GENOME EVOLUTION AND SPECIALIZED METABOLIC GENE INNOVATION IN THE MEDICINAL PLANT *LITHOSPERMUM ERYTHRORHIZON* AND THE TOXIC ALGA *PRYMNESIUM PARVUM*

by

Robert P. Auber

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Biochemistry

West Lafayette, Indiana

May 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Jennifer H. Wisecaver, Chair

Department of Biochemistry

Dr. Clint Chapple

Department of Biochemistry

Dr. Brian P. Dilkes

Department of Biochemistry

Dr. Joshua R. Widhalm

Department of Horticulture and Landscape Architecture

Approved by:

Dr. Andrew D. Mesecar

Dedicated to my mentors, friends, and family

ACKNOWLEDGMENTS

I would like to thank my research advisor Dr. Jen Wisecaver for her patience and guidance during my development as a scientist. I also would like to acknowledge my advisory committee members Dr. Clint Chapple, Dr. Brian Dilkes, and Dr. Josh Widhalm for their support throughout this process as well as their advice for my future direction.

I would like to thank all my fellow laboratory members for providing insight and support towards my dissertation.

I would like to thank the Purdue RCAC for their support in providing the computational resources necessary to conduct my research.

I would like to thank my collaborators Thiti Suttiyut and Dr. Josh Widhalm for their work with the shikonin project. The great effort, discussion, and communication they've provided have made working on the project an exciting and rewarding experience.

I would like to thank all the contributors of my work on *Prymnesium parvum*, including Will Driscoll, Greg Southard, and Olga Yurchenko.

Lastly, I would like to thank my friends and family for their support and their understanding of the commitments of my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Specialized metabolites are chemical tools produced by organisms to aid in their interaction with the surrounding environment. These diverse compounds can often function as metabolic weapons (*e.g.* antibiotics), structural components (*e.g.* lignins), or even attractants (*e.g.* flavonoids). Because of their frequent utilization in niche environments, specialized metabolite production is often lineage- or even species-specific. Therefore, knowledge between specialized metabolic systems is often nontransferable, which poses a major obstacle in the characterization of these bioactive and commercially relevant compounds. Beyond resolving the chemical composition of a specialized metabolite, the identification of responsible pathway genes and the evolutionary processes responsible for their formation is an arduous task. These gaps in knowledge are further widened by the lack of genomic resources available for specialized metabolite producing species. In this work, we present the genome assemblies of two organisms, each with unique specialized metabolic pathways: the Chinese medicinal plant *Lithospermum erythrorhizon* and the toxic golden alga *Prymnesium parvum.* Leveraging the predicted proteome of *L. erythrorhizon*, we investigated the evolutionary history of specialized metabolic genes responsible for the production of shikonin, a 1,4-naphthoquinone specialized metabolite. We identified a retrotransposition-mediated duplication event responsible for the creation of the core shikonin biosynthesis gene, *PGT*. In addition, we performed a global coexpression network analysis to identify regulatory and enzymatic gene candidates involved in the shikonin biosynthesis pathway. We also built phylogenetic trees of known and candidate shikonin genes to reveal patterns of lineage-specific gene duplication and retroduplication. Like plants, unicellular algae are known for their production of diverse, often toxic, specialized metabolites. However, these species are often enigmatic. For example, previous studies have documented large phenotypic variation in both toxin chemotypes and levels among different strains of *P. parvum*. To investigate the genetic basis of this variation, we generated near chromosome level assemblies of two *P. parvum* strains and performed a broad genome survey of thirteen additional strains. As a result, we identified a commonly studied reference strain, UTEX 2797, as a hybrid with two distinct subgenomes. We also provide evidence of significant variation in haploid genome size across the species. Collectively, these studies supply genetic resources for the future study of these organisms, as well as provide insight into the evolution of their specialized metabolic pathways.

CHAPTER 1. INTRODUCTION

1.1 Evolution of the genome

1.1.1 Mechanisms of gene innovation

Gene innovation is the process whereby a new gene is introduced into a genome and evolves a novel function. Two major mechanisms of gene innovation are horizontal gene transfer and gene duplication[1].

The transfer of genetic material from one species into another through non-sexual processes is referred to as horizontal gene transfer (HGT). HGT is accepted as a common mode of gene transfer in prokaryotes, and also in eukaryotes, but to a lesser extent[2]. Within eukaryotes, fungal lineages appear to harbor the most instances of observable HGT, though the sampling bias of fungal genomes relative to other lineages (particularly microbial, i.e., protist lineages) may contribute to this finding[2]. Regardless of lineage, HGT events can confer immediate evolutionary benefits and often offer ecologically-related advantages by introducing genes novel to the recipient[3].

Gene duplication has been postulated as a major mechanism in producing novel genetic material since the work of Susumu Ohno in 1970[4]. Duplication events can vary dramatically in terms of the portion of the genome affected, from fragments of a single gene (e.g. exon duplication[5]) to the entire genome (i.e. whole genome duplication or WGD). WGDs can be the result of autopolyploidization, in which multiple copies of the same genome are created and maintained, as well as allopolyploidization, in which two distinct species are hybridized together to create two subgenomes. The effect of genome doubling *per se* compared to the impact of merging pre-existing genetic diversity via hybridization in polyploids remains unclear[6,7]. Hundreds of whole genome duplications (WGDs), or polyploidization events, have been identified in eukaryotes, some dating back to 500 million years ago[8]. Local or tandem duplications are also prevalent mechanisms of gene duplication. Often resulting from unequal crossover events[9], these duplications can create additional gene copies proximal to the genomic region of origin. Tandem duplicates are identifiable in a large proportion of sequenced genomes and are thought to be a major contributor to single gene family expansions[10,11]. Lastly, transposon-mediated gene duplication is also an established source of gene duplication. Class I elements are transposons

capable of retrotranscribing an RNA intermediate and reintegrating it into the genome[12]. Unlike class II elements, which excise and insert through a 'cut-and-paste' mechanism, class I elements produce duplicates of the original element. In addition to creating a copy of the original element itself, class I elements are also capable of retrotranscribing and integrating mRNA transcripts of other unassociated genes[13].

Following HGT or gene duplication, new genes are typically subjected to one of four possible evolutionary outcomes: retention of ancestral function, nonfunctionalization (i.e. pseudogenization), subfunctionalization, or neofunctionalization. In the case of nonfunctionalization, one gene copy accumulates deleterious mutations while the other continues performing the function of the ancestral gene. Subfunctionalization describes an event in which both genes lose partial functionality of the ancestral gene but complement each other to retain the original function. Lastly, neofunctionalization results in the acquisition of a novel function, i.e., innovation.

Different evolutionary pressures dictate the fate of gene duplicates and vary dependent upon the mode of duplication. Gene dosage, i.e., the number of copies of a gene in a genome, is hypothesized as a major factor impacting fate of new genes, as some molecular processes and pathways are more sensitive to changes in gene dosage than others[14,15]. For example, some duplications leading to increased accumulation of gene product may be functionally beneficial[16]. In contrast, imbalances of transcriptional and translational levels in more sensitive systems could lead to instability, such as those of multi-subunit protein complexes[17]. Therefore, gene retention is also thought to be influenced by physiological and metabolic factors.

Duplications impacting the entire genome have different effects on relative gene dosage compared to duplications at the single gene level. By effectively multiplying the entire genome after WGD, gene stoichiometry is still maintained. Therefore, the loss of a WGD-derived gene could initiate an imbalance. This is consistent with the observation of genes involved in dosage sensitive processes are often over-retained after WGDs[15]. It has also been observed in allopolyploids that gene balance can be maintained through dominance in expression or retention of one gene copy over another[14]. Instances of subgenome dominance are often observed in allopolyploids, in which gene retention and expression are biased towards one subgenome over the other[18].

1.1.2 Biochemical outcomes of gene innovation

An organism's metabolism is often informally classified as primary or specialized. Primary metabolism refers to metabolic pathways and processes directly involved in growth, reproduction, and development. Specialized metabolism, on the other hand, consists of processes not absolutely involved in the survival of an organism. Specialized metabolic processes are often unique to specific lineages and thus pathways are not universally conserved across eukaryotes or sometimes within a single species[19]. Compared to animals, specialized metabolism is expanded in plants, fungi, and many protist lineages. The specialized metabolites produced by these lineages can vary considerably in their function; ranging from structural components (*e.g.* lignins) to allelopathic toxins (*e.g.* juglone). Many specialized metabolites function at the interface of an organism and its environment[20].

While primary and specialized metabolism may differ in their essentiality, they share a biochemical connection. Specialized metabolites are often derived from primary metabolite precursors[20], creating metabolic linkages between pathways classified as primary and specialized. Enzymes in specialized metabolic pathways can evolve from enzymes in primary metabolism via changes in regulation, enzyme promiscuity, and protein-protein interaction[21]. These shared metabolic and evolutionary connections blur the distinctions between primary and specialized metabolic pathways.

The relative contribution of tandem gene duplication and WGD in the evolution of specialized metabolic genes is unresolved with different analyses of different species finding different patterns. For example, Chae et al.[22] found specialized metabolic genes were enriched in tandem duplicates and depleted in WGD-derived genes in land plants. In contrast, expansion of specialized metabolic gene families was significantly associated with whole genome duplication events in angiosperms[23].The exploration of the relative contributions between tandem and WGD has only been employed in a small number of species and far less effort has been made to measure the impact of other mechanisms such as retroduplication. Further work identifying the relevant modes of duplication in more species and diverse lineages would provide crucial insight into possible lineage-specific biases.

Examples of gene innovation are abundant in specialized metabolism. Many popular cultivated plant species (*e.g.* strawberry, potato, cabbage) have undergone recent polyploidization events which have resulted in expanded metabolic capacities[24]. Additionally, subsequent

neofunctionalization of *MAM* tandem gene duplicates involved in glucosinolate biosynthesis (a specialized metabolite involved in plant defense in Brassicales) has resulted in significant differences in the substrate specificity and inhibition dynamics of the encoded enzyme[25].

Quantitative changes in the existing metabolic profile of an organism can also facilitate the generation of novel chemical reactions. A change in intracellular localization or regulation of a metabolic process can also yield significant changes in metabolism. For example, the accumulation of betalain in the Caryophyllales was mediated by the evolution of a feedback-insensitive ADH gene, which produced an abundance of betalain precursor[26]. Evolution of gene regulation by trans-acting factors can also facilitate metabolic innovation through expression divergence[27]. For example, members of the WRKY transcription factor family in plants[28] and velvet family of regulatory proteins in fungi[29] have been demonstrated to directly regulate specialized metabolic pathways.

One proposed mechanism responsible for the regulation of specialized metabolism genes is the formation of metabolic gene clusters (MGCs). MGCs are collections of functionally related genes located proximally to each other in the genome. Genes encoding an entire metabolic pathway can be co-located in a genome including enzymatic genes, transcription factors, and transporters[30]. MGCs are frequently observed in fungi and are often transferred between fungal lineages via HGT[31]. MGCs are less common in plants[32], though several plant MGCs have been identified[33]. While the mechanisms leading to the formation of MGCs are not fully understood, the convergent evolution of similar MGCs from unclustered pathways suggests that clustering is evolutionary advantageous[30]. Hypotheses for the benefit of MGCs are that they promote coregulation[34] and co-inheritance via genetic linkage[35]. For example, the toxicity avoidance hypothesis, which states that gene coregulation and genetic linkage of MGCs minimize the impact of toxic intermediates produced by mis-regulation or unbalanced gene loss, respectively[36].

1.2 Genomic signatures of evolution

1.2.1 Synteny

Synteny is a measurement of the conserved order of loci along a chromosome. Comparing synteny is a powerful approach used to identify structural variation between genomes. The use of synteny in genetics predates whole genome sequencing, when genetic mapping was relied upon to

determine the order and linkage between loci[37]. Currently, synteny is integral to comparative genomics, as it is frequently utilized to study changes in genome architecture[38,39]. Genome sequence alignments can identify disruptions in synteny such as the absence of genetic sequence in one region compared to another (indel), the reversal of sequence order (inversion), the relocation of sequence (transposition), or sequence redundancy (duplication). Further, synteny has been a useful tool in estimating the age of WGD events[40] and tracing the evolution of MGCs[41].

1.2.2 Sequence divergence

Comparing the accumulation of observable mutations throughout a genome and their retention via vertical inheritance is used as a proxy of relatedness and divergence. Variation between gene sequences can be used to infer phylogenetic relationships between species. For example, comparisons of variation in the highly variable internal transcribed spacer (ITS) sequence from ribosomal DNA loci between species can be used to estimate species relationships and construct phylogenies[42]. Multi-gene phylogenies, complimentary to synteny analyses, can resolve more phylogenetically complex relationships such as incomplete lineage sorting, allopolyploidy[43] and introgression[44]. Phylogenetics has also assisted in the tracing of a gene's evolutionary history by dating evolutionary events such as gene duplications and HGT.

1.2.3 Homology

Homology, or similarity between genes because of shared ancestry, is frequently used to predict shared functional identity between homologous genes. Homologous genes that have descended from a single gene copy via speciation are denoted as orthologous[45]. Paralogs on the other hand, are homologous gene sets that are derived from a duplication event. The ortholog conjecture is a common assumption that orthologs are more functionally similar than paralogs[46], which has served as the foundation of many functional genome annotation pipelines[47]. This conjecture is based on the reasoning that orthologs are under stronger selective pressures to maintain ancestral function, while paralogs are more likely to develop novel functionality due to relaxed selective pressures. While the validity of this conjecture is being actively tested[48,49], the majority of predicted functional annotations remain based upon sequence homology[50]. Thus, due to the major

contribution of gene duplication, homology-based approaches have limited utility when attempting to precisely classify genes involved in specific specialized metabolic pathways.

1.2.4 Expression modularity

Another common feature of many specialized metabolic pathways is the presence of a rapidly inducible and tightly coordinated regulatory network[27,51,52]. It is hypothesized that the tight transcriptional regulation of specialized metabolic pathways reflects the specific spatial and temporal roles specialized metabolites play in ecological interactions. Gene coexpression can be an effective tool to identify genes in specialized metabolic pathways. Network analysis can be used to model complex interactions within cellular and metabolic systems[53]. Coexpression networks have recovered known specialized metabolic pathway genes as coexpressed modules and successfully predicted novel roles for genes in specialized metabolism based on coexpression associations[54–57]. However, approaches that use a global coexpression network typically rely on hundreds to thousands of expression datasets derived from a diverse set of conditions. Because of the limited taxonomic distribution of such expression datasets, global coexpression networks have not been applied to non-model species. Dataset limitations aside, coexpression network analysis is effective in recovering metabolic pathways regardless of gene clustering or colocalization[22,57]. Therefore, this approach is more applicable for identifying pathway genes compared to MGC mining approaches[58], especially in plants where MGCs are less common[32].

1.3 Methodological considerations of comparative genomics

The greater the number and taxonomic distribution of sequenced genomes and transcriptomes, the stronger the foundation for comparative genomic studies. The ongoing genomics revolution has been facilitated by advancements in sequencing throughput and cost efficiency brought by next generation sequencing (NGS) technologies[59]. Large scale sequencing efforts, such as the 1000 plants initiative (1KP)[60], have increased available genetic resources for a diverse selection of species. However, there remains a strong sampling bias towards model species[61]. This biases our understanding towards biological processes most relevant in model systems. Further, the relatively narrow taxonomic distribution of specialized metabolism genes compounds these limitations, emphasizing the need for capturing under-sampled lineages in these datasets.

The study of specialized metabolic pathways in non-model organisms often requires the generation of sequencing data. *De novo* transcriptome assembly has emerged as a popular approach to identify specialized metabolic pathways for multiple reasons. First, the approach is cheaper than carrying out a genome assembly. Secondly, the data used to construct the transcriptome assembly can be of dual purpose and reused to perform expression analyses. Such repurposing is particularly useful as expression analyses can be used to identify gene candidates involved in specialized metabolic pathways. However, transcriptome assemblies produce notable statistical errors[62] and lack the genomic context of assembled sequences. Such information is critical when studying specialized metabolic genes and their evolution such as duplication state, products of retrotransposition, and gene clustering which are only identifiable from genome assembly information.

The quality of a genomic resource dictates the power of subsequent analyses. The arrival of third generation sequencing technology has greatly facilitated the creation of more quality genomic resources. The longer read lengths produced from these technologies relative to NGS facilitate the improved assembly of formerly unresolvable regions, such as repetitive sequence and haplotypes[63]. These new capabilities, therefore, aid in the identification of key features in specialized metabolic gene innovation such as tandem duplications or polyploid subgenomes.

1.4 References

1 Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res 2010; 20: 1313–1326.

2 Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 2008; 9: 605–618.

3 Keeling PJ. Functional and ecological impacts of horizontal gene transfer in eukaryotes. Curr Opin Genet Dev 2009; 19: 613–619.

4 Ohno S. Evolution by gene duplication. Springer-Verlag: Berlin, 1970.

5 Letunic I, Copley RR, Bork P. Common exon duplication in animals and its role in alternative splicing. Hum Mol Genet 2002; 11: 1561–1567.

6 Doyle JJ, Flagel LE, Paterson AH et al. Evolutionary Genetics of Genome Merger and Doubling in Plants. Annu Rev Genet 2008; 42: 443–461.

7 Parisod C, Holderegger R, Brochmann C. Evolutionary consequences of autopolyploidy. New Phytol 2010; 186: 5–17.

8 Van De Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. Nat Rev Genet 2017; 18: 411–424.

9 Achaz G, Coissac E, Viari A, Netter P. Analysis of intrachromosomal duplications in yeast Saccharomyces cerevisiae: A possible model for their origin. Mol Biol Evol 2000; 17: 1268–1275.

10 Cannon SB, Mitra A, Baumgarten A, Young ND, May G. The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. BMC Plant Biol 2004; 4: 1–21.

11 Osuna-Cruz CM, Bilcke G, Vancaester E et al. The Seminavis robusta genome provides insights into the evolutionary adaptations of benthic diatoms. Nat Commun 2020; 11. doi:10.1038/s41467-020-17191-8.

12 Boeke JD, Garfinkel DJ, Styles CA, Fink GR. Ty elements transpose through an RNA intermediate. Cell 1985; 40: 491–500.

13 Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. Nat Genet 2000; 24: 363–367.

14 Papp B, Pál C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. Nature 2003; 424: 194–197.

15 Edger PP, Pires JC. Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. Chromosom Res 2009; 17: 699–717.

16 Perry GH, Dominy NJ, Claw KG et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet 2007; 39: 1256–1260.

17 Veitia RA, Bottani S, Birchler JA. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. Trends Genet 2008; 24: 390–397.

18 Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M, Wang X. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. Proc Natl Acad Sci U S A 2014; 111: 5283–5288.

19 Mithen R, Raybould AF, Giamoustaris A. Divergent selection for secondary metabolites between wild populations of Brassica oleracea and its implications for plant-herbivore interactions. Heredity (Edinb) 1995; 75: 472–484.

20 Maeda HA. Evolutionary diversification of primary metabolism and its contribution to plant chemical diversity. Front Plant Sci 2019; 10: 1–8.

21 Moghe GD, Last RL. Something old, something new: Conserved enzymes and the evolution of novelty in plant specialized metabolism. Plant Physiol 2015; 169: 1512–1523.

22 Chae L, Kim T, Nilo-Poyanco R, Rhee SY. Genomic signatures of specialized metabolism in plants. Science (80-) 2014; 344: 510–513.

23 Kawai Y, Ono E, Mizutani M. Expansion of specialized metabolism-related superfamily genes via whole genome duplications during angiosperm evolution. Plant Biotechnol 2014; 31: 579–584.

24 Iannicelli J, Guariniello J, Tossi VE et al. The "polyploid effect" in the breeding of aromatic and medicinal species. Sci Hortic (Amsterdam) 2020; 260: 108854.

25 de Kraker JW, Gershenzon J. From amino acid to glucosinolate biosynthesis: Protein sequence changes in the evolution of methylthioalkylmalate synthase in Arabidopsis. Plant Cell 2011; 23: 38–53.

26 Lopez-Nieves S, Yang Y, Timoneda A et al. Relaxation of tyrosine pathway regulation underlies the evolution of betalain pigmentation in Caryophyllales. New Phytol 2018; 217: 896–908.

27 Grotewold E. Plant metabolic diversity: A regulatory perspective. Trends Plant Sci 2005; 10: 57–62.

28 Schluttenhofer C, Yuan L. Regulation of specialized metabolism by WRKY transcription factors. Plant Physiol 2015; 167: 295–306.

29 Bayram Ö, Braus GH. Coordination of secondary metabolism and development in fungi: The velvet family of regulatory proteins. FEMS Microbiol Rev 2012; 36: 1–24.

30 Rokas A, Wisecaver JH, Lind AL. The birth, evolution and death of metabolic gene clusters in fungi. Nat Rev Microbiol 2018. doi:10.1038/s41579-018-0075-3.

31 Wisecaver JH, Rokas A. Fungal metabolic gene clusters-caravans traveling across genomes and environments. Front Microbiol 2015; 6: 1–11.

32 Kliebenstein DJ, Osbourn A. Making new molecules - evolution of pathways for novel metabolites in plants. Curr Opin Plant Biol 2012; 15: 415–423.

33 Nützmann HW, Huang A, Osbourn A. Plant metabolic clusters – from genetics to genomics. New Phytol 2016; 211: 771–789.

34 Hurst LD, Pál C, Lercher MJ. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 2004; 5: 299–310.

35 Wong S, Wolfe KH. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. Nat Genet 2005; 37: 777–782.

36 McGary KL, Slot JC, Rokas A. Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds. Proc Natl Acad Sci 2013; 110: 11481–11486.

37 Morgan T. An attempt to analyze the constitution of the chromosomes on the basis of sex-limited inheritance in Drosophila. J Exp Zool 1911; 11: 365–412.

38 McCouch SR. Genomics and synteny. Plant Physiol 2001; 125: 152–155.

39 Haibao T, E. BJ, Xiyin W, Ray M, Maqsudul A, H. PA. Synteny and Collinearity in Plant Genomes. Science (80-) 2008; 320: 486–488.

40 Kuraku S, Meyer A. Detection and Phylogenetic Assessment of Conserved Synteny Derived from Whole Genome Duplications BT - Evolutionary Genomics: Statistical and Computational Methods, Volume 1. In: Anisimova M (ed). . Humana Press: Totowa, NJ, 2012, pp 385–395.

41 Lind AL, Wisecaver JH, Lameiras C et al. Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. PLoS Biol 2017; 15: 1–26.

42 Baldwin BG, Sanderson MJ, Porter JM et al. The its Region of Nuclear Ribosomal DNA : A Valuable Source of Evidence on Angiosperm Phylogeny Source : Annals of the Missouri Botanical Garden , Vol . 82 , No . 2 (1995), pp . 247-277 Published by : Missouri Botanical Garden Press Stable URL : http://. Ann Missouri Bot Gard 1995; 82: 247–277.

43 Edger PP, McKain MR, Bird KA, VanBuren R. Subgenome assignment in allopolyploids: challenges and future directions. Curr Opin Plant Biol 2018; 42: 76–80.

44 Hibbins MS, Hahn MW. Corrigendum to: Phylogenomic approaches to detecting and characterizing introgression. Genetics 2022; 220. doi:10.1093/genetics/iyab220.

45 Fitch WM. Distinguishing Homologous from Analogous Proteins. Syst Biol 1970; 19: 99–113.

46 Koonin E V. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 2005; 39: 309–338.

47 Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. Protein Molecular Function Prediction by Bayesian Phylogenomics. PLoS Comput Biol 2005; 1: e45.

48 Nehrt NL, Clark WT, Radivojac P, Hahn MW. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol 2011; 7. doi:10.1371/journal.pcbi.1002073.

49 Chen X, Zhang J. The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data. PLoS Comput Biol 2012; 8. doi:10.1371/journal.pcbi.1002784.

50 Rhee SY, Mutwil M. Towards revealing the functions of all genes in plants. Trends Plant Sci 2014; 19: 212–221.

51 Tohge T, Fernie AR. Co-expression and co-responses: within and beyond transcription. Front Plant Sci 2012; 3: 1–6.

52 Hartmann T. From waste products to ecochemicals: Fifty years research of plant secondary metabolism. Phytochemistry 2007; 68: 2831–2846.

53 Barabási AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nat Rev Genet 2004; 5: 101–113.

54 Horan K, Jang C, Bailey-Serres J et al. Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiol 2008; 147: 41–57.

55 Mentzen WI, Wurtele ES. Regulon organization of arabidopsis. BMC Plant Biol 2008; 8: 1–22.

56 Mao L, Van Hemert JL, Dash S, Dickerson JA. Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics 2009; 10: 1–24.

57 Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A. A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. Plant Cell 2017; 29: 944–959.

58 Blin K, Shaw S, Steinke K et al. AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 2019; 47: W81–W87.

59 Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell 2013; 155: 27.

60 Leebens-Mack JH, Barker MS, Carpenter EJ et al. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 2019; 574: 679–685.

61 Sibbald SJ, Archibald JM. More protist genomes needed. Nat Ecol Evol 2017; 1: 145.

62 Freedman AH, Clamp M, Sackton TB. Error, noise and bias in de novo transcriptome assemblies. Mol Ecol Resour 2021; 21: 18–29.

63 Li C, Lin F, An D, Wang W, Huang R. Genome sequencing and assembly by long reads in plants. Genes (Basel) 2018; 9. doi:10.3390/genes9010006.

CHAPTER 2. HYBRID *DE NOVO* GENOME ASSEMBLY OF THE RED GROMWELL (*LITHOSPERMUM ERYTHRORHIZON*) REVEALS EVOLUTIONARY INSIGHT INTO SHIKONIN BIOSYNTHESIS

*Originally published in *Horticulture Research*

Robert P. Auber, Thiti Suttiyut, Rachel M. McCoy, Manoj Ghaste, Joseph W. Crook, Amanda L. Pendleton, Joshua R. Widhalm & Jennifer H. Wisecaver Hybrid de novo genome assembly of red gromwell (*Lithospermum erythrorhizon*) reveals evolutionary insight into shikonin biosynthesis. *Horticulture Research* 7, 82 (2020). https://doi.org/10.1038/s41438-020-0301-9

2.1 Declaration of collaborative work

Robert Auber performed the Oxford Nanopore sequencing, generated the genome and transcriptome assemblies, and performed the gene expression analysis. Robert Auber and Dr. Jennifer Wisecaver performed the gene family analyses and synteny analyses. Dr. Amanda Pendleton assisted with genome-level analyses. Thiti Suttiyut, Rachel McCoy, Manoj Ghaste, and Joseph Crook performed the molecular cloning, RNA-seq experiments, and qPCR. Robert Auber, Dr. Joshua Widhalm, and Dr. Jennifer Wisecaver wrote the manuscript with input from all co-authors. Robert Auber, Thiti Suttiyut, Dr. Joshua Widhalm, and Dr. Jennifer Wisecaver conceived the project and were involved in experimental design.

2.2 Abstract

Lithospermum erythrorhizon (red gromwell; zicao) is a medicinal and economically valuable plant belonging to the Boraginaceae family. Roots from *L. erythrorhizon* have been used for centuries based on the antiviral and wound-healing properties produced from the bioactive compound shikonin and its derivatives. More recently, shikonin, its enantiomer alkannin, and several other shikonin/alkannin derivatives have collectively emerged as valuable natural colorants and as novel drug scaffolds. Despite several transcriptomes and proteomes having been generated from *L. erythrorhizon*, a reference genome is still unavailable. This has limited investigations into elucidating the shikonin/alkannin pathway and on understanding its evolutionary and ecological

significance. In this study, we obtained a de novo genome assembly for *L. erythrorhizon* using a combination of Oxford Nanopore long-read and Illumina short-read sequencing technologies. The resulting genome is ~367.41 Mb long, with a contig N50 size of 314.31 kb and 27,720 predicted protein-coding genes. Using the *L. erythrorhizon* genome, we identified several additional *p*-hydroxybenzoate:geranyltransferase (PGT) homologs and provide insight into their evolutionary history. Phylogenetic analysis of prenyltransferases suggests that PGTs originated in a common ancestor of modern shikonin/alkannin-producing Boraginaceous species, likely from a retrotransposition-derived duplication event of an ancestral prenyltransferase gene. Furthermore, knocking down expression of LePGT1 in *L. erythrorhizon* hairy root lines revealed that LePGT1 is predominantly responsible for shikonin production early in culture establishment. Taken together, the reference genome reported in this study and the provided analysis on the evolutionary origin of the shikonin/alkannin biosynthesis will be useful for guiding the elucidation of the remainder of the pathway.

2.3 Introduction

The purple-colored roots of red gromwell (*Lithospermum erythrorhizon*; Fig. 2-1a,b), also known as "zicao" in Chinese, "jichi" in Korean, and "murasaki" in Japanese, have been used as part of traditional medicines, as a dyestuff, and in cosmetics across many cultures for centuries. The responsible bioactive and pigmented compounds, shikonin–or its enantiomer, alkannin–and dozens of other acylated shikonin/alkannin derivatives (Fig. 2-1c), are synthesized in the root periderm of *L. erythrorhizon* and several other Boraginaceae species[1,2]. Shikonins/alkannins are deposited into the rhizosphere where they function in plant-microbe interactions and interfere with the growth of competing plants (allelopathy), roles suggested to have contributed to the invasion success of species like *Echium plantagineum*[3]. The presence of alkannins was also reported in *Plagiobothrys arizonicus* leaves[4], though the physiological and/or ecological significance of their presence in aerial tissues is unclear.

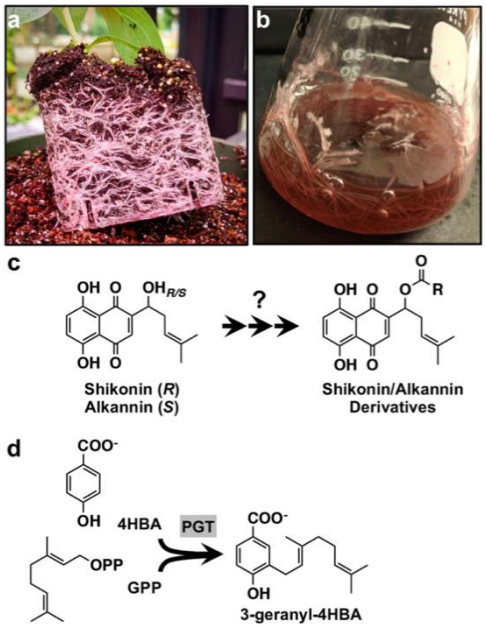

Figure 2-1 Shikonin is produced in the roots of *Lithospermum erythrorhizon*. a Intact roots of *L. erythrorhizon* producing shikonin. b Hairy root culture of *L. erythrorhizon* producing shikonin. c The structures of shikonin and its enantiomer, alkannin. Shikonin and alkannin are precursor to dozens of acylated derivatives collectively produced by members of the Boraginaceae. d The shikonin pathway starts with the conjugation of 4-hydroxybenzoic acid (4HBA) and geranyl diphosphate (GPP) catalyzed by *p*-hydroxybenzoate:geranyltransferase (PGT).

