UNIVERSITY OF CALIFORNIA SAN DIEGO

A genomic approach to accessing and characterizing secondary metabolite biosynthetic pathways from marine red macroalgae

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

in

Chemistry

by

Taylor Sydney Steele

Committee in charge:

Professor Bradley S. Moore, Chair
Professor Elizabeth Komives, Co-Chair
Professor Michael Burkart
Professor William Fenical
Professor Alexis Komor

2023

The Dissertation of Taylor Sydney Steele is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my grandmothers, Norma Sanfilippo and Pamela Steele, thank you for sharing your curious minds and open hearts.

EPIGRAPH

Every new discovery is just a reminder – We're all small and stupid.

Everything Everywhere All at Once

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE iii

DEDICATION iv

EPIGRAPH v

TABLE OF CONTENTS vi

LIST OF FIGURES ix

LIST OF TABLES xv

LIST OF SCHEMES xvi

ACKNOWLEDGEMENTS xvii

VITA xx

ABSTRACT OF THE DISSERTATION xxi

CHAPTER 1. Specialized terpenoid biosynthesis across domains of life 1

1.1 Introduction and Context for Chapter 1 2

1.2 References for Chapter 1 Introduction 4

1.3 Abstract 5

1.4 Introduction 5

1.5 Basic Biosynthetic Principles 7

1.6 Terpenoid Distribution and Genome Mining 12

1.7 Plants 16

1.8 Fungi 23

1.9 Bacteria 28

1.10 Algae 32

1.11 Insects 38

1.12 Amoeba 42

1.13 Sponges 44

1.14 Discussion 45

1.15 References 49

1.16 Acknowledgements 61

CHAPTER 2. Domoic acid biosynthesis in the red alga *Chondria armata* suggests a complex evolutionary history for toxin production 62

2.1 Introduction and Context for Chapter 2 63

2.2 References for Chapter 2 Introduction 65

2.3 Reprint of “Domoic acid biosynthesis in the red alga *Chondria armata* suggests a complex evolutionary history for toxin production” 66

2.4 Acknowledgements 105

CHAPTER 3. Biosynthesis of haloterpenoids in red algae via microbial-like type I terpene synthases 106

3.1 Introduction and Context for Chapter 3 107

3.2 References for Chapter 3 introduction 109

3.3 Abstract 111

3.4 Introduction 112

3.5 Results and Discussion 115

3.6 Conclusions 124

3.7 Main Text References 126

3.8 Supplementary Information 129

3.9 Supplementary Information References 170

3.10 Acknowledgements 171

CHAPTER 4. Ongoing investigations of red algal genomics and haloterpenoid biosynthesis downstream tailoring reactions 172

4.1 Introduction and Context for Chapter 4 173

4.2 References for Chapter 4 introduction 175

4.3 Introduction 176

4.4 Red Algal DNA Barcoding 181
4.5 High Molecular Weight DNA Extraction for Marine Macroalgal Tissue 183
4.6 Genome Mining and Co-occurrence Networks 187
4.7 Vanadium-Dependent Haloperoxidase Expression Efforts Introduction 191
4.8 General Strain Methodology 195
4.9 Attempted Expression of VHPOs in *Escherichia coli* 197
4.10 Attempted Characterization of VHPOs in *Streptomyces coelicolor* CH999 198
4.11 Attempted Characterization of VHPOs in *Saccharomyces cerevisiae* 201
4.12 Discussion 203
4.13 References 207
4.14 Acknowledgements 211
CHAPTER 5. Perspectives and outlook 212

LIST OF FIGURES

Figure 1.1 Structural overview of TS enzymes 11

Figure 1.2 Counts of terpenoids by compound category 12

Figure 1.3 Phylogenetic tree of representative class 1 TS sequences 14

Figure 1.4 Function and architecture of plant TS enzymes 17

Figure 1.5 Diterpene-scaffold generation by tri-domain TSs and their connection with oxidized diterpenoids 19

Figure 1.6 Sesqui- and monoterpenoids and their connected hydrocarbons produced by sesqui- and mono-TS 21

Figure 1.7 Several monoterpenes produced by a *Nicotiana suaveolens* mono-TS 23

Figure 1.8 Fungal TS scaffold generation 25

Figure 1.9 Cyclooctatin biosynthesis. GGPP synthase CotB1 catalyzes the production of GGPP (5) from IPP (1) and DMAPP 29

Figure 1.10 A. Geosmin (**44**) and B. 2-methylisoborneol (2-MIB) (**45**) biosynthesis 30

Figure 1.11 Algal TS scaffold generation 33

Figure 1.12 Crystal structure of DabA (PDB ID:6VL1). Panel highlights modified active site residues DDXXD is NDXXA and NSE is NTE. Orange is the DabA product *N*-geranyl-l-glutamate 36

Figure 1.13: Cladogram of red algae by Order (A) by Class (B) within Florideophyceae 37

Figure 1.14 Insect terpenoid scaffolds 39

Figure 1.15 Discodiene biosynthesis in *Dictyostelium discoideum*. FPP (**4**) is converted to discoidol (**55**) by the class 1 TS DdTPS8, which is then converted to discoidiene (**56**) by the CYP450 CYP521A1 42

Figure 1.16 A. Sponge nitrogenous terpenes. B. Products of sponge terpene synthases 44

Figure 2.1 Kainoid biosynthetic genes and producing organisms 68

Figure 2.2 Visualization of syntenic comparisons between *dab*, *rad*, and *kab* gene clusters 69

Figure 2.3 Activities and predicted structures of the known kainoid synthase enzymes. 70

Figure 2.4 Phylogenetic analysis of (A) kainoid synthase and (B) coclustered DA CYP450 enzymes (circle denotes key branch points) 71

Figure 2.S1 Cladogram of DA and KA producers by order in Rhodophyta. Blue and red dots indicate KA and DA producing orders, respectively. See Table S1 for a list of producing organisms 89

Figure 2.S2 K-mer analysis of the *C. armata* genome. Tetraploid (41x), diploid (83x), and haploid (166x) peaks are labeled. Sequences to the left of the dashed line were omitted. Haploid genome size was estimated to be 480 Mb 90

Figure 2.S3 BUSCO assessment of red macroalgal genomes. *C. armata* BUSCO assessment highlighted in black box 91

Figure 2.S4. The three copies of the rad cluster within the *C. armata* genome. Mean coverage across each copy in parentheses. The sequences of *rad1*, *rad2*, and *rad3* are deposited in GenBank, accession numbers OK169902, OK169903, and OK169904, respectively 92

Figure 2.S5 Domoic acid isomers in *C. armata*. Negative mode LC-HRMS chromatograms using LC Method A of domoic acid (DA), isodomoic acids (isoA, isoB, isoC), and cNGG standards compared with *C. armata* extract, showing extracted ion chromatograms (EIC) for anticipated DA isomer and cNGG 93

Figure 2.S6 10% SDS-PAGE gel. Sample order is as follows: Ladder, MBP-RadA1, RadC1, RadC2. Note the additional band in the RadA lane is the MBP tag without RadA 94

Figure 2.S7 Kainoid synthase reactions were set up as previously described using NGG (1 mM). (A) NGG and dainic acid isomers. (B) Negative mode LC-HRMS chromatograms using LC Method A of kainoid synthase reactions are shown together with dainic acid standards 95

