EVOLUTION, GENE EXPRESSION, AND EDUCATION: MODELING THE EVOLUTION OF POPPY (*PAPAVER,* PAPAVERACEAE) GENE EXPRESSION AND GRADUATE STUDENT PROFESSIONAL IDENTITY

by

AMANDA KELLY LANE

(Under the Direction of Jim Leebens-Mack)

ABSTRACT

This dissertation includes two parts. The first part, Chapters 2 and 3, address questions about evolution and transcript expression in the plant genus *Papaver* and relatives within the Ranunculales*.* Benzylisoquinoline alkaloids (BIAs) are an important class of plant secondary metabolites because they are varied and many are used as pharmaceuticals or are being explored for their pharmaceutical uses. BIAs are primarily produced by the order Ranunculales. A few of these alkaloids are only found in *Papaver* including some used as pharmaceuticals. In these chapters I infer the phylogenetic relationships of species within the Ranunculales including important representatives of *Papaver.* I resolve the relationships of major families within the order and explore the gene tree discordance across the phylogeny. RNA-seq analysis were then used to examine the expression of genes in BIA biosynthesis and also to describe gene co-expression networks in *Papaver somniferum* and *Papaver setigerum*. Comparative transcriptomics is useful for identifying genes of interest and understanding the evolution of pathways and processes between species. The second part of this dissertation, Chapter

4, acknowledges the importance of graduate training and undergraduate education for scientific progress. Science faculty who see themselves as teachers or have an identity as a teacher may be more likely to adopt evidence-based teaching practices. In this chapter I developed a mechanistic model of the factors that influence professional identity in graduate students. Specifically, I studied factors that hindered or promoted professional identity as a college teacher. Independent teaching experiences, teaching professional development, and teaching mentors contributed to salient and stable teaching identities among doctoral students. Being recognized by faculty as a teacher was also important, but rare. Participants observed that the professional culture of life sciences strongly valued research over teaching, resulting in a sometimes cold and isolating environment for students interested in teaching.

INDEX WORDS: *Papaver*, transcriptomics, phylogenomics, benzylisoquinoline alkaloids, co-expression networks, professional identity, graduate students, teaching identity

EVOLUTION, GENE EXPRESSION, AND EDUCATION: MODELING THE EVOLUTION OF POPPY (*PAPAVER*, PAPAVERACEAE) GENE EXPRESSION AND GRADUATE STUDENT PROFESSIONAL IDENTITY

by

AMANDA KELLY LANE

BS, Washington and Lee University, 2013

BA, Washington and Lee University, 2013

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2018

EVOLUTION, GENE EXPRESSION, AND EDUCATION: MODELING THE EVOLUTION OF POPPY (*PAPAVER*, PAPAVERACEAE) GENE EXPRESSION AND GRADUATE STUDENT PROFESSIONAL IDENTITY

by

AMANDA KELLY LANE

Major Professor:	Jim Leebens-Mack
Committee:	Tessa Andrews
	Chung-Jui Tsai
	Jessica C. Kissinger
	Jonathan Arnold

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2018

DEDICATION

This dissertation is dedicated to my parents, Coy and Kathy Lane. Without them this would not be possible. It is also dedicated to my friends Elizabeth “Trish” Lamb, Danielle “Miguel” Maurer, and Joanna “Meredith” Stanchek. Without them this would not have been nearly as much fun.

ACKNOWLEDGEMENTS

There are many people to thank for their support in writing this dissertation and their support throughout my doctoral training. In addition to my advisors and committee there are many people who helped me with my research and personal development.

Thank you to John Kerry who performed initial exploratory analyses with the *Papaver somniferum* data, and to collaborators Megan Augustin and Toni Kutchan for RNA-seq materials, data, and metabolite abundance data. Thanks go to the collaborators of the 1,000 Plant Transcriptomes project for transcriptomic data used throughout chapters 2 and 3. Additional thanks for my poppy-related work go to current and former members of the Leebens-Mack lab including Alex Harkess, Lauren Eserman, Michelle Hwang, Raj Ayyampalayam, and especially Karolina Heyduk for analysis advice, ideas and scripts.

Many members of the Andrews Lab gave significant assistance and advice on my professional identity research including Ansley Thomas, Carlton Hardison and Ariana Simon. Also, thank you to Michelle Ziadie and Anna Jo Auerbach for your support throughout it all.

The University of Georgia Biology Education Research Group (BERG) has been an incredible resource for my research, professional development, and personal growth. BERG is a tirelessly supportive community and I consider all of its members my mentors, even those members who are also my peers. Each member has aided me in some way and has my gratitude. Special thank you to Julie Stanton, Paula Lemons, and Jenn

Thompson who provided feedback. Thank you to Erin Dolan and Peggy Brickman for aiding in my professional development. Also, thank you to Lisa Limeri who has been an incredible sounding board while writing my dissertation.

Growing as an educator has been a key part of my graduate career and I will be forever thankful to those who have helped me along the way including Kris Miller, Lindsey Harding, Kim Brown, and Zoe Morris.

Most of the people already mentioned are not only colleagues or mentors but also my friends. However, I must also acknowledge a few more, Rachel Kerwin, Megan Behringer, Caitlin Conn, and Dan Frailey have all been my peer mentors and friends and have helped with the emotional elements of graduate school.

Finally, thank you to John Wares, your support early on is a key reason I made it to the end, and to Susan White, who helped me cross the t's and dot the i's.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Dissemination and public understanding of scientific discoveries rely on the combination of well-performed scientific inquiry and science communication. Science communication that likely impacts the greatest audience occurs in both K-12 and college classrooms. The work described here comes in two parts and acknowledges the importance of undergraduate science, technology, engineering and math (STEM) education in the process of science and for the future of scientific discovery. Chapters 2 and 3 of my dissertation address questions about evolution and transcript profiles in poppies (*Papaver*) and relatives within the order Ranunculales. Poppies and their relatives produce several pharmaceutically important plant secondary metabolites. The objectives of my work are to characterize gene co-expression networks associated with the production of these metabolites and hypothesize the species relationships in the Ranunculales, which will provide a foundation for future work.

Chapter 4 of my dissertation is discipline-based education research (DBER) that investigates the teaching identities of graduate students. This work culminates in a model of the factors that contribute to the professional identities of life sciences graduate students, specifically those factors that affect their identities as undergraduate instructors. In this introduction I describe background information that is relevant to both components of my dissertation.

Evolution and roles of plant secondary metabolism

Plant secondary metabolism can be defined as the byproducts of primary metabolism or as those plant compounds that do not contribute to cell division and growth (Hartmann 2007). Many studies have shown that these metabolites play a number of vital roles for plants including defense from herbivory, abiotic stress response, and attracting pollinators (Ziegler & Facchini 2008, Piasecka et al. 2015).

There are hundreds of thousands of identified secondary metabolites that have been classified based upon their chemical structure (Piasecka et al. 2015). Often, metabolites with similar structure will serve similar functions between and within species (Piasecka et al. 2015). For example, alkaloids are one large category of metabolites that are defined by having a nitrogen atom in a heterocyclic ring (Ziegler & Facchini 2008). Alkaloids often function as defense compounds, combating against both herbivores and pathogens (Ziegler & Facchini 2008).

Since plant secondary metabolites facilitate information flow between plants and their environment, it is no surprise that they are diverse and most likely constantly evolving in response to changing environments (Hartmann 2007). Plants can have morphological adaptations that are related to the metabolites they produce. Some species including many in *Papaver* have specialized structures and cell types used for storing and secreting metabolites. Laticifers are one such structure. They are important for plant-insect interactions because when laticifers are damaged they secrete a sticky, viscous substance called latex, which includes various metabolites that can combat herbivory such as alkaloids (Lange et al. 2015). Laticifers are only found in the angiosperm fossils younger than ~50mya (Lange et al. 2015).

Gene duplication followed by neofunctionalization is considered the main genetic mechanism that drives the diversification of plant secondary metabolism (Ober 2005). Metabolic genes often duplicate as tandem repeats and one or more of these tandem copies undergo neofunctionalization resulting in a cluster of genes that all produce enzymes that are part of the same biochemical pathway (e.g. Winzer et al. 2012, Nützmann & Osbourn 2014); however, some metabolic gene clusters also include non-paralogous genes (Ober 2005). A 10-gene cluster was recently discovered in *Papaver somniferum* that results in the production of the alkaloid noscapine and includes both genes from the same family and single genes from other families (Winzer et al. 2012). Therefore, understanding how these pathways evolve necessitates understanding how genes duplicate, neofunctionalize, and the evolutionary pressures that maintain these clusters (Ober 2005).

Understanding plant secondary metabolism diversification and evolution requires using a variety of study systems as well as interdisciplinary thinking that includes research from phylogenetics, genomics, biochemical, and ecological perspectives. In chapters 2 and 3 we focus on improving understanding species evolution and gene expression in *Papaver* species, with special attention to alkaloid biosynthesis.

Benzylisoquinoline alkaloids: natural purpose and agricultural interest

There are at least 2500 different alkaloids classified as benzylisoquinoline alkaloids (BIAs) including many of pharmaceutical relevance (Hagel & Facchini 2013, Beaudoin & Facchini 2014). BIAs are primarily produced by species in the Ranunculales, which includes *Papaver* (Papaveraceae), but they also are found in other orders such as the Nulumbonaceae and the Piperales (Liscombe et al. 2005). Liscombe et al. (2005)

propose a single evolutionary origin of BIA biosynthesis that occurred prior to the divergence of eudicots. However, this inference was based solely on biochemical activity of the enzyme (*S*)-norcoclaurine synthase (NCS), which is necessary for BIA production, and gene trees of a few BIA genes. While it may be true that synthesis of BIAs have a single origin, this hypothesis has not yet been rigorously tested.

Many commercial pharmaceuticals are derived from BIAs including the antitussive codeine, the analgesic morphine, and the antimicrobial and possible tumor suppressor sanguinarine, which has increased interest in studying BIAs (Ahmad et al. 2000, Ziegler & Facchini 2008, Hagel & Facchini 2013, Beaudoin & Facchini 2014). Morphine and codeine are thought to be primarily produced in a single species, *Papaver somniferum* (opium poppy) (Beaudoin & Facchini 2014). However, a closely related species *Papaver setigerum* is also known to produce morphine, but it produces markedly less morphine than *P. somniferum* and some populations do not produce detectable amounts of morphine at all (e.g. Chap. 3, La Valva et al. 1985). The many chiral centers that must be traversed for production of morphine make it commercially intractable to synthesize in a laboratory (Beaudoin & Facchini 2014). Instead purposeful breeding has resulted in cultivars of *P. somniferum* that produce increased amounts of particular metabolites including a high morphine producing line (Winzer et al. 2012).

Researchers have begun engineering *Saccharomyces cerevisiae* to produce both morphine and sanguinarine with the long-term goal of reducing production costs (Fossati et al. 2014, Galanie et al. 2015). However, *ex situ* function of both of these synthetic pathways require optimization before they equal productivity in plants. Further understanding of the regulation of BIA biosynthesis could inform researchers on how to improve these yields (Fossati et al. 2014, Galanie et al. 2015). Additionally, characterizing regulators of BIA biosynthesis could aid in targeted breeding strategies in opium poppy.

Benzylisoquinoline alkaloids: biosynthesis and regulation

The degree to which we know the enzymes and genes responsible for the production of any particular BIA varies. My work focuses on two pathways of interest the morphine biosynthesis pathway and the sanguinarine biosynthesis pathway. Morphine and sanguinarine biosynthesis share a precursor, (S)-reticuline, and contain many alkaloids of pharmaceutical interest (Hagel & Facchini 2013).

Whereas there have been great advances in understanding of BIA biosynthetic pathways, the regulation of genes involved in these pathways is not well characterized. Multiple tissues and cell types are involved in production of BIAs. Efficient biosynthesis requires coordinated transcriptional regulation across tissues and cell types (Beaudoin & Facchini 2014). There are several families of transcription factors that have been shown to regulate different pathways in plant secondary metabolism including bHLH, WRKY, MADS-BOX and MYB (Vom Endt et al. 2002). However, only a few transcription factors have been investigated for their role in BIA production such as psWRKY and cjWRKY1 (Mishra et al. 2013, Yamada et al. 2017).

Comparative transcriptomics have been successful in leading discovery of new genes involved in BIA synthesis, but so far these have been mostly limited to EST libraries and many have focused on cell cultures rather than untreated plant tissues (e.g. Zulak et al. 2007, Farrow et al. 2012). With decreasing costs, next-generation sequencing methods can be used in BIA producing species, including *P. somniferum,* to continue to discover proteins and genes that contribute to the production of BIAs. In chapter 3 we compare the transcriptomics of various tissues from *P. somniferum* and *P. setigerum* using RNA-seq.

Evolutionary history of Ranunculales and the Papaveraceae

In order to understand the evolution and diversification of BIA biosynthesis we must be well-grounded in our understanding of species evolution for those that produce BIAs, especially in the Ranunculales. Ranunculales also merit evolutionary interest outside of BIA production because they are sister to all other eudicots therefore holding a keystone position in the greater angiosperm phylogeny (Worberg et al. 2007, Wickett et al. 2014). Many of the family-level relationships within the Ranunculales have been previously resolved, but the placement of some lineages is still ambiguous and relationships among many species in the genus *Papaver* are uncharacterized with some analyses unable to resolve a polytomy in *Papaver* (Hoot et al. 1997, Carolan et al. 2006). Prior studies have relied on a limited number of genes to deduce their phylogenies, which might contribute to the challenge of delineating species relationships (e.g Hoot et al. 1997, Carolan et al. 2006). In chapter 2 we used a phylogenomics approach that relies on transcriptomic data to infer the phylogeny of the Ranunculales including *Papaver* species such as *P. somniferum*. By using transcriptomic data we were able to leverage nearly 900 genes to infer the species phylogeny and assess our confidence of how well it represents the genetic history of these species.

Professional identity may affect instructor choices

Despite high-profile calls for the adoption of evidence-based instructional practices in undergraduate STEM education, these practices are still not commonly used

(National Research Council, 2012). One hypothesis regarding the slow adoption of evidence-based strategies is that science faculty often do not see teaching as an important part of their identities as professionals, and thus do not prioritize changing their teaching (Brownell & Tanner 2012). A paper proposing this idea has been cited over 200 times, but there have been few studies of professional identity among current and future science faculty. Graduate school is a key time for professional identity development because it is when students are socialized into the profession (Austin 2002). Currently we know little about how doctoral students develop identities as college teachers including what factors may promote or hinder this development.

Studying professional identity in graduate students

In chapter 4 I aim to characterize the factors that promote and hinder teaching identity among life sciences doctoral students. I interviewed 33 life sciences doctoral students who had a variety of career interests and analyzed the transcripts of these interviews using iterative and collaborative qualitative content analysis (e.g. Saldaña 2013). From this analysis we developed a mechanistic model of the factors that influenced teaching identity in our participants. Independent teaching experiences, teaching professional development, and teaching mentors contributed to salient and stable teaching identities among doctoral students. Being recognized by faculty as a teacher was also important, but rare. Participants observed that the professional culture of life sciences strongly valued research over teaching, resulting in a sometimes cold and isolating environment for students interested in teaching. The culture also made it harder to find opportunities for teaching development and made it more challenging to take advantage of these opportunities. The mechanistic model described in this work is an important first step in understanding how doctoral training influences teaching identity and it should be tested and refined through additional empirical work.

References

Ahmad, N., Gupta, S., Husain, M. M., Heiskanen, K. M., & Mukhtar, H. (2000). Differential antiproliferative and apoptotic response of sanguinarine for cancer cells versus normal cells. *Clinical Cancer Research*, *6*(4), 1524-1528.

Austin, A. E. (2002). Preparing the Next Generation of Faculty. *The Journal of Higher Education*, 73(1), 94-122.

Beaudoin, G. A. W., & Facchini, P. J. (2014). Benzylisoquinoline alkaloid biosynthesis in opium poppy. *Planta, 240*(1), 19–32.

Brownell, S. E., & Tanner, K. D. (2012). Barriers to Faculty Pedagogical Change: Lack of Training, Time, Incentives, and...Tensions with Professional Identity? *CBE-Life Sciences Education*, 11(4), 339-346.

Carolan, J. C., Hook, I. L. I., Chase, M.W., Kadereit, J. W., & Hodkinson, T. R. (2006). Phylogenetics of Papaver and related genera based on DNA sequences from ITS nuclear ribosomal DNA and plastid trnL intron and trnL-F intergenic spacers. *Annals of Botany*, 98(1), 141–155.

Farrow, S. C., Hagel, J. M., & Facchini, P. J. (2012). Transcript and metabolite profiling in cell cultures of 18 plant species that produce benzylisoquinoline alkaloids. *Phytochemistry, 77,* 79-88.

Fossati, E., Ekins, A., Narcross, L., Zhu, Y., Falgueyret, J. P., Beaudoin, G. A., ... & Martin, V. J. (2014). Reconstitution of a 10-gene pathway for synthesis of the plant alkaloid dihydrosanguinarine in *Saccharomyces cerevisiae*. *Nature communications*, *5*, 3283.

Galanie, S., Thodey, K., Trenchard, I. J., Interrante, M. F., & Smolke, C. D. (2015). Complete biosynthesis of opioids in yeast. *Science*, *349*(6252), 1095-1100.

Hagel, J. M., & Facchini, P. J. (2013). Benzylisoquinoline alkaloid metabolism: a century of discovery and a brave new world. *Plant and Cell Physiology*, *54*(5), 647-672.

Hartmann, T. (2007). From waste products to ecochemicals: fifty years research of plant secondary metabolism. *Phytochemistry*, *68*(22-24), 2831-2846.

Hoot, S. B., Kadereit, J. W., Blattner, F. R., Schwarzbach, A. E., & Crane, P. R. (1997). Data congruence and phylogeny of the Papaveraceae s.l. based on four data sets: atpB and rbcL sequences, trnK restriction sites, and morphological characters. *Systematic Botany*, 22(3), 575–590.

Lange, B. M. (2015). The evolution of plant secretory structures and emergence of terpenoid chemical diversity. *Annual review of plant biology*, *66*, 139-159.

La Valva, V., Sabato, S., & Gigliano, G. S. (1985). Morphology and Alkaloid Chemistry of *Papaver setigerum DC*. (Papaveraceae). *Taxon, 34*(2), 191-196.

Liscombe, D. K., MacLeod, B. P., Loukanina, N., Nandi, O. I., & Facchini, P. J. (2005). Erratum to "Evidence for the monophyletic evolution of benzylisoquinoline alkaloid biosynthesis in angiosperms"[Phytochemistry 66 (2005) 1374–1393]. *Phytochemistry*, *66*(20), 2500-2520.

Mishra, S., Triptahi, V., Singh, S., Phukan, U. J., Gupta, M. M., Shanker, K. *et al.* (2013). Wound Induced Transcriptional Regulation of Benzylisoquinoline Pathway and Characterization of Wound Inducible PsWRKY Transcription Factor from *Papaver somniferum. PLOS One, 8*(1), e52784.

National Research Council (2012). *Discipline-based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*. Washington, DC: National Academies Press.

Nützmann, H. W., & Osbourn, A. (2014). Gene clustering in plant specialized metabolism. *Current opinion in biotechnology*, *26*, 91-99.

Ober, D. (2005). Seeing double: gene duplication and diversification in plant secondary metabolism. *Trends in Plant Science, 10*(9), 444-449.

Onoyovwe, A., Hagel, J. M., Chen, X., Khan, M. F., Schriemer, D. C., & Facchini, P. J. (2013). Morphine Biosynthesis in Opium Poppy Involves Two Cell Types: Sieve Elements and Laticifers. *The Plant Cell*, *25*(10), 4110–4122.

Piasecka, A., Jedrzejczak‐Rey, N., & Bednarek, P. (2015). Secondary metabolites in plant innate immunity: conserved function of divergent chemicals. *New Phytologist*, *206*(3), 948-964.

Saldaña, J. (2013). *The Coding Manual for Qualitative Researchers* (2nd ed.). Washington, DC: Sage.

Treutter, D. (2005). Significance of flavonoids in plant resistance and enhancement of their biosynthesis. *Plant biology*, *7*(6), 581-591.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45), E4859–E4868.

Winzer, T., Gazda, V., He, Z., Kaminski, F., Kern, M., Larson, T. R. *et al.* (2012). A *Papaver somniferum* 10-Gene Cluster for Synthesis of the Anticancer Alkaloid Noscapine. *Science*, *336*(6089): 1704-1708.

Worberg, A., Quandt, D., Barniske, A. M., Löhne, C., Hilu, K.W., & Borsch, T. (2007). Phylogeny of basal eudicots: Insights from non-coding and rapidly evolving DNA. *Organisms Diversity & Evolution*, 7(1), 55–77.

Yamada, Y., Shimada, T., Motomura, Y., & Sato, F. (2017). Modulation of benzylisoquinoline alkaloid biosynthesis by heterologous expression of CjWRKY1 in *Eschscholzia californica* cells. *PLOS One, 12*(10), e0186953.

Ziegler, J., & Facchini, P. J. (2008). Alkaloid biosynthesis: metabolism and trafficking. *Annu. Rev. Plant Biol.*, *59*, 735-769.

Zulak, K. G., Cornish, A., Daskalchuk, T. E., Deyholos, M. K., Goodenowe, D. B., Gordon, P. M., et al. (2007). Gene transcript and metabolite profiling of elicitor-induced opium poppy cell cultures reveals the coordinate regulation of primary and secondary metabolism. *Planta, 225*(5), 1085-1106.

CHAPTER 2

PHYLOGENOMIC ANALYSIS OF THE RANUNCULALES RESOLVES BRANCHING EVENTS ACROSS THE ORDER[1]

[1] Lane, A.K., Augustin, M.A., Ayyampalayam, S., Plant, A., Gleissberg, S., Di Stilio, V., dePamphilis, C. W., Wong, G.K., Kutchan, T.M, Leebens-Mack, J.H. Phylogenomic Analysis of the Ranunculales Resolves Branching Events Across the Order. Accepted by Botanical Journal of the Linnean Society. Reprinted here with permission of the publisher, 04/11/2018.

Abstract

Poppies (*Papaver*), columbines (*Aquilegia*), buttercups (*Ranunculus*) and related species, are members of the order Ranunculales, the sister lineage to all other eudicots. Using coalescent and concatenated methods to analyze alignments of 882 putatively single copy genes, we have developed a robust phylogeny including 27 species representing all major lineages within the Ranunculales and all tribes within the Papaveraceae. This first phylogenomic analysis including exemplar Ranunculales, Asterid, Caryophillid and Rosid species provides necessary foundation for future investigations of character evolution in the Ranunculales. For example, the production of benzylisoquinoline alkaloids with known and potential pharmaceutical applications is largely restricted to the Ranunculales. Understanding species relationships within the order is critical for understanding the evolution of benzylisoquinoline alkaloid biosynthesis across the Ranunculales including the production of morphine and codeine in opium poppy (*Papaver somniferum*). Analysis of gene tree discordance within select portions of the phylogeny suggests that the few observed differences between trees derived from supermatrix and coalescent-based summary analyses are attributable to incomplete lineage sorting. Discordance between gene tree and species tree inferences should be taken into account in future comparative analyses of character evolution within the Ranunculales.

Introduction

Eudicots comprise 75% of all flowering plant diversity. The Ranunculales holds a key position in the angiosperm phylogeny as the sister lineage to all other extant eudicots. Evolutionary analyses of the Ranunculales have contributed to elucidation of the diversification of important traits including perianth morphology, biochemical pathways, and woody vs. herbaceous habits (Kim et al. 2004, Liscombe et al. 2005, Sharma et al. 2011, Bartholmes et al. 2012, Pabon-Mora et al. 2013). The Ranunculales has significant pharmaceutical importance due to its unique production of benzylisoquinoline alkaloids (BIAs), including analgesics such as morphine and codeine, antibacterials such as sanguinarine and anticancer drugs such as noscapine (Liscombe et al. 2005, Hagel & Facchini 2013, Beaudoin & Facchini 2014). In order to better understand the evolution of BIA biosynthesis and other traits, a well-supported phylogeny of the Ranunculales is required. While many of the family level relationships within the Ranunculales have been resolved in previous studies, there are still ambiguities remaining for the placement of some lineages and resolution of species level relationships within families (Hoot et al. 1997, Soltis et al. 2000, Kim et al. 2004, Anderson et al. 2005, Carolan et al. 2006, Worberg et al. 2007, Hilu et al. 2008, Wang et al. 2009, Soza et al. 2013, Lehtonen et al. 2016). For example, the placement of Eupteleaceae is controversial and little work has been done to resolve relationships among *Papaver* L. species. Whereas much of what we know about the Ranunculales phylogeny is based on a few genes, many of which are encoded by the plastid genome, this study utilizes transcriptome data to provide a phylogenomic reconstruction of relationships across the order.

We use a phylogenomic approach to analyze species relationships for 27 species including exemplars for Ranunculids, Asterids, Caryophillids and Rosids. Putatively single copy loci have been identified from available genome sequences and transcriptomic data. We estimated and compared both concatenation and coalescent-based species tree inferences. Importantly, we also characterize nodes on the species phylogeny exhibiting possible conflict between gene trees and the species tree. Studying gene tree discordance provides additional insights for understanding ancestral character states and evolution of important traits such as BIA biosynthesis.

Methods

Transcriptome assembly

Illumina 100 bp paired-end RNA-seq reads were assembled for 20 ranunculid species and acquired from the 1000 Plant Transcriptomes (1KP) project (Matasci et al. 2014, Wickett et al. 2014). These 20 species represent five out of seven families within the Ranunculales including six poppy species (members of the subfamily Papaveroideae of the family Papaveraceae) exhibiting a broad range of BIA profiles. Tissues were collected by collaborators (Table 2.1), and RNA was extracted and sequencing was performed using previously described protocols (Jiao et al. 2012, Johnson et al. 2012). RNA Seq libraries were prepared with an insert size of ~200 base pairs and at least 2 Gb were sequenced for each sample as described in Wickett et al. (2014).

De novo transcriptome assemblies were built for each species using the Trinity *de novo* Assembler (v 2.0.6) platform including the *in silico* normalization pipeline with default parameters following a procedure described in Haas et al. 2013 (Grabherr et al. 2011, Haas et al. 2013). Read mapping and abundance estimation was then performed with RSEM (v 1.2.20) and all transcript assemblies with less than 1% of the per-component read mappings were removed (Li & Dewey 2011). Transdecoder was used to translate nucleotide sequences into amino acid sequences (Haas et al. 2013).

Gene Selection

Sequences from the resulting transcriptomes were assigned to orthogroups circumscribed in a global gene family classification for land plant genomes developed by the Amborella Genome Project (2013) Transcript assemblies were sorted into orthogroups using BLAST and gene families identified as single copy among the 14 eudicots included in the orthogroup circumscription were evaluated further. As described by Wickett et al. (2014) in cases where multiple transcripts from a species sorted into a single copy orthogroup, a scaffolding procedure was performed to collapse sequences that had 95% identity or greater. After scaffolding transcripts sorted into the putatively single copy gene families, additional copy number assessment and filtering were performed. If three or more of the 20 species contributed more than one sequence to an orthogroup that entire orthogroup was removed from the analysis. For the remaining orthogroups, sequences from species with more than one copy were removed from the gene family. Finally, only gene families with 80% of the 20 species included were retained for phylogenetic analysis.

After filtering species, occupancy in the gene families ranged from 58.7% to 100%. The differences in these percentages can be partially explained by differences in RNA-Seq sampling and sequence depth, which varied across taxa, as well as taxon specific duplication events, which would exclude sequences from consideration. Orthologs from *Amborella trichopoda* Baill*., Vitis vinifera* L*., Musa acuminata* Colla*,* *Solanum lycopersicum* L*.*, *Phoenix dactylifera* L., *Populus trichocarpa* Torr. & A. Gray ex Hook and *Aquilegia coerulea* E. James were added to the gene family datasets before multiple sequence alignment and gene tree estimation.

Phylogenetic analyses

For each orthogroup (i.e. estimated gene family cluster), peptide sequences were aligned using MAFFT (v7.215-e), and then nucleotides were mapped to the amino acids alignment using PAL2NAL (v. 14) (Katoh 2002, Suyama et al. 2006). Maximum-likelihood (ML) analyses were performed using RAxML (v. 7.3.0) with GTRGAMMA or PROTGTRGAMMA models for nucleotide and amino acid alignments respectively (Stamatakis 2006). *A. trichopoda*, the sister lineage to the clade containing all other extant angiosperms (e.g. Amborella Genome Project 2013), was used to root all gene trees. Therefore, gene families that were missing *A. trichopoda* from the alignment were dropped from further analyses. This resulted in 882 genes being used for phylogenetic analyses.

A coalescent-based analysis was conducted using RAxML bootstrap gene trees as input for ASTRAL (v. 4.7.6), which utilizes a multi-locus bootstrapping approach to assess species tree clade support while accounting for conflict within and among gene family alignments (Mirarab et al. 2014). An analysis of 882 concatenated gene alignments was also performed using RAxML with GTRGAMMA and PROTGTRGAMMA models for nucleotide and amino acid alignments, respectively, with 100 bootstrap replicates. In addition, gene tree quartet frequencies were calculated for the estimated species trees (Mirarab & Warnow 2015) and local posterior probabilities were calculated (Sayyari & Mirarab 2016a,b).

Inferred species trees included three regions with questionable resolution as reflected by low bootstrap support or conflict among trees estimated using contrasting analyses (i.e. AA vs nucleotide alignments, and concatenated vs ASTRAL analyses). Gene tree incongruence was assessed for these select relationships using custom scripts (github.com/kheyduk/Phylogenomics). For example, *Euptelea pleiosperma* Hook f. & Thompson was placed sister to the rest of the Ranunculales in some gene trees or sister to Papaveraceae in others. The numbers of trees with 50% or 80% bootstrap support for alternative placements for *E. pleiosperma* were counted using the getConflict.pl script (Heyduk et al. 2015, github.com/kheyduk/Phylogenomics). Relationships among the four *Papaver* species and species in the tribe Fumarioideae were also assessed by determining the number of gene trees that supported alternative relationships among the relevant lineages.

Results

Concatenated and coalescent analyses returned phylogenetic estimates with topologies that were largely congruent (Figure 2.1). Euptelaceae is sister to a clade including all other lineages within the Ranunculales. The Papaveraceae is sister to a group containing Lardizabalaceae and the Ranunculaceae, which is sister to Berberidaceae. Optimizing concordance among quartets in gene trees and the species tree inference, ASTRAL has been shown to converge on the true species phylogeny in the face of incomplete lineage sorting (Mirarab et al. 2014). The ASTRAL tree was supported by 87.7% of the quartets present in the gene trees estimated on peptide alignments and 92.3% of the quartets in gene trees estimated on the nucleotide alignments. This indicates a low level of incomplete lineage sorting across the majority of the species tree.

Our analyses recovered many of the previously described familial relationships within the Ranuculales, supported the circumscription of the Fumarioid lineage as a subfamily within the Papaveraceae, and resolved previously unresolved relationships between the sampled *Papaver* species (Figure 2.1) (Hoot et al. 1997, Kim et al. 2004, Wang et al. 2009, Pérez-Gutiérrez et al. 2012, Hoot et al. 2015, Sauquet et al. 2015). Estimated branch lengths in the concatenated alignment analysis revealed several points on the tree with rapid rates of speciation including the early diversification of the Ranunculales and *Papaver* (Figure 2.2). As expected (Degnan & Rosenberg 2009, Liu et al. 2009), incongruence among quartets estimated from gene trees was highly elevated at these points in the species phylogeny (Figure 2.1). Euptelaceae was resolved as sister to a clade including Ranunculaceae and Papaveraceae, but the abundances of gene tree quartets supporting alternative resolutions were nearly equal (Figure 2.1). We further investigated discordance among gene trees with respect to alternative relationships among the Euptelaceae, Ranunculaceae and Papaveraceae (Figure 2.3A). Hypotheses one and two shown in Figure 2.3A have both been recovered in the phylogenetic literature (Hoot et al. 1997, Kim et al. 2004, Wang et al. 2009, Pérez-Gutiérrez et al. 2012, Hoot et al. 2015), but hypothesis one clearly has stronger support in the gene trees and our species tree estimate (Figure 2.1). A second point on the tree included three species for which topology varied depending on the alignment type used for reconstruction (amino acid vs. nucleotide) (Figure 2.3B). Finally, three potential relationships between *Papaver* species were investigated due to variation in topology and bootstrap support in the different analyses (Figure 2.3C). *Papaver somniferum* L. and *Papaver setigerum* DC. have been circumscribed as sub-species (Malik et al. 1979, Garnock-Jones & Scholes 1990), and they reliably form a clade in our gene tree estimates.