Shikonins/alkannins have more recently been discovered to exhibit a range of pharmacological properties[5]. Shikonin has been found to suppress human immunodeficiency virus (HIV) type 1[6] and to display anti-tumor effects in breast cancer cells via multiple signaling pathways[7]. Thus, combined with their traditional medicinal and cosmetic value, there has been wide interest for many decades in scaling shikonin/alkannin production. Early efforts back in the 1970s and 1980s centered on producing shikonin in *L. erythrorhizon* cell cultures, which was also the first industrial scale platform for producing a secondary metabolite in dedifferentiated plant cells[8]. With advances in understanding of pathway precursors[4,9], strategies to increase shikonin production in *L. erythrorhizon* through metabolic engineering were developed (*e.g.*[10,11]). Current efforts have extended to include synthetic chemistry for producing shikonin, alkannin, and derivatives with higher specificity and potency[12,13]. Moreover, the use of comparative transcriptomics (*e.g.*[14–16]) and proteomics[17] approaches for elucidating the shikonin/alkannin pathway has uncovered several gene candidates as well as to the identification of the geranylhydroquinone hydroxylase (GHQH; CYP76B74)[18]. Comparatively less attention has focused on the evolutionary origin of shikonin/alkannin pathway genes[16,19] and on the ecological

significance of producing the compounds[20]. It was only recently that the *E. plantagineum* genome was published, the first from a shikonin/alkannin-producing species or a member of the Boraginales[21].

Despite being among the longest-studied plant natural products, there are still many gaps in knowledge about shikonin/alkannin pathway genes, architecture, and regulation. In this study, we report the first *de novo* genome for *L. erythrorhizon*, generated by combining Oxford Nanopore Technology (ONT) long reads with Illumina HiSeq short reads. From this assembly, we identified the existence of 11 previously unreported *p*-hydroxybenzoate:geranyltransferase (PGT; Fig. 2-1c) homologs and provide insight into their contribution to shikonin biosynthesis based on (i) their distribution amongst shikonin/alkannin-producing species within the Boraginaceae and (ii) *L. erythrorhizon* knock-down hairy root lines with reduced expression of *LePGT1*. Taken together, the results of our study provide evolutionary insight into the origin of the shikonin/alkannin pathway, and the genome assembly offers a major resource for exploring outstanding questions in shikonin/alkannin metabolism.

2.4 Results

2.4.1 Genome assembly and annotation

To create a reference genome, we combined *L. erythrorhizon* ONT genomic DNA (gDNA) reads generated in-house from Siebold & Zucc. plants with publicly available Illumina gDNA reads sequenced by Nanjing University in 2018 from an unknown accession (SRR5644206). The Illumina data consisted of ~21.7 Gb Illumina HiSeq paired-end short reads (150 bp) with an estimated heterozygosity of 0.39% and projected genome size of 369.34 Mb (Fig. S1a). Our in-house ONT data consisted of ~7.6 Gb long-reads (N50 = 15.03 kb) providing roughly 20-fold genome coverage. The short and long reads assembled into 2465 contigs using the DBG2OLC hybrid assembler[22], yielding a 367.41 Mb genome with a longest contig of 3.44 Mb and an N50 contig length of 314.31 kb (Table 2-1).

Table 2-1 Summary of *L. erythrorhizon* genome assembly and gene models

Genome assembly statistics	
Total length	367,405,101
No. contigs	2,465
Largest contig length	3,439,996
N50 contig length	314,306
N90 contig length	61,630
Counts of N50 (no. contigs)	233
Counts of N90 (no. contigs)	1,370
Genome GC content	35.17%
Gene model statistics	
Gene number	27,720
Gene density (kb/gene)	13.25
Mean gene length	3,772
Avg. no. exons per gene	7
Mean exon length	320
Exon GC content	39.32%

Using a *de novo* repeat modeler, 51.78% of assembly bases were denoted as repetitive elements and were subsequently masked (Table S1). Of these elements, the majority were long terminal repeats (LTRs) which comprised 23.43% of the genome assembly. Unclassified elements were the second most common, accounting for 21% of the genome. DNA repeat elements comprised 4.45% of the genome. The repeat content in *L. erythrorhizon* is comparable to the repeat content observed in the other sequenced Boraginaceae, *E. plantagineum* (43.3% repetitive; 23.08% LTRs)[23].

Protein-coding genes were identified through a combination of *ab initio,* homology-based, and transcriptome-based prediction methods. A total of 27,720 genes encoding 39,395 proteins were predicted (Table 2-1). The average protein-coding gene was 3,772 bp long and contained 7 exons. Functional annotations were assigned to 80.02%, 72.89%, 59.30%, 23.57%, 7.53%, 5.85% of genes using the InterPro[24], Pfam[25], GO[26], Trans Membrane (TMHMM), KEGG, and MetaCyc[27] databases, respectively (Supplementary Dataset S1).

2.4.2 Quality assessment

To evaluate the completeness and coverage of the assembly, we aligned the ONT gDNA, Illumina gDNA, and Illumina RNA reads to the *L. erythrorhizon* genome assembly. Coverage histograms of the ONT and Illumina gDNA reads indicated a single peak at ~18-fold and ~45-fold coverage, respectively (Fig. S1b,c), indicative of the genome being largely homozygous. The alignment rates of the Illumina gDNA reads was high at 95.97%; however 46% of the assembled genome lacked gDNA read support (Fig. S1c). This is likely due to the Illumina gDNA reads being generated via a PCR amplified library, leading to inconsistent coverage across the genome. The amount of coding-regions with Illumina gDNA read support was 89.6% (≥ 10 mapped reads). The alignment rate of the RNA reads ranged from 59.29% - 72.71% (in the case of libraries prepared via ribosomal depletion) to 86.74% - 90.15% (in the case of libraries prepared via polyA capture) (Table S2).

We then used BUSCO[28] to assess the completeness of the predicted proteome. Within the *L. erythrorhizon* protein-coding gene set, 1142 of 1400 conserved embryophyta genes (79.3%) were identified as complete, of those 93.43% were present in single-copy and 6.57% were duplicated (Table S3). Furthermore, 279 of 303 conserved eukaryota genes (92.08%) were identified as complete, of those 78.14% were present in single-copy and 21.86% were duplicated (Table S3). Lastly, we calculated the Alien Index (AI) for all predicted proteins of the genome assembly to assess possible contamination[29]. No assembly contig had a majority of their predicted proteins with AI scores > 0, indicating no detectible contamination in the assembly. Only 50 of the 39,395 total proteins (0.13%) had AI scores > 0.05, which could be indicative of horizontal gene transfer (HGT) (Supplementary Dataset S2). However, manual inspection of each of these proteins did not yield any strong HGT candidates.

2.4.3 Gene family analysis

To investigate the evolution of different gene families, including those involved in the production of shikonin, we performed an OrthoFinder[30] analysis using the protein-coding genes of *L. erythrorhizon* and 31 other eudicot species (Table S4). Incorporated into our analysis were four additional Boraginaceae species, including three species known to produce shikonin (*E. plantagineum, Arnebia euchroma,* and *Lithospermum officinale*) as well as *Mertensia paniculate*,

a Boraginaceae whose transcriptome was sequenced by the 1000 Plants Initiative (oneKP[31]) and whose ability to produce shikonin is unknown. We also included 14 additional Boraginales species sequenced by the oneKP project (Table S4). Our OrthoFinder-inferred species tree had the Boraginales sister to a large clade consisting of the Solanales, Gentianales, and Laminales (Fig. 2-2). This placement is in disagreement with the analysis by Tang et al.[23] that showed the Boraginales sister to the Solanales. This conflict is unsurprising as the evolutionary relationships of these lamiid orders remains uncertain[32]. Additional work is needed to resolve these relationships and determine the source of the phylogenetic discordance.

The OrthoFinder analysis identified 100,874 orthogroups (predicted gene families), of which 24,346 consisted of two or more species in the analysis (Table S5). Of the 14,885 orthogroups containing one or more *L. erythrorhizon* sequences, 3441 orthogroups (23.13%) were present in all species, 3003 (20.17%) were *L. erythrorhizon* specific, and 8441 (56.71%) consisted of *L. erythrorhizon* and one or more additional species (Fig. 2-2; Table S6). In total, 36,392 of 39,395 *L. erythrorhizon* proteins (92.38%) were assigned to an orthogroup containing sequence(s) from one or more additional species (Table S6; Supplementary Dataset S3). The total number of *L. erythrorhizon* orthogroups (14,885) was comparable to the other sequenced genomes in the analysis (average 15,227), which ranged from 12,666 in *Solanum lycopersicum* (tomato) to 19,349 in *Actinidia chinensis* (kiwi). Furthermore, the percentage of *L. erythrorhizon* orthogroups that was species-specific (20.17%) was also comparable to the other sequenced genomes (average 22.47%), ranging from 5.05% in tomato to 46.55% in *Utricularia gibba* (a carnivorous aquatic bladderwort). The other three species known to produce shikonin showed an increased number of orthogroups (average 26,040.33), which is likely a result of their predicted proteomes being derived from *de novo* transcriptome assemblies rather than sequenced genomes. The average percentage of species-specific orthogroups found in the oneKP transcriptomes (0.50%) was noticeably lower that the other species in the analysis (Fig. 2) due to these proteomes being filtered prior to publication[31].

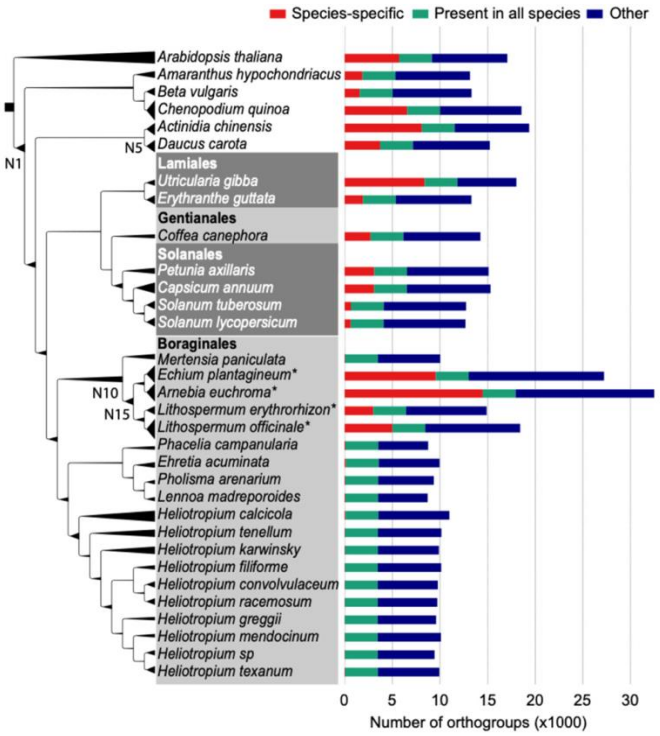

Figure 2-2 OrthoFinder gene family analysis. The OrthoFinder inferred species phylogeny is displayed on the left. The branch thickness is scaled based on the number of predicted duplicated events to have occurred at the descendent node; thinner branches indicate fewer duplications, thicker branches indicate more (see Table S7). Internodes discussed in the text (N1, N5, N10, and N15) are labeled (the species tree with all labeled internodes can be accessed in Table S7). Asterisks (*) indicate known shikonin/alkannin-producing species. Horizontal bar plots (right) indicate the number of orthogroups that are species-specific (red), maintained in all 32 species (green), or present in more than one but less than all species in the analysis (blue).

To identify orthogroups that had expanded in one or more ancestors of *L. erythrorhizon*, we parsed the number of OrthoFinder-predicted gene duplications at each node of the inferred species tree (Fig. 2-2; Table S7). The average number of orthogroups that duplicated one or more times at a given internode (*i.e.* non-leaf node) was 1489.55 and ranged from 41 duplications at internode N5 (the common ancestor of *Daucus carota* and *A. chinensis)* to 4489 duplications at internode N1 (the common ancestor of Caryophyllales and asterids) (Table S7). A total of 2818 orthogroups duplicated at internode N10 (the common ancestor of the five Boraginaceae species), and the *L. erythrorhizon* genes that duplicated at this internode were enriched in 21 Gene Ontology (GO) categories (Benjamini-Hochberg adjusted *p*-value < 0.1; Table S8) including transferase activity (GO:0016740; p = 1.10e-3), signal transduction (GO:0007165; p = 6.45e-4), and transmembrane transporter activity (GO:0022857; p = 0.08). Similarly, the *L. erythrorhizon* genes from the 3908 orthogroups that duplicated at internode N15 (the common ancestor of the four Boraginaceae species known to produce shikonin) were enriched in 14 GO categories (Benjamini-

Hochberg adjusted *p*-value < 0.1; Table S8) including transferring of acyl groups transferase activity (GO:0016746; $p = 0.3$), DNA-binding transcription factor activity (GO:0003700; $p = 3.79e\text{-}3$), and transmembrane transport (GO:0055085; $p = 0.07$).

2.4.4 Evolution of p-hydroxybenzoate:geranyltransferase (PGT) genes for shikonin biosynthesis

One orthogroup predicted to have undergone multiple duplication events in the last common ancestor of shikonin/alkannin-producing species was OG0000509. This orthogroup was comprised of genes that code for prenyltransferases, including the characterized ubiquinone prenyltransferase in *Arabidopsis thaliana,* AtPPT1 (Coq2; At4g23660)[33]. In addition, OG0000509 contained 13 *L. erythrorhizon* genes (Table S9), including Leryth_015068 (hereafter referred to as *LePGT1*), which encodes a protein 97.06% identical to LePGT1, and Leryth_002561 (hereafter referred to as *LePGT2*), which encodes a protein 100% identical to LePGT2[34]. The *p*-hydroxybenzoate:geranyltransferases (PGTs) catalyze the conjugation of 4-hydroxybenzoic acid (4HBA) and geranyl diphosphate (GPP), the first and committed step of the shikonin pathway (Fig. 2-1d)[34–36], and have been identified from several shikonin/alkannin-producing species[19,37–39].

To better understand the evolutionary history of LePGT1 and LePGT2, we constructed a robust phylogenetic tree of OG0000509 prenyltransferases (Fig. 2-3a). Of the 13 sequences in *L. erythrorhizon,* one was excluded from the phylogenetic analysis due to it being a suspected pseudogene (Leryth_015069). This suspected pseudogene is located ~10 kb away from *LePGT1* and was likely the result of a tandem duplication event (Fig. S2). Pseudogenization is suspected due to the fact that the sequence appears truncated and is missing the conserved NDXXD motif indicative of putative prenyl diphosphate binding (Fig. S3)[34]. In agreement with the OrthoFinder analysis, the phylogeny shows a large radiation of prenyltransferase genes in the Boraginales followed by an additional radiation in the Boraginaceae. The Leryth_011786 gene copy is notable in that it is on a small branch relative to the other *L. erythrorhizon* sequences and groups closest to homologs in the other lamiids (Fig. 2-3a), suggesting it is likely the "missing" ubiquinone prenyltransferase[40] (hereafter referred to LePPT1).

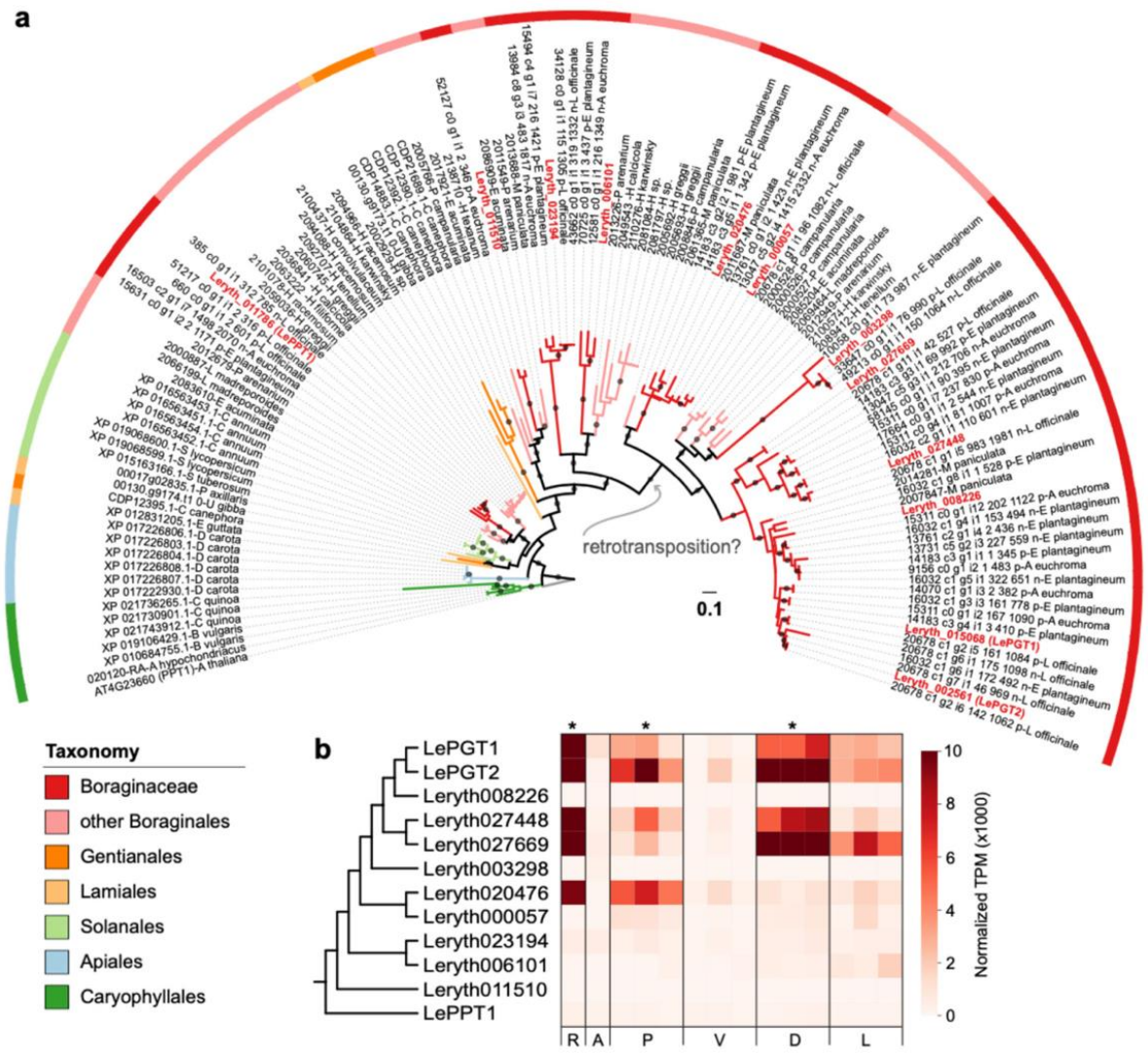

Figure 2-3 Phylogenetic analysis of prenyltransferase homologs in orthogroup OG0000509. a Maximum likelihood tree of orthogroup OG0000509 rooted on *Arabidopsis thaliana* ubiquinone polyprenyltransferase, AtPPT1. Nodes with IQ-TREE ultrafast parametric support values > 0.95 are indicated by grey circles on the preceding branch. The branches and outer color bar are color-coded to match the taxonomic classification of each sequence. *Lithospermum erythrorhizon* sequences are indicated by the red font. The hypothetical location of the inferred duplication via retrotransposition in the ancestor of genes encoding PGTs and PGT-like homologs is indicated by the grey arrow. b Heatmap showing the gene expression pattern of *L. erythrorhizon* prenyltransferase genes in whole roots (R), aerial tissue (A), root periderm (P), root vascular (V), hairy root grown in the dark (D), and hairy root grown in the light (L). Conditions where shikonin is most abundant are indicated with an aterisk (*).The cladogram (left) shows the evolutionary relationship between prenyltransferase genes according to the overall maximum likelihood phylogeny in part a.

As recognized by Kusano et al.[40], *LePGT1* and *LePGT2* are both comprised of a single exon (Fig. S2). The six *L. erythrorhizon* genes most closely related to *LePGT1* and *LePGT2* (Leryth_008226, Leryth_027669, Leryth_027448, Leryth_000057, Leryth_020476, Leryth_003298) are also single exonic. In contrast, *AtPPT1*, as well as its predicted *L. erythrorhizon* ortholog, *LePPT1*, are multi-exonic, each containing eight exons. The three remaining unclassified *L. erythrorhizon* genes in the phylogeny were found to contain a variable number of exons: Leryth_023194 with ten exons, Leryth_011510 with eight exons, and Leryth_006101 with ten exons. The loss of exons in *LePGT1* and *LePGT2*, and the six additional *PGT-like* genes, suggests that a retrotranspotition event is responsible for the duplication that gave rise to the specialized prenyltransferase genes involved in the production of shikonin.

We performed a syntenic block analysis to investigate possible whole genome dupliation (WGD) in the ancestor of *L. erythrorhizon*. The distribution of synonymous substitutions (Ks) indicates a peak at roughly ~0.45, which is suggestive of a potential polyploidy event[41] (Fig.S4a). This peak roughly matches the results of Tang et al (2020), who performed a larger analysis of WGD in the Boraginaceae. In the Tang analysis, the Ks peak of ~0.417 was proposed to have arisen via a WGD in the ancestor of the Boraginaceae roughly 25 MYA[23]. Only two syntenic blocks containing PGT and PGT-like genes were identified (Table S10). The first syntenic block had a median Ks of 0.534 and contained PGT1 (Leryth_015168) and another single exonic PGT-like gene (Leryth_008226)(Fig. S4b). The second syntenic block had a median Ks of 0.466 and contained two multi-exonic PGT-like genes (Leryth_011510 and Leryth006101) (Fig. S4c). Given that the median Ks of the two syntenic blocks containing PGT and PGT-like genes lies near the peak of ~0.45 in the Ks distribution (Fig. S4a), it is possible that these duplications arose via the WGD event. The fact that there is shared synteny between two multi-exonic homologs as well as between two single-exonic homologs—but that there is no shared synteny between a multi-exonic gene and a single-exonic gene—suggests that the retrotransposition event occurred prior to the whole genome duplication.

Aside from the suspected pseudogene, all other encoded prenyltransferases in orthogroup OG0000509 contain the conserved NDXXD motif indicative of putative prenyl diphosphate binding (Fig. S3)[34]. All but two unclassified homologs (Leryth_023194 and Leryth_006101) also maintained the GX(K/Y)STAL sequence motif conserved in this subfamily of 4HB:prenyltransferases (Fig. S3)[34]. One of these unclassified homologs, Leryth_023194, also

contained an N-terminal chloroplast targeting sequence (Table S11). None of the other prenyltransferase homologs contained a detectable signal or transit peptide. Lastly, we checked the relative gene expression of all prenyltransferase homologs in three tissue/growth condition comparisons in which shikonin was variably abundant: *L. erythrorhizon* whole root tissue versus above ground tissue, root outer periderm versus inner vascular tissue, and hairy root tissue cultures grown under dark versus lighted conditions. Shikonin production is higher in the former of all three comparisons. *LePGT1 and LePGT2*, along with *PGT-like* homologs Leryth_027448 and Leryth_027669 were significantly overrepresented (adjusted pvalue < 0.05, logfold change > 1) in conditions associated with increased shikonin production (whole root tissue, root periderm tissue, and hairy root grown in the dark; Fig. 2-3b; Table S9). Additionally, *PGT-like* homolog Leryth_020476 was significantly overrepresented in whole root and root periderm tissue (Fig. 2-3b; Table S9). Lastly, Leryth_000057 and Leyrth_023194 were significantly overrepresented in root periderm tissue. The other sequences showed zero to low expression in all samples.

2.4.5 LePGT1 is the predominant PGT functioning in the shikonin pathway

LePGT1 is considered as the key regulatory enzyme in the shikonin pathway[42,43]. As there are no reported genetic studies with *PGTs*, and in light of the newly identified PGT-like encoding genes found in the *L. erythrorhizon* genome (Fig. 2-4), we knocked down *LePGT1* expression in *L. erythrorhizon* hairy roots to investigate if LePGT1 is indeed the predominant PGT controlling shikonin production. Several independent *PGT1*-RNAi (*PGT1i*) lines were generated, excised, transferred to B5 media plates for selection, and then screened based on total shikonins (the sum of shikonin plus its derivatives) production in liquid culture using HPLC coupled with diode array detection (DAD). Individual lines were cultured in liquid B5 in constant light without selection for 14 d and then transferred to M9 and constant darkness to induce shikonin production. Analysis of culture media 3 d after transfer to M9 and darkness revealed 17 *PGT1i* lines producing between 1% and 59% of the total shikonins synthesized by control hairy roots sampled at the same time (Fig. 2-4a). Two lines, *PGT1i*-21 and *PGT1i*-79, were further analyzed by qRT-PCR. Both lines were found to have greater than 95% reduced *LePGT1* levels while expression of *LePGT2* remained statistically unchanged compared to the control (Fig. 2-4b). Because the gene encoding LePGT1 is more similar to LePGT2 than any of the PGT-likes (Fig. 2-3a), these data indicate that the RNAi construct specifically targeted *LePGT1*. Analysis of the culture media from the same

hairy roots used to perform qRT-PCR revealed that the absence of *LePGT1* expression (Fig. 2-4b) correlated with a more than 95% decrease in total shikonin content (Fig. 2-4c). These results provide further support for LePGT1 being predominantly responsible for the formation of 3-geranyl-4HBA. They also imply that the PGT-like proteins encoded in the *L. erythrorhizon* genome likely do not play a major role in shikonin production.

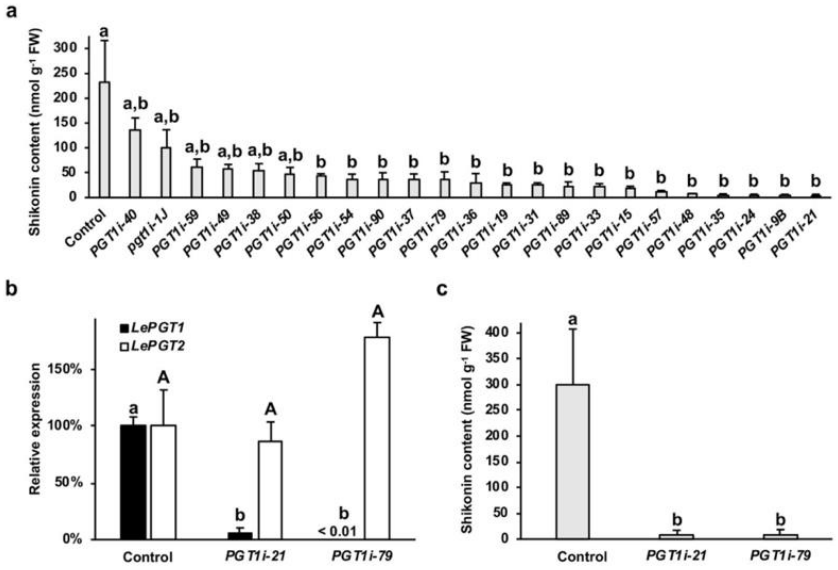

Figure 2-4 *In vivo* characterization of LePGT1. a Screening of *LePGT1*-RNAi (*PGT1i*) lines based on total shikonin levels present in liquid culture media 3 d after transfer of 14-d-old hairy roots to M9 and darkness. b Expression levels of *LePGT1* and *LePGT2* in hairy roots of two independent *PGT1i* lines compared to control. c Analysis of total shikonin in same lines used to measure expression in panel b. All data are means ± SEM (n = 3-4 biological replicates). Different letters indicate significant differences via analysis of variance (ANOVA) followed by post-hoc Tukey test (α = 0.05). In panel b, lowercase and capital letters correspond to statistical comparisons for *LePGT1* and *LePGT2* expression, respectively.

2.5 Discussion

In this study, we report the first *de novo* assembly of a *L. erythrorhizon* genome obtained through a combination of ONT long-read and Illumina HiSeq short-read sequencing technologies. The ~367.41 Mb assembly contained 27,720 predicted protein-coding genes that were clustered into 14,885 orthologous gene families, 79.84% of which were also found in other species in our comparative phylogenetic analysis (Fig. 2-2). A significant number of these orthogroups (2818) appear to have duplicated in the last common ancestor of the five Boraginaceae species in our

analysis, and another 3908 orthogroups appear to have duplicated in the last common ancestor of the four known shikonin/alkannin-producing species. Among the 6726 orthogroups duplicated at these internodes is OG0000509, the orthogroup containing *AtPPT1* and *LePPT1* (genes involved in the primary metabolic process of ubiquinone biosynthesis) as well as the specialized genes *LePGT1* and *LePGT2* involved in shikonin biosynthesis. An additional nine homologous prenyltransferases of unknown function were also identified in the *L. erythrorhizon* genome. This analysis illustrates the importance of gene duplication in the evolution of the shikonin/alkannin pathway, in agreement with other analyses of specialized metabolism in plants (*e.g.*[44–46]). In the case of the *PGT* and *PGT-like* sequences in *L. erythrorhizon*, the initial duplication event appears to have occurred via retrotransposition based the fact that these sequences appear to have lost all introns (Fig. 2-3a; Table S9). The relative contribution of retrotransposition compared to other processes that generate gene duplicates (*e.g.* whole genome duplication) to metabolic innovation in plants remains to be investigated. Overall, this genomic resource will complement the available *E. plantagineum* genome[21] and the extensive sets of transcriptomes (*e.g.*[14–16]) and proteomes[17] published from Boraginaceae species for elucidating the shikonin/alkannin pathway and its evolutionary origin, as well as the evolution of specialized metabolism more generally in this family.

The connection between shikonin/alkannin and ubiquinone biosynthesis is not limited to homologous prenyltransferases. The hydroxybenzene ring, ring *A* (Fig. 2-5), of shikonin/alkannin's naphthazarin moiety is derived from L-phenylalanine via cinnamic acid and 4HBA[4,47]–the same route that is partially responsible for forming the benzoquinone ring of ubiquinone (coenzyme Q) in plants[48,49]. Like the shikonin/alkannin pathway, ubiquinone biosynthesis starts off with the conjugation of 4HBA with a polyprenyldiphosphate, catalyzed by a PPT (Fig. 2-5). In addition to the analogous PPT- and PGT-catalyzed reactions, the ubiquinone and shikonin/alkannin pathways share other similar ring modification reactions that occur early in their respective pathways (Fig. 2-5). Both ubiquinone and shikonin/alkannin biosynthesis require the prenylated 4HBA ring to be decarboxylated and hydroxylated at the C1 position, though the sequence of reactions and necessary enzymes are unknown in both pathways. In the *Escherichia coli* ubiquinone pathway, the non-oxidative ring decarboxylation of 3-polyprenyl-4HBA is catalyzed by UbiD in concert with a UbiX chaperone for substrate reorientation[50]. Orthologs of genes encoding UbiD and UbiX are absent from plant genomes. Therefore, if plants also use a

series of non-oxidative decarboxylation and hydroxylation steps to modify the C1 position of the 4HBA ring, they do so using other evolved enzymes. It is also possible that plants have evolved to achieve the decarboxylation and hydroxylation via an oxidative decarboxylase that would carry out both reactions. Whatever the mechanism for C1 decarboxylation and hydroxylation, if it is indeed shared between shikonin/alkannin and ubiquinone biosynthesis, it is possible that there is an evolutionary linkage among the enzymes involved given the other metabolic connections between the two pathways.

Downregulation of *LePGT1* expression by 95% was sufficient to reduce shikonin production by more than 97% 3 d after transfer of hairy roots to M9 media and darkness (Fig. 2-4b,c). While this provides strong evidence that LePGT1 is principally controlling shikonin formation, it does not rule out that LePGT2 still plays a significant role. Like *LePGT1*, *LePGT2* is highly expressed under shikonin producing conditions (Fig. 2-3b; Table S9)[34,51]. One explanation for achieving near abolishment of shikonin by only knocking down *LePGT1* could be that LePGT1 and LePGT2 form heteromers *in vivo*. Biochemical studies, however, suggest that LePGT1 is capable of functioning homomericly[42,43], so this appears unlikely. Another possibility is that LePGT1 functions early in shikonin formation and LePGT2 takes over later. LePGT1 has 5- and 10-fold higher affinities for 4HBA and GPP, respectively, compared to LePGT2[34]. It can be envisioned that early in hairy root culture establishment, when precursor concentrations are *a priori* low, that LePGT1 is the *de facto* PGT responsible for forming 3-geranyl-4HBA. In this study, shikonin levels were measured at day 3, the first day that visible production occurred in control cultures. It is possible that over time, as precursor pools increase, the relative contribution of LePGT2 would become greater. If any of the PGT-like proteins encoded in the *L. erythrorhizon* genome also contribute PGT activity, their relative contributions would also increase over the culture period. In light of the current study, reports of multiple *PGT* unigenes and their variable expression patterns in other transcriptomic studies[19,39] should be re-investigated to determine if any encode PGT-like proteins. More detailed investigations looking at shikonin formation over time using metabolic flux analysis with stable isotopic labeling are needed to determine the temporal contributions of PGTs.