Figure 2.S8 Prekainic acid substrate screen. Kainoid synthase reactions were set up as previously described using prekainic acid (1 mM). Negative mode LC-HRMS chromatograms using LC Method B of kainoid synthase reactions are shown together with prekainic acid and kainic acid standards 96

Figure 2.S9 CASTp output of modeled pocket volumes for Pseudo-nitzschia multiseries DabC, C. armata RadC1, and Digenea simplex KabC. Enzyme binding pockets highlighted in red, measurements of pocket area and volume listed below each model label 97

Figure 2.S10 Overnight assays with purified MBP-□7-RadA1. Substrate screening suggest both GPP and DMAPP can be accepted as substrates to make NGG and

prekainic acid, respectively. RadA1 N-prenyltransferase reactions were set up as previously described using L-glutamate acid (20 mM) and DMAPP or GPP prenyl donors (1 mM) 98

Figure 2.S11 N-prenyltransferase enzyme maximum-likelihood phylogenetic tree. RadA sequences mapped to previously reported tree. They clade with the other N-prenyltransferases and are very distantly related to other terpene cyclases 99

Figure 2.S12 Expanded diatom CYP450 maximum-likelihood phylogenetic tree. Diatom sequences used in Figure 4B highlighted in blue and red. Numeric identifiers are Joint Genome Institute Protein IDs (JGI PIDs) from the respective publicly available diatom genomes through JGI 100

Figure 2.S13 Expanded *C. armata* CYP450 maximum-likelihood phylogenetic tree. RadD highlighted in red 101

Figure 2.S14 1H-NMR spectrum of isodomoic acid C. Analysis performed in D_2O/600 MHz. The signal of the residual MeOD was adjusted at δ 3.30 ppm as the internal reference in D_2O 102

Figure 3.1 Haloterpenoid producing red macroalgae and representative associated metabolites 112

Figure 3.2 Phylogenetic analysis of red algal TS sequences 116

Figure 3.3 Algal TS biochemical characterization 119

Figure 3.4 Assembled contigs from in *L. pacifica* and *L subopposita* show the physical co-localization of TSs with genes encoding for other TSs (red) or a VHPO (blue) 122

Figure 3.5. Vanadium-dependent haloperoxidase phylogeny including representative bacterial, fungal, algal, and plant VHPO sequences. Putative red algal VHPO sequences highlighted in red, LsVHPO-1 is labeled. The scale measures evolutionary distances in substitutions per amino acid 123

Figure 3.S1 The distribution of TS sequences by class and order. Purple circle indicates the presence of a TS and a white circle indicates sequencing data is available, but no TS sequence has been identified. The classes Porphyridiophyceae and Florideophyceae harbor MTSs or 7 out of 49 investigated species to-date 144

Figure 3.S2 BUSCO assessment of red macroalgal genomes. *P. pacificum, L. subopposita L. pacifica, and P. hornemannii* BUSCO assessment highlighted in black box 145

Figure 3.S3 Sequence alignment of ten new red algal TS sequences, key active site residues are highlighted in yellow and summarized in blue in sequence logo (bottom) 146

Figure 3.S3 Cont. Sequence alignment of ten new red algal TS sequences, key active site residues are highlighted in yellow and summarized in blue in sequence logo (bottom) 147

Figure 3.S4 10% SDS-PAGE gel. Sample order is as follows: Ladder, PpTS-CA, PhTS-1, PhTS-2, PhTS-3, LsTS-1, LsTS-2, LsTS-3, LsTS-4, LsTS-5, LsTS-6, Ladder, LpTS-B, Ladder, LpTS-C 148

Figure 3.S5 GCMS analysis of PpTS-C2 with the substrates GPP (top) FPP (middle top) GGPP (middle bottom) and standard of ocimene isomers (bottom). Mass spectra for major peaks are shown 149

Figure 3.S6 GCMS analysis of PhTS-1 with the substrates GPP (top) FPP (middle top) GGPP (middle bottom) and standard of myrcene isomers (bottom). Mass spectra for major peaks are shown 150

Figure 3.S7 GCMS analysis of PhTS-2 with the substrates GPP (top) FPP (middle top) and GGPP (middle bottom) and germacrene D (**14**) standard (bottom). Mass spectra for major peaks are shown. 151

Figure 3.S8 GCMS analysis of PhTS-3 with the substrates GPP (top) FPP (middle top) and GGPP (middle bottom) and valencene (**15**) (bottom). Mass spectra for major peaks are shown 152

Figure 3.S9 GCMS analysis of LsTS-1 with the substrates GPP (top) FPP (middle) and GGPP (bottom). Mass spectra for major peaks are shown 153

Figure 3.S10 GCMS analysis of LsTS-2 with the substrates GPP (top) FPP (middle) and GGPP (bottom). Mass spectra for major peaks are shown 154

Figure 3.S11 GCMS analysis of LsTS-3 with the substrates GPP (top) FPP (middle top) and GGPP (middle bottom) and germacrene B (**12**) (bottom). Mass spectra for major peaks are shown 155

Figure 3.S12 GCMS analysis of LsTS-4 with the substrates GPP (top) FPP (middle top) GGPP (middle bottom) and ledene (**9**) standard (bottom). Mass spectra for major peaks are shown 156

Figure 3.S13 GCMS analysis of LsTS-5 with the substrates GPP (top) FPP (middle) and GGPP (bottom). Mass spectra for major peaks are shown 157

Figure 3.S14 GCMS analysis of LsTS-6 with the substrates GPP (top) FPP (middle) and GGPP (bottom) 158

Figure 3.S15 GCMS analysis of LphTPS-B with the substrates GPP (top) FPP (middle) and bicylogermacrene standard (**10**) (bottom). Mass spectra for major peaks are shown. 159

Figure 3.S16 GCMS analysis of LphTPS-C with the substrates GPP (top) FPP (middle) and GGPP (bottom). Mass spectra for major peaks are shown 160

Figure 3.S17 ^{1}H-NMR spectrum of **8** in C_6D_6 161

Figure 3.S18 ^{13}C-NMR spectrum of **8** in C_6D_6 162

Figure 3.S19 ^{1}H-NMR spectrum of **13** in $CDCl_3$ 163

Figure 3.S20 ^{13}C-NMR spectrum of **13** in $CDCl_3$ 164

Figure 3.S21 ^{1}H- ^{13}C-HSQC spectrum of **13** in $CDCl_3$ 165

Figure 3.S22 1H- 13C-HMBC spectrum of **13** in CDCl3 166

Figure 3.S23 ^{1}H- ^{1}H--COSY spectrum of **13** in $CDCl_3$ 167

Figure 3.S24 ^{1}H- ^{1}H-NOESY spectrum of **13** in $CDCl_3$ 168

Figure 3.S25 Proposed bromonium-induced halocyclization cascade by LsVHPO-1 in oppositol (**4**) biosynthesis. LsVHPO-1 is co-clustered with the germacrene D-4-ol (**8**) producing TS, LsTS-2 169

Figure 4.1 Supply limited bioactive marine natural products and their host organisms. 176