The majority of gene trees support most of the relationships recovered in the ASTRAL analysis (Figure 2.1), but with respect to relationships among *Papaver* species, more gene trees supported the topology recovered in the concatenated analysis of the amino acid alignments (Figure 2.3C hypothesis 1). This result may implicate an anomaly zone wherein the most probable gene trees do not match the underlying species tree (Degnan & Rosenberg 2006, Rosenberg & Tao 2008). However, the impact of uncertainty in gene tree inference cannot be discounted since the majority of well-supported (bootstrap > 80%) gene trees were concordant with the ASTRAL tree.

Discussion

In the first phylogenomic study of the Ranunculales, we investigate the species phylogeny as well as the underlying gene tree incongruences in order to understand the evolutionary framework of this important clade in the angiosperm phylogeny. Here we illustrate the importance of understanding the nature of the input data including the incongruence among gene trees and among species tree estimates based on different methods of analysis. We identified lineages with significant gene tree incongruence that may be impossible to resolve even with large amounts of data.

The relationship between *Euptelea* Siebold & Zucc. and the rest of the Ranunculales has been difficult to predict due to low bootstrap support in previous anaylses (Wang et al. 2009). Understanding the phylogenetic position of *Euptelea* is important for understanding character state evolution within the Ranunculales (Kim et al.

2004). For example, the placement of *Euptelea* influences reconstruction of woody versus herbaceous habit within the Ranunculales (Kim et al. 2004, Ren et al. 2007). The placement also has implications for understanding the evolution of BIA biosynthesis because there is no evidence for BIA biosynthesis in *Euptelea,* whereas other lineages within the Ranunculales are known to produce BIAs (Liscombe et al. 2005).

In order to investigate the source of conflicting inferences for placement of *Euptelea* in previous studies (Soltis et al. 2000, Kim et al. 2004, Anderson et al. 2005, Worberg et al. 2007, Hilu et al. 2008, Wang et al. 2009) we assessed variation in phylogenetic signal among our gene trees (Figure 2.3, Figure S2.1). Among the 795 gene trees that included *Euptelea*, the greatest number of trees supported *Euptelea* as basal to the remaining Ranuculales, which coordinates with our phylogeny (Figure 2.3A tree 1). Additionally, the internode branch for *Euptelea* is short, one condition for incomplete lineage sorting, which could explain why some gene trees are incongruent with the species tree (Maddison & Knowles 2006).

Not all areas of conflict in the species tree showed clear gene tree support for one topology over another. The relationships between *Corydalis linstowiana* Fedde*, Capnoides sempervirens* (L.) Borkh.*,* and *Cysticapnos vesicaria* (L.) Fedde were examined because they were the only relationships that changed between the amino acid and nucleotide-based species trees in the coalescent analysis (Figure 2.2 node C). Both analyses yielded high bootstrap support for their respective branching orders, and the gene trees show some support for both topologies (Figure 2.3C).

Conflicting signal among gene trees was also observed in the resolution of relationships among sampled *Papaver* species (Figure 2.3C). The trees estimated from the concatenated analyses differ from those inferred in the ASTRAL analyses (compare Figures 2.1 and 2.2). Previous studies have failed to resolve relationships among *Papaver bracteatum* Lindl.*, P. somniferum,* and *Papaver rhoeas* L. (Hoot et al. 1997, Carolan et al. 2006). Relationships between *Papaver* species are important for understanding the evolution of morphinan alkaloids because *P. somniferum,* and its potential subspecies *P. setigerum,* are the only plant species to produce these alkaloids (Malik et al. 1979, Garnock-Jones & Scholes 1990, Liscombe et al. 2005). Additionally, *P. bracteatum* has been shown to produce close morphinan precursors, such as thebaine, whereas *P. rhoeas* has not been shown to produce these compounds (e.g. Sharghi & Lalezari 1967). Resolving the relationships among these three species is important for understanding the evolution of BIA biosynthesis. We determined the number of gene trees that supported each of three viable hypotheses with *P. setigerum* and *P. somniferum* forming a well-supported clade (Figure 2.3C). The distribution of gene tree resolutions suggests that it is least likely for *P. bracteatum* to be sister to *P. somniferum.* However, the relationship observed in the majority of estimated gene trees matches the topology recovered by the concatenated amino acid analysis (169 total for amino acid and 185 total for nucleotide, Figure 2.3C) with a balanced tree placing *P. bracteatum* and *P. rhoeas* in one clade sister to the *P. setigerum* + *P. somniferum* clade (Figure 2.3C tree1). This result is consistent with the expectation that with short intervals between speciation events, balanced gene trees may be more common than an unbalanced species phylogeny, a situation described as the anomaly zone (Degnan & Rosenberg 2006, Rosenberg & Tao 2008).

Nonetheless, the observed discordance among inferred gene trees must be considered in comparative analysis of trait evolution. The histories of genes underlying traits of interest may not be consistent with the species tree. Ancestral character state reconstructions and comparative analyses of trait evolution should consider the consequences of incomplete sorting and ancestral genes, and all alternative gene topologies should be considered. For example, we find that *P. rhoeas* is sister to the *P. setigerum* + *P. somniferum* clade in the ASTRAL trees. However, *P. rhoeas* is the only sampled *Papaver* species that does not produce thebaine. Genes underlying production of thebaine may have been lost in the *P. rhoeas* lineage or null alleles for one or more of these genes may have existed in the ancestral *Papaver* population and subsequently became fixed in the *P. rhoeas* lineage. This second scenario is clearly possible given the high frequency of inferred gene trees with *P. bracteatum* as sister to the *P. setigerum* + *P. somniferum* clade.

Clear understanding of gene tree discordance and processes that contribute to discordance are necessary in order to guide inferences about species relationships and character evolution. For example, a whole genome duplication event has been inferred at the base of the Ranunculales and has been shown to contribute significantly to the evolution of certain gene families (Cui et al. 2006, Pabón-Mora et al. 2013). In addition to incomplete lineage sorting and interspecific gene flow, polyploidization and subsequent loss of paralogous genes in diverging lineage could contribute to gene tree discordance even in studies such as this one that focus on putatively single copy genes. Comprehensive analyses of large numbers of gene alignments can help produce the data required to resolve species relationships in the face of ILS and other sources of gene tree discordance at various depths of a phylogenetic tree (e.g. Wickett et al. 2014, Smith et al. 2015).

Inferred phylogeny for the Ranunculales (Figure 2.1, Figure S2.1) support previous work on the order while resolving nodes that were difficult to resolve with fewer genes. We were able to resolve a polytomy for four *Papaver* species, which is useful for understanding BIA biosynthesis. A phylogenomic approach was able to elucidate species relationships while accounting for gene tree incongruence due to incomplete lineage sorting.

References

Amborella Genome Project. (2013). The Amborella Genome and the Evolution of Flowering Plants. *Science, 342*(6165), 1241089.

Anderson, C. L., Bremer, K., & Fris, E. M. (2005). Dating phylogenetically basal eudicots using rbcL sequences and multiple fossil reference points. *American Journal of Botany, 92*(10), 1737–1748.

Bartholmes, C., Hidalgo, O., & Gleissberg, S. (2012). Evolution of the YABBY gene family with emphasis on the basal eudicot *Eschscholzia californica* (Papaveraceae). *Plant Biology*, *14*(1), 11–23.

Beaudoin, G. A. W., & Facchini, P. J. (2014). Benzylisoquinoline alkaloid biosynthesis in opium poppy. *Planta*, *240*(1), 19–32.

Carolan, J. C., Hook, I. L. I., Chase, M.W., Kadereit, J. W., & Hodkinson, T. R. (2006). Phylogenetics of Papaver and related genera based on DNA sequences from ITS nuclear ribosomal DNA and plastid trnL intron and trnL-F intergenic spacers. *Annals of Botany*, *98*(1), 141–155.

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. Washington, DC: Sage.

Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J. et al. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Research, 16*, 738-749.

Degnan, J. H., & Rosenber, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution, 24*(6), 332–340.

Degnan, J. H., & Rosenberg, N. A. (2006). Discordance of Species Trees with Their Most Likely Gene Trees. *PLoS Genetics, 2*(5), e68

Garnock-Jones, P. J., & Scholes, P. (1990). Alkaloid content of Papaver somniferum subsp. setigerum from New Zealand. *New Zealand Journal of Botany, 28*(3), 367-369.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology, 29*(7), 644–652.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J. et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols, 8*(8), 1494–512.

Hagel, J. M., & Facchini, P. J. (2013). Benzylisoquinoline alkaloid metabolism: A century of discovery and a brave new world. *Plant and Cell Physiology, 54*(5), 647–672.

Heyduk, K., Trapnell, D. W., Barrett, C. F., & Leebens-Mack, J. (2015). Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biological Journal of the Linnean Society, 117*(1), 106-120.

Hilu, K. W., Black C., Diouf D., & Burleigh, J. G. (2008). Phylogenetic signal in matK vs. trnK: A case study in early diverging eudicots (angiosperms). *Molecular Phylogenetics and Evolution, 48*(3), 1120–1130.

Hoot, S. B., Kadereit, J. W., Blattner, F. R., Schwarzbach, A. E., & Crane, P. R. (1997). Data congruence and phylogeny of the Papaveraceae s.l. based on four data sets: atpB and rbcL sequences, trnK restriction sites, and morphological characters. *Systematic Botany, 22*(3), 575–590.

Hoot, S. B., Wefferling, K. M., & Wulff, J. A. (2015). Phylogeny and Character Evolution of Papaveraceae s. l. (Ranunculales). *Systematic Botany, 40*(2), 474–488.

Jiao, Y. Leebens-Mack, J. Ayyampalayam, S., Bowers, J. E., McKain, M. R., McNeal, J. et al. (2012). A genome triplication associated with early diversification of the core eudicots. *Genome Biology, 13*, R3.

Johnson, M. T. J., Carpenter, E. J., Tian, Z., Bruskiewich, R., Burris, J. N., Carrigan, C.T., et al. (2012). Evaluating Methods for Isolating Total RNA and Predicting the Success of Sequencing Phylogenetically Diverse Plant Transcriptomes *PLoS ONE, 7*(11), e50226.

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.

Kim, S., Soltis, D. E., Soltis, P.S., Zanis, M.J., & Suh, Y. (2004). Phylogenetic relationships among early-diverging eudicots based on four genes: Were the eudicots ancestrally woody? *Molecular Phylogenetics and Evolution, 31*(1), 16–30.

Lehtonen, S., Christenhusz, M. J. M., & Falck, D. (2016). Sensitive phylogenetics of *Clematis* and its position in Ranunculaceae. *Botanical Journal of the Linnean Society, 182*(4), 825-867.

Liu, L., Yu, L., Pearl, D. K., & Edwards, S. V. (2009). Estimating species phylogenies using coalescence times among sequences. *Systematic Biology, 58*(5), 468–477.

Maddison, W. P., & Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic biology, 55*(1), 21–30.

Malik, C. P., Mary, T. N., & Grover, I. S. (1979). Cytogenetic Studies in *Papaver* V. Cytogenetic studies on *P. somniferum* X *P. setigerum* hybrids and amphiploids. *Cytologia, 44*(1), 59–69.

Matasci, N., Hung, L. H., Yan, Z., Carpenter, E. J., Wickett, N. J., Mirarab, S., et al. (2014). Data access for the 1,000 Plants (1KP) project. *GigaScience, 3*, 17.

Mays, N., & Pope, C. (1995). Rigour and qualitative research. *British Medical Journal*, *311*(6997), 109-112.

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics, 31*(12), i44-i52.

Mirarab, S., Reaz R., Bayzid, M. S., Zimmerman, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics, 30*(17), i541–i548.

Pabón-Mora, N., Hidalgo, O., Gleissberg, S., & Litt, A. (2013). Assessing duplication and loss of APETALA1/FRUITFULL homologs in Ranunculales. *Frontiers in Plant Science, 4*, 358.

Pérez-Gutiérrez, M. A., Romerio-García, A. T., Salinas, M. J., Blanca, G., Carment, Fernández, M., & Suárez-Santiago, V.N. (2012). Phylogeny of the tribe fumarieae (papaveraceae s.l.) based on chloroplast and nuclear DNA sequences: Evolutionary and biogeographic implications. *American Journal of Botany, 99*(3), 517–528.

Ren, Y., Li, H. F., Zhao, L., & Endress, P. K. (2007). Floral morphogenesis in Euptelea (Eupteleaceae, Ranunculales). *Annals of Botany, 100*(2), 185–193.

Rosenberg, N. A., & Tao, R. (2008). Discordance of Species Trees with Their Most Likely Gene Trees: The Case of Five Taxa. *Systematic Biology, 57*(1), 131-140.

Sauquet, H., Carrive, L., Poullain, N., Sannier, J., Damerval, C., & Nadot, S. (2015). Zygomorphy evolved from disymmetry in Fumarioideae (Papaveraceae, Ranunculales): new evidence from an expanded molecular phylogenetic framework. *Annals of Botany, 115(*6), 895–914.

Sayyari, E., & Mirarab, S. (2016a). Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction. *BMC Genomics*, *17*(Suppl 10)**,** 783.

Sayyari, E., & Mirarab, S. (2016b). Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution, 33*(7), 1654-1668.

Shargi, N., & Lalezari, I. (1967). Papaver bracteatum Lindl., a Highly Rich Source of Thebaine. *Nature*, *213*(5082), 1244.

Sharma, B., Guo, C. Kong, H., & Kramer, E. M. (2011). Petal-specific subfunctionalization of an APETALA3 paralog in the Ranunculales and its implications for petal evolution. *New Phytologist, 191*(3), 870–883.

Smith, S. A., Moore, M. J., Brown, J. W., & Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance and gene duplications with examples from animals and plants. *BMC Evolutionary Biology, 15*(1), 150.

Soltis, D. E., Soltis, P. S., Chase, M. W., Mort, M.E., Albach, D. C., Zanis, M. et al. (2000). Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. *Botanical Journal of the Linnean Society, 133*(4), 381–461.

Soza, V. L., Haworth, K. L., & Di Stilio, V. S. (2013). Timing and Consequences of Recurrent Polyploidy in Meadow-Rues (*Thalictrum*, Ranunculaceae). *Molecular Biology and Evolution, 30*(8), 1940-1954.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics, 22*(21), 2688–2690.

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research, 34*(suppl_2), W609-W612.

Wang, W., Lu, A. M., Ren, Y., Endress, M. E., & Chen, Z. D. (2009). Phylogeny and classification of Ranunculales: Evidence from four molecular loci and morphological data. *Perspectives in Plant Ecology, Evolution and Systematics, 11*(2), 81–110.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, *111*(45), E4859–E4868.

Worberg, A., Quandt, D., Barniske, A. M., Löhne, C., Hilu, K.W., & Borsch, T. (2007). Phylogeny of basal eudicots: Insights from non-coding and rapidly evolving DNA. *Organisms Diversity & Evolution, 7*(1), 55–77.

Tables

Table 2.1: Voucher Table for RNA-Seq and Genomic Samples. Listed is relevant information for identifying sample sources for both the RNA-Seq samples (all Ranunculales) and the acquired genomic data (Ranunculales and various angiosperm outgroups). Samples used from 1KP include the 1KP identifier. Genomic data are indicated with GENOME and are further identified in Wickett et. al. 2014.

Identifier	**Family**	**Species**	**Tissue Type**	**Voucher Data**
WFBF	Berberidaceae	*Podophyllum peltatum*	leaves and root	Deyholos2013-11
YHFG	Berberidaceae	*Nandina domestica*	young leaves and flower buds	Soltis and Miles 2972
QTJY	Eupteleaceae	*Euptelea pleiosperma*	leaves	Chase 33135
CCID	Lardizabalaceae	*Akebia trifoliata*	leaves	Soltis and Miles 2751
NMGG	Papaveraceae	*Hypecoum procumbens*		GDAC 43575
AUGV	Papaveraceae	*Capnoides sempervirens*		NBGB 19871844
UDHA	Papaveraceae	*Cysticapnos vesicaria*		Goldblatt and Porter 12396 (MO)
XHKT	Papaveraceae	*Sanguinaria canadensis*	young shoot and flower buds	JLM2013-65 (GA)
YDEH, GOQJ, SMZF, CCHG, BFMT	Papaveraceae	*Argemone mexicana*	leaf, stem, root, flower bud, developing fruit	T. Kutchan 6173585 (MO)
UNPT, NJKC, RKGT, TUHA, ERXG	Papaveraceae	*Eschscholzia californica*	leaf, stem, root, flower bud, developing fruit	T. Kutchan 6173583 (MO)
RRID, TMWO, ZSNV	Papaveraceae	*Papaver bracteatum*	leaf, bulb, root, stem	M. Augustin MO 5452578
ACYX, GMAM, QZBA, IORZ, MVTZ	Papaveraceae	*Papaver rhoeas*	leaf, stem, root, flower bud, developing fruit	T. Kutchan 6173586 (MO)
FNXH, MLPX, JSVC, STDO, EPRK	Papaveraceae	*Papaver setigerum*	leaf, stem, root, flower bud, developing fruit	T. Kutchan 6173582 (MO)
BMRX, SUFP, MIKW, FPYZ, KKCW	Papaveraceae	*Papaver somniferum*	leaf, stem, root, flower bud, developing fruit	T. Kutchan 6173586 (MO)

ZGQD	Papaveraceae	*Corydalis linstowiana*	leaves	Chase 34442 K
XMVD	Papaveraceae	*Chelidonium majus*	leaf	M.K. Deyholos 2016-100
VGHH	Ranunculaceae	*Hydrastis canadensis*	young shoots and flowers	ABG19970771
ZUHO	Ranunculaceae	*Anemone hupehensis*	leaf	DWS NYBG 99/89A
GBVZ	Ranunculaceae	*Thalictrum thalictroides*	floral buds	V. Di Stilio 123 (WTU)
UPOG	Ranunculaceae	*Anemone pulsatilla*	leaf	M.K. Deyholos 2016-101
GENOME	Amborellaceae	*Amborella trichopoda*		
GENOME	Vitaceae	*Vitis vinifera*		
GENOME	Salicaceae	*Populus trichocarpa*		
GENOME	Ranunculaceae	*Aquilegia coerulea*		
GENOME	Solanaceae	*Solanum lycopersicum*		
GENOME	Arecaceae	*Phoenix dactylifera*		
GENOME	Musaceae	*Musa acuminata*		

Figures

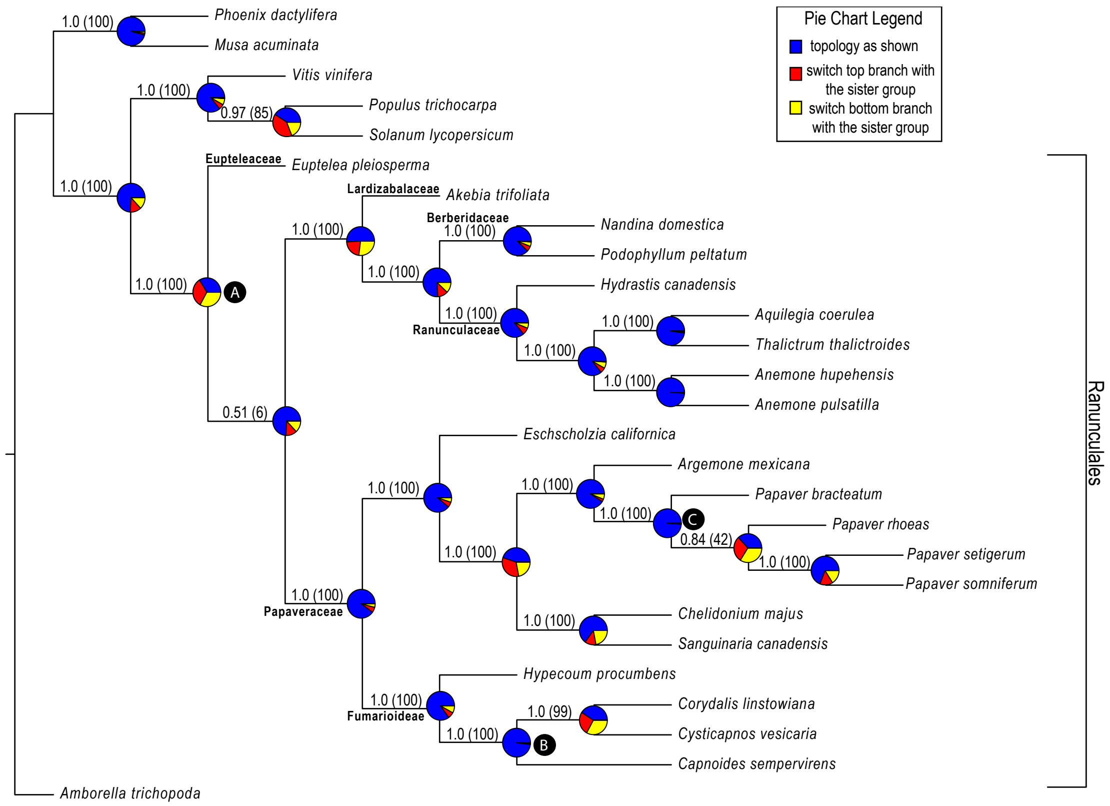


Figure 2.1: ASTRAL Species-Tree from 882 gene trees estimated from amino acid alignments. Bootstrap values are in parentheses and local posterior probabilities are adjacent. Pie charts depict the percentage of quartets from all gene trees that support one of three topologies: Q1 (blue) is the topology as shown, Q2 (red) is the lower child with the sister group and the upper child with the outgroup, Q3 (yellow) is the upper child with the sister group and the lower child with the outgroup (Mirarab & Warnow, 2015). Letters correspond to nodes discussed in the other figures. Branches are annotated with relevant family and tribe names in bold.

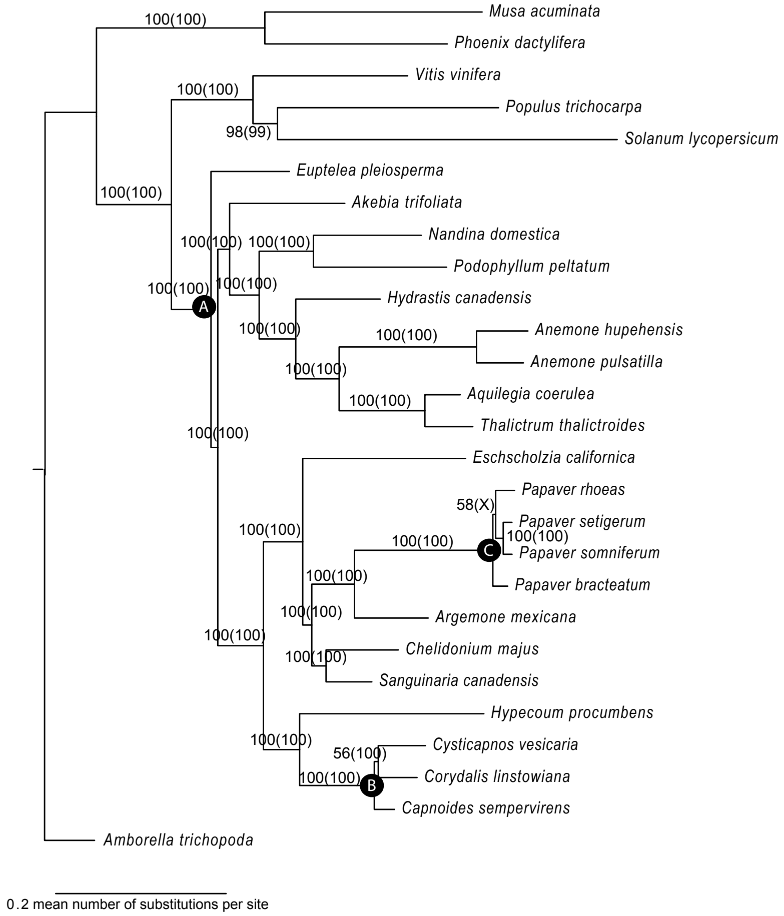


Figure 2.2: Concatenated Species Tree. This is a species-level tree was constructed using RAxML from a concatenated sequence from alignments of all genes included in the ASTRAL analysis. Both nucleotide and amino acid reconstructions' bootstrap values are included with the nucleotide topology shown and the corresponding bootstrap values listed first. In parentheses are the bootstrap values from the amino acid-based analysis. X indicates partitions not included in the amino acid topology.

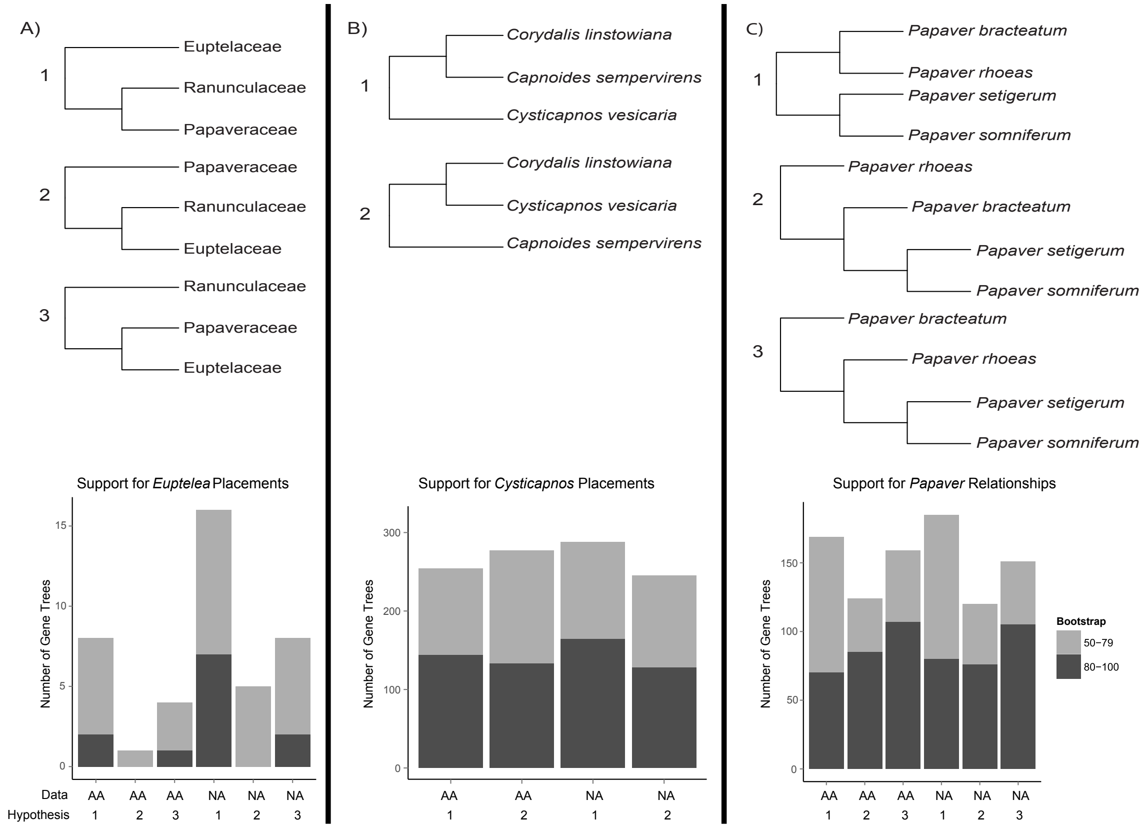


Figure 2.3: (Top) Hypotheses for Unresolved Species Relationships. The letters of these hypotheses correspond to the nodes shown in Figs. 1 and 2. **(Bottom) Bar Graphs Showing Number of Gene Trees that Support Varying Hypotheses.** The Figure Legend applies to parts A-C. The hypotheses numbers reference the hypotheses introduced in Figure 2.1 and the data type represents if nucleotide (NA) or amino acid (AA) alignments were used to make the trees. Gene trees had to have a bootstrap support for each hypothesis either from 50-79 or above 80 as indicated by shading. A) Input gene trees were required to have Ranunculaceae and Papaveraceae families to each be monophyletic with a bootstrap value of 50 or more, but did allow for missing taxa. *Euptelea pleiosperma* had to be present in all trees. B) Input trees were required to have all three sampled members of the Fumarioideae *Corydalis linstowiana, Cysticapnos vesicaria,* and *Capnoides sempervirens*. C) Input gene trees were required to have *Papaver somniferum* and/or *Papaver setigerum* present. If both species were present they were assumed to be sister.

CHAPTER 3

CO-EXPRESSION ANALYSIS AND COMPARATIVE TRANSCRIPTOMICS OF *PAPAVER SOMNIFERUM* AND THE CLOSELY RELATED *PAPAVER SETIGERUM*

Introduction

The secondary compound biochemistry of Papaver *somniferum, Papaver setigerum* and other species in the Ranunculales has been rigorously investigated for over two hundred years, but genome and transcriptome resources for the order remain limited (Hagel & Facchini 2013). Primary interest in *P. somniferum* and other poppy species is motivated by their production of a class of plant secondary metabolites, benzylisoquinoline alkaloids (BIAs), because many of these alkaloids are currently used as or are being explored for their use as pharmaceuticals (Hagel & Facchini 2013, Beaudoin & Facchini 2014). BIAs are only produced in certain plant species, most of which are in the angiosperm order Ranunculales, and each species varies in which BIAs it produces as well as the quanitity produced (Ziegler et al. 2005, Ziegler et al. 2006, Ziegler & Facchini 2008). In particular, *P. somniferum* and *P. setigerum* are the only plant species known to produce two important BIAs, the analgesic morphine and the antitussive codeine. The production of unique, pharmaceutically important alkaloids makes *Papaver* an interesting system in which to investigate the evolution of plant secondary metabolites (e.g. Pichersky & Gang 2000, Theis & Lerdau 2003, Ober 2005). However, the lack of genomic resources for *Papaver* has limited our ability to dissect evolutionary relationships between these species. Here we explore the transcriptomes of both *P. setigerum* and *P. somniferum*, and advance investigations of genetic divergence between these species, particularly with respect to BIA biosynthesis.

Plant secondary metabolism involves large, complex pathways that vary between species (Hartmann 2007, Ziegler & Facchini 2008, Farrow et al. 2012). Comparative transcriptomics is one way to investigate variation in plant secondary metabolism (Ziegler et al. 2005, Ziegler et al. 2006, Yonekura-Sakakibara et al. 2008, Usadel et al. 2009). Thus far, transcriptome sequencing in BIA producing species has mostly been limited to sequencing of low coverage EST libraries (e.g. Zulak et al. 2007, Farrow et al. 2012, Desgagné-Penix et al. 2012). However, developing BIA biosynthesis into a model for investigating evolution of plant secondary metabolism requires deep understanding of the pathway, regulatory networks, and the evolutionary relationships between study species. Therefore, we performed deeper replicated RNA-seq analysis of variation in gene expression among *Papaver* tissues.

The phylogenetic relationships of *P. somniferum* and *P. setigerum* with other members of the Papaveraveae and Ranunculales have recently been elucidated using phylogenomic approaches (Chap. 2). It has been long held that *P. somniferum* and *P. setigerum* are at least close sister species if not subspecies (e.g. Chap. 2, La Valva et al. 1985, Garnock-Jones et al. 1990, Hosokawa et al. 2004, Carolan et al. 2006). The taxonomic delineations are important from a legal and conservation perspective (La Valva et al. 1985, Garnock-Jones et al. 1990, Hosokawa et al. 2004), but understanding the degree of divergence between *P. somniferum* and *P. setigerum* is also important for comparative analyses of phenotypic differences.

The degree to which we know the enzymes and genes responsible for the production of any particular BIA varies. In this study we focus on two pathways of interest in *P. somniferum* and *P. setigerum*, the morphine biosynthesis and sanguinarine biosynthesis pathways. Here we refer to enzymes and genes known to be part of these pathways as morphine- and sanguinarine-related. These pathways are of interest because sanguinarine is an antimicrobial agent and morphine is a common analgesic and is used to create many other analgesics (Hagel & Facchini 2013, Beaudoin & Facchini 2014). Figure 3.2 summarizes these pathways that are described here.