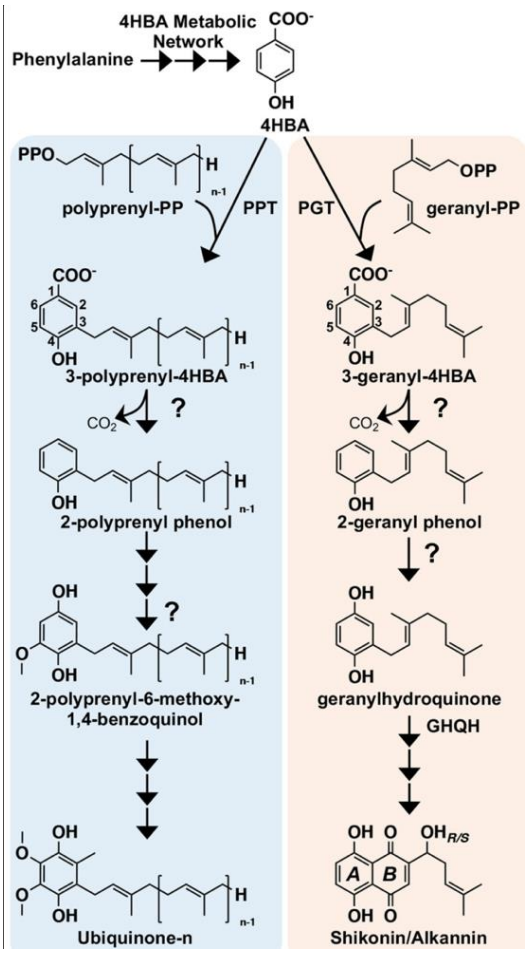

Figure 2-5 Similarities between shikonin and ubiquinone biosynthesis. The committed steps of the shikonin and ubiquinone pathways rely on homologous prenyltransferases that conjugate 4-hydroxybenzoic acid (4HBA), derived from phenylalanine, with a prenyl diphosphate precursor. Subsequent decarboxylation and hydroxylation at the C1 position of the 3-prenylated/geranylated 4HBA ring is required in both pathways, although the responsible enzymes (depicted by "?") in each route remain unknown. Non-oxidative decarboxylation of the prenylated-4HBA ring is shown as it occurs in bacteria. It is possible that plants use an oxidative decarboxylation mechanism, which would result in concomitant decarboxylation and hydroxylation at the C1 position, bypassing the phenolic intermediates. Abbreviations: GHQH, geranylhydroquinone hydroxylase; PGT, *p*-hydroxybenzoate:geranyltransferase; PPT, polyprenyltransferase.

2.6 Conclusions

The first genome assembly from the medicinally and economically important plant *L. erythrorhizon* is expected to advance understanding of the evolutionary history of the Boraginales. As just the second genome from a member of this order, the other coming from *E. plantagineum*[21], the *L. erythrorhizon* genome will provide another piece of the puzzle needed to reconstruct the

phylogenetic relationships among the Boraginales, Solanales, Gentianales, and Laminales (Fig. 2-2). The *L. erythrorhizon* genome will also serve as a novel tool for elucidating the remaining missing steps in the shikonin/alkannin pathway and for filling gaps in knowledge about its metabolic origin. Our phylogenetic analysis of prenyltransferases encoded in the *L. erythrorhizon* genome (Fig. 2-3a) has already led to the remarkable discovery of several additional encoded PGTs. It also provided evidence to suggest that *PGTs* arose in a common ancestor of modern shikonin/alkannin-producing Boraginaceae species via a retrotransposition-derived duplication event and subsequent neofunctionalization of an ancestral prenyltransferase gene. Based on homology between PGTs and PPTs it is possible that this points to an evolutionary link between ubiquinone and shikonin/alkannin biosynthesis, especially considering the other metabolic similarities shared between the two pathways (Fig. 2-5). This would not be the first connection found between primary and specialized quinone metabolism in plants as it was recently reported that the pathway to synthesize the naphthoquinone moiety of juglone in black walnut trees (*Juglans nigra*) is shared with the phylloquinone (vitamin K1) pathway[52]. Taken together, the results from our study provide several new leads for investigating the evolution of specialized metabolism in the Boraginaceae.

2.7 Materials and methods

2.7.1 Plant materials, growth conditions, and general experimental procedures

Seeds of *L. erythrorhizon* (accession Siebold & Zucc.) were obtained from the seed bank at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) in Germany and plants were propagated under standard greenhouse conditions to bulk seeds.

For all hairy root culture work, *L. erythrorhizon* Siebold & Zucc. seeds were sterilized by chlorine gas according to Lindsey *et al.*[53]. After exposure to chlorine gas for 2 h, seeds were rinsed five times with sterile water and shaken in 200 µg/mL cefotaxime overnight. They were then rinsed with sterile water and plated on half-strength Murashige and Skoog (MS, Phytotech Labs) media with 0.1% Plant Preservation Mixture (PPM^{TM}, Plant Cell Technology). Seeds were stratified at 4^{o}C for 2 weeks and moved to room temperature under 12:12 light:dark cycle. Once germinated, seed coats were removed, and each seedling was transferred to a magenta box with half-strength MS media. Plants with their second pairs of true leaves were used in hairy root transformation.

Hairy roots were maintained in solid Gamborg B5 media (Phytotech Labs) containing 3% sucrose and 10mg/L Basta (PlantMedia™). Hairy roots were subcultured every two weeks. For quantification of shikonin, 1-cm hairy root fragments were transferred to 20 mL liquid Gamborg B5 media and grown under constant light at 28°C for two weeks without Basta. Shikonin production was then induced by transferring hairy roots to M9 media[54] and culturing at 28°C in constant darkness without Basta.

2.7.2 Nanopore sequencing

For nanopore sequencing, leaves from an *in vitro* cultured 3-month-old *L. erythrorhizon* Siebold & Zucc. plant were frozen in liquid nitrogen, ground by mortar and pestle, and high molecular weight genomic DNA was extracted using a CTAB phenol chloroform extraction protocol (available on protocols.io: dx.doi.org/10.17504/protocols.io.bamnic5e) and purified using a Genomic DNA Clean and Concentrator kit (Zymo Research). At least 2 μg of gDNA was used as input for an Oxford Nanopore LSK-109 library ligation kit and sequenced on R9 MinION flow cells. Base calling was performed with Guppy v2.3.5[55]. Reads less than 3 kilobase pairs long or with quality scores less than 7 were discarded. Reads are available for download at the NCBI Sequence Read Archive (PRJNA596998).

2.7.3 Genome assembly

Lithospermum erythrorhizon Illumina gDNA PE reads sequenced by Nanjing University in 2018 from an unknown accession were downloaded from NCBI SRA database experiment SRR5644206 and assembled with Abyss v2.1.5[56] using a k-mer size of 75. The assembled Abyss contigs and the ONT long reads served as input for the DBG2OLC hybrid assembler using the following parameters: KmerCovTh 2, AdaptiveTh 0.0001, MinOverlap 20, RemoveChimera 1, k 17[22]. The resulting hybrid assembly was error corrected via five rounds of polishing with the Illumina gDNA reads using Pilon v1.23[57]. Five additional rounds of Pilon polishing were performed using *L. erythrorhizon* Siebold & Zucc. stranded RNA-seq reads (see RNA-seq experiments section below) as input to fix single nucleotide errors in transcribed regions. Genome size was estimated using GenomeScope[58] with a k-mer (k=21) depth distribution of the Illumina gDNA reads calculated using Jellyfish v2.2.10[59]. Illumina gDNA reads were mapped back to the final assembly using

BWA v0.7.15[60], RNAseq reads were mapped using STAR v2.5.4[61], and ONT reads were mapped using minimap2 v2.13[62]. Histograms of Illumina and ONT read depth when mapped to the final assembly were generated by the program purge_haplotigs v1.1.1 to assess the level of assembly heterozygosity[63].

2.7.4 Genome annotation

De novo repeat identification was performed using RepeatModeler v1.0.9 and masked using RepeatMasker v4.0.7 (http://repeatmasker.org). Gene model and protein prediction was conducted with MAKER2 v2.31.10[64] by supplying protein homology-based evidence, transcriptomic evidence in the form of a genome guided transcriptome assembly generated from in-house Siebold & Zucc. stranded RNA-seq reads (see RNA-seq experiments section below) using Trinity v2.5.1[65], and *ab initio* gene calling using SNAP[66] and BRAKER2[67]. Gene models with an AED score < 0.9 or ones that encoded a predicted protein < 30 amino acids long were excluded from the final gene set. Additional information on the full annotation pipeline is presented in Fig. S6. Conservation of core genes was performed using BUSCO v2.0[28]. Functional annotation of the final protein set was performed using InterProScan[24] and TargetP[68].

We checked the genome for possible contamination using the Alien Index (AI) pipeline (https://github.rcac.purdue.edu/jwisecav/phylo-pipe; last updated August 26, 2019) as previously described[29]. Briefly, each predicted protein sequence was queried against the NCBI RefSeq database (release 97) using Diamond v0.9.22.123[69], and the AI score was calculated based on the output. The AI score is given by the formula: AI=nbsO-nbsE, where *nbsO* is the normalized bit score of the best hit to a species outside of the eudicot lineage, *nbsE* is the normalized bit score of the best hit to a species within the eudicot lineage (skipping all hits to *L. erythrorhizon* present in the RefSeq database). AI scores range from -1 to 1, being greater than zero if the predicted protein sequence had a better hit to a non-eudicot species, suggestive of either HGT or contamination[29].

2.7.5 RNA-seq experiments

For the *L. erythrorhizon* root periderm and vascular tissues RNA-seq experiment, 3-month-old Siebold & Zucc. plants grown in soil under standard greenhouse conditions were harvested. Roots were collected from nine individual plants and divided into three groups, each containing three

unique individuals. The periderm and vascular tissues were isolated by peeling the periderm from the roots (Fig. S5a), and the prepared portions from the three individuals in each group were pooled. Tissues were frozen in liquid nitrogen, ground by mortar and pestle, and 100 mg was used to analyze total shikonins content each sample (Fig. S5b). From the same sets of samples, RNA was extracted as described below, quantified, and DNase-treated (NEB) according to the manufacturer's instructions. A total of six cDNA libraries from the three biological replicates prepared from each of the *L. erythrorhizon* periderm and vascular tissue pools, were constructed using a ribominus TruSeq Stranded Total RNA library prep kit (Illumina, San Diego, CA), and 101-bp paired-end reads were generated via Illumina HiSeq 2500 at the Purdue Genomics Center, with at least 67 million reads per library. Sequence quality was assessed by FastQC (v. 0.10.0; http://www.bioinformatics.babraham. ac.uk). The raw data were submitted to the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra/) and are available at the NCBI Sequence Read Archive (PRJNA596998)

The experimental design for the RNA-seq experiment comparing *L. erythrorhizon* hairy roots sampled in B5 in the light and M9 in the dark was based on a previous report of observed rapid increases in expression of shikonin precursor pathway genes, and in *PGT*, within 2 hrs after switching *L. erythrorhizon* cell cultures from growth in B5 in the light to growth in M9 in darkness[70]. In this study, several cultures from three independently generated *L. erythrorhizon* hairy root lines were started in liquid Gamborg B5 media containing 3% sucrose at 28^{o}C in the light (~100 µE m^{-2} s^{-1}). After 2 weeks, hairy roots from three cultures for each of the three lines (n = 3 biological replicates per line) were harvested and pooled to represent the B5 light-treated samples. The remaining hairy root cultures were transferred to M9 media and darkness. After 2 hrs, hairy roots from three cultures for each of the three lines (n = 3 biological replicates per line) were harvested and pooled to represent the M9 dark-treated samples. Samples were frozen in liquid nitrogen, ground by mortar and pestle, and RNA was extracted as described below. Six cDNA libraries were generated with a TruSeq Stranded mRNA library prep kit (Illumina, San Diego, CA) and were sequenced on an Illumina NovaSeq 6000 at the Purdue Genomics Center. Sequence quality assessment were performed as described above for the periderm and vascular tissues RNA-seq experiment. The raw data were submitted to the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra/) and are available at the NCBI Sequence Read Archive (PRJNA596998).

Additionally, unstranded RNA-seq data of *L. erythrorhizon* whole roots and aerial tissue from an unknown accession was downloaded from the NCBI SRA (experiments SRR3957230 and SRR3957231) to include in the gene expression analysis. Gene abundance estimates of PGT and PGT-like genes (Fig. 2-3b, Table S9) were measured using Kallisto v0.45.0[71] and normalized for library depth using DESeq2[72]. Differential expression status was determined using the EdgeR v3.24.3[73] package. For the EdgeR analysis, raw counts were normalized into effective library sizes using the trimmed mean of M-values (TMM) method[74], and exact tests were conducted using a trended dispersion value and a double tail reject region. A false discovery rate was calculated using the Benjamini–Hochberg procedure[75]. Genes with a logfold change in abundance greater than 1 and false discovery rate less than 0.05 were considered as differentially represented.

2.7.6 *De novo* transcriptome assemblies of additional Boraginaceae

Illumina RNA-seq reads from additional shikonin-producing species *L. officinale*, *Arnebia euchroma* and *Echium plantagineum*[19,39] were downloaded from the following NCBI SRA experiments: SRR4034889, SRR4034892, SRR4034890, SRR4034891, SRR6799516, SRR6799517, and SRR6799518. Raw RNA-seq reads were normalized and error corrected with the BBnorm using a target size of 40 and a minimum depth of 2 (software last modified October 19, 2017) and Tadpole using default parameters (software last modified June 27, 2017), programs from the BBMap software package[76]. The resulting clean reads were assembled *de novo* using Trinity v2.5.1[65] with default parameters for stranded (*L. officinale*) and unstranded (*E. plantagineum* and *A. euchroma*) libraries. Coding regions were inferred using TransDecoder v3.0.1[77].

2.7.7 Identification of orthologous gene families

Homology between the predicted proteomes of *L. erythrorhizon* and 31 other eudicot species was determined with OrthoFinder v2.1.2 using the following parameters: -S diamond -M msa -T fasttree[30]. The species tree in Fig. 2-2 was generated using TreeGraph2 v2.15.0[78]. Hypergeometric tests were performed in python using the SciPy library hypergeom, and p-values were adjusted for multiple comparisons using the StatsModels library multitest with the Benjamini & Hochberg (BH) method[75].

2.7.8 Phylogenetic analysis

We performed a separate phylogenetic analysis of the orthogroup containing LePGT1 and LePGT2 (OG0000509). Three sequences were excluded due their long branches, including the suspected pseudogene Leryth_015069. The remaining sequences were aligned with MAFFT v7.407 using the E-INS-I strategy and following parameters: --maxiterate 1000 --bl 45 --op 1.0 --retree 3[79]. The maximum likelihood phylogeny was constructed using IQ-TREE version 1.6.10[80] using the built in ModelFinder to determine the best-fit substitution model[81] and performing SH-aLRT and ultrafast bootstrapping analyses with 1,000 replicates each. The gene tree in Fig. 2-3a was generated using iTOL v4[82].

2.7.9 Synteny analysis

Regions of shared synteny within the genome of *L. erythrorhizon* were detected using SynMap2 on the online Comparative Genomics Platform (CoGe) using default settings with the exception that the merge syntenic blocks algorithm was set to Quota Align Merge, syntenic depth algorithm was set to Quota Align, and the CodeML option was activated to calculate substitution rates between syntenic CDS pairs. For syntenic blocks containing PGT genes and their homologs, the encompasing contigs were aligned using promer (v3.07) of the MUMmer4 alignment system[83].

2.7.10 Cloning and generation of LePGT1i hairy root lines

For the *LePGT1*-RNAi construct, DNA containing two spliced *LePGT1* cDNA fragments of the coding region corresponding to nucleotides 179-698 and 179-502, the latter in antisense orientation to create a hairpin structure, was synthesized (Genscript, Piscataway, NJ). 5'-CACC was added for subcloning into pENTR™/D-TOPO (Invitrogen™, Carlsbad, CA), sequence verified, and transferred into the destination vector, pB2GW7[84], by recombination using LR Clonase Enzyme Mix™ (Invitrogen). The final construct, pB2GW7-*PGT1i*, was transformed into *Agrobacterium rhizogenes* strain ATCC 15834 competent cells by freeze-thaw transformation[85]. Briefly, competent cells were incubated with 100 ng of pB2GW7-*PGT1i* for 15 min on ice. Then, cells were snap frozen in liquid nitrogen for 5 min and consecutively thawed at 37°C for 5 min. Nutrient Broth (NB) media was added to the culture and kept in 37°C with shaking for 2 h before being plated on NB agar containing 50 µg/mL spectinomycin for selection.

L. erythrorhizon hairy root *PGT1i* lines were generated based on the protocol from Fang et al.[86] with slight modification. *A. rhizogenes* containing pB2GW7-*PGT1i* was inoculated in NB with 50 µg/mL spectinomycin and kept in shaking incubator at 28°C until reached OD_{600} = 1. Then, acetosyringone was added to the media to a final concentration of 0.1 mM and the culture was grown further for 4 h in dark followed by centrifugation. The pellet was then washed and resuspended in half-strength MS containing 0.1 mM acetosyringone. Each stem of sterile plants grown in tissue culture was wounded by surgical blade and the prepared culture was applied to the wounded area by cotton swab. The plants were then kept in darkness for 1 d and returned to normal growth conditions. The hairy roots emerged between 10-28 d post infection. Emergent roots were excised and placed on Gamborg B5 media with 3% sucrose and 200 µg/mL cefotaxime to eliminate *A. rhizogenes.* After 2 weeks, hairy roots were transferred to Gamborg B5 media containing 3% sucrose and 10 mg/L Basta for selection for 2 weeks. Hairy root lines transformed by *A. rhizogenes* without pB2GW7-*PGT1i* were generated to use as control.

2.7.11 RNA extraction and qRT-PCR

The total RNA was extracted from approximately 20 mg of hairy root tissues according to the protocol from Ghawana *et al*.[87] Briefly, the samples were frozen by liquid nitrogen and ground by mortar and pestle. 2 mL of RNA extraction buffer (phenol containing 0.1% sodium dodecyl sulfate, 0.32M sodium acetate, and 0.1 M ethylenediaminetetra acetic acid) was added to the sample in the mortar and mixed, followed by addition 0.8 mL of RNAse-free water. After mixing, the mixture was incubated for 5 min before transferring to microtubes. 0.3 mL of chloroform was added, and the sample was vortexed and centrifuged at 4°C, 13,000 rpm for 10 min. The supernatant was then transferred to the new tube containing 0.6 mL isopropanol. Next, the sample was mixed by inverting, and nucleic acids were precipitated at -20°C for 10 min. After precipitation, the sample was centrifuged at 4°C for 10 min. The pellet was washed with 70% ethanol before air drying. The RNA pellet was dissolved in RNase free water. The total RNA was concentrated and purified using an RNA Clean & Concentrator Kit (Zymo Research) with on-column DNase treatment (Zymo Research) using the manufacturer protocol. cDNA synthesis was performed by 5X All-In-One RT MasterMix (abm) according to manufacturer instructions using 500 ng of total RNA.

Expression of *LePGT1* and *LePGT2* was measured by qRT-PCR with comparative quantification using the $2^{-\Delta\Delta CT}$ method[88]. Primers were designed using Primer-BLAST on NCBI[89].

Due to the sequence similarity of members of the *LePGT* and *LePGT-like* (Table S9) gene family, each primer was checked against all members for possible off-target matches. To minimize off-target amplification, primer pairs were selected that had either (i) four or more mismatches in a primer to all other *LePGT* and *LePGT-like* family genes or (ii) two mismatches in one primer and three mismatches in the other primer to all other *LePGT* and *LePGT-like* family genes[90]. qRT-PCR reactions were performed using a QuantStudio™ 6 (ThermoFisher) in a 10 µL reaction as follows: 5 µL of 5x Fast SYBR Green PCR master mix (ThermoFisher), 1 µL each of the forward and reverse primers (50-900 nM final concentration; Table S12), and 3 µL of diluted cDNA. Expression was normalized to *L. erythrorhizon* glyceraldehyde 3-phosphate dehydrogenase (*LeGAPDH*) using primers from Zhao *et al* (2015)[91].

2.7.12 Shikonin extraction and quantification

The extraction of total shikonins was modified from Boehm *et al*.[92]. Briefly, a 4 mL sample of growth media from each hairy root line was sampled at day 3 after transfer to M9 and darkness, extracted with 4 mL of chloroform, and then the chloroform layer was separated and dried under a gentle stream of N_2 at 40°C. Base hydrolysis was performed on the remaining residue in the tube by adding 2 mL of 1 M NaOH and shaking for 1 h at room temperature. The solution was neutralized by adding 1 mL of 6 M HCl and vortexed. Shikonin was extracted by adding 3 mL of ethyl acetate by liquid-liquid extraction. The ethyl acetate layer was separated and dried under N_2, dissolved in 250 µL methanol, and 20 µL was used for detection by high performance liquid chromatography with diode array detection (HPLC-DAD). The extraction procedure was performed in reduced light to minimize the photo-degradation of shikonin.

HPLC-DAD analyses were performed with an Agilent 1260 Infinity HPLC system (Agilent Technologies). Chromatographic separation of shikonin was achieved using a Zorbax SB-C18 column (4.6 × 250 mm, Agilent) kept at 25°C. The mobile phase gradient started at 60% A (30:70 acetonitrile and water with 0.1% formic acid) and 40% B (30:70 isopropanol and acetonitrile with 0.1% formic acid) with 1 min hold and then linearly increased to 99% B over 15 min with a hold of 4 min, and then returned to 40% B from 19 to 20 min with a hold of 1 min. Shikonin eluted at 8.4 min and was detected at 520nm by DAD. Instrument operation and data analysis steps were performed through the Agilent ChemStation software. Shikonin quantitation by DAD was done by running a linear range of 1.25, 2.5, 5, 10 and 20 nmol calibration standards,

followed by a linear regression formula calculation. Differences between shikonin content in each line (n = 3-4 biological replicates) was analyzed using one-way ANOVA and the means were compared with Tukey's HSD post hoc test at 95% significant level.

For analysis of shikonins from *L. erythrorhizon* root periderm and vascular tissues by HPLC coupled with fluorescence detection (HPLC-FLD; Fig. S5), chromatographic separation was conducted using a Zorbax Eclipse XDB-C18 column (4.6 x 150 mm, Agilent). The column was eluted at 25ºC using a linear gradient starting from 70% A (water with 0.1% formic acid) and 30% B (acetonitrile with 0.1% formic acid), to 1% A and 99% B over 40 min at a flow rate of 0.5 mL min^{-1}, followed by a 10 min re-equilibration step. Shikonin eluted at 25.9 min and was detected by fluorescence using λ_{ex} = 228 nm and λ_{em} = 390 nm after passing through an in-line post-column dry reactor packed with zinc dust, which was previously used for detection of the 1,4-naphthoquinone juglone[52].

2.8 Supplemental data

Supplemental tables, figures, and datasets are available at https://doi.org/10.1038/s41438-020-0301-9

2.9 References

1 Papageorgiou VP, Assimopoulou AN, Couladouros EA, Hepworth D, Nicolaou KC. The chemistry and biology of alkannin, shikonin, and related naphthazarin natural products. Angew Chemie - Int Ed 1999; 38: 270–300.

2 Skoneczny D, Weston P, Zhu X et al. Metabolic Profiling and Identification of Shikonins in Root Periderm of Two Invasive Echium spp. Weeds in Australia. Molecules 2017; 22: 330.

3 Zhu X, Skoneczny D, Weidenhamer JD et al. Identification and localization of bioactive naphthoquinones in the roots and rhizosphere of Paterson's curse (Echium plantagineum), a noxious invader. J Exp Bot 2016; 67: 3777–3788.

4 Schmid HV, Zenk MH. p-hydroxybenzoic acid and mevalonic acid as precursors of the plant naphthoquinone alkannin. Tetrahedron Lett 1971; 12: 4151–4155.

5 Widhalm JR, Rhodes D. Biosynthesis and molecular actions of specialized 1,4-naphthoquinone natural products produced by horticultural plants. Hortic Res 2016; 3: 16046.

6 Chen X, Yang L, Zhang N et al. Shikonin, a component of chinese herbal medicine, inhibits chemokine receptor function and suppresses human immunodeficiency virus type 1. Antimicrob Agents Chemother 2003; 47: 2810–6.

7 Duru N, Gernapudi R, Zhou Q. Chemopreventive activities of shikonin in breast cancer. Biochem Pharmacol 2014; 3: e163.

8 Yazaki K. Lithospermum erythrorhizon cell cultures: Present and future aspects. Plant Biotechnol 2017; 34: 131–142.

9 Gaisser S, Heide L. Inhibition and Regulation of Shikonin Biosynthesis in Suspension Cultures of Lithospermum. Phytochemistry 1996; 41: 1065–1072.

10 Boehm R, Sommer S, Li SM, Heide L. Genetic engineering on shikonin biosynthesis: expression of the bacterial ubiA gene in Lithospermum erythrorhizon. Plant Cell Physiol 2000; 41: 911–919.

11 Köhle A, Sommer S, Yazaki K et al. High level expression of chorismate pyruvate-lyase (UbiC) and HMG-CoA reductase in hairy root cultures of Lithospermum erythrorhizon. Plant Cell Physiol 2002; 43: 894–902.

12 Wang R, Zhou S, Jiang H, Zheng X, Zhou W, Li S. An Efficient Multigram Synthesis of Alkannin and Shikonin. European J Org Chem 2012; 2012: 1373–1379.

13 Wang F, Yao X, Zhang Y, Tang J. Synthesis, biological function and evaluation of Shikonin in cancer therapy. Fitoterapia 2019; 134: 329–339.

14 Wang S, Ping L, Teng G. Different secondary metabolic responses to MeJA treatment in shikonin-proficient and shikonin-deficient cell lines from Arnebia euchroma (Royle) Johnst. 2014; : 587–598.

15 Zhu Y, Lu GH, Bian ZW et al. Involvement of LeMDR, an ATP-binding cassette protein gene, in shikonin transport and biosynthesis in Lithospermum erythrorhizon. BMC Plant Biol 2017; 17: 1–10.

16 Wu F-Y, Tang C-Y, Guo Y-M et al. Transcriptome analysis explores genes related to shikonin biosynthesis in Lithospermeae plants and provides insights into Boraginales? evolutionary history. Sci Rep 2017; 7: 4477.

17 Takanashi K, Nakagawa Y, Aburaya S et al. Comparative Proteomic Analysis of Lithospermum erythrorhizon Reveals Regulation of a Variety of Metabolic Enzymes Leading to Comprehensive Understanding of the Shikonin Biosynthetic Pathway. Plant Cell Physiol 2019; 60: 19–28.

18 Wang S, Wang R, Liu T et al. CYP76B74 catalyzes the 3’’-hydroxylation of geranylhydroquinone in shikonin biosynthesis. 2018 doi:10.1104/pp.18.01056.

19 Rai A, Nakaya T, Shimizu Y et al. De Novo Transcriptome Assembly and Characterization of Lithospermum officinale to Discover Putative Genes Involved in Specialized Metabolites Biosynthesis *. 2018.

20 Zhu X, Skoneczny D, Weidenhamer JD et al. Identification and localization of bioactive naphthoquinones in the roots and rhizosphere of Paterson’s curse (Echium plantagineum), a noxious invader. J Exp Bot 2016; 67: 3777–3788.

21 Tang C, Li S, Wang Y, Wang X. Comparative genome/transcriptome analysis probes Boraginales’ phylogenetic position, WGDs in Boraginales, and key enzyme genes in the alkannin/shikonin core pathway. Mol Ecol Resour 2019; : 1–14.

22 Ye C, Hill CM, Wu S, Ruan J, Ma Z. DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci Rep 2016; 6: 1–9.

23 Tang CY, Li S, Wang YT, Wang X. Comparative genome/transcriptome analysis probes Boraginales’ phylogenetic position, WGDs in Boraginales, and key enzyme genes in the alkannin/shikonin core pathway. Mol Ecol Resour 2019. doi:10.1111/1755-0998.13104.

24 Jones P, Binns D, Chang HY et al. InterProScan 5: Genome-scale protein function classification. Bioinformatics 2014; 30: 1236–1240.

25 El-Gebali S, Mistry J, Bateman A et al. The Pfam protein families database in 2019. Nucleic Acids Res 2019. doi:10.1093/nar/gky995.

26 Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004; 32: 258D – 261.

27 Caspi R, Billington R, Fulcher CA et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res 2018. doi:10.1093/nar/gkx935.

28 Waterhouse RM, Seppey M, Simao FA et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 2018; 35: 543–548.

29 Wisecaver JH, Alexander WG, King SB, Todd Hittinger C, Rokas A. Dynamic Evolution of Nitric Oxide Detoxifying Flavohemoglobins, a Family of Single-Protein Metabolic Modules in Bacteria and Eukaryotes. Mol Biol Evol 2016; 33. doi:10.1093/molbev/msw073.

30 Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 2015; 16: 1–14.

31 Leebens-Mack JH, Barker MS, Carpenter EJ et al. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 2019; 574: 679–685.

32 Chase MW, Christenhusz MJM, Fay MF et al. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc 2016. doi:10.1111/boj.12385.

33 Okada K, Ohara K, Yazaki K et al. The AtPPT1 gene encoding 4-hydroxybenzoate polyprenyl diphosphate transferase in ubiquinone biosynthesis is required for embryo development in Arabidopsis thaliana. Plant Mol Biol 2004; 55: 567–577.

34 Yazaki K, Kunihisa M, Fujisaki T, Sato F. Geranyl Diphosphate:4-Hydroxybenzoate Geranyltransferase fromLithospermum erythrorhizon: Cloning and characterization of a key enzyme in shikonin biosynthesis. J Biol Chem 2002; 277: 6240–6246.

35 Yazaki K, Fukui H, Tabata M. Isolation of the intermediates and related metabolites of shikonin biosynthesis from Lithospermum erythrorhizon cell cultures. Chem Pharm Bull (Tokyo) 1986; 34: 2290–2293.

36 Heide L, Tabata M. Geranylpyrophosphate: p-hydroxybenzoate geranyltransferase activity in extracts of Lithospermum erythrorhizon cell cultures. Phytochemistry 1987; 26: 1651–1655.

37 Singh RS, Gara RK, Bhardwaj PK et al. Expression of 3-hydroxy-3-methylglutaryl-CoA reductase, p-hydroxybenzoate-m-geranyltransferase and genes of phenylpropanoid pathway exhibits positive correlation with shikonins content in arnebia [Arnebia euchroma (Royle) Johnston]. BMC Mol Biol 2010; 11: 88.

38 Wu SJS-J, Qi JLJ-L, Zhang W-JWJ et al. Nitric Oxide Regulates Shikonin Formation in Suspension-Cultured Onosma paniculatum Cells. Plant Cell Physiol 2009; 50: 118–128.

39 Wu FY, Tang CY, Guo YM et al. Transcriptome analysis explores genes related to shikonin biosynthesis in Lithospermeae plants and provides insights into Boraginales' evolutionary history. Sci Rep 2017; 7: 1–11.

40 Kusano H, Li H, Minami H, Kato Y, Tabata H, Yazaki K. Evolutionary developments in plant specialized metabolism, exemplified by two transferase families. Front Plant Sci 2019. doi:10.3389/fpls.2019.00794.