Figure 4.2 Gene co-occurrence network. Each dot represents a gene, and each connection indicates physical co-occurrence on a contiguous sequence without respect to order or distance 188

Figure 4.3 Gene clusters in *Laurencia subopposita* and *Laurencia pacifica*. A. TS gene clusters, red – algal TS, blue – VHPO. B. Syntenic comparison of UbiA gene clusters. Pink – UbiA, orange – IDS, blue – VHPO, green – short chain dehydrogenase 189

Figure 4.4 Representative phylogeny of vanadium-dependent haloperoxidases (VHPO) across various taxa and detailed view of new 'microbial-like' algal VHPO clade. Sequences attempted to be biochemically characterized are bolded and highlighted with an asterisk 192

Figure 4.5 Model of LsVHPO-1 the co-clustered VHPO with a terpene synthase in the Laurencia subopposita genome 194

Figure 4.6 *Streptomyces coelicolor* CH999 expression results of PhVHPO-3. A. FPLC trace of Ni-NTA affinity chromatography run B. 10% SDS-PAGE gel of PhVHPO-3 purification 199

Figure 4.7 *Saccharomyces cerevisiae* expression of LsVHPO-1. A. Western blot of LsVHPO-1, from left to right: ladder, supernatant, and pellet. B. Fluorescence of GFP in induced and uninduced cultures 202

Figure 4.8 Proposed dihalogenation reaction to incorporate fluorine hypothesized could be catalyzed by the newly identified class of red algal vanadium-dependent haloperoxidases 205

LIST OF TABLES

Table 1.1 Raw data associated with Figure 1.2. Counts of terpenoids by compound category in the Natural Products Database. 13

Table 2.S1 Summary of known KA and DA red algal producers from the literature 82

Table 2.S2 Primers used in this study 83

Table 2.S3 QUAST statistics of the final assembled *C. armata* genome 84

Table 2.S4 BUSCO assessment of publicly available red macroalgal genomes 85

Table 2.S5 Kainoid synthase representative sequences 86

Table 2.S6 CYP450 representative sequences 87

Table 2.S7 Accession numbers of sequences deposited and used in this study 88

Table 3.S1 NMR data for LpTPS-C product (500MHz, $CDCl_3$), in line with literature .. 137

Table 3.S2 Distribution of terpene synthase sequences in Rhodophyta 139

Table 3.S2 Cont. Distribution of terpene synthase sequences in Rhodophyta 140

Table 3.S3 Genome sequencing summary statistics 141

Table 3.S4 UniProt accession numbers of red algal (RA) IDS sequences supplementing previously published TS list and VHPO sequences for constructing TS and VHPO phylogenies. 142

Table 3.S5 Top 30 BlastP hits of VHPO-TS co-localized glycosyltransferase 143

Table 4.1 DNA barcoding primers used in this study: RuBisCo: Ribulose-1,5-bisphosphate carboxylase large subunit (rbcL, ~1,350 bp) 182

Table 4.2 DNA barcoding primers used in this study: CO1: Cytochrome c oxidase subunit 1 extended fragment (COI, ~1,232 bp) 182

Table 4.3 *Streptomyces coelicolor* CH999 *in vitro* PhVHPO-3 assay conditions 200

Table 4.4 *Saccharomyces cerevisiae* assay conditions for both *in vivo* and crude lysate assays 203

LIST OF SCHEMES

Scheme 1.1 Overview of isoprenoid biosynthesis pathways from across the tree of life 8

Scheme 4.1 Vanadium-dependent haloperoxidase (VHPO) mechanism. Specific microbial VHPOs are proposed to form a lysine chloroamine intermediate, which directs a regiospecific chlorination reaction. All other known VHPOs lack an additional lysine in the active site and perform non-specific halogenation reactions..... 193

ACKNOWLEDGEMENTS

I would like to thank my advisor and committee chair, Bradley S. Moore, for his thoughtfulness and steady leadership over the past five years. He models excellence both in, and out, of the lab. Thank you for teaching me to show up with kindness and enthusiasm in every aspect of life. When all else fails, go jump in the ocean and try again. I would also like to thank my co-advisor and co-chair Elizabeth Komives for her mentorship and assistance navigating graduate school between two departments. I am very grateful for your advocacy and support. To my committee, thank you for your encouragement and feedback, particularly to William Fenical for the many enlightening natural products discussions.

I would also like to acknowledge my undergraduate research advisors, Professors Ariane Jansma, David E. Cummings, and Ryan Botts. Thank you, Ariane, for instilling in me a deep love of proteins and your patient guidance teaching me the fundamentals of biochemistry. To Dr. Cummings and Dr. Botts, thank you both for your trust when you handed me a MinION sequencer and said, ‘go figure it out’. You both changed the course of my career. Dr. Botts, the loss of your mentorship feels like a conversation cut short, thank you for your sharing your love of bikes and bioinformatics.

To Professor Jon Chekan, whether you know it or not, your patience and humor set me up for success in graduate school. Thank you for teaching me which details are important, and how to (mostly) keep myself organized. I would also like to thank Dr. Immo Burkhardt, Dr. April Lukowski, Dr. Vikram Shende, and Dr. Stella de Lima Camargo. Your fingerprints will always be on my work, and I am grateful for your friendship and mentorship throughout the years. Thank you Rebbi, or Dr. Rebecca Schäfer, I am so

grateful for the many miles we spent together on our bikes. You showed me I could climb scientific and athletic mountains I never thought I could conquer. To all past and future members of the Moore lab, thank you for the foundations you have laid and for the stories to come. I am so grateful for the privilege of being a Moore lab member.

To Sonia Teder-Moore and Malia Moore thank you for the times you have included me in your family and shared a story over coffee, you have both made graduate school, and the Moore lab, a warmer, better place to be. To Kai, thank you for reminding me why I love science and, especially, thank you for sharing my love of seaweed and including me in your journey. Learning to dive together and watching you grow as a scientist was one of my favorite parts of graduate school. You are an excellent human, and I am so proud of you.

Most importantly, thank you to my family. To my parents, Carolyn and John, my aunt Joanne, and my grandparents Norma and Joe, thank you for your unwavering love and support. Grandma, thank you for reminding me every day is a new day worth pursuing. To Luna Bean, for every late night and quiet moment in the sunshine, every Ph.D. student deserves a cat like you. Finally, to Gabe and Otter, thank you for your consistent love and support, and reminding me to find joy in each day.

Chapter 1 contains material currently being prepared for publication, coauthored with Wilson, K.A., Moore, B.S., and Burkhardt, I. The dissertation author was the primary co-author of this chapter with Wilson, K.A.

Chapter 2, in full, is a reprint of the material as it appears in *The Proceedings of the National Academy of Sciences*, Steele, T.S., Brunson, J.K., Maeno, Y., Terada, R.,

Allen, A.E., Yotsu-Yamashita, M., Chekan, J.R., and Moore, B.S., 2022. The dissertation author was the primary co-researcher and co-author of this chapter with Brunson, J.K.

Chapter 3 contains material currently being prepared for publication, coauthored with Burkhardt, I., Moore, M., de Rond, T., Michael, T., and Moore, B.S. We are grateful to the Joint Genome Institute, Department of Energy for their ongoing support and sequencing data generated for this project (CSP-504335). The dissertation author was the primary researcher and author of this material.