Morphine and sanguinarine biosynthesis share a precursor, (S)-reticuline (Hagel & Facchini 2013). The morphinian biosynthesis pathway begins with 1,2-dehydroreticulinium synthase (DRS) transforming (S)-reticuline into 1,2-dehydroreticulinium (Hirata et al. 2004). This is then converted to (R)-reticultine by 1,2, dehydroreticulinium reductase (DRR) (De-Eknamkul & Zenk 1992). (R)-reticuline is converted to salutaridine by salutaridine synthase (SalSyn) and then reduced to salutaridinol by salutaridine reductase (SalR) (Gesell et al. 2009, Ziegler et al. 2006). Salutaridinol 7-O-acetyltransferase (SalAT) acts upon salutaridinol converting it to 7-O-acetylsalutaridinol (Grothe te al, 2001). There can be a spontaneous reaction transforming 7-O-acetylsalutaridinol into thebaine (Lenz & Zenk 1994). Alternatively, 7-O-acetylsalutaridinol may be converted to thebaine through interaction with thebaine synthase (THS) (Fisinger et al. 2007). Thebaine is then synthesized into neopinone by thebaine 6-O-demethylase (T6ODM) (Hagel & Facchini 2010). Neopinone can either be converted into oripavine by codeine demethylase (CODM) or spontaneously alter to become codeinone. T6ODM then catalyzed the conversion of oripavine to morphinone, which is subsequently interacts with codeinone reductase (COR) resulting in morphine. Codeinone can also be converted to morphine by first interacting with COR to become codeine and then demethylated by CODM (Hagel & Facchini 2010).

Sanguinarine biosynthesis begins with (S)-reticuline reacting with the berbine bridge enzyme (BBE) producing (S)-scoulerine (Dittrich & Kutchan 1991). (S)-chelanthifoline synthase (CheSyn) converts (S)-scoulerine to (S)-chilanthifoline, which is then converted to (S)-stylopine by (S)-stylophine synthase (StySyn) (Bauer & Zenk 1991, Ikezawa et al. 2007, Diaz Chavez et al. 2011). (S)-stylopine interacts with tetrahydroprotoberberine cis-N-methyltransferase (TNMT) creating (S)-cis-N-methylstylopine (Liscombe & Facchini 2007). This is coverted to protopine by protopine 6-hydoxylase (P6H) (Takemura et al. 2012). Protopine interacts with methylsylstylopine 14-hydroxylase (MSH) producing 6-hydoxyprotopine, which spontaneously converts to dihydosanguinarine (Tanahashi & Zenk 1988, Beaudoin & Facchini 2013). Finally, dihydrosanguinarine is oxidized by dihydropbenzophenanthridine oxidase (DBOX) resulting in sanguinarine, which can be converted back to dihydrosanguinarine by sanguinarine reductase (SanR) (Vogel et al. 2010, Hagel et al. 2012)

Genetic regulation of these pathways is less well understood than their biochemical counterparts. A few gene families have been implicated in regulation of BIA production, most notably WRKY transcription factors, but their regulatory mechanisms are yet to be discovered (Kato et al. 2007, Phukan et al. 2013, Beaudoin & Facchini 2014, Yamada et al. 2017). We do know that BIAs and BIA-related genes have tissue-specific expression patterns that may drive variation in alkaloid biosynthesis among species, tissues, and specific cell types (Bird et al. 2013, Onoyovwe et al. 2013, Chen & Facchini 2013).

Comparative transcriptomics is one method that can open new avenues for research into BIA regulation, grow research resources for *P. somniferum,* and elucidate genetic and metabolic divergence between *P. somniferum* and *P. setigerum.* We elected to use both differential expression and Weighted Gene Correlation Network Analysis (WGCNA) to compare transcriptomes from *P. somniferum* and *P. setigerum* (Langfelder & Hovarth 2008). In general, co-expression analyses such as WGCNA aim to identify clusters of genes that have correlated expression patterns across samples, which may be various tissues, individuals, time points, or conditions. Clusters of co-expressed genes often consist of genes that share regulation or play a role in similar pathways or mechanisms (Wolfe et al. 2005). WGCNA weights the correlations between genes exponentially such that strong correlations are accentuated in analyses aimed at defining modules (i.e. clusters) of co-expressed genes (Langfelder & Hovarth 2008, Zhao et al. 2010). We focus on identification of transcripts that are co-expressed with morphinine- and sanguinarine-related genes.

In summary, we use comparative transcriptomics and specifically co-expression network analysis to compare and contrast *P. somniferum* and *P. setigerum*. We describe these contrasts between the two potential subspecies' transcriptomes and co-expression networks. In doing so we investigate the genetic divergence between the two species and lay a foundation for further research into the regulation of morphine and sanguinarine biosynthesis.

Methods

Transcriptome assembly and assessment

RNA-seq libraries were made for three individuals for each species and five tissues for each individual (flower bud, stem, root, leaf, and capsule). These libraries are a part of the 1000 Plant Transcriptomes (1kp) project (Matasci et al. 2014, Wickett et al. 2014). Collaborators of 1kp performed RNA extraction and sequencing as described previously in Jiao et al. (2012) and Johnson et al. (2012). Libraries were sequenced with Illumina 100bp, paired-end reads with an insert size of ~200 bp, and at least 2Gb were sequenced for each sample as further described in Wickett et al. (2014). The same tissues that were used for RNA-seq were also analyzed for alkaloid content using liquid chromatography tandem mass spectrometry (LC-MS/MS). The data were analyzed for a specific subset of BIAs.

A *de novo* transcriptome assembly was built for *P. somniferum* using Trinity *de novo* Assembler (v2.0.6) with *in silico* normalization and default parameters following the procedures described by Haas et al. (2013) (Grabherr et al. 2011). A reference transcriptome was constructed from *P. somniferum* RNA-seq data and reads from both species were mapped to the *P. somniferum* reference. TransDecoder (2.0.1) was used with default parameters to predict full-length sequences and subsequently filter the transcriptome for transcripts with open reading frames (https://github.com/TransDecoder). Read mapping and abundance estimation using was conducted using RSEM (v1.3.0), which employs transcripts per million (TPM) as the unit of transcript expression and performs normalization of transcript expression across libraries using trimmed mean of M-values (TMM) (Robinson & Oschlack 2010, Li &

Dewey 2011, Wagner et al. 2012). We chose to use Trinity isoforms rather than genes or components for our analyses to increase our chances to characterize expression differences between different copies of the morphine- and sanguinarine-related genes.

Expression estimation for *P. somniferum* and *P. setigerum* libraries transcripts was performed and isoforms with less than 10% per component read mapping were removed (Li & Dewey 2011). We also used a TPM threshold of 1 across all libraries in order to distinguish expressed transcripts. Transcripts were annotated by performing blastn to the NCBI nucleotide collection with an e-value cutoff of 1e-5 and the result with the top bitscore was used (Altschul et al. 1990, Camacho et al. 2008).

We assessed the how well the transcriptome represented both the *P. somniferum* RNA-seq libraries as well as the *P. setigerum* libraries. TransRate (v1.0.3) was used following their established protocol as an additional assessment of open reading frames, read mapping using SNAP, N50, and how well the transcriptome represented *P. setigerum* relative to *P. somniferum* using the TransRate score as a metric (Zaharia et al. 2011, Smith-Unna et al. 2016). We ran TransRate for each species separately. TransRate outputs several statistics including percentage of read pairs that mapped back to the assembly and a TransRate assembly score which considers the quality of all contigs and read mapping. Assemblies within a single analysis can be compared using this score, but the authors of TransRate also analyzed numerous publicly available assemblies to compare against (Smith-Unna et al. 2016).

Network and differential expression analyses

We analyzed transcript expression by studying both patterns of differential expression between all samples as well as performing Weighted Gene Correlation

Network Analysis (WGCNA) (Langfelder & Horvath 2008). Differentially expressed transcripts were determined using default parameters in EdgeR (version 3.20.6) executed through the Trinity (v2.4.0) pipeline with samples from the same tissue in the same species designated as biological replicates (Robinson et al. 2010, Robinson & Oshlack 2010, McCarthy et al. 2012).

Unlike many other co-expression analyses that only consider correlations between transcripts that are beyond a given threshold, WGCNA considers all correlations. Rather than using a cutoff, correlations are weighted through a power transformation where the power is a provided parameter referred to as beta. In this way, strong correlations are weighted higher compared to weaker correlations (Langfelder & Horvath 2008).

We ran WGCNA using the corresponding R package (v1.4.9). Analyses focused on the 15,000 transcripts that were most heterogeneously expressed across all libraries for use in the WGCNAs for both *P. somniferum* and *P. setigerum*. The EdgeR pipeline from Trinity (v2.0.6) was used to identify the most heterogeneously expressed transcripts without using the biological replication option. Pairwise comparisons between all libraries were done, producing logFC information for all transcripts between all libraries. This log transformation of the data decreases the impact of outliers as well as the impact of expression in only a few libraries when determining variation. We then calculated the variation of the logFC and selected the 15,000 transcripts with the most variation. The remaining transcripts exhibited little variation in expression and thus provide no information about gene expression correlations (Langfelder & Horvath 2008). We also followed this same procedure to determine what would be the 15,000 most heterogeneously expressed genes when considering each species separately rather than combined.

To perform WGCNA, first we had to calculate the correlation between transcripts using expression in each library from a single species. We chose to use the Gini Correlation Coefficient (GCC) because it is robust to non-normal distributions, is stable to outliers, is less dependent on sample sizes, and does not rely on an *a priori* distribution unlike the more commonly used Pearson Correlation Coefficient (Ma & Wang 2012, https://github.com/liangjiaoxue/PythonCalculation). We followed the protocols outlined in Langfelder & Horvath (2008) and related work for an unsigned network (Dudoit et al. 2002, Langfelder et al. 2011, https://labs.genetics.ucla.edu/horvath/co-expressionNetwork/Rpackages/WGCNA/Tutorials/). Following recommendations in these references, we selected a beta value of 9 and a dynamic merge value of 0.2 for both the *P. somniferum* network and the *P. setigerum* network (Figure S3.1). We also constrained module sizes to be at least 20 transcripts. Additional analyses of WGCNA outputs utilized code from the online tutorials, self-designed scripts, and took inspiration and some code from Oldham et al. (2006). These included calculations of eigengene correlations, calculations of the number of transcripts that overlapped per module, and comparisons of transcript expression and transcript connectivity. Modules were visualized using Cytoscape utilizing subsets of transcripts and only viewing connections of 0.8 or greater due to computing limitations (Shannon et al. 2003, Smoot et al. 2011). Layouts of the modules were arranged for visualization using a Prefuse Force Directed Layout, which is the default for Cytoscape.

Results

Transcriptome assembly and assessment

After filtering the *P. somniferum* transcriptome assembly using Transdecoder, TransRate was used to assess the transcriptome quality (https://github.com/TransDecoder, Smith-Unna et al. 2016). Almost 61,000 transcripts were determined by both TransRate and Transdecoder to have an open reading frame, similar to the final number of transcripts that remained after filtering for lowly expressed transcripts and well-represented isoforms (Table S3.1). Approximately 69% of *P. somniferum* read pairs and 74% of *P. setigerum* read pairs mapped back to the reference transcriptome assembly (Table 3.1). TransRate scores for the reference transcriptome were assessed with each *P. somniferum* (0.257) and *P. setigerum* (0.214) read sets. Assessment statistics for the *P. somniferum* and *P. setigerum* mapped reads were not widely different from each other and were on par with many published *de novo* transcriptomes (>0.22) (Table 3.1) (Smith-Unna et al. 2016).

Differential expression

Clustering samples based on correlation of differentially expressed transcripts showed that the same tissues from different species share more similar transcript expression than different tissues from the same species (Figure 3.2). Large-scale patterns among tissues were evident for both species. For example, while *P. somniferum* and *P. setigerum* root samples showed similar expression patterns with each other, they are distinct from all other tissues. In comparison, leaf and stem tissues had similar expression patterns within and across species (Figure 3.2). Root tissues were again noticeably distinct in that they have the largest group of transcripts showing decreased expression compared to that in other tissues (Figure S3.2).

Assessing WGCNA for each species

The top 15,000 most heterogeneously expressed transcripts across all *P. somniferum* and *P. setigerum* libraries were used as the transcript set in the WGCNA. Over 11,000 of these transcripts were in the most heterogeneously expressed set for both species and the combined set. The combined set only included 256 transcripts that were not in most heterogeneously expressed for either species individually (Figure S3.3). In addition, we compared how many different isoforms of a single gene were included in the 15,000 transcript set to the number of isoforms per gene in the transcriptome. Both the transcriptome (~80,000 transcripts) and the transcript set used for WGNCA (15,000 transcripts) were primarily composed of only one isoform per gene (Figure S3.4). Our method of selecting most heterogeneously expressed genes resulted in a dataset that did not appear to violate the assumptions made by WGCNA for *P. somniferum* (Figure S3.1). For *P. somniferum* data, weighting the correlation values to a power of nine resulted in a power-law degree distribution as expected in the scale-free topology estimated by WGCNA (R>=0.8). For *P. setigerum*, there was no transformation exponent that resulted in a well-defined power-law degree distribution. Violating the WGCNA assumption of a scale-free topology can result in large modules with less biological relevance (i.e. over clustering) (Langfelder & Horvath 2008). However, the *P. setigerum* module sizes were similar to those estimated for *P. somniferum* suggesting that any over clustering that did occur in the *P. setigerum* analysis was not severe.

WGCNA recovered 19 modules for the *P. somniferum* network and 15 modules for the *P. setigerum network.* These modules varied in size from 21-3310 transcripts and 26-2853 transcripts respectively. The dynamic merge value is one user-provided parameter that is used to define modules. We qualitatively explored multiple dynamic merge values for use here and selected a value (0.2) that limited correlation between modules but that also did not provide so many modules that they became challenging to interpret and concisely discuss. The value used for each network resulted in some modules with correlated eigengenes. This could indicate a need to merge more modules together, but the majority of module eigengene pairwise correlations were less than 0.7 (95% *P. somniferum*, 92% *P. setigerum*) (Figures 3.3 & S3.5).

Comparing and contrasting WGCNA between species

Both transcript expression and connectivity of transcripts in a network can be compared to describe differences between species that are independent of studying individual modules. Connectivity is the sum of adjacencies (i.e. connection strengths) between a transcript and all other transcripts. Transcript expression in *P. somniferum* and *P. setigerum* were correlated when averaged across all libraries from a single tissue as well as averaged across all libraries (rho range from 0.75 to 0.82). However, comparing the total connectivity for each transcript had less correlation between species (rho=0.68) (Figure S3.6).

Identifying the number of modules with significant overlap in transcript content and highly correlated module eigengenes between *P. somniferum* and *P. setigerum* are additional methods of assessing the similarity between the two networks and thus the two species. An eigengene is a mathematic representation of the average expression pattern of the transcripts in a given module. As such each module has an eigengene. If two modules have similar (correlated) eigengenes, then transcripts from one module may have strong co-expression with transcripts from the other. With an increased dynamic merge value, those modules are likely to merge together indicating that they are likely to share biological functions. The eigengenes for 13 of the *P. somniferum* modules are correlated (>0.7) with at least one module eigengene from *P. setigerum* (Figure S3.7). However, of those *P. somniferum* modules, four had more than one correlated (>0.7) module eigengene from *P. setigerum*. Several *P. somniferum* modules also showed statistically significant numbers of shared transcripts with one or two modules from *P. setigerum* (Figure S3.8*)*. Those modules that had little to no overlap in transcript content with any other module were often the smaller modules. Comparing the transcript overlap with the eigengene correlations showed that most modules with highly correlated eigengenes also share large numbers of homologous transcripts. However, there are a few exceptions such as *P. somniferum* module *C* and *P. setigerum* module *h,* which had correlated eigengenes but only one transcript in common. Again, only smaller modules appeared to have a mismatch in strength of eigengene correlation and proportion of shared transcripts (Figure S3.8).

BIA enzymes in the network

Some genes in the morphine and sanguinarine biosynthesis pathways were identified in the transcripts used in the WGCNA (Figures 3.1 & 3.4). However, not all genes in these pathways were identified in the data set and multiple transcripts were annotated as the same gene. In cases where multiple copies of the same gene were identified we arbitrarily assigned them numbers that are consistent throughout. *CheSyn* and *StySyn*, which are both part of the sanguinarine biosynthesis pathway, had similar expression patterns to each other, are most greatly expressed in roots, and had stronger expression in *P. somniferum* than in *P. setigerum.* Transcripts ascribed to genes in the morphine biosynthesis pathway such as *CODM*, *COR*, and *SalAt* also showed greater expression in *P. somniferum* than *P. setigerum*. Finally, *BBE*, which converts (S)-reticuline to (S)-scoulerine but also acts on a variety of other BIAs, had more consistently high expression across tissues and individuals than the other described genes (Figures 3.1 & 3.4).

These genes were sorted into a limited number of modules in both species. *BBE* did not share a module with any other described biosynthesis gene. Every other transcript was in one of two modules in *P. somniferum.* There were also two modules in *P. setigerum* that contained most of these transcripts (Figure 3.1). The four modules that contained the majority of the transcripts in both species (modules O, P, n & o) all had eigengene expression that had noticeably different expression patterns in root samples compared to all other tissues (Figure S3.9).

Metabolites measured from each tissue sample showed patterns consistent with previous studies (e.g. Garnock-Jones & Scholes 1990, Frick et al. 2005). Sanguinarine and protopine are more abundant in root tissues whereas morphinian alkaloids (e.g. codeine, morphine, morphinone) are most abundant in capsules. These patterns hold in both *P. somniferum* and *P. setigerum,* however *P. somniferum* samples had greater abundance of morphinian alkaloids overall (Table S3.2). Predictably, samples clustered based on species and showed some clustering based on tissue when plotting a principle component analysis of the abundances of all measured metabolites (Figure S3.10).

Most alkaloid metabolite abundances did not have strong (nearing 1 or -1) correlations with the module eigengenes for the modules containing their related BIA genes. Sanguinarine and protopine concentrations were both highly correlated with *P. somniferum* module O but not with module P (Figure S3.11). Morphinan alkaloids (e.g. codeine, morphine, codeinone) do not have strong correlations with either of these module eigengenes, but do have strong correlations with other modules that do not contain the relevant BIA genes. Eigengenes for *P. setigerum* modules o and n had weak correlations with all of the metabolites (Figure S3.11).

There were numerous transcripts annotated as transcription factors in the modules that also contained morphine- and sanguinarine-related transcripts. In order to identify those transcription factors most likely to regulate morphine- and sanguinarine-related genes we identified those transcription factors that had strong connectivity to the relevant transcripts (Table S3.3, Figure S3.12). Several transcription factors in the WRKY family including those sequenced from *P. somniferum* and from other species were strongly connected. MYB and MADS-box transcription factor families were also represented, among others. (Table S3.3). Often, both species would have some relevant transcripts strongly co-expressed with each transcription factor, but the relevant transcripts differ between species. Transcription factors that are highly co-expressed with the morphine- and sanguinarine-related transcripts are not always highly co-expressed in both species (Figure S3.12). *P. somniferum* also showed a higher density of transcription factors that were highly co-expressed with at least one morphine- or sanguinarine-related transcript (Figure S3.12 A&B). Furthermore, few of these transcription factors appeared to be strong hub genes in these modules. Hub genes are those genes or transcripts that show high total within module connectivity meaning that they are strongly co-expressed with many genes in the module. They are also often evolutionarily conserved (Masalia et al. 2017).

Cytochrome P450s are one of the largest families of genes in plants and have functions ranging from stress response to development and several cytochrome P450s are involved in the production of various BIAs including CheSyn and StySyn (e.g. Hori et al. 2017, Winzer et al. 2012). The modules that contained morphinian and sanguinarine related genes also contained several transcripts annotated as cytochrome P450s including two that may be part of a 10-gene cluster that produces noscapine one of which has been proposed to encode canadine synthase, another BIA-related enzyme (Winzer et al. 2012) (Table S3.4). Many of the other annotated P450s have been previously undescribed in *Papaver*.

Discussion

While WGCNA is a useful tool to explore co-expression of genes and suggest new lines of inquiry, application to comparing non-model species provides new challenges including identifying orthologous sequences, acquiring significant numbers of libraries, and drawing biologically relevant conclusions with limited gene annotations. Comparing co-expression network analyses between species is easiest in species with strong genomic resources allowing for identification of one-to-one orthologs. Species without abundant genomic resources have had to accept comparisons only at the gene family or orthogroup level. Here we circumvent this issue by utilizing the same transcriptome for both species. Deep description of the expression data from *P. somniferum* and *P. setigerum* revealed new insights into co-expressions of genes from different parts of the large BIA biosynthesis pathway, implicating a family of transcription factors that warrant further exploration for their role in BIA biosynthesis.

Expression of most the morphine- and the sanguinarine-related transcripts show similar patterns to the accumulation of the metabolites. *T6ODM* is a notable exception with higher than anticipated expression in root tissues and strong differences in expression between individuals from the same species. *BBE* is also an exception, however, it is known to play important roles in the production of many other BIAs and, as a result, is expected to be expressed in other tissues. Both *T6ODM* and *BBE* have been implicated in other parts of the larger BIA pathway, which may account for their unique expression patterns. Focused efforts towards understanding the regulation of *T6ODM* and *BBE* may help discover if regulation of morphine or sanguinarine biosynthesis was co-opted from regulation of other BIAs.

WGCNA revealed that, while the two species are similar, they do not show identical co-expression patterns indicating that information learned from *P. somniferum* must be separately investigated in *P. set* before assumed. Modules often circumscribe transcripts that share pathways or regulation (Wolfe et al. 2005). Even large modules with highly correlated eigengenes show variation in transcript content, which suggests that gene regulation may differ between the two species. This is logical given that they vary in alkaloid productivity.

Several transcription factors were strongly co-expressed with morphine- and sanguinarine-related transcripts including many from the WRKY family, which have previously been shown to regulate BIA biosynthesis (Winzer et al. 2012, Beaudoin & Facchini 2014, Agarwal et al. 2016). MYB transcription factors were also represented.

MYB and WRKY families play roles in plant stress responses while other transcription factors with strong co-expression, such as a few MADS-box genes, are most well known for their role in plant development (Airolidi & Davies 2012, Ambawat et al. 2013, Phukan et al. 2016). Members of MYB and WRKY families have been shown to be differentially expressed across *P. somniferum* cultivars that produce varying amounts of BIAs (Agarwal et al. 2016). However, WRKY transcription factors have received more research attention because specific genes have been confirmed to regulate or been implicated in regulation of BIAs (e.g. PsWRKY, CjWRKY1) (Kato et al. 2007, Mishra et al. 2013, Yamada et al. 2017). MYB transcription factors warrant future study for their potential to regulate BIA biosynthesis. Intriguingly, many of these transcription factors were strongly co-expressed with both morphinine- and sanguinarine-related transcripts. However, they did not act as hub genes in the modules. One characteristic of hub genes in angiosperms is that they are evolutionarily conserved (Masalia et al. 2017). It is therefore less surprising that possible BIA regulators are not hub genes because they are likely to be novel and/or rapidly evolving in order to accommodate for the variation in alkaloid biosynthesis. Shared modules between morphine- and sanguinarine-related genes are one line of evidence that suggests regulation of both pathways is overlapping or in some way co-dependent beyond the regulation of their shared precursor (S)-reticuline. However, we see different transcription factors strongly co-expressed with morphine- and sanguinarine-related transcripts from one species to the next with *P. somniferum* having a larger number of relevant transcription factors. This could indicate that BIA production is more integrated with other pathways and processes in *P. somniferum* than in *P. setigerum.*

Transcript expression in root tissues appeared to drive formation of the larger modules with the eigenegenes showing strong root-based expression patterns. Differential expression analyses show that roots show transcript expression patterns that are distinct from all aboveground tissues. Future studies in plants should separate above ground and below ground tissues into individual networks. For future work in *Papaver* and BIA biosynthesis, comparisons with networks from additional BIA-producing species could provide further insight into BIA regulation, highlighting transcription factors that share or differ in their co-expression between species. Comparisons with other species would also indicate the variation of co-expression across *Papaver* species against which the variation between *P. somniferum* and *P. setigerum* can be judged. Finally, WRKY and MYB transcription factors co-expressed with morphine- and sanguinarine-related transcripts should undergo functional testing such as knockdown with viral induced gene silencing (VIGS) to explore their relationship with BIAs.

Comparing expression and co-expression between *P. somniferum* and *P. setigerum* revealed that the species share transcript expression patterns across tissues and have groups of co-expressed genes that highly overlap in transcript content and eigengene expression. Narrowing in on morphine- and sanguinarine-related genes showed that these pathways shared co-expression modules and both pathways had genes highly co-expressed with the same transcription factors. However, co-expression did vary between species such that morphine- and sanguinarine-related genes were not always co-expressed with the same transcription factors from one species to the next. Co-expression network analysis revealed a family of transcription factors that may play a previously unexplored role in regulation BIA biosynthesis in poppies. This descriptive analysis not only identified candidate regulators, but also provided understanding on the relationship of these two species and suggest we exercise caution in assuming shared co-expression and regulation even between closely related species.

References

Airolidi, C. A. & Davies, B. (2012). Gene Duplication and the Evolution of Plant MADS-box Transcription Factors. *Journal of Genetics and Genomics. 39*(4): 157-165.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology, 215*(3), 403-410.

Ambawat, S. Sharm, P. Yadav, N. R., & Yadav, R. C. (2013). MYB transcription factor genes as regulators for plant responses: an overview. *Physiology and Molecular Biology of Plants*, *19*(3): 307-321.

Bauer, W., & Zenk, M. H. (1991). Two methylenedioxy bridge forming cytochrome P-450 dependent enzymes are involved in (S)-stylopine biosynthesis. *Phytochemistry, 30*(9), 2953-2961.

Beaudoin, G. A. W., & Facchini, P. J. (2013). Isolation and characterization of a cDNA encoding (S)-*cis*-N-methylstylopine 14-hydoxylase from opium poppy, a key enzyme in sanguinarine biosynthesis. *Biochemical and biophysical research communications, 431*(3), 597-603.

Beaudoin, G. A. W., & Facchini, P. J. (2014). Benzylisoquinoline alkaloid biosynthesis in opium poppy. *Planta, 240*(1), 19–32.

Bird, D. A., Franceschi, V. R., & Facchini, P. J. (2003). A Tale of Three Cell Types: Alkaloid Biosynthesis Is Localized to Sieve Elements in Opium Poppy. *The Plant Cell*, *15*(11), 2626-2635.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., *et al.* (2008). BLAST+: architecture and applications. *BMC Bioinformatics, 10*(1), 421.

Carolan, J. C., Hook, I. L. I., Chase, M.W., Kadereit, J. W., & Hodkinson, T. R. (2006). Phylogenetics of Papaver and related genera based on DNA sequences from ITS nuclear ribosomal DNA and plastid trnL intron and trnL-F intergenic spacers. *Annals of Botany*, *98*(1), 141–155.

Chen, X., & Facchini, P. J. (2013). Short-chain dehydrogenase/reductase catalyzing the final step of noscapine biosynthesis is localized to laticifers in opium poppy. *The Plant Journal, 77*(2), 173-184.

De-Eknamukul, W. & Zenk, M. H. (1992). Purification and properties of 1,2,-dehydroreticuline reductase from *Papaver somniferum* seedlings. *Phytochemistry, 31*(3), 813-821.

Desgagné-Penix, I., Farrow, S. C., Cram, D., Nowak, J., & Facchini, P. J. (2012). Integration of deep transcript and targeted metabolite profiles for eight cultivars of opium poppy. *Plant Molecular Biology, 79*(3), 295-313.

Díaz Chávez, M. L., Rolf, M., Gesell, A., & Kutchan, T. M. (2011). Characterization of two methylenedioxy bridge-forming cytochrome P450-dependent enzymes of alkaloid formation in the Mexican prickly poppy *Argemone mexicana*. *Archives of biochemistry and biophysics, 507*(1), 186-193.

Dittrich, H. & Kuchan, T. M. (1991). Molecular cloning, expression, and induction of berberine bridge enzyme, an enzyme essential to the formation of benxophenanthridine alklaoids in the response of plants to pathogenic attack. *Proceedings of the National Academy of Sciences, 88*(22), 9969-9973.

Dudoit, S., Yang, Y. H., Callow, M. J., & Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica sinica*, 111-139.

Farrow, S. C., Hagel, J. M., & Facchini, P. J. (2012). Transcript and metabolite profiling in cell cultures of 18 plant species that produce benzylisoquinoline alkaloids. *Phytochemistry, 77,* 79-88.

Fisinger, U., Grobe, N., & Zenk, M. H. (2007). Thebaine synthase: a new enzyme in the morphine pathway in *Papaver somniferum. Natural Product Communications, 2*(3), 249-253.

Frick, S., Kramell, R., Schmidt, J., Fist, A. J., & Kutchan, T. M. (2005). Comparative Qualitative and Quantitative Determination of Alkaloids in Narcotic and Condiment *Papaver somniferum* Cultivars. *Journal of Natural Products, 68*(5), 666-673.

Garnock-Jones, P. J. , & Scholes, P. (1990). Alkaloid content of *Papaver somniferum subsp. setigerum* from New Zealand, *New Zealand Journal of Botany, 28*(3), 367-369.

Gesell, A. Rolf, M., Ziegler, J. Diaz Chavez, M. L, Huang, F. C., & Kutchan, T. M. (2009). CYP719B1 is salutaridine synthase, the C-C phenol-coupling enzyme of morphine biosynthesis in opium poppy. *Journal of Biological Chemistry, 284*(36),24432-24442.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., *et al.* (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644-52.

Grothe, T., Lenz, R., & Kutchan, T. M. (2001). Molecular characterization of the salutaridinol 7-O-acetyltransferase involved in morphine biosynthesis in opium poppy, *Papaver somniferum*. *Journal of Biological Chemistry, 276*(33),30717-30723.
Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., *et al*. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494-512.

Hagel, J. M., Beaudoin, G. A., Fossati, E., Ekins A., Martin, V. J., & Facchini, P. J. (2012). Characerization of a flavoprotein oxidase from opium poppy catalyzing the final steps in sanguinarine and papaverine biosynthesis. *Journal of Biological Chemistry, 287*(51), 42972-42983.

Hagel, J. M., & Facchini, P. J. (2010). Dioxygenases catalyze the O-demethylation steps of morphine biosynthesis in opium poppy. *Nature Chemical Biology, 6*(4), 273-275.

Hagel, J. M., & Facchini, P. J. (2013). Benzylisoquinoline alkaloid metabolism: A century of discovery and a brave new world. *Plant and Cell Physiology, 54*(5), 647–672.

Hartmann, T. (2007). From waste products to ecochemicals: Fifty years research of plant secondary metabolism. *Phytochemistry, 68*, 2831-2846.

Hirata, K. Poeaknapo, C. Shmidt, J., & Zenk, M. H. (2004). 1,2-Dehydroreticuline synthase, the branch point enzyme opening the morphinan biosynthetic pathway. *Phytochemistry, 65*(8), 1039-1046.

Hori, K., Yamada Y., Ratmoyo, P., Minakuchi, Y., Toyoda A., Hirakawa, H., & Sato, F. (2017). Mining of the uncharacterized cytochrome P450 genes involved in alkaloid biosynthesis in California poppy using a draft genome sequence. *Plant and Cell Physiology, 59*(2), 222-233.

Hosokawa, K., Shibata, T., Nakamura, I., & Hishida, A. (2004). Discrimination among species of *Papaver* based on the plastid *rpl16* gene and the *rpl16-rpl14* spacer sequence. *Forensic Science International, 139*(2-3), 195-199.

Ikezawa, N. Iwasa, K. & Sato, F. (2007). Molecular cloning and characterization of methlenedioxy bridge-forming enzymes involved in stylopine biosynthesis in *Eschscholzia californica*. *The FEBS journal, 274*(4),1019-1035.