41 Lyons E, Freeling M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J 2008. doi:10.1111/j.1365-313X.2007.03326.x.

42 Ohara K, Muroya A, Fukushima N, Yazaki K. Functional characterization of LePGT1, a membrane-bound prenyltransferase involved in the geranylation of p-hydroxybenzoic acid. Biochem J 2009; 421: 231–241.

43 Ohara K, Mito K, Yazaki K. Homogeneous purification and characterization of LePGT1 - A membrane-bound aromatic substrate prenyltransferase involved in secondary metabolism of Lithospermum erythrorhizon. FEBS J 2013; 280: 2572–2580.

44 Lopez-Nieves S, Yang Y, Timoneda A et al. Relaxation of tyrosine pathway regulation underlies the evolution of betalain pigmentation in Caryophyllales. New Phytol 2018. doi:10.1111/nph.14822.

45 Edger PP, Heidel-Fischer HM, Bekaert M et al. The butterfly plant arms-race escalated by gene and genome duplications. Proc Natl Acad Sci 2015; 112: 8362–8366.

46 Moghe GD, Leong BJ, Hurney SM, Jones AD, Last RL. Evolutionary routes to biochemical innovation revealed by integrative analysis of a plant-defense related specialized metabolic pathway. Elife 2017; 6: e28468.

47 Loscher R, Heide L. Biosynthesis of p-Hydroxybenzoate from p-Coumarate and p-Coumaroyl-Coenzyme A in Cell-Free Extracts of Lithospermum erythrorhizon Cell Cultures. Plant Physiol 1994; 106: 271–279.

48 Block A, Widhalm JR, Fatihi A et al. The Origin and Biosynthesis of the Benzenoid Moiety of Ubiquinone (Coenzyme Q) in Arabidopsis. Plant Cell 2014; 26: 1938–1948.

49 Soubeyrand E, Johnson TS, Latimer S et al. The Peroxidative Cleavage of Kaempferol Contributes to the Biosynthesis of the Benzenoid Moiety of Ubiquinone in Plants. Plant Cell 2018; 30: 2910–2921.

50 White MD, Payne K a. P, Fisher K et al. UbiX is a flavin prenyltransferase required for bacterial ubiquinone biosynthesis. Nature 2015. doi:10.1038/nature14559.

51 Zhu Y, Chu S-J, Luo Y-L et al. Involvement of LeMRP, an ATP-binding cassette transporter, in shikonin transport and biosynthesis in Lithospermum erythrorhizon. Plant Biol 2017; 17: 1–9.

52 McCoy RM, Utturkar SM, Crook JW, Thimmapuram J, Widhalm JR. The origin and biosynthesis of the naphthalenoid moiety of juglone in black walnut. Hortic Res 2018; 5. doi:10.1038/s41438-018-0067-5.

53 Lindsey I, Rivero BE, Calhoun LS, Grotewold CS, Brkljacic E. Standardized Method for High-throughput Sterilization of Arabidopsis Seeds. J Vis Exp 2017; : 56587.

54 Fujita Y, Hara Y, Ogino T, Suga C. Production of shikonin derivatives by cell suspension cultures of Lithospermum erythrorhizon. Plant Cell Rep 1981; 1: 59–60.

55 Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol 2019; 20: 1–10.

56 Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: A parallel assembler for short read sequence data. Genome Res 2009; 19: 1117–1123.

57 Walker BJ, Abeel T, Shea T et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014; 9. doi:10.1371/journal.pone.0112963.

58 Vurture GW, Sedlazeck FJ, Nattestad M et al. GenomeScope: Fast reference-free genome profiling from short reads. Bioinformatics 2017; 33: 2202–2204.

59 Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 2011; 27: 764–770.

60 Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013; 00: 1–3.

61 Dobin A, Davis CA, Schlesinger F et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 2013; 29: 15–21.

62 Li H. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. Bioinformatics 2016; 32: 2103–2110.

63 Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 2018; 19: 1–10.

64 Holt C, Yandell M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 2011; 12. doi:10.1186/1471-2105-12-491.

65 Haas BJ, Papanicolaou A, Yassour M et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 2013; 8: 1494–512.

66 Korf I. Gene finding in novel genomes. BMC Bioinformatics 2004; 5: 1–9.

67 Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-Genome Annotation with BRAKER. In: Kollmar M (ed). Gene Prediction: Methods and Protocols. Springer New York: New York, NY, 2019, pp 65–95.

68 Armenteros JJA, Salvatore M, Emanuelsson O et al. Detecting sequence signals in targeting peptides using deep learning. Life Sci Alliance 2019. doi:10.26508/lsa.201900429.

69 Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015; 12: 59–60.

70 Zhang WJ, Su J, Tan MY et al. Expression analysis of shikonin-biosynthetic genes in response to M9 medium and light in Lithospermum erythrorhizon cell cultures. Plant Cell Tissue Organ Cult 2010; 101: 135–142.

71 Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 2016; 34: 525–527.

72 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014. doi:10.1186/s13059-014-0550-8.

73 Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2009; 26: 139–140.

74 Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010; 11. doi:10.1186/gb-2010-11-3-r25.

75 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 1995; 57: 289–300.

76 Bushnell B. BBTools Software Package. 2017.

77 Haas B. TransDecoder (Find Coding Regions within Transcripts). 2012; : http://transdecoder.sourceforge.net.

78 Stöver BC, Müller KF. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 2010. doi:10.1186/1471-2105-11-7.

79 Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 2013; 30: 772–780.

80 Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 2015; 32: 268–274.

81 Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods 2017. doi:10.1038/nmeth.4285.

82 Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 2019. doi:10.1093/nar/gkz239.

83 Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol 2018. doi:10.1371/journal.pcbi.1005944.

84 Karimi M, Inzé D, Depicker A. GATEWAYTM vectors for Agrobacterium-mediated plant transformation. Trends Plant Sci 2002; 7: 193–195.

85 Cui W, Liu W, Wu G. A simple method for the transformation of Agrobacterium tumefaciens by foreign DNA. Chin J Biotechnol 1995; 11: 267–274.

86 Fang R, Wu F, Zou A et al. Transgenic analysis reveals LeACS-1 as a positive regulator of ethylene-induced shikonin biosynthesis in Lithospermum erythrorhizon hairy roots. Plant Mol Biol 2016. doi:10.1007/s11103-015-0421-z.

87 Ghawana S, Paul A, Kumar H et al. An RNA isolation system for plant tissues rich in secondary metabolites. BMC Res Notes 2011; 4: 85.

88 Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. Methods 2001; 25: 402–408.

89 Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 2012; 13: 134.

90 Lefever S, Pattyn F, Hellemans J, Vandesompele J. Single-nucleotide polymorphisms and other mismatches reduce performance of quantitative PCR assays. Clin Chem 2013; 59: 1470–1480.

91 Zhao H, Chang QS, Zhang DX et al. Overexpression of LeMYB1 enhances shikonin formation by up-regulating key shikonin biosynthesis-related genes in Lithospermum erythrorhizon. Biol Plant 2015; 59: 429–435.

92 Boehm R, Sommer S, Li S-M, Heide L. Genetic Engineering on Shikonin Biosynthesis: Expression of the Bacterial ubiA Gene in Lithospermum erythrorhizon. Plant Cell Physiol 2000; 41: 911–919.

CHAPTER 3. INTEGRATIVE ANALYSIS OF THE SHIKONIN METABOLIC NETWORK IDENTITFIES NEW GENE CONNECTIONS AND REVEALS EVOLUTIONARY INSIGHT INTO SHIKONIN BIOSYNTHESIS

*Originally published in *Horticulture Research*

Thiti Suttiyut & Robert P Auber, Manoj Ghaste, Cade N Kane, Scott A M McAdam, Jennifer H Wisecaver, Joshua R Widhalm, Integrative analysis of the shikonin metabolic network identifies new gene connections and reveals evolutionary insight into shikonin biosynthesis, Horticulture Research, Volume 9, 2022, uhab087, https://doi.org/10.1093/hr/uhab087

3.1 Declaration of collaborative work

Robert Auber performed the gene coexpression network analysis, promoter analysis, phylogenetic analyses, and synteny analysis. Thiti Suttiyut generated the RNAi knockdown lines and performed the subsequent inhibitor experiments and RNAseq analysis of gene expression. Cade Kane and Dr. Scott McAdam performed the analysis of abscisic acid. Thiti Suttiyut and Manoj Ghaste performed the analysis of additional metabolites. Thiti Suttiyut, Robert Auber, Dr. Jennifer Wisecaver, and Dr. Joshua Widhalm wrote the manuscript with input from all coauthors. Thiti Suttiyut, Robert Auber, Dr. Jennifer Wisecaver, and Dr. Joshua Widhalm conceived the project and were involved in experimental design.

3.2 Abstract

Plant specialized 1,4-naphthoquinones present a remarkable case of convergent evolution. Species across multiple discrete orders of vascular plants produce diverse 1,4-naphthoquinones via one of several pathways using different metabolic precursors. Evolution of these pathways was preceded by events of metabolic innovation and many appear to share connections with biosynthesis of photosynthetic or respiratory quinones. Here, we sought to shed light on the metabolic connections linking shikonin biosynthesis with its precursor pathways and on the origins of shikonin metabolic genes. Downregulation of *Lithospermum erythrorhizon* geranyl diphosphate synthase (LeGPPS),

recently shown to have been recruited from a cytoplasmic farnesyl diphosphate synthase (FPPS), resulted in reduced shikonin production and a decrease in expression of mevalonic acid and phenylpropanoid pathway genes. Next, we used *LeGPPS* and other known shikonin pathway genes to build a coexpression network model for identifying new gene connections to shikonin metabolism. Integrative in silico analyses of network genes revealed candidates for biochemical steps in the shikonin pathway arising from Boraginales-specific gene family expansion. Multiple genes in the shikonin coexpression network were also discovered to have originated from duplication of ubiquinone pathway genes. Taken together, our study provides evidence for transcriptional crosstalk between shikonin biosynthesis and its precursor pathways, identifies several shikonin pathway gene candidates and their evolutionary histories, and establishes additional evolutionary links between shikonin and ubiquinone metabolism. Moreover, we demonstrate that global coexpression analysis using limited transcriptomic data obtained from targeted experiments is effective for identifying gene connections within a defined metabolic network.

3.3 Introduction

The shikonins are a group of red-pigmented naphthoquinones produced in the root periderm of many members of Boraginaceae[1,2]. They include shikonin (Fig. 3-1), its enantiomer alkannin, and several shikonin/alkannin derivatives that are excreted into the rhizosphere, where they function in defense, mediate plant-microbe interactions, and/or elicit allelopathic effects on other plants. For example, the invasion success of Paterson's curse (*Echium plantagineum*) in southeast Australia is attributed, at least in part, to the synthesis and release of shikonins[3]. Shikonins are also the bioactive compounds responsible for the various pharmacological properties of medicinal plants like red gromwell (*Lithospermum erythrorhizon*)[4] and have emerged as scaffolds for semi-synthesis of novel cancer therapeutics[5]. The structure of shikonin is comprised of a redoxactive naphthazarin (5,6-dihydroxy-1,4-naphthoquinone) ring fused with a 1-hydroxy-4-methyl-3-pentenyl side chain (Fig. 3-1). The hydroxybenzene ring, ring A, of shikonin's naphthazarin moiety is derived from Lphenylalanine via cinnamic acid and 4-hydroxybenzoate (4-HBA)[6,7]. This is the same route predominantly responsible for forming the benzoquinone ring of ubiquinone (coenzyme Q) in plants[8,9]. Many of the genes responsible for synthesis of the 4-HBA precursor of shikonin have already been cloned and investigated (e.g.[10-12]). In contrast, the genetic basis and

regulation underlying the unique formation of the prenyl diphosphate precursor providing shikonin's quinone ring, ring B, and its six-carbon atom isoprenoid side chain is not as well characterized.

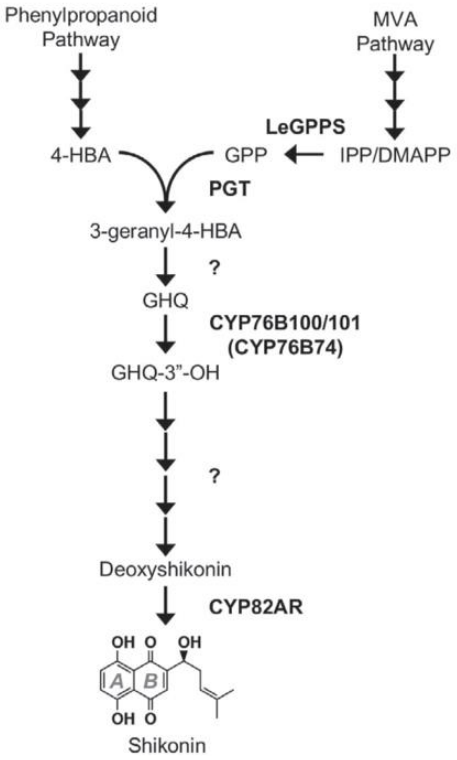

Figure 3-1 The shikonin metabolic network. Depicted is the current understanding of the enzymes and intermediates involved in synthesizing shikonin from precursors of the phenylpropanoid (4-HBA) and the MVA (GPP) pathways. Question marks indicate proposed steps lacking experimental evidence. Abbreviations: 4-HBA, 4-hydroxybenzoate; CYP76B74 (*Arnebia euchroma*) and CYP76B100/101 (*Lithospermum erythrorhizon*), GHQ 3″-hydroxylase; CYP82AR, deoxyshikonin hydroxylase; DMAPP, dimethylallyl diphosphate; GHQ, geranylhydroquinone; GPP, geranyl diphosphate; GPPS, geranyl diphosphate synthase; IPP, isopentenyl diphosphate; MVA, mevalonic acid; PGT, p-hydroxybenzoate:geranyltransferase.

The shikonin pathway begins with the conjugation of 4-hydroxybenzoate (4-HBA) and geranyl diphosphate (GPP) catalyzed by p-hydroxybenzoate:geranyltransferase (PGT)[13] to produce 3-geranyl-4-HBA[14] (Fig. 3-1). GPP and other prenyl diphosphates are synthesized from the condensation of the five-carbon building blocks isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). In plants, GPP is typically produced by plastidial GPP synthases (GPPSs) that catalyze the condensation of one IPP and one DMAPP derived from the methylerythritol phosphate (MEP) pathway localized in plastids[15]. Plants also produce IPP and

DMAPP via the mevalonic acid (MVA) pathway, a route separated from the MEP pathway that is compartmentalized across the cytoplasm, endoplasmic reticulum, and peroxisomes[15-17]. The MVA pathway is generally considered to generate isoprenoid precursors for farnesyl diphosphate (FPP) synthases (FPPSs), which catalyze the condensation of one DMAPP with two IPP molecules to produce FPP and two molecules of pyrophosphate in the cytoplasm. Experimental evidence by Gaisser and Heide[18] and others (reviewed in Widhalm and Rhodes[4]) long suggested that shikonin biosynthesis unconventionally relies on GPP produced by a cytoplasmic GPPS. The recent discovery and biochemical characterization of *L. erythrorhizon* GPPS (LeGPPS, Fig. 3-1)[9] revealed that it is a neofunctionalized cytoplasmic farnesyl diphosphate synthase (FPPS) and that mutation(s) adjacent to the first aspartate-rich motif resulted in acquisition of GPPS activity[19]. Previous analysis by our group of the *L. erythrorhizon* genome uncovered an evolutionary link between PGTs and the ubiquinone prenyltransferase gene, demonstrating that retrotransposition-derived gene duplication and subsequent neofunctionalization contributed to the evolution of *PGT* genes[20]. Coupled with the evolution of *LeGPPS* from a cytoplasmic FPPS[19] and whole genome duplication (WGD) in the Boraginaceae[20,21], the evolutionary history of the shikonin pathway appears to be marked by several events of metabolic innovation. Taken together, this raises the prospect of additional evolutionary links between the shikonin and ubiquinone pathways and opens new questions about the metabolic intersection of the isoprenoid, phenylpropanoid, ubiquinone, and shikonin pathways in the Boraginaceae. In this study, we investigated the metabolic connections linking shikonin biosynthesis with its precursor pathways by downregulating expression of *LeGPPS* and testing the capacity of the MEP and MVA pathways to supply GPP for shikonin production. We also explored whether network analysis of transcript abundances could identify genes coexpressed with *LeGPPS* and other established shikonin pathway genes. Integrative computational analyses of candidate genes identified by the model suggest likely metabolic roles for these genes and give insight into the evolution of metabolic innovation in the shikonin pathway. Our study provides evidence of crosstalk between the MVA, MEP, and phenylpropanoid pathways and reveals additional evolutionary links between shikonin and ubiquinone biosynthesis. Given the other links between specialized and primary quinone metabolism[22,23], the mechanistic insights uncovered here are expected to broadly guide investigation into the convergent evolution of specialized 1,4-naphthoquinone metabolism in plants.

3.4 Results

3.4.1 Cytoplasmic *LeGPPS* supplies GPP to the shikonin pathway using MVA pathway-derived IPP/DMAPP

To investigate the in vivo role of *LeGPPS*, which sits at the interface between the MVA, phenylpropanoid, and shikonin pathways, we knocked down expression of its encoding gene in *L. erythrorhizon* hairy roots. Several independent *LeGPPS*-RNAi (*LeGPPSi*) lines were generated, excised, and transferred to B5 selection media plates and subsequently screened for levels of total shikonins excreted into the growth media 4 d after transfer to M9 and darkness. Analysis of 10 independent *LeGPPS*-RNAi lines revealed that total shikonins were reduced by more than 95% compared to the lowest producing emptyvector control line (Fig. 3-2a). Further analysis of two independent lines, *LeGPPSi*-45 and *LeGPPSi*-75, revealed that *LeGPPS* expression was reduced by more than 95% compared to empty-vector control EV-26 without any affect on expression of the canonical plastidial GPPS gene *LepGPPS* (Fig. 3-2b). Re-analysis of total shikonins excreted from *LeGPPSi* lines 45 and 75 confirmed nearly 95% reduction compared to EV-26 (Fig. 3-2c), thus indicating that cytoplasmic *LeGPPS* is predominantly responsible for supplying GPP precursor to the shikonin pathway. Concievably, MEP pathway-derived GPP could contribute to the shikonin pathway if it or MEP pathwayderived IPP/DMAPP were exported from the plastid to the cytoplasm and used as substrate by LeGPPS. To test for MEP pathway involvement in shikonin production we carried out two inhibitor experiments on the EV26 and *LeGPPSi*-45 lines (Fig. 3-1). We predicted that if the MVA pathway is predominantly responsible for supplying IPP/DMAPP to LeGPPS, then treatment with the MVA pathway inhibitor mevinolin should decrease shikonin accumulation in EV-26 lines but not in the *LeGPPSi*-45 RNAi line. Indeed, total shikonins produced by mevinolin-treated EV-26 lines were reduced by 76% compared to those in the EV-26 control lines (Fig. 3-3a), while shikonins in mevinolin-treated *LeGPPSi*-45 lines were unchanged compared to the *LeGPPSi*-45 control lines (Fig. 3-3b). If the MEP pathway does not supply IPP/DMAPP precursor to the shikonin pathway, we expected no change in shikonin accumulation in EV-26 lines treated with the MEP pathway inhibitor fosmidomycin compared to controls. If, however, the MEP pathway is contributing to the remaining shikonin produced by *LeGPPSi*-45 lines, treatment with fosmidomycin should further reduce shikonin accumulation compared to *LeGPPSi*-45 controls. Instead, we observed that shikonin production increased by 73%

and 108%, respectively, in EV-26 and *LeGPPSi*45 lines treated with fosmidomycin compared to their corresponding controls (Fig. 3-3a,b). This result points to crosstalk between the MEP and MVA pathways such that when flux through the MEP pathway is impaired, flux through the MVA pathway is increased. Taken together, our genetic and inhibitor studies support the work of Ueoka et al.[19] by showing that *LeGPPS* is required for shikonin formation and it shows that the MEP pathway does not supply IPP or DMAPP substrates, or direct GPP precursor to the shikonin pathway.

3.4.2 Downregulation of *LeGPPS* reveals crosstalk between phenylpropanoid and isoprenoid metabolism

The observed increase in shikonin content in *LeGPPSi*-45 RNAi lines treated with the MEP pathway inhibitor fosmidomycin (Fig. 3-3b) led us to hypothesize that the smaller pool size of shikonin in LeGPPS-RNAi lines (Fig. 3-2) may be due, in part, to an upstream effect on the MVA pathway. To investigate if MVA pathway gene expression is changed, we performed RNA-seq analysis of *LeGPPSi*-45 lines compared to EV-26 control. This analysis confirmed that *LeGPPS* but not *LepGPPS* is significantly downregulated the *LeGPPSi*-45 line (Fig. S1a). Our analysis showed 6115 differentially expressed genes (DEGs); 2903 genes were significantly overexpressed in *LeGPPSi*-45 lines compared to EV-26 while 3212 were significantly underexpressed (Table S1), including shikonin pathway genes *LePGT1*, *LePGT2*, *CYP76B100*, and *CYP82AR* (Fig. S1b). Kyoto Encyclopedia of Genes and Genomes (KEGG) term enrichment analysis of genes underexpressed in the *LeGPPSi*-45 line revealed an enrichment of genes involved in various metabolic pathways connected to shikonin metabolism (BH-adjusted p-value <0.05; Fig. 3-4a, Fig. S2). The category "monoterpenoid biosynthesis," which encompases metabolic genes downstream of GPP was significantly enriched among underexpressed genes. The KEGG category "terpenoid backbone biosynthesis," which contains the MVA and MEP pathway genes, was not significantly enriched (BH-adjusted p-value = 0.097; Fig. S2). Yet, 11 of the 17 MVA pathway genes involved in IPP biosynthesis were found to be significantly underexpressed in *LeGPPSi*-45. This included six of the eight genes encoding 3-hydroxy-3-methylglutaryl-CoA reductase (HMGR), which is generally considered to catalyze the rate-limiting step of the MVA pathway (Fig. 3-4a, Table S1)[24]. This suggests that lower expression of upstream MVA pathway genes may have contributed to reduced shikonin production in LeGPPS-RNAi lines (Fig. 3-2). These data also point to an

unknown factor connecting downregulation of *LeGPPS* with reduced expression of upstream MVA pathway genes.

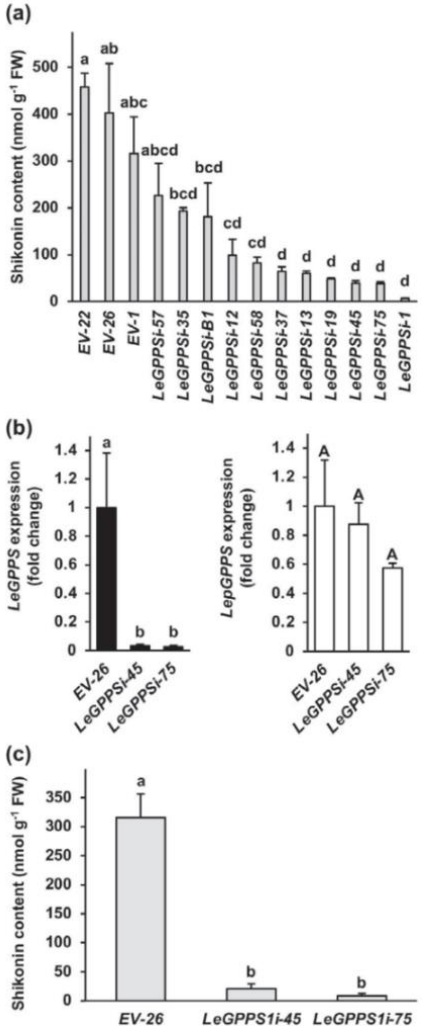

Figure 3-2 In vivo characterization of *LeGPPS*. Screening of *LeGPPS*-RNAi (*LeGPPSi*) lines based on total shikonin levels present in liquid culture media 3 d after transfer of 14-d-old hairy roots to M9 and darkness (a). Expression levels of *LeGPPS* and the canonical plastid-localized GPPS gene (*LepGPPS*) in hairy roots of two independent *LeGPPSi* lines compared to an empty-vector control line (*EV*-26) (b). Analysis of total shikonin in the same cultures used to measure expression in panel b (c). All data are means ± SEM (n = 3–4 biological replicates). Different letters indicate significant differences via analysis of variance (ANOVA) followed by post-hoc Tukey test ($\alpha = 0.05$). In panel b, lowercase and capital letters correspond to statistical comparisons for *LeGPPS* and *LepGPPS* expression, respectively.

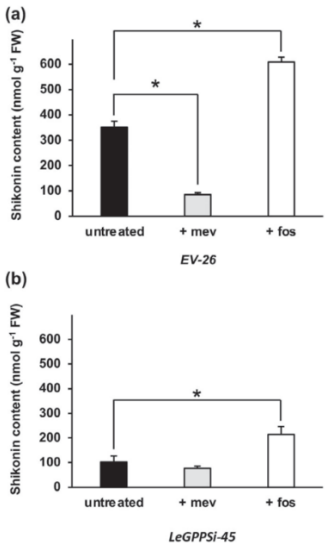

Figure 3-3 Effect of MVA and MEP pathway-specific inhibitors on formation of total shikonins. Total shikonin levels present in liquid culture media were measured in empty-vector control line 26 (*EV*-26) (a) and *LeGPPS* RNAi line 45 (*LeGPPSi*-45) (b) following mock treatment or treatment with 100 μM of the MVA pathway inhibitor mevinolin (+ mev) or the MEP pathway inhibitor fosmidomycin (+ fos). Inhibitor treatments were administered immediately upon transfer of 14-d-old hairy roots to M9 and darkness. Total shikonins were measure at 6 d after transfer of 14-d-old hairy roots to M9 and darkness. All data are means ± SEM (n = 3–4 biological replicates). Statistically significant differences are indicated (*P < 0.05, Student's *t* test).

The KEGG pathway analysis also revealed that genes involved in "phenylpropanoid biosynthesis" and "ubiquinone and other terpenoid-quinone biosynthesis" were enriched in those underexpressed in *LeGPPSi*-45 (Fig. S2). This is noteworthy because the phenylpropanoid pathway supplies p-coumaroyl-CoA to make the 4-HBA precursor that becomes the hydroxybenzene ring, ring A, of shikonin's naphthazarin moiety (Fig. 3-1) and of ubiquinone's benzenoid moiety[8]. Further examination of genes underexpressed in *LeGPPSi*-45 showed that several genes in the core phenylpropanoid pathway are underexpressed, including multiple genes encoding phenylalanine ammonia-lyases (PALs) (Fig. 3-4b, Table S1). Moreover, one copy of the At4g19010-like peroxisomal p-coumarate-CoA ligase genes (Leryth_018919) was significantly underexpressed (Fig. 3-4b, Table S1). In *Arabidopsis thaliana*, it was demonstrated that At4g19010 is responsible for activating the propyl side chain of p-coumarate for β-oxidative shortening to supply 4-HBA precursor for ubiquinone biosynthesis[8]. These results suggest that, like the MVA pathway, an unknown factor links downregulation of *LeGPPS* to reduced expression of phenylpropanoid and benzenoid pathway genes.

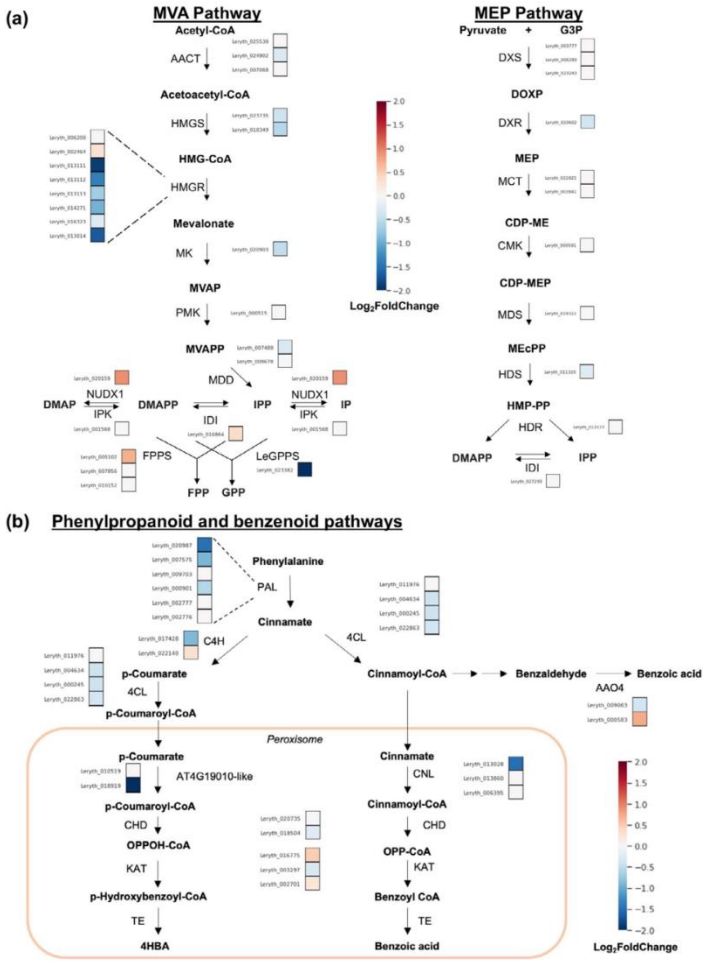

Figure 3-4 Effect of *LeGPPS* RNAi downregulation on expression of MVA, MEP, phenylpropanoid, and benzenoid pathway genes. The average log_2fold-change in expression for each gene in *LeGPPSi*-45 lines compared to *EV-26* lines in the mevalonic acid (MVA) and methylerythritol phosphate (MEP) pathways (a) and in the phenylpropanoid and benzenoid pathways (b) are shown. Abbreviations: 4CL, 4-coumarate CoA-ligase; AACT, acetoacetyl-CoA thiolase; AA04, Arabidopsis Aldehyde Oxidase 4; C4H, cinnamate 4-hydroxylase; CDP-ME, 4-Diphosphocytidyl-2-C-methylerythritol; CDP-MEP, 4-Diphosphocytidyl-2-C-methylerythritol 2-phosphate; CHD, cinnamoyl-CoA hydratase/dehydrogenase; CMK, 4-(cytidine 5′ -diphospho)-2-C-methyl-D-erythritol kinase; CoA, coenzyme A; DMAP, dimethylallyl phosphate; DMAPP, dimethylallyl diphosphate; DOXP, 1-deoxy-D-xylulose 5-phosphate; DXR, 1-deoxy-D-xylulose 5-phosphate reductoisomerase; DXS, 1-deoxy-D-xylulose 5-phosphate synthase; FPP, farnesyl diphosphate; FPPS, farnesyl diphosphate synthase; G3P, D-glyceraldehyde 3-phosphate; GPP, geranyl diphosphate; GPPS, geranyl diphosphate synthase; HDR, (E)-4-hydroxy-3-methylbut-2-enyl diphosphate reductase; HDS, (E)-4- hydroxy-3-methylbut-2-enyl diphosphate synthase; HMG-CoA, 3-hydroxy-3-methylglutaryl-CoA; HMGR, 3-hydroxy-3-methylglutaryl-CoA reductase; HMGS, 3-hydroxy-3-methylglutaryl-CoA synthase; HMP-PP, (E)-1-hydroxy-2-methylbut-2-enyl 4-diphosphate; IDI, isopentenyl diphosphate isomerase; IP, isopentenyl phosphate; IPK, isopentyl phosphate kinase; IPP, isopentenyl diphosphate; KAT, 3-ketoacylthiolase 1; MCT, 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase; MDD, mevalonate diphosphate decarboxylase; MDS, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; MEcPP, methylerythritol cyclodiphosphate; MK, mevalonate kinase; MPD, phosphomevalonate decarboxylase; MVAP, mevalonate 5-phosphate; MVAPP, mevalonate diphosphate; NUDX1, Nudix enzyme 1; OPP-CoA, 3-oxo-3-phenylpropionoyl-CoA; PAL, L-phenylalanine ammonia lyase; PMK, phosphomevalonate kinase; PXA1, peroxisomal ABC transporter 1; TE, thioesterase.