Chapter 4 contains unpublished material coauthored with Moore, M., Burkhardt, I., and Moore, B.S. The dissertation author was the primary researcher and author of this chapter.

VITA

2017 Bachelor of Science in Biology-Chemistry, Point Loma Nazarene University

2019 Master of Science in Chemistry, University of California San Diego

2019-2022 National Science Foundation, Graduate Research Fellow

2023 Doctor of Philosophy in Chemistry, University of California San Diego

PUBLICATIONS

Steele, T.S., Brunson, J.K., Maeno, Y., Terada, R., Allen, A.E., Yotsu-Yamashita, M., Chekan, J.R., and Moore, B.S. Domoic acid biosynthesis in the red alga *Chondria armata* suggests a complex evolutionary history for toxin production. *Proc. Natl. Acad. Sci. U.S.A*, 119**,** (6) e2117407119 (2022).

Chen, P.Y.T., Adak, S., Chekan, J.R., Liscombe, D.K., Miyanaga, A., Bernhardt, P., Diethelm, S., Fielding, E.N., George, J.H., Miles, Z.D., Murray, L.A.M., **Steele, T.S.,** Winter, J.M., Noel, J.P, and Moore, B.S. Structural basis of stereospecific vanadium-dependent haloperoxidase family enzymes in napyradiomycin biosynthesis. *Biochemistry*. 61, 1844–1852 (2022).

Wilson, K.A., de Rond, T., Burkhardt, I., **Steele, T.S.,** Schäfer, R., Podell, S., Allen, E.E., Moore, B.S. Terpene biosynthesis in marine sponge animals. *Proc. Natl. Acad. Sci. U.S.A*. 120**,** (9) e2220934120 (2023).

ABSTRACT OF THE DISSERTATION

A genomic approach to accessing and characterizing secondary metabolite biosynthetic pathways from marine red macroalgae

by

Taylor Sydney Steele

Doctor of Philosophy in Chemistry

University of California San Diego, 2023

Professor Bradley S. Moore, Chair

Professor Elizabeth Komives, Co-Chair

Nature is a gifted chemist, and marine organisms are capable producers of structurally and functionally diverse small molecules. Many of these genetically encoded small molecules, or natural products, are extremely bioactive and have inspired the development of new classes of therapeutic agents. However, the availability and scalability of marine systems has limited efforts to bring promising compounds to the clinic. Synthetic and biosynthetic research in natural products offers complementary strategies to preserve sensitive marine organisms while gaining access to their unique

chemistry. Similarly, advancements in computing, and the emergence of technologies like long-read, third-generation sequencing, have made it cheaper, faster, and more efficient to connect genes to chemistry – a critical step in circumventing supply issues through the development of sustainable, biocatalytic routes to a target compound. The goal of this dissertation is to uncover the genetic basis for how marine red algae, or seaweeds produce anthropologically relevant chemicals, such as the neurotoxin, domoic acid, and the preclinical antitumor agent, halomon.

Both molecular families discussed in this dissertation belong to the broader chemical classification of terpenoids. Chapter 1, which serves as an introduction to the dissertation, is a taxonomy guided review of terpenoid natural products. Chapter 2 of the dissertation describes the sequencing and assembly of the draft genome of the red macroalga *Chondria armata*. Here we identify and compare domoic acid biosynthetic gene clusters and describe the *in vitro* validation of key biosynthesis enzymes. This work fills a gap in understanding of kainoid biosynthesis in red macroalgae. Chapter 3 describes the sequencing of four haloterpenoid producing red algae, *Plocamium pacificum*, *Laurencia subopposita*, *Laurencia pacifica*, and *Portieria hornemannii*, the notable producer of the haloterpenoid antitumor agent, halomon. Through this work, I identify haloterpenoid biosynthesis pathways from each organism and report the *in vitro* reconstitution of algal class 1 microbial-type terpene synthases. Finally, Chapter 4 discusses genome sequencing and mining strategies in red algae. This work also addresses downstream tailoring steps of algal haloterpenoids. Overall, the work here describes genomic approaches to accessing and characterizing secondary metabolite biosynthetic pathways from marine red algae.

CHAPTER 1. Specialized terpenoid biosynthesis across domains of life

1.1 Introduction and Context for Chapter 1

As the largest and most widely distributed class of natural products, terpenoids are the core backbones of many medicinally and industrially important compounds.[1] Terpenoids perform various essential physiological roles ranging from mediating growth and reproduction to facilitating complex interkingdom communication.[2] Terpenoids and isoprenoids are both synonymous for natural products constructed from two foundational 5-carbon (C5) building blocks: dimethylallyl diphosphate (DMAPP) and isopentenyl diphosphate (IPP). DMAPP and IPP derive from two metabolic pathways, either the mevalonate pathway (MEV) or the methylerythritol 4-phosphate (MEP) pathway. The MEV pathway, which utilizes acetyl CoA and is found in most eukaryotes and archaea, was the first and (for several decades) only known pathway for isoprenoid biosynthesis.[3] The MEP pathway, which utilizes 3-glyceral phosphate and pyruvate, was discovered later, and is found in most bacteria, as well as in plants, which contain both the MEV and MEP pathways.[4] The existence of two independent, nonhomologous pathways to produce terpenoid biosynthetic precursors reinforces the essential nature of terpenoids in both primary and secondary metabolism. Terpene scaffolds are constructed primarily using class 1 terpene synthases, which lay the foundation for the vast structural diversity of terpenoids. Both types of molecules discussed in this dissertation, the kainoids in Chapter 2 and the haloterpenoids in Chapters 3 and 4, are considered terpenoids, and are discussed in this review. The work discussed in Chapters 2-4, along with the discoveries by my coauthors, Kayla Wilson and Immo Burkhardt, in sponges and corals, respectively, provided the inspiration for the following review article.[5,6]

This chapter is currently in preparation for submission to Natural Product Reports, co-authored with Kayla Wilson, Immo Burkhardt, and Bradley S. Moore. The purpose of this article is to review class 1 terpene synthases in natural product biosynthesis. Terpene synthases have been extensively studied in plants, fungi, and bacteria, and traditionally review articles have covered each taxon individually. Recent discoveries in terpenoid biochemistry in insects, algae, corals, and sponges has rapidly expanded our understanding of terpene biosynthesis in these distinct evolutionary lineages. Therefore, this review is intended to compile terpenoid research across the tree of life and discuss the diversity, enzymology, and evolution of class 1 terpene synthases. We provide key examples from each well-studied taxon (plants, fungi, and bacteria) and provide an in-depth discussion of newly discovered families of terpene synthases in insects, algae, corals, and sponges. If available, we describe the genomic context of each terpene synthase and provide insight into the dynamic role of downstream tailoring enzymes in terpenoid biosynthesis. We also highlight the apparent modes of evolution of terpene synthases, 1. gene duplication and neofunctionalization and 2. horizontal gene transfer. Finally, we tie together the similarities, differences, and overarching principles of terpenoid secondary metabolism and make connections to the evident evolutionary success of class 1 terpene synthases.

1.2 References for Chapter 1 Introduction

1. D. W. Christianson, Structural and Chemical Biology of Terpenoid Cyclases. *Chem. Rev.* **117**, 11570–11648 (2017).

2. J. M. Galindo-Solís, F. J. Fernández, Endophytic Fungal Terpenoids: Natural Role and Bioactivities. *Microorganisms* **10** (2022).