Kato, N., Dubouzet, E., Kokabu, Y., Yoshida S., Taniguchi Y., Dubouzet, J. G., *et al.* (2007). Identification of a WRKY protein as a transcriptional regulator of benzylisoquinoline alkaloid biosynthesis in *Coptis japonica*. *Plant and cell physiology, 48*(1), 8-18.

Kerry, J. W. (2013). System-based transcriptomic and metabolomic network analyses in *Papaver somniferum*. Unpublished master's thesis, University of Georgia, Athens, Georgia.

La Valva, V., Sabato, S., & Gigliano, G. S. (1985). Morphology and Alkaloid Chemistry of *Papaver setigerum DC*. (Papaveraceae). *Taxon, 34*(2), 191-196.

Langfelder, P., & Hovarth, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559.

Langfelder, P., Luo, R., Oldham, M. C., & Hovarth, S. (2011). Is my network module preserved and reproducible? *PloS Computational Biology*, *7*(1), e1001057.

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics, 24*(5), 719-720.

Lenz, R., & Zenk, M. H. (1994). Closure of the oxide bridge in morphine biosynthesis. *Tetrahedron Letters, 35*(23), 3897-3900.

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323.

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2009). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, *26*(4), 493–500.

Liscombe, D. K. & Facchini, P. J. (2007). Molecular cloning and characterization of tetrahydroprotoberberine cis-N-methyltransferase, an enzyme involved in alkaloid biosynthesis in opium poppy. *Journal of Biological Chemistry, 282*(20), 14741-14751.

Liscombe, D. K., MacLeod, B. P., Loukanina, N., Nandi, O. I., & Facchini, P.J. (2005). Erratum to "Evidence for the monophyletic evolution of benzylisoquinoline alkaloid biosynthesis in angiosperms. " [Phytochemistry 66 (2005) 1374-1393]. *Phytochemistry, 66*(20), 2500-2520.

Ma, C. & Wang, X. (2012). Application of the Gini Correlation Coefficient to Infer Regulatory Relationships in Transcriptome Analysis. *Plant physiology, 160*(1), 192-203.

Malik, C. P., Mary, T. N., & Grover, I. S. (1979). Cytogenetic Studies in *Papaver* V. Cytogenetic studies on *P. somniferum* X *P. setigerum* hybrids and amphiploids. *Cytologia*, *44*(1), 59-69.

Masalia, R. R., Bewick, A. J., & Burke, J. M. (2017). Connectivity in gene coexpression networks negatively correlates with rates of molecular evolution in flowering plants. *PloS one*, *12*(7), e0182289.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., et al. (2006). ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics, 7*(Suppl 1), S7.

Matasci, N., Hung, L. H., Yan, Z., Carpenter, E. J., Wickett, N. J., Mirarab, S., *et al.* (2014). Data access for the 1,000 Plants (1KP) project. *GigaScience*, *3*(1), 17.

McCarthy, J. D, Chen, Y., & Smyth, K. G. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research, 40*(10), 4288-4297.

Mishra, S., Triptahi, V., Singh, S., Phukan, U. J., Gupta, M. M., Shanker, K. *et al.* (2013). Wound Induced Transcriptional Regulation of Benzylisoquinoline Pathway and Characterization of Wound Inducible PsWRKY Transcription Factor from *Papaver somniferum. PLOS One, 8*(1), e52784.

Morishige, T., Tsujita, T., Yamada, Y. & Sato, F. (2000). Molecular characterization of the S-adenosyl-L-methionine:3'-hydroxy-N-methylcoclaurine 4'-O-methyltransferase involved in isoquinoline alkaloid biosynthesis in *Coptis japonica*. *Journal of Biological Chemistry, 275*(30), 23398-23405.

Ober, D. (2005). Seeing double: gene duplication and diversification in plant secondary metabolism. *Trends in Plant Science, 10*(9), 444-449.

Oldham, M. C., Hovarth, S., & Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences, 103*(47), 17973-17978.

Onoyovwe, A., Hagel, J. M., Chen, X., Khan, M. F., Schriemer, D. C., & Facchini, P. J. (2013). Morphine Biosynthesis in Opium Poppy Involves Two Cell Types: Sieve Elements and Laticifers. *The Plant Cell*, *25*(10), 4110–4122.

Phukan, U. J., Jeena, G. S., & Shukla, R. K. (2016). WRKY transcription factors: molecular regulation and stress responses in plants. *Frontiers in plant science, 7*, 760.

Pichersky, E., & Gang, D. R. (2000). Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends in Plant Sciences, 5*,(10), 1360-1385.

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140.

Robinson, M. D. & Oshlack, A. (2010). A scaling normalization method for differential expression of RNA-seq data. *Genome Biology, 11*(3), R25.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research, 13*, 2498-2504.

Sharghi, N., & Lalezari, I. (1967). Papaver bracteatum Lindl., a Highly Rich Source of Thebaine. *Nature, 213*(5082), 1244.

Smith-Unna, R., Boursnell, C., Patro, R., Habbard, J., & Kelly, S. (2016). TransRate: reference free quality assessment of *de novo* transcriptome assemblies. *Genome Research*, *26*(8), 1134–1144.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., & Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics, 27*(3), 431-432.

Takemura, T., Ikezawa, N., Iwasa, K., & Sato, F. (2012). Molecular cloning and characterization of a cytochrome P450 in sanguinarine biosynthesis form *Eschscholzia californica* cells. *Phytochemistry, 91*, 100-108.

Tanahashi, T., & Zenk, M. H. (1988). One step enzymatic synthsis of dihydrosanguinarine from protopine. *Tetrahedron letters, 29*(44), 5625-5628.
Theis, N., & Lerdau, M. (2003). The evolution of function in plant secondary metabolites. *International Journal of Plant Sciences, 164*(3 Suppl.), S93-S102.

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W, Tanimoto, M. *et al.* (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, Cell & Environment, 32*(12), 1633-1651.

Vogel, M., Lawson, M., Sippl, W., Conrad, U. & Roos, W. (2010). Structure and mechanism of sanguinarine reductase, an enzyme of alkaloid detoxification. *Journal of Biological Chemistry, 285*(24), 18397-18406.

Vom Endt, D., Kijne, J. W., Memelink, J. (2002). Transcription factors controlling plant secondary metabolism: what regulates the regulators? *Phytochemistry, 61*(2), 107-114.

Wagner, G. P. , Kin, K., & Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-se data: RPKM measure is inconsistent among samples. *Theory in Bioscience, 131*(4): 281-285.

Winzer, T., Gazda, V., He, Z., Kaminski, F., Kern, M., Larson, T. R. *et al.* (2012). A *Papaver somniferum* 10-Gene Cluster for Synthesis of the Anticancer Alkaloid Noscapine. *Science*, *336*(6089): 1704-1708.

Wolfe, C. J., Kohane, I. S., & Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics, 6*(1), 227.

Yamada, Y., Shimada, T., Motomura, Y., & Sato, F. (2017). Modulation of benzylisoquinoline alkaloid biosynthesis by heterologous expression of CjWRKY1 in *Eschscholzia californica* cells. *PLOS One, 12*(10), e0186953.

Yonekura-Sakakibara, K., Tohge, T., Matsuda, F., Nakabayashi, R., Takayama, H., Niida, R., et al. (2008). Comprehensive Flavonol Profiling and Transcriptome Coexpression Analysis Leading to Decoding Gene-Metabolite Correlations in *Arabidopsis*. *The Plant Cell, 20*(8), 2160-2176.

Zaharia, M., Bolosky, W. J., Curtis, K., Fox, A., Patterson, D., Shenker, S. *et al*. (2011). Faster and More Accurate Sequence Alignment with SNAP. arXiv:1111.5572v1,

Zhao, W., Langfelder, P., Fuller, T., Dong, J., Li, A., & Hovarth, S. (2010) Weighted Gene Coexpression Network Analysis: Stat of the Art. *Journal of Biopharmaceutical Statistics, 20*(2), 281-300.

Ziegler, J., Diaz-Chávez, M. L., Kramell, R., Ammer, C., & Kutchan, K. M. (2005). Comparative macroarray analysis of morphine containing *Papaver somniferum* and eight morphine free *Papaver* species identifies an *O*-methyltransferase involved in benzylisoquinoline biosynthesis. *Planta, 222*(3), 458-471.
Ziegler, J., & Facchini, P. J. (2008). Alkaloid biosynthesis: metabolism and trafficking. *Annual Review of Plant Biology, 59*, 735-769.

Ziegler, J., Voigtländer, S., Schmidt, J. Kramell, R. Miersch, O., Ammer, C. *et al.* (2006). Comparative transcript and alkaloid profiling in *Papaver* species identifies a short chain dehydrogenase/reductase involved in morphine biosynthesis. *The Plant Journal, 48*(2), 177-192.

Zulak, K. G., Cornish, A., Daskalchuk, T. E., Deyholos, M. K., Goodenowe, D. B., Gordon, P. M., et al. (2007). Gene transcript and metabolite profiling of elicitor-induced opium poppy cell cultures reveals the coordinate regulation of primary and secondary metabolism. *Planta, 225*(5), 1085-1106.

Tables

Table 3.1: TransRate assembly assessment data.

Species	**Total # Transcripts**	**Total # Read Pairs**	**% Mapped Read Pairs**	**# of Transcripts with ORF**	**N50**	**Transrate Assembly Score**
P. somniferum	84,368	~279 million	69%	60,813	1987	0.257
P. setigerum	-	~183 million	78%	_-	-	0.214

Dashes indicate values that are dependent only on the transcriptome and therefore are shared by both species.

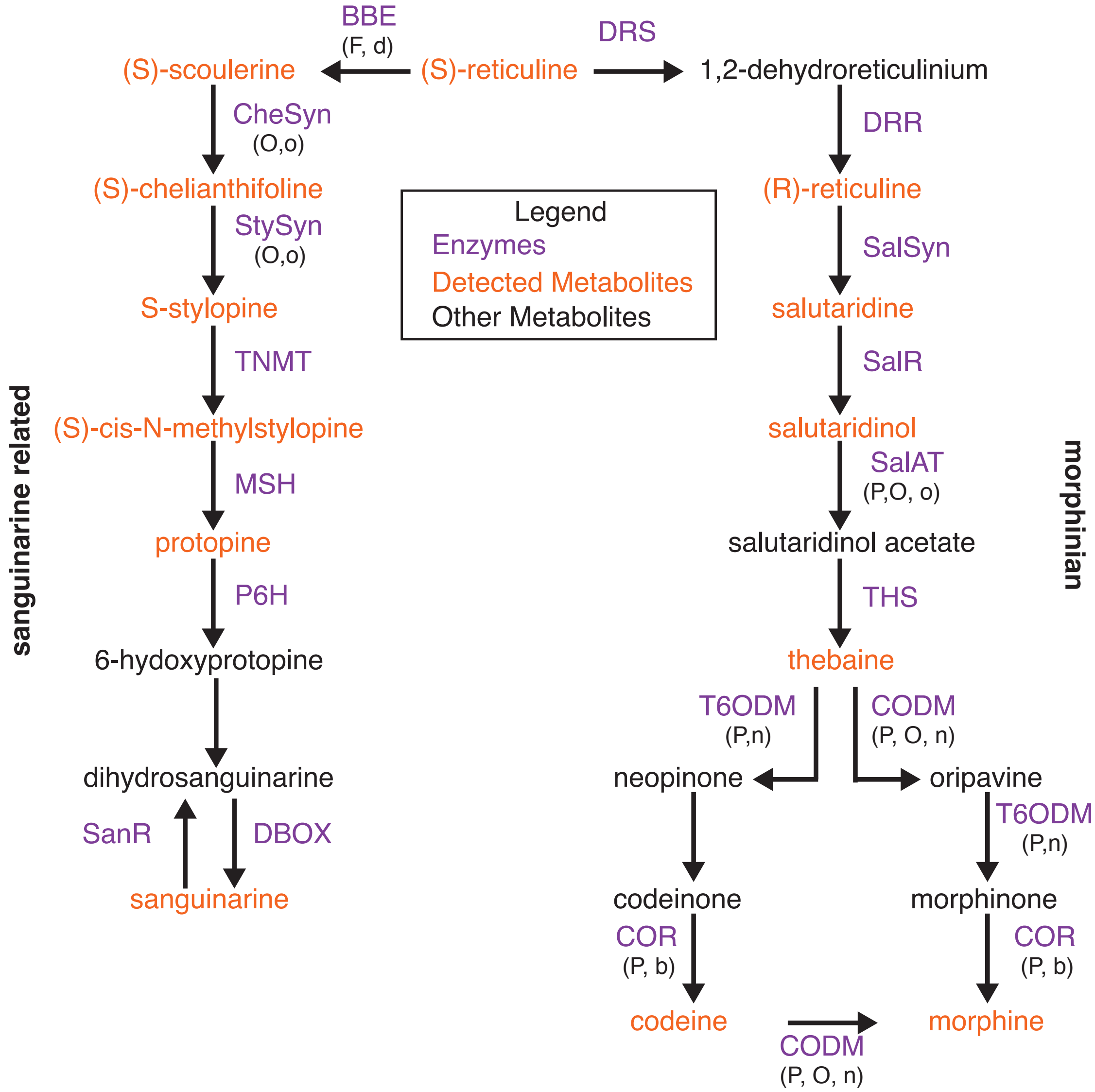


Figure 3.1: Pathway for morphine and sanguinarine biosynthesis. The pathways included here show morphine and sanguinarine biosynthesis beginning with their shared precursor (S)-reticuline, which is also acted upon by enzymes beyond those depicted here. Enzymes are shown in purple. Modules are shown in black below the enzymes. Capital letters indicate *P. somniferum* modules and lower case indicates *P. setigerum* modules. Metabolites that were detected in at least one species at any amount are in orange. Metabolites that were either not measured or not detected are in black. Labels to the left and the right indicate the descriptor for that part of the pathway. (See Kerry 2013).

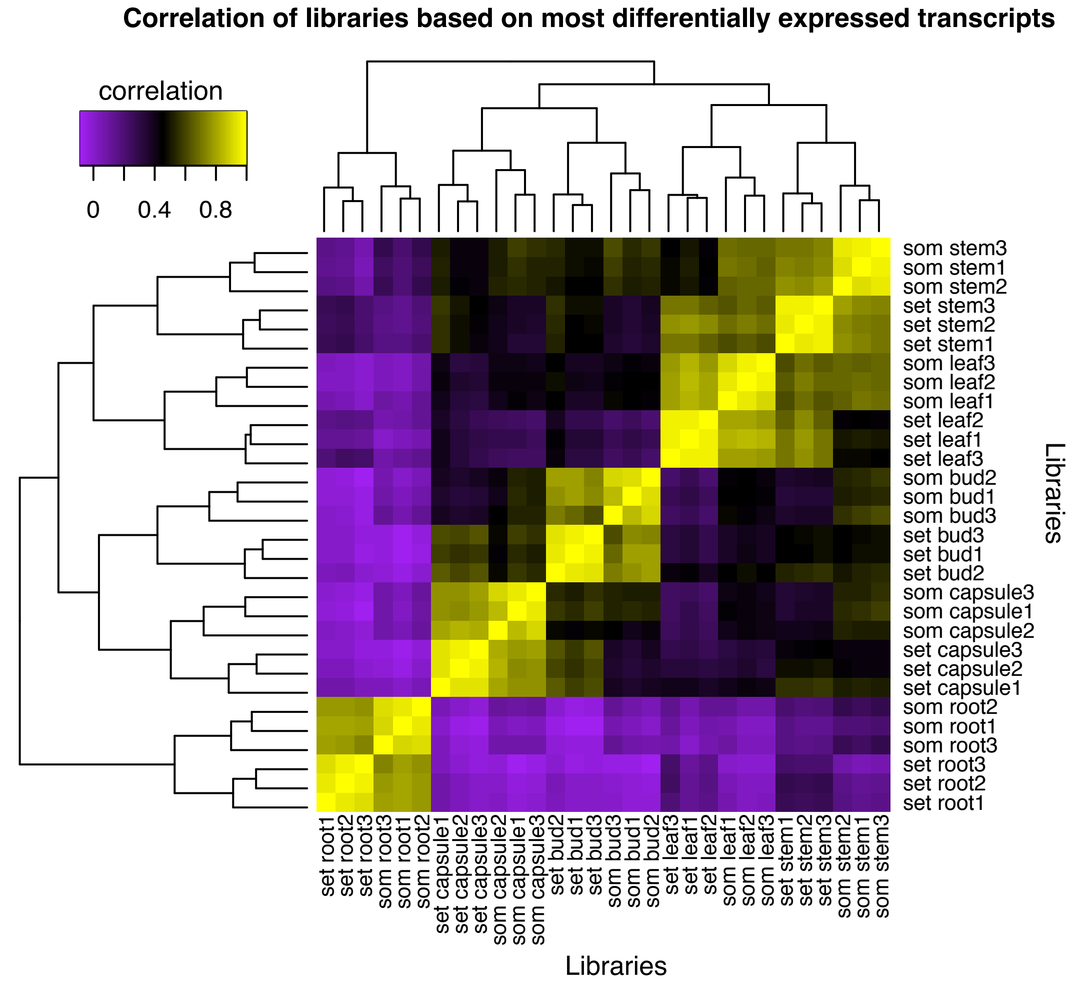


Figure 3.2: Correlation of libraries based on most differentially expressed transcripts. Dendrograms show hierarchical clustering of libraries. Colors indicate the absolute Pearson correlation coefficient between libraries, which is based expression of all transcripts that are at least 4-fold differentially expressed between at least one pair of libraries with a p-value of 1e-03 or lower. "Som" denotes *P. somniferum* RNA-Seq library and "set" denotes *P. setigerum*. Numbers in library names denote individual.

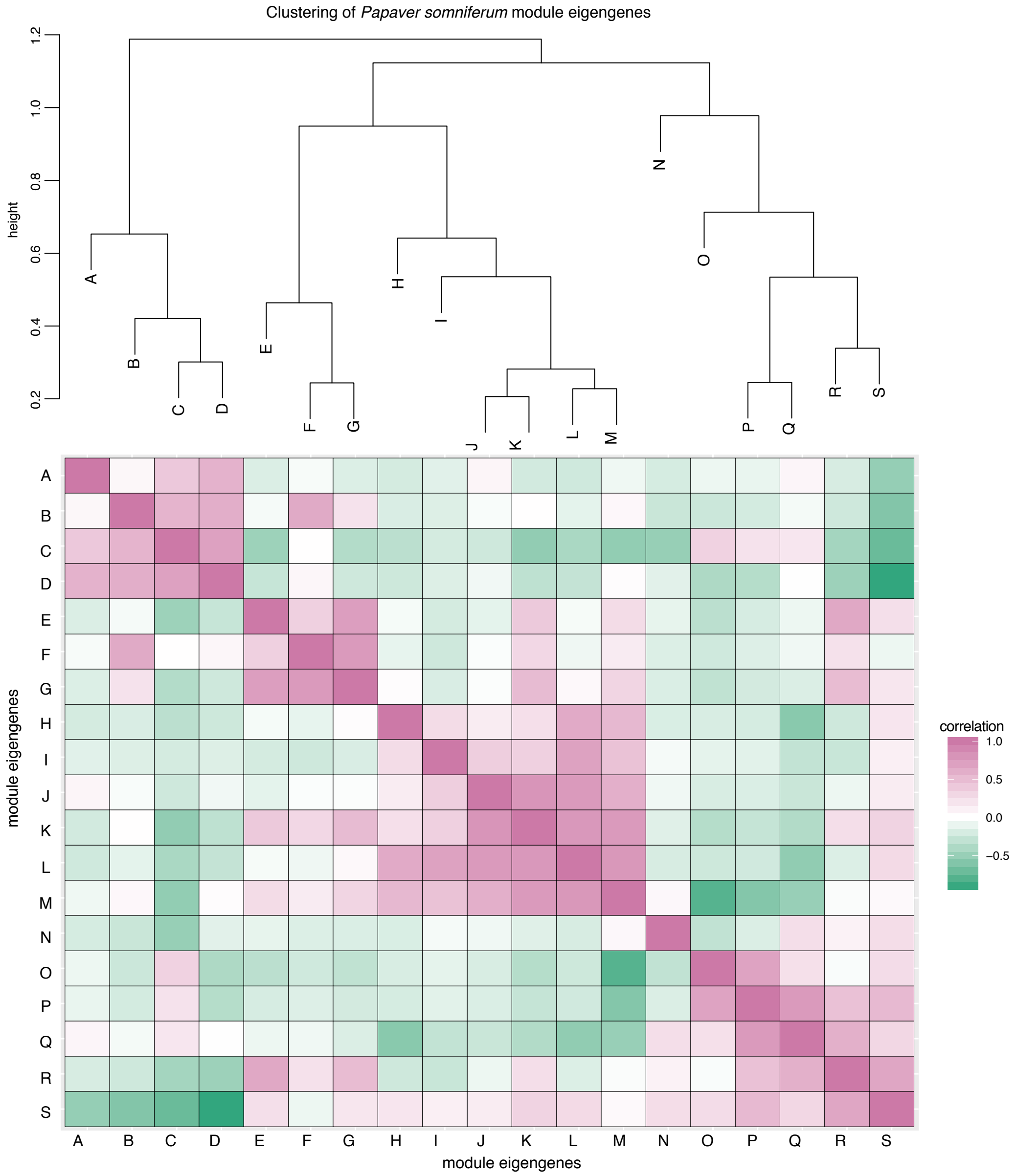


Figure 3.3: Clustering of *P. somniferum* module eigengenes by Pearson correlation. The dendrogram shows the hierarchical clustering of module eigengenes. The heatmap is colored by Pearson correlation coefficient of the module eigengenes comparing all modules from the *P. somniferum* WGCNA with each other.

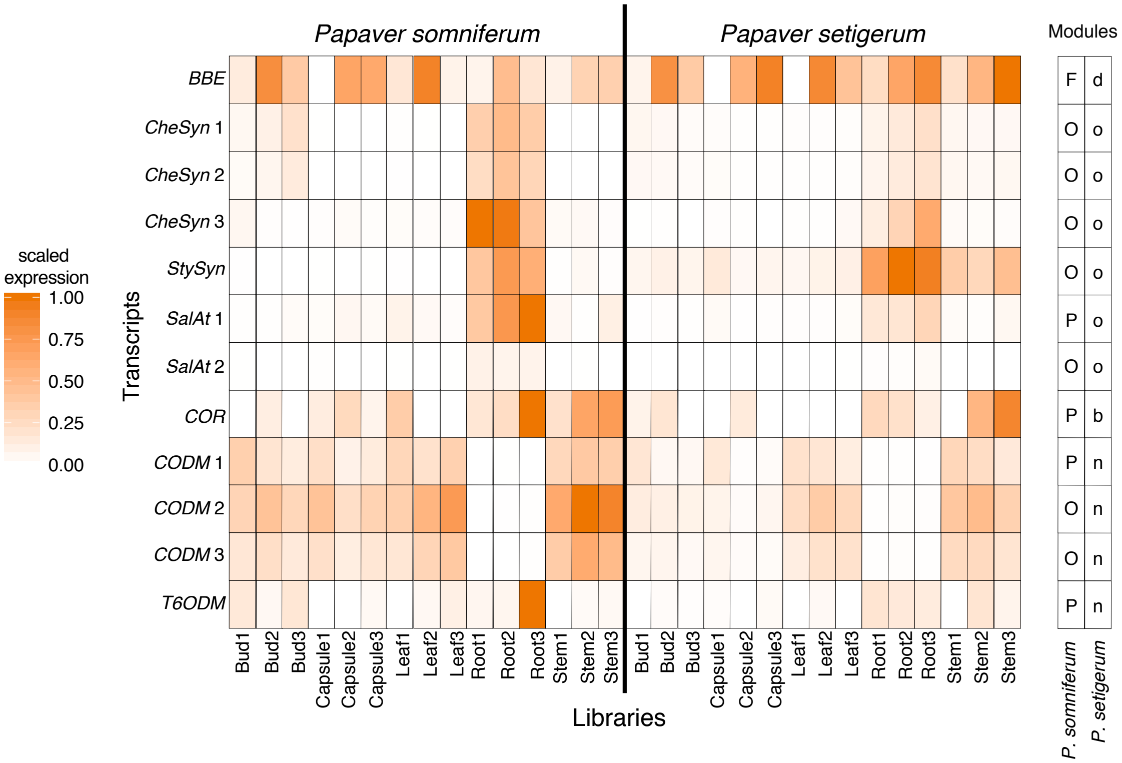


Figure 3.4: Transcript expression and module assignment for BIA-related enzymes. Heatmap shows transcript expression in TPM across all libraries. Expression is scaled within a gene such that the expression values are divided by the maximum expression of any transcript annotated as the same gene. The x-axis shows the blastn annotation of the transcripts where the numbers are assigned to denote different transcripts annotated as the same gene. The y-axis denotes library. Library names include tissue type and numbers corresponding to the individual. Letters on the right indicate the module from each species that includes the transcript.

CHAPTER 4

A MODEL OF THE FACTORS INFLUENCING TEACHING IDENTITY AMONG LIFE SCIENCES DOCTORAL STUDENTS[1]

[1] Lane, A. K., Hardison, C., Simon, A., & Andrews, T. C. A model of the factors influencing teaching identity among life sciences doctoral students. Accepted by Journal of Research in Science Teaching. Reprinted here with permission of the publisher, 04/11/2018.

Abstract

One barrier to the adoption of evidenced-based teaching practices may be that faculty do not see teaching as an important part of their identities as professionals. Graduate school is a key time for professional identity development, and currently we know little about how doctoral students develop identities as college teachers. In this qualitative study we aim to characterize the factors that promote and hinder teaching identity among 33 life sciences doctoral students with diverse career interests at one research university. We collected data using semi-structured interviews and analyzed it using qualitative content analysis. Our analysis involved iteratively and collaboratively analyzing interview transcripts while considering existing literature about socialization and professional identity and remaining open to novel ideas in the data. From this analysis we developed a mechanistic model of the factors that influenced teaching identity in our participants. Independent teaching experiences, teaching professional development, and teaching mentors contributed to salient and stable teaching identities among doctoral students. Being recognized by faculty as a teacher was also important, but rare. The professional culture of life sciences acted like a blizzard that doctoral students had to navigate through to develop a teaching identity. This culture strongly valued research over teaching, resulting in a sometimes cold and isolating environment for students interested in teaching. The culture also made it harder to see existing opportunities for teaching development and made it more challenging to move toward these opportunities, much like the deep snow and driving winds of a blizzard. The mechanistic model described in this work is an important first step in understanding how

doctoral training influences teaching identity. This model serves as a hypothesis that should be tested and refined through additional empirical work across contexts.

Introduction

Undergraduate science instructors have been slow to adopt evidence-based instructional practices (National Research Council, 2012). Among instructors who make attempts to adopt these practices, many abandon their efforts or implement the practices ineffectively (e.g., Andrews et al. 2011, Henderson et al. 2012). Thus, researchers are carefully considering what is necessary to support undergraduate instructors in changing their teaching (e.g., Ebert-May et al. 2011, Henderson et al. 2011). Instructors often describe a lack of time, training, and incentives as barriers to using evidence-based practices (e.g., Dancy & Henderson 2010, Andrews & Lemons 2015, Shadle, Marker, & Earl 2017). However, Brownell and Tanner (2012) hypothesize that addressing these barriers is insufficient to foster meaningful change within a professional culture of science that promotes the development of research identities but not teaching identities. Graduate school is a key time for professional identity development (Austin 2002), but currently we know little about how doctoral students develop professional identities as college teachers. As a first step in addressing this gap, this study aims to better understand what can promote and hinder a professional identity as college teacher among life sciences doctoral students. This study was informed by theory and empirical research related to socialization, identity theory, and professional identity development.

Guiding Theoretical Frameworks

Socialization is an ongoing process of becoming a member of a community of practice (e.g., Lave & Wenger 1991), and graduate students are simultaneously being socialized in multiple communities because they are joining a community of graduate students, a department, a discipline, and more (Austin & McDaniels 2006). In a department, faculty serve as full members of a community of practice and graduate students are newcomers who learn about and make sense of the community by participating in it (Lave & Wenger 1991, Wegner & Nückles 2015). Newcomers to a community learn what is expected and what is needed to succeed in the community (Turner & Thompson 1993). This involves learning about and coming to adopt the culture of the community, including the practices, norms, values, and discourse of the community (Sfard 1998, Bieber & Worley 2006, Wegner & Nückles 2015).

While the culture of a community has great influence, individuals also have agency to push against this influence (Bourdieu 1977, Barker 2005). Graduate students may develop values and engage in behaviors that are not highly valued in their department or discipline (e.g., Thiry et al. 2007). However, individuals who demonstrate characteristics of more valued identities may receive more resources from the community (Hall & Burns 2009). For example, Thiry et al. (2007) propose that graduate students in the sciences anticipated that disapproval from their advisor about participating in science outreach activities might result in loss of research support, delays in dissertation approval, and less support for job and fellowship applications.

The culture of a community of practice, such as an academic department, is communicated formally through degree expectations, and also tacitly as students observe and participate in routine interactions in the department (Turner et al. 2002). Doctoral students figure out what it means to be a faculty member in the discipline through observations and experiences with faculty before and during graduate school, but often fail to develop a full understanding of what faculty work involves (e.g. Austin 2002, Bieber & Worley 2006). Doctoral students report receiving mixed messages about faculty responsibilities and priorities, such as rhetoric from upper administration about the importance of undergraduate teaching contrasted with their personal observations of the work for which faculty are rewarded (Austin 2002). Graduate students continuously work to make sense of the culture of the communities in which they participate and to compare the culture to their own values and interests (Austin 2002, Bieber & Worley 2006). By participating in communities of practice, students come to construct and enact their own professional identities (Weidman et al. 2001, Szelényi et al. 2016).

Professional identity includes how a person defines themselves professionally, including workplace values, roles, and responsibilities, and how the person is seen by those around them (Hall & Burns 2009). An individual's professional identity is multifaceted and is one of many identities an individual uses to make sense of themselves (e.g. Coldron & Smith 1999, Beijaard et al. 2004, Stets & Serpe 2013). For example a faculty member may have professional identities as an undergraduate teacher, a mentor, and a geneticist, as well as identities as a Latina, a mother, and a political liberal (Stets & Serpe 2013). Some identities are more salient than others. A salient identity is consistently invoked across situations and is more likely to be invoked in any given situation (Stets & Serpe 2013). Professional identities can be tenuous in that different identities may be invoked depending on the situation, but they can also become stable over time if they are repeatedly and habitually invoked across contexts (Carlone & Johnson 2007).

Constructing a professional identity is an ongoing process where an individual observes and interprets experiences in order to make sense of oneself (Coldron & Smith 1999). Work from business management has examined how people transition into new professional roles. Newcomers observe role models to see what makes them successful and to determine in what ways they see themselves as similar and dissimilar from role models (e.g. Ibarra 1999, Ronfeldt & Grossman 2008). They also experiment with professional identities by engaging in authentic professional activities, then assessing and modifying their enactment of the identity. For example, graduate school can offer the opportunity to take on roles and responsibilities of a researcher, an undergraduate instructor, and a research mentor. A graduate student evaluates themselves in these roles and pays attention to external assessments. By trying out an identity, a newcomer can consider the degree to which an identity aligns with how they see themselves and how others see them. They then make decisions about what parts or whole professional identities to retain and what to discard (Ibarra 1999).

Evaluating a provisional identity hinges on implicit and explicit feedback from individuals who are firmly established in the profession. These meaningful others indicate how well a provisional identity is aligned with the values and expectations of the profession, and thus whether the student has potential in the profession (Ibarra 1999, Carlone & Johnson 2007). Positive recognition from meaningful others can be key to persisting in seeking out opportunities to enact an identity. When meaningful others in a community do not provide recognition, identity development may be disrupted or stagnated, ultimately influencing career intentions (Carlone & Johnson 2007). Female doctoral students in science and engineering who were interested in becoming community educators or entrepreneurs felt that their doctoral programs and advisors did not support or value for these pursuits, and this led some participants to suppress these identities (Szelényi et al. 2016).