3.4.3 Coexpression network analysis recovers known shikonin pathway gene associations and predicts new connections

We hypothesized that *LeGPPS* and other known shikonin biosynthesis genes, including *LePGT1*, would appear as hub genes that we could use to identify coexpressed genes with roles in shikonin biosynthesis. To construct a transcriptional network model with a high likelihood of recovering the shikonin biosynthetic pathway as a module, we used publicly available comparative RNA-seq experiments from tissues and conditions divergent in their shikonin levels (NCBI Sequence Read Archives PRJNA596998[20] and PRJNA331015). These included whole *L. erythrorhizon* root tissue versus above ground tissue; root periderm versus root vascular (inner) tissue; and hairy root cultures grown in M9 media in the dark versus roots grown in B5 media in light conditions. In all three experiments, the former tissue or condition in each comparison was previously shown to contain higher *LePGT1* expression and shikonin content[20]. Like *LeGPPS*, *LePGT1* functions at the interface of the phenylpropanoid, MVA, and shikonin pathways (Fig. 3-1). Therefore, we constructed the model based on the hypothesis that genes involved in the shikonin pathway and upstream metabolism would also be more highly expressed in the same tissues or conditions as LePGT1.

One potential source of noise in coexpression analyses is the inclusion of genes that are either not expressed or are constitutively expressed at a constant level across all conditions. These genes may appear significantly coexpressed with other genes in the dataset artifactually[25]. Although this is less of a concern when working with dozens or hundreds or RNA-seq samples[26], our dataset consisted of only 14 RNA-seq samples across six total tissues or conditions. To control for this source of false positive coexpression, we only included genes likely to be DEGs in at least one of the three comparisons at a false discovery rate (FDR) cutoff of ≤ 0.1. A total of 8680 transcripts were included using this approach (Table S2). Within this set of transcripts, 23.9% (2077) were overexpressed in whole root; 37.9% (3290) were overexpressed in root periderm; and 21.6% (1876) were overexpressed in hairy roots sampled in the dark (Table S3). The overlap of all three comparisons contained 374 genes that were overexpressed in the shikonin accumulating condition (Fig. 3-5a). These included several genes already implicated in shikonin biosynthesis: *LePGT1*, *LePGT2*, two additional PGT-like genes[20], *LeGPPS*[19], *CYP82AR2*[27], *L. erythrorhizon* pigment callus-specific gene 2 (*LePS-2*)[28], and *LeMYB1*[29] (Table S4).

Decreased accumulation of transcripts encoding core phenylpropanoid and β-oxidative benzenoid biosynthetic genes (Fig. 3-4b) raises the possibility that 4-HBA availability might also limit shikonin production in GPPSi-RNAi lines (Fig. 3-2). To test this, shikonin levels were determined in EV-26 and *LeGPPSi*-45 lines supplied with exogenous 4-HBA. The amount of shikonin produced, however, remained unchanged compared to the unfed controls (Fig. S3) suggesting that 4-HBA availability does not limit shikonin production in *LeGPPS* knockdown lines. Taken together, the in vivo investigation of *LeGPPS* demonstrates that in addition to *LeGPPS* being involved in shikonin biosynthesis, the expression of *LeGPPS* is highly connected to other genes in the larger shikonin metabolic network including those in the shikonin, MVA, phenylpropanoid, and benzenoid pathways.

The 8680 DEGs were used as input for a global coexpression network analysis (Table S5). Pairwise measurements of gene coexpression were specified as mutual ranks (MRs), which are calculated as the geometric mean of the rank of the Pearson's correlation coefficient (PCC) of gene A to gene B and the PCC rank of gene B to gene A[30]. Ranking the PCCs in this manner has been shown to improve the recovery of known pathways as discrete subgraphs in global coexpression networks[31]. We constructed four MR-based networks (N1-N4), using different coexpression thresholds for assigning edge weights (i.e. connections) between nodes (i.e. genes) in the network. Networks were ordered by size (i.e. total number of edges between nodes), such that N1 represents the smallest network and N4 represents the largest network. Graph-clustering implemented by ClusterONE[32] was used to discover coexpressed subgraphs (hereafter referred to as gene modules) within the global networks (Dataset S1). The benefit of using ClusterONE over other graph-clustering methods, e.g. MCL[33] is its capacity to assign genes to multiple overlapping modules, which is more reflective of complex biological networks. We chose to focus our analysis on four target genes based on evidence of their involvement in shikonin metabolism: *LeGPPS*[19] (Fig.3-2), *LePGT1*[20], *LeCYP76B101*[34,35], and *LeMDR*[36]. Because ClusterONE modules can overlap, each target gene was assigned to multiple modules within the larger networks. For example, *LePGT1* was found in 3, 2, 3, and 6 different modules in network N1, N2, N3, and N4, respectively (Dataset S1). To address this redundancy, we collapsed all modules within a network that contained one or more of the four target shikonin metabolic genes into non-intersecting metamodules (Fig. 3-5b; Fig. S4)[26]. Collectively, these metamodules are models, which we refer to as shikonin metabolic subnetworks.

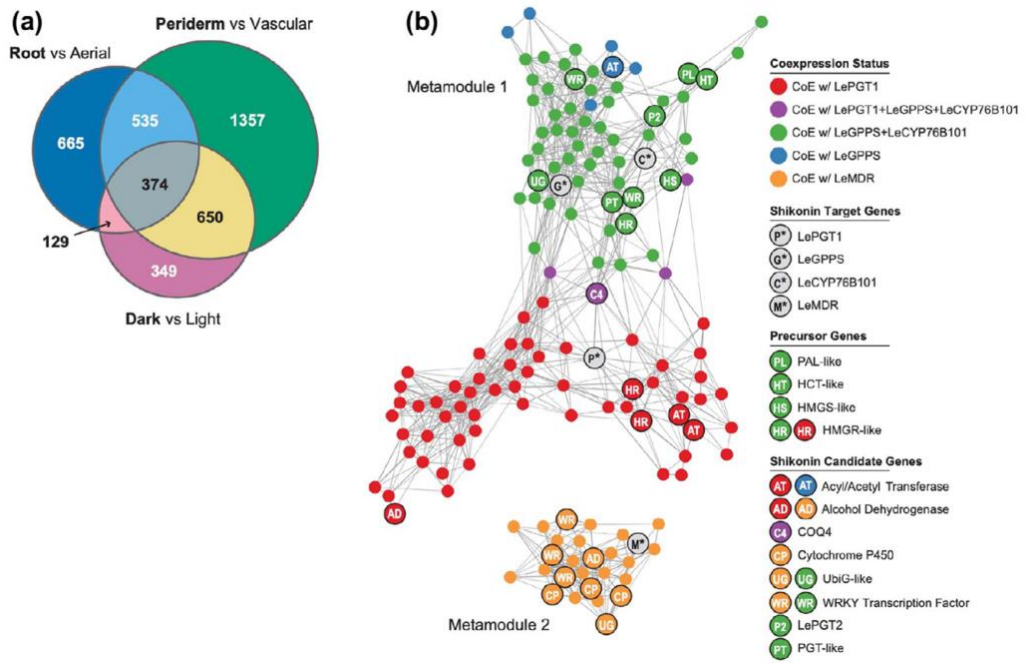

Figure 3-5 Analysis of gene expression in *Lithospermum erythrorhizon*. Venn diagram showing the overlap of genes that are significantly overexpressed in conditions where shikonin is abundant (bold) (a). Network map of genes coexpressed with target genes *LePGT*, *LeGPPS*, *LeCYP76B101*, and *LeMDR* using the N2 global coexpression network (b). Nodes in the map represent genes, and edges connecting two genes represent the weight (transformed MR score) for the association. Genes are colored according to its coexpression status with known shikonin genes (grey). Network maps were drawn using a Fruchterman-Reingold force-directed layout using the edge-weighted spring embedded layout in cytoscape (https://cytoscape.org)

The number of genes recovered in the shikonin metabolic subnetworks varied from 102, 152, 359, and 1268 genes in networks N1, N2, N3, and N4, respectively (Tables S6–S9). We focused our subsequent analyses on the N2 network, which contained a large number of candidate genes to investigate while also limiting the number of peripheral genes that appeared only weakly connected to shikonin biosynthesis (Fig. S4). The N2 shikonin metabolic subnetwork was comprised of two metamodules (Fig. 3-5b). The first N2 metamodule contained 125 genes including *LeGPPS*, *LePGT1*, and *LeCYP76B101*; whereas, the second metamodule contained 27 genes including *LeMDR*. To be considered coexpressed in our analysis two genes must have at least one shared module within the larger metamodule. For example, *LeGPPS* and *CYP76B101* were coexpressed with one another, being members of three shared modules: N2M94, N2M298, and N2M317 (Dataset S1). Within metamodule 1, 60 genes were coexpressed with *LeGPPS* and *CYP76B101*; 6 genes were uniquely coexpressed with *LeGPPS*; and 59 genes were uniquely coexpressed with *LePGT1* (Fig. 3-5b; Table S7). Four genes (Leryth_014746, Leryth_025160,

Leyrth_004583, Leryth_002195) coexpressed with all three *LeGPPS*, *LePGT1*, and *LeCYP76B101* (Fig. 3-5b; Table S7). Although the genes of metamodule 2, including *LeMDR*, were not coexpressed with the three other target shikonin genes in network N2, the two larger networks N3 and N4 did show a small amount of overlap (Fig. S4).

Table 3-2 Shikonin pathway gene candidates identified via coexpression network analysis

Gene	InterPro	Network	Coexpressed with	DE DvL	DE PvV	DE RvL
Leryth_021809	Cytochrome P450	N2	*LeMDR*	yes	yes	yes
Leryth_001242	Cytochrome P450	N2	*LeMDR*		yes	yes
Leryth_000257	Cytochrome P450	N2	*LeMDR*		yes	
Leryth_002195	Ubiquinone biosynthesis protein COQ4	N2	*LeGPPS*, *LePGT*, *CYP76B101*	yes	yes	yes
Leryth_019821	Ubiquinone biosynthesis O-methyltransferase COQ3	N2	*LeMDR*		yes	yes
Leryth_021171	Ubiquinone biosynthesis O-methyltransferase COQ3	N2	*LeGPPS*, *CYP76B101*	yes	yes	
Leryth_001358	Prephenate dehydrogenase	N2	*LePGT*	yes	yes	
Leryth_020454	Quinoprotein glucose/sorbosone dehydrogenase	N4	*LeGPPS*, *LePGT*, *CYP76B101*	yes	yes	yes
Leryth_012925	Chloramphenicol acetyltransferase-like	N2	*LePGT*		yes	yes
Leryth_015823	Chloramphenicol acetyltransferase-like	N2	*LeGPPS*, *CYP76B101*		yes	yes

In agreement with previous studies[37,38], six genes were recovered in N2 metamodule 1 encoding enzymes with annotations related to the phenylpropanoid and MVA pathways including PAL, HCT, HMGS, and HMGR (Fig. 3-5b; Table S7). Past studies have identified additional candidate genes possibly involved in the shikonin pathway including LePS-2[28], LeACS-1[39], LeMYB1[29], and LeDI-2[40]. Of these, only LePS-2 was coexpressed with any validated shikonin biosynthetic genes, being coexpressed with LePGT1, LeGPPS, and LeCYP76B101 in the larger N3 and N4 networks (Tables S8,S9).

The N2 shikonin subnetwork was enriched in 62 Gene Ontology (GO) categories (BH-adjusted p-value <0.05; Table S10) including broad enzymatic categories such as GO:0016491 oxidoreductase activity (18 genes), and GO:0016740 transferase activity (36 genes). Another enriched category, ATPase-coupled intramembrane lipid transport activity (2 genes;

Leryth_023505, Leryth_019206) is of high interest because the previous implication of an ARF/GEF-like system required for shikonin transport[41].

To identify shared 5′ cis regulatory regions among the coexpressed genes, we performed a motif enrichment analysis on the genes of the N2 shikonin subnetwork using Motif Indexer[42]. The most overrepresented motif within the upstream region of shikonin subnetwork genes was AmrGTCwA (p-value = 9.67x10−10; FDR = 0.007; Table S11), the reverse compliment of which (TwGACykT) is similar to the canonical W-box element sequence motif (T)TGAC(C/T) recognized by the WRKY family of transcription factors[43]. Of the 152 genes in the N2 shikonin subnetwork, 47.37% of genes (N = 72) contained this motif including all four target genes: LePGT1, LeGPPS, LeMDR, and CYP76B101. Five WRKY transcription factors were identified in the N2 shikonin subnetwork (Fig. 3-5b) two of which (Leryth_027519 and Leryth_002564) were also significantly overexpressed in all three conditions where shikonin was abundant (Table S7).

3.4.4 Expansion of the LeFPPS gene family in the Boraginales gave rise to LeGPPS

We next performed a phylogenetic analysis to gain insight into the evolutionary events giving rise to genes in the shikonin metabolic network. Previous work demonstrated that *LeGPPS* encodes an enzyme having GPPS-like activity but is a member of the FPPS gene family[19]. To better understand the evolutionary history of *LeGPPS*, we reconstructed the phylogeny of the FPPS gene family using homologous sequence groups downloaded from the PLAZA 4.0 database (Table S12)[44]. In addition to *L. erythrorhizon*, we included in our analysis de novo transcriptome-based proteomes from 18 additional Boraginales species including three other shikonin producing plants (*E. plantagineum, Arnebia euchroma, and Lithospermum officinale*), one additional Boraginaceae that does not produce shikonin (*Mertensia paniculata*), and 14 additional Boraginales species that do not produce shikonin (Table S13)[20,45]. The Boraginales contain two distinct subfamilies in the FPPS gene family phylogeny (Fig. S7). Subfamily I contains *LeGPPS* and was present in 17 of the 19 Boraginales in the analysis, including all four shikonin-producing species (Fig. S7). Subfamily I was absent in *Heliotropium karwinsky* and *Heliotropium sp.* Subfamily II, the canonical FPPS group, contains LeFPPS1 (Leryth_005102) and 29 other sequences. Subfamily II was present in all four shikonin-producing species and absent in *Heliotropium calcicole* and *Heliotropium*

texanum. The absence of subfamily I or II in some *Heliotropium* species is likely artifactual due to these gene sets being transcriptome derived. Subfamily II also contained two additional genes from *L. erythrorhizon*, Leryth_007856 (referred to as LeFPPS2 by Ueoka et al.[20]) and Leryth_010152 (hereafter LeFPPS3) (Fig. S7).

We searched for shared synteny between genome assembly contigs containing FPPS genes in *L. erythrorhizon* to investigate whether whole genome duplication (WGD) was involved in the evolution of the subfamily I. A WGD is proposed for the Boraginaceae roughtly 25 MYA[21] and *L. erythrorhizon* and *E. plantagineum* have similar distributions of synonymous substitution (Ks) between syntenic paralogs at 0.45 and 0.417, respectively[20,21]. The contigs containing *LeFPPS1* and *LeFPPS3* (Fig. S5) were syntenic and the syntelogs in these two contigs have a median Ks value of 0.484 (Table S14). The median Ks of this syntenic block is similar to the peaks in Ks distribution described by Auber et al.[20] and Tang et al.[21], consistent with the Boraginaceae WGD giving rise to *LeFPPS1* and *LeFPPS3*. In contrast, the lack of shared synteny between *LeGPPS* and any of the three genes in the FPPS group suggests that *LeGPPS* did not arise via WGD. Intron position is conserved between *LeGPPS* and the three FPPS genes, with only *LeFPPS3* showing some divergent intron positioning toward its 3′ end (Fig. S6), which is consistent with segmental duplication or DNA transposition giving rise to the *LeGPPS* homolog rather than retrotransposition.

Previous work by Ueoka et al.[19] demonstrated that the histidine (His) residue adjacent to the first aspartate-rich motif in *LeGPPS* was responsible for its GPPS-like activity. Examination of our FPPS gene family sequence alignment shows that this His residue is present in all sequences of the GPPS group (subfamily I), with the exception of two transcriptome-derived sequences from two non shikonin-producing species *Ehretia acuminata* and *Heliotropium greggii* that are both missing this region (Fig. S7). In contrast, all sequences in the FPPS group (subfamily II) contain the canonical leucine (Leu) residue adjacent to the Asp-rich motif, with the exception of three transcriptome-derived sequences that are missing the region (Fig. S7). We identified a His residue in place of Leu in three additional sequences from *Fragaria vesca* (woodland strawberry*), Pyrus bretschneideri* (Chinese white pear), and *Solanum tuberosum* (potato) (Figs. S7,S8). Like the Boraginales, each of these three species maintained a second FPPS gene that retains the canonical Leu adjacent to the Asp-rich motif (Fig. S8). A similar observation was made with FPPS homologs from *Fragaria x ananassa* (strawberry), *Malus domestica* (apple), and *Prunus persica* (peach)[19].

Thus, the recruitment of a cytoplasmic FPPS to function as a GPPS convergently evolved multiple times in plants and has likely contributed to the diversification of plant terpenoid metabolism.

3.4.5 Shikonin pathway gene candidates provide insights into specialized metabolic innovation in the Boraginaceae

We extended our phylogenetic analysis to additional shikonin gene candidates (Table 3-1). We first considered gene candidates that could be responsible for missing enzymes in the shikonin pathway. It is estimated that 97% of cytochromes P450 in plants are associated with specialized metabolic pathways[46]. Therefore, considering that missing steps in the shikonin pathway require decarboxylation, hydroxylations, or carbon–carbon ring closure, we examined cytochromes P450 in the coexpression network. In addition to *LeCYP76B101*, which was used as a known target in metamodule construction, three additional cytochromes P450 were recovered in the N2 shikonin subnetwork, all three of which were coexpressed with *LeMDR* metamodule 2 (Fig. 3-5b; Table S7). None of the three additional cytochromes P450 correspond to the *LeCYP82AR2* recently described to catalyze deoxyshikonin hydroxylation in vitro[27]. Although *LeCYP82AR2* (Leryth_026973, Table S3) was not recovered as a candidate in the N1 or N2 shikonin subnetworks, it was coexpressed with *LePGT1*, *LeGPPS*, and *LeCYP76B101* in our N3 network (Fig. S4c) and was also overexpressed in all shikonin-abundant conditions (Table S3).

One of the three cytochromes P450 identified was Leryth_021809, which encodes a CYP76B6-like enzyme and was significantly overexpressed in all shikonin-abundant conditions (Table S3). Other CYP76B genes, including *CYP76B74* in A. euchroma[35] and *CYP76B100*/*101*[34] in *L. erythrorhizon* (Fig. 3-1), have already been implicated in oxidative reactions in shikonin biosynthesis, but the evolutionary relationships between these genes has been unclear. A phylogeny of CYP76B6-like genes reveals that *AeCYP76B74* and *LeCYP76B101* are orthologs (Fig. S9). *LeCYP76B100* is the mostly closely related paralog to *LeCYP76B101* but groups more closely to other sequences in *A. euchroma*, *E. plantagineum*, and *Mertensia paniculata* (Fig. S9). This indicates that the gene duplication event that gave rise to *LeCYP76B100/101* occurred in the last common ancestor of these Boraginaceae species. Leryth_021809, the additional CYP76B6-like gene recovered in the coexpression analysis, is within a separate clade that has expanded within shikonin producing species. The closest homolog in *M. paniculata* (the only Boraginaceae

in our analysis that does not produce shikonin) groups closer to a different cytochrome P450 in *L. erythrorhizon* (Leryth_021691; Fig. S9).

A phylogeny of sequences homologous to the second cytochrome P450 candidate (Leryth_001242), which encodes a CYP76A2-like enzyme, also shows an expansion of gene copies in the Boraginaceae (Fig. S10). This gene is one of three homologs in a tandem repeat, including Leryth_001243 and Leryth_001244 indicating that tandem gene duplication has expanded this cytochrome P450 subfamily in *L. erythrorhizon* (Fig. S11). Leryth_001243 and Leryth_001244 were not captured in the shikonin subnetwork but their expression is greater in whole root versus above ground tissue (Table S3). Lastly, the phylogeny of sequences homologous to the third cytochrome P450 candidate (Leryth_000257), which encodes a CYP89A2-like enzyme, shows a smaller group of Boraginales sequences without the rounds of expansion present in the other two trees (Fig. S12).

The production of geranylhydroquinone (GHQ) from 3-geranyl-4HBA (Fig. 3-1) may occur via decarboxylation and subsequent hydroxylation or a single oxidative decarboxylation event[4]. In addition to cytochromes P450, we examined the generated shikonin network for non-cytochrome P450 candidate genes that may function in either hypothesized mechanism. One candidate to consider is a prephenate dehydrogenase-like (PDH-like) gene. PDH catalyzes oxidative decarboxylation of prephenate to 4-hydroxyphenylpyruvate for synthesis of tyrosine[47]. Leryth_001358 encodes a PDH-like protein and is coexpressed with *LePGT1* in the N2 subnetwork (Fig. 3-5b; Table S7). Similar to other genes in our analysis, the phylogeny of PDHs shows an expansion of this gene family within the Boraginaceae (Fig. S13). Coexpression of the PDH-like gene may simply be related to the connection between shikonin and aromatic amino acid metabolism via phenylpropanoid metabolism, further research is needed to determine if a duplicated PDH could evolve to utilize another 4-hydroxylated substrate.

Given the dozens of shikonin and alkannin derivatives collectively present in the Boraginaceae[48], we looked for genes in the shikonin subnetwork that may encode tailoring enzymes involved in the synthesis of shikonin derivatives. Recently, two BAHD acyltransferases, shikonin O-acyltransferase (*LeSAT1*) and alkannin O-acyltransferase (*LeAAT1*), were discovered to mediate enantiomer-specific acylation in *L. erythrorhizon*[49]. Neither *LeSAT1* nor *LeAAT1* were recovered in the coexpression networks but expression of both is more abundant in at least one shikonin-abundant condition (Table S3). In our N2 shikonin subnetwork (Fig. 3-5b; Table S7),

two additional genes encoding putative transferases (Leryth_012925, Leryth_015823) were recovered. Phylogenetic analysis of the Leryth_015823 transferase and its homologs places Leryth_015823 in a group that contains all Boraginaceae species in our analysis (Fig. S14). In contrast, phylogenetic analysis of Leryth_012925 and its homologs shows Leryth_012925 on a long branch and lacking closely related homologs in other Boraginaceae (Fig. S15), which may make it a potential candidate for a *L. erythrorhizon*-specific shikonin/alkannin tailoring enzyme that is absent in the other shikonin-producing species in our analysis.

3.4.6 Coexpression network analysis reveals candidates with links to ubiquinone biosynthesis

It has already been demonstrated that LePGT1 and LePGT2 evolved via duplication of a primary metabolic prenyltransferase involved in ubiquinone biosynthesis[20]. Given this previously observed connection, the coexpression of a COQ4 ubiquinone biosynthesis-like gene with *LePGT1*, *LeGPPS*, and *LeCYP76B101* in the N2 network appeared remarkable (Fig. 3-5b; Table S7). Although the precise biochemical function of COQ4 is unknown it is thought to function as a scaffold protein binding proteins and lipids required for efficient ubiquinone biosynthesis[50]. The phylogenetic tree of the COQ4 gene family (Fig. S16) is strikingly similar to that of the ubiquinone prenyltransferase gene family[20]. Both phylogenies contain two subfamilies of Boraginales sequences. One subfamily has shorter branch lengths and contains a single sequence per species, suggesting that this subfamily has retained the ancestral COQ4 ubiquinone biosynthesis activity (Fig. S16). The second Boraginales subfamily has longer branches and shows a radiation of COQ4 paralogs and includes the candidate gene (Leryth_002195; Fig. S16), which was overexpressed in all shikonin-abundant conditions (Table S3). Given the similarities in the precursors and biosynthetic steps in the ubiquinone and shikonin pathways[20], this COQ4 paralog (Leryth_002195) could fulfill an analogus function and participate in assembling a shikonin biosynthesis metabolon.

In addition to the COQ4-like genes, we al identified two COQ3-like O-methyltransferase genes in the N2 shikonin subnetwork (Leryth_019821 and Leryth_021171). Leryth_019821 was coexpressed with *LeMDR* and Leryth_021171 was coexpressed with *LeGPPS* and *LeCYP76B101* (Fig. 3-5b). A phylogenetic tree of the COQ3 gene family suggests that the two copies in *L. erythrorhizon* diverged in an ancestor of the Boraginaceae; the Leryth_021171 subfamily contained a sequence from *M. paniculata*, whereas the Leryth_019821 subfamily appears to be

unique to shikonin producing species (Fig. S17). The metabolic significance of this network connection remains enigmatic, though it is possible that these enzyme could function in formation of shikonin derivatives.

A final connection to ubiquinone metabolism uncovered in the coexpression analysis was the recovery of a quinoprotein dehydrogenase gene (Leryth_020454) in the largest N4 subnetwork that coexpressed with *LeGPPS, LePGT1*, and *LeCYP76B101* (Fig. S4d; Table S9). Leryth_020454 was also significantly overexpressed in all shikonin-abundant conditions (Table S3). Quinoprotein dehydrogenases catalyze the oxidation of glucose to gluconate with concomitant reduction of ubiquinone to ubiquinol[51]. It is conceivable that such an enzyme could function to maintain shikonins and/or pathway intermediates in reduced states to protect the cell. Alternatively, it could function to ensure a pathway intermediate(s) remains in its reduced form. A similar chemical prerequisite is necessary for transmethylation of the 1,4-naphthoquinone ring of demethylphylloquinone in the vitamin K1 pathway[52]. The phylogeny of quinoprotein dehydrogenases shows two copies of this gene in the Boraginaceae (Fig. S18). The clade that contains Leryth_020454 appears unique to shikonin producers and is absent in *M. panticulata* (Fig. S18). Collectively, the analyses provided here suggest there are multiple genes in the shikonin coexpression network that originated from duplication of ubiquinone pathway genes.

3.5 Discussion

In this study, we downregulated expression of *LeGPPS* to explore the connections linking the shikonin pathway with the pathways supplying its metabolic precursors. In doing so, we showed that the recently discovered *LeGPPS*, an FPPS with evolved GPPS activity[19], is required for shikonin production (Fig. 3-2) and that *LeGPPS* supplies GPP precursor to the shikonin pathway using MVA-pathway derived IPP/DMAPP (Fig. 3-3). We also performed a series of computational analyses to investigate the evolutionary history of metabolic innovation in the shikonin pathway. Synteny analysis of the *L. erythrorhizon* genome revealed one syntenic block in contigs containing *LeFPPS1* and *LeFPPS3* (Fig. S5) suggesting that WGD in the Boraginaceae was responsible for a duplication giving rise to these canonical FPPS paralogs (Fig. S7). However, the absence of shared synteny between *LeGPPS* and any other FPPS genes, suggests that *LeGPPS* did not arise via WGD. There is also no clear evidence of tandem duplication, and the presence of introns likely rules out retro duplication similar to what occurred with PGT evolution[20]. Instead, conservation of

intron positions between *LeGPPS* and other FPPS genes (Fig. S6) is consistent with a segmental or DNA transposition event.

Wisecaver et al.[26] previously showed that network analysis based on abundant coexpression data (i.e. hundreds of RNA-seq and/or microarray samples) is a powerful strategy for high-throughput discovery of genes involved in specialized metabolic pathways in plants. We utilized a similar computational approach here with a limited but strategically selected set of transcriptome samples (N = 14) to construct a shikonin metabolic network model. We chose to focus our analysis on *LeGPPS*[19] (Fig. 3-2), *LePGT1*[20], *LeCYP76B101*[34,35], and *LeMDR*[36] given their demonstrated roles in shikonin metabolism. Using conventional differential gene expression analysis to refine the gene coexpression matrix, we uncovered a *L. erythrorhizon* shikonin gene network model that predicts strong associations between MVA pathway genes and known shikonin biosynthesis genes, as well as links between shikonin genes and several uncharacterized enzyme-coding genes (Fig. 3-5) that present new candidates for missing shikonin biosynthesis steps (Fig. 3-1). Moreover, *L. erythrorhizon* produces high amounts of rosmarinic acid and other specialized metabolites[53]. It is therefore important to note that the gene connections uncovered in the shikonin coexpression subnetworks may extend beyond shikonin biosynthesis. However, examining expression of the rosmarinic acid biosynthesis gene CYP98A654 (Leryth_006600) and five other CYP98A6-like genes present in the *L. erythrorhizon* genome (see Table S1 and S3) showed that none were recovered in any of our shikonin subnetworks, including the largest (N4). Plotting the spearman's correlation of each CYP98A6 homolog against the eigengene[54] for each module within the N2 shikonin subnetwork shows that CYP98A6 homologs are poorly correlated with the shikonin subnetwork (Fig. S19). This is consistent with results from hierarchical clustering analysis of proteomic data showing that CYP98A6 clusters separately from shikonin biosynthesis proteins[55].

Similar to *LeGPPS* (Fig. S7) and the PGTs[20], Boraginales-specific gene family expansions were observed in the phylogenies (Figs. S8,S9,S11–S17) of the genes identified by coexpression network modeling (Table 3-1). Therefore, gene duplication appears to be the primary mechanism contributing to metabolic innovation in the Boraginales. Synteny analysis suggests that WGD was unlikely to be responsible for the expansion in the gene families for these candidates (data not shown). Furthermore, examination of the genomic regions surrounding these candidates suggests that tandem duplication did not contribute to their respective gene family expansions either, except

for the cytochrome P450 encoded by Leryth_001242 (Fig. S11). Though Leryth_001243 and Leryth_001244 were not candidates identified in the coexpression network, their transcript abundance is higher in roots than in aboveground tissues (Table S3). These cytochromes P450 are predicted to encode CYP76A2-like enzymes. Other CYP76A members have been found to catalyze oxidation cascades involved in formation of terpenoid-derived specialized metabolites[56,57], thus making Leryth_001242 and its paralogs intriguing shikonin pathway candidate genes.

The shikonin pathway relies on precursors from both isoprenoid and phenylpropanoid metabolism. Inhibitor experiments with LeGPPS-RNAi lines led us to discover an additional layer of regulatory complexity coordinating flux between the phenylpropanoid, MVA, and MEP pathways. Inhibition with the MEP pathway inhibitor fosmidomycin, for example, unexpectedly led to increased shikonin levels in both the EV-26 and *LeGPPSi*-45 lines (Fig. 3-3). This not only provides further evidence that neither IPP/DMAPP derived from the MEP pathway, nor GPP produced from MEP pathway-derived IPP/DMAPP, is exported to the cytoplasm for shikonin biosynthesis but it likely points to an increase in flux through the MVA pathway due to the impairment of the MEP pathway.

To test if impairment of the MVA pathway affects regulation of the MEP pathway, we examined expression of MVA and MEP pathway genes in EV-26 lines treated with mevinolin (Fig. S20). Treatment with mevinolin increased expression of MVA pathway genes and decreased expression of early MEP pathway genes, including one of the copies encoding the first and rate-limiting enzyme 1-deoxy-D-xylulose-5-phosphate synthase (DXS). These data implicate the existence of unknown factors coordinating flux from central carbon metabolism into the MVA and MEP pathways, adding another level of control to the complex regulation of these parallel routes in plants[15].