3. J. Lombard, D. Moreira, Origins and early evolution of the mevalonate pathway of isoprenoid biosynthesis in the three domains of life. *Mol Biol Evol* **28**, 87–99 (2011).

4. G. Flesch, M. Rohmer, Prokaryotic hopanoids: the biosynthesis of the bacteriohopane skeleton. *Eur. J. Biochem.* **175**, 405–411 (1988).

5. K. Wilson, T. de Rond, I. Burkhardt, T. S. Steele, R. J. B. Schäfer, S. Podell, E. E. Allen, B. S. Moore, Terpene biosynthesis in marine sponge animals. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2220934120 (2023).

6. I. Burkhardt, T. de Rond, P. Y.-T. Chen, B. S. Moore, Ancient plant-like terpene biosynthesis in corals. *Nat Chem Biol* **18**, 664–669 (2022).

1.3 Abstract

Terpenoids are widely distributed and considered the largest class of natural product chemicals. Found across all domains of life, terpenoids encompass both primary and secondary, or specialized metabolic products. The general biosynthetic machinery for terpenoid precursor supply is ubiquitous across all organisms, though only certain groups are known to produce specialized terpenoids. While initial work on terpenoid biosynthesis focused on plants, the field has rapidly expanded in recent decades, first with bacterial and fungal terpene synthases, and later with those found in insects, algae, and marine animals. The rapid expansion of terpenoid biosynthetic knowledge has been fueled by DNA sequencing and genomic advances that reveal genes encoding terpenoid biosynthesis are often co-clustered, from microbes to plants and animals. These organisms construct terpene scaffolds primarily using class 1 terpene synthases, which lay the foundation for the vast structural diversity of terpenoids. In this review we aim to highlight terpenoid diversity in secondary metabolism across different branches of the tree of life and discuss the apparent evolutionary success of underlying enzymatic principles.

1.4 Introduction

Terpenoids are widely considered to be the largest and most diverse class of naturally occurring compounds. Estimates of characterized terpenoids currently number more than 100,000, excluding steroids, and encompass both primary and secondary metabolism. In this review, we defer to work by Chevrette and co-authors to define primary and secondary metabolism.[1] Primary, or central metabolism, refers to pathways with a wide taxonomic distribution and whose products are required for autonomous

growth, sometimes supported by redundant, more promiscuous pathways. Secondary metabolism refers to pathways with limited taxonomic distribution and are often the result of enzyme promiscuity and recruitment. Secondary metabolism evolution is typically dynamic, undergoing diversifying or purifying selection to offer a competitive advantage in a given environment, which may have a profound effect on survival. Within secondary metabolism, specialized metabolism is defined as a secondary metabolic pathway in which the metabolite is experimentally described to contribute to niche-specific functions. In primary metabolism, terpenoids, such as steroids and carotenoids, perform various essential physiological roles like mediating growth and reproduction.[2] In secondary metabolism, terpenoids help an organism adapt to changes in their environment and serve key functions, such as signaling, defense, and attraction. Initial work on secondary terpenoids often focused on terrestrial plants, which are well known for production of terpenoid-rich essential oils,[3] however, these 'secondary' terpenoids are now understood to be ubiquitous across the tree of life.[4]

Beyond terrestrial plants, terpenoids outside of primary metabolism have been characterized in fungi,[5] bacteria,[6] and algae[7] as well as many marine animals, like octocorals,[8] and sponges.[9] Terpenoids also serve many human purposes, for example as key components in the production of widely used fragrances, flavorings, and resins. The medicinal properties of terpenoids have been recognized by humans for centuries in traditional medicines, as well as in modern treatment. Many notable pharmaceuticals are, or are inspired by terpenoid natural products, such as taxol and artemisinin.[10, 11] Similarly, because of their role in plant defense mechanisms, many terpenoids are toxic to herbivores and other pests, making them a valuable tool for natural pest control.

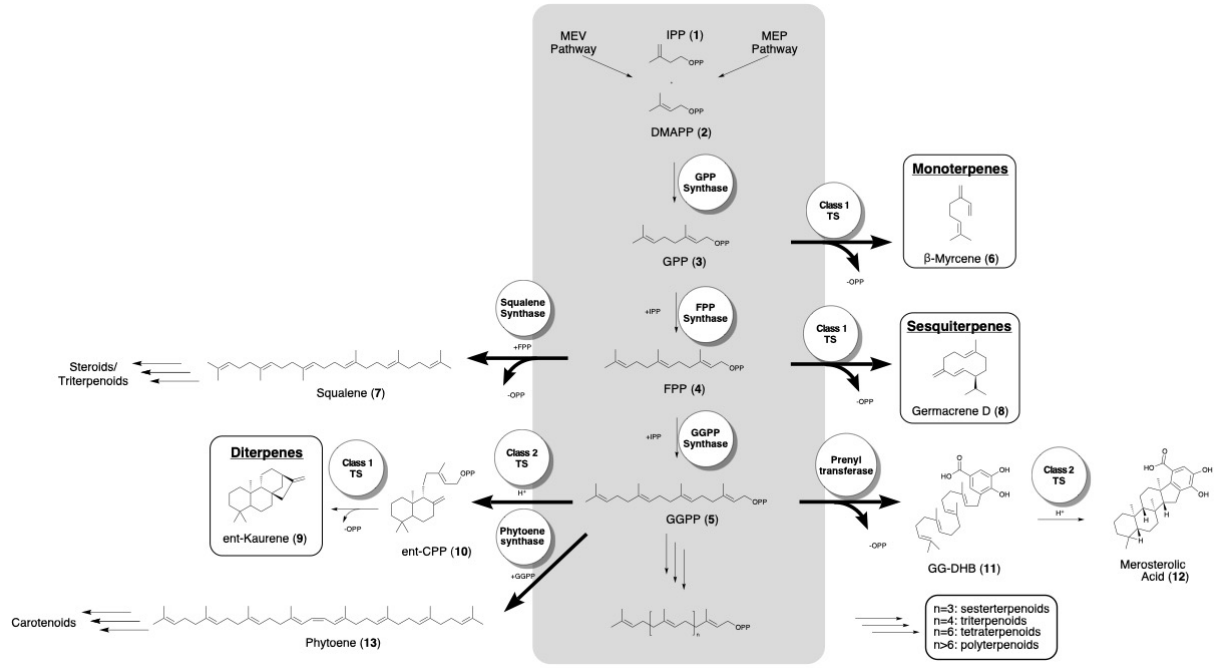
Overall, terpenoids play a crucial role in the natural world, contributing to the survival and reproduction of many species. While many aspects of terpenoid chemistry, biosynthesis, evolution and ecology have been reviewed separately, the recent discoveries of class 1 terpene synthases in many different types of organisms with distinct lifestyles, producing diverse chemical structures with a virtually identical set of enzymes, has raised the questions: 1. What renders the terpene biosynthetic logic and underlying enzymology suitable for these different lifeforms? and 2. What could explain the apparent interkingdom success of terpenoids compared to other classes of secondary metabolites? Thus, in this review we first define class 1 terpenoid diversity and distribution around the tree of life and discuss the enzymatic principles and biosynthetic logic for the assembly of these compounds. Then, we highlight elucidated biosynthetic pathways, bioactivities, and ecological aspects of class 1 terpenoids produced in taxa with identified terpene synthase (TS) genes. Finally, we discuss similarities, differences and overarching principles that might indicate the apparent success of this secondary metabolite machinery.

