v4-0-7) to generate and soft mask a preliminary set of repetitive elements in the assembled genome (Smit et al., 2013; Smit & Hubley, 2008). This set of repetitive elements was excluded from the subsequent gene annotation.

We applied the funannotate pipeline (v1.6.0) to annotate the assembled genome for protein coding genes and tRNAs (J. M. Palmer, 2019). Funannotate is a wrapper for several evidence-based and *ab initio* gene prediction softwares but also includes convenience scripts to simplify submission of genome annotations to data repositories such as NCBI. To train the gene predictors, we provided publicly available RNA sequencing data from NCBI SRA accession SRR9888534, along with the *D. stramonium* reads from medplantrnaseq.org, and mRNA-seq reads generated for the differential gene expression analyses (below). Following the training step, funannotate ran AUGUSTUS

(v3.3), GeneMark-ETS (v4.38), SNAP, and GlimmerHMM (v3.0.4) (Korf, 2004; Lomsadze et al., 2005; Majoros et al., 2004; Stanke et al., 2006). Funannotate combined these gene prediction outputs with alignments of transcripts, generated by Trinity (v2.8.4) and PASA (v2.3.3), and protein evidence and passed them to EVidenceModeler (v1.1.1) which produced a well-supported annotation of protein coding genes (Grabherr et al., 2011; Haas et al., 2003, 2008). Separately, tRNAscan-SE (v2.0.3) searched for and annotated tRNA loci in the assembled genome (Chan & Lowe, 2019).

Once the annotation of protein coding genes and tRNA loci was completed, we used the Extensive *de novo* TE Annotator (EDTA) pipeline to create a more thorough annotation of TIR, LTR, and helitron transposable elements (Ou et al., 2019). This analysis made use of the gene annotation information to remove potentially protein coding loci from the transposable element inventory.

We used GetOrganelle (v1.7.1) to assemble both organellar genomes (Jin et al., 2020). For the plastid genome, we used the previously published *D. stramonium* plastid assembly (GenBank accession NC_018117) as an alignment seed (Yang et al., 2014). To annotate genes as well as the large and small single copy regions and inverted repeat regions, we used GeSeq (Tillich et al., 2017). For the mitochondrial genome, we used the *S. lycopersicum* mitochondrial genome (Genbank accession NC_035963) as the seed. To determine the similarity to the reference plastid genome, we aligned with the full-length plastid genomes with MAFFT (Katoh & Standley, 2013).

We deposited the raw sequencing reads used to assemble this genome in the SRA under NCBI Bioproject PRJNA612504. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JACEIK000000000.

Summaries of gene features and transposable elements proceeded with custom R scripts that are available in the public GitHub repository.

Ortholog Analyses

To determine orthologues among the 13 species, we used OrthoFinder2 (Emms & Kelly, 2019). This analysis included three members of the subfamily Solanoideae, *D. stramonium*, *Solanum lycopersicum*, and *Capsicum annuum*; two more distantly related members of Solanaceae, *Nicotiana attenuata* and *Petunia axillaris*; two non-solanaceous asterids, *Helianthus annuus* and *Lactuca sativa*; three rosids, *Vitis vinifera*, and *Arabidopsis thaliana*; two grasses, *Zea mays* and *Oryza sativa*, one non-grass monocot, *Asparagus officinalis*; and finally, the early-diverging angiosperm *Aquilegia coerulea (Badouin et al., 2017; Bombarely et al., 2016; Filiault et al., 2018; Harkess et al., 2017; Hosmani et al., 2019; Jaillon et al., 2007; Jiao et al., 2017; Kim et al., 2017; Lamesch et al., 2012; Ouyang et al., 2007; Reyes-Chin-Wo et al., 2017; Xu et al., 2017)*. For non-solanaceous species, we downloaded the reference proteomes and reference transcriptomes from Phytozome (v13). References for *D. stramonium* were generated in this study, those for *S. lycopersicum*, *C. annuum* and *P. axillaris* were downloaded from Sol Genomics Network (http://solgenomics.net/), and those for *N. attenuata* were downloaded from the *Nicotiana attenuata* Data Hub (http://nadh.ice.mpg.de/NaDH/).

For gene tree construction, we used either the loci from the OrthoFinder2 clustering, or, in the case of H6H, added additional loci based on BLAST searches with OrthoFinder2 output protein sequences as queries (Altschul et al., 1997). We then aligned these protein sequences with MAFFT (v7.471) and constructed phylogenetic trees with

RAxML-NG (v0.9.0) using the JTT+Γ+I model and 1000 bootstraps (Katoh & Standley, 2013; Kozlov et al., 2019).

For Ks estimates between and within *D. stramonium* and *S. lycopersicum*, we used the wgd software suite's tools for all-vs-all protein searches, MCL clustering, and Ks distribution calculation (Zwaenepoel & Van de Peer, 2019). We included the options *--nostrictcds* and *--ignorestop* during the all-vs-all protein searches to avoid various formatting issues with the publically available transcriptome sequence files. In the Ks distribution calculations, we also passed a proteome sequence file instead of relying on automatically translated transcriptomes. We plotted output data from the Ks distributions using a custom R script available on our public GitHub repository. To obtain estimates of constituent Ks peaks within the Ks distributions we also used the *wgd mix* program's Bayesian Gaussian mixture model function to decompose the distributions, determine peak Ks values, and Ks peak weights.

For lineage specific duplication events in *D. stramonium*, *S. lycopersicum*, and *A. thaliana*, we conducted GO enrichment analyses of duplicated genes using a custom R script. For consistency, this script used custom GO annotations for the proteome of each of the three species, which we generated using InterProScan (v5.45-80.0) (Jones et al., 2014).

We used custom R scripts with help from the phytools and ggtree packages to plot and annotate phylogenetic trees (Revell, 2012; Yu, 2020).

Genome Resequencing and Polymorphism Analysis

The UCR Genomics Core constructed DNA-sequencing libraries for genomic DNA from the three GFP transgene-containing plants using the NEBNext Ultra II FS DNA Library Prep Kit for Illumina and sequenced them to approximately 5x haploid genome coverage in a 2x75bp Illumina NextSeq run. The raw DNA-seq reads were deposited in the SRA under BioProject PRJNA648005.

We mapped the reads for each plant back to the reference genome using BWA MEM, then removed duplicates and flagged discordant or split reads with SAMBLASTER (Faust & Hall, 2014; Li & Durbin, 2009). We then used FreeBayes and LUMPY as implemented by SpeedSeq to call SNPs and structural variants, respectively, between the reference genome and the resequenced transformants (Chiang et al., 2015; Garrison & Marth, 2012; Layer et al., 2014). Subsequently, we applied both snpEff and bcftools to summarize the variants detected (Cingolani et al., 2012).

Gene Expression Analysis

We prepared mRNA-sequencing libraries from RNA of the GFP transgene-containing plants and three wild-type plants using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina, and sequenced these on the same 2x75bp Illumina NextSeq run as the DNA sequencing libraries. This produced approximately 33 million reads for each plant. The raw RNA-seq reads were deposited in the SRA under BioProject PRJNA648005.

The demultiplexed RNA-seq data were trimmed with TrimGalore. The trimmed reads were mapped to the assembled reference genome using STAR (v2.5.3a) in a single

pass using splice junctions annotated from the reference genomes (Dobin et al., 2013; Krueger, 2012). We then passed these counts directly into DESeq2 to identify differentially expressed genes between the two genotypes of plants (Love et al., 2014). A list of these genes is provided in an additional file (Additional File 1) and the script used to perform the differential expression analysis is included in our public GitHub repository. We generated a subset of genes with proximal transposable elements using our transposable element annotation and bedtools intersect. A list of these genes and the distance to the nearest transposable element is provided in an additional file (Additional File 2). All correlational analyses between gene expression and polymorphisms or transposable elements, including linear regressions and hypergeometric tests were conducted, summarized, and plotted using custom R scripts available in our public GitHub repository.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated during the current study are available under the NCBI Bioproject accessions PRJNA612504 and PRJNA648005. Additionally, *D. stramonium* RNA-seq files used for genome annotation are available under NCBI SRA accession SRR9888534

and at https://medplantrnaseq.org/. Scripts to analyze this data are available in a GitHub repository at https://github.com/rajewski/Datura-Genome.

Competing interests

The authors declare that they have no competing interests.

Funding

Funding for material and personnel was provided by a National Science Foundation grant to AL (IOS 1456109). A Department of Education Graduate Assistance in Areas of National Need (GAANN) grant to the University of California, Riverside partially funded AR. Neither funding organization contributed to the design of the study, analysis, interpretation of data, or in writing this manuscript.

Authors' contributions

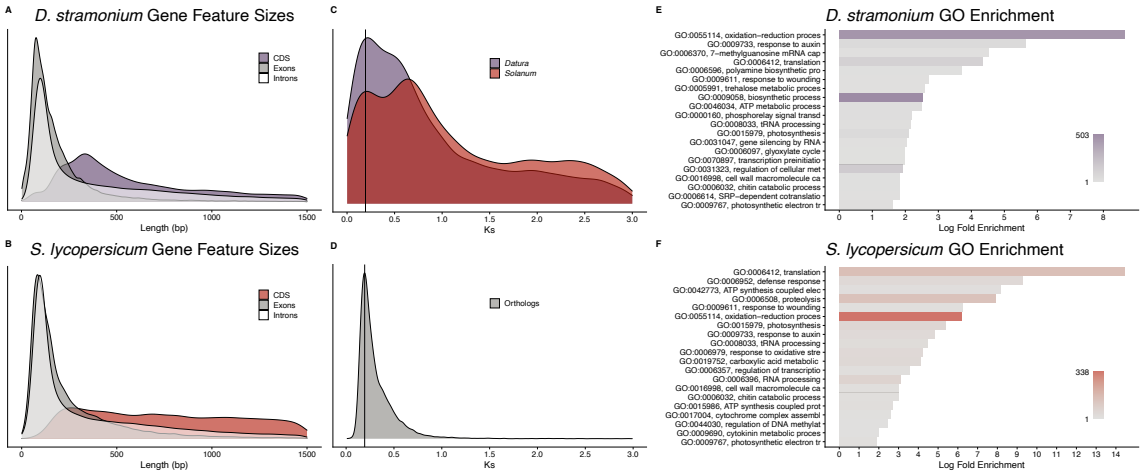
AR, JS, and AL designed the analyses. AR and DCH performed the analyses. AR, DCH, JS, and AL wrote the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgment

The sequencing was carried out at the Genomics Core at the Institute of Integrative Genome Biology (IIGB), University of California, Riverside.

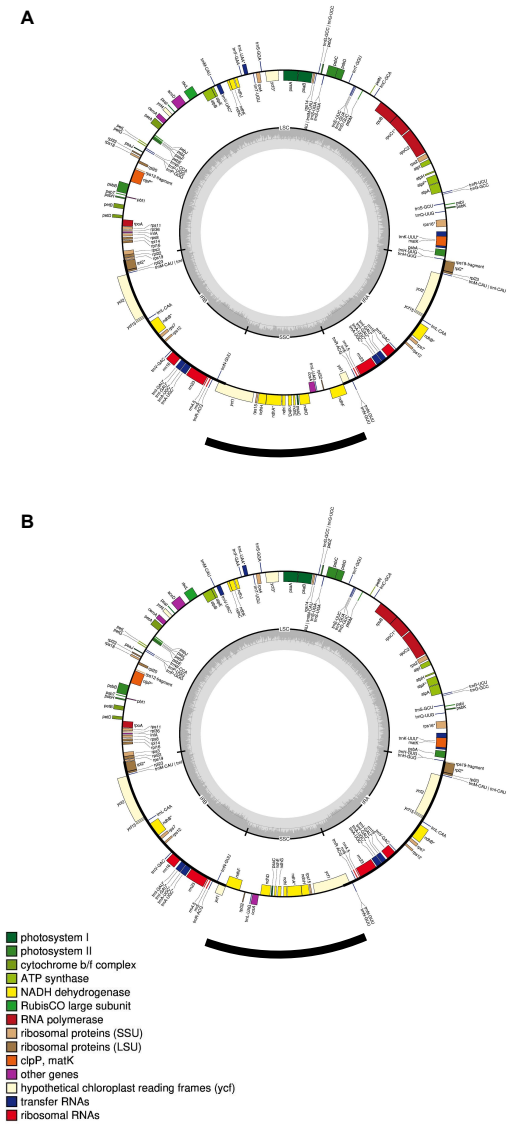
Figures and Tables

Figure 3.1 – Genome Annotation Features Summary



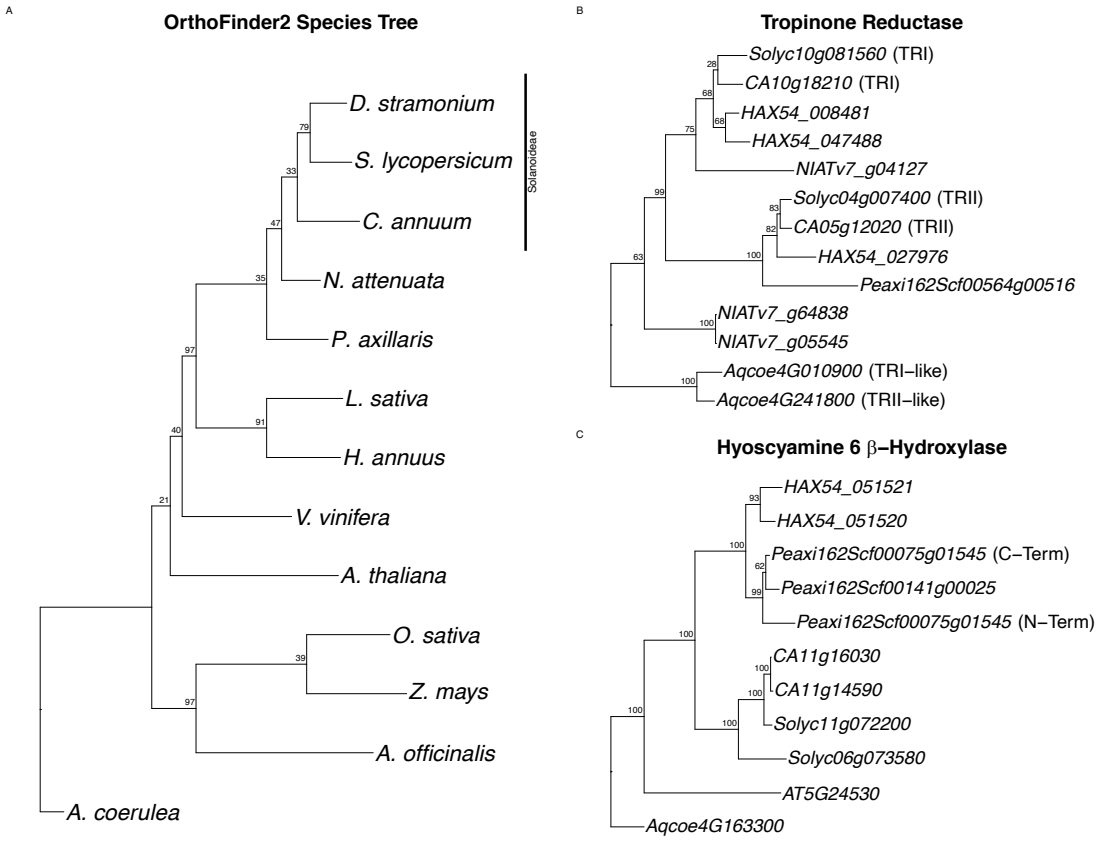
Summary of gene annotations. Density plots (A-B) of the sizes for total coding sequence lengths, individual exon lengths, and individual intron lengths for D. stramonium (A) and S. lycopersicum (B). Ks plots (C-D) showing the smoothed density of Ks values for paralogous genes (C) within D. stramonium (purple) or S. lycopersicum (red) and orthologous genes (D) between D. stramonium and S. lycopersicum. GO term enrichments for genes duplicated at the terminal branch of the phylogeny in Figure 3A for D. stramonium (E) and S. lycopersicum (F). GO term names have been truncated to fit available space, and bar colors correspond to the number of genes assigned to the given GO term, with a color scale shown in the lower right of each plot.

Figure 3.2 – Maps of Plastid Genome Conformations



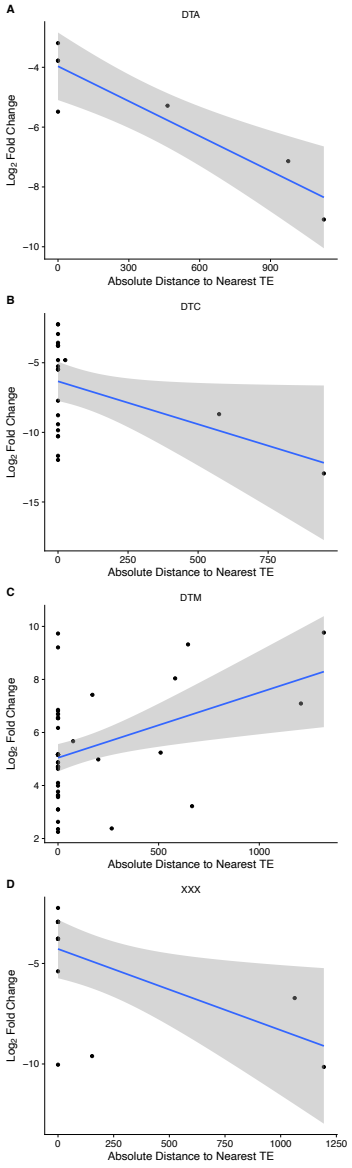
Assembled and annotated chloroplast genomes for D. stramonium showing the two inversion polymorphisms (A and B). The inverted small single-copy region is highlighted by the black sector below each circular genome. Annotated loci are plotted and labeled along the interior and exterior of the outermost circle. Loci are color coded by function as described in the legend in the lower left corner. The small single-copy, large single-copy, and inverted-repeat regions are delineated in the interior grey circles. Adapted from GeSeq output.

Figure 3.3 – Phylogenies of Species for Orthology Search and Selected Duplicated Genes



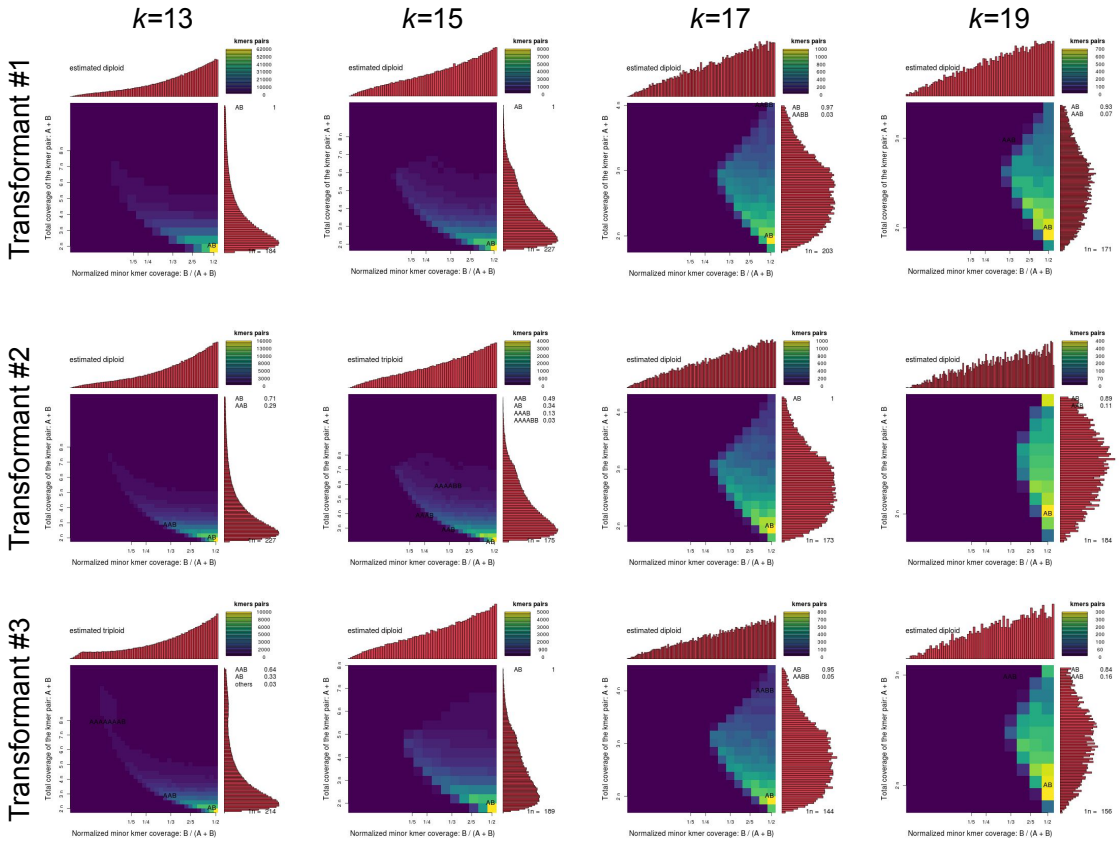
Phylogenetic trees representing (A) the species relationships inferred by OrthoFinder2, with the erroneously arranged Solanoideae clade highlighted by the black bar. The gene tree of putative tropinone reductase protein sequences (B), with previously published functional annotations of proteins in parentheses. The gene tree of putative hyoscamine 6 β-hydroxylase protein sequences (C) with the N-terminal and C-terminal domains of the petunia fusion protein annotated in parentheses.

Figure 3.4 – Correlations Gene Expression and Distance to Nearest Transposons



Linear relationship between absolute distance from the gene body to the nearest transposable element (in bp) and the log2 fold change of expression between the GFP transformants and wild-type plants. Downregulated differentially expressed genes for TIR/hAT (A), TIR/CACTA (B), and unknown (D) superfamilies, and upregulated differentially expressed genes for TIR/Mutator (C) superfamily elements.

Figure 3.5 – Smudgeplots of Resequenced Plants



Smudgeplots using increasing k-mer lengths 13, 15, 17, and 17 (columns) on GFP transformants 1-3 (rows). Each smudegeplot is a 2D heat map showing the total coverage for a pair of k-mers differing by 1 bp versus the coverage of the minor k-mer in the pair as a fraction of the total coverage for the pair. Estimated ploidies are shown in the top left corner of each graph and the probability of various ploidies is shown on the right.

Table 3.1 – Statistics for Genome Assemblies

Genome Assembly Statistics										
	Contigs (n)	**Scaffolds** (n)	**Ungapped Size** (Gbp)	**Gapped Size** (Gbp)	**Ambiguous Bases** (%)	**Contig N50** (kbp)	**Scaffold N50** (kbp)	**Largest Contig** (kbp)	**Largest Scaffold** (kbp)	**BUSCO Complete Genes** (%)
Short-Read Assembly	3,860,052	3,860,072	1.9	1.9	0.0	1.89	1.89	54	54	67.7
Scaffolding and Gap Filling[1]	3,662,748	3,458,610	2.2	2.7	19.2	5.42	103.51	232	1,489	94.0
Length Filtering[2]	436,743	232,605	1.6	2.1	24.6	11.29	164.61	232	1,489	94.0
Gap Filling and Polishing[3]	407,409	232,605	1.6	2.1	24.4	13.11	164.10	235	1,484	94.7

[1] With ONT long reads

[2] ≤ 500bp

[3] With Illumina short reads

Summary statistics for the reference genome of Datura stramonium. Final version of the genome is shown on the last line. Contig and scaffold are shown as a count. Ungapped and Gapped sizes represent the total length in gigabasepairs of the assembled genome without or with ambiguous bases (Ns), respectively, introduced during scaffolding. Ambiguous bases are shown as a percentage of the total gapped genome size. Contig and scaffold N50 are shown in kilobase pairs as are the largest contig and scaffold.

Table 3.2 – Ortholog And Gene Duplication Summary

			Orthofinder Genes		
	Number of genes	Lineage-specific Gene Duplication Events	Assigned Orthogroup	Unassigned	Lineage-specific
A. coerulea	30,023	9,138	26,367 (87.8%)	3,656 (12.2%)	4,030 (13.4%)
A. officinalis	27,395	7,595	23,222 (84.8%)	4,173 (15.2%)	4,259 (15.5%)
A. thaliana	27,416	8,820	24,864 (90.7%)	2,552 (9.3%)	3,493 (12.7%)
C. annuum	34,899	8,750	33,649 (96.4%)	1,250 (3.6%)	1,209 (3.5%)
D. stramonium	52,149	14,057	44,248 (84.8%)	7,901 (15.2%)	10,906 (20.9%)
H. annuus	52,243	18,131	44,759 (85.7%)	7,484 (14.3%)	7,802 (14.9%)
L. sativa	38,910	11,161	35,339 (90.8%)	3,571 (9.2%)	5,834 (15.0%)
N. attenuata	33,449	7,371	32,205 (96.3%)	1,244 (3.7%)	2,825 (8.4%)
O. sativa	39,049	9,256	31,424 (80.5%)	7,625 (19.5%)	5,802 (14.9%)
P. axillaris	32,928	6,490	30,962 (94.0%)	1,966 (6.0%)	1,767 (5.4%)
S. lycopersicum	34,075	4,830	29,990 (88.0%)	4,085 (12.0%)	1,671 (4.9%)
V. vinifera	31,845	7,671	27,416 (86.1%)	4,429 (13.9%)	2,606 (8.2%)
Z. mays	39,498	13,407	36,260 (91.8%)	3,238 (8.2%)	5,703 (14.4%)

Summary of ortholog search of 13 angiosperm taxa. Number of protein-coding genes used in the analysis, number of gene duplication events in this taxon not present at higher taxonomic levels, number of genes successfully assigned to an orthogroup (percent), number of genes not assigned to an orthogroup (percent), number of genes assigned to a lineage-specific orthogroup.

Table 3.3 – Transposon Summary by Family

Transposable Elements by Superfamily			
	Num. Elements	**Total Length** (bp)	**% of Genome**[1]
Class I: DNA Transposons			
RIX	1,621	558,395	0.03
RLC	47,835	28,841,830	1.81
RLG	746,734	514,699,841	32.25
RLR	2,812	1,217,204	0.08
RLX	319,025	160,715,085	10.07
Class II: Retrotransposons			
DHH	137,640	44,191,157	2.77
DMM	92	20,518	0.00
DTA	45,358	10,968,642	0.69
DTC	172,857	46,682,112	2.92
DTH	89,050	19,392,473	1.21
DTM	329,935	94,406,038	5.91
DTT	75,352	15,645,736	0.98
Unknown			
XXX	214,698	45,228,639	2.83
Total	2,183,009	982,567,670	61.55

[1] Based on ungapped genome size of 1.6Gbp

Superfamilies named according to Wicker et al, 2007

Transposable elements are broken down first by class then by superfamily (abbreviated according to Wicker et al, 2007).

Table 3.4 – Summary of Mutations in Resequenced Plants

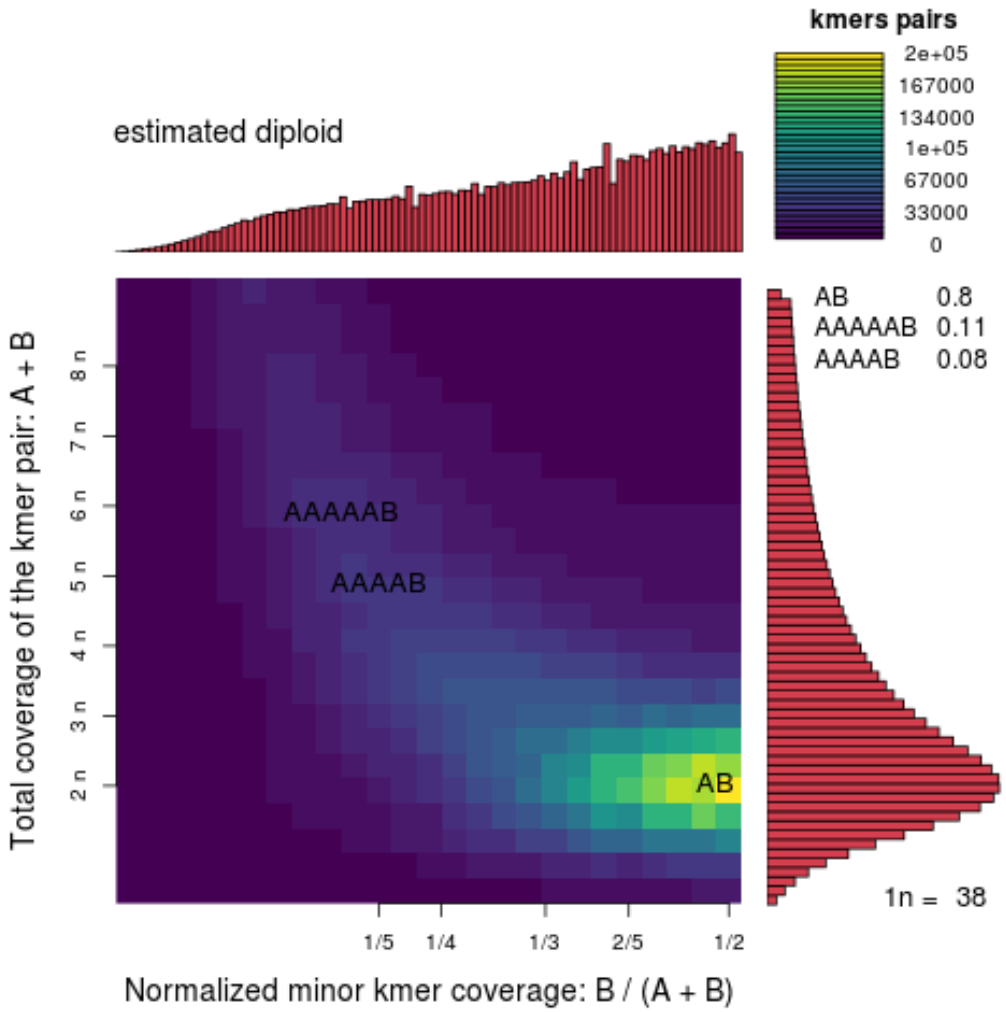
Polymorphisms by Location		
	Num. Polymorphisms	**Percent**
Upstream	330,511	12.15
Exon	27,168	1.00
Splice Junction	1,536	0.06
Intron	75,716	2.78
Downstream	261,937	9.63
Intergenic[1]	2,020,042	74.29
Other[2]	2,312	0.09
Total	2,719,222	100

[1] >5kb from gene body

[2] Including non-coding genes

Total polymorphisms in the three resequenced GFP transformants classified by their location with respect to specific gene regions or intergenic regions.