Comparative RNA-seq analysis of EV-26 and *LeGPPSi*-45 hairy root lines revealed that downregulation of *LeGPPS* results in transcriptional changes of genes throughout the terpenoid and phenylpropanoid metabolic networks (Fig. 3-4). The decreased expression of upstream MVA pathway genes and increased expression of genes encoding cytoplasmic enzymes utilizing IPP/DMAPP (i.e. NUDX1[17] and FPPS) may indicate that IPP/DMAPP accumulates when the LeGPPS step is limiting. The increased pool of IPP/DMAPP may then be sensed by the cell leading to transcriptional reprogramming of isoprenoid metabolism to redirect the C5 building blocks

toward other products. While levels of sterols (cytoplasmic IPP/DMAPP-derived product) and abscisic acid (plastidial IPP/DMAPP-derived product) were not significantly different, the levels of ubiquinones (mitochondrial IPP/DMAPP-derived product) were increased by 36% in *LeGPPSi*-45 lines compared to EV-26 lines (Fig. S21). The observed increase in ubiquinone levels is also noteworthy because it further suggests that its precursor pools are shared with the shikonin pathway.

The WRKYs are strong candidates for factors coordinately regulating expression of phenylpropanoid and terpenoid metabolic genes. As one of the largest classes of plant transcription factors, they are involved in regulating processes in response to a number of developmental cues and environmental stimuli. Moreover, they can act as activators or repressors and in doing so they create a regulatory network modulating signaling events from organelles and the cytoplasm to the nucleus[58]. Here, we found that 72 of the 152 genes in the N2 shikonin subnetwork, including *LePGT1*, *LeGPPS*, *LeMDR*, and *CYP76B101*, contain a canonical W-box element sequence motif (T)TGAC(C/T) (Table S11) recognized by the WRKY family of transcription factors[43]. From our analyses we identified five candidate transcription factors containing WRKY domains in the N2 shikonin subnetwork (Fig. 3-5b) including two, Leryth_027519 and Leryth_002564, which were both overexpressed in all shikonin-abundant conditions in the analyzed RNA-seq datasets (Table S3).

In addition to sharing 4-HBA and MVA-derived prenyl diphosphate metabolic precursors and having a common origin of their prenyltransferase genes, the shikonin and ubiquinone pathways rely on multiple analogous biochemical ring modifications[20]. This raises the prospect that neofunctionalization of duplicated ubiquinone biosynthesis genes facilitated evolution of the shikonin pathway. Considering this hypothesis, we explored the shikonin coexpression subnetworks for other connections to ubiquinone biosynthesis-like genes. Interestingly, COQ3-like O-methyltransferase and a quinoprotein dehydrogenase genes were found in the coexpression network that are unique to shikonin-producing species (Figs. S16 and S17). Whether these genes function in shikonin metabolism or point to another functional connection between shikonin and ubiquinone remains unclear. Moreover, we identified that in addition to encoding a canonical COQ4, *L. erythrorhizon* has a COQ4-like gene that was coexpressed with *LePGT1*, *LeGPPS*, and *LeCYP76B101* in the N2 network and was overexpressed in shikonin-abundant conditions (Fig. 3-5b; Tables S3 and S7). COQ4 is a scaffold protein found in plants, fungi, and animals, including

humans, that is required for ubiquinone biosynthesis. While its specific function is unknown, it binds proteins and lipids and thus likely assembles a metabolon for efficient ubiquinone biosynthesis[50]. Whether COQ4-like functions similarly in shikonin biosynthesis is an open question that should be explored, especially considering any insight may inform the function of the canonical COQ4 found throughout eukaryotes. Given that shikonin is abundant and non-vital, it may provide a better model for genetically studying the COQ4 gene family.

In summary, our study has i) indicated transcriptional and metabolic connections linking the shikonin pathway with it precursor pathways; ii) established a shikonin coexpression network model that includes genes encoding candidates for missing shikonin pathway steps and regulatory factors; iii) revealed instances of Boraginales-specific gene family expansion facilitated by duplication events for genes in the shikonin metabolic network; and iv) uncovered evolutionary links between shikonin metabolic network genes and ubiquinone pathway genes. The evolution of other plant specialized 1,4-naphthoquinone pathways appears to be linked to primary metabolic quinone pathways[22,23]. Thus, we expect that the evolutionary mechanistic insights gained here, combined with the demonstration that a robust coexpression network can be built from a small set of RNA-Seq experiments relying on spatial- and condition-specific metabolite correlations, can be used to guide further investigation into the convergent evolution of specialized 1,4-naphthoquinone metabolism in plants.

3.6 Materials and methods

3.6.1 Plant materials and hairy root culturing

L. erythrorhizon (accession Siebold & Zucc.) seeds were obtained from the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) seed bank (Gatersleben, Germany). Propagation of plants to bulk seeds and the generation and maintenance of hairy roots were performed as done previously[20].

3.6.2 Generation of *LeGPPSi* and empty-vector control hairy root lines

The LeGPPS-RNAi (*LeGPPSi*) construct was created by synthesizing (Genscript, Piscataway, NJ) spliced fragments of the LeGPPS coding region corresponding to nucleotides 165–727 and 165–519, the latter in antisense orientation to create a hairpin structure. A 5’-CACC sequence was

added for subcloning into pENTR™/D-TOPO (Invitrogen™, Carlsbad, CA) and subsequent transfer into the destination vector, pB2GW7[59], by recombination using LR Clonase Enzyme Mix™ (Invitrogen). The final construct, pB2GW7-GPPSi, was transformed into Agrobacterium rhizogenes strain ATCC 15834 competent cells by freeze–thaw transformation[60] and plated on Nutrient Broth (NB) agar containing 50 μg/mL spectinomycin for selection.

L. erythrorhizon hairy root GPPSi lines were generated by applying prepared cultures of A. rhizogenes containing the pB2GW7-GPPSi construct to wounded stems of *L. erythrorhizon* plants in tissue culture as previously described[20]. Emergent roots from plants 2–4 weeks after infection were excised and transferred to Gamborg B5 media plates containing 3% sucrose and 200 μg/mL cefotaxime to eliminate A. rhizogenes. After 2 weeks, hairy roots were transferred to Gamborg B5 media containing 3% sucrose and 10 mg/L Basta for selection for 2 weeks. Hairy root lines transformed by A. rhizogenes carrying an empty pB2GW7 vector were generated in parallel as controls.

3.6.3 RNA extraction and qRT-PCR analysis

Total RNA was extracted from ~100 mg of flash-frozen hairy root tissue and qRT-PCR reactions were performed using a QuantStudio™ 6 (ThermoFisher) as previously described[20]. Expression of LeGPPS and LeGPPS2 was measured with comparative quantification using the $2^{-\Delta\Delta CT}$ method[61]. Primers were designed using Primer-BLAST on NCBI[62] (Table S15). Expression was normalized to *L. erythrorhizon* glyceraldehyde 3-phosphate dehydrogenase (LeGAPDH)[63].

3.6.4 Metabolite extraction and quantification

Extraction and analysis of ABA by liquid chromatography coupled with tandem mass spectrometry (LC–MS/MS) was performed as previously described[64]. Extraction of total shikonins from growth media of hairy root cultures and quantification on an Agilent 1260 Infinity high performance liquid chromatography with diode array detection (HPLC-DAD) system (Agilent Technologies) was done as previously described[20]. Sterols were extracted from 100–200 mg of ground flash-frozen hairy root tissue, derivatized with BSTFA, and analyzed on an Agilent 7890B gas chromatograph (GC) coupled with a 5977A mass spectrometer (MS) equipped with a DB-5MS column (30 m × 0.25 mm × 0.25 μm film; Agilent Technologies) and employing Chemstation

software as previously described[16]. Ubiquinones were extracted from 100–200 mg of ground flash-frozen fresh tissue in 3 mL of 95% ethanol spiked with 4 nmol ubiquinone-4 internal standard and incubated overnight with shaking at 4°C. The next day, samples were centrifuged at 500 x g to pellet debris. Then, 1.5 mL of water was added to supernatant and partitioned twice with 4.5 mL hexane. The hexane layers were combined and concentrated under nitrogen gas at 37°C. Nearly dry samples were resuspended in 1 ml 90:10 methanol:dichloromethane and filtered through 0.2 μm PTFE syringe filters. Care was taken throughout the extraction process to protect samples from light. Samples were analyzed by HPLC-DAD on an Agilent Zorbax SB-C18 column (5 μm, 250 x 4.6 mm) thermostatted at 25°C and eluted in isocratic mode with 30% 60:40 isopropanol:hexanes and 70% 80:20 methanol:hexanes[8]. Ubiquinones were detected spectrophotometrically at 255 nm and had retention times of 4.8 min for ubiquinone-4, 11.4 min for ubiquinone-9, and 14.3 min for ubiquinone-10. Instrument operation and data analysis steps were performed through the Agilent ChemStation software. Quantification of ubiquinones was done by DAD using signals obtained in the linear range of calibration standards (0.0313, 0.0625, 0.125, 0.250 and 0.500 nmol). The data were corrected for recovery according to the ubiquinone-4 internal standard, and final quantifications were made using linear regression. Differences in total shikonin and ubiquinone-9 and ubiquinone-10 content produced by empty-vector control and *LeGPPSi* lines (n = 4 biological replicates) were analyzed using one-way ANOVA and means were compared with Tukey's HSD post-hoc test at a 95% significance level.

3.6.5 RNA-sequencing analysis of *LeGPPSi* and empty-vector control lines

For RNA-seq analysis of *L. erythrorhizon* EV-26 and *LeGPPSi*-45, three independent hairy root cultures of each line were started in liquid Gamborg B5 media containing 3% sucrose and grown at 28°C in 100 μE m−2 s−1 light. After two weeks, the hairy roots were transferred to M9 media containing 3% sucrose and darkness for six days. The hairy roots were then frozen in liquid nitrogen, ground by mortar and pestle, and RNA was extracted from ~100 mg of tissue as described above. For RNA-seq analysis of *L. erythrorhizon* EV-26 lines, three independent hairy root cultures were grown as just described. Mock (control) and 100 μM mevinolin treatments were administered immediately upon transfer of 14-d-old hairy roots to M9 and darkness. Total RNA was extracted at 6 d after transfer of 14-d-old hairy roots to M9 and darkness.

Library construction (NEBNext Ultra RNA Library Prep Kit, New England Biolabs Inc.) from 1 μg RNA, Illumina sequencing, and analyses of DEGs were performed by Novogene Corporation Inc. (Sacramento, CA). Paired-end clean reads were mapped to the *L. erythrorhizon* reference genome[20] using HISAT2 software [65] For each sequenced library, read counts were adjusted by TMM [66] and DEG analysis was performed using DESeq2[67] with p-value adjusted using an FDR calculated with Benjamini–Hochberg (BH) methods[68]. Genes were considered significantly differentially expressed if they had a BH-adjusted p-value of 0.005 and a log2 fold change of 1. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses of DEGs was implemented by the clusterProfiler R package[69] and KEGG pathways with BH-adjusted p-value <0.05 were considered significantly enriched. The raw data were submitted to the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra/) and are available at the NCBI Sequence Read Archive (PRJNA811172).

3.6.6 Analysis of transcriptomes used to build shikonin gene coexpression networks

Illumina RNA-seq reads of *L. erythrorhizon* root periderm, root vascular, and hairy root cultures were generated as described[20] and are available at the NCBI Sequence Read Archive (PRJNA596998). Additional Illumina RNA-seq reads of *L. erythrorhizon* whole roots and above ground tissue (pooled leaves and stems) were downloaded from the NCBI SRA database, experiments SRR3957230 and SRR3957231 respectively. *L. erythrorhizon* gene functional annotations were downloaded from Auber et al.[20].

L. erythrorhizon RNA-seq raw reads were error corrected using the Tadpole (default parameters; software last modified June 27, 2017) program from the BBMap software package (https://sourceforge.net/projects/bbmap/). Gene expression was quantified with Kallisto[70] by aligning the error corrected reads to a collection of the longest transcript per gene of the *L. erythrorhizon* genome. The *LePS-2* gene was previously implicated in shikonin biosynthesis[28] but was not present in v1.0 of the *L. erythrorhizon* gene set[20]. Therefore, we identified a putative coding sequence for *LePS-2* in the *L. erythrorhizon* genome assembly manually and added its sequence to the total gene set prior to gene expression quantification.

Analyses of differential gene expression was performed using the edgeR package[71]. Gene expression counts were normalized using the TMM (trimmed mean of M values) method[66]. Exact tests were conducted using a trended dispersion value and a double tail reject region. FDRs were

calculated using the BH procedure[68]. Genes that did not have a significant differential expression status in at least one comparison (FDR < 0.1) were excluded from downstream coexpression analyses.

3.6.7 Coexpression network analysis

Raw gene expression counts were normalized using the transcripts per million method and transformed using the variance-stabilizing transformation method in DESeq2[67], and global gene coexpression networks were constructed as previously described[26]. Briefly, a Pearson's correlation coefficient (PCC) was calculated between gene pairs and converted into a mutual rank (MR) using scripts available for download on GitHub (https://github.rcac.purdue.edu/jwisecav/coexp-pipe). MR scores were transformed to network edge weights using the exponential decay function e−(MR−1/x); four different networks were constructed with x set to 5, 10, 25, and 50, respectively. Edges with a weight < 0.01 were trimmed from the global network. Modules of coexpressed genes were detected using ClusterOne v1.0 using default parameters[32]. Module eigengenes were calculated using WGCNA[54,72]. Overlapping modules within each coexpression network were combined by collapsing all modules containing the known Shikonin pathway genes *LeGPPS*[19], *LePGT1*[13], *LeCYP76B101*[35] and *LeMDR*[73] into a subnetwork. Modules were visualized in Cytoscape using the spring embedded layout. Tests for functional enrichment of Gene Ontology (GO) terms in the different shikonin subnetworks (Table S10) were performed using hypergeometric tests using the SciPy library hypergeom, and p-values were adjusted for multiple comparisons using the StatsModels library multitest using the BH procedure[68]. GO terms and other gene functional annotations were taken from Auber et al. [20].

3.6.8 Promoter analysis

Nucleic acid sequence motifs enriched in promoter regions of genes in the N2 shikonin subnetwork (N = 152) were identified with Motif Indexer[42] using a 1000 base pair window upstream of all transcriptional start sites using the same upstream region of all *L. erythrorhizon* genes as background. Identified motifs were consolidated and ranked using the KeyMotifs.pl perl script provided by Motif Indexer. To calculate a false discovery rate, 1000 random sets of 152 genes

were run through Motif Indexer determine a p-value threshold. No motif identified from a random gene set had a p-value less than 1x10−9.

3.6.9 Phylogenetic analysis

To construct gene phylogenies, the gene family containing the best A. thaliana BLAST hit to the query gene was downloaded from the PLAZA 4.0 Dicots comparative genomics database[44]. Homology between the predicted proteomes of *L. erythrorhizon* and 18 additional Borginales was determined with OrthoFinder v2.1.2 using the following parameters: -S diamond -M msa -T fasttree[74]. OrthoFinder orthogroups containing the query gene were combined with the Plaza 4.0 gene family to obtain the final sequence sets. Sequences were aligned with MAFFT[75] using the E-INS-I strategy and following parameters: —maxiterate 1000 —bl 45 —op 1.0 —retree 3. The maximum likelihood phylogeny was constructed using IQ-TREE[76] using the built in ModelFinder to determine the best-fit substitution model[77] and performing SH-aLRT and the ultrafast bootstrapping analyses with 1000 replicates each. For the cytochrome P450 and acetyltransferase gene candidates, because the PLAZA 4.0 gene families were so large, a quick guide tree of the entire gene family was built using FastTree[78]. Regions of the guide tree that contained candidate genes of interest were identified; sequences within these regions were realigned using MAFFT, and phylogenies were built using IQ-TREE as described above.

3.6.10 Synteny analysis

Regions of shared synteny within the genome of *L. erythrorhizon* were detected using SynMap2 on the online Comparative Genomics Platform (CoGe) using default settings with the exception that the merge syntenic blocks algorithm was set to Quota Align Merge, syntenic depth algorithm was set to Quota Align, and the CodeML option was activated to calculate substitution rates between syntenic CDS pairs. For syntenic blocks containing genes of interest and their homologs, the encompassing contigs were aligned using promer of the MUMmer4 alignment system[79].

3.7 Supplemental data

Supplemental tables, figures, and datasets are available at https://doi.org/10.1093/hr/uhab087

3.8 References

1 Papageorgiou VP, Assimopoulou AN, Couladouros EAet al. . The chemistry and biology of alkannin, shikonin, and related naphthazarin natural products. Chemistry (Weinheim an der Bergstrasse, Germany). 1999;38:270–301.

2 Skoneczny D, Weston P, Zhu Xet al. . Metabolic profiling and identification of Shikonins in root periderm of two invasive Echium spp. weeds in Australia. Molecules. 2017;22:330.

3 Zhu X, Skoneczny D, Weidenhamer JDet al. . Identification and localization of bioactive naphthoquinones in the roots and rhizosphere of Paterson's curse (Echium plantagineum), a noxious invader. J Exp Bot. 2016;67:3777–88.

4 Widhalm JR, Rhodes D. Biosynthesis and molecular actions of specialized 1,4-naphthoquinone natural products produced by horticultural plants. Hortic Res. 2016;3:16046.

5 Wang F, Yao X, Zhang Yet al. . Synthesis, biological function and evaluation of Shikonin in cancer therapy. Fitoterapia. 2019;134:329–39.

6 Schmid HV, Zenk MH. P-hydroxybenzoic acid and mevalonic acid as precursors of the plant naphthoquinone alkannin. Tetrahedron Lett. 1971;12:4151–5.

7 Loscher R, Heide L. Biosynthesis of p-Hydroxybenzoate from p-Coumarate and p-Coumaroyl-coenzyme a in cell-free extracts of Lithospermum erythrorhizon cell cultures. Plant Physiol. 1994;106:271–9.

8 Block A, Widhalm JR, Fatihi Aet al. . The origin and biosynthesis of the Benzenoid moiety of ubiquinone (coenzyme Q) in Arabidopsis. Plant Cell. 2014;26:1938–48.

9 Soubeyrand E, Johnson TS, Latimer Set al. . The Peroxidative cleavage of Kaempferol contributes to the biosynthesis of the Benzenoid moiety of ubiquinone in plants. Plant Cell. 2018;30:2910–21.

10 Yazaki K, Kataoka M, Honda Get al. . cDNA cloning and gene expression of phenylalanine ammonia-Lyase in Lithospermum erythrorhizon. Biosci Biotechnol Biochem. 1997;61:1995–2003.

11 Yamamura Y, Ogihara Y, Mizukami H. Cinnamic acid 4-hydroxylase from Lithospermum erythrorhizon: cDNA cloning and gene expression. Plant Cell Rep. 2001;20:655–62.

12 Singh RS, Gara RK, Bhardwaj PKet al. . Expression of 3-hydroxy-3-methylglutaryl-CoA reductase, p-hydroxybenzoate-m-geranyltransferase and genes of phenylpropanoid pathway exhibits positive correlation with shikonins content in arnebia

13 Yazaki K, Kunihisa M, Fujisaki Tet al. . Geranyl Diphosphate:4-Hydroxybenzoate Geranyltransferase fromLithospermum erythrorhizon: cloning and characterization of a key enzyme in shikonin biosynthesis. J Biol Chem. 2002;277:6240–6.

14 Yazaki K, Fukui H, Tabata M. Isolation of the intermediates and related metabolites of shikonin biosynthesis from Lithospermum erythrorhizon cell cultures. Chem Pharm Bull (Tokyo). 1986;34:2290–3.

15 Vranová E, Coman D, Gruissem W. Network analysis of the MVA and MEP pathways for isoprenoid synthesis. Annu Rev Plant Biol. 2013;64:665–700.

16 Henry LK, Gutensohn M, Thomas STet al. . Orthologs of the archaeal isopentenyl phosphate kinase regulate terpenoid production in plants. Proc Natl Acad Sci. 2015;112:10050–5.

17 Henry LK, Thomas ST, Widhalm JRet al. . Contribution of isopentenyl phosphate to plant terpenoid metabolism. Nat Plants. 2018;4:721–9.

18 Gaisser S, Heide L. Inhibition and regulation of shikonin biosynthesis in suspension cultures of Lithospermum. Phytochemistry. 1996;41:1065–72.

19 Ueoka H, Sasaki K, Miyawaki Tet al. . A cytosol-localized Geranyl Diphosphate synthase from Lithospermum erythrorhizon and its molecular evolution. Plant Physiol. 2020;182:1933–45.

20 Auber RP, Suttiyut T, McCoy RMet al. . Hybrid de novo genome assembly of red gromwell (Lithospermum erythrorhizon) reveals evolutionary insight into shikonin biosynthesis. Hortic Res. 2020;7:82.

21 Tang CY, Li S, Wang YTet al. . Comparative genome/transcriptome analysis probes Boraginales' phylogenetic position, WGDs in Boraginales, and key enzyme genes in the alkannin/shikonin core pathway. Mol Ecol Resour. 2020;20:228–41.

22 Meyer GW, Bahamon Naranjo MA, Widhalm JR. Convergent evolution of plant specialized 1,4-naphthoquinones: metabolism, trafficking, and resistance to their allelopathic effects. J Exp Bot. 2021;72:167–76.

23 McCoy RM, Utturkar SM, Crook JWet al. . The origin and biosynthesis of the naphthalenoid moiety of juglone in black walnut. Hortic Res. 2018;5:67.

24 Tholl D. Biosynthesis and Biological Functions of Terpenoids in Plants. Adv Biochem Eng Biotechnol. 2015;148:63–106.

25 Usadel B, Obayashi T, Mutwil Met al. . Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ. 2009;32:1633–51.

26 Wisecaver JH, Borowsky AT, Tzin Vet al. . A global Coexpression network approach for connecting genes to specialized metabolic pathways in plants. Plant Cell. 2017;29:944–59.

27 Song W, Zhuang Y, Liu T. CYP82AR subfamily proteins catalyze C-1′ Hydroxylations of Deoxyshikonin in the biosynthesis of Shikonin and Alkannin. Org Lett. 2021;23:2455–9.

28 Yamamura Y, Sahin FP, Nagatsu Aet al. . Molecular cloning and characterization of a cDNA encoding a novel apoplastic protein preferentially expressed in a shikonin-producing callus strain of Lithospermum erythrorhizon. Plant Cell Physiol. 2003;44:437–46.

29 Zhao H, Baloch SK, Kong LRet al. . Molecular cloning, characterization, and expression analysis of LeMYB1 from Lithospermum erythrorhizon. Biol Plant. 2014;58:436–44.

30 Obayashi T, Kinoshita K. Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. DNA Res. 2009;16:249–60.

31 Liesecke F, Daudu D, De Bernonville RDet al. . Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. Sci Rep. 2018;8: Article number 10885. 10.1038/s41598-018-29077-3.

32 Wu H, Gao L, Dong J, Yang X. Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein interaction networks. PLoS One. 2014;9:471–2.

33 Dongen S, Abreu-Goodger C. Using MCL to Extract Clusters from Networks. In: Bacterial Molecular Networks. Springer: New York, NY, 2012,281–95.

34 Song W, Zhuang Y, Liu T. Potential role of two cytochrome P450s obtained from Lithospermum erythrorhizon in catalyzing the oxidation of geranylhydroquinone during Shikonin biosynthesis. Phytochemistry. 2020;175:112375.

35 Wang S, Wang R, Liu Tet al. . CYP76B74 catalyzes the 3″-hydroxylation of geranylhydroquinone in shikonin biosynthesis. Plant Physiol. 2019;179:402–14.

36 Zhu Y, Chu S-J, Luo Y-Let al. . Involvement of LeMRP, an ATP-binding cassette transporter, in shikonin transport and biosynthesis in Lithospermum erythrorhizon. Plant Biol. 2017;17:1–9.

37 Rai A, Nakaya T, Shimizu Yet al. . De novo Transcriptome assembly and characterization of Lithospermum officinale to discover putative genes involved in specialized metabolites. Planta Med. 2018;84:920–34.

38 Wu FY, Tang CY, Guo YMet al. . Transcriptome analysis explores genes related to shikonin biosynthesis in Lithospermeae plants and provides insights into Boraginales' evolutionary history. Sci Rep. 2017;7: Article number 4477. 10.1038/s41598-017-04750-1.

39 Fang R, Wu F, Zou Aet al. . Transgenic analysis reveals LeACS-1 as a positive regulator of ethylene-induced shikonin biosynthesis in Lithospermum erythrorhizon hairy roots. Plant Mol Biol. 2016;90:345–58.

40 Yazaki K, Matsuoka H, Shimomura Ket al. . A novel dark-inducible protein, LeDI-2, and its involvement in root-specific secondary metabolism in Lithospermum erythrorhizon. Plant Physiol. 2001;125:1831–41.

41 Tatsumi K, Yano M, Kaminade Ket al. . Characterization of Shikonin derivative secretion in Lithospermum erythrorhizon hairy roots as a model of lipid-soluble metabolite secretion from plants. Front Plant Sci. 2016;7:1–11.

42 Ma S, Bachan S, Porto Met al. . Discovery of stress responsive DNA regulatory motifs in arabidopsis. PLoS One. 2012;7:e43198. 10.1371/journal.pone.0043198.

43 Rushton PJ, Torres JT, Parniske Met al. . Interaction of elicitor-induced DNA-binding proteins with elicitor response elements in the promoters of parsley PR1 genes. EMBO J. 1996;15:5690–700.

44 Van Bel M, Diels T, Vancaester Eet al. . PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. 2018;46:D1190–6.

45 Leebens-Mack JH, Barker MS, Carpenter EJet al. . One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019;574:679–85.

46 Moore BM, Wang P, Fan Pet al. . Robust predictions of specialized metabolism genes through machine learning. Proc Natl Acad Sci U S A. 2019;116:2344–53.

47 Maeda H, Dudareva N. The shikimate pathway and aromatic amino acid biosynthesis in plants. Annu Rev Plant Biol. 2012;63:73–105.

48 Papageorgiou V, Assimopoulou A, Samanidou Vet al. . Recent advances in chemistry, biology and biotechnology of Alkannins and Shikonins. Curr Org Chem. 2006;10:2123–42.

49 Oshikiri H, Watanabe B, Yamamoto Het al. . Two BAHD acyltransferases catalyze the last step in the shikonin/alkannin biosynthetic pathway. Plant Physiol. 2020;184:753–61.

50 Stefely JA, Pagliarini DJ. Biochemistry of mitochondrial coenzyme Q biosynthesis. Trends Biochem Sci. 2017;42:824–43.

51 Oubrie A, Rozeboom HJ, Kalk KHet al. . Structure and mechanism of soluble quinoprotein glucose dehydrogenase. EMBO J. 1999;18:5187–94.

52 Fatihi A, Latimer S, Schmollinger Set al. . A dedicated type II NADPH dehydrogenase performs the penultimate step in the biosynthesis of vitamin K1 in Synechocystis and Arabidopsis. Plant Cell. 2015;27:1730–41.

53 Dresler S, Szymczak G, Wójcik M. Comparison of some secondary metabolite content in the seventeen species of the Boraginaceae family. Pharm Biol. 2017;55:691–5.

54 Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. BMC Syst Biol. 2007;1:54.

55 Takanashi K, Nakagawa Y, Aburaya Set al. . Comparative proteomic analysis of Lithospermum erythrorhizon reveals regulation of a variety of metabolic enzymes leading to comprehensive understanding of the Shikonin biosynthetic pathway. Plant Cell Physiol. 2019;60:19–28.

56 Guo J, Zhou YJ, Hillwig MLet al. . CYP76AH1 catalyzes turnover of miltiradiene in tanshinones biosynthesis and enables heterologous production of ferruginol in yeasts. Proc Natl Acad Sci U S A. 2013;110:12108–13.

57 Miettinen K, Dong L, Navrot Net al. . The seco-iridoid pathway from Catharanthus roseus. Nat Commun. 2014;5:3606. 10.1038/ncomms4606.

58 Bakshi M, Oelmüller R. Wrky transcription factors jack of many trades in plants. Plant Signal Behav. 2014;9:1–18.

59 Karimi M, Inzé D, Depicker A. GATEWAY™ vectors for agrobacterium-mediated plant transformation. Trends Plant Sci. 2002;7:193–5.

60 Cui W, Liu W, Wu G. A simple method for the transformation of agrobacterium tumefaciens by foreign DNA. Chin J Biotechnol. 1995;11:267–74.

61 Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. Methods. 2001;25:402–8.

62 Ye J, Coulouris G, Zaretskaya Iet al. . Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012;13:134.

63 Zhao H, Chang QS, Zhang DXet al. . Overexpression of LeMYB1 enhances shikonin formation by up-regulating key shikonin biosynthesis-related genes in Lithospermum erythrorhizon. Biol Plant. 2015;59:429–35.

64 McAdam SAM, Brodribb TJ. Mesophyll cells are the main site of abscisic acid biosynthesis in water-stressed leaves. Plant Physiol. 2018;177:911–7.

65 Kim D, Paggi JM, Park Cet al. . Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.

66 Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25. 10.1186/gb-2010-11-3-r25.

67 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15:1–21.

68 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57:289–300.

69 Yu G, Wang L-G, Han Yet al. . clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.

70 Bray NL, Pimentel H, Melsted Pet al. . Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.

71 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

72 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. 10.1186/1471-2105-9-559.

73 Zhu Y, Lu GH, Bian ZWet al. . Involvement of LeMDR, an ATP-binding cassette protein gene, in shikonin transport and biosynthesis in Lithospermum erythrorhizon. BMC Plant Biol. 2017;17:1–10.

74 Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:1–14.

75 Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.

76 Nguyen LT, Schmidt HA, Von Haeseler Aet al. . IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.

77 Kalyaanamoorthy S, Minh BQ, Wong TKFet al. . ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9.

78 Price MN, Dehal PS, Arkin AP. FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490.

79 Marçais G, Delcher AL, Phillippy AMet al. . MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol. 2018;14:e1005944. 10.1371/journal.pcbi.1005944.

CHAPTER 4. HYBRIDIZATION, PLOIDY, AND GENOME SIZE VARIATION IN THE TOXIC ALGA *PRYMNESIUM PARVUM*

4.1 Declaration of collaborative work

Robert Auber performed the Oxford Nanopore sequencing, generated the genome assemblies, and performed all genome-level analyses. Olga Yurchenko assisted with DNA isolation for Illumina sequencing. Dr. Jennifer Wisecaver assisted with genome-level analyses. Robert Auber and Dr. Jennifer Wisecaver prepared figures.

4.2 Introduction

Prymnesium parvum, also known as the golden alga, is notorious for forming toxic algal blooms around the world. This eukaryotic microalga is globally distributed and capable of acclimating to both marine and brackish water systems[1]. In addition to being an obligate autotroph, *P. parvum* can also display mixotrophic behaviors[2–4]. Micropredation in *P. parvum* is thought to be facilitated by the production of toxic metabolites known as prymnesins[5–7] which also function as allelochemicals against grazers and competitors[8,9]. Additionally, prymnesins are potent ichthyotoxins that induce massive fish kills during bloom events, which heavily disrupt ecosystems[10,11].

A remarkable amount of phenotypic variation has been observed among *P. parvum* strains. Abiotic factors including temperature[1,12,13], salinity[1,12,13], pH[14], and light[1,15] have all been demonstrated to have distinct effects on the growth of different strains. Further, rates of mixotrophy and prymnesin production in response to environmental factors also vary between strains[1,12,15,16]. Even strains isolated from the same bloom event have displayed significant variations in growth and mixotrophy[9]. Beyond these physical measurements, strains of *P. parvum* also vary in the types of prymnesin they produce. Three major classes of the prymnesins, A-type, B-type, and C-type, have been shown to be variably produced between *P. parvum* strains[6,7]. Together, our knowledge of the wide phenotypic variation observed in *P. parvum* is extensive, however, our understanding of the underlying genotypic variation is comparatively limited.