1.5 Basic Biosynthetic Principles

Terpenoids, terpenes, and isoprenoids are all synonymous for natural products constructed from two basic 5-carbon (C5) building blocks: isopentenyl diphosphate (IPP, **1**) and dimethylallyl diphosphate (DMAPP, **2**) (**Scheme 1.1**). Compounds **1** and **2** can derive from two independent nonhomologous metabolic pathways, the mevalonate pathway (MEV) or the methylerythritol 4-phosphate (MEP) pathway. The MEV pathway, which utilizes acetyl CoA and is found in most eukaryotes and archaea, was the first and (for several decades) only known pathway for isoprenoid biosynthesis.[12] Originally

described in yeast and animals in the 1950s, it was assumed to be the only route for IPP and DMAPP biosynthesis until the discovery of the MEP pathway in the 1990s.[13] The MEP pathway, which utilizes 3-glyceral phosphate and pyruvate, is found in most bacteria, as well as in plants, which contain both the MEV and MEP pathways. Interestingly, the uncharacterized Candidate Phyla Radiation (CPR) bacterial phylum seems to have a MEV pathway equivalent.[14, 15] In pl[11,12]ants and other photosynthetic eukaryotes like algae, the MEP pathway is located exclusively in the plastid, suggesting the pathway dates to the ancestral cyanobacterial endosymbiont from which the plastid originated, though some phylogenetic evidence contradicts this theory.[12] Overall, the evolutionary history and conservation of these two isoprenoid biosynthesis pathways reinforces the importance of their biological functions.

Compound **2** is extended with **1** units by an isopentenyl diphosphate synthase (IDS) in a 'head-to-tail' reaction to yield the linear isoprenoid diphosphate precursors,



Scheme 1.1 Overview of isoprenoid biosynthesis pathways from across the tree of life.

beginning with C10 geranyl diphosphate (GPP, **3**) and then C15 farnesyl diphosphate (FPP, **4**). Further extension of **4** with **1** yields C20 geranylgeranyl diphosphate (GGPP, **5**) and finally C25 geranylfarnesyl diphosphate (GFPP) (**Scheme 1.1**).

Every known organism possesses the biosynthetic machinery to produce these linear oligoprenyl precursor and there are numerous classes of enzymes that utilize isoprenyl diphosphate substrates and can be transformed into a variety of terpenoid compounds. These terpenoid compounds are typically named according to the number of C5 isoprene units incorporated into their structure, beginning with hemiterpenes (one), monoterpenes (two), sesquiterpenes (three), diterpenes (four), and sesterterpenes (five), triterpenes (six) and lastly tetraterpenes (eight). In a majority of organisms, two equivalents of FPP are converted by a squalene synthase in a head-to-head reaction to squalene (**Scheme 1.1**), which is the precursor to many primary metabolites like steroids and hopanoids. GGPP can be dimerized head-to-head as well to produce phytoene, the universal precursor to all phytoenoids and carotenoids. Isoprenyl diphosphate substrates can be acted upon by class 2 TSs, which catalyze a protonation-induced cyclization, and often leaves the pyrophosphate group intact for further biosynthetic steps. Isoprenyl diphosphates are also utilized by prenyltransferases, which transfer the allylic prenyl group onto a receptor molecule. This prenyl chain can be further cyclized by terpene synthases, often containing multiple rings and stereocenters. The resulting terpenoid products from these described reactions are key structural scaffolds to the immense and remarkable structural diversity of terpenoids.

The enzymes that convert isoprenyl diphosphates to secondary/specialized terpenoids directly are known as class 1 terpene cyclases (TSs) and will be the focus of

this review. Unlike class 2 TS enzymes, which generate carbocations via protonation induced cyclization, class 1 TSs rely on dephosphorylation to catalyze allylic cation formation. Consequently, class 1 TSs can act either directly on linear IDS products like GPP and FPP or on the products of class 2 TSs that still have an intact pyrophosphate group, like *ent*-CPP (**Scheme 1.1**).

The structure of a class 1 TS stabilizes the allylic cation and facilitates the many intramolecular reactions notable in terpenoid biochemistry, including hydride and methyl/alkyl shifts, cyclization, and carbocation quenching. These cationic cascades result in polycyclic and complex hydrocarbon products. Structurally, all class 1 TS enzymes are composed of a metal-binding alpha-helical bundle, a feature that is also seen in, and likely arose from IDS enzymes (**Figure 1.1A**).[16] Most IDSs and non-plant TSs are single domain enzymes consisting of one alpha domain (**Figure 1.1B**), while plant terpene cyclases are di-, or tri-domain enzymes (**Figure 1.1C**). Class 1 plant TSs always contain an alpha domain, which shares the same fold as monodomain enzymes. Didomain enzymes consist of this alpha domain and an additional beta domain, which shares the conserved fold of squalene-cyclase like class 2 TSs but is often catalytically inactive. Three domain enzymes consist of this alpha- and beta-domain assembly and contain an additional gamma domain, which does not participate in catalysis but most likely has a structural role in these proteins. While the amino acid sequences of class 1 TSs are often quite variable, the overall structure and key active site motifs are conserved. The dephosphorylation reaction is induced by a magnesium cluster, which is bound by highly conserved active site residues, namely the aspartate rich DDXXD motif (**Figure 1.1D**) and NSD/DTE motif (**Figure 1.1D**). In microbial-type class 1 TSs, a highly

conserved RY dimer near the C-terminus also plays a key role in substrate recognition, though this dimer is missing in plant TSs (**Figure 1.1D**). Many TSs from bacteria, plants, fungi, coral, and sponges also have a conserved arginine pyrophosphate sensor which forms a hydrogen bond with the pyrophosphate moiety of the substrate (**Figure 1.1D**).