Supplemental Figure 3.1 – Smudgeplot of Genome Sequenced Plant



A smudgeplot of raw Illumina short reads from the reference genome plant prior to assembly. This plot is a 2D heat map showing the total coverage for a pair of k-mers differing by 1 bp versus the coverage of the minor k-mer in the pair as a fraction of the total coverage for the pair. Estimated ploidy is shown in the top left corner and the probability of various ploidies is shown to the right.

Additional File 1

Link

Additional File 2

Link

References

Alizadeh, A., Moshiri, M., Alizadeh, J., & Balali-Mood, M. (2014). Black henbane and its toxicity - a descriptive review. *Avicenna Journal of Phytomedicine*, *4*(5), 297–311.

Alonso, J. M., & Stepanova, A. N. (2003). T-DNA Mutagenesis in Arabidopsis. In E. Grotewold (Ed.), *Plant Functional Genomics* (pp. 177–187). Humana Press.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., Lelandais-Brière, C., Owens, G. L., Carrère, S., Mayjonade, B., Legrand, L., Gill, N., Kane, N. C., Bowers, J. E., Hubner, S., Bellec, A., Bérard, A., Bergès, H., Blanchet, N., … Langlade, N. B. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature*, *546*(7656), 148–152.

Barchi, L., Pietrella, M., Venturini, L., Minio, A., Toppino, L., Acquadro, A., Andolfo, G., Aprea, G., Avanzato, C., Bassolino, L., Comino, C., Molin, A. D., Ferrarini, A., Maor, L. C., Portis, E., Reyes-Chin-Wo, S., Rinaldi, R., Sala, T., Scaglione, D., … Rotino, G. L. (2019). A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. *Scientific Reports*, *9*(1), 11769.

Bardini, M., Labra, M., Winfield, M., & Sala, F. (2003). Antibiotic-induced DNA methylation changes in calluses of Arabidopsis thaliana. *Plant Cell, Tissue and Organ Culture*, *72*(2), 157–162.

Belling, J., & Blakeslee, A. F. (1922). The Assortment of Chromosomes in Triploid Daturas. *The American Naturalist*, *56*(645), 339–346.

Benfey, P. N., & Chua, N.-H. (1985). Combinatorial Regulation of Transcription in Plants. *P. K. Chanda, M. Ono, M. Kuwano, H. F. Kung, J. Bacteriol.*, *161*, 446.

Blakeslee, A. F. (1921). The Globe Mutant in the Jimson Weed (Datura stramonium). *Genetics*, *6*(3), 241–264.

Blakeslee, A. F. (1922). Variations in Datura Due to Changes in Chromosome Number. *The American Naturalist*, *56*(642), 16–31.

Blakeslee, A. F., & Avery, B. T. (1917). Adzuki Beans and Jimson Weeds: Favorable Class Material for Illustrating the Ratios of Mendel's Law—Actual Practice in Making Counts Is Necessary Before the Student Can Fully Grasp Modern Ideas of Heredity. *The Journal of Heredity*, *8*(3), 125–131.

Blakeslee, A. F., & Avery, B. T. (1919). Mutations in the Jimson Weed. *The Journal of Heredity*, *10*(3), 111–120.

Blakeslee, A. F., Belling, J., Farnham, M. E., & Bergner, A. D. (1922). A Haploid Mutant in the Jimson Weed, "Datura stramonium." *Science*, *55*(1433), 646–647.

Bolger, A., Scossa, F., Bolger, M. E., Lanz, C., Maumus, F., Tohge, T., Quesneville, H., Alseekh, S., Sørensen, I., Lichtenstein, G., Fich, E. A., Conte, M., Keller, H., Schneeberger, K., Schwacke, R., Ofner, I., Vrebalov, J., Xu, Y., Osorio, S., … Fernie, A. R. (2014). The genome of the stress-tolerant wild tomato species Solanum pennellii. *Nature Genetics*, *46*(9), 1034–1038.

Bombarely, A., Moser, M., Amrad, A., Bao, M., Bapaume, L., Barry, C. S., Bliek, M., Boersma, M. R., Borghi, L., Bruggmann, R., Bucher, M., D'Agostino, N., Davies, K., Druege, U., Dudareva, N., Egea-Cortines, M., Delledonne, M., Fernandez-Pozo, N.,

Franken, P., … Kuhlemeier, C. (2016). Insight into the evolution of the Solanaceae from the parental genomes of Petunia hybrida. *Nature Plants*, *2*(6), 16074.
Chan, P. P., & Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods in Molecular Biology* , *1962*, 1–14.
Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R., & Hall, I. M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, *12*(10), 966–968.
Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92.
Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., Soltis, P. S., Carlson, J. E., Arumuganathan, K., Barakat, A., Albert, V. A., Ma, H., & dePamphilis, C. W. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Research*, *16*(6), 738–749.
De-la-Cruz, I. M., Cruz, L. L., Martínez-García, L., Valverde, P. L., Flores-Ortiz, C. M., Hernández-Portilla, L. B., & Núñez-Farfán, J. (2020). Evolutionary response to herbivory: population differentiation in microsatellite loci, tropane alkaloids and leaf trichome density in Datura stramonium. *Arthropod-Plant Interactions*, *14*(1), 21–30.
De-la-Cruz, I. M., & Núñez-Farfán, J. (2020). The complete chloroplast genomes of two Mexican plants of the annual herb Datura stramonium (Solanaceae). *Mitochondrial DNA Part B*, *5*(3), 2829–2831.
Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , *29*(1), 15–21.
Doyle, J. J., & Doyle, J. L. (1987). *A rapid DNA isolation procedure for small quantities of fresh leaf tissue*. https://worldveg.tind.io/record/33886/
Dupin, J., & Smith, S. D. (2018). Phylogenetics of Datureae (Solanaceae), including description of the new genus Trompettia and re–circumscription of the tribe. *Taxon*, *67*(2), 359–375.
Eichten, S. R., Ellis, N. A., Makarevitch, I., Yeh, C.-T., Gent, J. I., Guo, L., McGinnis, K. M., Zhang, X., Schnable, P. S., Vaughn, M. W., Dawe, R. K., & Springer, N. M. (2012). Spreading of heterochromatin is limited to specific families of maize retrotransposons. *PLoS Genetics*, *8*(12), e1003127.
Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238.
Faust, G. G., & Hall, I. M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* , *30*(17), 2503–2505.
Filiault, D. L., Ballerini, E. S., Mandáková, T., Aköz, G., Derieg, N. J., Schmutz, J., Jenkins, J., Grimwood, J., Shu, S., Hayes, R. D., Hellsten, U., Barry, K., Yan, J., Mihaltcheva, S., Karafiátová, M., Nizhynska, V., Kramer, E. M., Lysak, M. A., Hodges, S. A., & Nordborg, M. (2018). The Aquilegia genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *eLife*, *7*. https://doi.org/10.7554/elife.36426
Filipecki, M., & Malepszy, S. (2006). Unintended consequences of plant transformation: a molecular insight. *Journal of Applied Genetics*, *47*(4), 277–286.
Fitch, M. M. M., Manshardt, R. M., Gonsalves, D., Slightom, J. L., & Sanford, J. C.

(1992). Virus Resistant Papaya Plants Derived from Tissues Bombarded with the Coat Protein Gene of Papaya Ringspot Virus. *Biotechnology* , *10*(11), 1466–1472.
Forsbach, A., Schubert, D., Lechtenberg, B., Gils, M., & Schmidt, R. (2003). A comprehensive characterization of single-copy T-DNA insertions in the Arabidopsis thaliana genome. *Plant Molecular Biology*, *52*(1), 161–176.
Gaire, B. P., & Subedi, L. (2013). A review on the pharmacological and toxicological aspects of Datura stramonium L. *Journal of Integrative Medicine*, *11*(2), 73–79.
Garbarino, J. E., & Gibbons, I. R. (2002). Expression and genomic analysis of midasin, a novel and highly conserved AAA protein distantly related to dynein. *BMC Genomics*, *3*, 18.
Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. In *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1207.3907
Gelvin, S. B. (2017). Integration of Agrobacterium T-DNA into the Plant Genome. *Annual Review of Genetics*, *51*, 195–217.
Georgiev, V., Marchev, A., Berkov, S., & Pavlov, A. (2013). Plant In Vitro Systems as Sources of Tropane Alkaloids. In K. G. Ramawat & J.-M. Mérillon (Eds.), *Natural Products: Phytochemistry, Botany and Metabolism of Alkaloids, Phenolics and Terpenes* (pp. 173–211). Springer Berlin Heidelberg.
Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652.
Grynkiewicz, G., & Gadzikowska, M. (2008). Tropane alkaloids as medicinally useful natural products and their synthetic derivatives as new drugs. *Pharmacological Reports: PR*, *60*(4), 439–463.
Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr, Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L., & White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*(19), 5654–5666.
Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, *9*(1), R7.
Hang, A., & Bregitzer, P. (1993). Chromosomal Variations in Immature Embryo-Derived Calli from Six Barley Cultivars. *The Journal of Heredity*, *84*(2), 105–108.
Harkess, A. (2017, November 7). *Modified Samberg Phenol:Chloforom HMW DNA prep for (some) plants*. Protocols.io. https://doi.org/10.17504/protocols.io.kpwcvpe
Harkess, A., Zhou, J., Xu, C., Bowers, J. E., Van der Hulst, R., Ayyampalayam, S., Mercati, F., Riccardi, P., McKain, M. R., Kakrana, A., Tang, H., Ray, J., Groenendijk, J., Arikit, S., Mathioni, S. M., Nakano, M., Shan, H., Telgmann-Rauber, A., Kanno, A., … Chen, G. (2017). The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nature Communications*, *8*(1), 1279.
Hashimoto, T., Nakajima, K., Ongena, G., & Yamada, Y. (1992). Two Tropinone Reductases with Distinct Stereospecificities from Cultured Roots of Hyoscyamus niger. *Plant Physiology*, *100*(2), 836–845.
Heinz, D. J., & Mee, G. W. P. (1971). Morphologic, Cytogenetic, and Enzymatic

Variation in Saccharum Species Hybrid Clones Derived from Callus Tissue. *American Journal of Botany*, *58*(3), 257–262.
Herman, L., Jacobs, A., Van Montagu, M., & Depicker, A. (1990). Plant chromosome/marker gene fusion assay for study of normal and truncated T-DNA integration events. *Molecular & General Genetics: MGG*, *224*(2), 248–256.
Hirakawa, H., Shirasawa, K., Miyatake, K., Nunome, T., Negoro, S., Ohyama, A., Yamaguchi, H., Sato, S., Isobe, S., Tabata, S., & Others. (2014). Draft genome sequence of eggplant (Solanum melongena L.): the representative solanum species indigenous to the old world. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *21*(6), 649–660.
Hirochika, H., Otsuki, H., Yoshikawa, M., Otsuki, Y., Sugimoto, K., & Takeda, S. (1996). Autonomous transposition of the tobacco retrotransposon Tto1 in rice. *The Plant Cell*, *8*(4), 725–734.
Hirochika, H., Sugimoto, K., Otsuki, Y., Tsugawa, H., & Kanda, M. (1996). Retrotransposons of rice involved in mutations induced by tissue culture. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(15), 7783–7788.
Hollister, J. D., & Gaut, B. S. (2009). Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research*, *19*(8), 1419–1428.
Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S., Fei, Z., Giovannoni, J. J., Mueller, L. A., & Saha, S. (2019). An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. In *bioRxiv* (p. 767764). https://doi.org/10.1101/767764
Huang, J., Zhang, K., Shen, Y., Huang, Z., Li, M., Tang, D., Gu, M., & Cheng, Z. (2009). Identification of a high frequency transposon induced by tissue culture, nDaiZ, a member of the hAT family in rice. *Genomics*, *93*(3), 274–281.
Ichikawa, T., Nakazawa, M., Kawashima, M., Muto, S., Gohda, K., Suzuki, K., Ishikawa, A., Kobayashi, H., Yoshizumi, T., Tsumoto, Y., & Others. (2003). Sequence database of 1172 T-DNA insertion sites in Arabidopsis activation-tagging lines that showed phenotypes in T1 generation. *The Plant Journal: For Cell and Molecular Biology*, *36*(3), 421–429.
Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., Jahesh, G., Khan, H., Coombe, L., Warren, R. L., & Birol, I. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, *27*(5), 768–777.
Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyère, C., … French-Italian Public Consortium for Grapevine Genome Characterization. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*(7161), 463–467.
Jiang, C., Mithani, A., Gan, X., Belfield, E. J., Klingler, J. P., Zhu, J.-K., Ragoussis, J., Mott, R., & Harberd, N. P. (2011). Regenerant Arabidopsis lineages display a distinct genome-wide spectrum of mutations conferring variant phenotypes. *Current*

Biology: CB, *21*(16), 1385–1390.
Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., Campbell, M. S., Stein, J. C., Wei, X., Chin, C.-S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K. L., Wolfgruber, T. K., May, M. R., Springer, N. M., … Ware, D. (2017). Improved maize reference genome with single-molecule technologies. *Nature*, *546*(7659), 524–527.
Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C. W., Yi, T.-S., & Li, D.-Z. (2020). GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*(1), 241.
Johnson, S. S., Phillips, R. L., & Rines, H. W. (1987). Possible role of heterochromatin in chromosome breakage induced by tissue culture in oats (Avena sativa L.). *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada*, *29*(3), 439–446.
Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* , *30*(9), 1236–1240.
Joyce, S. M., Cassells, A. C., & Mohan Jain, S. (2003). Stress and aberrant phenotypes in vitro culture. *Plant Cell, Tissue and Organ Culture*, *74*(2), 103–121.
Kaeppler, S. M., Kaeppler, H. F., & Rhee, Y. (2000). Epigenetic aspects of somaclonal variation in plants. In M. A. Matzke & A. J. M. Matzke (Eds.), *Plant Gene Silencing* (pp. 59–68). Springer Netherlands.
Kaeppler, S. M., & Phillips, R. L. (1993). DNA methylation and tissue culture-induced variation in plants. *In Vitro Cellular & Developmental Biology-Plant*, *29*(3), 125–130.
Karp, A. (1991). On the current understanding of somaclonal variation. *Oxford Surveys of Plant Molecular and Cell Biology*, *7*. http://agris.fao.org/agris-search/search.do?recordID=US201301745745
Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.
Kikuchi, K., Terauchi, K., Wada, M., & Hirano, H.-Y. (2003). The plant MITE mPing is mobilized in anther culture. *Nature*, *421*(6919), 167–170.
Kim, S., Park, J., Yeom, S.-I., Kim, Y.-M., Seo, E., Kim, K.-T., Kim, M.-S., Lee, J. M., Cheong, K., Shin, H.-S., Kim, S.-B., Han, K., Lee, J., Park, M., Lee, H.-A., Lee, H.-Y., Lee, Y., Oh, S., Lee, J. H., … Choi, D. (2017). New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. *Genome Biology*, *18*(1), 210.
Kim, S., Park, M., Yeom, S.-I., Kim, Y.-M., Lee, J. M., Lee, H.-A., Seo, E., Choi, J., Cheong, K., Kim, K.-T., Jung, K., Lee, G.-W., Oh, S.-K., Bae, C., Kim, S.-B., Lee, H.-Y., Kim, S.-Y., Kim, M.-S., Kang, B.-C., … Choi, D. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature Genetics*, *46*(3), 270–278.
Kobayashi, S., Goto-Yamamoto, N., & Hirochika, H. (2004). Retrotransposon-induced mutations in grape skin color. *Science*, *304*(5673), 982.
Ko, D. K., Nadakuduti, S. S., Douches, D. S., & Buell, C. R. (2018). Transcriptome profiling of transgenic potato plants provides insights into variability caused by plant transformation. *PloS One*, *13*(11), e0206055.

Kohnen-Johannsen, K. L., & Kayser, O. (2019). Tropane Alkaloids: Chemistry, Pharmacology, Biosynthesis and Production. *Molecules* , *24*(4). https://doi.org/10.3390/molecules24040796

Koncz, C., Németh, K., Rédei, G. P., & Schell, J. (1992). T-DNA insertional mutagenesis in Arabidopsis. *Plant Molecular Biology*, *20*(5), 963–976.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btz305

Krueger, F. (2012). *Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries*. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Kubešová, M., Moravcova, L., Suda, J., Jarošík, V., Pyšek, P., & Others. (2010). Naturalized plants have smaller genomes than their non-invading relatives: a flow cytometric analysis of the Czech alien flora. *Preslia*, *82*(1), 81–96.

Labra, M., Vannini, C., Grassi, F., Bracale, M., Balsemin, M., Basso, B., & Sala, F. (2004). Genomic stability in Arabidopsis thaliana transgenic plants obtained by floral dip. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *109*(7), 1512–1518.

Lakstygal, A. M., Kolesnikova, T. O., Khatsko, S. L., Zabegalov, K. N., Volgin, A. D., Demin, K. A., Shevyrin, V. A., Wappler-Guzzetta, E. A., & Kalueff, A. V. (2019). DARK Classics in Chemical Neuroscience: Atropine, Scopolamine, and Other Anticholinergic Deliriant Hallucinogens. *ACS Chemical Neuroscience*, *10*(5), 2144–2159.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., & Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, *40*(Database issue), D1202–D1210.

Larkin, P. J., & Scowcroft, W. R. (1981). Somaclonal variation — a novel source of variability from cell cultures for plant improvement. In *Theoretical and Applied Genetics* (Vol. 60, Issue 4, pp. 197–214). https://doi.org/10.1007/bf02342540

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), R84.

Lee, M., & Phillips, R. L. (1987). Genomic rearrangements in maize induced by tissue culture. In *Genome* (Vol. 29, Issue 1, pp. 122–128). https://doi.org/10.1139/g87-021

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* , *25*(14), 1754–1760.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., & Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, *33*(20), 6494–6506.

LoSchiavo, F., Pitto, L., Giuliano, G., Torti, G., Nuti-Ronchi, V., Marazziti, D., Vergara, R., Orselli, S., & Terzi, M. (1989). DNA methylation of embryogenic carrot cell cultures and its variations as caused by mutation, differentiation, hormones and hypomethylating drugs. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *77*(3), 325–331.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.
Lucht, J. M., Mauch-Mani, B., Steiner, H.-Y., Metraux, J.-P., Ryals, J., & Hohn, B. (2002). Pathogen stress increases somatic recombination frequency in Arabidopsis. *Nature Genetics*, *30*(3), 311–314.
Lu, L., Chen, J., Robb, S. M. C., Okumoto, Y., Stajich, J. E., & Wessler, S. R. (2017). Tracking the genome-wide outcomes of a transposable element burst over decades of amplification. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(49), E10550–E10559.
Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* , *20*(16), 2878–2879.
Marcotrigiano, M., & Jagannathan, L. (1988). Paulownia tomentosa somaclonal snowstorm. *HortScience: A Publication of the American Society for Horticultural Science*, *23*(1), 226–227.
McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, *226*(4676), 792–801.
Miyao, A., Nakagome, M., Ohnuma, T., Yamagata, H., Kanamori, H., Katayose, Y., Takahashi, A., Matsumoto, T., & Hirochika, H. (2012). Molecular spectrum of somaclonal variation in regenerated rice revealed by whole-genome sequencing. *Plant & Cell Physiology*, *53*(1), 256–264.
Naito, K., Cho, E., Yang, G., Campbell, M. A., Yano, K., Okumoto, Y., Tanisaka, T., & Wessler, S. R. (2006). Dramatic amplification of a rice transposable element during recent domestication. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(47), 17620–17625.
Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C. N., Richardson, A. O., Okumoto, Y., Tanisaka, T., & Wessler, S. R. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*, *461*(7267), 1130–1134.
Nocquet, P.-A., & Opatz, T. (2016). Total Synthesis of (±)-Scopolamine: Challenges of the Tropane Ring. *European Journal of Organic Chemistry*, *2016*(6), 1156–1164.
Oono, K. (1985). Putative homozygous mutations in regenerated plants of rice. *Molecular & General Genetics: MGG*, *198*(3), 377–384.
Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N., & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, *20*(1), 275.
Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R. L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., & Buell, C. R. (2007). The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Research*, *35*(Database issue), D883–D887.
Palmer, J. D. (1983). Chloroplast DNA exists in two orientations. *Nature*, *301*(5895), 92–93.
Palmer, J. M. (2019). *funannotate* (Version 1.6.0) [Computer software]. https://github.com/nextgenusfs/funannotate/releases/tag/1.6.0
Parr, A. J., Payne, J., Eagles, J., Chapman, B. T., Robins, R. J., & Rhodes, M. J. C. (1990). Variation in tropane alkaloid accumulation within the solanaceae and strategies for its exploitation. *Phytochemistry*, *29*(8), 2545–2550.

Rajewski, A. C., Elkins, K. B., Henry, A., Van Eck, J., & Litt, A. (2019). In vitro plant regeneration and Agrobacterium tumefaciens--mediated transformation of Datura stramonium (Solanaceae). *Applications in Plant Sciences*, e01220.

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1432.

Rang, A., Linke, B., & Jansen, B. (2005). Detection of RNA variants transcribed from the transgene in Roundup Ready soybean. *European Food Research and Technology = Zeitschrift Fur Lebensmittel-Untersuchung Und -Forschung. A*, *220*(3-4), 438–443.

Razali, R., Bougouffa, S., Morton, M. J. L., Lightfoot, D. J., Alam, I., Essack, M., Arold, S. T., Kamau, A. A., Schmöckel, S. M., Pailles, Y., Shahid, M., Michell, C. T., Al-Babili, S., Ho, Y. S., Tester, M., Bajic, V. B., & Negrão, S. (2018). The Genome Sequence of the Wild Tomato Solanum pimpinellifolium Provides Insights Into Salinity Tolerance. *Frontiers in Plant Science*, *9*, 1402.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution / British Ecological Society*, *3*(2), 217–223.

Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikit, S., Song, C., Xia, L., Froenicke, L., Lavelle, D. O., Truco, M.-J., Xia, R., Zhu, S., Xu, C., Xu, H., Xu, X., Cox, K., Korf, I., Meyers, B. C., & Michelmore, R. W. (2017). Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature Communications*, *8*, 14953.

Salmela, L., & Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics* , *30*(24), 3506–3514.

Sanford, J. C., Klein, T. M., Wolf, E. D., & Allen, N. (1987). Delivery of Substances into Cell and Tissues Using a Particle Bombardment Process. *Particulate Science and Technology*, *5*(1), 27–37.

Särkinen, T., Bohs, L., Olmstead, R. G., & Knapp, S. (2013). A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. *BMC Evolutionary Biology*, *13*, 214.

Schiavinato, M., Strasser, R., Mach, L., Dohm, J. C., & Himmelbauer, H. (2019). Genome and transcriptome characterization of the glycoengineered Nicotiana benthamiana line ΔXT/FT. *BMC Genomics*, *20*(1), 594.

Schmitt, F., Oakeley, E. J., & Jost, J. P. (1997). Antibiotics induce genome-wide hypermethylation in cultured Nicotiana tabacum plants. *The Journal of Biological Chemistry*, *272*(3), 1534–1540.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., … Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, *463*(7278), 178–183.

Sha, Y., Li, S., Pei, Z., Luo, L., Tian, Y., & He, C. (2004). Generation and flanking sequence analysis of a rice T-DNA tagged population. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *108*(2), 306–314.

Sierro, N., Battey, J. N. D., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., Peitsch, M. C., & Ivanov, N. V. (2013). Reference genomes and transcriptomes of Nicotiana sylvestris and Nicotiana tomentosiformis. *Genome Biology*, *14*(6), R60.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* , *31*(19), 3210–3212.

Smit, A. F. A., & Hubley, R. (2008). *RepeatModeler* (Version 1.0.3) [Computer software]. http://www.repeatmasker.org

Smit, A. F. A., Hubley, R., & Green, P. (2013). *RepeatMasker* (Version 4.0.3) [Computer software]. http://www.repeatmasker.org

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(Web Server issue), W435–W439.

Stroud, H., Ding, B., Simon, S. A., Feng, S., Bellizzi, M., Pellegrini, M., Wang, G.-L., Meyers, B. C., & Jacobsen, S. E. (2013). Plants regenerated from tissue culture contain stable epigenome changes in rice. *eLife*, *2*, e00354.

The Angiosperm Phylogeny Group, Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., Soltis, D. E., Mabberley, D. J., Sennikov, A. N., Soltis, P. S., & Stevens, P. F. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society. Linnean Society of London*, *181*(1), 1–20.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, *45*(W1), W6–W11.

Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*(7400), 635–641.

Veilleux, R. E., & Johnson, A. A. T. (1998). Somaclonal variation: molecular analysis, transformation interaction, and utilization. *Plant Breeding Reviews*, *16*, 229–266.

Warren, R. L. (2016). RAILS and Cobbler: Scaffolding and automated finishing of draft genomes using long DNA sequences. *Journal of Open Source Software*, *1*(7), 116.

Warren, R. L., Coombe, L., Mohamadi, H., Zhang, J., Jaquish, B., Isabel, N., Jones, S. J. M., Bousquet, J., Bohlmann, J., & Birol, I. (2019). ntEdit: scalable genome sequence polishing. *Bioinformatics* , *35*(21), 4430–4432.

Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J. M., & Birol, I. (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*, *4*, 35.

Wenck, A., Czakó, M., Kanevski, I., & Márton, L. (1997). Frequent collinear long transfer of DNA inclusive of the whole binary vector during Agrobacterium-mediated transformation. *Plant Molecular Biology*, *34*(6), 913–922.

Weng, M.-L., Becker, C., Hildebrandt, J., Neumann, M., Rutter, M. T., Shaw, R. G., Weigel, D., & Fenster, C. B. (2019). Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in Arabidopsis thaliana. *Genetics*, *211*(2), 703–714.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews. Genetics*, *8*(12), 973–982.

Xia, K., Liu, X., Zhang, Q., Qiang, W., Guo, J., Lan, X., Chen, M., & Liao, Z. (2016). Promoting scopolamine biosynthesis in transgenic Atropa belladonna plants with pmt and h6h overexpression under field conditions. *Plant Physiology and Biochemistry: PPB / Societe Francaise de Physiologie Vegetale*, *106*, 46–53.

Xu, S., Brockmöller, T., Navarro-Quezada, A., Kuhl, H., Gase, K., Ling, Z., Zhou, W., Kreitzer, C., Stanke, M., Tang, H., Lyons, E., Pandey, P., Pandey, S. P., Timmermann, B., Gaquerel, E., & Baldwin, I. T. (2017). Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(23), 6133–6138.

Yang, Y., Dang, Y., Li, Q., Lu, J., Li, X., & Wang, Y. (2014). Complete chloroplast genome sequence of poisonous and medicinal plant Datura stramonium: organizations and implications for genetic engineering. *PloS One*, *9*(11), e110656.

Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *69*(1). https://doi.org/10.1002/cpbi.96

Zhang, D., Wang, Z., Wang, N., Gao, Y., Liu, Y., Wu, Y., Bai, Y., Zhang, Z., Lin, X., Dong, Y., Ou, X., Xu, C., & Liu, B. (2014). Tissue culture-induced heritable genomic variation in rice, and their phenotypic implications. *PloS One*, *9*(5), e96879.

Zhao, X., Meng, Z., Wang, Y., Chen, W., Sun, C., Cui, B., Cui, J., Yu, M., Zeng, Z., Guo, S., Luo, D., Cheng, J. Q., Zhang, R., & Cui, H. (2017). Pollen magnetofection for genetic modification with magnetic nanoparticles as gene carriers. *Nature Plants*, *3*(12), 956–964.

Zwaenepoel, A., & Van de Peer, Y. (2019). wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* , *35*(12), 2153–2155.

Chapter 4: Multispecies Transcriptomes Reveal Core Fruit Development Genes

Introduction

Seed-bearing fruits are the hallmark feature uniting the angiosperms, and this innovation has contributed to the enormous success of the group in terms of both species richness and economic importance for humans. Indeed, 82% of daily calories eaten by humans are derived directly from angiosperm plants (FAO, 2017) and 80% of those calories are from the fruits themselves. When indirect sources are taken into account, effectively all calories eaten by humans derive from angiosperms.

From a diversity standpoint, angiosperms also represent an unparalleled evolutionary success story. Since their initial split with gymnosperms, angiosperms diversified prolifically to comprise approximately 90% of all extant land plant species and now occupy key positions in nearly every biome on the planet (Crepet & Niklas, 2009). Although the precise reasons for the evolutionary diversification and success of angiosperms are still debated (Armbruster, 2014), certainly the complex interplay between flowers and their pollinators and the ability to further use animals as a seed dispersal vectors has contributed significantly to this (Regal, 1977).

Although the molecular mechanisms underlying fruit development and evolution are not thoroughly understood, morphological changes are well documented and provide a conceptual framework to examine molecular mechanisms. Fruits can broadly be classified as either dry or fleshy. The true berry of cultivated tomato (*Solanum lycopersicum*) and the pepo of melon (*Cucumis melo*) are examples of fleshy fruits, while the capsules of desert tobacco (*Nicotiana obtusifolia*) and the silique of the

model plant *Arabidopsis thaliana* (hereafter Arabidopsis) are both dry fruits. Despite the very different appearances of these fruits, their developmental progression can each be divided into common stages with similar processes occurring at each stage across all four species (Table 1) (Gillaspy et al., 1993).