Teaching Identity Among STEM Graduate Students

Brownell & Tanner (2012) hypothesized that the culture of science gets in the way of developing a professional identity as a college teacher and that faculty without this identity may be less willing to engage in instructional change. They proposed three tensions between a professional identity as a scientist and the adoption of evidence-based teaching practices. The first tension is that training as a scientist promotes the development of a professional identity as a researcher, but not as a college teacher (Brownell & Tanner 2012). Despite the fact that about half of students in STEM start and end doctoral training interested in teaching undergraduates (Connolly et al. 2016), many programs provide little or no training in pedagogy and evidence-based teaching to graduate teaching assistants, or offer only voluntary training in these areas (e.g., Schussler et al. 2015). The second tension is that scientists, including doctoral students, may be afraid to "come out" as teachers because they fear they will not be taken seriously by the larger scientific community (Connolly 2010, Brownell & Tanner 2012). The third tension is that the professional culture of science considers teaching to be lower status than research, positioning scientists to feel they have to choose between the two (Brownell & Tanner 2012). Current and future faculty may feel that to be seen as "real" scientists they need to shy away from spending time on their teaching and teaching development (e.g. Thiry et al. 2007, Brownell & Tanner 2012).

Though Brownell and Tanner's (2012) essay is one of the most commonly-cited papers published in *CBE-Life Sciences Education*, indicating that their ideas resonate with the community, few studies have empirically examined professional identities among future and current science faculty (Kendall et al. 2013). One study examined the professional identities of 24 graduate students from groups underrepresented in science and engineering who participated in a K12 education outreach program (Thiry et al. 2007). These students perceived that their involvement in science outreach was in opposition to being considered a "real scientist" by peers and faculty. Their advisors encouraged them to minimize time spent away from research and they experienced negative reactions from other faculty and peers about their involvement in the science outreach program. These experiences led them to feel like they were outsiders within their department and discipline (Thiry et al. 2007). The outreach program provided an alternative community of like-minded individuals whom they saw as sharing their values and interests. Even though many of these students were personally interested in outreach and teaching, they "found it difficult to let go of academic research as a career goal" because they recognized "the lower status granted to teaching as compared to research" (Thiry et al. 2007 pg 407). They also worried that the science research community might be closed to them as their identities as science educators became more salient.

Other research has focused on teaching development and teaching experiences for graduate students, which may be related to teaching identity. A longitudinal study of STEM doctoral students at three research universities revealed that most (84.6%) students participated in at least some teaching professional development. These experiences improved their feelings of competence as a teacher and their sense of community with their peers (Connolly et al. 2016). Furthermore, participation in teaching development programs did not negatively affect time to degree completion. Teaching experiences during doctoral training helped students explore teaching as a career option, and resulted in less bias against faculty teaching jobs, confirmed interest in faculty teaching jobs, or students deciding that they did not want positions that included teaching. Despite these benefits, doctoral students were often discouraged from spending any more than the minimum required time on teaching and teaching development (Connolly et al. 2016). Thus, the experiences of these STEM graduate students seem to align with much of what Brownell & Tanner (2012) hypothesized. Further research is needed to understand the mechanisms of how doctoral training influences teaching identity (Kendall et al. 2013).

Our objective in this research was to undertake a qualitative investigation of the experiences and perceptions of life sciences doctoral students who had diverse career interests to better understand the factors influencing professional identity as a college teacher. Specifically, we aimed to elucidate what promotes and hinders professional identity as a college teacher among doctoral students in the life sciences, and to develop a mechanistic model of how doctoral training influences identity as a college teacher. The model produced by this work is grounded in the experiences of our participants, and can serve as a hypothesis to be tested and refined through additional qualitative investigations. It also lays the groundwork for larger-scale quantitative investigations.

Methods

Context

Investigating the role of socialization necessitates clearly describing the context in which this research was situated. We investigated the experiences of participants in several life sciences departments at a large university classified by Carnegie as "highest research activity." Participants' major advisor acted as a research advisor, dissertation chair, academic advisor, and often a key source of stipend funding. All departments guaranteed the same stipend funding to all students for at least five years, either through research or teaching assistantships, and students received full tuition waivers. Most students receiving research assistantships were paid for research that would be a part of their dissertation. Doctoral students in these departments were expected to work at least the equivalent of a 40-hour work week, including time spent in classes, conducting research, and working as a graduate teaching assistant.

The nature of research in the life sciences and how it is funded influences doctoral training. Most research projects are collaborative. Doctoral students typically work on research ideas developed by their advisors, and research is funded to be completed by a particular date. They generally conduct research within a laboratory, and may have a designated "bench" space and desk space. The term "lab" is commonly used to refer to the collective space that includes laboratory benches, research equipment, and student desks. It is also used to refer to the hierarchically-organized group of individuals overseen by the research advisor, including laboratory technicians, postdoctoral researchers, graduate students, and undergraduate researchers, all of whom work in close physical proximity.

The university where this research was situated has both discipline-based education researchers employed as tenure-track faculty in science departments and a Center for Teaching and Learning that may support teaching endeavors of graduate students. Most life sciences departments have at least one tenure-track faculty member who is a discipline-based education researcher and these individuals are seen as opinion leaders for undergraduate education by faculty (Andrews et al. 2016). The Center for

Teaching and Learning at this institution employs six or more full-time staff to support teaching development and recognition.

The experiences and perspectives of the lead researchers are also relevant to interpreting our results. AKL and TCA led all data collection and analysis. Both AKL and TCA have been graduate students in life sciences departments and TCA is faculty in a life sciences department. Both were members of life sciences departments within the focus university.

Participants and Recruitment

We interviewed a wide range of graduate students, including students from multiple departments, with diverse career interests, and at different stages of graduate training. We aimed to maximize the variation in our sample so that we could identify patterns across students, while also revealing the unique experiences of different individuals (Patton 1990). We started by recruiting participants from one life sciences department to control for effects of departmental culture and to allow us to recruit multiple students with the same major advisor. However, initial analysis revealed that only a few of these participants identified as college teachers. Therefore, we recruited participants from additional departments, including some with a demonstrated interest in teaching and some who were not interested in teaching. We recruited all participants by email, including up to three reminder emails and also gave a brief presentation at a student association meeting. We asked potential participants to complete a short survey asking about their career interests, years in school, and semesters as a teaching assistant. We offered a $25 gift card as incentive for interview participation.

Our final sample included 33 doctoral students earning degrees in four life sciences departments at one university. Over half of the students (n = 18) were from the first department in which we sampled. These students represented about 39% of the graduate student population in that department and 59% of the research groups with graduate students. Additional sampling yielded 15 students working toward doctoral degrees across three other departments.

Data Collection

We conducted semi-structured interviews that lasted 30 to 90 minutes. AKL completed and audio recorded all interviews. We used one interviewer to ensure consistency across data and because we anticipated that students would be more forthcoming in their answers with a peer.

We developed an interview protocol based on prior research on professional identity (e.g., Austin 2002, Bieber & Worley 2006, Pratt, Rockmann, & Kaufmann 2006) and refined it iteratively using pilot interviews. Pilot interviews are not included in the final data set. We designed our final protocol to learn about participants' thinking and experiences related to (a) their career plans; (b) teaching, development as a teacher, and training for teaching; (c) whether they saw themselves as researchers and teachers and why; (d) their advisors' views on research and teaching; (e) their opinions about tensions proposed by Brownell & Tanner (2012). Based on the interviews conducted in the first round of data collection, we added three questions to the protocol for participants who expressed interest in college teaching. These questions probed more directly about how participants felt others viewed their interest in teaching. Interviews were transcribed verbatim and checked for accuracy. All interview questions are available in Appendix C.

Qualitative Data Analysis

Our data analysis aimed to identify and describe factors that promote or hinder professional identities as college teachers, and to organize these factors into a theorized mechanistic model depicting relationships among factors. We used qualitative content analysis organized within Atlas.ti to accomplish this. All four authors contributed to the data analysis.

Identifying and describing all potentially relevant ideas.

We systematically identified and described participants' thinking and experiences relevant to our research questions. Each researcher listened to the same subset of interviews while considering guiding questions about the participant's career goals and teaching identity. The full research team then met and discussed these questions. This process produced an initial list of ideas and experiences related to teaching identity. Next we formally analyzed a single transcript by dividing it into sections of text that communicated a discrete idea and labeling these sections with codes to summarize their content. Codes consisted of short phrases that described the content of the text and a summary of the ideas captured by the code, including nuance across participants. Codes emerged from our analysis, rather than from an *a priori* list (Charmaz 2006, Saldaña 2013). We applied codes to any section of interview transcript that contained the corresponding idea. These sections are quotes and ranged in length from sentences to paragraphs. We continued with the remaining interviews, adding, omitting, combining, and splitting codes to best capture the thinking and experiences of our participants. As codes changed, we re-analyzed data we had previously coded (Saldaña 2013). We completed each part of this process in pairs or small groups to allow for constant comparison of our interpretations of the data.

During this phase, we also read all quotes within each code numerous times. This allowed us to refine codes so that all quotes within code addressed the same core idea. We simultaneously worked on groups of related codes in order to compare and contrast codes (Saldaña 2013). When this step was completed, we had catalogued and extensively described the thinking and experiences recounted by our participants. Many themes in the data were evident at this point, but subsequent analysis helped reassemble the data to understand connections among ideas.

Elucidating relationships to develop a mechanistic model.

The second phase of our analysis aimed to identify emergent themes and describe relationships between themes. This was done iteratively and in concert with revisiting the literature to view our results through multiple lenses. We categorized participants based on their identities as teachers in order to make systematic comparisons among them to better understand the experiences associated with a teaching identity. Developing identity categories was iterative and involved making comparisons among participants to elucidate their level of interest in teaching, how they saw themselves (i.e., their thinking), and the experiences they had pursued related to teaching (i.e., their actions). We grouped participants into three identity categories, which are described in more detail in the results.

This phase also involved creating iterative drafts of a visual representation of the factors influencing teaching identity and the relationships among these factors. On multiple occasions, each researcher independently developed and presented a visual representation. Discussing and synthesizing these representations revealed similarities and differences in our thinking, which then guided additional analysis (Charmaz 2006). We constantly compared drafts of the model to written descriptions of results and to full transcripts to build a final model representative of the experiences of our participants.

Trustworthiness in qualitative analysis.

We employed multiple strategies to maximize the trustworthiness of our qualitative work. We used deliberate sampling to address our research questions (e.g., Mays & Pope 1995, Charmaz 2006). We also aimed to be transparent in describing our methods in order to increase the confidence in our interpretations and make our analysis as understandable and replicable as possible (Denzin 1978). Additionally, we completed all analyses in a team. This adds trustworthiness because it requires extensive discussion and consensus making and counters individual biases towards the data. All authors were fully immersed in the data over a long period of time, providing us with ample knowledge to critically consider the data and interpretations (Eby et al. 2009). Lastly, we sought feedback from experts who were both experienced qualitative researchers and faculty within life sciences departments.

Results

We classified our participants into three categories: graduate students who had a salient and stable teaching identity (n = 12), those with a nascent teaching identity (n = 7), and those who did not have a teaching identity (n = 14) (Table 4.1). Participants with a salient and stable teaching identity were interested in college teaching and had repeatedly pursued the chance to teach undergraduate courses and to improve in their teaching, which meant seeking opportunities beyond what was required by their doctoral programs. Participants with a nascent teaching identity were interested in college teaching, and had the potential to develop a salient and stable teaching identity, but may or may not seek opportunities that would foster this identity. Participants without a teaching identity were not interested in college teaching and were unlikely to seek additional opportunities related to teaching due to their lack of interest.

There were a few differences between these groups besides their identities as teachers. Participants with nascent teaching identities had been in a doctoral program for less time than individuals in the other categories and had taught fewer semesters as graduate teaching assistants. In contrast, all participants with a salient and stable teaching identity were in at least their third year of doctoral studies. Additionally, most students with nascent or salient and stable teaching identities were from the United States, but over half of our participants without a teaching identity were citizens of countries other than the United States (Table 4.1).

The remainder of the results describe what influenced teaching identity among our participants. These factors are summarized in a visual depiction of the factors influencing teaching identity (Figure 4.1), and described in detail below. We draw heavily on the words of our participants to present our results. We have lightly edited some quotes for grammar and clarity, but always endeavored to maintain the intended meaning. We use pseudonyms to refer to all participants and to other individuals.

The professional culture in the life sciences

The professional culture was a pervasive influence in the experiences of our participants. We found it productive to use a metaphor to describe the role of professional culture in the development of a teaching identity (Figure 4.1). The professional culture of life sciences acted like a blizzard that doctoral students had to navigate through to develop a teaching identity. Students commonly experienced a negative response to their interest in teaching from individuals in their departments, which made the environment feel cold and isolating like a blizzard. The culture resisted movement toward opportunities to develop as teachers, much like deep snow and driving winds push back on people trying to navigate in a blizzard. The culture also made it harder to see the opportunities that existed related to teaching, much like blizzard conditions make it hard to visually discern objects, even those that are close by. Thus, we propose that doctoral students in the life sciences have to brave a blizzard to develop an identity as a college teacher (Figure 4.1).

Participants widely perceived that teaching was less valued than research in the life sciences and in some of their labs and departments. This perception was informed by their observations and experiences within and beyond their institution, and was not limited to those with salient and stable teaching identities. Importantly, these perceptions were not the result of single interactions or the influence of any one individual. Rather, they developed over time and were a general impression of the culture.

Many participants had formed perceptions of what was valued in the work environment by observing faculty. Participants noticed that many faculty minimized the time they spent on teaching and improving as teachers because they garnered prestige and rewards as a result of their research, rather than their teaching. For example, Priya reasoned that tenure-track faculty at research-intensive universities want to focus on their research, even if they like teaching, because it is more important to their career success:

> “So, I think I understand why [tenure-track faculty in the life sciences] really want to just be focused on their lab work because that is what is giving them credibility

> as a scientist because it’s either publish or perish, that’s unfortunately the culture right now... even if they enjoy interacting with students, is not going to give them the grants.”

Participants also perceived that faculty saw teaching as a focus for people who were less successful in research. Andrew explained it this way:

> “I feel like there is a certain attitude that if you’re into [teaching] then you’re no good at anything else that you do...it is just something that we have to do and it should not be an area of focus for a real scientist, right? A real scientist only cares about their research and they teach because they have to.”

Andrew described how this perspective made him doubt himself:

> “So that’s something that kind of makes you want to--when you worry about whether you’re weird, or you’re wrong, or you’re just not good at actual science? And maybe that’s why you’re into teaching...So I must’ve been bad at what I’m doing because I like doing this other thing, so it must mean that I’m not good at research. Certainly I’ve felt that.”

Some participants explained that their advisor or lab subscribed to these beliefs. For example, Julia said:

> “I think in general in my lab, what is important is the science, the bench part. We are not supposed to become teachers. That is secondary. That’s something that we have to do, but it’s not what we are supposed to do.”

Similarly, when Joshua was asked if it would be looked upon favorably in his lab for a student to dedicate time to being a good teacher, he responded “No, not at all. The work we do takes a lot of time, and teaching, at least in all of our point of view, takes a lot of time out of our actual work.” This idea and others may be learned from advisors. For example, Rahul explained “I think I should really mention that a lot of my viewpoints that I have about teaching and science comes from discussions with [my advisor] or listening to him speak.”

The attitude that teaching is less valued that research may dissuade students from seeking teaching opportunities even if they have an interest in teaching. Maria was aiming to be a faculty member at a large, research-intensive university and had a nascent teaching identity. She "really enjoys" teaching and sees it as one motivator for her career plans. Like other participants, Maria saw teaching experiences where she has "autonomy over the presentations" as important training for her future career. However, time spent away from research is seen unfavorably by her advisor:

> "Actually, my mentor, he's very adamant about -- as a graduate student, you do your research. And really what matters to get you to the next level is the research. Your teaching experiences kind of come second. So, if anything negatively affects the research, then it should be something that gets placed second priority."

Maria believed she is unlikely to get any more teaching experiences, and even less likely to have teaching experiences that include ownership while in graduate school.

Participants also got the message that specialized training was unnecessary to be an effective teacher. Some advisors explained that learning to teach occurs when a faculty member teaches their first class, making any formal preparation a poor use of time. Karen described conversations with her advisor about preparing for teaching responsibilities:

> "He doesn't see the value in taking classes or TA-ing a lab for the experience or anything like that. For him, it's all about the science. And so, I'm sure that I have adapted some of that throughout my career."

Some participants struggled to figure out what experiences they needed to pursue to prepare them for a career involving teaching. For example, Matthew wanted to teach at a community college, but found that most of the people he spoke to about it "didn't really know 100% what they were talking about." Faculty and peers at his institution communicated that "all you should be doing is publishing and that's what's going to get you a job," whereas individuals who worked at smaller colleges told him, "that's important, but if you don't do the teaching experience…you're not going to be getting a teaching job." This "difference in voices" made him feel conflicted. Should he prioritize the more dominant and common perspective communicated by faculty he sees every day or the perspective of a few faculty at the type of institution to which he aspired?

One consequence of the blizzard-like culture was that students who were interested in teaching questioned their identities as scientists. Matthew's evolving thoughts exemplify this struggle. He questioned how his choice to teach at a community college would impact his science identity, "Do I still get to call myself a scientist once I become a full time teacher?" To Matthew, the day-to-day responsibilities of being a college teacher were divorced from those of a scientist. He explained:

> "I still say that I would be a scientist, but I feel like I'm almost relying on my degrees to call myself a scientist rather than what I'm practicing, you know?...I'm kind of up in the air on it to be perfectly honest."

Later in the interview Matthew seemed to expand his definition of what it means to be a scientist to include teaching. He said, "I would say pretty firmly that I'm still conducting some critical inquiry and I'm not going to stop being a scientist just because I'm not conducting publishable research."

Though many participants with salient and stable teaching identities described questioning their decisions to pursue a career focused on teaching, they also described personal resiliency which helped them persist in an environment where their aspirations were discouraged. Catherine recalled interactions with her committee in which they tried to dissuade her from taking on teaching responsibilities:

> "My committee members have not been the most supportive of my teaching. One has called it PhD suicide and another has really just told me that I need to stop teaching and focus on research."

However, Catherine continued to pursue what she thought was best for her own professional development. These participants were not unaffected by the professional culture, but they worked to minimize the effect of others' values on their own decisions. Anna explained that she had developed "thick skin" and that discouragement from faculty had made her "even more determined." Similarly, Justin recognized that teaching-focused careers were less respected by some, but was steadfast in his career goals:

> "So I guess...public perception that these types of jobs are a step-down has kind of gone through my head, but then I think about it more and then I'm like, 'I don't really care what people think. This is what I want to do.'"

Participants with salient and stable teaching identities almost all intended to pursue careers at institutions that focused more heavily on undergraduate education than their graduate institution, including primarily undergraduate institutions and community colleges (Table S4.1). They anticipated that the institutions where they aspired to work would value teaching more than their current institutions and this contributed to their resiliency. Robert explained that he wanted to work at a small, primarily undergraduate institution because he thought these jobs were more likely to prioritize "teaching and developing your art as a teacher." He loved teaching, wanted to improve at teaching, and saw research-intensive universities as placing little value on teaching and a scholarly approach to teaching.

Interest in college teaching

Interest in college teaching was key to a teaching identity. Opportunities to teach before and during graduate school were often responsible for piquing interest in teaching. Interest in college teaching only transformed into interest in a career involving teaching when doctoral students were aware of viable career options (Figure 4.1). Some participants discovered their interest in teaching as graduate teaching assistants. Justin came to graduate school because he was interested in research. He worked as a teaching assistant during his third year to fulfill a departmental requirement. After this experience, he was again supported on research assistantships and he missed teaching. He felt that it had been the most rewarding part of graduate school, and recalled thinking, "Well, if that's what I value and that's what brings me happiness, then I should pursue a career where that's kind of the goal." Other participants became interested in college teaching prior to graduate school. As an undergraduate, Elizabeth aspired to become a dentist, but interacting with an inspirational professor and serving as an undergraduate teaching assistant showed her other possibilities (Figure 4.1).

Though some participants, like Elizabeth, went to graduate school because they were interested in a teaching career, most participants did not begin graduate training aware that they wanted a career involving college teaching. Therefore, opportunities to teach as a graduate student were critical for exploring career interests. Participants were best positioned to make informed decisions if they had the opportunity to teach undergraduate courses early in graduate school because this gave them time to develop as teachers. Recognizing their own interest in teaching motivated students to seek teaching professional development and independent teaching experiences (Figure 4.1). Kelsey explained:

> "I hear from other people in other departments, where they don't even have teaching [requirements] at all, and they will teach their last semester and they're like, 'Oh, I actually really like this and now I'm screwed because I haven't developed my teaching skills'."

Participants with salient and stable teaching identities described how much they enjoyed working with students and watching them learn. They discovered this through teaching experiences. For example, Justin found that he appreciated the chance to "have a really close connection" with students, and push students to think more deeply until they had a “wow moment.” Kelsey explained, “I really enjoy the students. I enjoy interacting with them. I really enjoy like in lab when they're all talking to each other and I'm walking around and it's just fun.”

Teaching experiences helped some participants confirm they were not interested in careers involving teaching. Emma had taught for three semesters. Teaching experiences confirmed her belief that she was not good at teaching and did not enjoy it. She explained: “I haven’t had any bad experiences really. It’s not like I saw something and was scarred for my whole life. I just have never liked being a teacher...I’m not good at it.” Not all students who were uninterested in teaching saw themselves as bad teachers. Others felt frustrated by undergraduates, especially when “dealing with people who aren’t intellectually invested in what’s going on in the course.” Another explained that “teaching comes with a lot of responsibilities which I don’t like.” These students often intended to pursue positions outside academia because they did not want any teaching responsibilities.

Other participants were interested in teaching, but may not pursue opportunities to develop as a teacher because they do not see viable career options that included teaching (Figure 4.1). Marco, a first year student, discovered an interest in teaching during his first graduate teaching experience, which was mandatory, “I’ve been learning like, ‘Oh teaching is also probably not a bad idea’ because I kind of enjoy it.” Despite having a nascent teaching identity, Marco was not seriously considering a career involving teaching because he was unaware of careers involving teaching beyond being faculty at a research-intensive university, and he was not interested in that career path. His lack of awareness of other career options left him thinking that he would be unable to pursue dual interests in research and teaching, and therefore may hinder him from pursuing opportunities that would strengthen his teaching identity. Robert had the same perspective until he learned, through a teaching professional development program, about the wide variety of positions available in higher education that involve teaching.

International participants did not always see opportunities for careers involving teaching in their home country. Some expected positions that involved teaching undergraduates to be rare or unavailable, or anticipated that they would be overqualified for these positions after earning a doctorate. Not surprisingly, doctoral students who did not see teaching in their future chose not to pursue additional chances to teach or participate in teaching professional development, both of which were important for developing coming to have a salient and stable teaching identity.

Teaching professional development

Teaching professional development fostered teaching identities among our participants by helping them gain teaching knowledge and skills, helping them identify teaching mentors, and introducing them to other like-minded peers who were also interested in teaching (Figure 4.1). Additionally, formal teaching professional development helped students prioritize teaching preparation, even when they felt too busy to find time for additional work.

Not much teaching professional development was required for doctoral training in the life sciences at this institution, so students interested in developing as teachers had to pursue opportunities themselves. Depending on their graduate teaching assignments, doctoral students were required to take one or two 1-credit courses related to pedagogy.

Participants received little or no teaching professional development within their teaching assignments, Anna explained:

> "Like when we TA, the professors do not do anything with teaching development when you're TAing. You're there to TA. You're there to do whatever it is that you're supposed to do. It's not about you gaining teaching experience or teaching skills."

Some doctoral students pursued teaching professional development beyond what was required. The university provided two formal, elective teaching professional development programs for graduate students across disciplines: a certificate in undergraduate teaching and a future faculty program. Any student could earn the certificate, but there was a yearly selection process for 15 students to be part of the future faculty program. Of the twelve participants with salient and stable teaching identities, nine were working toward the certificate and five had participated in the future faculty program. Participants without salient and stable teaching identities were not pursuing either of these programs (Table S4.1). The teaching certificate program was self-guided, and generally required over a year to complete. It entailed teaching four sections as a graduate teaching assistant, designing a teaching project and showing its effectiveness in the classroom under the guidance of a self-selected mentor, disseminating the results of the teaching project, and completing three elective pedagogy courses. The future faculty program included a twice monthly meetings facilitated by a staff member from the Center for Teaching and Learning to talk about and improve their teaching. All participants had previously won a departmental or institutional teaching award.

Participants described benefits from the teaching certificate. It helped participants hold themselves accountable for participating in teaching professional development. Ryan was glad that he had committed to the teaching certificate early in graduate school because, “You start getting to your second and third year and you start losing motivation for anything extra. It was good that I had sort of made up my mind early on.” Other participants said the certificate forced them to take pedagogy classes and work on a teaching project. They welcomed this push because they might otherwise have prioritized research. Interestingly, the certificate program provides no external accountability. Students could have stopped working toward the requirements at any time without consequence other than not earning the certificate. They used the structure of the requirements to hold themselves accountable for participating in training they saw as valuable.

Participants especially discussed the value of the pedagogy courses they completed to meet certificate requirements. The courses provided the chance to build knowledge and skills, as well as access to teaching mentors. These courses were generally taught using evidence-based instructional practices and participants saw benefits in taking classes that were taught using the methods promoted in the courses. Participants described learning about a range of topics in these courses including writing teaching philosophies, active-learning strategies, “how to actually grade writing and science,” course design, and creating a syllabus.

Some participants were surprised by how much they gained from pedagogy courses. At first Catherine “wasn’t super excited about taking the courses” because she thought that “being in the classroom is probably the best way to learn about teaching.” However, she also said, “I got far more than I expected to get out of these courses.” Many participants who took these courses described building connections with the instructors as an important benefit. For example, one of Catherine's instructors became her teaching mentor (Figure 4.1).

The future faculty program helped students identify like-minded peers (Figure 4.1). Robert asserted that the future faculty program was a primary factor in his decision to pursue a teaching-focused career because it "opened my eyes to the vast amount of teaching opportunities" beyond large research-intensive universities. Discussing teaching with people who shared his passion helped Robert realize that he wanted a teaching career. He recalled, "their enthusiasm and love for teaching really kind of rubbed off on me to where I was like, 'This is what I want to do.'" Like the teaching certificate program, the future faculty program provided a chance to build relationships with teaching mentors. Robert described the future faculty program leader as "the ultimate role model" for teaching because he "cares about his students more than anyone I've ever seen."

Some participants did not intend to participate in either of these formal teaching professional development programs because they were unaware of what was required and the potential benefits. For example, Rahul described the teaching certificate this way, "I'm sure I'm missing something, but I do not know what I am missing." Although the program is advertised in one of the required pedagogy courses, some participants did not have the information they needed to make a decision.

Teaching mentors and like-minded peers

Participants benefited from relationships with faculty who acted as mentors and advocates, and from relationships with like-minded peers. Participants commonly identified mentors and built these relationships through teaching professional development programs (Figure 4.1). Mentors provided advice, materials, encouragement, and inspiration, which helped foster their identities as teachers. Teaching mentors and like-minded peers also encouraged participants to engage in independent teaching experiences. Catherine explained what her mentor provided her, “I’ve been in some scenarios in which I really needed her mentoring voice to kind of guide me through those scenarios in teaching and [she] really has kind of developed me as an instructor.” Stephanie developed a mentoring relationship with the primary instructor for one of her teaching assignments. Stephanie explained that her mentor provided more than just advice and opportunities, “It’s nice to know that I’m passionate about something and she kind of reciprocates that and is encouraging my passion.”

Teaching mentors, unlike research mentors, are not a standard part of doctoral training in the life sciences. Participants only found these mentors if they sought them out or sought out experiences that brought them into contact with potential mentors. Kelsey explained that graduate students may not know faculty or other students who are interested in teaching, “I think a lot of students don’t know about that little [teaching] community and where it is, especially graduate students.” A few participants found teaching mentors by identifying faculty in their department who had a reputation for being invested in teaching. However, Megan described teaching mentorship as a limited resource. Megan identified only one potential mentor in her department and explained that one person may not be enough to serve all the interested graduate students. Notably, major advisors typically were not seen as teaching mentors. Participants rarely described their advisors providing teaching feedback or directing them to teaching opportunities. A few participants even wanted advocacy from other faculty to convince their advisors to support them in their teaching development. Furthermore, many faculty supervise teaching assistants, but most were not seen as providing teaching training or mentorship.

Advocates helped participants convince their advisors that opportunities to develop as a teacher were worthwhile. One faculty member, who we will refer to as Alison, was described as an advocate by two participants from the same department. They saw Alison as willing to talk to their advisors on their behalf to advocate for more teaching opportunities. Specifically, Alison spoke to other faculty about the importance of teaching experience on the job market. Catherine described Alison as being:

> "A voice for me in the department as well, making sure that I have an opportunity to do all the teaching that I want to do and have the best opportunities that I can. Alison has encouraged the department to allow me to do things and encouraged my PI to allow me to do more teaching-related things."

Doing more teaching was "not necessarily encouraged" in Catherine's department, making Alison's advocacy important to legitimize Catherine's pursuits.

Peers who shared a passion for teaching provided participants with a forum to talk about teaching and teaching challenges. Participants met like-minded peers within their department and through teaching professional development. Elizabeth described the importance of being with people who can "support and relate to the good and the bads" of teaching, "when you're in the trenches, nothing helps more than people who understand what you're going through because they're going through it too." Matthew explained that his peers inspired him to become more involved with teaching opportunities, "having a cohort of people who are interested and passionate about [teaching]... pushes you because you feel left behind if you're not pursuing the same opportunities."

Independent teaching experiences

Participants commonly desired independent teaching experiences. Such experiences provided opportunities to be recognized by others as a teacher and made participants more likely to recognize themselves as a teacher (Figure 4.1). Even participants without a teaching identity expressed interest in the chance to teaching independently. Students who had not felt a sense of independence in their teaching did not get as much satisfaction from their teaching, and often were not excited about future teaching opportunities. Therefore, students who had experienced independence were motivated to seek more opportunities and students who had not experienced independence were less likely to know these opportunities existed or to seek them out (Figure 4.1).

Participants felt a teaching experience was independent when they had the chance to plan and implement their own class activities and have authentic, autonomous interactions with students. Different types of experiences could provide this, including guest lecturing, serving as instructor of record, and having freedom to make teaching-related choices in typical teaching assignments. However, this autonomy was not consistently available in teaching assignments. Andrew, who had a salient and stable teaching identity, explained that the standard teaching assignment was often insufficient for students to develop as instructors:

> “I think that a teaching assignment requirement is a start and not necessarily a full complement of training, especially if that teaching assignment is something like grading for a lecture course. I think student interaction and then independent student interaction are key experiences that some grad students miss out on because they're either, like I said, a grader or they're in a lab with a professor there. They're not making their own decisions about how to be in a class with students."

Many participants with salient and stable teaching identities felt motivated to search for new teaching opportunities because they enjoyed previous independence in their teaching or found their prior teaching assignments too limiting. For example, Anna wanted more opportunities for independence because she had not been "asked to give lectures or design activities that often" in prior teaching assignments. Instead she often felt that her primary responsibilities were "babysitting," making "sure students don't light themselves on fire," and helping with grading. When Anna expressed her interest in a teaching career, her major advisor invited her to guest lecture in his 200-student course. He was one of just a few advisors who helped participants find independent teaching opportunities. Anna described this guest lecturing opportunity as "absolutely one of the best teaching things I could have done just because it gave me experience giving not only just a lecture but a lecture to a larger class."

After this successful experience, Anna wanted more opportunities to teach independently. She felt that having ownership in her teaching helped her learn about course design and gathering and using feedback from students. At the time of the interview, Anna was preparing to teach a course as instructor of record with mentorship from a faculty member. At this institution, the instructor of record has ultimate responsibility for planning and teaching a course, including developing the course plan and syllabus, designing each class period, and creating and grading assessments. Anna had just begun preparing for this role and was pleased that it was already a "valuable learning experience."

Participants with nascent teaching identities or who did not have teaching identities also desired more ownership in their teaching. Participants who were not interested in teaching wondered if they would feel differently had they had more independence. Joshua explained:

> "I do not share the passion of being a teacher. Which means, although I get excited while teaching, I don't feel accomplished much afterwards. Maybe it has something to do with I have been teaching dependently. Dependently means I answer to another actual educator or teacher, that I work for him or her instead of having my own classes. So I don't feel responsible to the students, instead I feel responsible to the teacher."