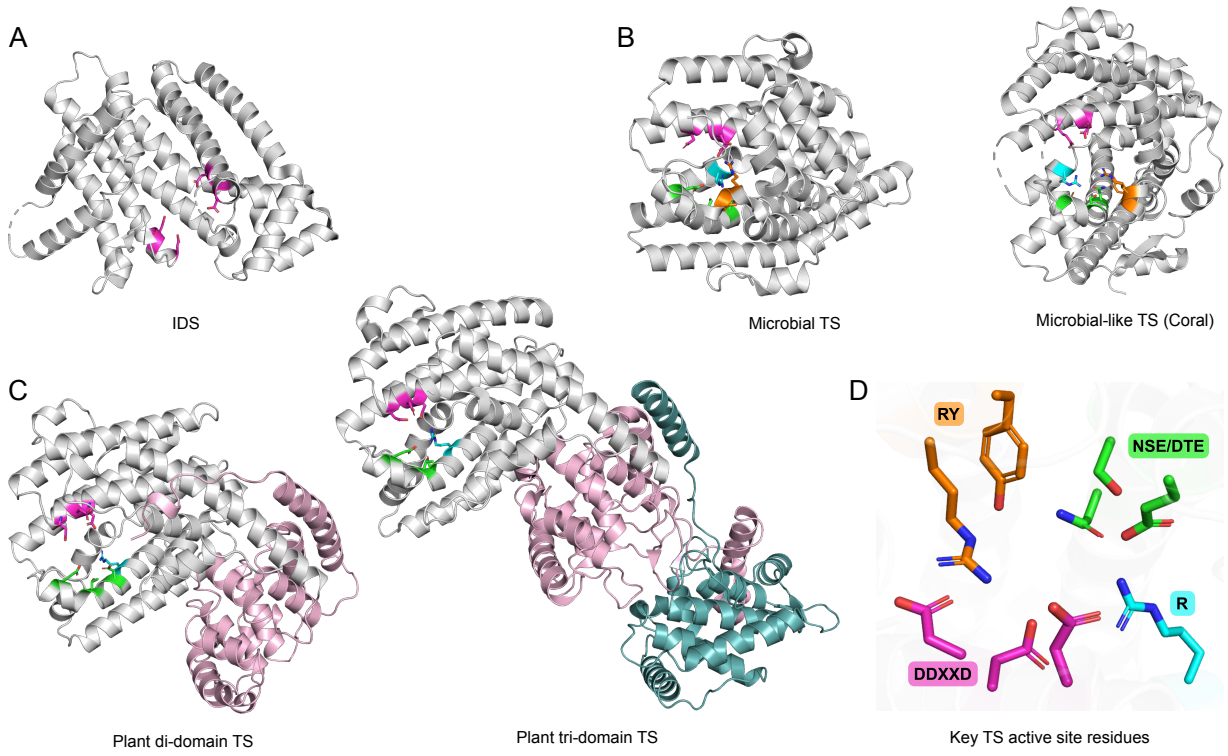


Figure 1.1 Structural overview of TS enzymes **A.** Isopentenyl diphosphate synthase (IDS) (FPP synthase from *Homo sapiens* PDB ID: 5JA0) with first and second aspartate-rich motifs highlighted (FARM and SARM, respectively) **B.** Microbial TS (epi-isozizaene synthase from *Streptomyces coelicolor*; PDB ID: 4LZ0) and microbial-type TS (cembrane A synthase from *Eleutherobia rubra*; PDB ID: 7S5L) **C.** Plant di-domain TS (5-epi-aristolochene synthase from *Nicotiana tabacum*; PDB ID: 5EAS) and plant tri-domain TS (taxadiene synthase from *Taxus brevifolia*; PDB ID: 3P5R) Color code: gray = alpha-domain, light pink = beta-domain, teal blue = gamma-domain. **D.** Expansion of TS active site from the microbial TS, showing DDXXD (neon pink), NSE/DTE (green), R (neon blue), and RY (orange) motifs. Key active site motifs highlighted in each structure according to this color scheme.

1.6 Terpenoid Distribution and Genome Mining

Terpenoids are the most abundant and widely distributed class of secondary metabolites in nature. In the Dictionary of Natural Products[17] terpenoids from class 1 TS reactions are the largest category of characterized isoprenoid compounds (**Figure 1.2A**), representing about 40% of the known metabolites, followed closely by squalenoid/triterpenoid derived compounds (~38%) (**Figure 1.2, Table 1.1**) Plants are exceptional producers of class 1 terpenoids, however, it is worth considering the current data presented in this review may be a result of sampling bias in early natural products discovery (**Figure 1.2B**). It is clear plants dominate characterized terpenoid metabolites. Together, seed plants and nonseed plants represent about 73% of described class 1 derived terpenoids. Outside of plants, the phylum Cnidaria is the second most productive source of class 1 TS terpenoid products (~9%) most of which are characterized from octocorals, closely followed by fungi (~8%). Sponges and red algae are other notable marine sources of terpenoids, with recent discoveries of class 1 TS sequences occurring

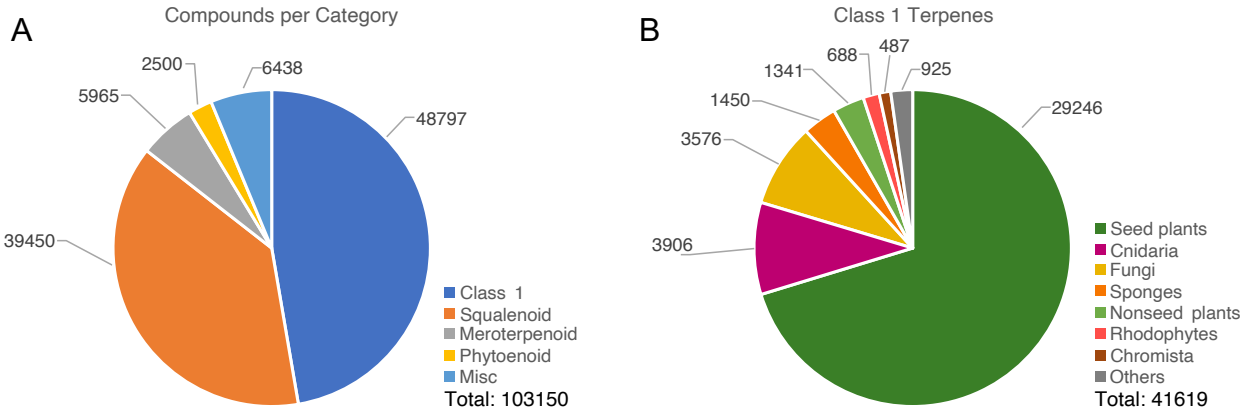


Figure 1.2 Counts of terpenoids by compound category **A**. Counts of class 1 terpenoids by taxonomy **B.** Number of terpenoids per category. Number of terpenoids per organism category. The 'Others' category encompasses 15 organisms including mollusks, arthropods, and bacteria. Full details depicting organism and terpenoid compound type are listed in **Table 1.1**.

in these taxa. Overall, terpenoid natural products are well dispersed across the tree of life and class 1 enzymology, the subject of this review, accounts for a significant percentage of known terpenoid metabolites.

Table 1.1 Raw data associated with **Figure 1.2.** Counts of terpenoids by compound category in the Natural Products Database. Letters after each organism name denote Dictionary of Natural Products Database identifiers.