All fruits are derived from one or multiple ovaries. The earliest stage of fruit development (Stage 1) occurs before the ovules have been fertilized and comprises a stage of ovary patterning that is common to all species. Although specific terminology differs, the ovaries of all four species previously mentioned are divided into multiple chambers. In the cases of desert tobacco, Arabidopsis, and the wild relative of tomato (*S. pimpinellifolium*), the ovary is divided into two chambers. The fruits of wild melon species have 2-5 chambers, while both cultivated melon and cultivated tomato have a variable number of chambers (Monforte et al., 2014). Following fertilization of the ovules, the ovary transitions to a fruit and enters into a stage of rapid cell division (Stage 2). The length of this phase differs, with both Arabidopsis and desert tobacco undergoing cell division phases of 1-3 days, while tomato and melon cell division phases can occur over 1-2 weeks (Chayut et al., 2015; Pabón-Mora & Litt, 2011; Ripoll et al., 2019). Additionally, the orientation of these cell divisions in the pericarp (outer fruit wall) varies. Pericarp cell divisions in desert tobacco are primarily anticlinal and maintain 7-8 pericarp cell layers, but pericarp divisions in tomato, and likely melon, are both anticlinal and periclinal and increase the number of cell layers dramatically (Pabón-Mora & Litt, 2011).

Following this burst of cell division, the fruit enters a phase of cell differentiation (Stage 3). In this stage, the fruits of each species begin to morphologically diverge from one another more drastically. Among the dry-fruited species Arabidopsis and desert tobacco, Stage 3 is characterized primarily by the deposition of lignin in the secondary cell

walls of the pericarp. Because both of these fruits are dehiscent, pericarp lignification is tightly spatially controlled to allow for the formation of dehiscence zones where the mature pericarp will split open to allow seed dispersal (Ferrándiz et al., 2000; Smykal et al., 2007). In tomato and melon, Stage 3 is one of pronounced pericarp cell expansion and contributes strongly to the mature fruit size. Concomitant with the increase in cell volume is also an increase in cell ploidy, with endoreduplication up to 256x (Bourdon et al., 2010). Endoreduplication has also been reported in Arabidopsis pericarp cells undergoing cell expansion and may be a more general feature of Stage 3 across fruit types (Ripoll et al., 2019).

Having reached their final size, these fruits transition to physiological maturity (Stage 4). In the case of the dry fruits presented here, Stage 4 involves a senescence, drying down, and dehiscence of the pericarp along the previously patterned dehiscence zones. During dehiscence, tension created by drying of the lignified pericarp and autolysis of certain cells in the dehiscence zone allow the pericarp to split open and seeds to be dispersed. In contrast, Stage 4 in fleshy fruits generally involves accumulation of sugars, volatile flavor compounds, pigments, and nutrients in the pericarp, along with softening of pericarp cell walls. In the climacteric fruits tomato and melon, this process coincides with a burst in production of the gaseous hormone ethylene, but non-climacteric fruits undergo similar processes in an ethylene-independent manner. Especially in tomatoes, an initial transition or “breaker” stage is also recognized between Stages 3 and 4. Breaker stage is characterized by the initial color change in the pericarp from green to pink or red.

The early morphological similarities and the similar developmental processes occurring across these diverse fruit types are likely related to their shared evolutionary origin. In fact across angiosperm evolution, there have been repeated shifts from an

ancestral dry fruit to a derived fleshy fruit (Bremer & Eriksson, 1992; Clausing et al., 2000; Cox, 1948; Givnish et al., 2005; Knapp, 2002; Plunkett et al., 1997; Spalik et al., 2001; Weber, 2004) The conservation of morphological, developmental, and evolutionary patterns led us to hypothesize that there might also be conservation of gene function and/or gene expression patterns in fruit development across species. Although many studies characterizing gene expression during fruit development have dramatically advanced our understanding within single species or between closely related species, a comparison at higher taxonomic levels could provide evidence for a set of “core” fruit development genes and shed light on the conserved pathways necessary to build a fruit.

We examined pericarp transcriptomes of two dry- and three fleshy-fruited species across developmental time. Our results draw upon 42 pericarp RNAseq libraries of three members of the nightshade family (Solanaceae) generated for this study as well as data from 30 additional publically available pericarp libraries of more distantly related dry- and fleshy-fruited species (Table 1). Integrating information about orthologous genes and using nested models to call differential gene expression, we uncovered a set of 121 differentially expressed genes with conserved patterns of expression among these species. These genes participate in many biological processes and may constitute a core set of genes whose expression patterns are necessary (but not sufficient) for fruit development. In addition, we found a much larger set of 1,795 genes with patterns of expression conserved within, but divergent between, dry and fleshy fruits. These genes with divergent patterns between fruit types, may represent accessory genes that act to specify the developmental patterns separating these fruit types.

Results

Expression patterns for polyamine and isoprenoid biosynthesis are conserved between wild and cultivated tomato species

In our investigation, we began with the commonly studied cultivated tomato (*S. lycopersicum*) but also included its closest wild relative (*S. pimpinellifolium*). We reasoned that the intentional and unintentional changes during the domestication of cultivated tomato could have an impact on gene expression patterns in the fruit, whose ripening, flavor, and structure have been targets of artificial selection.

Using RNAseq data from 5 developmental stages from fruit of both tomato species (Table 1), we first asked which differentially expressed genes across fruit development showed a conserved pattern of expression between the two species. We aligned reads from both tomato species to the most recent annotation of the cultivated tomato genome and called differential expression among developmental stages with a model that was blind to species (Fischer et al., 2018; Hosmani et al., 2019; Sander et al., 2017). This model required that the expression of a gene be statistically significantly different between at least two stages. We discovered 6,165 genes (of 34,075 total) with changes in pericarp expression level with the same pattern in cultivated and wild tomato. A GO term enrichment analysis of this cohort of genes revealed that they function in diverse general biological processes including glucose metabolism, transport, and responses to damage and stress (Fig. 4.1A). In addition, several lower-level GO terms were also enriched among this set of genes including spermidine biosynthetic processes, which play a role in the synthesis of polyamine compounds related to flavor and timing of fruit senescence (Nambeesan et al., 2010).

To uncover more fine-scale patterns among these differentially expressed genes, we clustered them by their expression profiles during fruit development and performed GO analyses on each of the 20 resulting clusters (Supplemental Figure 4.1). Several of these clusters showed informative enrichments. Cluster 4 contained genes with low and steady expression in early fruit development, peaking at the transition to Stage 3 and remaining high through the red ripe stage (Fig. 4.1B). This cluster showed enrichment for isoprenoid biosynthesis (GO:0008299), fatty acid biosynthesis, and potassium ion transport (Fig. 4.1C). Given the peak expression of this cluster prior to the breaker stage, it is likely that these terms relate to the accumulation of pigment and flavor compounds before and during ripening (Adams et al., 1978; Li et al., 2020; Tieman et al., 2012)(Li et al., 2020; Tieman et al., 2012). This cluster also showed enrichment for genes related to cell wall modification, consistent with the prominent changes in cell wall composition as the fruit ripens and softens. Cluster 10 showed a nearly opposite pattern to cluster 4, with low expression in later fruit development, and high to moderate expression at Stages 1-3 (Fig. 4.1D). These earlier stages of fruit development include bursts of cell division and DNA replication and this cluster contained significant hits for DNA replication, nucleotide biosynthesis and several cell wall biosynthetic terms (Fig. 4.1E).

Wild and cultivated tomato show subtle differences in expression patterns

One of the most notable effects of artificial selection between cultivated and wild tomato is fruit size. As the pericarp makes up a substantial portion of the fruit, we wanted to know the extent to which pericarp gene expression patterns differ between the two

species. We therefore called differentially expressed genes with a model that included the species as a covariate and used a likelihood ratio test to determine which genes showed a statistically significant difference in gene expression pattern between the two species. The resulting 1,472 genes that exhibited divergent expression patterns between cultivated and wild tomato showed GO term enrichment for plant-type cell wall organization and lipid biosynthetic processes, with 11 genes assigned to each term, the maximum number of genes for any GO term in this analysis (Fig. 4.1F). This enrichment likely reflects both the different flavor profiles of the two fruits as well as their conspicuous differences in pericarp size. A clustering and GO analysis of these 1,472 genes produced clusters with only very subtle differences in gene expression profiles between species and no apparently informative GO terms (Supplemental Figure 4.2). Potentially the differences in fruit phenotype between wild and cultivated tomato involve a small number of genes with slight changes in expression pattern, but we cannot rule out that these differences involve changes in timing or expression domains that were not included in our sampling regime.

Divergence in expression of ethylene and secondary metabolite synthesis genes following domestication

Because cultivated tomato is routinely used as a model to study climacteric fruit ripening, many genes have been identified as playing a role in this process. We asked to what extent the expression patterns of these well-studied ripening genes have changed following domestication. We used our combined wild and cultivated tomato dataset to examine the expression of 21 structural genes involved in ethylene biosynthesis, pigment production, and flavor compound biosynthesis (Supplemental Figure 4.3). Among these

structural genes, one ethylene-related gene and two flavor compound-related genes have a pattern of expression with statistically significant differences between cultivated and wild tomato (Fig. 4.2A-C).

The gene *ACO6* encodes an ethylene biosynthesis enzyme whose role has not been well characterized during fruit development (Houben & Van de Poel, 2019). In our analysis, *ACO6* was the only structural gene related to ethylene synthesis or perception with a statistically significant difference in expression pattern between the two tomato species (Fig. 4.2A). The other genes showed either no statistically significant change in expression across pericarp development or no statistically significant difference in pattern between the two species. In contrast, *ACO6* has higher expression at every stage we sampled in wild tomato compared to cultivated tomato. Additionally, *ACO6* reaches its maximum expression in cultivated tomato at stage 2, which is characterized primarily by cell division, whereas in wild tomato, peak expression is reached at stage 3, which is characterized largely by cell expansion (Table 1). The peak at stage 3 was not seen for any other *ACO* homologues, suggesting a divergent role for this enzyme during pericarp development (Supplemental Fig. 4.3A-G).

TomLoxC encodes a lipoxygenase and contributes to desirable flavor in tomato fruit (Chen et al., 2004; Shen et al., 2014). In both species, expression was not detected in stages 1-2 of pericarp development (Fig. 4.2B). In wild tomato, *TomLoxC* transcripts accumulated to moderate levels at stage 3 and breaker stage pericarps, but dropped to much lower levels in red ripe fruits. In cultivated tomato, however, we did not detect any *TomLoxC* transcripts until the breaker stage, where we observed maximum expression. The level dropped slightly at the red ripe stage, but still remained higher than the peak expression seen in wild tomato. Polymorphism in *TomLoxC* expression was recently

observed in a large study of wild and cultivated tomato accessions, and found to correlate with a large deletion in the promoter of *TomLoxC* that was selected against during domestication (Gao et al., 2019).

Finally, *GAD1* encodes one of three known tomato glutamate decarboxylases, which are responsible for the production of **γ**-aminobutyric acid (GABA) (Akihiro et al., 2008). In our analysis both tomato species displayed a similar trend for *GAD1* expression during pericarp development, which was consistent with previous studies (Akihiro et al., 2008) (Fig. 4.2C). However the two species showed a statistically significant difference in the magnitude of expression, with wild tomato showing approximately 3x higher peak expression of *GAD1* at the red ripe stage. GABA can accumulate to very high levels in tomato fruit and is thought to be involved with stress responses and defense (Bouché et al., 2003; MacGregor et al., 2003). Given that wild tomato is a widely recognized resource for introgression of stress tolerance, this difference in a key GABA biosynthesis enzyme represents a potential future avenue for plant breeders (Razali et al., 2018).

Fruit size-, firmness-, and lignification-related transcription factors differ in expression between wild and cultivated tomato

Because changes in the expression of transcription factors can influence the expression of many target genes simultaneously, we wanted to know the extent to which such regulatory genes differed in expression pattern between these two species. We selected 18 transcription factors with prominent roles in fruit and flower development and used our combined wild and cultivated tomato data set to ask if any of these genes showed

statistically significant differences in expression between the two species (Supplemental Figure 4.4).

Although many of the selected genes showed statistically significant differential expression across pericarp development with a pattern common to both species, only four had statistically significant support for a difference in expression between the two species. This included three type-II MADS-box genes *MBP3*, *TAG1*, and *TAGL1*, along with the *SQUAMOSA* promoter-binding protein-like transcription factor, *SPL-CNR* (Fig. 4.2D-G).

MBP3 and *AGL11* are orthologous to the Arabidopsis gene *SEEDSTICK*, which helps specify ovule identity (Ocarez & Mejía, 2016; Pinyopich et al., 2003). *AGL11* does not show statistically significant differential expression between tomato species; however its paralog, *MBP3*, does (Supplemental Figure 4.4A and Fig. 4.2D). Our dataset shows that in cultivated tomato, *MBP3* expression is low in stages 1 and 2 before becoming nearly undetectable for the rest of fruit development. In contrast, wild tomato *MBP3* is similar to cultivated tomato in expression at stages 1 and 2 but peaks at stage 3 with a roughly 3-fold increase compared to stage 1. Several functional characterizations suggest that *AGL11* helps specify ovule identity in tomato, but we could find no functional characterizations of *MBP3* (Huang et al., 2017; Ocarez & Mejía, 2016).

The genes *TAG1* and *TAGL1* are orthologs of the Arabidopsis genes *AGAMOUS* and *SHATTERPROOF1/2*, respectively (Pan et al., 2010; Pnueli et al., 1994). Both tomato genes have been shown to control several aspects of fruit development and to help specify the identity of stamens and carpels (Gimenez et al., 2016; Pan et al., 2010). Comparing wild and cultivated tomato, *TAG1* shows a more extreme difference in expression than *TAGL1*, though both are statistically significant ($p << 0.01$, Fig. 4.2F-G). In wild tomato, *TAG1* expression increases linearly nearly 25-fold between stage 3 and the red ripe stage;

however, in cultivated tomato the increase is barely detectable. For *TAGL1* the departure in expression is more subtle but most obvious at the breaker stage where wild tomato *TAGL1* expression peaks and cultivated tomato *TAGL1* expression is at its lowest levels. Previous silencing experiments in cultivated tomato suggest that both genes contribute positively to pericarp thickness (Gimenez et al., 2016). Our result is therefore counterintuitive as cultivated tomato generally has a thicker pericarp than wild tomato, but wild tomato showed consistently higher expression of both genes in the pericarp.

The *SQUAMOSA* promoter-binding protein-like transcription factor *SPL-CNR* is thought to be the causative gene for the *Cnr* mutation that affects ripe tomato fruit color and firmness (Eriksson et al., 2004; Lai et al., 2020; Manning et al., 2006; Thompson et al., 1999). In our analysis, *SPL-CNR* showed a statistically significant difference in expression between the two tomato species (p=3.2x10^{-4}) with wild tomato showing higher expression in both stage 3 and breaker stage pericarps (Fig. 4.2E). Recently *SPL-CNR* expression has been shown to negatively affect cell-to-cell adhesion and to promote cell death (Lai et al., 2020), consistent with a model whereby low expression of *SPL-CNR* in the *Cnr* mutant could lead to a non-softening fruit due to increased cell adhesion or lower levels of cell death. The decreased firmness in mature wild tomato fruits coupled with their higher expression of *SPL-CNR* and the increased desirability of firmer cultivated tomato fruits suggests that the expression changes at the *SPL-CNR* locus could have been the result of domestication (Doganlar et al., 2002; Tanksley et al., 1996).

Desert tobacco pericarp transcriptome is enriched for secondary metabolite synthesis and shows fewer differentially expressed genes than tomato

In contrast to tomato, desert tobacco (*Nicotiana obtusifolia*) produces a dry capsular fruit. We were able to obtain RNA from pericarps at stages 1-3 as well as a "transition" stage as the fruit is maturing, analogous to breaker stage in tomato (Table 1). Physiologically mature desert tobacco fruits are dry and highly lignified, and we were unable to obtain RNA from this final stage.

Because fruit development in desert tobacco has not been molecularly characterized, we examined gene expression dynamics in desert tobacco pericarp development. We applied a similar model that required the expression of a gene be statistically significantly different between at least two stages in order to be considered differentially expressed. We uncovered 1,392 desert tobacco genes with differential expression across the four stages, much fewer than the 6,165 differentially expressed genes among the tomato stages. We performed a GO analysis on this cohort of genes and found that they largely relate to either DNA replication and synthesis or to the synthesis of secondary metabolites such as spermidine or terpenoids (Fig. 4.3A). Interestingly, the set of genes with conserved expression among the two tomato species also showed an enrichment for secondary metabolites including the polyamine, spermidine (Fig. 4.1).

We performed an analysis to sort the differentially expressed genes into clusters with similar expression profiles over time. This unsupervised method produced six profiles, and for each profile we performed a GO analysis (Fig. 4.3B-G and Supplemental Figure

4.5). Interestingly, clusters 1, 3, and 5 have roughly complementary patterns to clusters 2, 6, and 4, respectively. Clusters 1 and 3 both contain several terms related to protein modification or degradation, while cluster 5 is primarily enriched for lipid and fatty acid biosynthesis. Clusters 2, 4, and 6 generally have a pattern of decreasing expression over time, and these clusters are all enriched for very basic metabolic functions such as DNA replication, translation, and biological processes. This decrease in expression could reflect the beginning of senescence and a general cessation of active metabolic processes.

Solanaceae Expression Patterns Align with Prominent Developmental Processes

The tomato species differ in fruit type from desert tobacco, and we wanted to know the extent to which expression patterns are conserved (or not) among the fruit of these phenotypically diverse, but relatively closely related taxa. To answer this we used OrthoFinder2 to find single-copy orthologous genes from dry-fruited desert tobacco and both fleshy-fruited tomato species together (Emms & Kelly, 2019). Because we were unable to obtain RNA from mature desert tobacco capsules, these datasets are sampled at four comparable developmental stages (Table 1). We then applied two nested statistical models to test for differential expression over time that was conserved among all species or divergent between fruit types.

Only 1,235 single-copy orthologs showed a statistically significant conservation of expression pattern across all three species. As a cohort, this comparatively small number of genes was enriched for five GO terms, including DNA replication and protein phosphorylation (Fig. 4.4A). To examine finer scale patterns among these genes, we

performed unsupervised clustering followed by a GO analysis of the genes in each cluster. This revealed seven profiles of gene expression patterns over time (Supplemental Figure 4.6). The expression patterns and GO term enrichments for the clusters largely agree with prominent developmental processes at various stages. For instance, cluster 3 has highest expression at stages 1 and 2 and is enriched for several terms related to DNA replication, which is known to occur early in fruit development (Fig. 4.4B-C) (Gillaspy et al., 1993; Pabón-Mora & Litt, 2011; Tanksley, 2004).

Our search for single-copy orthologs that have statistically significant differences in expression pattern between fruit types yielded 4,647 genes. A GO term analysis of this set of genes revealed terms underlying known phenotypic differences between these two fruit types including terpenoid biosynthetic processes, which are likely related to flavor compound production, as well as polysaccharide catabolism, cellulose biosynthesis, glycolytic processes, and carbohydrate derivative metabolism, which could relate to the differential accumulation of sugars and/or cell wall composition between these fruit types (Fig. 4.4D). Unsupervised clustering and GO analyses were also carried out on this dataset; however, this did not yield readily informative patterns or terms (Supplemental Figure 4.7).

Solanaceae Orthologs of Ripening-Related Genes Show Fruit Type-Specific Expression Patterns

Given the interesting differences between wild and cultivated tomato in expression of the ripening related structural and regulatory genes , we asked to what extent the expression pattern of these genes has diverged between the fleshy-fruited tomato species

and the dry-fruited desert tobacco. We restricted our analysis to genes that had a single unambiguous ortholog in all three species and found orthologs for four of 12 ethylene-related structural genes and five of 18 transcription factors (Table 2). We then pooled replicates from both tomato species as a single representative fleshy-fruited taxon and contrasted their expression values with those from desert tobacco. This effectively averages differences in expression that may have been apparent between wild and cultivated tomato but allows us to search for genes with strong signal of fruit-type specific expression over time. Using a likelihood ratio test, we were able to discern if the expression patterns show conservation between fruit types, within fruit types, or are divergent between fruit types.

Interestingly, all nine of the genes for which we determined orthology show a decrease in expression between stage 3 and the transition stage of the desert tobacco capsule (Fig. 4.5). This result echoes that seen in desert tobacco clusters 2, 4, and 6 from the entire cohort of 1,392 differentially expressed genes, suggesting again that there may be a trend toward gradual ramping down of metabolic processes as the fruit begins to senesce.

Among the ethylene-related structural genes, we found orthologs for *ACO4, ACO5*, *ACO6*, and *NR/ETR3* (Fig. 4.5A-D). *ACO4, ACO5*, and *NR/ETR3* each have statistically significant differences in their expression patterns between the fruit types (p=1.01x10^{-9}, 8,9x10^{-20}, and 1.8x10^{-5}, respectively). *ACO6* is differentially expressed over developmental time but this pattern is different in each of the three species. The lack of conservation for the *ACO6* expression pattern is likely due to the differences in expression among the two tomato species, which have nearly opposite patterns of expression over time. Interestingly, for *ACO5,* all desert tobacco timepoints show higher expression

magnitudes than in tomato, and for *ACO6* desert tobacco shows higher expression than cultivated tomato. However desert tobacco capsules are non-climacteric fruits and the high expression of these ethylene biosynthetic genes suggests that the involvement of ethylene in maturity of desert tobacco and other dry fruits deserves further study.

Among the transcription factors, we resolved unambiguous, single-copy orthologs across the three species for *AGL11, FYFL*, *SPL-CNR*, *TAG1*, and *TAGL1* (Fig. 4.5E-I). Only *FYFL* and *TAG1* lacked statistically significant conservation of expression pattern among the three species (Fig. 4.5F,H). In contrast to our tomato comparisons, *AGL11*, which did not show statistically significant differences between tomato species, does show statistically significant differences between fruit types (Fig. 4.5E, p=6.5x10^{-3}, 4.5x10^{-5}, and 1.7x10^{-5}). As mentioned previously, the role of *AGL11* and its paralog *MBP3* in the pericarp is unclear at present, but the statistically significant divergence in expression pattern of *AGL11* between fruit types and of *MBP3* among tomato species highlights the need for further study of these gene functions following their duplication.

Orthologs of *SPL-CNR* and *TAGL1* both showed statistically significant conservation in their expression patterns by fruit types (Fig. 4.5G-H, p=5.4x10^{-17} and 5.6x10^{-3}). The Arabidopsis ortholog of *TAGL1* promotes the formation of the dehiscence zone in the pericarp of that dry fruit (Ferrándiz et al., 2000). In our analysis, the pattern of expression for *TAGL1* is higher overall in dry fruited species and peaks at stage 3 as the dehiscence zone is forming. This provides some evidence for the functional conservation of this gene's role in dry fruit dehiscence. For *SPL-CNR*, we observe roughly opposing patterns of expression between dry and fleshy fruits. *SPL-CNR* increases in expression as fleshy fruits enter the breaker stage, before they have begun to soften. In contrast, we see a decrease in *SPL-CNR* expression as dry fruits approach dehiscence. Additional

functional studies of this gene's role across dry-fruited species could help extend its established role in cell-cell adhesion and clarify its potential role in dry fruit maturity.

Very Few Genes Show Conservation Of Expression Pattern Between Dry and Fleshy Fruit

Our analysis of the tomato species and desert tobacco revealed a number of informative patterns, but all three species belong to the same family. As a result, we cannot tell if common patterns of gene expression are due to shared phylogenetic history or represent trends across angiosperm fruit development. We wanted to find generalizable trends in gene expression that might underlie the divergence between dry and fleshy fruit development or support conservation of certain gene expression patterns between these two phenotypically diverse fruits. We therefore chose to add *Arabidopsis thaliana*, which produces a dry silique and melon (*Cucumis melo*), which produces a type of berry with a leathery rind known as a pepo.

In order to enable expression comparisons between and among species, we used Orthofinder2 to group genes from these species into orthologous groups based on protein sequence similarity and phylogenetic relationships (Emms & Kelly, 2019). Due to their high degree of similarity, and because we had mapped wild tomato RNAseq using the cultivated tomato genome, we used cultivated tomato protein sequences in the orthology search to represent both cultivated and wild tomato. For subsequent gene expression analyses, however, the two tomato species were not combined. We were able to group the genes from these species into 19,249 orthogroups (Fig. 4.6A); however, many orthogroups were not shared among all species, and even among universally shared

orthogroups, there were many cases of gene family expansion or loss within a single species. Because comparing transcript levels among unequal numbers of genes across species is not meaningful, we limited our interspecific expression analysis to only single-copy genes falling into universally present orthogroups. This filtering left 4,163 orthogenes for comparisons among both tomato species, desert tobacco, Arabidopsis, and melon (Fig. 4.6B).

For these five species, we wanted to use comparable developmental stages to see if any orthologous genes shared similar expression dynamics over time among all species or among species with similar fruit types. After integrating the publically available Arabidopsis and melon pericarp RNAseq data with our own tomato and desert tobacco datasets, we had comparable data for stage 2, stage 3, and transition stage in all species (Table 1).

We first assessed the extent to which any of the 4,163 orthologous genes were differentially expressed over time and shared a conserved pattern across all five species. To call differential expression across the three stages, we used a model (Model 1) that is blind to species but requires a gene to have a statistically significant change in expression between at least two stages in order to be differentially expressed. Surprisingly, this resulted in only 121 orthologous genes with a pattern of differential expression over time that is the same in all 5 species (Supplementary File 1). To determine if the expression data from these genes showed a detectable signal based on developmental stage, species, or fruit type, we conducted a principal component analysis using their expression values (Fig. 4.7A-C). Model 1 did not consider species in calling differentially expressed genes, and in fact the variance explained by the first five principal components appears not to have strong signal for interspecific differences. The notable exception to this is PC2,

which explains 15% of the variance and seems mostly to separate melon from the other four species; however, PC2 also separates stage 2 from later stages in both tomato species as well as stages 2 and 3 from the transition stage in Arabidopsis. PC1 explains 35% of the variance and largely distinguishes the breaker stage tomato samples from all other samples. PC3 serves to differentiate the three developmental stages of tomato from one another and also separates stage 2 samples from later stages in Arabidopsis. The developmental stages of melon are weakly distinguished by PC4 and more prominently by PC5, each of which explain 7% of the variance. PC5 also weakly separates the developmental stages of desert tobacco.

To categorize these 121 genes, we performed a GO term enrichment analysis and found a number of terms relating to prominent processes common across fruit development including cell proliferation, anatomical structure formation, cytokinesis, and cell wall modification (Supplemental Figure 4.8). Looking at shared expression patterns among these 121 genes showed only two clusters of expression profiles. Cluster 1 contains genes whose expression increases between stage 3 and the transition stage, while cluster 2 contains genes whose expression is generally decreasing during fruit development. GO terms associated with cluster 1 were too broad to be informative, however cluster 2 was enriched for several terms related to cell and cellular component organization, which is consistent with the high expression of this cluster during the early patterning and cell division phases of development.

The very small number of genes with conserved patterns across all five angiosperm species further suggests that it may be possible to define a core set of pericarp development-related genes that have a conserved function despite large divergences in both evolutionary time and in phenotype.

Divergence in Expression of Genes Related to Cell Division, Plastid Localization, and Secondary Cell Wall Composition Between Dry and Fleshy Fruits

Having established that few orthologous genes have conserved expression patterns across all five species, we next asked if and to what extent genes might show conservation of expression patterns within, but not between, fruit types. We reasoned that these fruit-type specific patterns could shed light on developmental processes shared by evolutionarily distant species with a common phenotype, dry or fleshy fruits. To answer this question, we created a model to call differentially expressed orthologous genes (Model 2) that is aware of fruit type for each of the five species but is blind to the species themselves. Like Model 1, which we used to find conserved patterns across all species, Model 2 also requires that a gene have a statistically significant change in expression between at least two of the three developmental stages. Because Models 1 and 2 are nested, genes are only differentially expressed by Model 2 if their expression pattern is better explained by Model 2 than by Model 1, as determined by a likelihood ratio test. This ensures that the difference in fruit type is driving the determination of differential expression.

Interestingly, Model 2 determined that nearly half of the 4,161 single-copy orthologous genes had divergent patterns of expression between dry and fleshy fruited species (Supplementary File 2). We performed a principal component analysis to see if any grouping by species, developmental stage, fruit type, or evolutionary distance might be driving this large number of differentially expressed genes (Fig. 4.8A-C). In this analysis, the first three principal components, which collectively explained 81% of the

variance, served primarily to distinguish among the species. PC1 accounted for the majority of the variance (54%) and separated the dry and fleshy fruited species. On PC1, desert tobacco was separated from the two tomato species, but not as dramatically as Arabidopsis from melon, suggesting that PC1 might also incorporate some amount of variance due to phylogenetic distance in addition to fruit type. Similarly, PC2, which explained 19% of the variance, did not separate the two dry-fruited species, but placed tomato and melon at two extremes. PC2 therefore combined both dry fruits but distinguished between two categories of fleshy fruits. PC3, which accounted for 8% of the variation, only seemed to separate desert tobacco from the other four species. PC4 and PC5 captured 3% and 2% of the variance, respectively, and showed a striking perpendicular separation of developmental stages in tomato and Arabidopsis, but placed both melon and desert tobacco at their intersection, roughly overlapping with stage 3 of tomato (Fig. 4.8C). Interestingly, in contrast to PC1-3, which primarily separated species, PC4 was the only principal component we examined that was able to separate the two tomato species, and even here the separation was only evident for the breaker stages samples.

To determine what sorts of genes were captured by this model, we performed a GO enrichment on all 1,795 genes (Fig. 4.8D). In contrast to the very focused enrichment seen in Model 1, the genes from Model 2 were enriched for more diverse terms. In fact, the enrichment of the very high-level metabolic processes term with 757 associated genes highlights the diversity of functions that separate pericarp development in dry- and fleshy-fruited species. Even lower-level enriched terms fall into very disparate categories such as protein trafficking, secondary metabolite synthesis and regulation of gene expression.