Another participant, Lucas, had hoped for a teaching assignment similar to that of a friend who he described as "doing full lectures." However his teaching assignment primarily involved cleaning and preparing materials for a lab course, which disappointed him. Lucas explained, "So I wasn't so excited about it and I just like gave up at the time and never thought about it again."

External recognition as a teacher

Only a few participants—all of whom had salient and stable identities—described instances when they felt like others recognized them as a college teacher. Participants felt recognized as teachers when they were offered additional teaching opportunities and when they received positive feedback on their teaching from mentors (Figure 4.1). When she guest lectured, Anna received positive feedback from her advisor and a staff member of the Center for Teaching and Learning. She described receiving positive feedback and recognition from someone whose job was "helping people with teaching" as a powerful experience that made her "feel like I was being successful in teaching." Catherine, a fifth-year graduate student, had more independent teaching experience than any other participant and therefore had the most opportunities to feel recognized as a teacher. She felt her department recognized her as a teacher when they asked her to teach a large lecture course, a responsibility not normally offered to graduate students. She was also asked to serve on a search committee for an instructor who would teach a course that Catherine had previously taught as instructor of record. Being recognized as a teacher by faculty in her department helped her recognize herself as a teacher. For example, she decided to mentor and share course materials with a new instructor in the department in order to "focus on making sure that she's a successful new hire to the university and mentoring her as she develops." This is evidence that she sees herself as a college teacher capable of supporting other college teachers. Anna and Catherine are unique. Most participants did not describe any experiences in which they felt recognized as college teachers.

Conclusions

Life sciences doctoral students navigated a professional culture that marginalized those interested in teaching. Students with salient and stable teaching identities persisted despite marginalization, but we suspect that not all students with nascent teaching identities will have the skills, resources, and agency to brave the blizzard of the professional culture. One result of this is that doctoral training at this institution, and probably many others, is propagating a narrow view of what it means to be a scientist. Improving the diversity and preparation of STEM professionals likely requires broadening our view of what it means to be a scientist. Researchers are scientists, and so are those who are primarily engaged in science outreach, national science policy, entrepreneurial science, and science teaching (e.g., Ecklund et al. 2012, Szelynyi et al. 2016, Thiry et al. 2007).

Faculty at research-intensive universities train the majority of doctoral students in the life sciences, yet we are often ill-equipped to prepare students for careers besides our own for at least two reasons. First, our professional experience is often limited to working at research universities, and we may not be knowledgeable about the expectations in other careers. For example, associate, baccalaureate, and master's institutions value teaching experience more highly than research experience or publication record, but faculty at doctoral institutions value research above anything else (Fleet et al. 2006). Second, there is an inherent conflict of interest between what a life sciences faculty member at a research university needs to accomplish for career advancement and the training students need for any role besides being a researcher. Life sciences faculty are expected to secure extramural funding for research and to fund the training of graduate students. When they are awarded this funding, they are responsible to the funding agencies for accomplishing the proposed research, and failing to do so will negatively impact their ability to secure funding in the future. In this way, the funding of graduate students is linked to the production of research, and the rationale path for faculty is to prioritize research productivity. Recognizing this conflict of interest will allow us to rethink how graduate training could be designed to prepare students for other careers, including those in industry, policy, and academic institutions that are different than research universities. For example, Bruce Alberts and colleagues proposed that we move toward funding graduate students on training grants and fellowships rather than through research grants because it allows for more peer-review and federal oversight of how students are trained (Alberts et al. 2014). Doctoral programs that include time for all students to engage in significant professional development other than development as a researcher could also address this conflict of interest. Additionally, we expect that life sciences faculty at research universities would welcome professional development aimed to help them better understand the various careers our graduate students pursue and what will make them competitive on the job market and successful in their job responsibilities. National funding agencies could lead the way by stating that competitive training and fellowship proposals will explicitly invest in supporting graduate student preparation for diverse careers.

Our findings suggest several factors may be important for supporting doctoral students who aim to have careers as college science faculty. Students benefited from the chance to teach early in graduate school because it allowed them to explore their career interests. Presenting teaching as an opportunity for career discovery, rather than as a distraction from research, might help students seriously consider multiple career paths (e.g., Gibbs & Griffin 2013). Doing so in the first years of graduate school will allow students to pursue additional training appropriate for their career interests.

Formal teaching development programs helped students gain knowledge and skills and were key to developing relationships that fostered their teaching identities, but were not widely accessed. The future faculty program was only available to students who had won a competitive teaching award, thereby selecting for students who likely already had a salient and stable teaching identity. The certificate program was viewed as too much of a time commitment by many students and faculty. Many students also lacked information about the program and its potential benefits. Thus, teaching professional development programs could better reach their potential by widely advertising directly to students, by assessing and advertising their outcomes, and through integration into typical doctoral training at an institution. Departments could play a role in this by encouraging students interested in teaching to participate in local or national teaching professional development programs. For example, the CIRTL Network (Center for the Integration of Research, Teaching, and Learning) offers online pedagogy courses, online discussions, and summer institutes (https://www.cirtl.net). Lastly, independent teaching experiences were highly valued by students. Graduate training recognizes the importance of increasing students' ability to conduct research independently, but life sciences doctoral programs are not typically designed to foster teaching independence. Creating easier access to teaching mentors and feedback could help students work toward independence and provide more opportunities for students to receive recognition as a teacher from faculty.

One compelling approach for redesigning a graduate program to better foster teaching knowledge, skills, and identity is dividing the graduate affairs committee into a graduate research committee and a graduate teaching committee (Kendall et al. 2013). This change, which was made by a life sciences department at a research-intensive university, recognized the unique training and support needed for teaching and ensured that these matters were not overshadowed by a focus on research. This graduate teaching committee makes teaching assignments with the professional development of graduate students in mind, providing a chance to scaffold students into more independent teaching experiences. The committee also oversees feedback for teaching assistants, ensuring that graduate students receive peer feedback and feedback from faculty on their teaching each year (Kendall et al. 2013). Positive evaluations open doors for more independent teaching experiences, and even becoming involved in departmental teaching reform efforts. Thus, there are formal avenues for recognizing students as teachers. Furthermore, the committee makes recommendations to the departments regarding courses that could be offered within the department to support students in their development as teachers (Kendall et al. 2013). This is just one possible way to change graduate training to better support teaching identities. It is promising that it addresses major factors that fostered teaching identity in our study.

Almost all of the students with salient and stable teaching identities intended to seek careers involving college teaching at primarily undergraduate institutions (Table S4.1). They had made this decision because they saw the values of a research university as out-of-line with their own values, and they expected other institution types to have a value system more aligned with their own. With one exception, the students in our sample who intended to seek careers as faculty at research-intensive institutions had nascent teaching identities or none at all (Table S4.1). Our sample is small, but these observations raise questions about what needs to change in doctoral training and the professional culture of the life sciences so that students aiming for careers in research-intensive universities see the opportunity to develop as a teacher as worth their time.

This study has several limitations that must be taken into account in interpreting these data. First, we investigated the experiences of students in the life sciences. Investigations of students in other science disciplines, including chemistry, physics, and geosciences, will be important to determine the degree to which our findings are discipline-specific. Ecklund et al. (2012) observed different perspectives about science outreach between biologists and physicists. We cannot assume that the same factors will influence identity in the same way across disciplines. Second, we collected data at one time point for each participant, requiring them to recall prior experiences. This retrospective approach may not fully capture what has influenced their professional identity and how. Longitudinal studies will be an important step for understanding the development of teaching identities. Lastly, this work examines a single university. Clustering participants within a single context allows for deeper insight regarding the local culture, which was important to this study. However, it also means that our findings may not be relevant to all contexts. An important next step will be comparing and contrasting cultures and training programs to better understand the complexity of what influences teaching identity.

We have described a model of the factors influencing teaching identity among life sciences doctoral students. This is important groundwork for understanding teaching identity among future and current faculty. Future empirical work can build on this model by examining teaching identity among new faculty and elucidating how teaching identity impacts instructional choices and the use of evidence-based instructional strategies over time. We wonder if our findings will be unsurprising to people who have been in a life sciences department at a research-intensive university. This leaves us asking: if this reality is already well-known, what will it take to motivate us to change our training approaches to prepare students for diverse and valuable careers beyond research?

References

Alberts, B., Kirschner, M. W., Tilghman, S., & Varmus, H. (2014). Rescuing US biomedical research from its systemic flaws. *Proceedings of the National Academy of Sciences, 111*(16), 5773-57777.

Andrews, T. C., Conaway, E. P., Zhao, J., & Dolan, E. L. (2016). Colleagues as change agents for undergraduate teaching. *CBE-Life Sciences Education, 15*(2), 1-17.

Andrews, T. M., & Lemons, P. P. (2015). It’s personal: Biology instructors prioritize personal evidence over empirical evidence in teaching decisions. *CBE-Life Sciences Education, 14*(1), 1-18.

Andrews, T. M., Leonard, M. J., Colgrove, C. A., & Kalinowski, S. T. (2011). Active Learning Not Associated with Student Learning in a Random Sample of College Biology Courses. *CBE-Life Sciences Education, 10*(4), 394-405.

Austin, A. E. (2002). Preparing the Next Generation of Faculty. *The Journal of Higher Education, 73*(1), 94-122.

Austin, A. E. & McDaniels, M. (2006). Preparing the professoriate of the future: Graduate student socialization for faculty roles. *Higher Education: Handbook of Theory and Research, 21*, 397-456.

Barker, C. (2005). *Cultural Studies: Theory and Practice*. London: Sage.

Beijaard, D., Meijer, P. C., & Verloop, N. (2004). Reconsidering research on teachers' professional identity. *Teaching and Teacher Education, 20*(2), 107-128.

Bieber, J. P., & Worley, L. K. (2006). Conceptualizing the Academic Life: Graduate Students' Perspectives. *The Journal of Higher Education, 77*(6), 1009-1035.

Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge, England: Cambridge University Press.

Brownell, S. E., & Tanner, K. D. (2012). Barriers to Faculty Pedagogical Change: Lack of Training, Time, Incentives, and...Tensions with Professional Identity? *CBE-Life Sciences Education, 11*(4), 339-346.

Carlone, H. B., & Johnson, A. (2007). Understanding the science experiences of successful women of color: Science identity as an analytic lens. *Journal of Research in Science Teaching, 44*(8), 1187-1218.

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. Washington, DC: Sage.

Coldron, J., & Smith, R. (1999). Active location in teachers' construction of their professional identities. *Journal of Curriculum Studies, 31*(6), 711-726.

Connolly, M. (2010). Helping future faculty come out as teachers. In: *Essays on Teaching Excellence: Toward the Best in the Academy* (Vol. 22). Nederland, CO: Professional & Organizational Development Network in Higher Education.

Connolly, M. R., Savoy, J. N., Lee, Y. -G., & Hill, L. B., (2016). Building a better future STEM faculty: How doctoral teaching programs can improve undergraduate education. Madison, WI: Wisconsin Center for Education Research, University of Wisconsin-Madison.

Dancy, M., & Henderson, C. (2010). Pedagogical practices and instructional change of physics faculty. *American Journal of Physics, 78*(10), 1056-1063.

Denzin, N. K. (1978). *The research act: A theoretical orientation to sociological methods* (2nd ed.). New York: McGraw-Hill.

Ebert-May, D., Derting, T. L., Hodder, J., Momsen, J. L., Long, T. M., & Jardeleza, S. E. (2011). What We Say Is Not What We Do: Effective Evaluation of Faculty Professional Development Programs. *Bioscience, 61*(7), 550-558.

Eby, L. T., Hurst, C. S., & Butts, M. M. (2009). Qualitative Research: The Redheaded Stepchild in Organizational and Social Science Research? In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences* (pp. 219-245). New York: Routledge.

Ecklund, E. H., James, S. A., & Lincoln, A. E. (2012). How Academic Biologists and Physicists View Science Outreach. *PLoS ONE, 7*(5), e36240.

Fleet, C. M., Rosser, M. F. N., Zufall, R. A., Pratt, M. C., Feldman, T. S., & Lemons, P. P. (2006). Hiring Criteria in Biology Departments of Academic Institutions. *BioScience, 56*(5), 430-436.

Gibbs, K. D., Jr., & Griffin, K. A. (2013). What Do I Want to Be with My PhD? The Roles of Personal Values and Structural Dynamics in Shaping the Career Interests of Recent Biomedical Science PhD Graduates. *CBE-Life Sciences Education, 12*(4), 711-723.

Hall, L., & Burns, L. (2009). Identity Development and Mentoring in Doctoral Education. *Harvard Educational Review, 79*(1), 49-70.

Henderson, C., Beach, A., & Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching*, *48*(8), 952-984.

Henderson, C., Dancy, M., & Niewiadomska-Bugaj, M. (2012). Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process? *Physical Review Physics Education Research, 8*(2), 020104.

Ibarra, H. (1999). Provisional Selves: Experimenting with Image and Identity in Professional Adaptation. Administrative Science Quarterly*, 44*(4), 764-791.

Kendall, K. D., Niemiller, M. L., Dittrich-Reed, D., Chick, L. D., Wilmoth, L., Milt, A., et al. (2013). Departments can develop teaching identities of graduate students. *CBE-Life Sciences Education, 12*(3), 316-317.

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge, United Kingdom: Cambridge University Press.

Mays, N., & Pope, C. (1995). Rigour and qualitative research. *British Medical Journal, 311*(6997), 109-112.

National Research Council (2012). *Discipline-based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*. Washington, DC: National Academies Press.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Thousand Oaks, CA: Sage.

Pratt, M. G., Rockmann, K. W., & Kaufmann, J. B. (2006). Constructing Professional Identity: The Role of Work and Identity Learning Cycles in the Customization of Identity Among Medical Residents. *Academy of Management Journal, 49*(2), 235-262.

Ronfeldt, M., & Grossman, P. (2008). Becoming a Professional: Experimenting with Possible Selves in Professional Preparation. *Teacher Education Quarterly, 35*(3), 41-60.

Saldaña, J. (2013). *The Coding Manual for Qualitative Researchers* (2nd ed.). Washington, DC: Sage.

Sfard, A. (1998). On Two Metaphors for Learning and the Dangers of Choosing Just One. *Educational Researcher, 27*(2), 4-13.

Shadle, S. E., Marker, A., & Earl, B. (2017). Faculty drivers and barriers: laying the groundwork for undergraduate STEM education reform in academic departments. *International Journal of STEM Education, 4*(1), 8.

Stets, J. E., & Serpe, R. T. (2013). Identity theory. In J. DeLamater & A. Ward (Eds.), *Handbook of Social Psychology* (pp. 31-60). New York: Springer.

Szelényi, K., Bresonis, K., & Mars, M. M. (2016). Who Am I versus Who Can I Become?: Exploring Women's Science Identities in STEM Ph.D. Programs. *The Review of Higher Education, 40*(1), 1-31.

Thiry, H., Laursen, S. L., & Liston, C. (2007). (De)valuing teaching in the academy: why are underrepresented graduate students overrepresented in teaching and outreach? *Journal of Women and Minorities in Science and Engineering, 13*(4), 391-419.

Turner, C. S. V., & Thompson, J. R. (1993). Socializing women doctoral students: Minority and majority experiences. *The Review of Higher Education, 16*(3), 355-370.

Turner, J. L., Miller, M., & Mitchell-Kernan, C. (2002). Disciplinary Cultures and Graduate Education. *Emergences: Journal for the Study of Media & Composite Cultures, 12*(1), 47-70.

Wegner, E., & Nückles, M. (2015). Knowledge acquisition or participation in communities of practice? Academics' metaphors of teaching and learning at the university. *Studies in Higher Education, 40*(4), 624-643.

Weidman, J. C., Twale, D. J., & Stein, E. L. (2001). *Socialization of graduate and professional students in higher education: A perilous passage?* San Francisco, CA: Jossey-Bass.

Tables

Table 4.1. Participant background and characteristics categorized by teaching identity.

Teaching identity	# of participants	% female	% international	Median (SD) year in PhD	Median (SD) terms as TA
Salient and stable	12	58%	0%	5 (1.2)	6 (4.6)
Nascent	7	86%	14%	1 (1.5)	1 (1.1)
Lacking	14	57%	64%	4 (1.5)	3 (4.5)
Total	**33**	**64%**	**30%**	**4 (1.7)**	**2 (4.4)**

SD = standard deviation

Figures

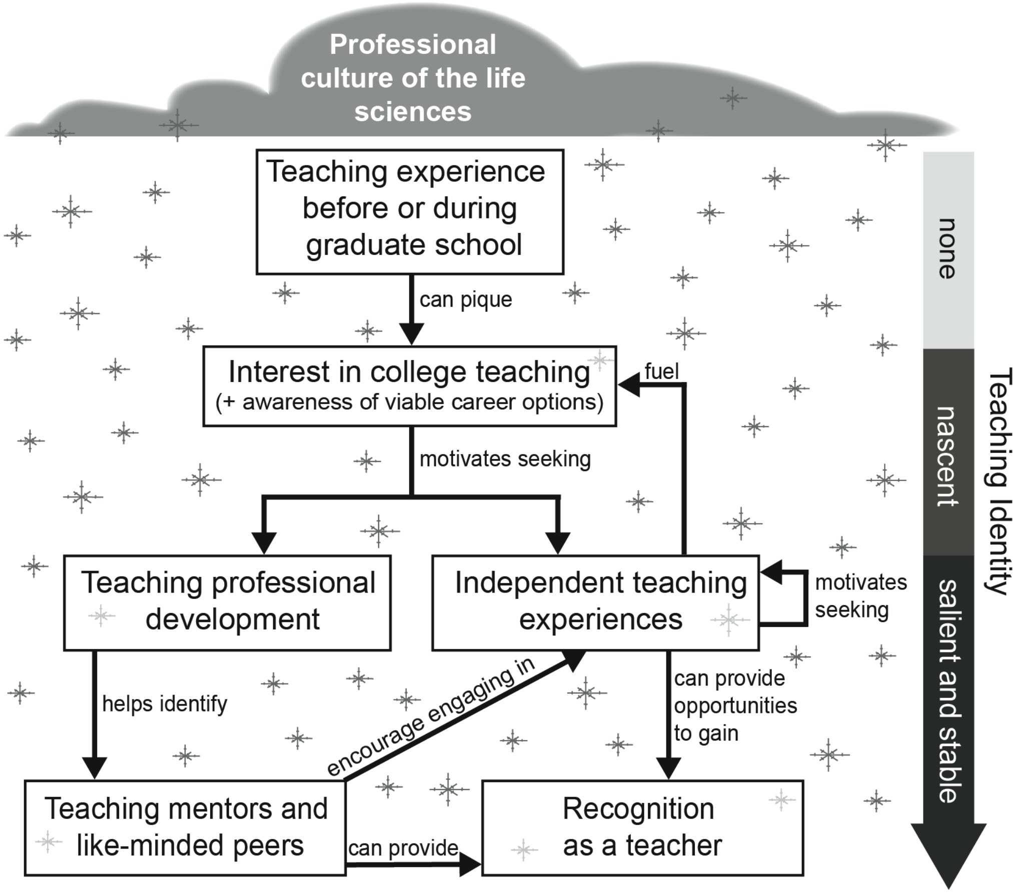

Professional culture of the life sciences
Teaching experience before or during graduate school
can pique
Interest in college teaching
(+ awareness of viable career options)
fuel
motivates seeking
Teaching professional development
Independent teaching experiences
motivates seeking
helps identify
encourage engaging in
can provide opportunities to gain
Teaching mentors and like-minded peers
can provide
Recognition as a teacher
none
nascent
salient and stable
Teaching Identity

Figure 4.1. A model of the factors influencing teaching identity among life sciences doctoral students. The professional culture of the life sciences is like a blizzard, obscuring the view of and acting as resistance against participating in opportunities that foster a teaching identity. Students must brave the blizzard to seek experiences that foster and affirm a teaching identity, including teaching professional development and independent teaching experiences. These, in turn, help students find teaching mentors and like-minded peers and provide the chance to be recognized as a teacher. The arrow to the right indicates the range of teaching identities observed, aligned with the corresponding experiences.

CHAPTER 5

CONCLUSIONS AND DISCUSSION

Foundations for studying BIA evolution

There is a great deal of work remaining to untangle the evolution of BIA biosynthesis, but investigation into species evolution described in Chapter 2 is necessary to ground future gene-centric analyses in the biology of the species. There are species relationships, such as between *Papaver bracteatum* and *Papaver rhoeas,* which may be important for understanding mechanisms of BIA evolution, but that are complicated by the presences of gene tree discordance possibly caused by incomplete lineage sorting. The potential presence of a rapid radiation within *Papaver* adds complexity to understanding how BIAs evolved in that genus especially morphinian alkaloids that are present only in *Papaver* species. The analyses in Chapter 2 unveil these added complexities so that they can be appropriately considered when hypothesizing evolution of BIAs.

Comparative transcriptomics for studying BIA biosynthesis

It is well known that genomic resources including gene co-expression analyses have increased our ability to understand plant genetics (Proost & Mutwil 2016). Morphine is considered one of the first compounds discovered in plants (Hartmann 2007). Centuries of study have been invested into BIAs*,* yet few genomics resources have been created for the *Papaver* species that produce morphinian alkaloids. In Chapter 3 I compared transcript expression in two closely related, BIA producing species, *P.*

somniferum and *P. setigerum*. I used both transcript expression and gene co-expression network analysis to describe the similarities and differences in these two species with special attention given to BIA biosynthesis genes. These methods indicated genes encoding BIA biosynthesis enzymes share similar co-expression patterns in these species such that they appear in modules together. However, further analysis of these modules indicated that the genes co-expressed with BIA biosynthesis genes differed between species. These analyses also suggested that the transcription factor family MYB should be investigated for a potential role in regulating BIA biosynthesis.

Future questions in *Papaver*

Chapters 2 and 3 of this dissertation lay important groundwork for further investigation into the evolution of BIA biosynthesis. Producing new phylogenies with increased taxon sampling in *Papaver* and other BIA producing genera will continue to grow this framework. These additional phylogenetic analyses would benefit from systematic consideration of gene tree discordance similar to that done in Chapter 2. Gene trees of BIA related enzymes should also be studied to elucidate the evolutionary origins of BIAs. Now that there is a well-supported species tree, this gene history can be compared to the species history.

There is significant opportunity for future investigations to build on the work in Chapter 3. The gene co-expression network analyses suggested transcription factors that warrant further study for their possible role in regulating BIA biosynthesis. These transcription factors can be tested for their role in BIA biosynthesis through the use of virus-induced gene silencing, a method shown to be successful in opium poppy (Hileman et al. 2005). Additionally, gene co-expression network analyses can be applied to other

BIA producing species and focus on additional sections of the BIA pathway. Finally, there has been some success in applying principles of co-expression analysis to metabolome data (e.g. DiLeo et al. 2011). This method could be fruitful if applied to BIAs. These co-expression analyses can provide additional insight for gene discovery and understanding regulation of the BIA pathway.

Modeling teaching identity in graduate students and future questions

There have been increasing discussions in the scientific community on changes that might be considered for improving graduate training (e.g. Alberts et al. 2014). One potential area of improvement is preparing students for a variety of career options (e.g. Gibbs & Griffin 2013). Chapter 4 of this dissertation focused on graduate student professional identity development. How graduate students come to see themselves as professionals may be important as they choose their future careers and come to feel prepared for those careers. Several types of experiences promoted students' professional identities as college teachers including early teaching experiences, teaching professional development, building mentoring relationships with those who valued teaching, and receiving recognition as a teacher. Graduate training programs could aim to provide students more of these types of experiences, encourage students to engage in these experiences, and promote these experiences as an important part of a complete graduate education. Finally, the culture of life sciences can hinder students in their pursuit of these experiences. By devaluing teaching, the culture can make it harder to find and participate in these experiences. Concerted efforts to include teacher as a possible role for a scientist could help alter this culture.

The model developed in Chapter 4 is an important first step to determining what influences graduate student development of a teaching identity. This model can be used as a hypothesis to be tested at additional institutions and departments. Longitudinal studies that repeatedly query students as they traverse graduate school will be key to understanding how these identities develop over time. Specifically, it will be important for future research to investigate how professional development, teaching experiences, and interactions with mentor affect student's perceptions immediately after the event as well as long term. Following up with students to determine their career trajectories and exploring the effect that department and institutions have on professional identity development will also add to our understanding.

References

Alberts, B., Kirschner, M. W., Tilghman, S., & Varmus, H. (2014). Rescuing US biomedical research from its systemic flaws. *Proceedings of the National Academy of Sciences, 111*(16), 5773-57777.

DiLeo, M. V., Strahan, G. D., den Bakker, M., & Hoekenga, O. A. (2011). Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. *PLoS One*, *6*(10), e26683.

Gibbs, K. D., Jr., & Griffin, K. A. (2013). What Do I Want to Be with My PhD? The Roles of Personal Values and Structural Dynamics in Shaping the Career Interests of Recent Biomedical Science PhD Graduates. *CBE-Life Sciences Education, 12*(4), 711-723.

Hartmann, T. (2007). From waste products to ecochemicals: Fifty years research of plant secondary metabolism. *Phytochemistry, 68*, 2831-2846.

Hileman, L. C., Drea, S., Martino, G., Litt, A., & Irish, V. F. (2005). Virus‐induced gene silencing is an effective tool for assaying gene function in the basal eudicot species Papaver somniferum (opium poppy). *The Plant Journal*, *44*(2), 334-341.

Proost, S., & Mutwil, M. (2016). Tools of the trade: studying molecular networks in plants. *Current opinion in plant biology*, *30*, 143-150.

APPENDIX A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

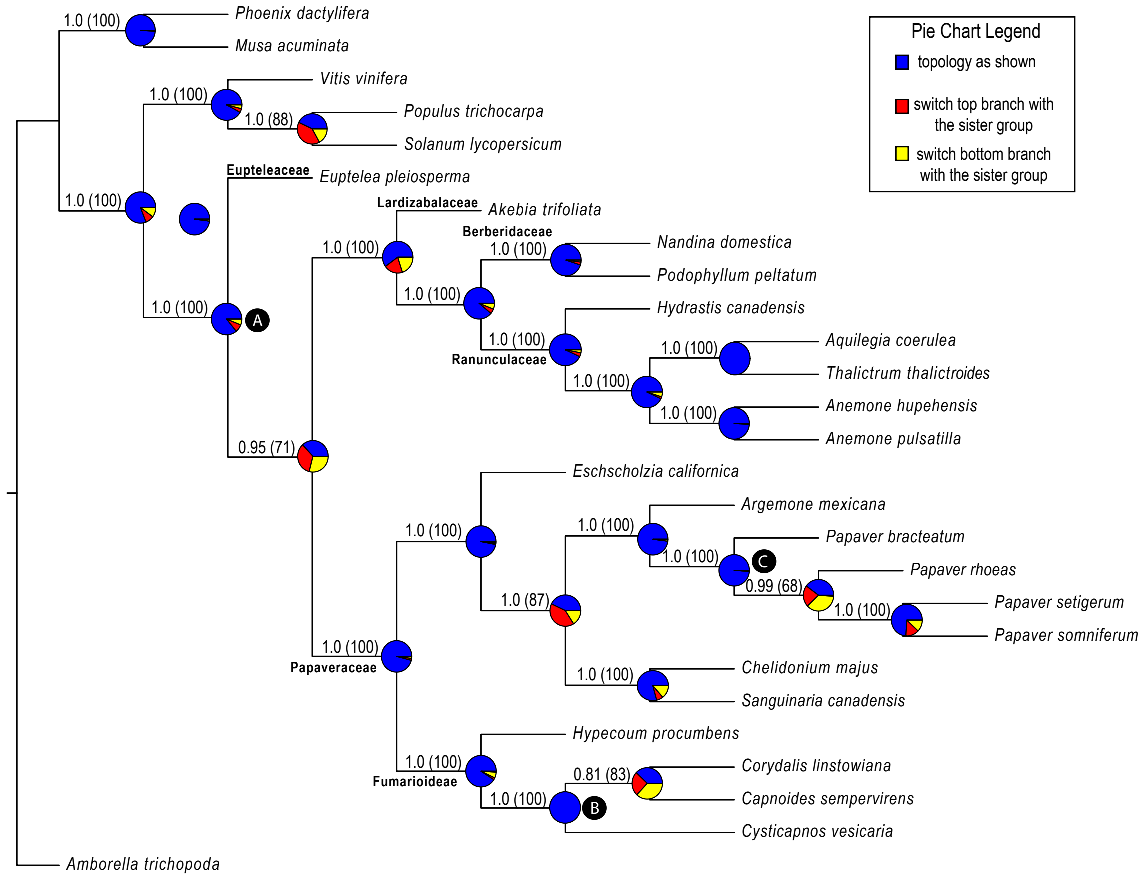


Figure S2.1: ASTRAL Species-Tree from 882 gene trees estimated from nucleic acid alignments. Bootstrap values are in parentheses and local posterior probabilities are adjacent. Pie charts depict the percentage of quartets from all gene trees that support one of three topologies: Q1 (blue) is the topology as shown, Q2 (red) is the lower child with the sister group and the upper child with the outgroup, Q3 (yellow) is the upper child with the sister group and the lower child with the outgroup (Mirarab & Warnow, 2015). Letters correspond to nodes discussed in the other figures. Branches are annotated with relevant family and tribe names in bold.

APPENDIX B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

Tables

Table S3.1: Number of transcripts after each filtering step.

Initial number of transcripts	249,363
After TransDecoder*	84,368
After TPM and % Isoform	69,156
Final number of used transcripts for WGCNA	15,000

* indicates transcriptome version used for TransRate analysis

Table S3.2: Metabolites detected via LC-MS/MS in all tissue samples. Values are in nmol metabolite/gram of tissue. Zeroes indicate that amounts were undetectable in this method.