Prior to molecular phylogenetics, scale morphology was the primary diagnostic approach used to classify species within the *Prymnesium* species complex[17]. However, molecular data later suggested the species complex was conspecific and morphology is not a proxy of genetic relationship[18,19]. Phylogenies built from ITS sequences revealed that the three major *P. parvum* clades each produce a unique prymnesin, indicating that these compounds are of monophyletic origin[7,20]. Dramatic genome size variation has also been reported in *P. parvum* and has been interpreted as differences in ploidy state[21]. Several species of Prymnesiophyceae are haplo-diploid and are capable of mitotic cell division in both haploid (1C) and diploid (2C) lifecycle stages[22]. The two stages may even be morphologically indistinguishable from one another, as has been proposed for *Chrysochromulina polylepsis*, a species closely related to *P. parvum*[23]. However, the life cycle of *P. parvum* has yet to be described. Our ability to further pursue questions pertaining to the phenotypic, metabolic, and genetic variation present in *P. parvum* is limited by a lack of genomic resources for the species.

Here, we report near chromosome-scale assemblies of two *P. parvum* strains and complementary phylogenomic analyses with thirteen additional strains to investigate the occurrence and scale of genome size variation in the species. In doing so, we provide evidence of hybridization and suggest the three phylogenetically and genomically distinct chemotypes of *P. parvum* actually represent three separate cryptic species.

4.3 Results

We obtained Hi-C scaffolded, highly contiguous genome assemblies of two *P. parvum* strains from Texas: UTEX2797 and 12B1. UTEX2797 was selected due to its status as a common reference strain used in numerous studies of *P. parvum e.g.*[7,13,14,20,24–26]. However, preliminary analysis of k-mer frequencies revealed that UTEX2797 displays high sequence level heterozygosity (Fig. 4-1A,B), which can complicate genome assembly. Therefore, we also selected strain 12B1[9] due to its small estimated haploid genome size and lower measure of heterozygosity (Fig. 4-1C,D). The nuclear DNA content of both strains was estimated using flow cytometry combined with propidium iodide staining. Preliminary data indicates 12B1 had 0.11 pg of DNA which corresponds to approximately 112 Mbp and UTEX2797 had 0.23 pg, corresponding to approximately 222 Mbp (Jennifer Wisecaver, personal communication).

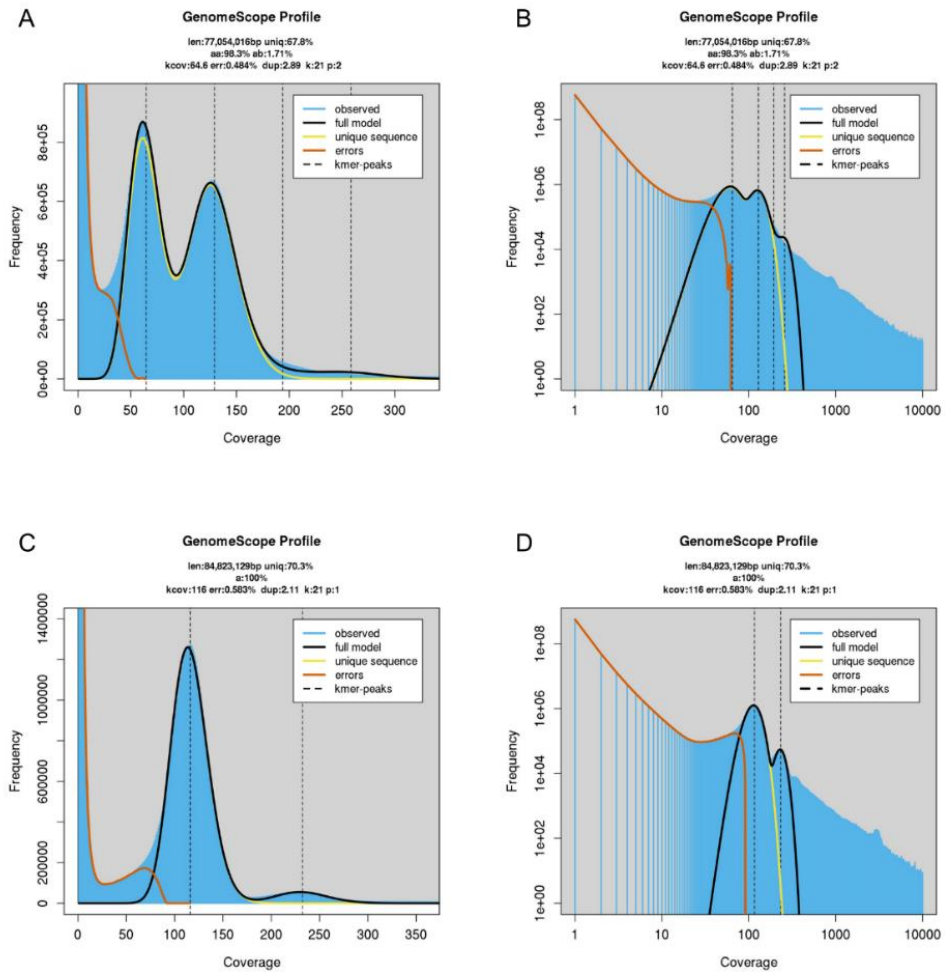

Figure 4-1. K-mer frequency plots showing estimated heterozygosity in *P. parvum* strains 12B1 and UTEX2797. A) Linear k-mer profile of UTEX2797 Illumina gDNA reads. Blue bars indicate observed k-mer frequencies and black lines indicate expected distribution for GenomeScope model for diploid genomes. B) Log transformed k-mer profile of UTEX2797. C) Linear k-mer profile of 12B1 with expected haploid GenomeScope model. D) Log transformed k-mer profile of 12B1.

The resulting nuclear genome assembly of 12B1 is 93.6 Mbp and consists of 34 scaffolds with a scaffold N50 of 3.2 Mbp (Table 4-1). The UTEX2797 assembly is 197.6 Mbp, over twice the length of the 12B1 assembly, and consists of 66 scaffolds with a scaffold N50 of 3.4 Mbp (Table 4-1). Compared to other sequenced haptophytes, the UTEX2797 and 12B1 assemblies are the second and third most contiguous (Fig. 4-2A); only the genome of *Isochrysis galbrata* is more complete with an N50 = 6.99 Mbp[27]. We predicted and annotated 23,820 and 45,535 genes in the genomes of 12B1 and UTEX2797, respectively (Table 4-1). The completeness of the annotation sets was assessed with BUSCO[28]; 216 of 255 conserved eukaryotic genes (84.7%) were recovered

as complete within both predicted proteomes (Table 4-1). This level of BUSCO recovery is the greatest of any currently available haptophyte assembly (Fig. 4-2A, Table 4-1). Whereas only 5.1% of BUSCO genes were duplicated in the 12B1 assembly, 75.5% were duplicated in UTEX2797.

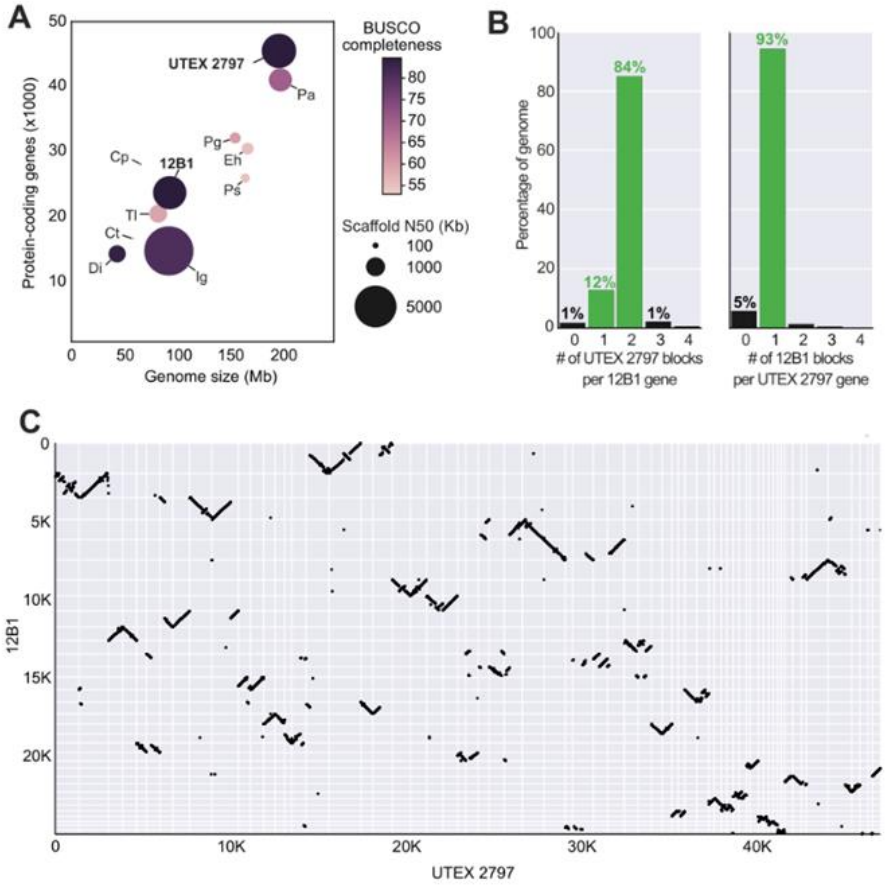

Figure 4-2 Summary of haptophyte genome assemblies and UTEX2797-12B1 synteny. A) Scatter plot indicating contiguity and completeness of sequenced haptophyte genome assemblies and gene annotations. B) Depth ratios of syntenic gene blocks in UTEX2797 to 12B1 (left) and 12B1 to UTEX2797 (right). C) Dot plot of synteny between UTEX2797 and 12B1 assemblies.

Table 4-3 Summary statistics of sequenced haptophyte genome assemblies and gene annotations.

Genome assembly	Genome Assembly Size (Mbp)	Protein Coding Genes	BUSCO Score (%)[a]	Scaffold N50 (Mbp)	Reference
Isochrysis galbrata	92.7	14,900	80	7	Chen *et al*. 2020 https://doi.org/10.22541/au.160881384.48495723/v1
Prymnesium parvum UTEX2797	197.6	45,535	85	3.4	This work
Prymnesium parvum 12B1	93.6	23,820	85	3.2	This work
Phaeocystis antarctica	198.9	41,088	70	1.6	https://phycocosm.jgi.doe.gov/Phaant1/Phaant1.home.html
Tisoschrysis lutea	82.6	20,582	59	0.93	Carrier *et al*. 2018 Algal Research; Berthelier, *et al*. 2018 SEANOE
Diacronema lutheri	43.5	14,446	84	0.85	Hulatt *et al*. 2021 Genome Biology and Evolution
Emiliania huxleyi	167.7	30,569	56	0.4	Read *et al*. 2013 Nature
Phaeocystis globosa	155.8	32,196	60	0.36	https://phycocosm.jgi.doe.gov/Phaglo1/Phaglo1.home.html
Pavlovales sp.	165.4	26,034	55	0.25	https://phycocosm.jgi.doe.gov/Pavlov2436_1/Pavlov2436_1.home.html
Chrysochromulina tobin	59.1	16,770	53	0.024	Hovde *et al*. 2015 PLoS Genetics
Chrysochromulina parva	65.8	28,185	71	0.016	Hovde *et al*. 2019 Algal Research

[a] BUSCO score represents the total percentage of conserved eukaryote genes identified in the predicted proteome of each species.

Synteny and collinearity analyses further highlight the duplicated state of the UTEX2797 assembly. A strong 2:1 synteny pattern is observed between the UTEX2797 and 12B1 assemblies, with 93% of UTEX2797 genes syntenic to one block in 12B1 while 84% of 12B1 genes are syntenic to two blocks in the UTEX2797 genome (Fig. 4-2B). Similarly, most 12B1 scaffolds are collinear with two syntenic regions in the UTEX2797 assembly, but an abundance of structural variants (*e.g.*, inversions, indels, translocations) were unique to each strain and even distinguished within the paired syntenic blocks of UTEX2797 (Fig. 4-2C). A cryptic haplo-diplonic life cycle has been proposed for *P. parvum*[21], raising the hypothesis that 12B1 and UTEX2797 represent haploid and diploid stages, respectively. However, neither syngamy nor meiosis has been observed in this species. Given the high levels of heterozygosity and structural differences between syntenic UTEX2797 scaffolds, an alternative hypothesis is that UTEX2797 has undergone allopolyploidization (whole genome doubling through hybridization).

In search of phylogenetic signals that would arise from distinct subgenomes, we performed an OrthoFinder analysis to construct gene families from the predicted proteomes of the UTEX2797 and 12B1 long read Hi-C guided assemblies as well as short read assemblies of eleven additional *P. parvum* strains (Table 4-2). Two *Chrysochromulina* species were used as outgroup taxa[29,30]. Two *P. parvum* strains were excluded from the analysis to reduce phylogenetic discordance that could arise by including genomes that are highly heterozygous (12A1; Figure 4-5) or non-clonal (K0081[1]). To construct a species tree, we employed a multi-labeled tree reconciliation approach capable of modeling polyploidy events[31]. Using 8,903 maximum likelihood nucleotide trees built from gene families containing 12B1:UTEX2797 syntelogs present in a 1:2 relationship (See Methods), we resolved a species tree in which UTEX2797 is present twice (Figure 4-3); grouping sister to 12B1 (inside subclade X) as well as sister to CCMP2941 (inside subclade Y). These phylogenomic results indicate that UTEX2797 possesses two distinct subgenomes resulting from a hybridization event that occurred between one subclade X parent and one subclade Y parent.

Table 4-4 Genome assembly statistics of all *P. parvum* strains used in this study

Strain	Clade	Assembly Type	Total Length	No. Scaffolds	No. Contigs	Scaffold N50 (kbp)	Scaffold L50	Contig N50 (kbp)	Contig L50
12A1	A-type	Abyss	88,266,609	42,808	44,207	3.82	6,575	3.636	6,864
12B1	A-type	Abyss	77,503,769	14,364	16,637	9.719	2,417	8.181	2,823
12B1	A-type	Phase	93,569,082	34	227	3,203.049	11	852.115	30
12B1	A-type	Masurca	94,718,270	274	275	1,058.297	23	1,058.297	23
CCMP2941	A-type	Abyss	78,590,816	22,671	24,881	5.965	3,859	5.378	4,244
CCMP3037	A-type	Abyss	77,045,749	17,653	19,643	7.339	3,113	6.578	3,433
K0081	B-type	Abyss	82,796,649	37,232	39,077	3.414	7,063	3.251	7,405
K0252	C-type	Abyss	86,921,338	27,208	29,111	5.553	4,609	5.137	4,957
K0374	B-type	Abyss	88,174,406	30,365	33,059	4.816	5,474	4.464	5,852
KAC39	B-type	Abyss	87,657,693	30,825	33,083	4.696	5,555	4.391	5,915
RCC1433	C-type	Abyss	84,595,500	17,325	19,275	8.872	2,793	7.902	3,114
RCC1436	C-type	Abyss	92,933,357	30,186	32,494	5.244	5,163	4.858	5,567
RCC191	C-type	Abyss	83,763,833	17,791	20,007	8.472	2,930	7.553	3,261
RCC3426	B-type	Abyss	85,734,123	32,591	34,736	4.25	5,911	4.002	6,261
RCC3703	A-type	Abyss	79,820,517	21,306	23,284	6.249	3,711	5.743	4,043
UTEX2797	A-type	Phase	197,592,770	66	585	3,431.116	21	548.273	106
UTEX2797	A-type	Canu	238,271,893	1,081	1,081	656.552	97	656.552	97
UTEX995	C-type	Abyss	87,218,798	33,070	35,047	4.297	5,869	4.104	6,101

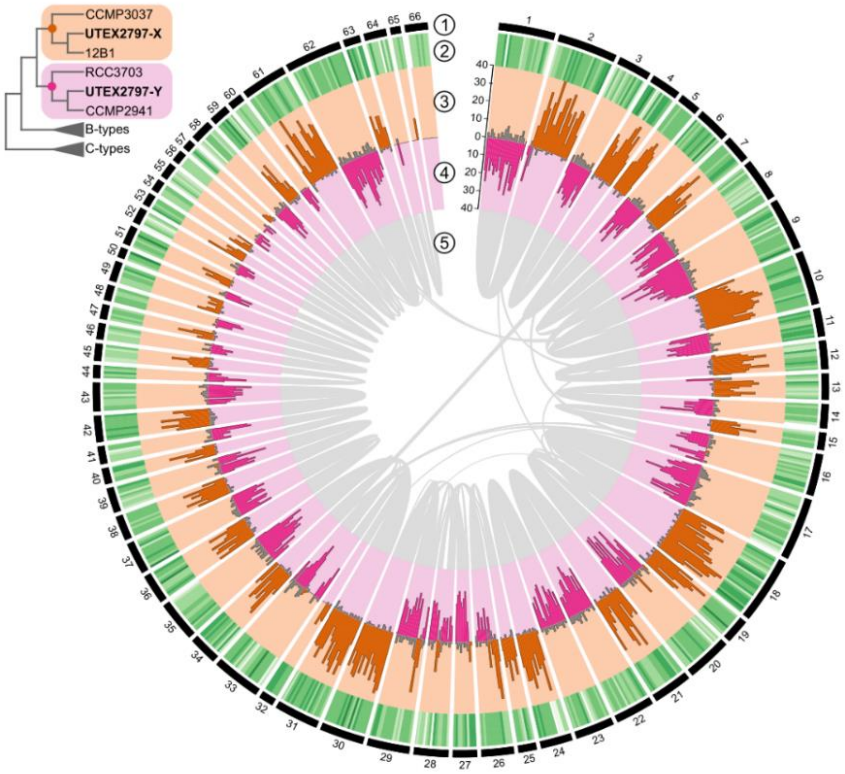

Figure 4-3 Hybrid genome structure of UTEX2797. Multi-label species tree of *P. parvum* strains with UTEX2797 modeled as an allopolyploid (top left) and Circos plot of UTEX2797 genome assembly (center). Track 1 represents assembled scaffolds. Track 2 is a heatmap displaying gene density along scaffolds in 250kb windows. Track 3 is a bar plot projected outwards displaying the number of genes in a 250kb window that phylogenetically group in clade Y in the species tree (orange) Track 4 is a bar plot projected inwards displaying the number of genes in a 250kb window that phylogenetically group in clade X in the species tree (pink). If the number of genes grouping in clade X are greater than clade Y, the bars of track 3 are colored orange, and vice versa for track 4. Track 5 displays syntenic blocks between scaffolds in the form of grey bands connecting syntenic regions.

To investigate if the two subgenomes can be distinguished at the assembly level, we identified the sister taxa of individual UTEX2797 genes. A clear pattern emerges between syntenic scaffolds in which one scaffold is primarily comprised of genes that group with subclade X, while the other is largely comprised of genes that group with subclade Y (Figure 4-3).

A second species tree was constructed to resolve phylogenomic relationships between strains with low heterozygosity, excluding strains 12A1, UTEX2797, and K0081 (Figure 4-4). The topology of this tree depicts the three chemotypes as three major phylogenetic clades. Additionally, the average nucleotide sequence identity between strains was calculated using 11,083 single copy gene families (Figure 4-4). Average identities ranged from 94.4% to 99.6% with the greatest dissimilarity occurring between A-type and C-type strains.

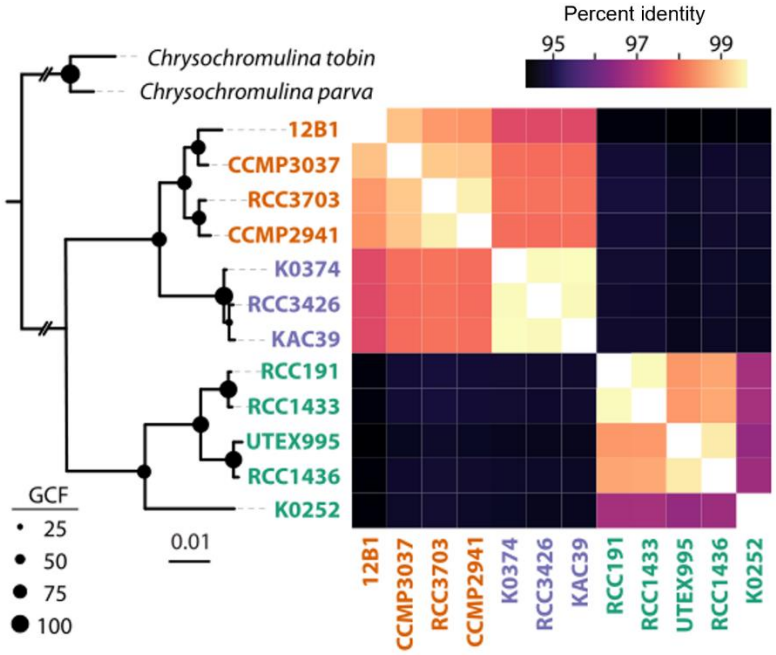

Figure 4-4 Species tree and average sequence identity of *P. parvum* strains. Species tree (left) of strains with low measures of heterozygosity and gene concordance factors (GCF) of each node displayed as proportional black circles. Heatmap (right) displaying average pairwise nucleotide identities between strain coding sequences.

To further explore genome variation in *P. parvum*, the haploid genome size and amount of heterozygosity for all strains was estimated *in silico* using the distribution of k-mer frequencies in whole genome Illumina sequencing data[32]. Illumina reads were first filtered to exclude contamination and normalized to permit inter-strain comparison of k-mer coverage (See Methods). Seven strains (12B1, CCMP3037 KAC-39, K-0374, RCC3426, RCC191, RCC1433) showed no evidence of heterozygosity, as illustrated by a single homozygous peak in k-mer plots (Fig.4-5). The coverage of maximal unique k-mers (CMUK) varied across these seven strains according to their chemotypes (Fig. 4-5, Table 4-3). Here, CMUK serves as a proxy for haploid genome size, with larger CMUKs indicative of smaller haploid genomes and vice versa. A-types, 12B1 and CCMP3037, had the largest CMUKs of 116 and 126, respectively. The two C-type strains, RCC191 and RCC1433, had intermediate CMUKs of 107 and 102, respectively. B-types had the smallest CMUKs (KAC-39 = 54, K-0372 = 54, and RCC3426 = 56), indicating that these strains had significantly larger haploid genomes compared to A- and C-type strains.

In addition to the seven strains that lacked discernable heterozygosity, five strains (CCMP2941, RCC3703, UTEX995, RCC1436, K-0252) had medium levels of genome heterozygosity, as illustrated by a second smaller heterozygous peak in k-mer plots at half the CMUK of the dominate homozygous peak (Fig. 4-5). The B-type strain K-0081 had a k-mer frequency plot without clear heterozygous and homozygous peaks, which could be due to the non-clonal nature of the strain as suggested in a previous study[1]. Lastly, two A-type strains 12A1 and UTEX2797 had pronounced heterozygous peaks indicative of very high levels of heterozygosity (Fig. 4-5). Despite high variations in heterozygosity, the homozygous peaks of A-type strains were consistent (mean CMUK = 120.2, Fig. 4-5, Table 4-3)

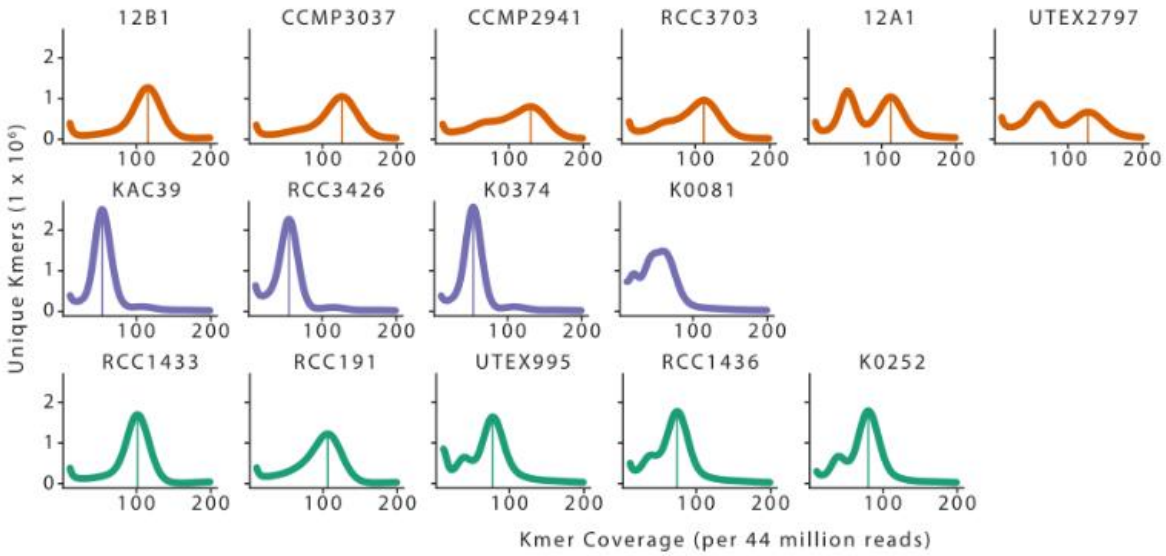

Figure 4-5 Normalized k-mer frequency plots of *P. parvum* strains. K-mer frequency plots of Illumina gDNA libraries following contamination filtration and depth normalization. Plots are arranged by A-type (top row), B-type (center row), and C-type (bottom row) as inferred through chemotyping and phylogenetic placement. Lines are drawn at coverages of maximal unique frequency (CMUK) at homozygous peaks.

Table 4-5 Haploid genome size ratios. CMUK = Coverage of maximal unique k-mers at homozygous peak in Illumina gDNA k-mer profile. An (*) denotes chemotypes that are inferred from phylogenetic placement

Strain	Chemotype	CMUK homozygous peak
12A1	A	112
12B1	A	116
CCMP2941	A	130
CCMP3037	A	126
RCC3703	A	111
UTEX2797	A	127
K-0081	B	59
K-0374	B	54
KAC-39	B	54
RCC3426	B*	56
K-0252	C	80
RCC1433	C	102
RCC1436	C	75
RCC191	C	107
UTEX995	C*	77

4.4 Discussion

In this study, we provide evidence of hybridization as well as large genome variation in *P. parvum*. Through the generation of two highly contiguous genome assemblies coupled with phylogenomic analyses, we identified the presence of two unique subgenomes in the widely studied *P. parvum* strain UTEX2797. Further, subgenomes are identifiable at the assembly level, though genes identified as subgenome X are more present than those of subgenome Y at a ratio of 1.36:1. Additionally, putative post-hybridization structural rearrangements do appear throughout the UTEX2797 genome as indicated by alternations in phylogenetic signal along a scaffold, though closer inspection of proper assembly at these regions is needed.

The two subgenomes of UTEX2797 group phylogenetically within the larger A clade are characterized by the production of A-type prymnesin. Notably, subclade X is composed of strains isolated from North American locations (12B1 from Texas and CCMP3037 from Wyoming), while subclade Y is composed of strains isolated from the United Kingdom (RCC3703) and Russia (CCMP2941). The introduction of *P. parvum* into North American waters has previously been suggested to be a result multiple introduction events from European populations[33]. If global transfer of *P. parvum* strains is a common enough occurrence, hybridization between formerly geographically distant strains could be possible, creating a hybrid vigor effect by equipping the

hybrid with adaptations to both environments. However, sampling of more A-type strains is needed to more accurately infer the evolutionary history and geographic origin of the two parental strains of UTEX2797 as well as the timing of hybridization.

As the first identification of hybridization in the haptophyte lineage, the prevalence of this phenomena still needs to be further assessed. Preliminary measurements of heterozygosity in *P. parvum* strain 12A1 also display high levels of heterozygosity (1.48%) similar to UTEX2797 (1.72%). Interestingly, both strains 12A1 and UTEX2797 were isolated eight years apart from the Colorado River Basin in Texas, USA. Additionally, putative haploid strain 12B1 was isolated from the same bloom event as 12A1, supporting the possibility bloom events are comprised of multiple genotypes and ploidy states.

Our analysis of the k-mer profiles of thirteen additional *P. parvum* strains reveals major differences in haploid genome sizes. Using CMUK as an inverse proxy of haploid genome size, we show over two-fold differences in haploid genome size between strains. Haploid genome sizes are generally consistent within chemotypes, with B-types showing the largest haploid genome size (mean CMUK=55.8), A-types the smallest size (mean CMUK=120.3), and C-types with an intermediate size (mean CMUK=88.2). Sexual recombination between strains with such large discrepancies in haploid genome size and nucleotide identity is likely not feasible and by this definition, supports the three chemotypes as discrete or cryptic species.

4.5 Methods

4.5.1 Culturing methods

Strains and their respective media types are summarized in Table 4-4. Cultures were kept at 20°C using a 12:12 light dark cycle in 200 μmol^2 of light.

Table 4-6 Summary of strains and their respective prymnesin types, culturing conditions. Reference 1[7], reference 2[6], and PC (personal communication with Timothy Fallon) denote sources of chemotype information

Strain	Prymnesin Type	PSU	Media	Location
12A1	A-type (PC)	8	L1-Si+NH4Cl	Texas, USA
12B1	A-type (PC)	8	L1-Si+NH4Cl	Texas, USA
CCMP2941	A-type (PC)	25	L1-Si	Lake Repnoye, Russia
CCMP3037	A-type (PC)	16	Black Sea	Wyoming, USA
K-0081	B-type (Ref. 1,2)	8	L1-Si+NH4Cl	Thy, Denmark
K-0252	C-type (Ref. 1), B-type (Ref. 2)	8	L1-Si+NH4Cl	Victoria, Australia
K-0374	B-type (Ref. 1,2)	8	L1-Si+NH4Cl	Norway
KAC-39	B-type (Ref. 1,2)	8	L1-Si+NH4Cl	Norway
RCC1433	C-type (Ref. 1,2)	32	L1-Si	Sallanelles, France
RCC1436	C-type (Ref. 1,2)	32	L1-Si	Plestin Les Grèves, France
RCC191	C-type (Ref. 1,2)	32	L1-Si	Dorset, UK
RCC3426	Unknown	32	L1-Si	Ryfylke, Norway
RCC3703	A-type (PC, Ref. 1)	32	L1-Si	Millport, UK
UTEX2797	A-type (Ref. 1,2)	8	L1-Si+NH4Cl	Texas, USA
UTEX995	Unknown	32	L1-Si	Essex, UK

4.5.2 Genome sequencing and assembly

Genomic DNA for Illumina sequencing was extracted from *P. parvum* cell pellets using the CTAB method according to the following protocol https://dx.doi.org/10.17504/protocols.io.b5qhq5t6[34]. Extracted DNA was purified using a Genomic DNA Clean and Concentrator kit (Zymo Research). Sequencing libraries were constructed and sequenced to produce 150-bp paired-end reads using one of two approaches: 1) libraries were prepared using a TruSeq DNA PCR-Free library prep kit (Illumina, San Diego, CA), and sequenced using an Illumina NovaSeq 6000 at the Purdue Genomics Center 2) libraries were prepared using an NEBNext DNA library prep kit (New England Biolabs Inc.) and sequenced using an Illumina NovaSeq 6000 by Novogene Corporation Inc. (Sacramento, CA). Illumina gDNA read quality was assessed by FastQC v0.10.0[35]. Short-read only genome assemblies were performed by Abyss v2.2.4[36] using a k-mer size of 96. Contigs less than 500 bp in length were discarded.