Because of the diversity of functional terms in the GO analysis of the entire cohort of genes, we next asked in what ways the patterns of expression diverged between fruit types and what sorts of genes displayed these patterns. Our clustering analysis resulted in eight expression profiles, and we performed a GO analysis on each cluster (Supplemental Figure 4.9). Interestingly many, but not all, of these clusters showed distinctive expression profiles with more focused enrichments. In cluster 4 the magnitude of expression diverges over time between dry and fleshy fruits, with fleshy fruits showing higher expression (Fig. 4.9A). This cluster was enriched for several terms relating to glucose and polysaccharide synthesis, which could correspond to the accumulation of sugars in fleshy fruits as they begin to ripen (Fig. 4.9B). Similarly, in cluster 6, dry fruits show the same pattern as cluster 4, but fleshy fruits show a slight drop in gene expression at stage 3 followed by a larger drop at the transition or breaker stage (Fig. 4.9C). This cluster is enriched for terms relating to DNA replication and cytokinesis, likely related to the burst of cell division in stage 2 of fruit development followed by the endoreduplication that occurs in stage 3 of tomato pericarps (Fig. 4.9D). At the transition or breaker stage of tomato fruit development, chloroplasts are known to reorganize and convert to chromoplasts, which store the conspicuous red pigments. This process is reflected in cluster 7 where dry fruits slowly drop in expression over time, but fleshy fruit show a jump in expression at the transition stage (Fig. 4.9E). This cluster is enriched for a number of terms relating to plastid remodeling and trafficking (Fig. 4.9F). Finally cluster 8 highlights the key feature of dry fruit pericarps, which deposit lignin polymers in their secondary cell walls as they develop. In cluster 8, dry fruit expression remains moderate, while fleshy fruit expression values drop and remain low following stage 2 (Fig. 4.9G). GO terms enriched in this cluster include a number of cell wall biogenesis terms (Fig. 4.9H). Overall

the profiles and enrichments seen in these clusters support a number of hypotheses regarding differential expression developmental processes separating dry and fleshy fruits and provide a basis for more direct studies of function divergence (or conservation) between these diverse fruit types.

Methods

Plant Materials

Seeds for *Solanum lycopersicum* cv. Ailsa Craig and *Solanum pimpinellifolium* (LA 2547) were provided by the UC Davis Tomato Genetics Resource Center, and those for *N. obtusifolia* (TW143) were obtained from the New York Botanical Garden. We grew all plants in a temperature controlled greenhouse at 26ºC on the campus of the University of California, Riverside.

Developmental Staging

For *Solanum spp*., we chose five developmental time points for sampling, corresponding to widely accepted stages in fruit development (Gillaspy et al., 1993): early ovary development until fruit set, initiation of cell division, initiation of cell differentiation, and ripening or maturity. For *Solanum spp*., we divided the ripening stage into a transition or “breaker” stage and true physiological maturity. The same schema was applied in the dry-fruited *N. obtusifolia*, except for physiological maturity, which is highly lignified and fully senesced. Because of the difficulty obtaining usable RNA from this stage, we did not include it for *N. obtusifolia* (Table 1).

To determine the timing of the early stages, we conducted serial sectioning and staining on a series of greenhouse-grown pericarps from each species. We collected fruit and ovary tissue from 0-15 DPA and trimmed them to roughly 1cm cubes as needed. We vacuum infiltrated (-0.08Mpa) these in FAA consisting of 10% formaldehyde, 50% ethanol, and 5% acetic acid in distilled water overnight and then stored them in 50% ethanol for later use. Before embedding the fixed tissue for sectioning, we first dehydrated them through an ethanol series ending with a final absolute ethanol dehydration overnight. Across two two-hour incubations at room temperature, we replaced the ethanol with 50% ethanol/50% Citrisolv (Decon Labs, King of Prussia, PA) followed by 100% Citrisolv. We then added paraffin chips, placed the samples in a 60ºC oven, and replaced the solution with liquid paraffin approximately 7 times over the next two days. After we could no longer smell the Citrisolv, we placed the tissue in aluminum crinkle dishes (VWR, Radnor, PA) to solidify before shaping and mounting them for sectioning the next day. We sectioned the blocks into 8-10μM thick ribbons and affixed them to microscope slides.

We stained high-quality, representative sections with Safranin O and Astra Blue. To deparaffinize the tissue slides we washed them twice for five minutes each in xylene, and followed this by rehydration through an ethanol series. We first stained in Safranin O (1% w/v in water) for 60 minutes, rinsed them twice with deionized water and then counterstained with Astra Blue (1% w/v in a 2% tartaric acid solution) for 10 minutes. We then rinsed the slides twice in water, and dehydrated them through the same ethanol series before rinsing twice with xylene. We then affixed a coverslip with permount and dried the slides at 40º overnight. We imaged the slides to count cell layers and observe cell size increases in the case of *Solanum spp.* and to observe lignification in the case of *N. obtusifolia*.

To determine the timing of stage 2 (cell division) in *N. obtusifolia* we observed fruits for a conspicuous jump in size and a shift in fruit apical shape from conical to blunted.

RNA Extraction and Sequencing

For all three species, we hand dissected pericarps on ice from developing fruits and, in the case of earlier developmental stages, pooled multiple pericarps from a single individual to obtain enough tissue for RNA isolation. Each biological replicate represents pericarps from a single plant. We snap froze dissected tissue in liquid nitrogen, ground each sample with a micropestle attached to a cordless drill, and isolated RNA with the RNeasy Plant Mini Kit (QIAGEN, Hilden, Germany) according to the manufacturer's protocol. For *N. obtusifolia* the lysis step of this protocol was modified to use buffer RLC instead of RLT and supplemented with 2.5% (w/v) polyvinylpyrrolidone (PVP). DNA contamination was removed with an on-column RNAse-Free DNAse kit (QIAGEN, Hilden, Germany) according to the manufacturer's protocol.

The UCR Genomics Core assessed the integrity of the isolated RNA using an Agilent 2100 Bioanalyzer. We prepared high-quality samples into Illumina RNA-sequencing libraries using the NEBNext Ultra II Directional RNA library Prep Kit for Illumina (New England BioLabs, Ipswich, MA, United States) and barcoded each library for multiplexing with the NEBNext Multiplex Oligos for Illumina (Index Primers Set 1) kit. Both protocols were undertaken according to the manufacturer's instructions.

Libraries for *S. lycopersicum*, *S. pimpinellifolium*, and *N. obtusifolia* were sequenced at the UCR Genomics Core. All *Solanum* libraries and the stage 1-3 libraries of *N. obtusifolia* were sequenced on the Illumina NextSeq V2 with a high-output 2x75bp run. The Stage 4 libraries were sequenced as part of an Illumina NextSeq 1x75bp run.

Raw sequence reads for all 42 pericarp libraries are available under NCBI BioProject PRJNA646747.

Bioinformatic Analyses

All scripts used to analyze RNA-seq data for this study are publically accessible in a GitHub repository (github.com/rajewski/SolTranscriptomes).

We downloaded the raw RNA-seq reads for the *Arabidopsis thaliana* and *Cucumis melon* experiments (PRJEB25745 and PRJNA314069, respectively, Table 1) from the Sequence Read Archive (Chayut et al., 2017; Mizzotti et al., 2018). We trimmed the demultiplexed RNA-seq data with TrimGalore (Krueger, 2012) and mapped reads using STAR v2.5.3a (Dobin et al., 2013). Because of the low continuity of the *S, pimpinellifolium* reference genome, we mapped RNA-seq reads for both *Solanum* species to the *S. lycopersicum* (SL4.0) genome assembly (Hosmani et al., 2019). For *N. obtusifolia,* we mapped the reads to version 1 of the *Nicotiana obtusifolia* reference genome assembly (Xu et al., 2017), for *Arabidopsis thaliana* data, we mapped reads to the TAIR10 assembly (Berardini et al., 2015), and for melon, we mapped read to the *Cucumis melo* cv. DHL92 genome (Garcia-Mas et al., 2012).

We used the program OrthoFinder2 (Emms & Kelly, 2019) to cluster the genes from the five species into orthologous groups based on protein sequence similarity. Within the framework of the OrthoFinder2 pipeline, we opted for gene tree estimation using multiple sequence alignments with MAFFT (Katoh & Standley, 2013) followed by IQ-Tree (Nguyen et al., 2015) instead of the default DendroBLAST algorithm (Kelly & Maini, 2013). To obtain a more tractable dataset for differential expression analyses, we eliminated

orthologous groups with paralogs and filtered the results for single copy genes common to all species.

Because our experimental design contained several sequential timepoints and multiple species, pairwise comparisons with time points coded as unrelated categorical variables would fail to intuitively capture the dynamic nature of gene expression and would suffer from a severe multiple testing problem. Similarly, treating time as a linear predictor of gene expression would fail to identify transiently up-regulated genes. To avoid this problem, we opted instead to implement a natural cubic spline basis transform of the time coordinates, as outlined in the supplemental material of (Fischer et al., 2018). For differential expression testing, a gene (or orthogene) is determined to be differentially expressed if its expression profile is better fit by this spline model than by a model incorporating only noise, as determined by a log ratio test. Additionally, for orthogene comparisons between fruit types, an orthogene may be differentially expressed if its expression profile is better fit by a model incorporating interaction between the fruit type (categorical) variable and the spline basis function coefficients than by a model with only the spline coefficients. We conducted these analyses in R using the DESeq2 and splines packages (Love et al., 2014; R Core Team, 2019). We then clustered genes determined to be differentially expressed using the DIANA algorithm of divisive clustering (Kaufman & Rousseeuw, 2005) as implemented by the R package DEGreport (Pantano, 2019). We interrogated groups of similarly expressed genes using several methods. To test for enrichment of Gene Ontology (GO) terms, we queried all protein sequences extracted from the reference genomes against the PFAM, ProSiteProfiles, TIGRFAM, and PRINTS databases (Attwood et al., 2012; El-Gebali et al., 2019; Haft et al., 2001; Sigrist et al., 2013) and aggregated all associated GO terms for each protein using a custom shell

script. We then used the R package topGO (Alexa & Rahnenfuhrer, 2016) to test for enrichment of GO terms using Fisher's Exact Test and the "weight01" algorithm against a background set of all GO terms in the genome (or in the set of orthologous genes) using a custom R script.

Discussion

Across angiosperm evolution there have been repeated transitions from ancestral dry fruits to derived fleshy fruits, often with dramatic consequences. Although the morphological and developmental basis of these transitions have been well-documented, the underlying molecular and genetic mechanisms that enable or hinder these transitions have received less attention. Here we present evidence for a small set of "core" genes whose patterns of differential expression during pericarp development are conserved across several angiosperm taxa. We also show that a much larger set of "accessory" genes exists with patterns of differential expression during pericarp development that are similar within but different between dry- and fleshy-fruited species. The expression patterns of these core and accessory genes corroborate a number of phenotypic observations regarding differences in dry and fleshy fruit cell wall composition, cell division, and secondary metabolite production. Interestingly, these expression patterns also raise new questions about the role of ethylene in dry fruit maturity as well as the role of additional transcription factors in dry fruit dehiscence.

At lower taxonomic levels, our data also highlight a number of gene expression differences correlated with the domestication of tomato (*S. lycopersicum*) from its wild ancestor (*S. pimpinellifolium*) and provide further genetic support for previously noted phenotypic differences in fruit size, firmness, and lignification.

Tomato Domestication-Related Genes

Although wild and cultivated tomato species share a number of genetic and morphological similarities, cultivated tomato has undergone quite strong artificial selection (Blanca et al., 2015). The effects of this artificial selection are quite pronounced on the fruits, which are larger, sweeter, and firmer in cultivated than in wild tomato. We detected signs of this domestication in our pericarp gene expression dataset.

Profiling the expression of 21 ethylene- and flavor compound-related structural genes as well as 18 regulatory genes implicated in fruit ripening, we found a few key differences in expression pattern between wild and cultivated tomato (Fig 4.2, Supplemental Fig. 4.4, and Supplemental Fig. 4.4). The gene *TomLoxC*, which encodes a lipoxygenase, contributes to desirable flavor in tomato fruit and showed different expression patterns between wild and cultivated tomato (Chen et al., 2004; Shen et al., 2014). This locus was previously identified as a target of selection during the domestication of tomato (Gao et al., 2019). The ethylene biosynthesis gene *ACO6* was the only ethylene-related gene in our dataset that showed different patterns of expression between wild and cultivated tomato, with expression of this gene higher at all stages of pericarp development in wild tomato (Fig. 4.2A). As we extended our analysis to include the dry-fruited desert tobacco pericarp transcriptome, we also saw comparatively high levels of *NoACO6* expression (Fig. 4.5C). In fact, the levels of *NoACO6* expression were higher than in cultivated tomato throughout pericarp development and also higher than wild tomato at Stages 1 and 2, which are characterized by ovary patterning and cell division. We also saw higher expression across pericarp development for another ethylene biosynthetic enzyme *ACO5* in desert tobacco as compared to the two tomato species (Fig.

4.5B). Higher expression of ethylene biosynthetic enzymes in this dry fruit is counterintuitive and highlights the need for further study of the roles these specific enzymes, and ethylene more generally, play in the ripening and maturity of dry fruits.

Among the regulatory genes, *MBP3* was expressed at higher levels in wild tomato following the stage of pericarp cell division (Fig 4.2D). The precise role of *MBP3* in tomato is unknown, but its paralog *AGL11* and their mutual ortholog in Arabidopsis both act to specify ovule identity (Huang et al., 2017; Ocarez & Mejía, 2016; Pinyopich et al., 2003). The role of these ovule identity genes in the pericarp is unclear at present, however the grape ortholog of these genes, *VvAGL11*, is adjacent to a QTL that controls both seedlessness and fruit size (Mejía et al., 2011). It could follow then that the differences in *MBP3* expression and in fruit size between wild and cultivated tomato, represent possible subfunctionalization following the duplication that produced *AGL11* and *MBP3*.

We also detected species-specific patterns of expression for the transcription factors *TAG1* and *TAGL1* between wild and cultivated tomato (Fig 4.2F-G). Beyond their roles in organ identity, both *TAG1* and *TAGL1* have been shown to contribute positively to pericarp thickness; however, our results show higher expression for these genes in wild tomato, which has a thinner pericarp (Gimenez et al., 2016). Apart from this role in pericarp thickness, numerous orthologs of *TAGL1* are well documented to promote lignification of the pericarp (Ferrándiz et al., 2000; Giménez et al., 2010; Gimenez et al., 2016). We were curious if this difference in *TAGL1* expression between our two tomato species also correlated with changes in expression of structural genes involved in lignin biosynthesis. We queried our results for interspecific expression differences in the first three enzymatic steps of lignin biosynthetic (*SlPAL*: *Solyc09g007920*, *SlC4H*: *Solyc06g150137*, *Sl4CL.1*: *Solyc03g117870*, *Sl4CL.2*: *Solyc06g068650*, and *Sl4CL.3*: *Solyc12g042460*) as well as

two enzymes at branch points of the pathway (*SlHCT*: *Solyc03g117600* and *SlF5H*: *Solyc02g084570*). We found that *SlHCT*, the first committed step in the formation of G- and S-type lignin, shows a statistically significant difference in expression pattern between wild and cultivated tomato (p=0.022, likelihood ratio test). This result suggests that, although neither fruit accumulates lignin to substantial levels, there may have been selection against pericarp lignification during tomato domestication. Extending the characterization of *TAGL1* to include desert tobacco, we also saw differences in expression for this gene between fruit types, with higher expression the desert tobacco *TAGL1* ortholog, *NoSHP* from Stages 1 through 3 of fruit development (Fig. 4.5I). This result supports potential conservation of the role *NoSHP* is expected to play in patterning dehiscence zones across evolutionarily divergent dry fruits (Ferrándiz et al., 2000).

Finally, we found support for expression differences in *SPL-CNR* between wild and cultivated tomato (Fig. 4.2E). Although the pattern of expression for both species shows an upward trend between Stage 2 and Breaker stage, the increase is more dramatic for wild tomato. *SPL-CNR* is believed to be the causative locus underlying the *Colorless non-ripening* (*Cnr*) mutant in tomato (Manning et al., 2006). Disruption of *SPL-CNR* in the *Cnr* mutant results in fruits that fail to soften or undergo color change at the ripening stage, and this has been related to changes in cell wall composition and cell-cell adhesion (Eriksson et al., 2004; Lai et al., 2020). Although both species of tomato turn red and soften at maturity, that is, neither species displays the extreme *Cnr* phenotype normally, there are quantitative differences in fruit firmness between them. Two large-scale QTL mapping studies of wild and cultivated tomato advanced backcrosses discovered six QTL for fruit firmness, and wild tomato alleles at four of those QTL are shown to decrease fruit firmness (Doganlar et al., 2002; Tanksley et al., 1996). Because soft fruits are more easily

damaged during harvest and less desirable to consumers, increasing fruit firmness for cultivated tomato is one target of breeding programs (Barrett et al., 2010). *SPL-CNR* might help increase fruit firmness though its role in cell-cell adhesion, and thus differences in *SPL-CNR* expression between these tomato species could be related to differences in fruit firmness, although many other loci are likely at play. Additionally, the established role of *SPL-CNR* in promoting cell-cell adhesion in tomato has led other authors to speculate that this gene might also play a role in dry fruit dehiscence (Eriksson et al., 2004). If this gene's function in cell-cell adhesion is conserved among diverse fruit types, then the difference in expression patterns for *SPL-CNR* between fruit types in our analysis is also suggestive of a potential role in dry fruit dehiscence. Including desert tobacco expression data, we observe roughly opposing patterns in *SPL-CNR* expression between dry and fleshy fruits (Fig. 4.5G). *SPL-CNR* increases in expression as fleshy fruits enter the breaker stage, before they begin to soften. In contrast, we see a decrease in *SPL-CNR* expression as dry fruits approach dehiscence, where loss of cell adhesion allows the fruit to split open. Additional functional studies of this gene's role across dry-fruited species could help extend its established role in cell-cell adhesion and clarify confirm its potential role in dry fruit maturity dehiscence and the potential conservation of function across fruits.

In this study, we mapped RNAseq reads from both wild and cultivated tomato to the cultivated tomato reference genome (Hosmani et al., 2019), and in our searches for orthologous genes, we used cultivated tomato sequences as a proxy for both wild and cultivated tomato. This simplified our interspecific comparisons, and mitigated the fact that the genome assemblies of wild tomato are not thoroughly annotated (Razali et al., 2018). Although these decisions enabled better interspecific comparisons, it means we are unable to examine the role of gene duplications and mutations that may have arisen since

wild and cultivated tomato split. However, this is unlikely to drastically affect our results since any gene duplications specific to a single species are filtered out of our interspecific comparisons.

Core and Accessory Pericarp Gene Expression Patterns

By examining the expression patterns of single-copy, orthologous genes across wild and cultivated tomato, desert tobacco, Arabidopsis, and melon, we were able to find evidence for two groups of genes in dry and fleshy fruit development, which we have termed the core and accessory genes. The core genes comprise a set of 121 orthologs whose expression patterns in the pericarp are conserved among all five species, while the accessory genome includes 1,795 orthologs whose expression patterns are each similar within fruit types but which show difference between fruit types.

Not all of the 121 core genes have been thoroughly characterized, so at present it is not possible to give a full inventory of functions, but the list suggests common developmental mechanisms that may be necessary for pericarp development. Orthologs for many of these core genes have annotated functions in processes of cell division and cell wall synthesis including the gene *KNOLLE* (*AT1G08560*), which helps pattern the rate and plane of cell divisions (Lukowitz et al., 1996). However other prominent structural genes for cellulose synthase, pectin methylesterase, and pectin lyase, and microtubule organizing proteins are also present (*CESA4*, *AT5G44030; PME5*, *AT5G47500; AT5G19730; CORD3, AT4G13370*; *CORD7*, *AT2G31920*; *FUSED*, *AT1G50240*). Other genes in this set have orthologs with annotated function in developmental patterning. For example, the Arabidopsis gene *ARABIDOPSIS CRINKLY4* (*ACR4*, *AT3G59420*) functions in pattern epidermal cells, root asymmetric cell divisions, and cuticle deposition, while

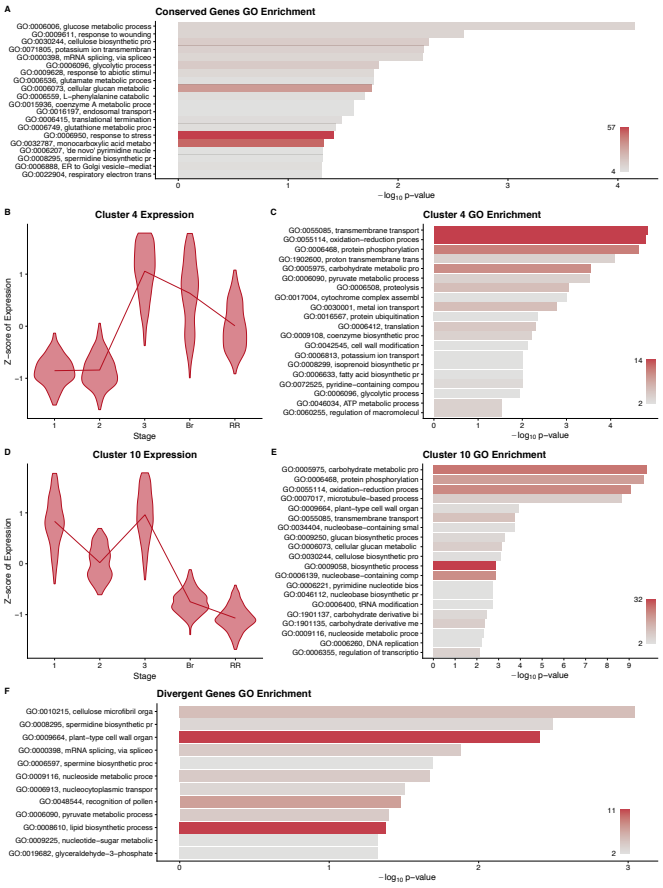
PERIANTHIA (*AT1G68640*) helps determine floral organ number (De Smet et al., 2008; Running & Meyerowitz, 1996; Watanabe et al., 2004). Beyond the expected cell division and pattern genes we also found several brassinosteroid-related genes as well as *ARGONAUTE7* (AGO7, *AT1G69440*) in this set of core genes. *AGO7* is involved in tasiRNA formation and ultimately helps to regulate development progression from vegetative to reproductive stages as well as leaf morphology in an auxin dependent manner (Adenot et al., 2006; Montgomery et al., 2008). The genes *DWARF4* (*DWF4*, *AT3G50660*) and *TITAN-LIKE* (*TTL, AT4G24900*) are involved in brassinosteroid biosynthesis and growth-responses, respectively (Azpiroz et al., 1998; Lu et al., 2012). The dwarfed phenotype of *dwf4* mutants is related to reduced cell elongation but not cell division, whereas the *ttl* mutant was first characterized based on an endosperm defect nuclear division defect. The dry and fleshy fruits studied here differ in a number of ways from one another, but overall size, especially in the pericarp tissues we sampled is one very conspicuous difference (Gillaspy et al., 1993). The overall size of a plant organ can be decomposed into the number of cells present and their sizes, so it is interesting that the brassinosteroid related genes in the core set of genes have complementary effects, modulating cell size and nuclear divisions, respectively.

Although our dataset includes eudicot plants from several families, we believe that the addition of more taxa could help refine this set of core and accessory genes. Because our method is based on patterns among shared, single-copy orthologs however, including additional very distantly related plants or plants with extremely reduced genomes would not be beneficial. We examined patterns of expression for approximately 5,000 orthologs in our five-species comparisons, and this number of orthologs is based not only on the presence of true orthologs among species, but also our ability to confidently identify

orthologs. Including more taxa would likely reduce the number of true single-copy orthologs, but because the determination of orthology is based upon finding clusters of proteins with similar sequence and resolving a phylogenetic relationship among them, additional genes could produce more informative gene trees and help increase ortholog numbers.

Figures and Tables

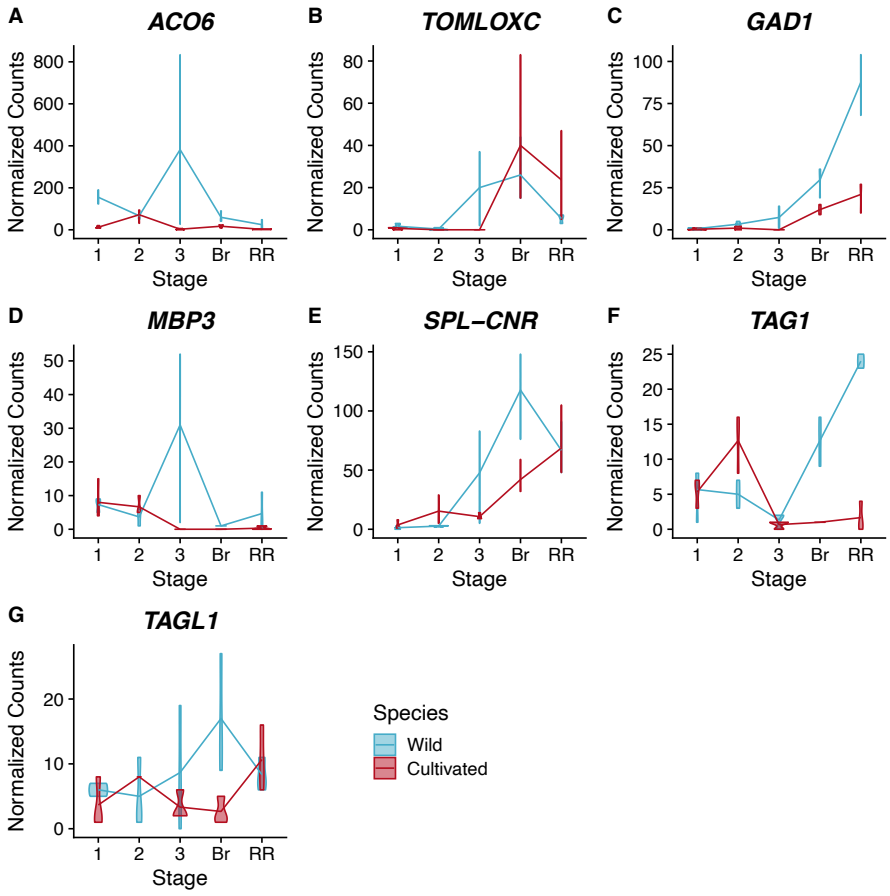
Figure 4.1 – GO Enrichment and Clustered Expression Patterns in Tomato Species



Summary of gene expression patterns conserved (A-E) or divergent (F) among cultivated and wild tomato. A gene ontology (GO) term enrichment analysis (A) performed on all differentially expressed genes without regard to species. Selected clusters of differentially expressed genes conserved among species are described with violin plots of normalized expression at each stage of development (B and D) and with GO enrichment analyses (C and E). For differentially expressed genes with divergent expression between the species, we performed a GO enrichment analysis (F). GO term descriptions to the left of the enrichment graphs are truncated for space and sorted by p-value. The bars are colored by the number of genes assigned to each GO term with legends in the lower right of each graph. Stages of fruit development in the axis of B and

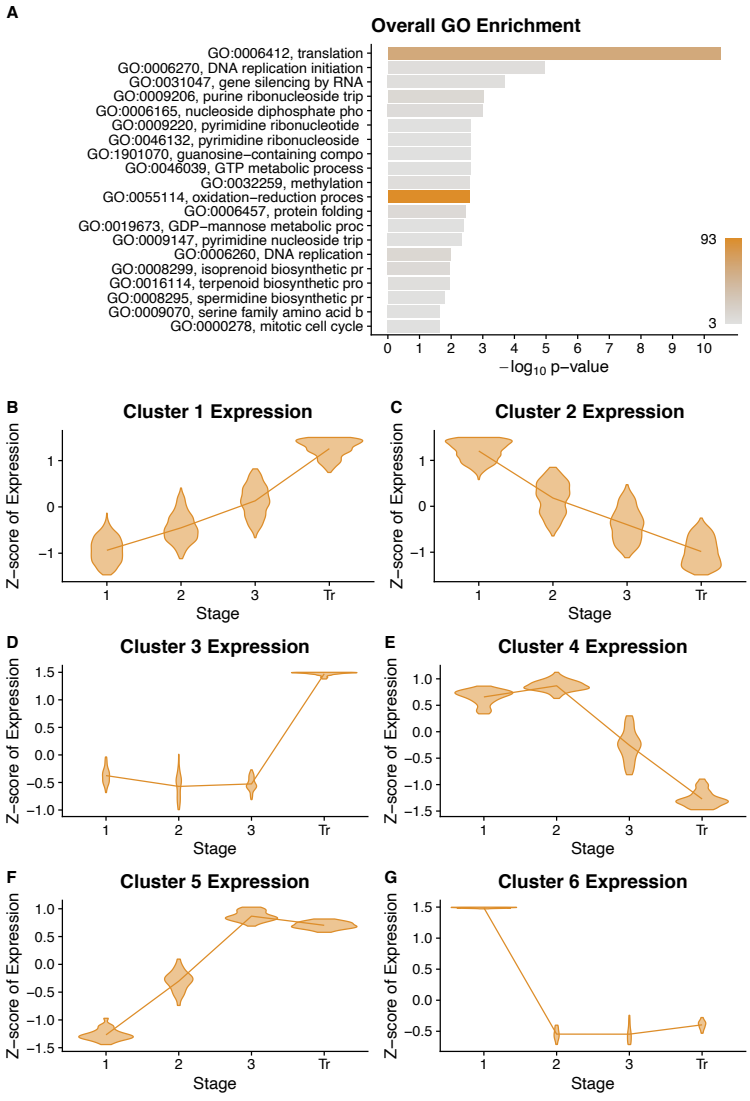
D are numbered sequentially followed by “Br” for breaker stage and “RR” for red ripe stage.