Sample[#]	Sanguinarine	Protopine	N-Methylstylopine	Scoulerine	Cheilanthifoline	Stylopine	Reticuline	Thebaine	Codeine	Morphine	Salutaridine	Rhoeadine	Salutaridinol	Berberine
som leaf 1	0	0.02935	0	0	0.00081	0	0.70714	32.22222	2.63244	12.03704	0.04938	0	0	0
som leaf 2	0	0	0	0	0.00094	0	0.03750	9.25000	1.25000	6.00000	0	0	0	0
som leaf 3	0	0	0	0	0.00260	0	0.04750	5.00000	0.62500	7.3000	0	0	0	0
som cap 1	0	0.00514	0	0	0.00071	0	0.88095	68.14815	3.89435	18.05556	0.12543	0	0	0
som cap 2	0	0	0	0.00793	0.00197	0	0.31000	14.25000	1.66667	6.00000	0.01667	0	0.00333	0
som cap 3	0	0	0	0.00333	0.00239	0	0.47500	25.00000	3.91667	20	0.08333	0	0.01000	0
som bud 1	0	0.00813	0	0	0.08458	0	0.15476	28.37037	1.75446	7.45833	0.02420	0	0	0
som bud 2	0	0	0	0.00167	0.07792	0	0.32500	4.50000	0.83333	7.50000	0.01389	0	0	0
som bud 3	0	0	0	0.00333	0.22078	0	0.37500	7.25000	0.58333	4.80000	0.00833	0	0	0
som stem1	0	0.00614	0	0	0.00054	0	0.73095	67.77778	2.88690	13.65741	0.05556	0	0	0
som stem 2	0	0	0	0.01333	0	0	0.47500	19.25000	1.66667	6.00000	0.05556	0	0.00833	0
som stem 3	0	0	0	0.00133	0	0	0.09250	5.00000	0.22500	2.70000	0	0	0	0
som root 1	1.47879	0.51410	0	0	0.00154	0	0.99286	1.29630	1.26042	1.10417	0.01506	0	0	0

som root 2	2.89855	0.23346	0	0.00333	0.00364	0	0.07500	1.50000	1.12500	1.50000	0.01389	0	0	0
som root 3	0.53623	0.15564	0	0.00667	0.00234	0	0.37500	1.00000	0.83333	1.26000	0.02778	0	0	0
set leaf 1	0	4.95238	0	0.05200	0.29692	0.02614	4.64000	0.43590	0.67407	0.82708	0.02586	7.13542	0	0.03636
set leaf 2	0	8.15217	0.42373	0.06667	0.13012	0.09000	0.66667	0.61333	1.42929	0.93333	0.01111	1.33333	0	0.00840
set leaf 3	0	3.26087	0.38136	0.02527	0.06988	0.05000	0.18000	0.26667	0.74242	0.26667	0.00370	1.26667	0	0.00400
set cap 1	0	5.25170	0	0.13600	0.18923	0.02000	12.25333	1.47179	2.27037	0.53422	0.03632	7.34375	0	0
set cap 2	0	4.29348	0.04068	0.02667	0.20482	0.03333	0.39333	0.26667	2.46465	0.40000	0.00370	1.26667	0	0.00400
set cap 3	0	8.15217	0.07924	0.06667	0.24096	0.06000	0.73333	1.00000	3.78788	0.66667	0.00370	1.86667	0	0.00720
set bud 1	0	5.44218	0	0.08400	1.06154	0.03590	9.33333	1.25128	2.46296	0.65489	0.03736	10.10417	0	0
set bud 2	0	5.43478	0.12712	0.06333	0.45783	0.08667	0.78667	2.00000	3.25758	0.49333	0.01852	1.33333	0	0.00940
set bud 3	0	4.89130	0.10593	0.06667	0.93976	0.06667	0.20000	0.52000	1.50000	0.86667	0.00741	1.93333	0	0.00600
set stem 1	0	8.84354	0	0.11200	0.06308	0.03897	22.66667	2.73077	4.07407	1.11553	0.06621	15.88542	0	0
set stem 2	0	5.38043	0.03093	0.01333	0.04819	0.03333	0.25333	0.17867	1.01010	0.78667	0	1.60000	0	0.00600
set stem 3	0	8.15217	0.06356	0.20000	0.07229	0.20000	2.00000	3.20000	3.78788	0.66667	0.03333	1.66667	0	0.01500
set root 1	0.18261	3.64626	0	0.02832	0.01058	0.00804	4.68000	1.10256	0.29556	0	0.02092	0.17188	0	0
set root 2	0.48852	1.52174	0.01610	0.00667	0.00231	0.00667	0.06373	0.32533	0	0	0.00370	0.01253	0	0.00300
set root 3	6.55738	1.03261	0.02924	0.00667	0.00241	0.00500	0.05267	0.18667	0.05051	0	0.00185	0.03147	0	0.00300

# som = *P. somniferum* , set = *P. setigerum*, cap = capsule

Table S3.3: Annotated transcription factors with strong connections to BIA-related genes. Adjacency or connection strength between each morphine- or sanguinarine-related transcript and every other transcript within the same module was ranked. Transcripts in the top 25% were designated as having strong connections with the corresponding morphinian or sanguinarine related transcript. Modules O and P are from *P. somniferum.* Modules n and o are from *P. setigerum.* In cases where the transcription factor is a hub in a module the word "Hub" is included in that module's column before the relevant transcripts. We define hubs as transcripts with total within module connectivity in the top 100 for that module.

Description of transcription factor[$]	**Accession**	**Module O**	**Module P**	**Module o**	**Module n**
Phytophthora infestans T30-4 transcription factor BTF3-like protein (PITG_09626) complete cds	XM_002903199.1	*CODM1, COR, SalAt1, T6ODM*	none	*CheSyn1, CheSyn2, CheSyn3, SalAT1, SalAT2, StySyn*	none
Malus x domestica transcription factor fer-like iron deficiency-induced transcription factor-like	XM_008349361.2	*CODM1, SalAt1*	none	none	none
Vigna angularis AP2-like ethylene-responsive transcription factor BBM1	XM_017561637.1	*CODM1*	none	none	*T6ODM*
Populus euphratica AP2-like ethylene-responsive transcription factor BBM2	XM_011008878.1	*COR, T6ODM*	none	none	*CODM1, CODM2, CODM3, T6ODM*
Solanum tuberosum ethylene-responsive transcription factor ERF096-like	XM_006351588.1	*COR, T6ODM*	none	none	none
Nicotiana sylvestris nuclear transcription factor Y subunit B-5-like	XM_009790971.1	*CODM1, SalAt1*	none	none	*CODM1, CODM3*
Nicotiana attenuata transcription factor MYB39-like	XM_019403612.1	*COR*	none	none	none
Phoenix dactylifera WRKY transcription factor 22-like	XM_008778118.1	Hub *CODM1, COR, SalAt1, T6ODM*	none	none	Hub *CODM1, CODM2, CODM3, T6ODM*
Camelina sativa AP2-like ethylene-responsive transcription factor AIL6	XM_010424533.2	*COR*	none	none	none
Cucumis melo transcription factor RAX3-like	XM_008446289.2	*CODM1, SalAt1*	none	*CheSyn1, CheSyn2, CheSyn3, SalAT1, SalAT2, StySyn*	none
Lupinus angustifolius probable WRKY transcription factor 61	XM_019582581.1	*CODM1, SalAt1*	none	*CheSyn1, CheSyn2, CheSyn3, SalAT1, SalAT2, StySyn*	none
Malus x domestica transcription factor MYB36-like	XM_008355036.2	*SalAt1*	none	none	*CODM2, CODM3, T6ODM*
Sesamum indicum transcription factor TGA9 transcript variant X3	XM_020696540.1	*T6ODM*	none	none	none

Nelumbo nucifera MADS-box transcription factor 23-like transcript variant X3	XM_010246397.2	*CODM1, SalAt1*	Hub *CheSyn, CheSyn2, CheSyn3, CODM2, CODM3, SalAt2*	none	*CODM1, CODM2, CODM3*
Solanum lycopersicum transcription factor bHLH18-like	XM_004230610.3	none	*CheSyn, CheSyn2, CODM2, CODM3, StySyn*	*CheSyn1, CheSyn2, SalAT1, StySyn*	*CODM1*
Gossypium arboreum transcription factor MYB24-like	XM_017777755.1	none	*CheSyn, CheSyn2, CheSyn3, CODM2, CODM3, SalAt2*	none	none
Cicer arietinum transcription factor WER-like transcript variant X1	XM_004495198.2	Hub none	Hub *CheSyn, CheSyn2, CheSyn3, CODM2, CODM3, SalAt2, StySyn*	none	Hub *CODM1, CODM2, CODM3, T6ODM*
Oryza sativa Japonica Group probable WRKY transcription factor 30	XM_015795764.1	none	*CheSyn, CheSyn2, CODM2, CODM3, StySyn*	none	none
Durio zibethinus transcription factor MYB14-like	XM_022914131.1	none	*CheSyn3, CODM2, CODM3, StySyn*	*CheSyn1, CheSyn2, CheSyn3, SalAT1, SalAT2, StySyn*	none
Pyrus x bretschneideri transcription factor MYB108	XM_009364828.2	none	*CODM2, CODM3*	none	*CODM1, CODM2, CODM3*
Nicotiana attenuata transcription factor MYB108-like	XM_019389729.1	none	*CheSyn3, CODM2, CODM3, SalAt2*	none	*CODM1*
Cucumis melo transcription factor JUNGBRUNNEN 1-like	XM_008458111.2	none	*CheSyn, CheSyn2, StySyn*	none	none
Papaver somniferum WRKY transcription factor 068_h09 partial cds	JN982462.1	none	*CheSyn, CheSyn2, CheSyn3, CODM2, CODM3*	*CheSyn1, CheSyn2, CheSyn3, SalAT1, SalAT2, StySyn*	none
Ananas comosus probable WRKY transcription factor 50	XM_020242336.1	none	*CheSyn, SalAt2*	none	none
Carica papaya probable WRKY transcription factor 25	XM_022054168.1	none	*CheSyn3, CODM2, CODM3*	none	none
Populus euphratica probable WRKY transcription factor 25	XM_011022705.1	none	*CODM3*	none	none

Brassica napus probable WRKY transcription factor 33	XM_013889648.2	none	*CheSyn3, CODM2, CODM3, SalAt2*	none	none
Phoenix dactylifera transcription factor HY5-like	XM_008809709.2	none	*CODM2, CODM3*	none	none
Hyacinthus orientalis bZIP family transcription factor partial cds	AY389637.1	none	*CODM2, CODM3*	none	none
Papaver somniferum WRKY transcription factor 80_f08 partial cds	JN982457.1	none	*CheSyn3, CODM2, CODM3, SalAt2*	none	none
Medicago truncatula BHLH transcription factor-like protein	XM_003593355.2	none	*CheSyn3, SalAt2*	none	none
Papaver somniferum WRKY transcription factor 58_f10 complete cds	JN982448.1	none	*CODM2, CODM3, SalAt2*	none	none
Prunus persica transcription factor MYB59	XM_007206201.2	none	*CheSyn3*	none	*T6ODM*
Tarenaya hassleriana transcription factor MYB59 transcript variant X4	XM_010537225.2	none	*CheSyn3*	none	*CODM1, T6ODM*
Nelumbo nucifera transcription factor MYB59-like transcript variant X2	XM_010261759.2	none	*CheSyn3, StySyn*	none	none
Tarenaya hassleriana putative Myb family transcription factor At1g14600	XM_010531534.2	none	*CheSyn, CheSyn2, StySyn*	none	none
Capsicum annuum heat stress transcription factor B-4b-like partial	XM_016685139.1	none	*StySyn*	none	none
Gossypium arboreum transcription factor RAX2-like	XM_017789779.1	none	*CheSyn, CheSyn2*	none	none
Papaver somniferum clone 3c WRKY transcription factor (WRKY) complete cds	JQ775582.1	none	*CheSyn3, CODM2, CODM3, SalAt2*	none	none
Nicotiana tomentosiformis probable WRKY transcription factor 31	XM_009616581.2	none	*CheSyn, CheSyn2, CheSyn3*	SalAT2	none
Beta vulgaris subsp. vulgaris probable WRKY transcription factor 14	XM_010667324.1	none	*CheSyn, CheSyn2, CheSyn3, CODM2, CODM3*	*CheSyn1, CheSyn2, CheSyn3, SalAT1, StySyn*	none
Vitis vinifera transcription factor bHLH112 transcript variant X1	XM_003635136.3	none	*StySyn*	*CheSyn1, CheSyn2, SalAT1, StySyn*	none
Papaver somniferum WRKY transcription factor 7_h10 partial cds	JN982454.1	none	*CheSyn3, StySyn*	none	none
Medicago truncatula MADS-box transcription factor	XM_003592305.2	none	*CODM3*	none	none
Torenia fournieri TfSEP3.1 for MADS-box transcription factor SEP3.1 complete cds	AB863627.1	none	*CODM2, CODM3*	*CheSyn1, CheSyn2, CheSyn3, SalAT1, StySyn*	none

Camelina sativa probable WRKY transcription factor 75	XM_010455061.1	none	Hub *CheSyn, CheSyn2, CheSyn3, CODM2, CODM3, SalAt2, StySyn*	none	none
Nelumbo nucifera B3 domain-containing transcription factor FUS3-like transcript variant X2	XM_010250870.2	none	none	*SalAT2*	none
Daucus carota subsp. sativus MADS-box transcription factor 6-like transcript variant X1	XM_017401616.1	none	none	*CheSyn2, CheSyn3, SalAT1, StySyn*	none
Gossypium raimondii ethylene-responsive transcription factor 1B-like	XM_012593182.1	none	none	*SalAT2*	none
Malus x domestica nuclear transcription factor Y subunit B-4-like	XM_008361186.2	none	none	none	*CODM2, CODM3, T6ODM*
Durio zibethinus transcription factor RAX3-like	XM_022918181.1	none	none	none	*CODM1, CODM2, CODM3*
Durio zibethinus transcription factor MYB48-like	XM_022875202.1	none	none	none	*T6ODM*

$ “Predicted:”, “mRNA” and “LOC” numbers have been removed from descriptions for brevity.

Table S3.4: Annotated Cytochrome P450s present in modules than contain morphine and sanguinarine related transcripts.

P. somniferum modules	*P. setigerum* modules	Annotations[#]	Accessions
D	b	*Nicotiana tabacum* cytochrome P450 84A1-like	NM_001325668.1
D	b	*Carica papaya* cytochrome P450 CYP736A12-like	XM_022047834.1
D	b	*Papaver somniferum* clone contig3 cytochrome P450 complete cds	JN185329.1
D	n	*Prunus mume* cytochrome P450 704B1	XM_008223505.1
D	o	*Hevea brasiliensis* cytochrome P450 86A22-like	XM_021785923.1
E	b	*Phalaenopsis equestris* cytochrome P450 72A15-like	XM_020734254.1
E	b	*Nelumbo nucifera* cytochrome P450 78A7-like	XM_010262808.2
E	b	*Theobroma cacao* cytochrome P450 704B1 transcript variant X2	XM_007045976.2
E	b	*Nelumbo nucifera* cytochrome P450 703A2	XM_010276788.1
E	b	*Nelumbo nucifera* cytochrome P450 CYP72A219-like	XM_010276879.2
F	g & j	*Papaver somniferum* clone contig7 cytochrome P450 complete cds	JN185333.1
F	l	*Papaver somniferum* cytochrome P450 (CYP719A21) gene complete cds	JQ659003.1
F, O & P	a,b,l & o	*Papaver somniferum* cytochrome P450 (CYP82Y1) gene complete cds	JQ659005.1
G	b	*Phalaenopsis equestris* cytochrome P450 86B1-like	XM_020743827.1
M	b	Lupinus angustifolius cytochrome P450 85A transcript variant X1	XM_019587215.1
M	b	*Gossypium hirsutum* cytochrome P450 CYP736A12-like	XM_016816988.1
M	o	*Prunus avium* cytochrome P450 78A5	XM_021954217.1
M	o	*Nelumbo nucifera* cytochrome P450 714C2-like	XM_010279888.2
M & O	b	*Papaver somniferum* clone contig10 cytochrome P450 complete cds	JN185335.1
M & O	e & o	*Papaver somniferum* clone contig6 cytochrome P450 complete cds	JN185332.1
N	b	*Hevea brasiliensis* cytochrome P450 86B1-like	XM_021783162.1
O	a	*Nelumbo nucifera* cytochrome P450 85A-like	XM_010262327.2
O	e	*Helianthus annuus* cytochrome P450 77A2-like	XM_022182664.1
O	h	*Ziziphus jujuba* cytochrome P450 71A1-like	XM_016027022.1
O	j	*Prunus avium* cytochrome P450 704B1	XM_021951784.1
O	l	*Ziziphus jujuba* cytochrome P450 71D8-like	XM_016023408.1
O	n	*Ricinus communis* cytochrome P450 94A2	XM_002520958.2
O	o	*Durio zibethinus* cytochrome P450 CYP82D47-like	XM_022884813.1

O	o	*Nicotiana tabacum* cytochrome P450 94A2-like	NM_001325581.1
O	o	*Manihot esculenta* cytochrome P450 90A1-like	XM_021769166.1
O & P	b & k	*Beta vulgaris subsp. vulgaris* cytochrome P450 90A	XM_010691495.2
O & P	o	*Ziziphus jujuba* cytochrome P450 71D8-like	XM_016023427.1
O & P	o	*Solanum pennellii* cytochrome P450 76A2-like	XM_015233876.1
O & P	o	*Jatropha curcas* cytochrome P450 71A1	XM_012224571.2
P	g	*Hevea brasiliensis* cytochrome P450 94A2-like	XM_021829956.1
P	g	*Daucus carota subsp. sativus* cytochrome P450 71D6-like	XM_017370643.1
P	j	*Chenopodium quinoa* cytochrome P450 71A26-like	XM_021864233.1
P	k	*Cajanus cajan* cytochrome P450 83B1-like	XM_020349379.1
P	l	*Theobroma cacao* cytochrome P450 87A3	XM_018116026.1
P	l	*Ziziphus jujuba* cytochrome P450 87A3	XM_016040789.1
P	n	*Juglans regia* cytochrome P450 CYP72A219-like transcript variant X2	XM_018972592.1
P	n	*Camelina sativa* cytochrome P450 86A1-like	XM_010485192.2
P	n & o	*Daucus carota subsp. sativus* cytochrome P450 71A8-like	XM_017362576.1

# The words “PREDICTED:” “mRNA” and the LOC numbers were removed from annotations

Figures

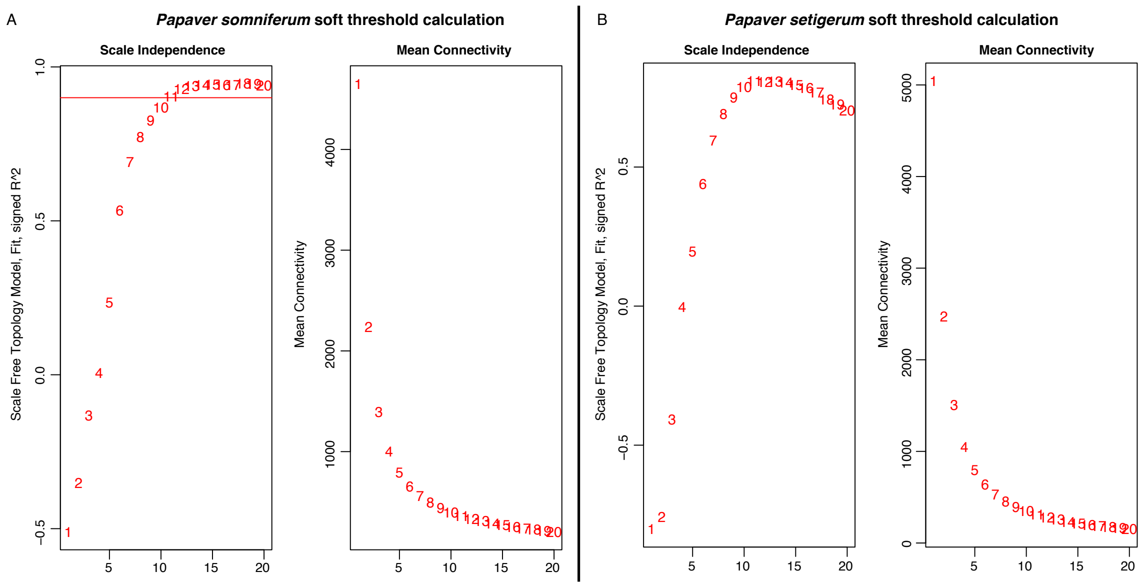


Figure S3.1: **Soft threshold calculation for determining beta for WGCNA.** For each panel the left graph shows scale independence for the scale free topology model with varying beta values. Red line indicates an R^2 of 0.9. Right shows the mean total connectivity for the transcripts given varying beta values. (A) calculations using *P. somniferum* data. (B) calculations using *P. setigerum* data.

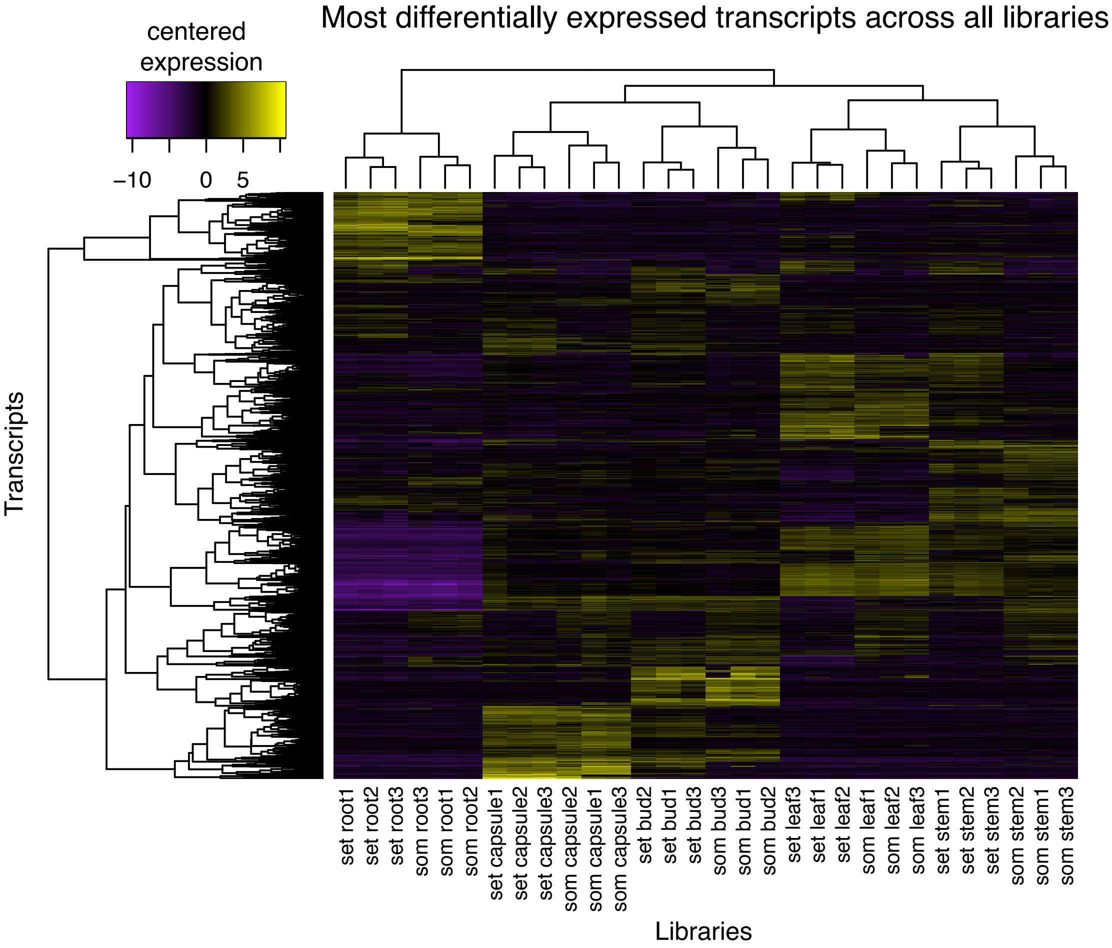


Figure S3.2: Heatmap of differentially expressed transcripts. Heatmap shows centered expression (value – column mean) for all genes 4-fold differentially expressed between at least one sample pair and with a p-value of 1e-3 or less. Left dendrogram is hierarchical clustering of all transcripts. Top dendrogram is hierarchical cluster of RNA-Seq libraries. Set denotes *P. setigerum* library. Som denotes *P. somniferum* library.

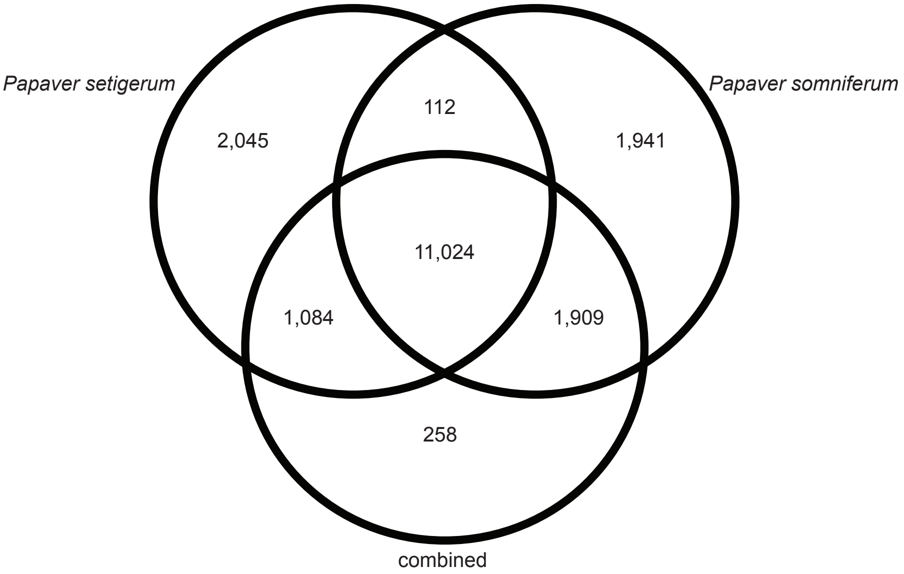


Figure S3.3: Venn diagram of most heterogeneously expressed transcripts determined for each species and when combining data from both species. Numbers indicate transcripts that were determined to be in the top 15,000 most heterogeneously expressed that were unique to each dataset or shared between dataset. Combined data included all libraries from both species.

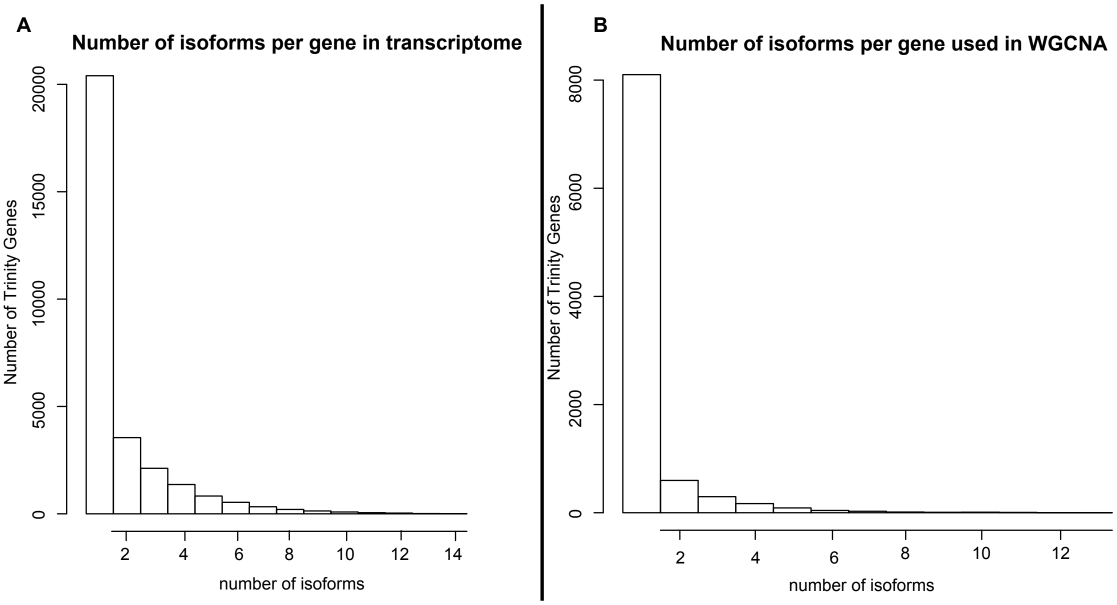


Figure S3.4: **Histograms of number of isoforms per gene included in the whole transcriptome or the transcripts used in the WGCNA.** A) histogram of the number of Trinity isoforms per Trinity gene in the entire *P. somniferum* transcriptome.. B) histogram of the number of Trinity isoforms per Trinity gene represented in the set of 15,000 transcripts used for WGCNA.

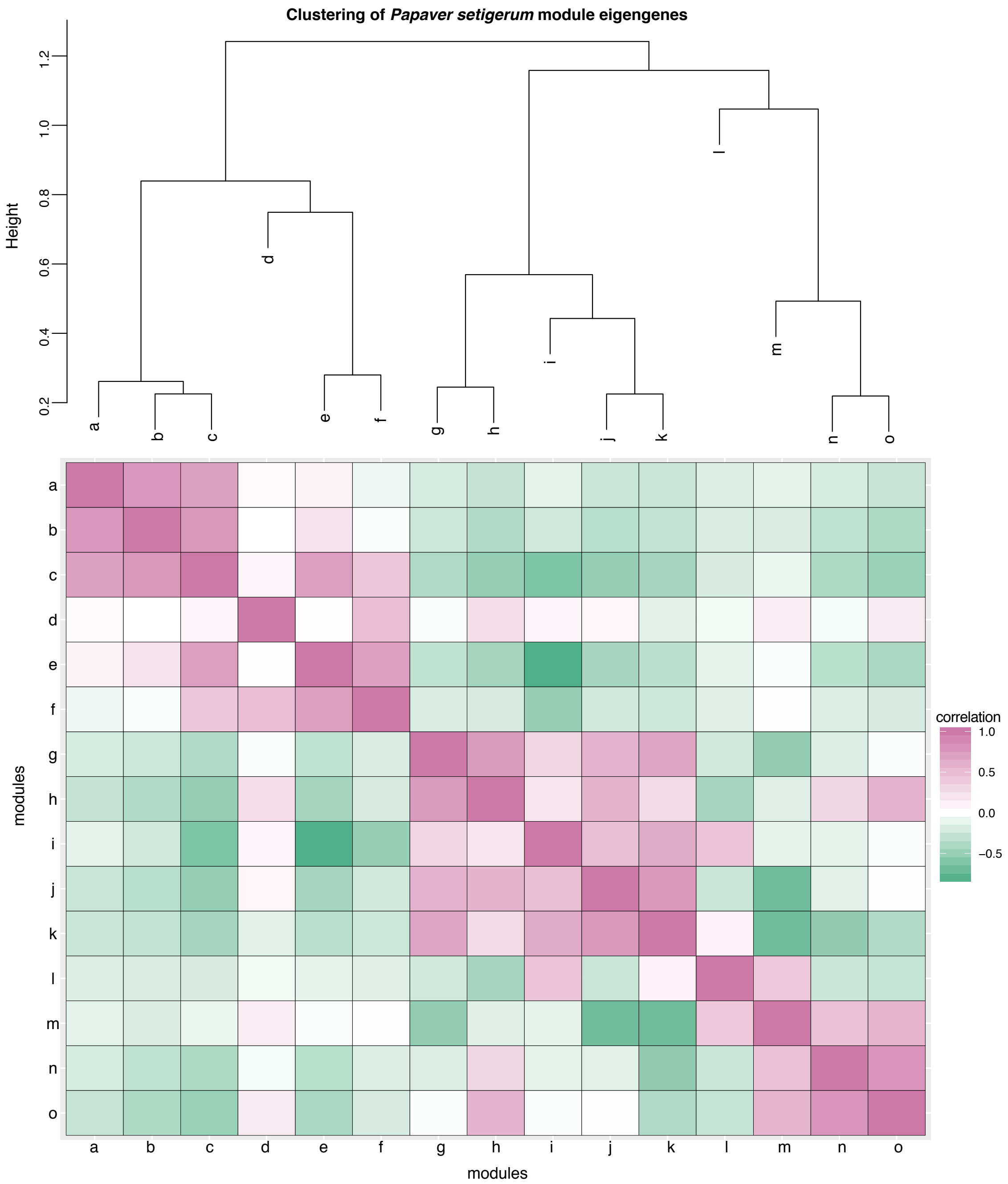


Figure S3.5: Clustering of *P. setigerum* module eigengenes by Pearson correlation. The dendrogram shows the hierchical clustering of module eigengenes. The heatmap is colored by Pearson correlation of the module eigengenes comparing all modules from the *P. setigerum* WGCNA with each other.

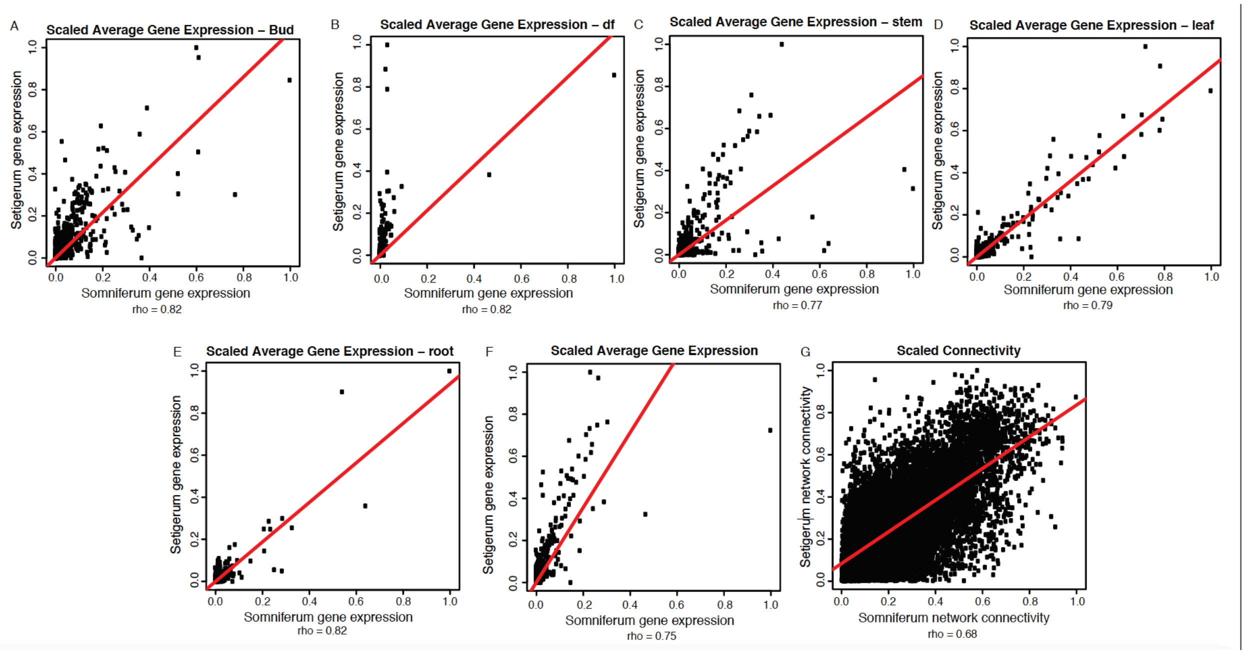


Figure S3.6: Correlations of transcript expression and total network connectivity. Red lines are lines of best fit and Spearman's rho is listed below each graph. For each graph the data is scaled by dividing by the maximum within a species (A-E) Transcript expression is averaged across all libraries from the same tissue for each species. (F) Transcript expression is averaged across all libraries. (G) Total connectivity is the sum of all connection strengths for each transcript.