For long-read sequencing with Oxford Nanopore Technologies (ONT), high molecular weight DNA was extracted from isolated *P. parvum* nuclei using the following protocol https://dx.doi.org/10.17504/protocols.io.7b7hirn[37]. At least 1.5 µg of gDNA was used as input for

an Oxford Nanopore LSK-109 library ligation kit and sequenced on R9 MinION flow cells. Base calling was performed with Guppy v2.3.5[38]. Reads less than 3 kbp long or with quality scores less than 7 were discarded. Different assembly approaches were selected to optimize for either assembly contiguity (in the case of low heterozygosity 12B1) or the amount of resolved haplotypes (in the case of high heterozygosity UTEX2797). The 12B1 long-read assembly was created using both Nanopore and Illumina gDNA data via MaSuRCA v3.3.1[39] with the following parameters: LHE_COVERAGE=60, CA_PARAMETERS=cgwErrorRate=0.15, K-MER_COUNT_THRESHOLD=2, CLOSE_GAPS=1, JF_SIZE=5000000000. The UTEX2797 long-read assembly was created using only Nanopore data via Canu v2.1.1[40] with an expected genome size of 200 Mbp. Both assembly types were error corrected via five rounds of polishing with Illumina gDNA reads using Pilon v1.23[41].

Chromatin conformation capture data was generated using a Phase Genomics (Seattle, WA) Proximo Hi-C 2.0 Kit, which is a commercially available version of the Hi-C protocol[42]. Following the manufacturer's instructions for the kit, intact cells were crosslinked using a formaldehyde solution, digested using the DPNII restriction enzyme, end repaired with biotinylated nucleotides, and proximity ligated to create chimeric molecules composed of fragments from different regions of the genome that were physically proximal in vivo. Molecules were pulled down with streptavidin beads and processed into an Illumina-compatible sequencing library. Illumina gDNA reads were mapped back to the Abyss assemblies using BWA v0.7.15[43]. Sequencing was performed on an Illumina HiSeq. Reads were aligned to the long-read assemblies (Canu assembly for UTEX2797 and Masurca assembly for 12B1) following the manufacturer's recommendations. Briefly, reads were aligned using BWA-MEM[43] with the -5SP and -t 8 options specified, and all other options default. SAMBLASTER[44] was used to flag PCR duplicates, which then excluded. Alignments were then filtered with samtools[45] using the -F 2304 filtering flag to remove non-primary and secondary alignments. Putative misjoined contigs were broken using Juicebox[46,47] based on the Hi-C alignments. Kraken v2[48] was used to identify eukaryotic contigs, which were separated from prokaryotic contaminants and selected for scaffolding. The same alignment procedure was repeated from the beginning on the resulting corrected assembly. Phase Genomics' Proximo Hi-C genome scaffolding platform was used to create chromosome-scale scaffolds from the corrected assembly as previously described[49]. As in the LACHESIS method[50], this process computes a contact frequency matrix from the aligned Hi-C read pairs, normalized by the number

of DPNII restriction sites (GATC) on each contig, and constructs scaffolds in such a way as to optimize expected contact frequency and other statistical patterns in Hi-C data. Approximately 60,000 separate Proximo runs were performed to optimize the number of scaffolds and scaffold construction to make the scaffolds as concordant with the observed Hi-C data as possible.

4.5.3 Gene prediction

De novo repeat identification was performed on the phase assemblies of strains 12B1 and UTEX2797 using RepeatModeler v2.0.1[51], and repeats were masked using RepeatMasker v4.0.7. For the short-read only assemblies, repeats were masked using the UTEX2797 repeat library.

To maximize capture of the *P. parvum* transcriptome for gene calling, we performed RNA-seq of UTEX2797 cultures growth in 6 different conditions and 4 diurnal timepoints (Table 4-5). Starting 100 mL cultures were inoculated at 10,000 cells/mL. Starting at five days post inoculation, cultures were maintained using semi-continuous replacement every three days by discarding 10% of the culture and replacing with fresh media. Cell density was measured every three days to track culture growth. Upon reaching densities of ~$1x10^6$ cells/mL, cultures were harvested by centrifugation at 4,500xg for 5 minutes and snap freezing in liquid nitrogen. RNA extracted from pelleted cells using the following protocol: dx.doi.org/10.17504/protocols.io.3byl4k6r8vo5/v1. Stranded RNAseq libraries were prepared with a NEBNext Ultra TM RNA Library Prep Kit (NEB, USA) following manufacturer's recommendations. Illumina RNA-seq reads were aligned to the UTEX2797 phased assembly using STAR v2.7.8a[52].

Table 4-6. Conditions of RNA-seq experiments. Micronutrient concentrations are in proportion to full L1-Si media concentrations

Condition	Salinity (psu)	Vitamins Conc.	Light (umol^2)	Phosphorus Conc.	Nitrogen Conc.	Replicates
Standard (12pm)	32	Full	200	Full	Full	5
Standard (6pm)	32	Full	200	Full	Full	1
Standard (12am)	32	Full	200	Full	Full	1
Standard (6am)	32	Full	200	Full	Full	1
Medium salinity	11	Full	200	Full	Full	1
Low salinity	2	Full	200	Full	Full	1
Low vitamin	32	1/10	200	Full	Full	1
Low phosphorus	32	Full	200	1/25	Full	1
Low nitrogen	32	Full	200	Full	1/25	1
Low light	32	Full	30	Full	Full	1

Gene model and protein prediction was first conducted on the UTEX2797 phased assembly with BRAKER2 v2.1.5[53,54]. BRAKER2 was supplied a repeat softmasked genome, a custom protein database comprised of Swiss-Prot and all haptophyte predicted proteins from the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP[55]), and the aligned UTEX2797 Illumina RNA-seq data. The resulting Augustus species-specific training configuration file[56] was saved and passed to the BRAKER2 runs for all other assemblies along with the custom protein database.

4.5.4 Bacterial contamination

Bacterial contamination in the short-read only assemblies was identified using Blobtools v1.1.1[57]. For each strain, Illumina gDNA reads were aligned to the Abyss assembly using BWA v0.7.15[43] to generate a coverage BAM file. Assembly contigs were queried against the NCBI nucleotide (nt) database using blastn v2.11.0[58]. DIAMOND v2.0.8.146[59] was used to query predicted proteins against a custom protein databases that consisted of NCBI RefSeq (release 98)[60] and was supplemented with additional predicted protein sequences from MMETSP[55] and the 1000 Plants transcriptome sequencing project (OneKP)[61]. The blobtools taxrule ‘bestsumorder’ was used to determine the taxonomic assignment of each contig prioritizing information from protein hits first. Contigs denoted as non-eukaryotic in origin were removed to produce the final filtered assembly.

Lastly, BBSplit v38.87[62] was used to exclude all Illumina reads that did not map to the filtered assembly.

4.5.5 Heterozygosity

Heterozygosity and coverage of maximal unique k-mers was estimated *in silico* from contamination-filtered Illumina gDNA reads using GenomeScope v2.0[32] with a k-mer (k=21) depth distribution calculated using KMC v3.1.1[63]. For generation of normalized k-mer profiles, the same number of reads were randomly subsampled from the Illumina gDNA libraries of each strain.

4.5.6 Functional annotation

Assessment of conserved single-copy eukaryotic genes was performed by BUSCO v4.0.6[28] by searching the predicted proteomes of each strain against the eukaryote_odb10 database. Functional annotation of predicted proteomes was performed using InterProScan v5.50-84.0[64].

4.5.7 Identification of orthologous gene families

Homology between the longest predicted proteins from each gene of select *P. parvum* strains and two other haptophyte species *Chrysochromulina tobin* and *Chrysochromulina parva* was determined with OrthoFinder v2.3.11[65] using the following parameters: -S diamond -M msa -T fasttree.

4.5.8 Phylogenetic tree building

For nucleotide sequences, the longest transcript of each gene was aligned with Guidance v2.02[66] using the following parameters: --msaProgram MAFFT –seqType codon. For peptide sequences, the longest encoded proteins of each gene were aligned with MAFFT v7.471 (Katoh and Standley 2013) using the E-INS-I strategy and following parameters: --maxiterate 1000 --bl 45 --op 1.0 --retree 3. Gene trees were built from alignments using IQ-TREE v1.6.12[67] and the built in ModelFinder[68] test to determine the best-fit substitution model and performing SH-aLRT and the

ultrafast bootstrapping analyses with 1000 replicates each. For the multi-labeled species tree, Orthogroups were filtered to retain gene families containing one 12B1 gene and two UTEX2797 genes that were identified as syntelogs from the CoGe analysis (see below). The number of sequences in the other strains was allowed to vary; however, orthogroups containing more than 20 sequences and or less than 13 strains were excluded. The resulting 8,903 nucleotide trees built from these gene families and the species tree provided by OrthoFinder were used to build a multi-labeled species trees with GRAMPA v1.3[31]. Sister taxa of each UTEX2797 gene was parsed from the nucleotide trees using custom python scripts and visualized using Circos v0.69-9[69]. Total subgenome gene content was assessed using resolved gene trees supplied by OrthoFinder. The single label species tree was constructed with 1,468 single copy gene trees built from amino acid alignments with average bootstrap values greater than 75 using the IQ-TREE concatenation approach with the edge-linked proportional partition model and ultrafast bootstrapping with 1000 replicates. Nucleotide alignments used to calculate strain pairwise sequence identity were first removed of gaps using Trimal v1.4.1[70] with the -nogaps parameter. Strain identity was determined as the average of trimmed nucleotide alignment identities calculated using the PhyKit pairwise_identity function[71].

4.5.9 Synteny analysis

Pairwise synteny between the Hi-C scaffolded genomes of UTEX2797 and 12B1 was identified and visualized with the JCVI pipeline[72]. Syntenic blocks within the UTEX2797 scaffolded genome were detected using SynMap2[73] on the online Comparative Genomics Platform (CoGe) using Quota Align and Quota Align Merge for syntenic depth and merge syntenic blocks algorithms.

4.6 References

1 Larsen A, Bryant S. Growth rate and toxicity of *Prymnesium parvum* and *Prymnesium patelliferum* (haptophyta) in response to changes in salinity, light and temperature. *Sarsia* 1998; **83**: 409–418.

2 Legrand C. Phagotrophy and toxicity variation in the mixotrophic Prymnesium patelliferum (Haptophyceae). *Limnol Oceanogr* 2001; **46**: 1208–1214.

3 Tillmann U. Kill and eat your predator: A winning strategy of the planktonic flagellate Prymnesium parvum. *Aquat Microb Ecol* 2003; **32**: 73–84.

4 Carvalho WF, Granéli E. Contribution of phagotrophy versus autotrophy to Prymnesium parvum growth under nitrogen and phosphorus sufficiency and deficiency. *Harmful Algae* 2010; **9**: 105–115.

5 Igarashi T, Satake M, Yasumoto T. Prymnesin-2: A Potent Ichthyotoxic and Hemolytic Glycoside Isolated from the Red Tide Alga Prymnesium parvum. *J Am Chem Soc* 1996; **118**: 479–480.

6 Rasmussen SA, Meier S, Andersen NG *et al*. Chemodiversity of Ladder-Frame Prymnesin Polyethers in Prymnesium parvum. *J Nat Prod* 2016; **79**: 2250–2256.

7 Binzer SB, Svenssen DK, Daugbjerg N *et al*. A-, B- and C-type prymnesins are clade specific compounds and chemotaxonomic markers in Prymnesium parvum. *Harmful Algae* 2019; **81**: 10–17.

8 Skovgaard A, Hansen PJ. Food uptake in the harmful alga Prymnesium parvum mediated by excreted toxins. *Limnol Oceanogr* 2003; **48**: 1161–1166.

9 Driscoll WW, Espinosa NJ, Eldakar OT, Hackett JD. Allelopathy as an emergent, exploitable public good in the bloom-forming microalga prymnesium parvum. *Evolution (N Y)* 2013; **67**: 1582–1590.

10 Johnsen TM, Eikrem W, Olseng CD, Tollefsen KE, Bjerknes V. Prymnesium parvum: The Norwegian Experience. *JAWRA J Am Water Resour Assoc* 2010; **46**: 6–13.

11 Southard GM, Fries LT, Barkoh A. Prymnesium parvum: The Texas Experience. *JAWRA J Am Water Resour Assoc* 2010; **46**: 14–23.

12 Larsen A, Eikrem W, Paasche E. Growth and toxicity in *Prymnesium patelliferum* (Prymnesiophyceae) isolated from Norwegian waters. *Can J Bot* 1993; **71**: 1357–1362.

13 Rashel RH, Patiño R. Influence of genetic background, salinity, and inoculum size on growth of the ichthyotoxic golden alga (Prymnesium parvum). *Harmful Algae* 2017; **66**: 97–104.

14 Lysgaard ML, Eckford-Soper L, Daugbjerg N. Growth rates of three geographically separated strains of the ichthyotoxic Prymnesium parvum (Prymnesiophyceae) in response to six different pH levels. *Estuar Coast Shelf Sci* 2018. doi:10.1016/j.ecss.2018.02.030.

15 Medić N, Varga E, Waal DB Van de, Larsen TO, Hansen PJ. The coupling between irradiance, growth, photosynthesis and prymnesin cell quota and production in two strains of the bloom-forming haptophyte, Prymnesium parvum. *Harmful Algae* 2022; **112**: 102173.

16 La Claire JW, Manning SR, Talarski AE. Semi-quantitative assay for polyketide prymnesins isolated from Prymnesium parvum (Haptophyta) cultures. Toxicon. 2015; **102**: 74–80.

17 Leadbeater BSC, Green JC. *The Haptophyte algae / edited by J.C. Green, B.S.C. Leadbeater*. Published for the Systematics Association by Clarendon Press ; Oxford University Press: Oxford : New York, 1994.

18 Larsen A, Medlin LK. Inter- and Intraspecific Genetic Variation in Twelve Prymaesium (haptophyceae) Clones1. *J Phycol* 1997; **33**: 1007–1015.

19 Larsen A. Prymnesium parvum and P. patelliferum (Haptophyta)—one species. *Phycologia* 1999; **38**: 541–543.

20 Anestis K, Kohli GS, Wohlrab S *et al.* Polyketide synthase genes and molecular trade-offs in the ichthyotoxic species Prymnesium parvum. *Sci Total Environ* 2021; **795**: 148878.

21 Larsen A, Edvardsen B. Relative ploidy levels in Prymnesium parvum and P. patelliferum (Haptophyta) analyzed by flow cytometry. *Phycologia* 1998; **37**: 412–424.

22 Green JC, Course PA, Tarran GA. The life-cycle of Emiliania huxleyi: A brief review and a study of relative ploidy levels analysed by flow cytometry. *EHUX (Emiliania huxleyi)* 1996; **9**: 33–44.

23 Edvardsen B, Vaulot D. Ploidy Analysis of the Two Motile Forms of Chrysochromulina Polylepis (prymnesiophyceae)1. *J Phycol* 1996; **32**: 94–102.

24 Talarski A, Manning SR, La Claire JW. Transcriptome analysis of the euryhaline alga, Prymnesium parvum (Prymnesiophyceae): Effects of salinity on differential gene expression. *Phycologia* 2016. doi:10.2216/15-74.1.

25 Taylor RB, Hill BN, Langan LM, Chambliss CK, Brooks BW. Sunlight concurrently reduces Prymnesium parvum elicited acute toxicity to fish and prymnesins. *Chemosphere* 2021; **263**: 127927.

26 Richardson ET, Patiño R. Growth of the harmful alga, Prymnesium parvum (Prymnesiophyceae), after gradual and abrupt increases in salinity. *J Phycol* 2021; **57**: 1335–1344.

27 Chen D, Yuan X, Zheng X *et al.* Insights into algae evolution for adapting to blue-green light garnered from the Isochrysis galbana genome. 2020. doi:10.22541/au.160881384.48495723/v1.

28 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015; **31**: 3210–3212.

29 Hovde BT, Deodato CR, Hunsperger HM *et al.* Genome Sequence and Transcriptome Analyses of Chrysochromulina tobin: Metabolic Tools for Enhanced Algal Fitness in the Prominent Order Prymnesiales (Haptophyceae). *PLOS Genet* 2015; **11**: e1005469.

30 Hovde BT, Deodato CR, Andersen RA, Starkenburg SR, Barlow SB, Cattolico RA. Chrysochromulina: Genomic assessment and taxonomic diagnosis of the type species for an oleaginous algal clade. *Algal Res* 2019; **37**: 307–319.

31 Gregg WCT, Ather SH, Hahn MW. Gene-Tree Reconciliation with MUL-Trees to Resolve Polyploidy Events. *Syst Biol* 2017; **66**: 1007–1018.

32 Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* 2020; **11**: 1432.

33 Lutz-Carrillo DJ, Southard GM, Fries LT. Global Genetic Relationships Among Isolates of Golden Alga (Prymnesium parvum) 1. *J Am Water Resour Assoc* 2010; **46**: 24–32.

34 Auber R. Total DNA extraction from plant tissue using CTAB method. protocols.io. 2022.https://dx.doi.org/10.17504/protocols.io.bamnic5e.

35 Bioinformatics B. FastQC A Quality Control tool for High Throughput Sequence Data. 2022.https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

36 Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: A parallel assembler for short read sequence data. *Genome Res* 2009; **19**: 1117–1123.

37 Auber R, Wisecaver J. Algal nuclei isolation for Nanopore sequencing of HMW DNA. protocols.io. 2022.https://dx.doi.org/10.17504/protocols.io.7b7hirn.

38 Technologies ON. Nanopore Community. 2022.http://nanoporetech.com/community.

39 Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. *Bioinformatics* 2013; **29**: 2669–2677.

40 Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. *Genome Res* 2017; **27**: 722–736.

41 Walker BJ, Abeel T, Shea T *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014; **9**. doi:10.1371/journal.pone.0112963.

42 Lieberman-Aiden E, van Berkum NL, Williams L *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (80-)* 2009; **326**: 289–293.

43 Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009; **25**: 1754–1760.

44 Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* 2014; **30**: 2503–2505.

45 Li H, Handsaker B, Wysoker A *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–2079.

46 Rao SSP, Huntley MH, Durand NC *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014; **159**: 1665–1680.

47 Durand NC, Robinson JT, Shamim MS *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* 2016; **3**: 99–101.

48 Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014; **15**: R46.

49 Bickhart DM, Rosen BD, Koren S *et al.* Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet* 2017; **49**: 643–650.

50 Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* 2013; **31**: 1119–1125.

51 Flynn JM, Hubley R, Goubert C *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci* 2020; **117**: 9451–9457.

52 Dobin A, Davis CA, Schlesinger F *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; **29**: 15–21.

53 Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-Genome Annotation with BRAKER. In: Kollmar M (ed). *Gene Prediction: Methods and Protocols*. Springer New York: New York, NY, 2019, pp 65–95.

54 Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinforma* 2021; **3**: lqaa108.

55 Keeling PJ, Burki F, Wilcox HM *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol* 2014; **12**: e1001889.

56 Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* 2006; **34**: W435–W439.

57 Laetsch DR, Blaxter ML. BlobTools: Interrogation of genome assemblies. *F1000Research* 2017; **6**: 1287.

58 Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009; **10**: 421.

59 Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015; **12**: 59–60.

60 O'Leary NA, Wright MW, Brister JR *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016; **44**: D733–D745.

61 Matasci N, Hung L-H, Yan Z *et al.* Data access for the 1,000 Plants (1KP) project. *Gigascience* 2014; **3**: 17.

62 Bushnell B. BBTools Software Package. 2017.

63 Kokot M, Długosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 2017; **33**: 2759–2761.

64 Jones P, Binns D, Chang HY *et al.* InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 2014; **30**: 1236–1240.

65 Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019; **20**: 238.

66 Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res* 2015; **43**: W7–W14.

67 Minh BQ, Schmidt HA, Chernomor O *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 2020; **37**: 1530–1534.

68 Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017. doi:10.1038/nmeth.4285.

69 Krzywinski M, Schein J, Birol İ *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res* 2009; **19**: 1639–1645.

70 Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009; **25**: 1972–1973.

71 Steenwyk JL, Buida TJ, Labella AL, Li Y, Shen X-X, Rokas A. PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics* 2021; **37**: 2325–2331.

72 Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and Collinearity in Plant Genomes. *Science (80-)* 2008; **320**: 486–488.

73 Haug-Baltzell A, Stephens SA, Davey S, Scheidegger CE, Lyons E. SynMap2 and SynMap3D: web-based whole-genome synteny browsers. *Bioinformatics* 2017; **33**: 2197–2198.

CHAPTER 5. PERSPECTIVES

5.1 Metabolic innovation in shikonin biosynthesis

1,4-Naphthoquinones are a major class of specialized metabolites that have independently evolved in a diverse set of plant lineages[1]. Through the study of the shikonin biosynthesis pathway in *L. erythrorhizon*, we investigated how metabolic innovation of 1,4-naphthoquinones arose in shikonin producers. The Boraginales order, which encompasses shikonin-producing plants, has been predicted to have undergone multiple rounds of whole genome duplication[2]. While WGD is recognized as a major source of gene duplication, our discovery of a retrotransposition-based duplication creating the core shikonin pathway gene PGT suggests that retrotransposition is also a relevant mechanism of metabolic gene innovation in this lineage. The relative contribution of retrotransposition to total genome-wide duplication levels is relatively unknown and further genome level analyses parsing exon structures of paralogous genes or target site duplications could assist in answering this question.

An evolutionary linkage between primary and specialized metabolic pathways was also revealed in our analysis of shikonin biosynthesis genes. In addition to sharing a 4HBA substrate, the proposed reactions in the biosynthetic pathways of shikonin and ubiquinone are highly analogous. Beyond this metabolic connection, our discovery of shared homology between the functioning prenyltransferases of shikonin (*PGT*) and ubiquinone (*PPT*) biosynthesis also provides an evolutionary connection. Thus, further elucidation of shikonin pathway genes could be guided by the evolutionary history of gene candidates. Interestingly, our coexpression analysis recovered several ubiquinone biosynthesis pathway gene homologs as coexpressed with core shikonin pathway genes. While the function of these homologs in ubiquinone biosynthesis are unknown (*e.g.* COQ4), functional characterization of their role in shikonin biosynthesis could reciprocally be used to inform their function in primary metabolism.

In addition to ubiquinone homologs, our analyses identified strong regulatory and enzymatic shikonin pathway gene candidates. Cytochrome P450 and acyltransferase gene candidates which could perform hydroxylation and acylation reactions on shikonin intermediates, respectively, are still to be experimentally tested. RNAi in hairy root cultures of *L. erythrorhizon* is a feasible approach to functionally confirm the roles of these genes in the shikonin pathway. Our

motif analyses also identified the enrichment of WRKY transcription factor binding domains upstream of genes coexpressing with known shikonin pathway genes. Notably, five genes annotated as WKRY transcription factors were recovered as gene candidates and serve as high priority candidates for validation experiments.

5.2 Harnessing global coexpression networks for specialized metabolic pathway elucidation

Global coexpression networks are recognized as useful tools to recover metabolic pathway genes[3]. However, this approach has typically been employed with large scale datasets with hundreds of conditions used as input. Because specialized metabolic pathway discovery is primarily conducted in non-model species, these datasets are usually not readily available. In our experimental design, we chose six conditions that varied shikonin content by differences in tissue localization, light, and media type. Our study demonstrates that global coexpression networks are still effective with a small number of datasets, given they are designed to perturb the pathway of interest. In theory, this approach can be employed to guide pathway discovery efforts in other specialized metabolic systems. Further, correlation of metabolite abundance data with gene expression data can also prove as a powerful approach[4]. Simultaneous sampling of both data types in *L. erythrorhizon* could prove effective in probing the shikonin biosynthesis pathway as well as other unelucidated metabolic pathways.

5.3 Genome variation in *Prymnesium parvum*

5.3.1 Hybridization

Prior to our study, allopolyploidzation had not yet been identified in the haptophyte lineage. Even across protist lineages more generally, few instances have been reported[5,6]. This could possibly be due to the lack of genomic resources for protists. The ability to fully resolve haplotypes or homoeologous subgenomes requires a multi-platform sequencing approach and robust assembly algorithms[7]. Now that hybridization has been identified in *P. parvum*, further work understanding the phenotypic implications of hybridization in *P. parvum* needs to be done. For example, the hybrid strain UTEX2797 grew faster than putative C-type strain UTEX995 in a series of growth experiments[8]. However, phenotypic measurements taken between UTEX2797 and other A-type

strains closely related to subgenome parental strains would be more appropriate to gauge the effect of hybridization. Further, metabolic quantification of prymnesin content would also provide insight into a possible dosage effect of prymnesin biosynthesis genes.

The geographical distance between the closest relatives of the parental strains of UTEX2797 warrants further investigation. While 12B1 and CCMP3037 were isolated from North American locations, the strains from subclade Y were isolated from the UK and Russia. Sampling of additional isolates from North America will likely assist in the identification of a closer relative of the ancestral parents. Additionally, another highly heterozygous strain, 12A1, was identified in our genome survey. Similar to UTEX2797, 12A1 was isolated from Texas and is a putative allopolyploid. However, the evolutionary history of 12A1 still needs to be investigated phylogenetically.

Given the hybrid genome structure of UTEX2797 and the significant retention of genes from both subgenomes (ratio of 1.36:1), questions remain concerning the expression dynamics between homeologous genes. In plants, hybridization often leads to one subgenome maintaining more genes and therefore becoming dominant over the other subgenome in order to maintain gene dosage balance[9]. To investigate expression dynamics in UTEX2797, I have performed RNA-seq experiments on both UTEX2797 and 12B1 (Table 5-1). Conditions were chosen to subject strains to as many micronutrient, abiotic, and biotic stressors possible, while also eliciting variation in prymnesin production[10]. For example, the addition of glycerol to cultures of *P. parvum* grown in darkness has been documented to affect growth rate and induce toxicity[11,12]. By identifying homoeologous gene pairs within UTEX2797 through synteny and homology, we can observe global trends in expression between subgenomes, as well as trends within modular pathways. Further, comparing expression between subgenome X and 12B1, a closely related strain to the ancestral parent of subgenome X, also allows us to observe the effects of hybridization on global gene expression. Beyond hybridization, UTEX2797 produces a significantly larger amount of prymnesin relative to 12B1 (Tim Fallon, personal communication,). Strain 12B1 has also been demonstrated to have a fitness advantage when grown in pure culture compared to putative hybrid 12A1, which has a fitness advantage when co-cultured with algal prey[13]. By comparing expression profiles of 12B1 and UTEX2797, we can search for candidate prymnesin biosynthesis genes that show preferential expression in conditions known to elicit toxicity or mixotrophy[10].

Table 5-1 Conditions of RNA-seq experiments performed to explore UTEX2797 and 12B1 gene expression dynamics. An (*) denotes conditions that are shock treatments.

Condition	Light ($umol^2$)	Temp (C)	Salinity (psu)	Phosphorus	Co-culture	Glycerol (M)
Standard	150	20	8	Full	None	0
Dark Toxicity	0	20	8	Full	None	0.5
Low Temperature*	150	10	8	Full	None	0
Low Salinity*	150	20	3.2	Full	None	0
High Salinity*	150	20	32	Full	None	0
Low Phosphorus	150	20	8	1/50	None	0
Prey feeding*	150	20	8	1/50	CCMP1179	0

5.3.2 Genome size variation

The large difference in estimated haploid genome size between phylogenetically distinct chemotypes of *P. parvum* is a strong indication that *P. parvum* is a cryptic species complex. Such genotypic divergence between chemotypes likely contributes to the major phenotypic differences observed between strains. The large putative genome size of B-types may be due to two scenarios. The common ancestor of A- and C-type strains may have experienced genome streamlining, or more parsimoniously, B-type strains may have undergone genome expansion. The amplification of transposable elements and other non-genic DNA in genomes has been a common mechanism of expansion in eukaryotic lineages[14–16]. To test this hypothesis in *P. parvum*, a more contiguous genome assembly of a B-type strain would permit a detailed inspection of transposable elements in the genome.

With new knowledge of the genetic differences between chemotypes, more informative phenotypic experiments can be designed. Further, identification of ploidy state may also contribute to phenotypic diversity. The presence of heterozygous peaks in A-type strains (CCMP2941, RCC3707), B-type strains (K0081), and C-type strains (RCC1433, UTEX995, K0252) is

indicative of ploidy states greater than one. Measurements of nuclear DNA content in these strains using propidium iodide staining and flow cytometry will complement our *in silico* approaches to determine if there are additional variations in ploidy states. Prior to the knowledge of this genetic variation, studies have identified differences in the scale structure of *P. parvum* strains using electron microscopy[17]. Perhaps such differences are unique to chemotypes or ploidy states, which has been observed in other haptophyte species[18]. Additionally, the proper assignment of phenotypic characteristics to each genotype or ploidy state could translate into more accurate modeling of natural *P. parvum* bloom events.

5.4 References

1 Meyer GW, Bahamon Naranjo MA, Widhalm JR. Convergent evolution of plant specialized 1,4-naphthoquinones: metabolism, trafficking, and resistance to their allelopathic effects. *J Exp Bot* 2021; **72**: 167–176.

2 Tang CY, Li S, Wang YT, Wang X. Comparative genome/transcriptome analysis probes Boraginales' phylogenetic position, WGDs in Boraginales, and key enzyme genes in the alkannin/shikonin core pathway. *Mol Ecol Resour* 2019. doi:10.1111/1755-0998.13104.

3 Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A. A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. *Plant Cell* 2017; **29**: 944–959.

4 Blazier A, Papin J. Integration of expression data in genome-scale metabolic network reconstructions . Front. Physiol. . 2012; **3**.https://www.frontiersin.org/article/10.3389/fphys.2012.00299.

5 Niwa K, Sakamoto T. Allopolyploidy in Natural and Cultivated Populations of Porphyra (bangiales, Rhodophyta)1. *J Phycol* 2010; **46**: 1097–1105.

6 Tanaka T, Maeda Y, Veluchamy A *et al.* Oil Accumulation by the Oleaginous Diatom Fistulifera solaris as Revealed by the Genome and Transcriptome . *Plant Cell Online* 2015; **27**: 162–176.

7 Ming R, Man Wai C. Assembling allopolyploid genomes: no longer formidable. *Genome Biol* 2015; **16**: 27.

8 Rashel RH, Patiño R. Influence of genetic background, salinity, and inoculum size on growth of the ichthyotoxic golden alga (Prymnesium parvum). *Harmful Algae* 2017; **66**:

97–104.

9 Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M, Wang X. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc Natl Acad Sci U S A* 2014; **111**: 5283–5288.

10 Granéli E, Salomon PS. Factors influencing allelopathy and toxicity in prymnesium parvum. *J Am Water Resour Assoc* 2010; **46**: 108–120.

11 RAHAT M, JAHN TL. Growth of Prymnesium parvum in the Dark; Note on Ichthyotoxin Formation*. *J Protozool* 1965; **12**: 246–250.

12 PADILLA GM. Growth and Toxigenesis of the Chrysomonad Prymnesium parvum as a Function of Salinity*. *J Protozool* 1970; **17**: 456–462.

13 Driscoll WW, Espinosa NJ, Eldakar OT, Hackett JD. Allelopathy as an emergent, exploitable public good in the bloom-forming microalga prymnesium parvum. *Evolution (N Y)* 2013; **67**: 1582–1590.

14 Bennetzen JL. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 2002; **115**: 29–36.

15 Kelkar YD, Ochman H. Causes and Consequences of Genome Expansion in Fungi. *Genome Biol Evol* 2012; **4**: 13–23.

16 Wanding Z, Gangning L, L. MP, A. JP. DNA methylation enables transposable element-driven genome expansion. *Proc Natl Acad Sci* 2020; **117**: 19359–19366.

17 Green JC, Hibberd DJ, Pienaar RN. The taxonomy of Prymnesium (Prymnesiophyceae) including a description of a new cosmopolitan species, P. Patellifera sp. nov., and further observations on P. parvum N. carter. *Br Phycol J* 1982; **17**: 363–382.

18 Houdan A, Billard C, Marie D *et al.* Holococcolithophore-heterococcolithophore (Haptophyta) life cycles: Flow cytometric analysis of relative ploidy levels. *Syst Biodivers* 2004; **1**: 453–465.