Figure 4.2 – Expression of Selected Tomato Genes



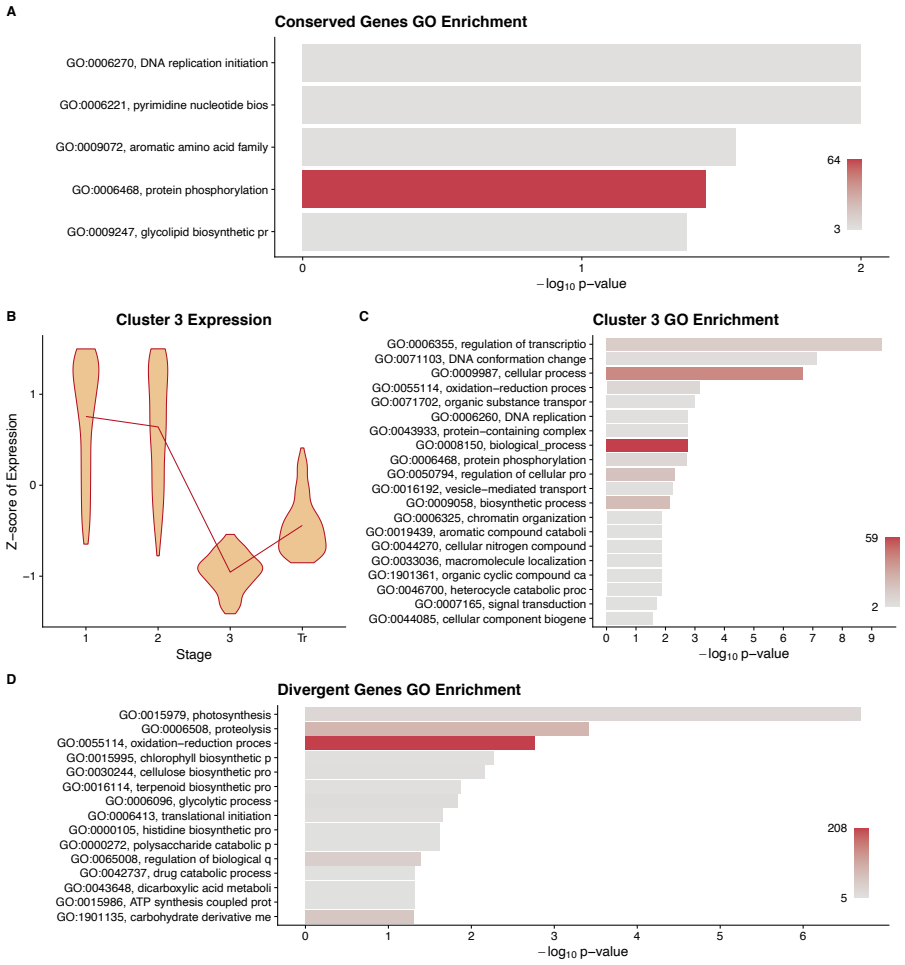
Expression profiles for ethylene-related (A), flavor compound-related (B-D, and regulatory (D-G) genes. Normalized counts of gene expression are represented by violin plots. Genes with statistically significant (FDR<0.01) differential expression across stages are shown in bold. Wild tomato is shown in blue and cultivated in red. Stages of fruit development on the X-axis are numbered sequentially followed by “Br” for breaker stage and "RR" for red ripe stage Note that panels have independent Y-axis to maximize readability.

Figure 4.3 – GO Enrichment and Tobacco Clustered Expression Patterns



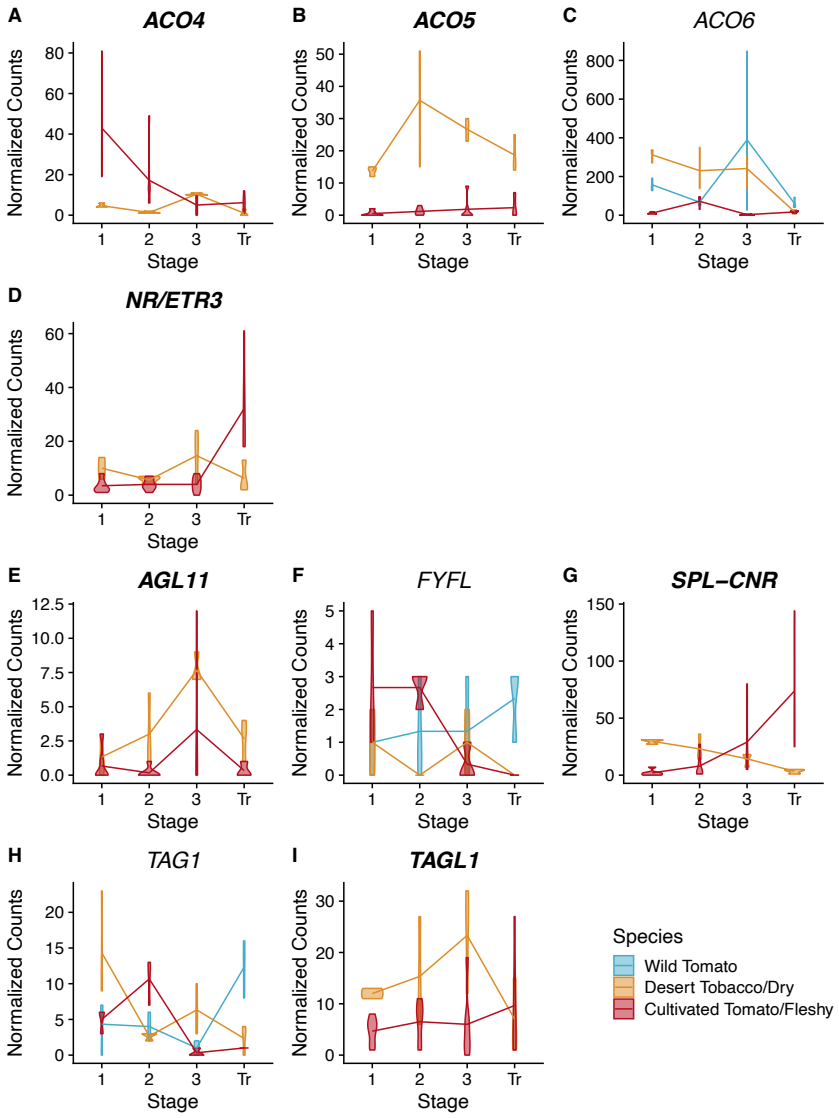
Summary of desert tobacco differentially expressed genes. A gene ontology (GO) term enrichment analysis (A) performed on all differentially expressed genes. All clusters of differentially expressed genes conserved among species are described with violin plots of normalized expression at each stage of development (B-G). Stages of fruit development in the axis of B-G are numbered sequentially followed by “Tr” for transition to mature stage.

Figure 4.4 – GO Enrichments and Clustered Solanaceae Gene Expression Patterns



Summary of differentially expressed orthologous genes. A gene ontology (GO) term enrichment analysis (A) performed on differentially expressed genes that had conserved patterns among the three species. A representative cluster of differentially expressed genes conserved among species is described with violin plots of normalized expression at each stage of development (B) along with a GO enrichment analysis (C) of the genes in that cluster. A gene ontology (GO) term enrichment analysis (D) performed on differentially expressed genes that had different patterns between fruit types. GO term descriptions to the left of the enrichment graphs are truncated for space and sorted by p-value. The bars are colored by the number of genes assigned to each GO term with legends in the lower right of the graph. Stages of fruit development in the axis of B-GD are numbered sequentially followed by “Tr” for transition to mature stage.

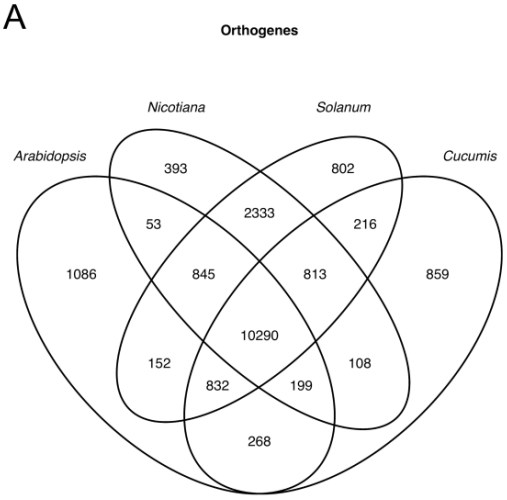
Figure 4.5 – Expression Patterns of Selected Solanaceae Genes



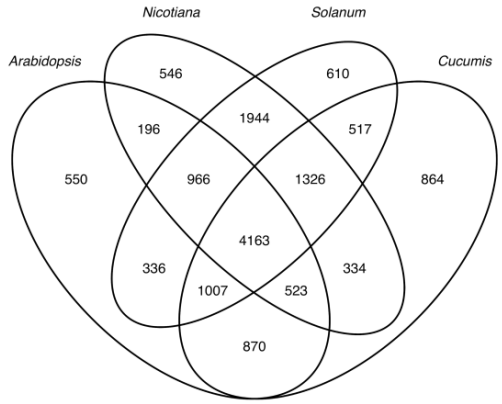
Expression profiles for ethylene-related (A-D) and regulatory (E-I) genes. Normalized counts of gene expression are represented by violin plots. Genes with statistically significant (FDR<0.01) differential expression across stages are shown in bold. Dry fruited desert tobacco values are shown in yellow. When expression pattern is better described by individual species trends (based on a log ratio test), wild tomato violin plots are shown in blue and cultivated tomato plots in red, otherwise both tomato species are shown together in red. Stages of fruit development on the X-axis are numbered sequentially followed by “Tr” for transition to maturity stage. Note that panels have independent Y-axis to maximize readability.

Figure 4.6 – Orthogene Numbers by Species

A



B

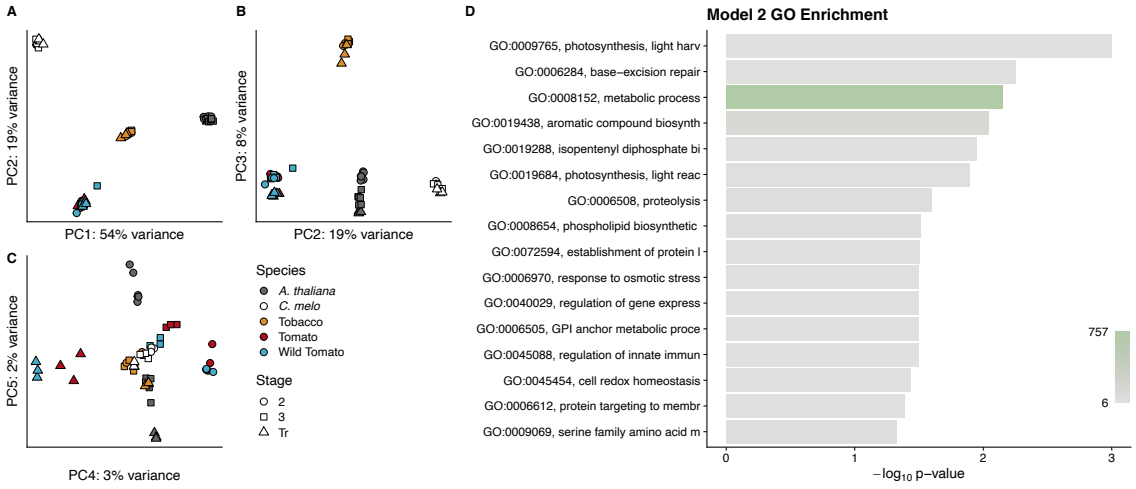


Venn diagram of orthologous genes (orthogenes) among the 4 species used in this study. All genes across the 4 species (A) and only single copy genes (B).

Figure 4.7 – Overview of Model 1 Genes

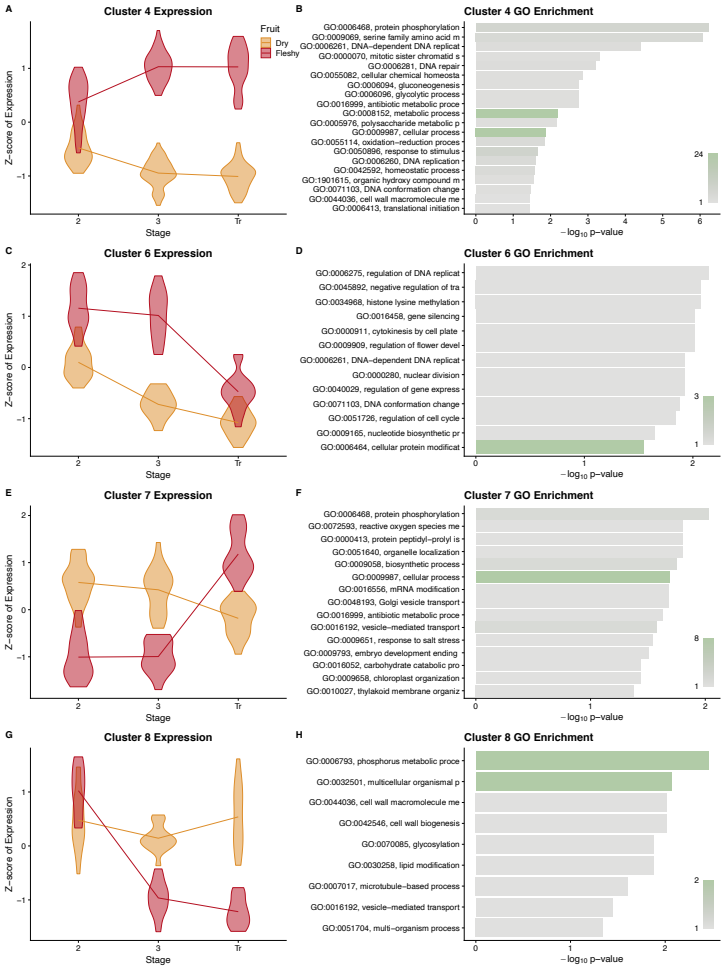
Summary of genes from Model 1. Principal components analysis (A-C) of gene expression values for each RNA-seq library. Points are colored by species and shaped by developmental stage as indicated in the legend. Principal components used for each graph are indicated on the axis along with the proportion of variance explained. A GO analysis (D) for the entire cohort of genes. GO term names to the left of the graph are truncated to available space. Terms are sorted by p-value, which is indicated by the bar height. Bars are colored by the number of genes annotated to that term, as indicated by the color scale in the lower right.

Figure 4.8 – Overview of Model 2 Genes



Summary of genes from Model 2. Principal components analysis (A-C) of gene expression values for each RNA-seq library. Points are colored by species and shaped by developmental stage as indicated in the legend. Principal components used for each graph are indicated on the axis along with the proportion of variance explained. A GO analysis (D) for the entire cohort of genes. GO term names to the left of the graph are truncated to available space. Terms are sorted by p-value, which is indicated by the bar height. Bars are colored by the number of genes annotated to that term, as indicated by the color scale in the lower right.

Figure 4.9 – GO Enrichments and Clustered Gene Expression Patterns for Model 2



Summary of differentially expressed orthologous genes. Representative clusters of differentially expressed genes with patterns that differ between dyr and fleshy fruited taxa are presented with violin plots of normalized expression at each stage of development (A,C,E,G) along with a GO enrichment analysis (B,D,F,H) of the genes in that cluster. GO term descriptions to the left of the enrichment graphs are truncated for space and sorted by p-value. The bars are colored by the number of genes assigned to each GO term with legends in the lower right of the graph. Stages of fruit development in the axis of A, C, E, and G are numbered sequentially followed by “Tr” for transition to mature stage.

Table 4.1 – Description of Developmental Stages

Description of Developmental Stages				
	Desert Tobacco	*A. thaliana*	Tomato	Melon
Stage 1	Ovary patterning (0 DAP)	Ovary patterning (3 DAP)	Ovary patterning (1 DAP)	Ovary patterning (—)[1]
Stage 2	Transverse anticlinal cell division (3 DAP)	Cell division and expansion (6 DAP)	Anticlinal and periclinal cell division (3 DAP)	Cell division (10 DAP)
Stage 3	Beginning basipetal lignification (6 DAP)	Beginning basipetal lignification (9 DAP)	Cell expansion and endoreduplication (15 DAP)	Cell expansion (20 DAP)
Transition	Color change from green to brown (11 DAP)	Color change to from green to yellow (12 DAP)	Initial color change from green to red, often termed 'breaker' stage (35 DAP)	Increase in sugar content, maximum firmness (30 DAP)
Stage 4	Senescence and dehiscence (—)[1]	Senescence and dehiscence (—)[1]	Cell wall softening, increase in sugar content, and full accumulation of pigments (45 DAP)	Fruit softening, mamximum sugar content (40 DAP)
Bioproject Accession	PRJNA646747 (This Study)	PRJEB25745 (Mizzotti et al, 2018)	PRJNA646747 (This Study)	PRJNA314069 (Chayut et al, 2017)

[1] Not sampled

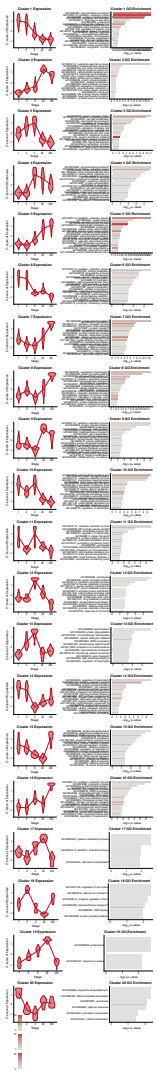
Pabón-Mora and Litt, 2011; Mizzotti et al, 2018; and Zhang et al, 2016

Table 4.2 – IDs and Names of Orthologous Genes

Orthology and Abbreviations			
Gene Name	*Solanum* Gene ID	*Nicotiana* Ortholog ID†	*Nicotiana* Abbreviation
ACO1	*Solyc07g049530.3.1*	—	—
ACO2	*Solyc12g005940.2.1*	—	—
ACO3	*Solyc07g049550.3.1*	—	—
ACO4	*Solyc02g081190.4.1*	*NIOBTv3_g13660.t1*	*NoACO4*
ACO5	*Solyc07g026650.3.1*	*NIOBTv3_g38689.t1*	*NoACO5*
ACO6	*Solyc02g036350.3.1*	*NIOBTv3_g02352.t1*	*NoACO6*
ACO7	*Solyc06g060070.3.1*	—	—
ACS2	*Solyc01g095080.3.1*	—	—
ACS4	*Solyc05g050010.3.1*	—	—
AGL11	*Solyc11g028020.3.1*	*NIOBTv3_g14436.t1*	*NoAGL11*
Cel2	*Solyc09g010210.3.1*	*NIOBTv3_g19880.t1*	*NoCel2*
Cel3	*Solyc07g005840.2.1*	*NIOBTv3_g12440.t1*	*NoCel3*
CHS-1	*Solyc09g091510.3.1*	—	—
CHS-2	*Solyc05g053550.3.1*	—	—
CTOMT1	*Solyc10g005060.4.1*	—	—
EIL1	*Solyc06g073720.3.1*	—	—
EIL2	*Solyc01g009170.4.1*	—	—
EIL4	*Solyc06g073730.2.1*	—	—
EJ2/MADS1	*Solyc03g114840.3.1*	—	—
EXP1	*Solyc06g051800.3.1*	*NIOBTv3_g17210.t1*	*NoEXP*
FUL1	*Solyc06g069430.3.1*	*NIOBTv3_g28929-D2.t1*	*NoFUL1*
FUL2	*Solyc03g114830.3.1*	*NIOBTv3_g39464.t1*	*NoFUL2*
FYFL	*Solyc03g006830.3.1*	*NIOBTv3_g10096.t1*	*NoFYFL*
GAD1	*Solyc03g098240.3.1*	*NIOBTv3_g11084.t1*	*NoGAD1*
GAD2	*Solyc11g011920.2.1*	—	—
GAD3	*Solyc01g005000.3.1*	—	—
J	*Solyc11g010570.2.1*	—	—
J2	*Solyc12g038510.2.1*	*NIOBTv3_g15806.t1*	*NoJ2*
MADS-RIN	*Solyc05g012020.4.1*	—	—
MBP10	*Solyc02g065730.2.1*	*NIOBTv3_g07845.t1*	*NoMBP10*
MBP20	*Solyc02g089210.4.1*	*NIOBT_gMBP20.t1*	*NoMBP20*
MBP3	*Solyc06g064840.4.1*	—	—
MC	*Solyc05g056620.2.1*	*NIOBTv3_g18077.t1*	*NoMC*
NAC-NOR	*Solyc10g006880.3.1*	*NIOBTv3_g08302.t1*	*NoNOR*
NR/ETR3	*Solyc09g075440.4.1*	*NIOBTv3_g10291.t1*	*NoETR3*
PGA2A	*Solyc10g080210.2.1*	—	—
PL1	*Solyc03g111690.4.1*	—	—
PSY1	*Solyc03g031860.3.1*	*NIOBTv3_g17569.t1*	*NoPSY1*
Solyc03g117740.3.1	*Solyc03g117740.3.1*	*NIOBTv3_g22270.t1*	*NIOBTv3_g22270.t1*
Solyc04g072038.1.1	*Solyc04g072038.1.1*	*NIOBTv3_g10008.t1*	*NIOBTv3_g10008.t1*
Solyc06g065310.3.1	*Solyc06g065310.3.1*	*NIOBTv3_g12238.t1*	*NIOBTv3_g12238.t1*
Solyc07g064300.3.1	*Solyc07g064300.3.1*	*NIOBTv3_g11662.t1*	*NIOBTv3_g11662.t1*
SPL-CNR	*Solyc02g077920.4.1*	*NIOBTv3_g27953.t1*	*NoSPL-CNR*
STM3	*Solyc01g092950.3.1*	—	—
TAG1	*Solyc02g071730.4.1*	*NIOBTv3_g22632-D2.t1*	*NoAG*
TAGL1	*Solyc07g055920.4.1*	*NIOBTv3_g13969.t1*	*NoSHP*
TM29	*Solyc02g089200.4.1*	*NIOBTv3_g14235.t1*	*NoSEP1*
TM3	*Solyc01g093965.2.1*	—	—
TM5	*Solyc05g015750.3.1*	—	—
TOMLOXC	*Solyc01g006540.4.1*	—	—

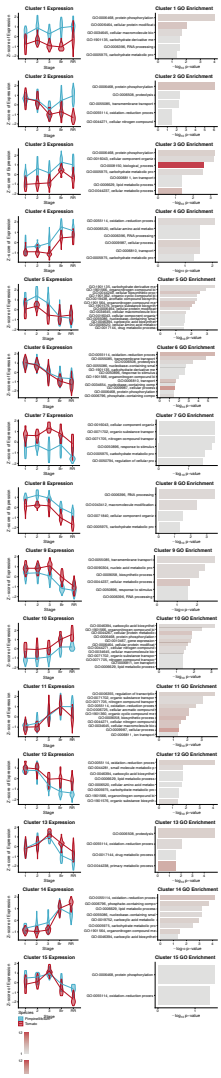
† Only one-to-one and many-to-one orthologs

Supplemental Figure 4.1 – GO Enrichment and Clustered Gene Expression Patterns for Tomato Conserved Clusters



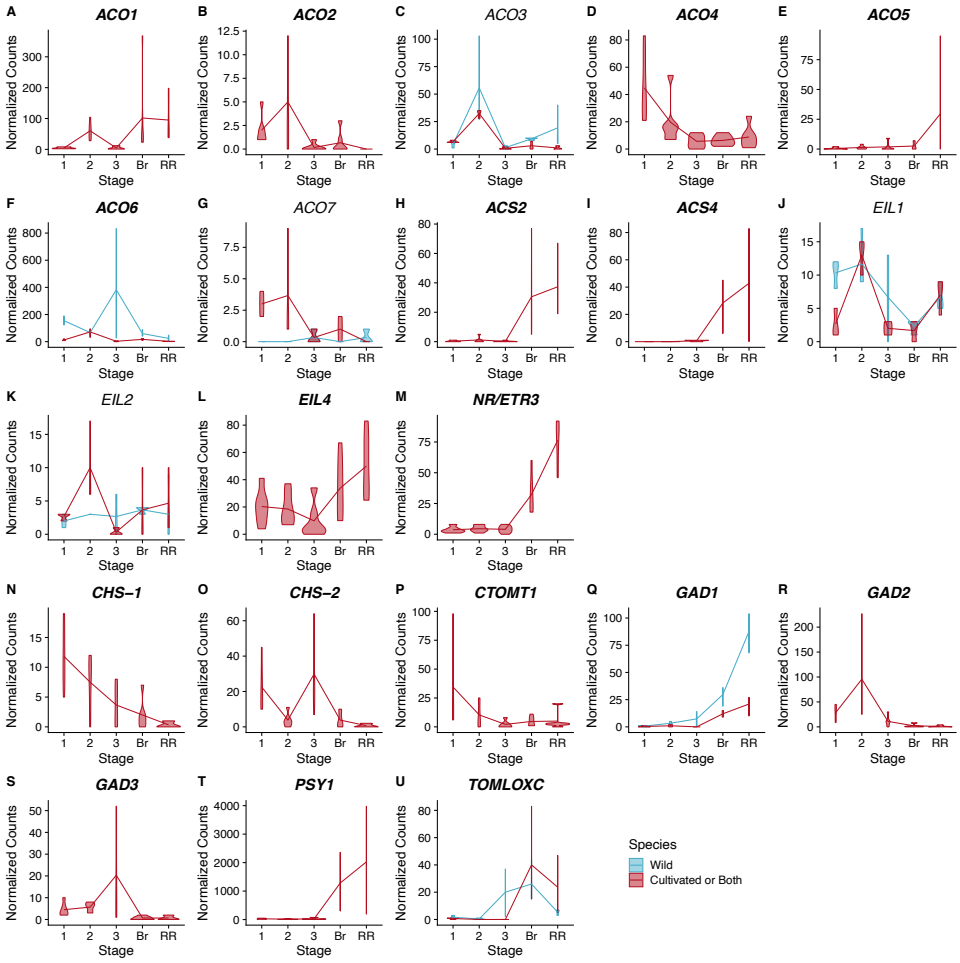
Summary of clustered gene expression profiles for genes with conserved patterns between wild and cultivated tomato. Violin plots of normalized expression by developmental stage for each cluster are shown on the left and gene ontology (GO) enrichment plots for the genes in the corresponding cluster are shown on the right. GO term descriptions to the left of the enrichment graphs are truncated for space and sorted by p-value. The bars are colored by the number of genes assigned to each GO term with legends in the lower right of each graph. Stages of fruit development in the axis of B and D are numbered sequentially followed by “Br” for breaker stage and “RR” for red ripe stage.

Supplemental Figure 4.2 – GO Enrichment and Clustered Gene Expression Patterns for Tomato Divergent Clusters



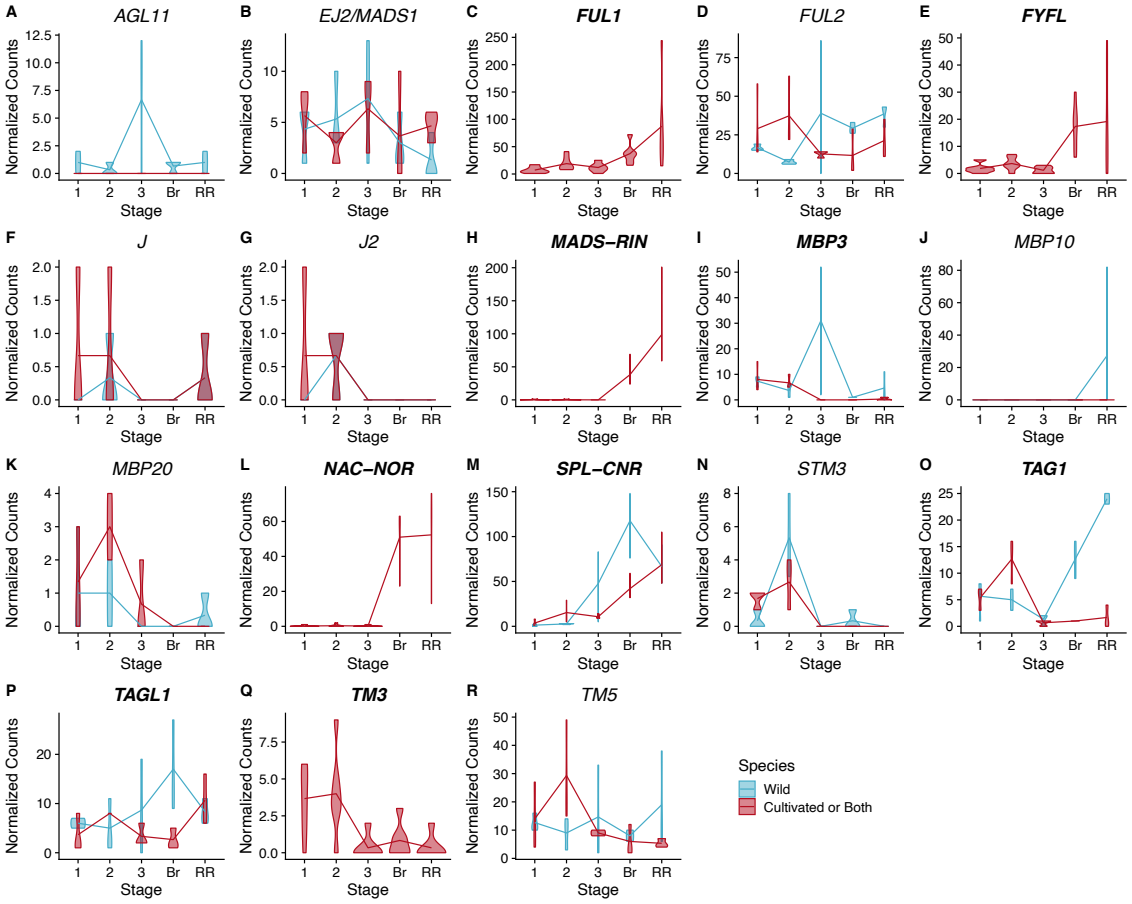
Summary of clustered gene expression profiles for genes with divergent patterns between wild and cultivated tomato. Violin plots of normalized expression by developmental stage for each cluster are shown on the left and gene ontology (GO) enrichment plots for the genes in the corresponding cluster are shown on the right. Profiles for wild tomato are shown in blue, while profiles for cultivated tomato are shown in red. GO term descriptions to the left of the enrichment graphs are truncated for space and sorted by p-value. The bars are colored by the number of genes assigned to each GO term with legends in the lower right of each graph. Stages of fruit development in the axis of B and D are numbered sequentially followed by “Br” for breaker stage and “RR” for red ripe stage.