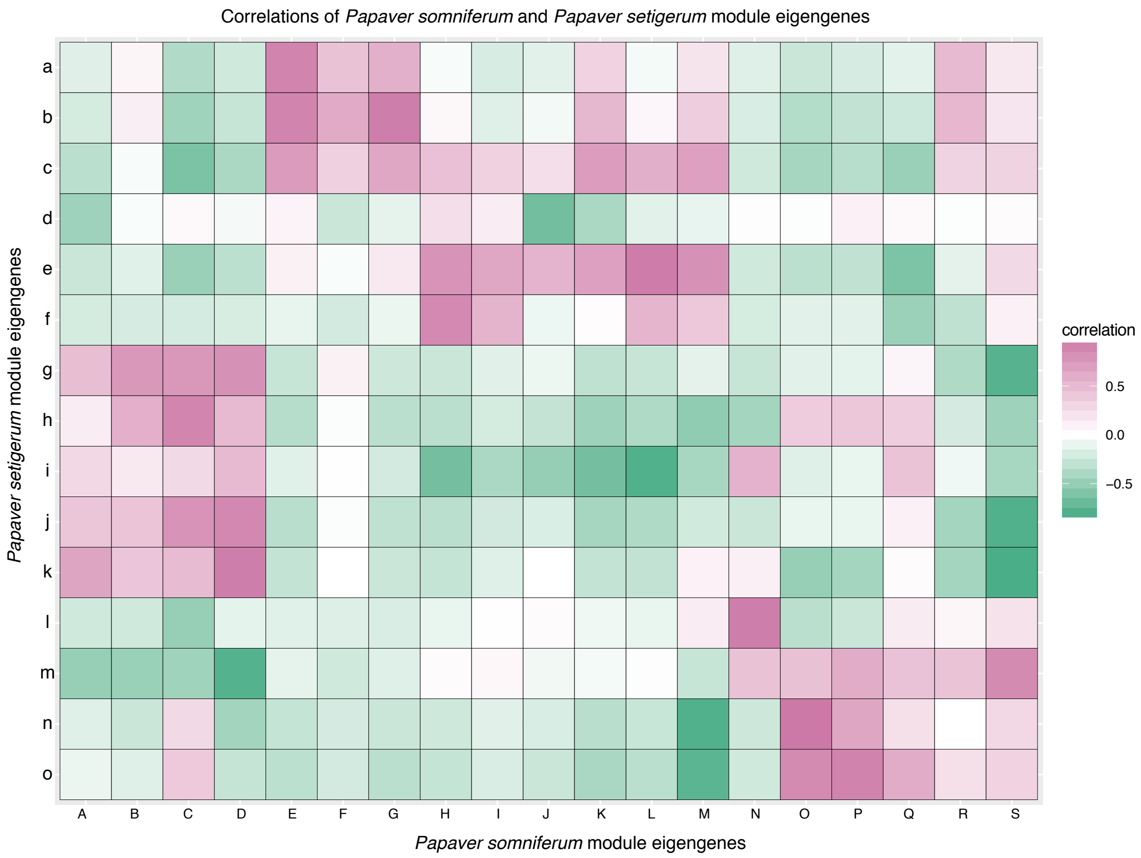
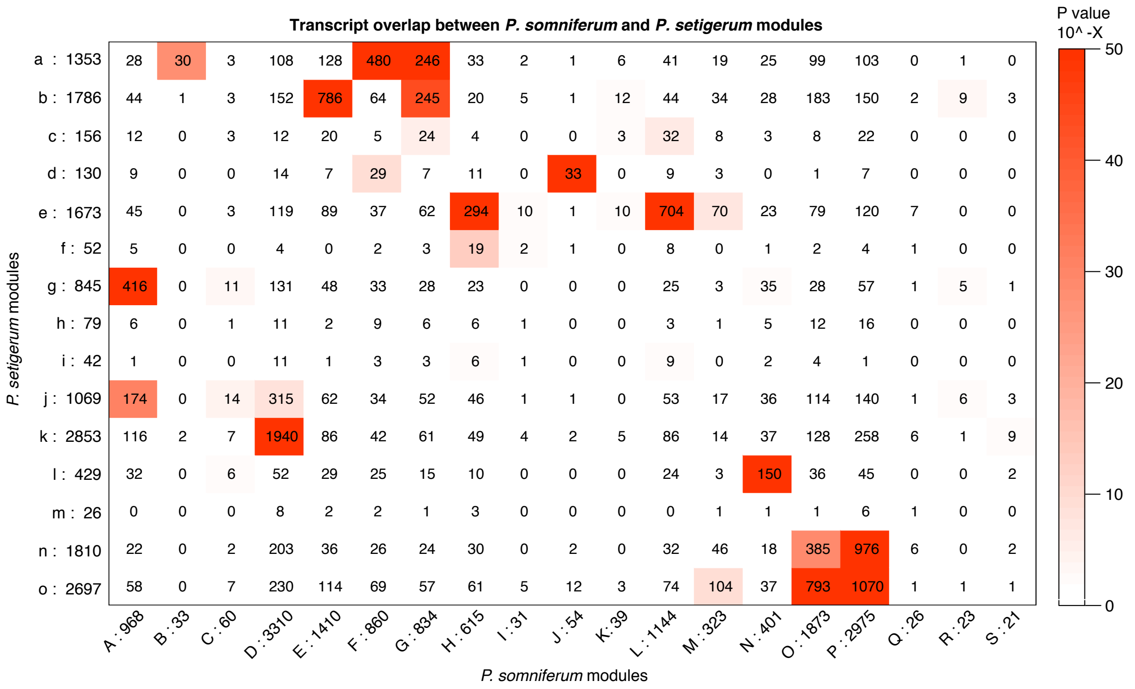


Figure S3.7: Correlations of module eigengenes of each species. *P. setigerum* module eigengenes are on the y-axis and *P. somniferum* on the x-axis. Colors indicate the Pearson correlation coefficient of module eigengenes.



	A : 968	B : 33	C : 60	D : 3310	E : 1410	F : 860	G : 834	H : 615	I : 31	J : 54	K : 39	L : 1144	M : 323	N : 401	O : 1873	P : 2975	Q : 26	R : 23	S : 21
a : 1353	28	30	3	108	128	480	246	33	2	1	6	41	19	25	99	103	0	1	0
b : 1786	44	1	3	152	786	64	245	20	5	1	12	44	34	28	183	150	2	9	3
c : 156	12	0	3	12	20	5	24	4	0	0	3	32	8	3	8	22	0	0	0
d : 130	9	0	0	14	7	29	7	11	0	33	0	9	3	0	1	7	0	0	0
e : 1673	45	0	3	119	89	37	62	294	10	1	10	704	70	23	79	120	7	0	0
f : 52	5	0	0	4	0	2	3	19	2	1	0	8	0	1	2	4	1	0	0
g : 845	416	0	11	131	48	33	28	23	0	0	0	25	3	35	28	57	1	5	1
h : 79	6	0	1	11	2	9	6	6	1	0	0	3	1	5	12	16	0	0	0
i : 42	1	0	0	11	1	3	3	6	1	0	0	9	0	2	4	1	0	0	0
j : 1069	174	0	14	315	62	34	52	46	1	1	0	53	17	36	114	140	1	6	3
k : 2853	116	2	7	1940	86	42	61	49	4	2	5	86	14	37	128	258	6	1	9
l : 429	32	0	6	52	29	25	15	10	0	0	0	24	3	150	36	45	0	0	2
m : 26	0	0	0	8	2	2	1	3	0	0	0	0	1	1	1	6	1	0	0
n : 1810	22	0	2	203	36	26	24	30	0	2	0	32	46	18	385	976	6	0	2
o : 2697	58	0	7	230	114	69	57	61	5	12	3	74	104	37	793	1070	1	1	1

Figure S3.8: Transcript overlap between *P. somniferum* and *P. setigerum* modules. X-axis are *P. somniferum* modules and the y-axis are *P. setigerum* modules. Numbers after semi-colons represent total number of transcripts in that module. Numbers of transcripts shared between two modules are shown in the heatmap. Color represents the –log(p-value) as determined through a one-sided, Fisher's Exact Test.

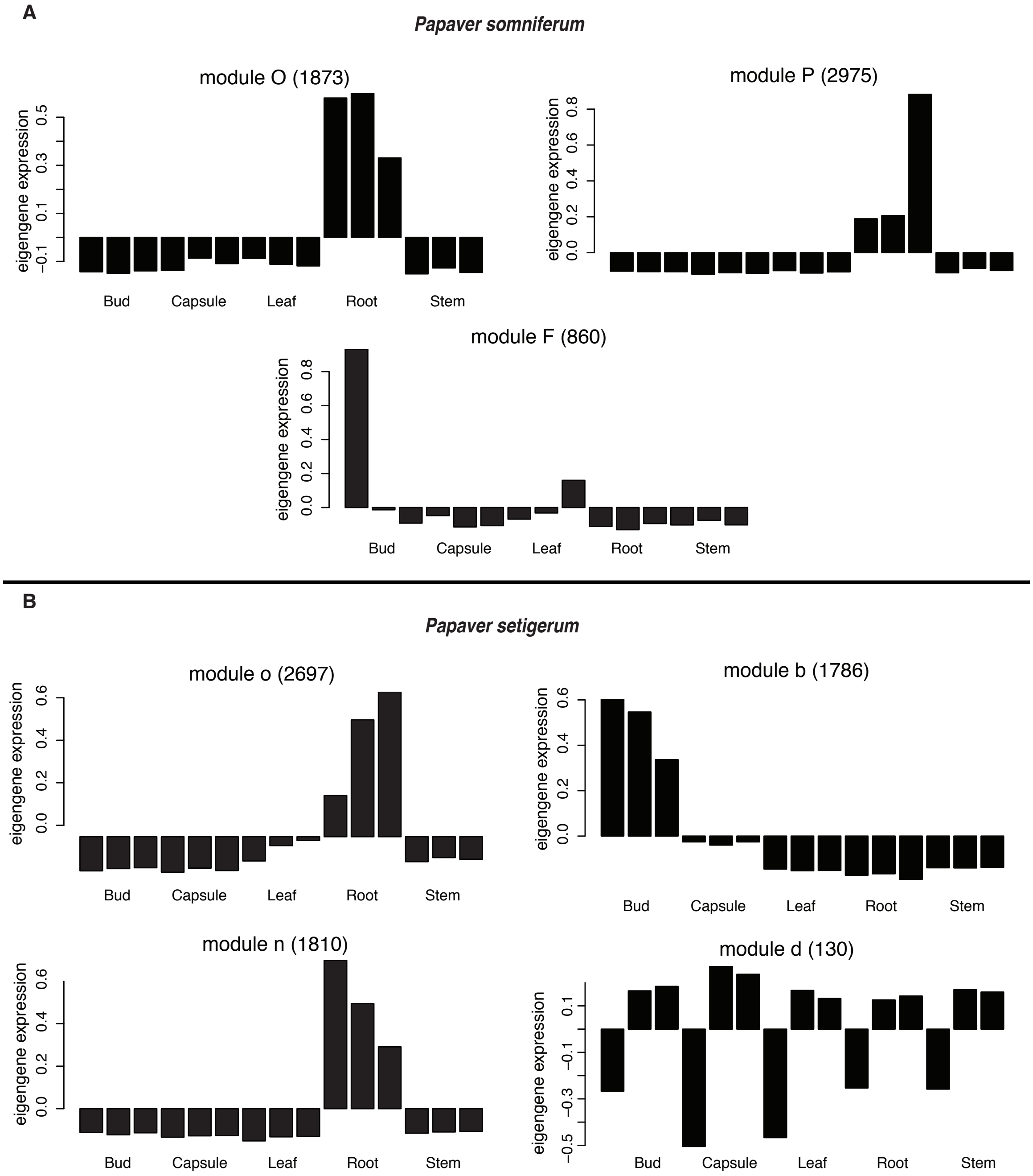


Figure S3.9: Module eigengenes for modules containing morphine & sanguinarine enzymes. (Top) Shown are modules from *P. somniferum* that contain at least one of the annotated enzymes from morphine or sanguinarine biosynthesis. (Bottom) Shown are modules from *P. seitgerum* that contain at least one of the annotated enzymes from morphine or sanguinarine biosynthesis. Libraries are on the y-axis with bars indicating eigengene expression on the y-axis. Module names are shown for each graph with the total number of transcripts per module in parentheses.

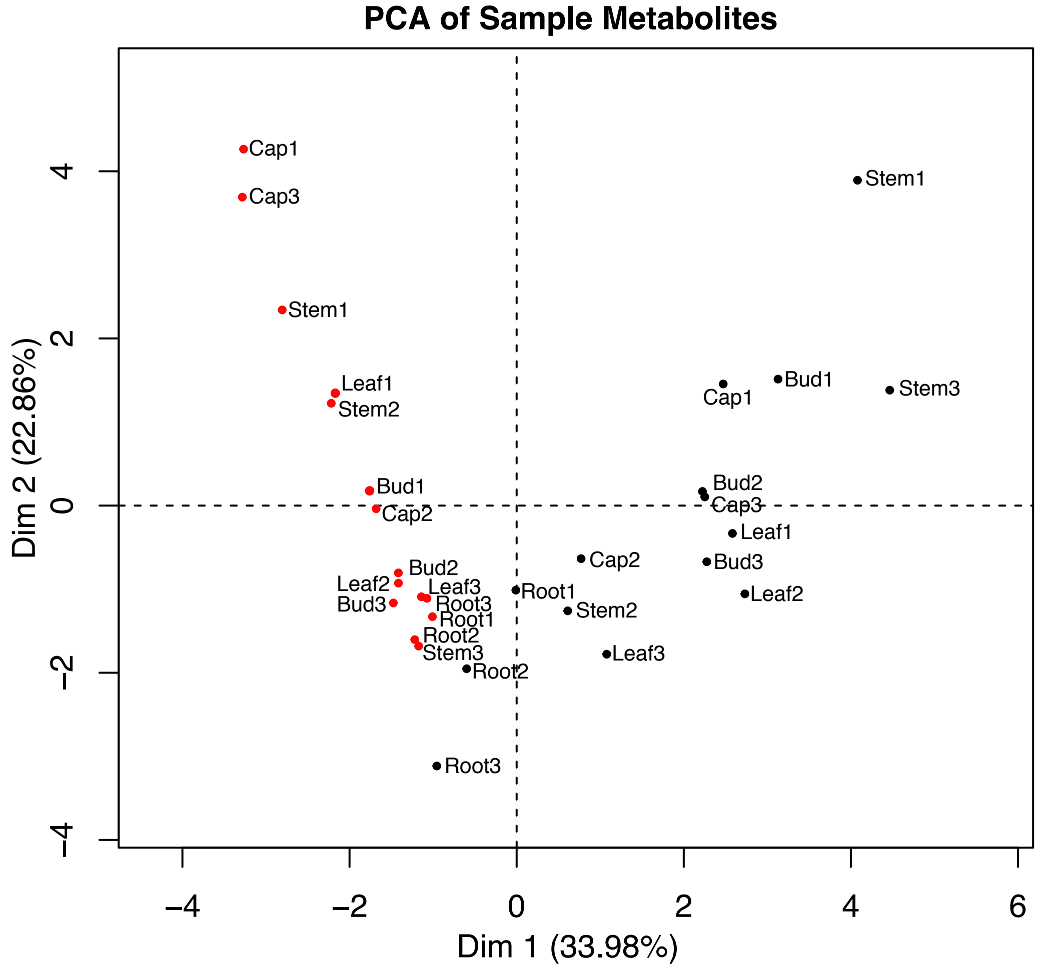


Figure S3.10: PCA of sample metabolites. This is a principal component analysis plot of the metabolite abundances per tissue sample. Axes represent the two dimensions that account for the most variation across samples with percent of variation accounted for in parentheses. Red dots represent *P. somniferum* samples and black dots represent *P. setigerum* samples.

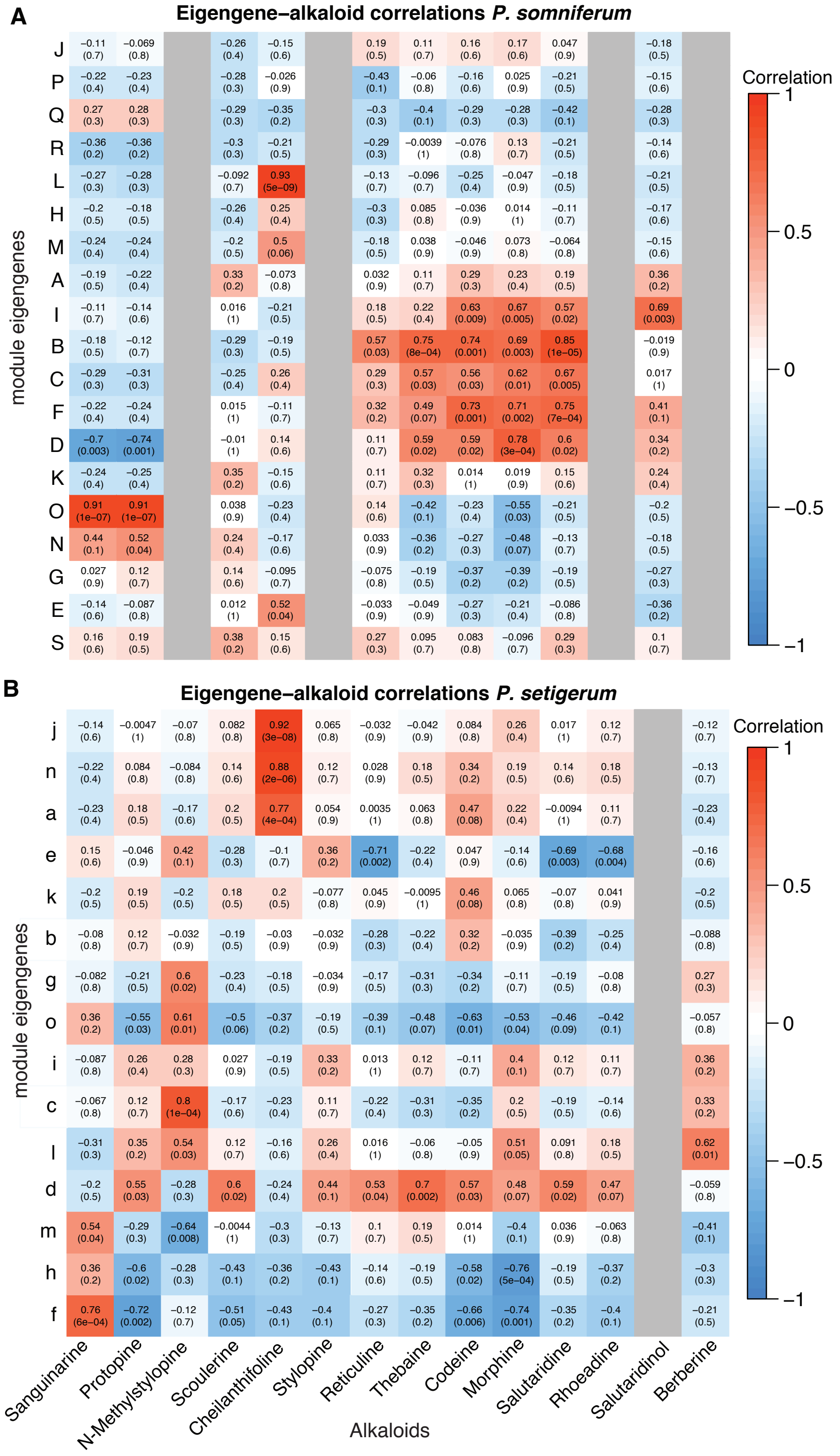


Figure S3.11: **Correlation of metabolite abundance and module eigengenes.** Colors indicate the Pearson Correlation between eigengenes and metabolites across samples. Grey columns indicate metabolites that were not detected in any sample for that species. For each comparison the correlation coefficient is shown with the p-value in parentheses.

module in parentheses. Note that the direction of eigengene expression is not representative of high and low expression due to the unsigned nature of the analysis. Modules are clustered within each species.

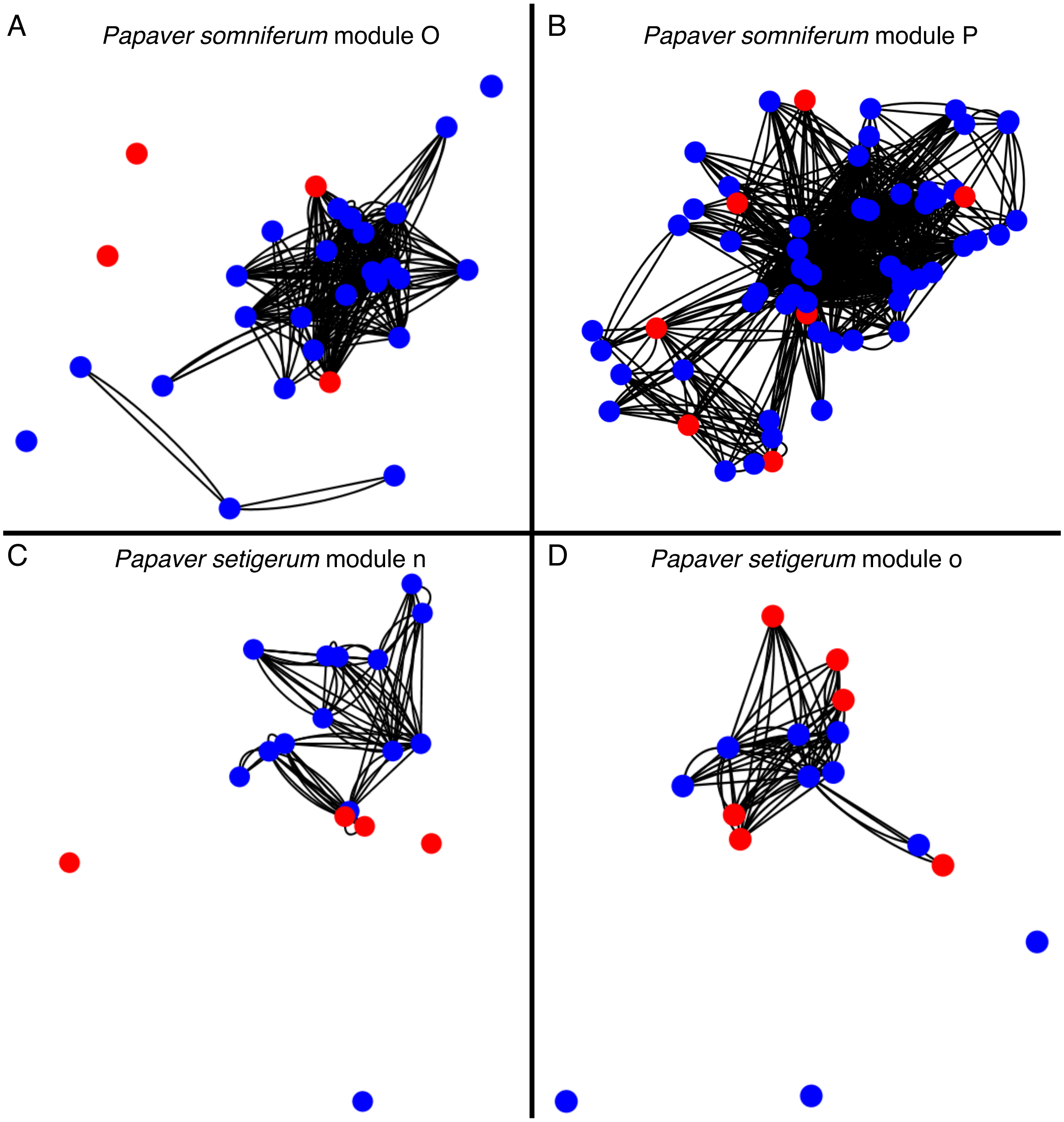

A
Papaver somniferum module O
B
Papaver somniferum module P
C
Papaver setigerum module n
D
Papaver setigerum module o

Figure S3.12: Visualizations of morphine- and sanguinarine-related transcripts and highly co-expressed transcription factors. All modules shown are those that contain the morphine- and sanguinarine-related transcripts from each species. Circles indicate transcripts (nodes) and lines indicate co-expression (edges). Edges are only shown if they represent an adjacency of at least 0.8. Red nodes represent morphine- and sanguinarine-related transcripts and blue nodes represent transcripts annotated as transcription factors that are highly correlated with at least one morphine- or sanguinarine-related transcript in at least one module from either species (also listed in Table S3.3). Many other transcripts are in these modules, but they are not shown in these diagrams. Panels A and B are modules from *P. somniferum* that were further described in the text. Panels C and D are modules from *P. setigerum* that are also further described in the text.

APPENDIX C

SUPPLEMENTAL MATERIALS FOR CHAPTER 3

Table S4.1. Background and characteristics of participants ordered by teaching identity and then department.

Pseudo-nym	PhD Year	Depart ment	Gen der	Citizenship	Teaching identity	# semesters as GTA[a]	Career aspirations[b]	Teachin g PD[c]
Amit	2	A	M	international	none	0	industry	-
Bhavna	4	A	F	international	none	6	industry	-
Emma	2	A	F	US	none	3	industry	-
Georgia	4	A	F	international	none	8	academic researcher *or* R1 faculty	-
Joshua	7	A	M	international	none	10	industry	-
Julia	4	A	F	international	none	1	government researcher	-
Kayla	7	A	F	US	none	1	industry	-
Liu	5	A	M	international	none	13	government researcher *or* R1 faculty	-
Lucas	4	A	M	international	none	1	industry *or* R1 faculty	-
Priya	3	A	F	international	none	5	unsure	-
Rahul	6	A	M	international	none	13	R1 faculty	-
Carly	4	B	F	US	none	3	unsure	-
Karen	4	B	F	US	none	2	industry	-
Stephen	5	B	M	US	none	2	industry *or* government researcher	-
Grace	1	A	F	US	nascent	0	industry *then* teaching	-
Mahira	2	A	F	international	nascent	3	R1 faculty *or* PUI faculty *or* teaching only	-

Marco	1	A	M	US	nascent	1	academic researcher	-
Nicole	1	A	F	US	nascent	0	R1 faculty	-
Sarah	1	A	F	US	nascent	0	government researcher *then* policy	-
Maria	3	B	F	US	nascent	1	R1 faculty	-
Rebecca	5	C	F	US	nascent	2	government researcher *or* industry	-
Robert	4	A	M	US	salient and stable	5	PUI faculty	FFP, TC
Stephanie	3	A	F	US	salient and stable	7	industry *then* teaching only	TC
Andrew	6	B	M	US	salient and stable	12	PUI faculty	FFP
Anna	4	B	F	US	salient and stable	2	PUI faculty	FFP, TC
Catherine	5	B	F	US	salient and stable	13	PUI faculty	FFP, TC
Ryan	3	B	M	US	salient and stable	1	industry *or* teaching only	TC
Shelby	3	B	F	US	salient and stable	2	PUI faculty	TC
Elizabeth	6	C	F	US	salient and stable	9	PUI faculty	FFP, TC
Justin	5	C	M	US	salient and stable	2	PUI faculty	-
Kelsey	5	C	F	US	salient and stable	13	R1 faculty *or* PUI faculty	TC
Matthew	5	C	M	US	salient and stable	7	CC faculty	TC
Megan	6	D	F	US	salient and stable	2	PUI faculty *or* CC faculty	-

[a]GTA = graduate teaching assistant

[b] In cases where the student expressed multiple possible career aspirations, we listed more than one. Unsure indicates they are considering many alternatives. *or* indicates that both options are being considered for their next position. *then* indicates that they plan to pursue one career and then switch to another in the distant future. Teaching indicates a full-time teaching position with institution type unspecified. PUI faculty = faculty member at a primarily undergraduate institution. CC faculty = faculty at a community college

Interview Protocol
Questions added for participants who expressed interest in college teaching - BLUE

Q1. Why did you come to graduate school?
Q2. What do you hope to accomplish as a graduate student?
Q3. What lab are you in?
a. Why did you choose to join that lab?
Q4. What accomplishments would you say are most valued in your lab?
a. Based on your interactions with others, how do you think that is similar or different from other labs in the department?
b. In what ways are those accomplishments acknowledged?
Q5. What do you want to do after you graduate?
Q6. What appeals to you about a position as an ___________?
Q7. What made you want to become ____________?
Q8. How did you communicate your teaching interests to your PI?
a. How was that received?
Q9. What training or experiences do you see as important for someone who wants a career that includes college teaching?
Q10. What, if anything, has caused you to question your career path?
Q11. Have you had any interactions with faculty or fellow students that have made you feel like pursuing teaching is not as respected as pursuing research?
a. What makes you think that?
b. How did that make you feel?
Q12. For you personally, what does it mean to be a scientist?
Q13. Do you see yourself as a scientist?
a. Why or why not?
b. At what point did you start to see yourself as a scientist?
c. OR What do you think would help you to see yourself as a scientist?
Q14. From your perspective, what does it mean to be a teacher?
Q15. Do you see yourself as teacher?
a. Why or why not?
b. At what point did you come to see yourself this way?
c. OR What do you think would help you to see yourself as a teacher?
Q16. Do you see yourself more as a teacher or a scientist? Why?
a. How do you see that changing as you graduate and take a job as a ___________? (do you want it to change?)
b. Reflecting a bit on your path, why do you think you identify more as a ________ and less as a ___________.
c. Has graduate school affected how you see yourself as a researcher and teacher? DON'T ask of new students.
Q17. What do you think it takes to become a good college teacher (focus on classroom teaching)?
Q18. What experiences and people have helped (will help) you develop as a college teacher?
a. Why? How?
b. What other experiences would you like to have?

c. Will you pursue that?
d. How?

Q19. UGA has a teaching certificate program to prepare people to be college teachers. Do you know about that?

a. Do you think you will participate in it?
b. If you were to participate, what do you expect to gain from it?
c. What might you lack if you don't participate in it?

Q20. Your advisor has responsibilities as both a teacher and a research. How does he/she see these dual roles?

a. Is it looked upon favorably for a student in your lab to dedicate time to being a good teacher?

b. Is it looked upon favorably for a student in your lab to prioritize teaching over research?

c. What if that person wants a job as a teacher in the future?

Q21. Do you think your PI would be as proud of a student who took a teaching job as s/he would be of a student who took a research position?

Q22. What relationships, positive or negative, have been most influential to you as a graduate student?

a. What does ________ provide for you?

Pause to read the participant a statement: It has been proposed that the professional culture of science considers teaching to be lower status than research. It may even be the case that faculty who want to be perceived as successful scientists have purposely avoided integrating teaching into their identities as professionals, because they feel it could undermine their status with colleagues and their department.

Q23. What do you think about that?

Q24. Does that align with your experiences?

Q25. My last question for you is: what graduate students do you know who are interested in having college teaching as part of their career?