Supplemental Figure 4.3 – Expression Patterns of Selected Tomato Structural Genes



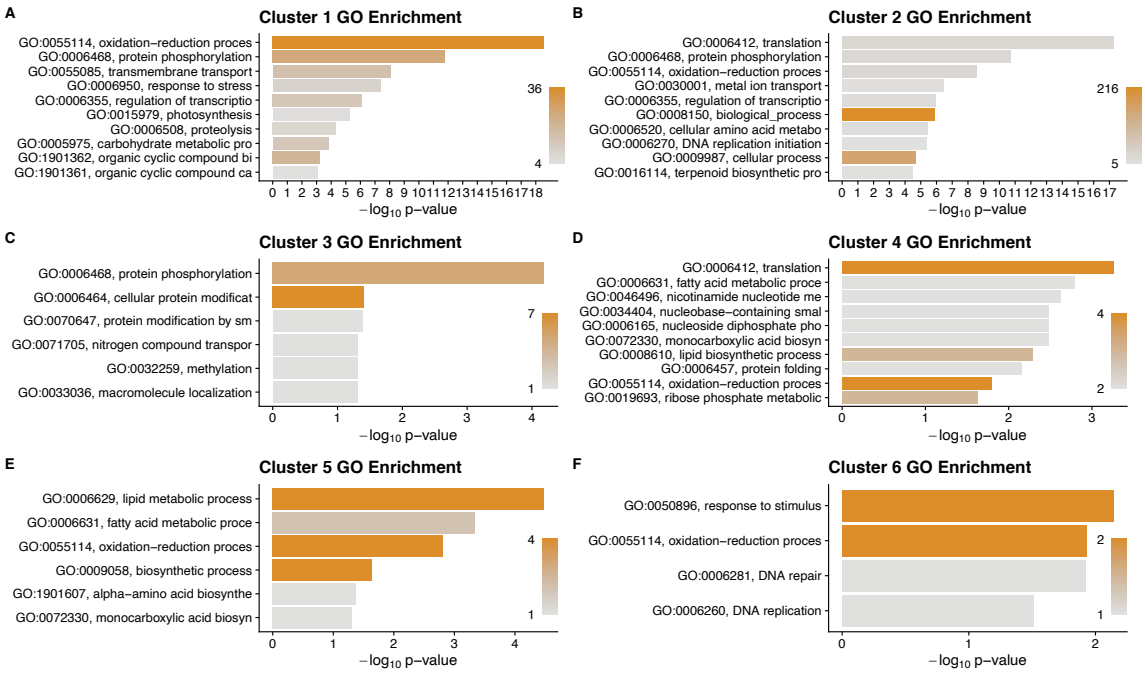
Summary of clustered gene expression profiles for genes with divergent patterns between wild and cultivated tomato. Violin plots of normalized expression by developmental stage for each cluster are shown on the left and gene ontology (GO) enrichment plots for the genes in the corresponding cluster are shown on the right. Profiles for wild tomato are shown in blue, while profiles for cultivated tomato are shown in red. GO term descriptions to the left of the enrichment graphs are truncated for space and sorted by p-value. The bars are colored by the number of genes assigned to each GO term with legends in the lower right of each graph. Stages of fruit development in the axis of B and D are numbered sequentially followed by “Br” for breaker stage and “RR” for red ripe stage.

Supplemental Figure 4.4 – Expression Patterns for Selected Tomato Regulatory Genes



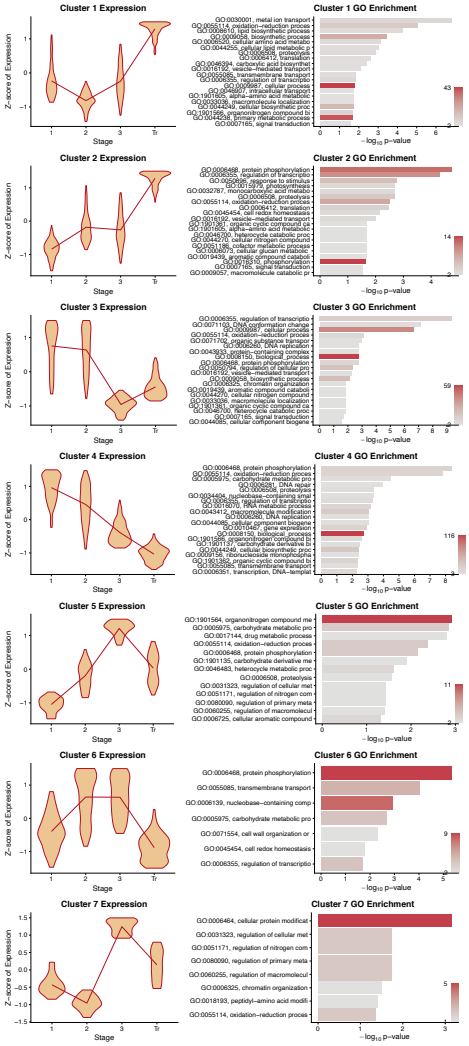
Expression profiles for selected regulatory genes. Normalized counts of gene expression are represented by violin plots. Genes with statistically significant (FDR<0.01) differential expression across stages are shown in bold. Where expression pattern is better described by individual species trends (based on a log ratio test), wild tomato violin plots are shown in blue and cultivated tomato plots are shown in red, otherwise the common pattern is shown in red. Stages of fruit development on the X-axis are numbered sequentially followed by “Br” for breaker stage and “RR” for red ripe stage. Note that panels have independent Y-axis to maximize readability.

Supplemental Figure 4.5 – GO Enrichments for Tobacco Genes Clusters



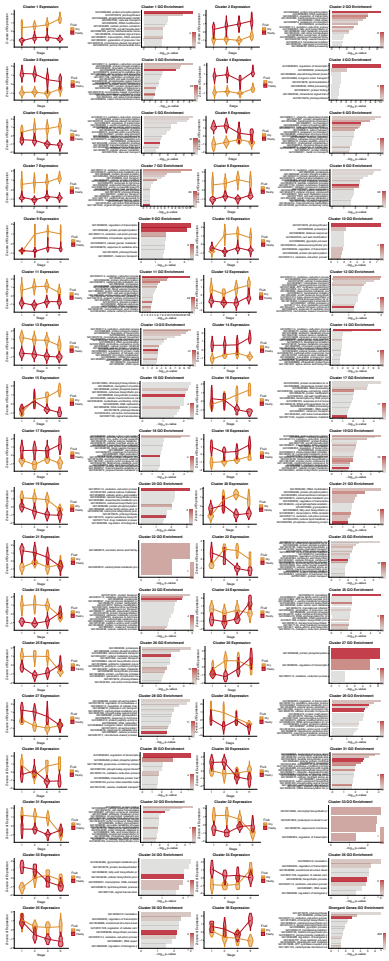
GO Enrichment analysis for desert tobacco gene expression clusters in Figure 4.3. GO term descriptions to the left of the enrichment graphs are truncated for space and sorted by p-value. The bars are colored by the number of genes assigned to each GO term with legends in the lower right of each graph. Stages of fruit development in the axis of B-GD are numbered sequentially followed by “Tr” for transition to mature stage.

Supplemental Figure 4.6 – GO Enrichment and Clustered Gene Expression Patterns for Solanaceae Conserved Clusters



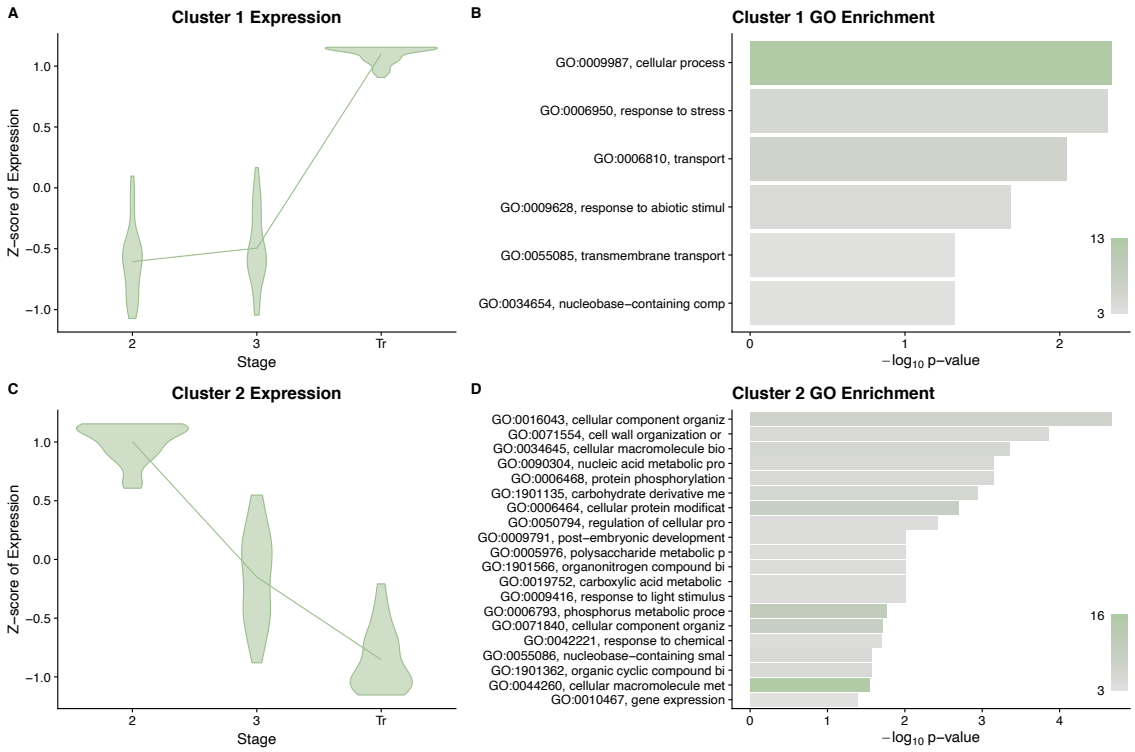
Summary of clustered gene expression profiles for genes with conserved patterns among the three solanaceous species. Violin plots of normalized expression by developmental stage for each cluster are shown on the left and gene ontology (GO) enrichment plots for the genes in the corresponding cluster are shown on the right. GO term descriptions to the left of the enrichment graphs are truncated for space and sorted by p-value. The bars are colored by the number of genes assigned to each GO term with legends in the lower right of each graph. Stages of fruit development in the axis are numbered sequentially followed by “Tr” transition to mature stage.

Supplemental Figure 4.7 – GO Enrichment and Clustered Gene Expression Patterns for Solanaceae Divergent Clusters



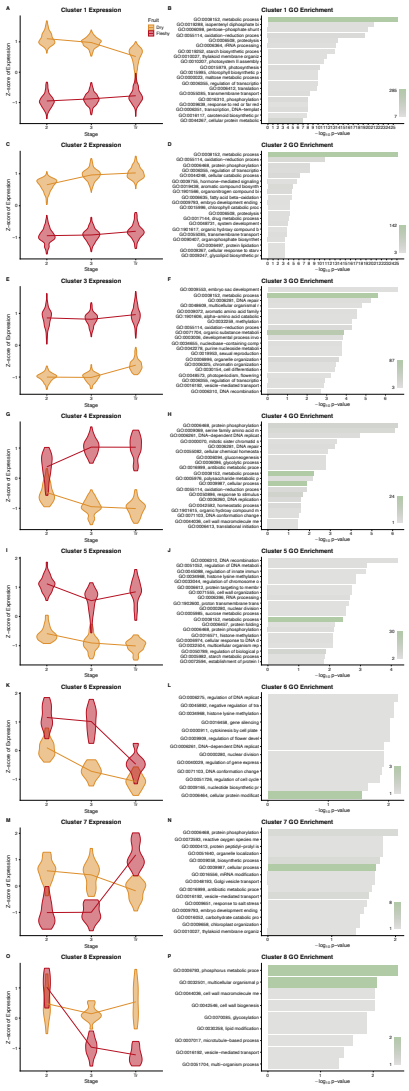
Summary of clustered gene expression profiles for genes with divergent patterns by fruit tpe among the three solanaceous species. Violin plots of normalized expression by developmental stage for each cluster are shown on the left and gene ontology (GO) enrichment plots for the genes in the corresponding cluster are shown on the right. Profiles for dry fruits are shown in yellow, while profiles for both tomato species are shown in red. GO term descriptions to the left of the enrichment graphs are truncated for space and sorted by p-value. The bars are colored by the number of genes assigned to each GO term with legends in the lower right of each graph. Stages of fruit development in the axis are numbered sequentially followed by “Tr” for transition to mature stage.

Supplemental Figure 4.8 – GO Enrichment and Clustered Gene Expression Patterns for Model 1 Clusters



Summary of clustered gene expression profiles for genes with conserved patterns among the five species. Violin plots of normalized expression by developmental stage for each cluster are shown on the left and gene ontology (GO) enrichment plots for the genes in the corresponding cluster are shown on the right. GO term descriptions to the left of the enrichment graphs are truncated for space and sorted by p-value. The bars are colored by the number of genes assigned to each GO term with legends in the lower right of each graph. Stages of fruit development in the axis are numbered sequentially followed by “Tr” transition to mature stage.

Supplemental Figure 4.9 – GO Enrichment and Clustered Gene Expression Patterns for Model 2 All Clusters



Summary of differentially expressed orthologous genes. Representative clusters of differentially expressed genes with patterns that differ between dry and fleshy fruited taxa are presented with violin plots of normalized expression at each stage of development along with a GO enrichment analysis of the genes in that cluster. GO term descriptions to the left of the enrichment graphs are truncated for space and sorted by p-value. The bars are colored by the number of genes assigned to each GO term with legends in the lower right of the graph. Stages of fruit development in the axis of A, C, E, and G are numbered sequentially followed by “Tr” for transition to mature stage.

Supplementary File 1

Link

List of gene names for conserved orthologous genes

Supplementary File 2

Link

List of gene names for divergent orthologous genes

References

Adams, P., Davies, J. N., & Winsor, G. W. (1978). Effects of Nitrogen, Potassium and Magnesium on the Quality and Chemical Composition of Tomatoes Grown in Peat. *The Journal of Horticultural Science*, *53*(2), 115–122.

Adenot, X., Elmayan, T., Lauressergues, D., Boutet, S., Bouché, N., Gasciolli, V., & Vaucheret, H. (2006). DRB4-dependent TAS3 trans-acting siRNAs control leaf morphology through AGO7. *Current Biology: CB*, *16*(9), 927–932.

Akihiro, T., Koike, S., Tani, R., Tominaga, T., Watanabe, S., Iijima, Y., Aoki, K., Shibata, D., Ashihara, H., Matsukura, C., Akama, K., Fujimura, T., & Ezura, H. (2008). Biochemical mechanism on GABA accumulation during fruit development in tomato. *Plant & Cell Physiology*, *49*(9), 1378–1389.

Alexa, A., & Rahnenfuhrer, J. (2016). topGO: Enrichment analysis for Gene Ontology. R package version 2.28. 0. *BioConductor. Published Online*.

Armbruster, W. S. (2014). Floral specialization and angiosperm diversity: phenotypic divergence, fitness trade-offs and realized pollination accuracy. *AoB Plants*, *6*(0). https://doi.org/10.1093/aobpla/plu003

Attwood, T. K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P. B., Popov, I., Romá-Mateo, C., Theodosiou, A., & Mitchell, A. L. (2012). The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. *Database: The Journal of Biological Databases and Curation*, *2012*, bas019.

Azpiroz, R., Wu, Y., LoCascio, J. C., & Feldmann, K. A. (1998). An Arabidopsis brassinosteroid-dependent mutant is blocked in cell elongation. *The Plant Cell*, *10*(2), 219–230.

Barrett, D. M., Beaulieu, J. C., & Shewfelt, R. (2010). Color, flavor, texture, and nutritional quality of fresh-cut fruits and vegetables: desirable levels, instrumental and sensory measurement, and the effects of processing. *Critical Reviews in Food Science and Nutrition*, *50*(5), 369–389.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The Arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. *Genesis* , *53*(8), 474–485.

Blanca, J., Montero-Pau, J., Sauvage, C., Bauchet, G., Illa, E., Díez, M. J., Francis, D., Causse, M., van der Knaap, E., & Cañizares, J. (2015). Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genomics*, *16*, 257.

Bouché, N., Lacombe, B., & Fromm, H. (2003). GABA signaling: a conserved and ubiquitous mechanism. *Trends in Cell Biology*, *13*(12), 607–610.

Bourdon, M., Frangne, N., Mathieu-Rivet, E., Nafati, M., Cheniclet, C., Renaudin, J.-P., & Chevalier, C. (2010). Endoreduplication and Growth of Fleshy Fruits. In U. Lüttge, W. Beyschlag, B. Büdel, & D. Francis (Eds.), *Progress in Botany 71* (pp. 101–132). Springer Berlin Heidelberg.

Bremer, B., & Eriksson, O. (1992). Evolution of fruit characters and dispersal modes in the tropical family Rubiaceae. *Biological Journal of the Linnean Society. Linnean Society of London*, *47*(1), 79–95.

Chayut, N., Yuan, H., Ohali, S., Meir, A., Sa’ar, U., Tzuri, G., Zheng, Y., Mazourek, M., Gepstein, S., Zhou, X., Portnoy, V., Lewinsohn, E., Schaffer, A. A., Katzir, N., Fei,

Z., Welsch, R., Li, L., Burger, J., & Tadmor, Y. (2017). Distinct Mechanisms of the ORANGE Protein in Controlling Carotenoid Flux. *Plant Physiology*, *173*(1), 376–389.

Chayut, N., Yuan, H., Ohali, S., Meir, A., Yeselson, Y., Portnoy, V., Zheng, Y., Fei, Z., Lewinsohn, E., Katzir, N., Schaffer, A. A., Gepstein, S., Burger, J., Li, L., & Tadmor, Y. (2015). A bulk segregant transcriptome analysis reveals metabolic and cellular processes associated with Orange allelic variation and fruit β-carotene accumulation in melon fruit. *BMC Plant Biology*, *15*(1), 274.

Chen, G., Hackett, R., Walker, D., Taylor, A., Lin, Z., & Grierson, D. (2004). Identification of a specific isoform of tomato lipoxygenase (TomloxC) involved in the generation of fatty acid-derived flavor compounds. *Plant Physiology*, *136*(1), 2641–2651.

Clausing, G., Meyer, K., & Renner, S. S. (2000). Correlations among fruit traits and evolution of different fruits within Melastomataceae. In *Botanical Journal of the Linnean Society* (Vol. 133, Issue 3, pp. 303–326). https://doi.org/10.1111/j.1095-8339.2000.tb01548.x

Cox, H. T. (1948). Studies in the Comparative Anatomy of the Ericales I. Ericaceae-Subfamily Rhododendroideae. In *American Midland Naturalist* (Vol. 39, Issue 1, p. 220). https://doi.org/10.2307/2421443

Crepet, W. L., & Niklas, K. J. (2009). Darwin's second "abominable mystery": Why are there so many angiosperm species? *American Journal of Botany*, *96*(1), 366–381.

De Smet, I., Vassileva, V., De Rybel, B., Levesque, M. P., Grunewald, W., Van Damme, D., Van Noorden, G., Naudts, M., Van Isterdael, G., De Clercq, R., Wang, J. Y., Meuli, N., Vanneste, S., Friml, J., Hilson, P., Jürgens, G., Ingram, G. C., Inzé, D., Benfey, P. N., & Beeckman, T. (2008). Receptor-like kinase ACR4 restricts formative cell divisions in the Arabidopsis root. *Science*, *322*(5901), 594–597.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , *29*(1), 15–21.

Doganlar, S., Frary, A., Ku, H.-M., & Tanksley, S. D. (2002). Mapping quantitative trait loci in inbred backcross lines of Lycopersicon pimpinellifolium (LA1589). In *Genome* (Vol. 45, Issue 6, pp. 1189–1202). https://doi.org/10.1139/g02-091

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238.

Eriksson, E. M., Bovy, A., Manning, K., Harrison, L., Andrews, J., De Silva, J., Tucker, G. A., & Seymour, G. B. (2004). Effect of the Colorless non-ripening mutation on cell wall biochemistry and gene expression during tomato fruit development and ripening. *Plant Physiology*, *136*(4), 4184–4197.

FAO. (2017). *Food Balance Sheet*. FAO Global Statistical Yearbook. http://www.fao.org/faostat/en/#data/FBS

Ferrándiz, C., Liljegren, S. J., & Yanofsky, M. F. (2000). Negative regulation of the *SHATTERPROOF* genes by FRUITFULL during *Arabidopsis* fruit development. *Science*, *289*(5478), 436–438.

Fischer, D. S., Theis, F. J., & Yosef, N. (2018). Impulse model-based differential

expression analysis of time course sequencing data. *Nucleic Acids Research*, *46*(20), e119.
Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., Thannhauser, T. W., Foolad, M. R., Diez, M. J., Blanca, J., Canizares, J., Xu, Y., van der Knaap, E., Huang, S., Klee, H. J., … Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, *51*(6), 1044–1051.
Garcia-Mas, J., Benjak, A., Sanseverino, W., Bourgeois, M., Mir, G., González, V. M., Hénaff, E., Câmara, F., Cozzuto, L., Lowy, E., Alioto, T., Capella-Gutiérrez, S., Blanca, J., Cañizares, J., Ziarsolo, P., Gonzalez-Ibeas, D., Rodríguez-Moreno, L., Droege, M., Du, L., … Puigdomènech, P. (2012). The genome of melon (Cucumis melo L.). *Proceedings of the National Academy of Sciences of the United States of America*, *109*(29), 11872–11877.
Gillaspy, G., Ben-David, H., & Gruissem, W. (1993). Fruits: A Developmental Perspective. *The Plant Cell*, *5*(10), 1439–1451.
Gimenez, E., Castañeda, L., Pineda, B., Pan, I. L., Moreno, V., Angosto, T., & Lozano, R. (2016). TOMATO AGAMOUS1 and ARLEQUIN/TOMATO AGAMOUS-LIKE1 MADS-box genes have redundant and divergent functions required for tomato reproductive development. *Plant Molecular Biology*, *91*(4-5), 513–531.
Giménez, E., Pineda, B., Capel, J., Antón, M. T., Atarés, A., Pérez-Martín, F., García-Sogo, B., Angosto, T., Moreno, V., & Lozano, R. (2010). Functional analysis of the Arlequin mutant corroborates the essential role of the Arlequin/TAGL1 gene during reproductive development of tomato. *PloS One*, *5*(12), e14427.
Givnish, T. J., Pires, J. C., Graham, S. W., McPherson, M. A., Prince, L. M., Patterson, T. B., Rai, H. S., Roalson, E. H., Evans, T. M., Hahn, W. J., Millam, K. C., Meerow, A. W., Molvray, M., Kores, P. J., O'Brien, H. E., Hall, J. C., Kress, W. J., & Sytsma, K. J. (2005). Repeated evolution of net venation and fleshy fruits among monocots in shaded habitats confirms a priori predictions: evidence from an ndhF phylogeny. *Proceedings. Biological Sciences / The Royal Society*, *272*(1571), 1481–1490.
Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, I. T., & White, O. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Research*, *29*(1), 41–43.
Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S., Fei, Z., Giovannoni, J. J., Mueller, L. A., & Saha, S. (2019). An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. In *bioRxiv* (p. 767764). https://doi.org/10.1101/767764
Houben, M., & Van de Poel, B. (2019). 1-Aminocyclopropane-1-Carboxylic Acid Oxidase (ACO): The Enzyme That Makes the Plant Hormone Ethylene. *Frontiers in Plant Science*, *10*, 695.
Huang, B., Routaboul, J.-M., Liu, M., Deng, W., Maza, E., Mila, I., Hu, G., Zouine, M., Frasse, P., Vrebalov, J. T., Giovannoni, J. J., Li, Z., van der Rest, B., & Bouzayen, M. (2017). Overexpression of the class D MADS-box gene Sl-AGL11 impacts fleshy tissue differentiation and structure in tomato fruits. *Journal of Experimental Botany*, *68*(17), 4869–4884.
Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software

version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.
Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.
Kelly, S., & Maini, P. K. (2013). DendroBLAST: approximate phylogenetic trees in the absence of multiple sequence alignments. *PloS One*, *8*(3), e58537.
Knapp, S. (2002). Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the Solanaceae. *Journal of Experimental Botany*, *53*(377), 2001–2022.
Krueger, F. (2012). *Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries*. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
Lai, T., Wang, X., Ye, B., Jin, M., Chen, W., Wang, Y., Zhou, Y., Blanks, A. M., Gu, M., Zhang, P., Zhang, X., Li, C., Wang, H., Liu, Y., Gallusci, P., Tör, M., & Hong, Y. (2020). Molecular and functional characterization of the SBP-box transcription factor SPL-CNR in tomato fruit ripening and cell death. *Journal of Experimental Botany*, *71*(10), 2995–3011.
Li, X., Tieman, D., Liu, Z., Chen, K., & Klee, H. J. (2020). Identification of a lipase gene with a role in tomato fruit short-chain fatty acid-derived flavor volatiles by genome-wide association. *The Plant Journal: For Cell and Molecular Biology*, *104*(3), 631–644.
Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.
Lukowitz, W., Mayer, U., & Jürgens, G. (1996). Cytokinesis in the Arabidopsis embryo involves the syntaxin-related KNOLLE gene product. *Cell*, *84*(1), 61–71.
Lu, X., Li, Y., Su, Y., Liang, Q., Meng, H., Li, S., Shen, S., Fan, Y., & Zhang, C. (2012). An Arabidopsis gene encoding a C2H2-domain protein with alternatively spliced transcripts is essential for endosperm development. *Journal of Experimental Botany*, *63*(16), 5935–5944.
MacGregor, K. B., Shelp, B. J., Peiris, S., & Bown, A. W. (2003). Overexpression of glutamate decarboxylase in transgenic tobacco plants deters feeding by phytophagous insect larvae. *Journal of Chemical Ecology*, *29*(9), 2177–2182.
Manning, K., Tör, M., Poole, M., Hong, Y., Thompson, A. J., King, G. J., Giovannoni, J. J., & Seymour, G. B. (2006). A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nature Genetics*, *38*(8), 948–952.
Mejía, N., Soto, B., Guerrero, M., Casanueva, X., Houel, C., Miccono, M. de L. Á., Ramos, R., Le Cunff, L., Boursiquot, J.-M., Hinrichsen, P., & Adam-Blondon, A.-F. (2011). Molecular, genetic and transcriptional evidence for a role of VvAGL11 in stenospermocarpic seedlessness in grapevine. *BMC Plant Biology*, *11*, 57.
Mizzotti, C., Rotasperti, L., Moretto, M., Tadini, L., Resentini, F., Galliani, B. M., Galbiati, M., Engelen, K., Pesaresi, P., & Masiero, S. (2018). Time-Course Transcriptome Analysis of Arabidopsis Siliques Discloses Genes Essential for Fruit Development and Maturation. *Plant Physiology*, *178*(3), 1249–1268.
Monforte, A. J., Diaz, A., Caño-Delgado, A., & van der Knaap, E. (2014). The genetic basis of fruit morphology in horticultural crops: lessons from tomato and melon. *Journal of Experimental Botany*, *65*(16), 4625–4637.

Montgomery, T. A., Howell, M. D., Cuperus, J. T., Li, D., Hansen, J. E., Alexander, A. L., Chapman, E. J., Fahlgren, N., Allen, E., & Carrington, J. C. (2008). Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. *Cell*, *133*(1), 128–141.

Nambeesan, S., Datsenka, T., Ferruzzi, M. G., Malladi, A., Mattoo, A. K., & Handa, A. K. (2010). Overexpression of yeast spermidine synthase impacts ripening, senescence and decay symptoms in tomato: Polyamines enhance shelf life in tomato. *The Plant Journal: For Cell and Molecular Biology*, *63*(5), 836–847.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274.

Ocarez, N., & Mejía, N. (2016). Suppression of the D-class MADS-box AGL11 gene triggers seedlessness in fleshy fruits. *Plant Cell Reports*, *35*(1), 239–254.

Pabón-Mora, N., & Litt, A. (2011). Comparative anatomical and developmental analysis of dry and fleshy fruits of Solanaceae. *American Journal of Botany*, *98*(9), 1415–1436.

Pan, I. L., McQuinn, R., Giovannoni, J. J., & Irish, V. F. (2010). Functional diversification of AGAMOUS lineage genes in regulating tomato flower and fruit development. *Journal of Experimental Botany*, *61*(6), 1795–1806.

Pantano, L. (2019). *DEGreport: Report of DEG analysis*. http://lpantano.github.io/DEGreport/

Pinyopich, A., Ditta, G. S., Savidge, B., Liljegren, S. J., Baumann, E., Wisman, E., & Yanofsky, M. F. (2003). Assessing the redundancy of MADS-box genes during carpel and ovule development. *Nature*, *424*(6944), 85–88.

Plunkett, G., Soltis, D., & Soltis, P. (1997). Clarification of the relationship beteen Apiaceae and Araliaceae based on matK and rbcL sequence data. *American Journal of Botany*, *84*(4), 565.

Pnueli, L., Hareven, D., Rounsley, S. D., Yanofsky, M. F., & Lifschitz, E. (1994). Isolation of the tomato AGAMOUS gene TAG1 and analysis of its homeotic role in transgenic plants. *The Plant Cell*, *6*(2), 163–173.

Razali, R., Bougouffa, S., Morton, M. J. L., Lightfoot, D. J., Alam, I., Essack, M., Arold, S. T., Kamau, A. A., Schmöckel, S. M., Pailles, Y., Shahid, M., Michell, C. T., Al-Babili, S., Ho, Y. S., Tester, M., Bajic, V. B., & Negrão, S. (2018). The Genome Sequence of the Wild Tomato Solanum pimpinellifolium Provides Insights Into Salinity Tolerance. *Frontiers in Plant Science*, *9*, 1402.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Regal, P. J. (1977). Ecology and evolution of flowering plant dominance. *Science*, *196*(4290), 622–629.

Ripoll, J.-J., Zhu, M., Brocke, S., Hon, C. T., Yanofsky, M. F., Boudaoud, A., & Roeder, A. H. K. (2019). Growth dynamics of the Arabidopsis fruit is mediated by cell expansion. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1914096116

Running, M. P., & Meyerowitz, E. M. (1996). Mutations in the PERIANTHIA gene of Arabidopsis specifically alter floral organ number and initiation pattern. *Development* , *122*(4), 1261–1269.

Sander, J., Schultze, J. L., & Yosef, N. (2017). ImpulseDE: detection of differentially

expressed genes in time series data using impulse models. *Bioinformatics* , *33*(5), 757–759.
Shen, J., Tieman, D., Jones, J. B., Taylor, M. G., Schmelz, E., Huffaker, A., Bies, D., Chen, K., & Klee, H. J. (2014). A 13-lipoxygenase, TomloxC, is essential for synthesis of C5 flavour volatiles in tomato. In *Journal of Experimental Botany* (Vol. 65, Issue 2, pp. 419–428). https://doi.org/10.1093/jxb/ert382
Sigrist, C. J. A., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L., & Xenarios, I. (2013). New and continuing developments at PROSITE. *Nucleic Acids Research*, *41*(Database issue), D344–D347.
Smykal, P., Gennen, J., De Bodt, S., Ranganath, V., & Melzer, S. (2007). Flowering of strict photoperiodic Nicotiana varieties in non-inductive conditions by transgenic approaches. *Plant Molecular Biology*, *65*(3), 233–242.
Spalik, K., Wojewódzka, A., & Downie, S. R. (2001). The evolution of fruit in Scandiceae subtribe Scandicinae (Apiaceae). *Canadian Journal of Botany. Journal Canadien de Botanique*, *79*(11), 1358–1374.
Tanksley, S. D. (2004). The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. *The Plant Cell*, *16 Suppl*, S181–S189.
Tanksley, S. D., Grandillo, S., Fulton, T. M., Zamir, D., Eshed, Y., Petiard, V., Lopez, J., & Beck-Bunn, T. (1996). Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative L. pimpinellifolium. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *92*(2), 213–224.
Thompson, A. J., Tor, M., Barry, C. S., Vrebalov, J., Orfila, C., Jarvis, M. C., Giovannoni, J. J., Grierson, D., & Seymour, G. B. (1999). Molecular and genetic characterization of a novel pleiotropic tomato-ripening mutant. *Plant Physiology*, *120*(2), 383–390.
Tieman, D., Bliss, P., McIntyre, L. M., Blandon-Ubeda, A., Bies, D., Odabasi, A. Z., Rodríguez, G. R., van der Knaap, E., Taylor, M. G., Goulet, C., Mageroy, M. H., Snyder, D. J., Colquhoun, T., Moskowitz, H., Clark, D. G., Sims, C., Bartoshuk, L., & Klee, H. J. (2012). The chemical interactions underlying tomato flavor preferences. *Current Biology: CB*, *22*(11), 1035–1039.
Watanabe, M., Tanaka, H., Watanabe, D., Machida, C., & Machida, Y. (2004). The ACR4 receptor-like kinase is required for surface formation of epidermis-related tissues in Arabidopsis thaliana. *The Plant Journal: For Cell and Molecular Biology*, *39*(3), 298–308.
Weber, A. (2004). Gesneriaceae. In *Flowering Plants · Dicotyledons* (pp. 63–158). https://doi.org/10.1007/978-3-642-18617-2_8
Xu, S., Brockmöller, T., Navarro-Quezada, A., Kuhl, H., Gase, K., Ling, Z., Zhou, W., Kreitzer, C., Stanke, M., Tang, H., Lyons, E., Pandey, P., Pandey, S. P., Timmermann, B., Gaquerel, E., & Baldwin, I. T. (2017). Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(23), 6133–6138.

Chapter 5: Conclusions

Datura stramonium Resources

I have created two resources, a genome assembly and a stable transformation protocol, for *Datura stramonium* that will enable more detailed studies of future questions related to genome evolution and functional genetics.

I demonstrated the first stable inheritance and expression of a transgene in the genus *Datura* using a *GREEN FLUORESCENT PROTEIN* (*GFP*) marker. This improves upon previously available methods for transient transformation and will allow the production of stable transgenic lines necessary for CRISPR/Cas9 experiments to study gene function. I also sequenced, assembled, and annotated the first draft genome assembly for this species. This assembly made use of Illumina short reads and Oxford Nanopore long reads. The assembly size agrees with previous estimates based on flow cytometry, and high BUSCO scores suggest that the assembly is representative of the genic portion of the genome. This draft assembly has many contigs (>200,000) and contains approximately 24% ambiguous bases (gaps). Because the assembly covers the genic portion of the genome well, it is likely that the gaps and fragmentation represent highly repetitive regions of the genome that cannot be sequenced by Illumina short reads and that could not be resolved with the current depth of long-read sequencing coverage (~7x). An additional consequence of this fragmentation is the high gene number in the annotation, currently over 52,000. Although this gene number is much higher than closely related species, I used a number of analyses to show that this number is unlikely to be due to polyploidy or bursts of recent gene duplications, but rather is an apparent overestimation. It is possible that the high apparent gene number is due to assembly gaps

fragmenting gene coding regions and splitting them into separate gene models. A similar issue was observed with the draft assembly of the eggplant (*Solanum melongena*) genome, which reported an initial gene number over 80,000 (Hirakawa et al., 2014). This number was later reduced to approximately 34,000 as genome contiguity was increased in subsequent assemblies (Barchi et al., 2019).

Using this genome assembly, I showed evidence for a series of gene duplications in the tropane alkaloid biosynthetic pathway. This pathway has many well-characterized steps, with a number of enzymes catalyzing reactions to shunt the pathway toward different products (Kohnen-Johannsen & Kayser, 2019). *Datura spp.* produce many tropane alkaloids, but the most prominent are hyoscyamine and scopolamine (Parr et al., 1990). The enzyme tropinone reductase I (TRI) is the first committed step in the synthesis of tropane alkaloids and the genome of *D. stramonium* possesses two copies of this gene in contrast to other closely related species like tomato (*Solanum lycopersicum*). The second duplication occurs much later in the pathway, in the gene encoding hyoscyamine 6 β-hydroxylase (H6H), which converts hyoscyamine to scopolamine, a much more potent and fast-acting hypnotic (Alizadeh et al., 2014).

With these resources, we can begin to formulate and answer new questions about gene and genome evolution within *D. stramonium* and at higher taxonomic levels. The duplications of the alkaloid biosynthetic genes deserve further study to determine their functional significance. Although I showed that the genome of *D. stramonium* encodes two copies of *TRI* and *H6H*, it is unknown if each copy encodes the same enzymatic functions. Such neofunctionalization following duplication has already been suggested for *TRI* and *TROPINONE REDUCTASE II* (*TRII*) following their divergence from a common ancestral protein and for several enzymes in the capsaicinoid biosynthesis pathway in pepper (Kim

et al., 2014; Nakajima et al., 1993). The possibility of neofunctionalization of *TRI* and *H6H* could be investigated by comparing amino acid sequences of paralogs with one another and with orthologs in other species that encode functional or nonfunctional versions of these enzymes. Metabolite profiling, enzyme kinetic studies, or substrate competition assays using purified versions of these duplicated enzymes would also determine if both copies are able to convert their respective substrates into products with equal efficiency.

Beyond their enzyme kinetics and substrate preference, the duplications could have resulted in different regulatory regions controlling spatial or temporal expression patterns for the paralogous genes. In the case of *H6H*, which is a tandem duplication separated by approximately 2 kilobase pairs, the divergence in expression and function may be less striking than for *TRI*, whose paralogs are separated by much larger distances (Haberer et al., 2004). The spatial and temporal expression patterns could be determined using *in situ* mRNA hybridization. Additionally, with the available tissue culture transformation protocol, it is now possible to use CRISPR to selectively knockout the expression of single paralogs and examine the impact each one has on alkaloid production independently. Combining tissue specific expression experiments, CRISPR-based mutagenesis, and metabolomic profiling of these alkaloids in various tissues of *D. stramonium* could help us to better engineer higher alkaloid content or change the profile of specific alkaloids that are present.

In a broader taxonomic context, the role of *H6H* is quite puzzling. My phylogenetic analysis of *H6H* among the various nightshades grouped the two paralogs in *D. stramonium* in a clade that is sister to two genes in *P. axillaris* (Fig. 3.3C). This *Datura* and *Petunia* clade was then sister to a clade containing sequences from *S. lycopersicum* and *C. annum*. Given the well established phylogeny of these species, one would expect

the gene tree to recapitulate the species trees, so that the *Datura*, *Solanum*, and *Capsicum* genes formed a clade that was sister to the *Petunia* genes, but it does not (Särkinen et al., 2013). The grouping of the *D. stramonium* genes separate from genes in other members of the Solanoideae could have several explanations. Potentially, orthologs of *H6H* do not exist in *S. lycopersicum* and *C. annum,* and these genes are simply the most similar genes to *DsH6H*. Given that H6H acts on a hypnotic alkaloid substrate to produce an even more potent hypnotic alkaloid product, the loss of *H6H* orthologs in these domesticated and edible species could be consistent with domestication-related selection against antinutritional compounds (Itkin et al., 2013). One further possibility is that the recovered *S. lycopersicum* and *C. annum* sequences are orthologs that have diverged in amino acid sequence since these species diverged from their common ancestor with *Datura spp.* This divergence could have been dramatic enough that they no longer group with the sequences from more closely related species. To examine this hypothesis further, denser taxonomic sampling of Solanoideae for the phylogenetic analysis could reveal a domestication-related pattern of mutations in *H6H* that is not present among non-domesticated species. Many non-domesticated relatives of both tomato and pepper exist, and by sampling these species, we could determine the extent to which this phylogenetic pattern is related to domestication versus other lineage-specific changes since the divergence of *Datura* from *Solanum* and *Capsicum.*

The most closely related genes to the *DsH6H* genes, two genes from petunia, show some evidence of a tandem duplication, albeit in a more complicated manner. The gene *Peaxi162Scf00075g01545* encodes a putative fusion protein of tandemly duplicated *H6H* homologues (Fig 3.3C). The C-terminal portion of this protein is most similar to another *H6H* homolog present on a different scaffold of the petunia genome assembly,

Peaxi162Scf00141g00025. One possibility to explain this complex relationship is a simple annotation error that incorrectly joins two adjacent *H6H* paralogs into a single coding sequence; however tandem gene duplications have been known to encode bona fide fusion proteins (Newman et al., 2015). Here again, denser taxonomic sampling may help to clarify the relationships among these genes by uncovering other species with similar arrangements of *H6H* genes; however PCR of this genomic region could confirm the tandem duplication, and RT-PCR assays could also be used to determine if the petunia genes encode fusion proteins as well.

Building off of the transgenic transformation study, I also used this genome assembly to characterize the impacts of tissue culture on mutation rate and gene expression in this species. Using mRNAseq and low-coverage genome resequencing of three T3 progeny from the initial GFP-transformant, I showed a dramatic impact on mutation rate of 1.16×10^{-3} mutations per site. Despite the elevated mutation rate, the impact of tissue culture on gene expression was negligible, with only 186 differentially expressed genes between leaves of transformed and untransformed plants. There are potentially other polymorphisms following tissue culture that could not be detected with the current methods. For instance, we did not detect large-scale genome rearrangements such as chromosomal translocations or inversions, but the low contiguity of the genome assembly and the short-read sequencing technology used to characterize the transformants prevents us from making conclusions about genome architecture at this scale. Relatedly, there may have been changes in transposon copy number in the genome following transformation, but mapping these changes is challenging given the high amount of ambiguous bases and gaps in the assembly, which likely correspond to transposon-rich, repetitive regions of the genome. There may have also been epigenetic changes

following transformation which could be detected with bisulfite sequencing, but we did not apply that sequencing technology here.

Multispecies Transcriptomes

I also undertook a comparative transcriptome study to look for conservation of gene expression patterns in the pericarps of five angiosperm species. Comparing gene expression between tomato with its wild relative (*S. pimpinellifolium*) revealed only a few subtle differences in expression patterns among the genes, but by examining specific genes implicated in fruit ripening, I found a number of potentially important divergences between the species that could be the result of domestication. These included changes in genes relating to fruit size, fruit firmness, and the deposition of woody tissue (lignin) in the pericarp. By incorporating gene expression data from the dry-fruited desert tobacco, I uncovered higher expression of several ethylene biosynthetic enzymes in this species than in the tomatoes. This result is puzzling given that tomato is a climacteric fruit that has been shown to undergo ethylene-dependent ripening, whereas desert tobacco is not. What is the role of ethylene in the maturity of dry fruits and to what extent is it conserved with fleshy fruits? This question could be addressed with functional characterizations of ethylene biosynthetic and response genes in desert tobacco and tomato fruits.

An additional interesting pattern was present in the expression data for the transcription factor *SPL-CNR*. This gene showed roughly opposite patterns between tomato and desert tobacco, with increasing expression over time in tomato and decreasing expression over time in desert tobacco. Functional studies of *SPL-CNR* in tomato showed that the protein decreases cell-cell adhesions and can promote cell death (Eriksson et al.,

2004; Lai et al., 2020). Does *SPL-CNR* function in the dehiscence zone of dry fruits with a conserved cell-cell adhesion function in both dry and fleshy fruits? The functional conservation (or not) of *SPL-CNR* orthologs between dry and fleshy fruits is especially intriguing as this gene's function in cell-cell adhesion could link the apparently disparate processes at maturity of fleshy fruit softening and dry fruit dehiscence. It would therefore be interesting to extend ectopic overexpression studies of *SPL-CNR* into a dry-fruited species such as tobacco and characterize pericarp development with a specific focus on the dehiscence zone.

It has previously been suggested that this gene's role in lowering cell-cell adhesion might allow it to function in both the softening of tomato fruits at maturity and separation of cells in the dehiscence zone of dry fruits at maturity (Eriksson et al., 2004). Increasing expression of *SPL-CNR* in tomato pericarps across development is consistent with this role (Eriksson et al., 2004). The role of *SPL-CNR* has not been well characterized in dry fruit development, but in our bulk RNA-seq from tissue of desert tobacco pericarps, we observed decreasing expression of *NoSPL-CNR*. This pattern of increasing expression in tobacco pericarps would seem to argue against a role for *SPL-CNR* in promoting dehiscence in dry fruits; however, if *NoSPL-CNR* has the hypothesized role in dry fruit dehiscence, we would expect the pattern of *NoSPL-CNR* expression to differ across cells of the pericarp with higher expression in cells of the dehiscence zone, which are separating at maturity, and lower expression in the rest of the cells of the pericarp outside of the dehiscence zone. Therefore, more detailed studies of *SPL-CNR* expression in different domains of the dry fruit pericarp would be necessary to follow up on this hypothesis.

Combining the pericarp expression data for all five species enabled the search for a set of single-copy, orthologous genes with conserved patterns of expression common to all species. I found a set of 121 such genes and termed them the “core” fruit development genes. This is in contrast to a much larger set of 1,795 single-copy, orthologous genes, the “accessory” fruit development genes, which have conserved patterns of expression within fruit types, but divergent patterns of expression between fruit types. Among the core genes, I found several genes that could function in cellular processes that are integral parts of fruit development including the gene *KNOLLE* (*AT1G08560*), which helps pattern the rate and plane of cell divisions (Lukowitz et al., 1996); and *ARABIDOPSIS CRINKLY4* (*ACR4*, *AT3G59420*), known to help pattern epidermal cells, root asymmetric cell divisions, and cuticle deposition (De Smet et al., 2008; Lukowitz et al., 1996; Running & Meyerowitz, 1996; Watanabe et al., 2004). The genes *DWARF4* (*DWF4*, *AT3G50660*) and *TITAN-LIKE* (*TTL, AT4G24900*) are involved in brassinosteroid biosynthesis and growth-responses, respectively, and were present among the core genes (Azpiroz et al., 1998; Lu et al., 2012; Montgomery et al., 2008). These two brassinosteroid-related genes have roles in cell-division and cell elongation, which are the prominent processes in Stages 1-3 of fruit development (Gillaspy et al., 1993). However several genes were also present in this data set for which a precise role in fruit development is less clear. For instance *ARGONAUTE7* (*AGO7*, *AT1G69440*) was present among the core genes and helps to regulate the developmental progression from vegetative to reproductive stages in an auxin-dependent manner, but also has roles in leaf development (Montgomery et al., 2008; Peng et al., 2017). Functions and classifications of the accessory genes are more diverse and this set could reflect differing subsets of genes that play a role in dry or fleshy fruit development but not both, or that have roles in

specific classes of fruits. For instance, these genes may have essential functions in the development of fleshy pepos such as melon, but not the fleshy berries of tomato.

Increased sampling of taxa could help both increase the number of one-to-one orthologs resolved in this data and also help refine the set of core fruit development genes further. Because the program we used for ortholog inference first groups protein sequences by similarity and then builds phylogenies within these groups, more thorough sampling of species could produce better resolved phylogenies and allow us to infer more one-to-one orthologs (Emms & Kelly, 2019). Many genes in the current datasets were not included in the five-species analysis as they were not one-to-one orthologs. Thus, increased taxonomic sampling could have resulted in more core genes that could play a conserved role in fruit development. There is also likely a point of diminishing returns with increased taxonomic sampling, as lineage-specific duplications or loss of genes would reduce the number of one-to-one orthologs. It would also be interesting to expand sampling to species with different types of fleshy and dry fruits. Although melon and tomato fruits are phenotypically different, they are both types of berries. What, for instance, would this set of core and accessory genes look like with wider sampling among Rosaceae? This family likewise contains a number of dry- and fleshy-fruited species, but there is a large diversity of fruit types and many of the fleshy fruited species in this family derive their fleshy parts from a hypanthium instead of an ovary proper (Xiang et al., 2017). Would this set of core genes remain relatively unchanged, with only the accessory set being affected or are the core/accessory gene sets strongly dependent on the evolutionary context?

Importantly, the fruit pericarp is a complex tissue, and the bulk RNA-seq techniques used here only allow for tissue-level insights (Roeder & Yanofsky, 2006). To what extent are certain cell populations in the dehiscence zone of a dry fruit similar to

pericarp cells of fleshy fruits? For instance, are the transcriptomes of fleshy fruit pericarp cells most similar to transcriptomes of cells in the dehiscence zone's separation layer, whose cell walls are also remodeled during dehiscence? Are the transcriptomes of fleshy fruit pericarp cells very distinct from the transcriptomes of all cell types in the dry fruit pericarp? Is the fleshy fruit pericarp subdivided into multiple domains analogous to the tissues of the dry fruit dehiscence zone, and, if so, do these domains share transcriptomic features with domains from the dry fruit pericarp? Combining rapidly maturing single-cell RNA-seq technology with the more recent advances in RNA velocity to transition/breaker-stage dry and fleshy fruits would be especially informative to map the fates of pericarp cells as they ripen. RNA velocity is a very recent advance in single-cell RNA sequencing technology that leverages levels of spliced and unspliced transcripts to predict the future transcriptomic state of a single cell (Bergen et al., 2020; La Manno et al., 2018). Using a similar strategy for comparing the expression patterns of orthologous genes among species that I applied here, this could uncover transcriptomic similarities within specific populations of pericarp cells between disparate fruit types. Because of the high costs of single-cell technologies, time course studies are often prohibitively expensive. However, sampling only transition/breaker stage pericarps and applying this method would allow us to capture the developmental trajectory of fleshy fruit pericarp cells as they ripen and compare this to dry fruit pericarp cells as they differentiate into the various tissues of the dehiscence zone. This and future analyses with emerging technologies will refine our knowledge of fruit development and hopefully uncover the evolutionary constraints, trajectories, and patterns enabling the success of angiosperms.

References

Alizadeh, A., Moshiri, M., Alizadeh, J., & Balali-Mood, M. (2014). Black henbane and its toxicity - a descriptive review. *Avicenna Journal of Phytomedicine*, *4*(5), 297–311.

Azpiroz, R., Wu, Y., LoCascio, J. C., & Feldmann, K. A. (1998). An Arabidopsis brassinosteroid-dependent mutant is blocked in cell elongation. *The Plant Cell*, *10*(2), 219–230.

Barchi, L., Pietrella, M., Venturini, L., Minio, A., Toppino, L., Acquadro, A., Andolfo, G., Aprea, G., Avanzato, C., Bassolino, L., Comino, C., Molin, A. D., Ferrarini, A., Maor, L. C., Portis, E., Reyes-Chin-Wo, S., Rinaldi, R., Sala, T., Scaglione, D., … Rotino, G. L. (2019). A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. *Scientific Reports*, *9*(1), 11769.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., & Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, *38*(12), 1408–1414.

Dehghan, E., Shahriari Ahmadi, F., Ghotbi Ravandi, E., Reed, D. W., Covello, P. S., & Bahrami, A. R. (2013). An atypical pattern of accumulation of scopolamine and other tropane alkaloids and expression of alkaloid pathway genes in Hyoscyamus senecionis. *Plant Physiology and Biochemistry: PPB / Societe Francaise de Physiologie Vegetale*, *70*, 188–194.

De Smet, I., Vassileva, V., De Rybel, B., Levesque, M. P., Grunewald, W., Van Damme, D., Van Noorden, G., Naudts, M., Van Isterdael, G., De Clercq, R., Wang, J. Y., Meuli, N., Vanneste, S., Friml, J., Hilson, P., Jürgens, G., Ingram, G. C., Inzé, D., Benfey, P. N., & Beeckman, T. (2008). Receptor-like kinase ACR4 restricts formative cell divisions in the Arabidopsis root. *Science*, *322*(5901), 594–597.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238.

Eriksson, E. M., Bovy, A., Manning, K., Harrison, L., Andrews, J., De Silva, J., Tucker, G. A., & Seymour, G. B. (2004). Effect of the Colorless non-ripening mutation on cell wall biochemistry and gene expression during tomato fruit development and ripening. *Plant Physiology*, *136*(4), 4184–4197.

Gillaspy, G., Ben-David, H., & Gruissem, W. (1993). Fruits: A Developmental Perspective. *The Plant Cell*, *5*(10), 1439–1451.

Haberer, G., Hindemitt, T., Meyers, B. C., & Mayer, K. F. X. (2004). Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of Arabidopsis. *Plant Physiology*, *136*(2), 3009–3022.

Hirakawa, H., Shirasawa, K., Miyatake, K., Nunome, T., Negoro, S., Ohyama, A., Yamaguchi, H., Sato, S., Isobe, S., Tabata, S., & Others. (2014). Draft genome sequence of eggplant (Solanum melongena L.): the representative solanum species indigenous to the old world. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *21*(6), 649–660.

Itkin, M., Heinig, U., Tzfadia, O., Bhide, A. J., Shinde, B., Cardenas, P. D., Bocobza, S. E., Unger, T., Malitsky, S., Finkers, R., Tikunov, Y., Bovy, A., Chikate, Y., Singh, P., Rogachev, I., Beekwilder, J., Giri, A. P., & Aharoni, A. (2013). Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science*, *341*(6142), 175–179.

Kim, S., Park, M., Yeom, S.-I., Kim, Y.-M., Lee, J. M., Lee, H.-A., Seo, E., Choi, J., Cheong, K., Kim, K.-T., Jung, K., Lee, G.-W., Oh, S.-K., Bae, C., Kim, S.-B., Lee, H.-Y., Kim, S.-Y., Kim, M.-S., Kang, B.-C., … Choi, D. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature Genetics*, *46*(3), 270–278.

Kohnen-Johannsen, K. L., & Kayser, O. (2019). Tropane Alkaloids: Chemistry, Pharmacology, Biosynthesis and Production. *Molecules* , *24*(4). https://doi.org/10.3390/molecules24040796

Lai, T., Wang, X., Ye, B., Jin, M., Chen, W., Wang, Y., Zhou, Y., Blanks, A. M., Gu, M., Zhang, P., Zhang, X., Li, C., Wang, H., Liu, Y., Gallusci, P., Tör, M., & Hong, Y. (2020). Molecular and functional characterization of the SBP-box transcription factor SPL-CNR in tomato fruit ripening and cell death. *Journal of Experimental Botany*, *71*(10), 2995–3011.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., … Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, *560*(7719), 494–498.

Lukowitz, W., Mayer, U., & Jürgens, G. (1996). Cytokinesis in the Arabidopsis embryo involves the syntaxin-related KNOLLE gene product. *Cell*, *84*(1), 61–71.

Lu, X., Li, Y., Su, Y., Liang, Q., Meng, H., Li, S., Shen, S., Fan, Y., & Zhang, C. (2012). An Arabidopsis gene encoding a C2H2-domain protein with alternatively spliced transcripts is essential for endosperm development. *Journal of Experimental Botany*, *63*(16), 5935–5944.

Montgomery, T. A., Howell, M. D., Cuperus, J. T., Li, D., Hansen, J. E., Alexander, A. L., Chapman, E. J., Fahlgren, N., Allen, E., & Carrington, J. C. (2008). Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. *Cell*, *133*(1), 128–141.

Nakajima, K., Hashimoto, T., & Yamada, Y. (1993). Two tropinone reductases with different stereospecificities are short-chain dehydrogenases evolved from a common ancestor. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(20), 9591–9595.

Newman, S., Hermetz, K. E., Weckselblatt, B., & Rudd, M. K. (2015). Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *American Journal of Human Genetics*, *96*(2), 208–220.

Parr, A. J., Payne, J., Eagles, J., Chapman, B. T., Robins, R. J., & Rhodes, M. J. C. (1990). Variation in tropane alkaloid accumulation within the solanaceae and strategies for its exploitation. *Phytochemistry*, *29*(8), 2545–2550.

Peng, J., Berbel, A., Madueño, F., & Chen, R. (2017). AUXIN RESPONSE FACTOR3 Regulates Compound Leaf Patterning by Directly Repressing PALMATE-LIKE PENTAFOLIATA1 Expression in Medicago truncatula. *Frontiers in Plant Science*, *8*, 1630.

Roeder, A. H. K., & Yanofsky, M. F. (2006). Fruit Development in Arabidopsis. In *The Arabidopsis Book* (Vol. 4, p. e0075). https://doi.org/10.1199/tab.0075

Running, M. P., & Meyerowitz, E. M. (1996). Mutations in the PERIANTHIA gene of Arabidopsis specifically alter floral organ number and initiation pattern. *Development* , *122*(4), 1261–1269.

Särkinen, T., Bohs, L., Olmstead, R. G., & Knapp, S. (2013). A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. *BMC Evolutionary Biology*, *13*, 214.

Watanabe, M., Tanaka, H., Watanabe, D., Machida, C., & Machida, Y. (2004). The ACR4 receptor-like kinase is required for surface formation of epidermis-related tissues in Arabidopsis thaliana. *The Plant Journal: For Cell and Molecular Biology*, *39*(3), 298–308.

Xiang, Y., Huang, C.-H., Hu, Y., Wen, J., Li, S., Yi, T., Chen, H., Xiang, J., & Ma, H. (2017). Evolution of Rosaceae Fruit Types Based on Nuclear Phylogeny in the Context of Geological Times and Genome Duplication. *Molecular Biology and Evolution*, *34*(2), 262–281.

Appendix 1: Virus-Induced Gene Silencing

Introduction

Previously, a pilot study was undertaken to look for candidate genes that function differently between dry and fleshy fruits. Litt and colleagues (unpublished data) sequenced the early fruit transcriptomes of cultivated tomato (*Solanum lycopersicum* cv. Micro-Tom), which has a fleshy berry, and woodland tobacco (*Nicotiana sylvestris*), which has a dry, capsular fruit. Analysis of these transcriptomes showed that an ortholog of the *Arabidopsis thaliana* transcription factor *FRUITFULL* (*FUL*) was more highly expressed in woodland tobacco than in tomato. *FUL* encodes a MADS-box transcription factor, which is known to pattern cell division, cell expansion, and lignification in Arabidopsis siliques (a dry dehiscent fruit) (Ferrándiz et al., 2000; Gu et al., 1998). In the Arabidopsis loss-of-function mutant, *ful-1*, cell division and differentiation defects in the pericarp lead to siliques that are shortened compared to wild type and fail to dehisce (Gu et al., 1998). A later study showed that *FUL* helps pattern lignification by repressing *SHATTERPROOF1/2*, and overexpression of *FUL* led to a loss of *SHP1/2* transcripts in the siliques and a total lack of a dehiscence zone at maturity (Ferrándiz et al., 2000). Similar overexpression phenotypes were seen in woodland tobacco fruits, suggesting a similar function in regulating patterning of this dehiscence zone (Smykal et al., 2007).

Further work by Pabón-Mora and Litt has shown that some differences between dry and fleshy nightshade fruits correspond to the same processes thought to be regulated by *FUL* (cell division, cell expansion, and lignification) (Pabon-Mora & Litt, 2011). Interestingly, despite the lack of dehiscence zones in fleshy fruits, *FUL* homologs are also expressed in tomato (Bemer et al., 2012; Fujisawa et al., 2014; Hileman et al., 2006; S.

Wang et al., 2014), suggesting a change in *FUL* gene function between these two fruit types, and making *FUL* and its orthologs candidates to play a role in fruit type transition.

FUL and its paralog, *AGL79*, which has diverged in expression pattern from both *FUL* and the other genes in its clade, arose from a duplication event in the core eudicots (Litt & Irish, 2003; Pařenicová et al., 2003). This duplication produced the *euFULI* clade, which contains *FUL*, and the *euFULII* clade, which contains *AGL79*. In Solanaceae there have been subsequent duplications, giving rise to a total of four genes, *FUL1* and *FUL2* in the *euFULI* clade, and *MBP10* and *MBP20* in the *euFULII* clade. The role of the *euFULII* genes has not been well investigated, but three studies suggest they might play a role in lateral root development, leaf morphology, and inflorescence branching (Berbel et al., 2012; Burko et al., 2013; Gao et al., 2018). In contrast, many studies have shown that the *euFULI* clade genes are widely involved in fruit development (Bemer et al., 2012; Dardick et al., 2010; Fujisawa et al., 2014; Gu et al., 1998; Jaakola et al., 2010; Østergaard et al., 2006; Shima et al., 2014; Smykal et al., 2007; Tani et al., 2007; S. Wang et al., 2014). Several groups have investigated the role of *euFULI* genes in fleshy fruit development, most notably, a series of three studies in tomato, which investigated the fruit phenotypes of *euFULI* knockdowns (Bemer et al., 2012; Shima et al., 2014; S. Wang et al., 2014). All of these studies showed defects in proper pigmentation of the ripe fruits, but they produced contradictory results regarding pericarp thickness, cuticle thickness, and ethylene response, among other traits. These contradictory results could be due to the use of different tomato cultivars, only partial silencing of the target genes, and potential off-target silencing of *euFULII* genes. Later, CRISPR knockouts for the tomato *euFULI* genes demonstrated that both *FUL1* and *FUL2* have overlapping roles in determining pericarp color and partially overlapping roles in ethylene production; however *FUL2* shows

additional roles in timing ripening and in pericarp morphology (R. Wang et al., 2019). Importantly, none of these studies have addressed the role of the *euFULII* genes, *SlMBP10* and *SlMBP20*. Thus, the role of *FUL* genes in tomato development remains unclear.

The role of these genes in the development of dry-fruited Solanaceae is almost totally unknown, but determining this role will help our understanding of fleshy fruit evolution. Given the sequence and expression conservation among *FUL* genes, especially within the same gene clade, they likely have at least partially redundant functions, which have yet to be clearly determined. For this study, we used desert tobacco (*Nicotiana obtusifolia*), a compact, profusely flowering, diploid congener of cultivated tobacco (*Nicotiana tabacum*) with a sequenced genome, as the representative dry-fruited species, in order to explore the function of *FUL* genes. Although stable transformation of desert tobacco is possible, the efficiency has proven quite low and unsuitable for generating higher-order CRISPR mutants. For this reason, we created a series of virus-induced gene silencing constructs targeting *euFULI*, *euFULII*, or all genes in the clade for desert tobacco and tracked several reproductive phenotypes.

Results

We generated VIGS constructs for the three groups of target genes (described below) and infiltrated a total of 135 plants for this experiment. These plants were divided across three experimental treatments and two developmental time points (Table 6.1). We used three experimental silencing constructs: euFULI targeting both *NoFUL1* and *NoFUL2*, euFULII targeting *NoMBP10* and *NoMBP20*, and All targeting *NoFUL1*, *NoFUL2*, *NoMBP10*, and *NoMBP20*. A BLAST search of the locus used to create the All silencing

construct suggested that this construct might also target the desert tobacco ortholog of *APETALA1*. In addition to these experimental treatments, wild-type plants that were not infiltrated at all were used as controls for normal development, and we included an empty vector control as well to assess the effect of viral replication and infiltration on the phenotypes. These plants were infiltrated with a version of TRV2 that did not contain a desert tobacco gene sequence for silencing.

For the *euFULI* and *euFULII* VIGS constructs, we infiltrated 15 plants for each construct at the rosette stage and 20 plants at the cauline leaf stage. This number was reduced to 15 plants in each stage for the All construct. For both the uninfiltrated plants and those infiltrated with the empty vector, we infiltrated 10 plants at the rosette stage and 8 at the cauline leaf stage; however, one of the uninfiltrated plants at the rosette leaf stage subsequently died (Table 6.1).

No Effect on Developmental Milestones

Our previous pilot study had shown some evidence of developmental delays when silencing multiple *FUL* genes in desert tobacco, and *FUL* genes have a reported role in the vegetative to reproductive transition (Balanzà et al., 2014, 2018; Pabón-Mora et al., 2012). To determine the relative contributions of *euFULI* and *euFULII* genes family members to developmental timing, we scored the plants in this study for their progress toward several reproductive milestones including bolting, visible flower buds, open flowers, and fruit production. Because these milestones are a proxy for plant developmental age, we analyzed our plants not only by the silencing construct used, but also by their stage at the onset of the experiment.

We scored plant stage at five timepoints and saw that plants generally progressed through development at similar speeds (Fig 6.1). To obtain a more rigorous analysis of this progression, we used an ANOVA to look for evidence that the silencing construct used might delay plant development. For days to bolting, days to first flower, and days to first fruit, we did not see a statistically significant difference ($p>>0.05$) between any of the silencing constructs for plants infiltrated either at the rosette leaf or cauline leaf stages (Fig 6.2).

No Effect on Plant Architecture

In several eudicot species, eu*FUL* genes have been demonstrated to affect the number of inflorescence branches (Berbel et al., 2012; Gao et al., 2018; Pabón-Mora et al., 2012, 2013). We also assessed the relative inflorescence branching among our silenced desert tobacco plants. We scored the number of actively growing inflorescence meristem branching at 27 days post infiltration.

Because developmental age can affect the degree of branching, we analyzed plants infiltrated at the cauline and rosette leaf stages separately. Using a negative binomial regression to account for this count data and a chi-squared statistic to test significance, we asked if any of the silencing constructs altered branch number (Fig 6.3). Neither the plants infiltrated on rosette leaves ($p=0.618$, Fig 6.3A) nor those infiltrated on cauline leaves ($p=0.555$, Fig 6.3B) showed a statistically significant change in branch number across silencing constructs.

Discussion and Conclusion

Despite preliminary evidence for the roles of *euFUL* genes in reproductive development and in plant architecture, our study was unable to find statistically significant evidence for an effect from our silencing constructs. Importantly, the current study was limited in scope and statistically underpowered, so caution should be taken in interpreting these results.

Our previous VIGS study, using the same target sequence as the All construct, showed that phenotypes were most dramatic when the plants were infiltrated at the four-leaf stage (unpublished data). This result aligns well with recommended protocols for other species (Coenen et al., 2018; Schultink et al., 2019; Senthil-Kumar & Mysore, 2014; Taheri-Dehkordi et al., 2018). Here, plants were infiltrated much later, roughly corresponding to the 10-leaf stage for rosette infiltration and the 16-leaf stage for cauline leaf infiltration. Because silencing often takes 2-3 weeks to occur, many of the experimental plants advanced to the final fruiting stage within 2 weeks of infiltration, and no changes in developmental progression would have been observed.

Because VIGS relies on post-transcriptional gene silencing to knockdown transcript levels, the efficiency of silencing can vary across replicates and across target genes. We were unable to confirm the presence or magnitude of silencing for the experimental plants, however in the future, this data could allow us to subset plants showing knockdown of the target gene transcript levels and analyze them separately from the non-silenced replicates.

The previous study, using the same target sequence as the All construct, showed two common phenotypes, developmental delay and non-senescent sepals on mature fruits. We were unable to track development long enough to ascertain changes in sepal

senescence at fruit maturity, but we also observed no delays in development. The lack of developmental delay could be due to a number of factors beyond the aforementioned age of the plants at infiltration. The growth environment for plants in the two studies was similar but not identical, specifically desert tobacco plants from the previous study were grown under short days (12h light/12h dark), whereas the current study used long days (16h light/8h dark) because of constraints beyond the scope of this experiment. Idiosyncrasies of phenotyping techniques could also vary across the two experiments. Importantly, the lack of phenotypic differences among the treatments is likely due to the lack of statistical power to detect subtle or moderate differences. For these developmental milestones, a power analysis, assuming a balanced replicate number of 13 across the five treatments, showed that a large phenotypic effect (f=0.4) could be detected with a probability of 0.696, while a small phenotypic effect (f=0.1) could only be detected with a probability of 0.082 (Cohen, 1988). To detect a large phenotypic effect with a reasonable power of 0.8, we would have needed more than 16 replicates per treatment, and more than 240 replicates per treatment to detect a small phenotypic effect. For the branching trait, our design was sufficiently powered to detect large (power=1) and medium (power=0.81) but not small effects (power=0.128). However, nearly 1,200 total plants would be needed to detect a small effect on branching with sufficient power (0.8).

The lack of expected phenotypic effects of silencing along with the accommodations made in the experimental design suggest that the results of this study are unlikely to be representative of the actual effects of *euFUL* gene silencing in desert tobacco. Future studies of *euFUL* gene function would benefit from larger replicate numbers, earlier infiltration of plants, and a longer observation period to detect later reproductive phenotypes. Additionally, quantitative RT-PCR confirmation of silencing for

each of the four *euFUL* genes in desert tobacco along with the potential off-target *NoAPETALA1* would provide more concrete evidence of the relative contributions of these genes to the potential phenotypes.

Methods

Target Selection and Cloning

Because the desert tobacco genome is not functionally annotated, we conducted a BLAST search using the *SlFUL1* (*Solyc06g069430*), *SlFUL2* (*Solyc03g114830*), *SlMBP10* (*Solyc02g065730*), and *SlMBP20* (*Solyc02g089210*) coding sequencing against the gene models from the desert tobacco genome (Altschul et al., 1997; Xu et al., 2017). This yielded three loci, *NoFUL1* (*NIOBTv3_g28929-D2*), *NoFUL2* (*NIOBTv3_g39464*), and *NoMBP10* (*NIOBTv3_g07845*), but no locus corresponded to *NoMBP20*. We therefore annotated this locus *de novo* in the desert tobacco genome using previously generated transcriptome data aligned to the genome. The *NoMBP20* sequence from this data aligned across a 5,338bp region of scaffold 1501 and comprised eight exons. We proceeded with this locus as the putative *NoMBP20* sequence.

Because MADS-box genes often share significant homology across the 5' MADS domain, we were unable to use this region to design specific VIGS silencing constructs (Litt & Irish, 2003). Instead, we focused on the more divergent 3' exons and 3' UTR to ensure specificity in construct design. To determine the approximate 3' UTR sequence, we aligned the same previously generated transcriptome sequences to the genome scaffolds flanking each FUL gene and annotated any transcript alignment that was not present in the existing exon annotation as a UTR.

For the *euFULI* and *euFULII* silencing constructs, we selected ~350bp regions encompassing a portion of exon 7, all of exon 8, and a portion of the 3' UTR for *NoFUL1*, *NoFUL2*, *NoMBP10*, and *NoMBP20*. We concatenated sequences of *NoFUL1* and *NoFUL2* or *NoMBP10* and *NoMBP20* to make the *euFULI* and *euFULII* silencing constructs, respectively. The construct to silence all genes comprised a 578bp region of *NoFUL2* exons 2-8 and 3' UTR with homology to all four genes. Each of these regions was synthesized by Genewiz (South Plainfield, NJ), cloned into TRV2 vectors in *Agrobacterium tumefaciens* GV3101 and stored as a glycerol stock for later use. The TRV1 (pYL192) and TRV2 (pYL156) vectors were provided as a generous gift from Dr. Dinesh-Kumar at UC Davis (Liu et al., 2002).

Plant Material

Seeds for desert tobacco were obtained from the New York Botanical Garden and germinated in a growth room at the University of California, Riverside maintained at 22°C for 24 h under 100 μmol m^{-2} s^{-1} light conditions at a 16h light and 8h dark photoperiod. We planted approximately 10-20 seeds per 4" pot and thinned to one seed per pot at the two-leaf stage. Prior to infiltration any plants with floral buds were discarded, and the remaining plants were divided into two groups based on the presence of an inflorescence stem.

Agrobacterium Infiltration

Inoculum for TRV1, TRV2-Empty, TRV2-euFULI, TRV2-euFULII, and TRV2-All was prepared and plants were infiltrated according to a previously established protocol for

Nicotiana benthamiana (Senthil-Kumar & Mysore, 2014). We infiltrated the top 3-4 leaves of each plant with a total of approximately 0.5mL of inoculum.

Phenotypic Measurements

One day after infiltration, the plants were individually numbered, randomized across flats, and returned to the original growth room. We recorded the developmental stage of the plants at five time points (11, 14, 17, 19, and 27 days post infiltration). Plants were in the vegetative stage if no elongated inflorescence was present. Once this inflorescence appeared, the plants were considered to have bolted (bolting stage). This stage persisted until floral primordia could be seen at the tip of the inflorescence (flower bud stage). The flower bud stage ended with the presence of the first flower with an open corolla limb (flower stage). Finally, any plant with a detached corolla tube was advanced to the fruit stage. We also scored plants at a single time point for the number of actively growing meristems on the inflorescence as a proxy for branching.

Statistical Analyses

Based on the five observation timepoints, we calculated the days to bolting, days to first flower, and days to first fruit. We considered plants infiltrated on rosette versus cauline leaves independently due to the confounding effect of developmental age on these response variables. For both the rosette and cauline leaf-infiltrated plants, we used a simple linear model in an ANOVA to determine if the silencing construct had a statistically significant effect on any of the three developmental milestones. Because plants infiltrated

at the cauline leaf stage had, by definition, already bolted, they were excluded from the days to bolting analysis.

For branch number, we also independently considered the rosette and cauline leaf-infiltrated plants, but because this count data did not follow a normal distribution, we applied a poisson generalized linear model and a Chi-squared test to determine if any of the constructs had a statistically significant effect on branch number.

All statistical analyses were conducted using custom R scripts and can viewed at a public GitHub repository (https://github.com/rajewski/VIGS). All figures for these analyses were also created in R using the packages ggplot2, patchwork, cowplot, and wesanderson (Pedersen, 2020; Ram & Wickham, 2018; Wickham, 2016; Wilke, 2019).

Figures and Tables

Figure A1.1 – Proportion of Plant at Developmental Stages by VIGS Construct over Time

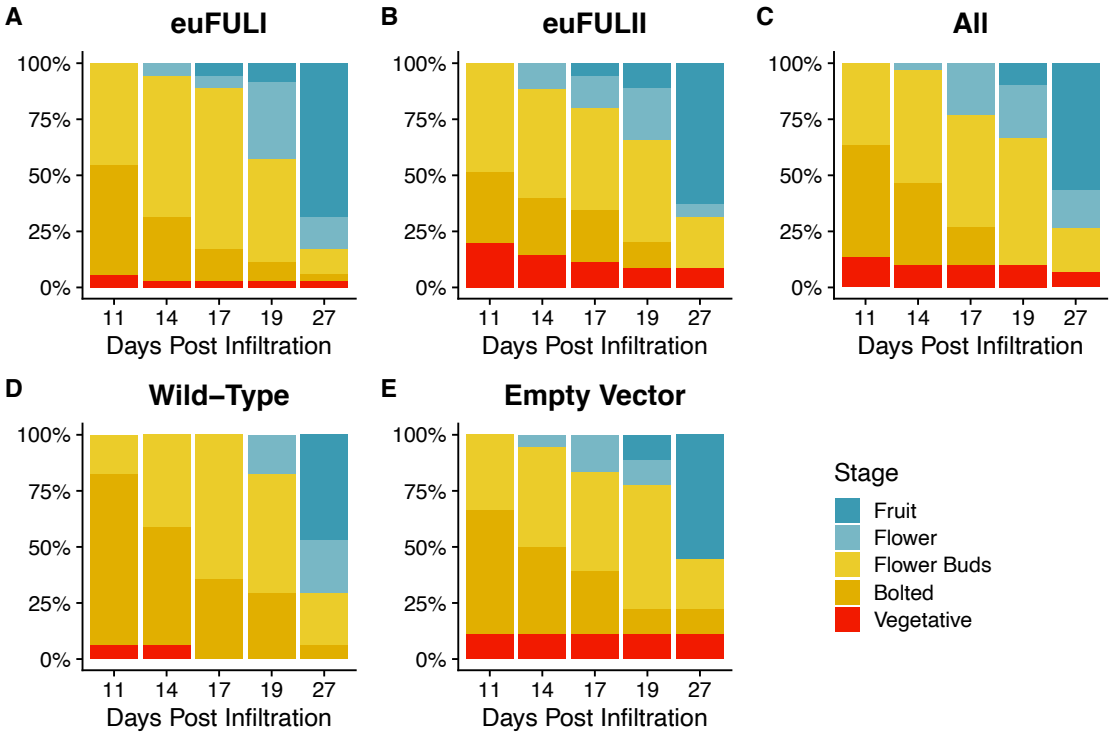


Figure A1.2 – Days to Developmental Milestones by VIGS Construct

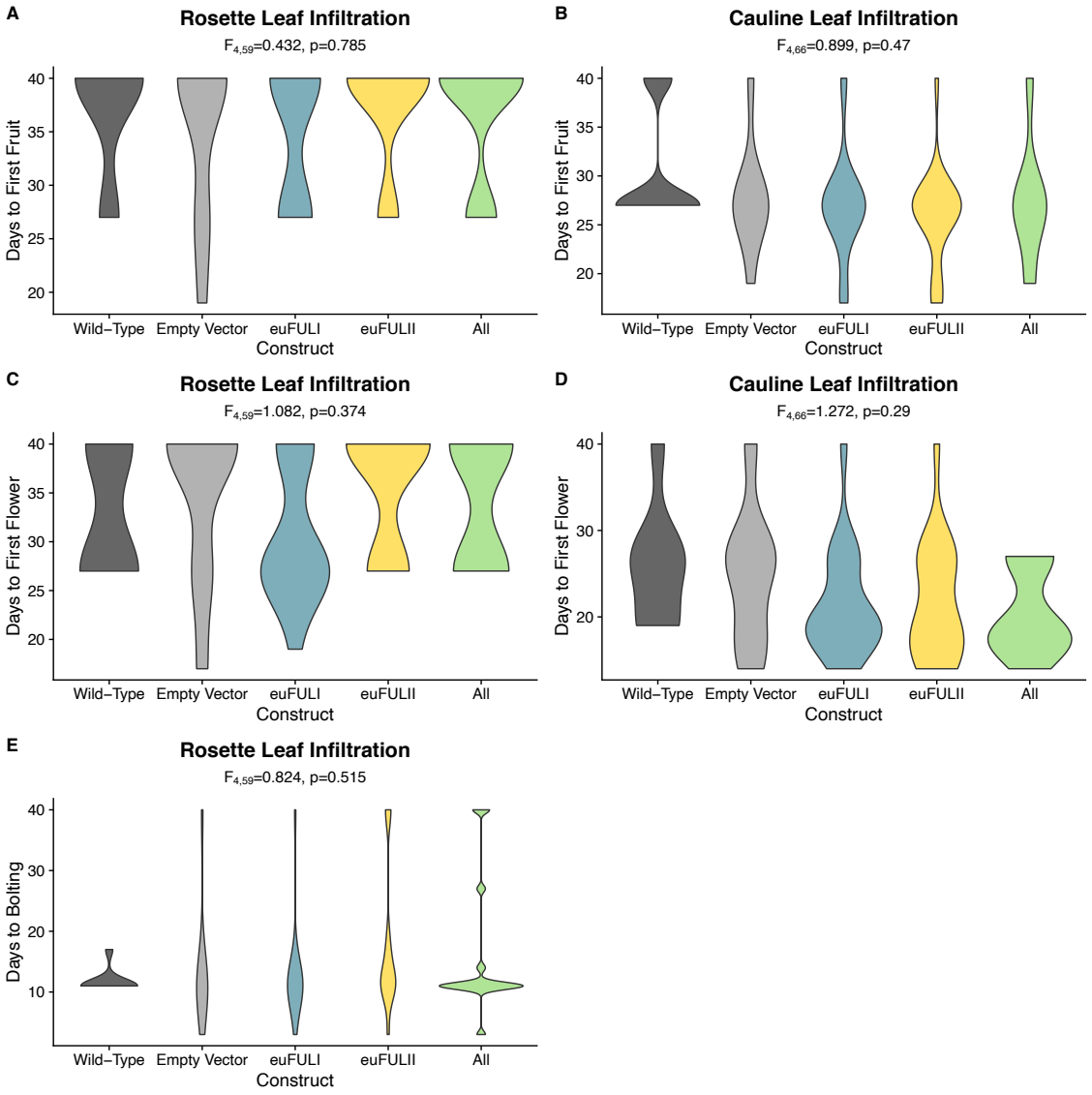


Figure A1.3 – Branching Summary by VIGS Construct

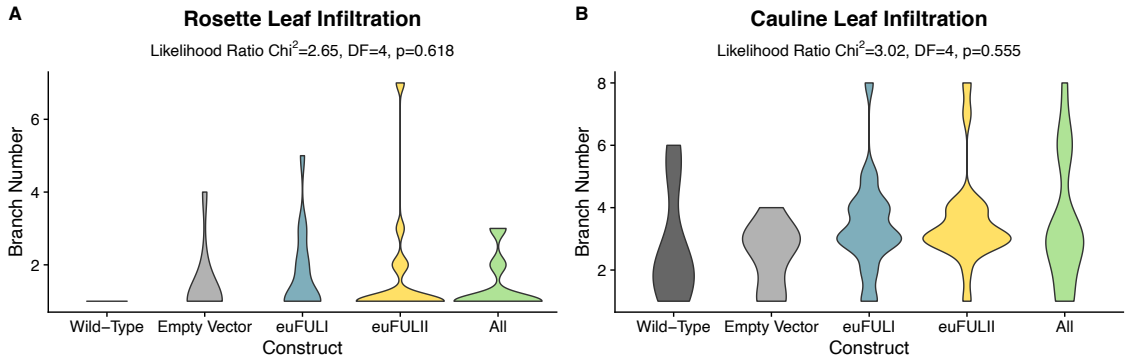


Table A1.1 – Replicate Summary by VIGS Construct and Developmental Stage

Summary of Replicates						
	Wild-Type	Empty Vector	euFULI	euFULII	All	Σ
Rosette	9	10	15	15	15	64
Cauline	8	8	20	20	15	71
Σ	17	18	35	35	30	135

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Balanzà, V., Martínez-Fernández, I., & Ferrándiz, C. (2014). Sequential action of FRUITFULL as a modulator of the activity of the floral regulators SVP and SOC1. *Journal of Experimental Botany*, *65*(4), 1193–1203.

Balanzà, V., Martínez-Fernández, I., Sato, S., Yanofsky, M. F., Kaufmann, K., Angenent, G. C., Bemer, M., & Ferrándiz, C. (2018). Genetic control of meristem arrest and life span in Arabidopsis by a FRUITFULL-APETALA2 pathway. *Nature Communications*, *9*(1), 565.

Bemer, M., Karlova, R., Ballester, A. R., Tikunov, Y. M., Bovy, A. G., Wolters-Arts, M., Rossetto, P. de B., Angenent, G. C., & de Maagd, R. A. (2012). The tomato FRUITFULL homologs TDR4/FUL1 and MBP7/FUL2 regulate ethylene-independent aspects of fruit ripening. *The Plant Cell*, *24*(11), 4437–4451.

Berbel, A., Ferrándiz, C., Hecht, V., Dalmais, M., Lund, O. S., Sussmilch, F. C., Taylor, S. A., Bendahmane, A., Ellis, T. H. N., Beltrán, J. P., Weller, J. L., & Madueño, F. (2012). VEGETATIVE1 is essential for development of the compound inflorescence in pea. *Nature Communications*, *3*, 797.

Burko, Y., Shleizer-Burko, S., Yanai, O., Shwartz, I., Zelnik, I. D., Jacob-Hirsch, J., Kela, I., Eshed-Williams, L., & Ori, N. (2013). A role for APETALA1/fruitfull transcription factors in tomato leaf development. *The Plant Cell*, *25*(6), 2070–2083.

Coenen, H., Viaene, T., Vandenbussche, M., & Geuten, K. (2018). TM8 represses developmental timing in Nicotiana benthamiana and has functionally diversified in angiosperms. *BMC Plant Biology*, *18*(1), 129.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Dardick, C. D., Callahan, A. M., Chiozzotto, R., Schaffer, R. J., Piagnani, M. C., & Scorza, R. (2010). Stone formation in peach fruit exhibits spatial coordination of the lignin and flavonoid pathways and similarity to Arabidopsis dehiscence. *BMC Biology*, *8*(1), 13.

Ferrándiz, C., Liljegren, S. J., & Yanofsky, M. F. (2000). Negative regulation of the *SHATTERPROOF* genes by FRUITFULL during *Arabidopsis* fruit development. *Science*, *289*(5478), 436–438.

Fujisawa, M., Shima, Y., Nakagawa, H., Kitagawa, M., Kimbara, J., Nakano, T., Kasumi, T., & Ito, Y. (2014). Transcriptional regulation of fruit ripening by tomato FRUITFULL homologs and associated MADS box proteins. *The Plant Cell*, *26*(1), 89–101.

Gao, R., Wang, Y., Gruber, M. Y., & Hannoufa, A. (2018). miR156/SPL10 Modulates Lateral Root Development, Branching and Leaf Morphology in Arabidopsis by Silencing AGAMOUS-LIKE 79. *Frontiers in Plant Science*, *8*, 2226.

Gu, Q., Ferrándiz, C., Yanofsky, M. F., & Martienssen, R. (1998). The *FRUITFULL* MADS-box gene mediates cell differentiation during Arabidopsis fruit development. *Development* , *125*(8), 1509–1517.

Hileman, L. C., Sundstrom, J. F., Litt, A., Chen, M., Shumba, T., & Irish, V. F. (2006). Molecular and phylogenetic analyses of the MADS-box gene family in tomato.

Molecular Biology and Evolution, *23*(11), 2245–2258.
Jaakola, L., Poole, M., Jones, M. O., Kämäräinen-Karppinen, T., Koskimäki, J. J., Hohtola, A., Häggman, H., Fraser, P. D., Manning, K., King, G. J., Thomson, H., & Seymour, G. B. (2010). A SQUAMOSA MADS box gene involved in the regulation of anthocyanin accumulation in bilberry fruits. *Plant Physiology*, *153*(4), 1619–1629.
Litt, A., & Irish, V. F. (2003). Duplication and diversification in the APETALA1/FRUITFULL floral homeotic gene lineage: implications for the evolution of floral development. *Genetics*, *165*(2), 821–833.
Liu, Y., Schiff, M., Marathe, R., & Dinesh-Kumar, S. P. (2002). Tobacco Rar1, EDS1 and NPR1/NIM1 like genes are required for N-mediated resistance to tobacco mosaic virus. *The Plant Journal: For Cell and Molecular Biology*, *30*(4), 415–429.
Østergaard, L., Kempin, S. A., & Bies, D. (2006). Pod shatter-resistant Brassica fruit produced by ectopic expression of the FRUITFULL gene. *Plant Biotechnology* , *4*(1), 45–51.
Pabón-Mora, N., Ambrose, B. A., & Litt, A. (2012). Poppy APETALA1/FRUITFULL orthologs control flowering time, branching, perianth identity, and fruit development. *Plant Physiology*, *158*(4), 1685–1704.
Pabon-Mora, N., & Litt, A. (2011). Comparative anatomical and developmental analysis of dry and fleshy fruits of Solanaceae. *American Journal of Botany*, *98*(9), 1415–1436.
Pabón-Mora, N., Sharma, B., Holappa, L. D., Kramer, E. M., & Litt, A. (2013). The Aquilegia FRUITFULL-like genes play key roles in leaf morphogenesis and inflorescence development. *The Plant Journal: For Cell and Molecular Biology*, *74*(2), 197–212.
Pařenicová, L., de Folter, S., Kieffer, M., & Horner, D. S. (2003). Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis new openings to the MADS world. *The Plant Cell*, *15*(7), 1538–1551.
Pedersen, T. L. (2020). *patchwork: The Composer of Plots*.
Ram, K., & Wickham, H. (2018). *wesanderson: A Wes Anderson Palette Generator*. https://CRAN.R-project.org/package=wesanderson
Schultink, A., Qi, T., Bally, J., & Staskawicz, B. (2019). Using forward genetics in Nicotiana benthamiana to uncover the immune signaling pathway mediating recognition of the Xanthomonas perforans effector XopJ4. *The New Phytologist*, *221*(2), 1001–1009.
Senthil-Kumar, M., & Mysore, K. S. (2014). Tobacco rattle virus-based virus-induced gene silencing in Nicotiana benthamiana. *Nature Protocols*, *9*(7), 1549–1562.
Shima, Y., Fujisawa, M., Kitagawa, M., Nakano, T., Kimbara, J., Nakamura, N., Shiina, T., Sugiyama, J., Nakamura, T., Kasumi, T., & Ito, Y. (2014). Tomato FRUITFULL homologs regulate fruit ripening via ethylene biosynthesis. *Bioscience, Biotechnology, and Biochemistry*, *78*(2), 231–237.
Smykal, P., Gennen, J., De Bodt, S., Ranganath, V., & Melzer, S. (2007). Flowering of strict photoperiodic Nicotiana varieties in non-inductive conditions by transgenic approaches. *Plant Molecular Biology*, *65*(3), 233–242.
Taheri-Dehkordi, A., Khandan-Mirkohi, A., Kafi, M., & Salami, S. A. (2018). Exploring and optimising the conditions for virus-induced gene silencing in an ornamental tobacco, Nicotiana alata. *The Journal of Horticultural Science & Biotechnology*, *93*(4), 377–384.

Tani, E., Polidoros, A. N., & Tsaftaris, A. S. (2007). Characterization and expression analysis of FRUITFULL- and SHATTERPROOF-like genes from peach (Prunus persica) and their role in split-pit formation. *Tree Physiology*, *27*(5), 649–659.

Wang, R., Tavano, E. C. da R., Lammers, M., Martinelli, A. P., Angenent, G. C., & de Maagd, R. A. (2019). Re-evaluation of transcription factor function in tomato fruit development and ripening with CRISPR/Cas9-mutagenesis. *Scientific Reports*, *9*(1), 1696.

Wang, S., Lu, G., Hou, Z., Luo, Z., Wang, T., Li, H., Zhang, J., & Ye, Z. (2014). Members of the tomato FRUITFULL MADS-box family regulate style abscission and fruit ripening. *Journal of Experimental Botany*, *65*(12), 3005–3014.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wilke, C. O. (2019). *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2."* https://CRAN.R-project.org/package=cowplot

Xu, S., Brockmoeller, T., Navarro-Quezada, A., Kuhl, H., Gase, K., Ling, Z., Zhou, W., Kreitzer, C., Stanke, M., Tang, H., Lyons, E., Pandey, P., Pandey, S. P., Timmermann, B., Gaquerel, E., & Baldwin, I. T. (2017). Wild tobacco genomes reveal the evolution of nicotine biosynthesis. In *bioRxiv* (p. 107565). https://doi.org/10.1101/107565