Organism	Class1	Class2	Meroterpenoid	Triterpene/ squalene based	Carotenoid/ phytoenoid	Misc. (iridoids etc.) or unclear
Archaea (Z.A.)	0	0	0	0	0	105
Bacteria (Z.B.)	227	34	229	139	210	39
Protozoa (Z.D.)	6	6	1	1	12	22
Chlorophytes (Z.E.)	34	12	26	120	99	75
Rhodophytes (Z.F.)	688	43	54	93	29	174
Fungi (Z.G.)	3576	368	1638	2135	156	133
Chromista (Z.H.)	487	2	285	151	55	83
Liverworts (Z.I.)	980	271	28	20	5	28
Mosses (Z.J.)	37	6	1	7	0	0
Lichens (Z.K.)	11	0	0	77	1	1
Fern allies (Z.L.)	27	0	8	134	3	4
Ferns (Z.M.)	297	76	39	384	7	37
Gymnosperms (Z.N.)	2187	298	11	400	36	65
Angiosperms (dic) (Z.Q.)	26113	5550	2737	27573	1484	4622
Angiosperms (mono) (Z.R.)	946	271	52	3392	115	132
Sponges (Z.S.)	1450	67	699	1492	52	506
Annelids, misc invert (Z.T.)	0	0	1	1166	0	0
Cnidaria (Z.U.)	3906	5	74	54	5	69
Mollusks (Z.V.)	311	135	16	125	94	138
Echinoderms (Z.W.	2	0	0	1428	20	3
Arthropods (Z.X.)	278	12	16	154	44	187
Hemichordates (Z.Y.)	13	21	49	58	7	5
Vertebrates (Z.Z.)	38	0	1	347	66	10
Bryozoa	5	1	0	0	0	0

The same or very similar terpenoid scaffolds appear throughout the tree of life, however the responsible TS sequences can be incredibly divergent. For example, (–)-germacrene D (**8**) is a common sesquiterpene biosynthesized and well characterized in many plants. Recently, the biosynthesis of **8** was characterized in sponges.[9] In a pairwise comparison between a representative plant **8**-synthase from oregano (UniProt ID: E2E2N8) and the new sponge **8**-synthase[9], the percent identity of shared amino acid sequence between the two TS sequences is 21%. This is true for many TS sequences

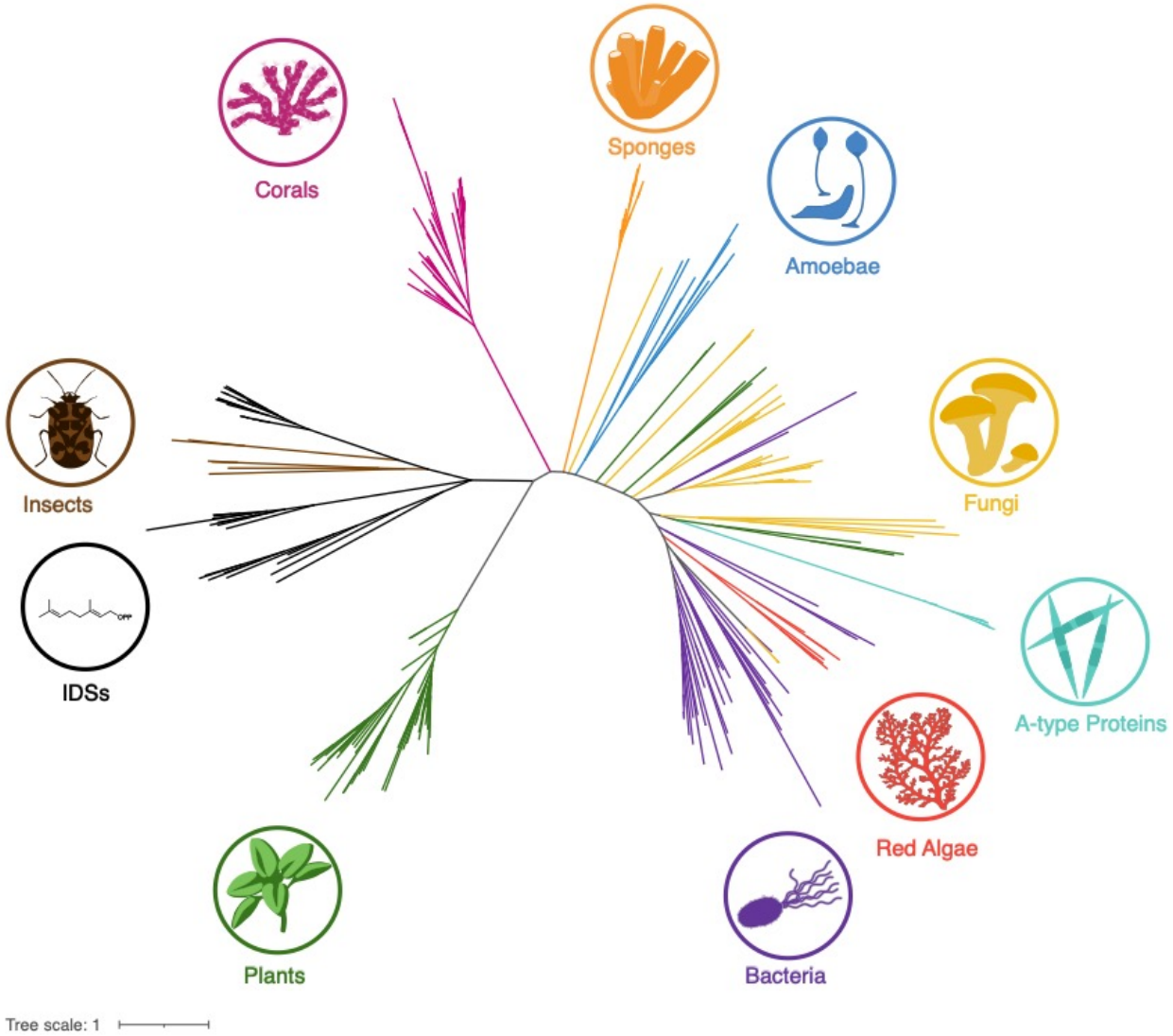


Figure 1.3 Phylogenetic tree of representative class 1 TS sequences. IDS sequences are included as an outgroup.

with shared chemistry, where the amino acid precent sequence similarity can fall into the low teens and represents a huge hurdle is TS discovery.

Despite overall low amino acid sequence similarity between class 1 TSs, the conservation of key active site motifs has made pattern-based genome mining efforts much more accessible. Conserved motifs are particularly amenable to discovery via a profile hidden Markov model (profile-HMM),[18] which is a statistical model and search method that ultimately relies on conserved residues or domains in a biological sequence to classify and annotate new TSs. With the advent of cheaper and more accessible sequencing, as well as huge growth in reference databases, we can readily access sequences to train profile-HMMs and discover TS sequences in previously unstudied organisms. For example, TSs were initially characterized in plants, bacteria, and fungi, and now, with the incredible growth of available sequencing data, in insects, algae, and marine animals. In a phylogenetic analysis (**Figure 1.3**) these enzymes typically clade by taxonomy rather than by function, for example, octocoral terpene cyclases form a distinct monophyletic clade. Microbial-type class 1 TSs are not as clearly separated from one another, forming smaller subclades containing bacteria, fungal, amoeba, and a handful of microbial-like plant terpene cyclases within in a large polyphyletic group, possibly resulting from more recent horizontal gene transfer (HGT) events. Other microbial-type TSs are found in sponges and corals, but display a distinct monophyly, which implies a more ancient HGT event. Finally, recently discovered insect terpene cyclases form a clade within the larger IDS family, suggesting recent neofunctionalization of an IDS to a TS. The expansion and evolution of TSs across Phyla is possible because of the ubiquitous nature of the underlying general biosynthetic machinery for terpenoid

precursor supply. Here, we aim to highlight terpenoid diversity across different branches of the tree of life and discuss the apparent evolutionary success of underlying enzymatic principles. In this review, we use a taxonomy guided approach to highlight terpenoid diversity in secondary metabolism across different branches of the tree of life. Taxa to be discussed are highlighted in the phylogeny of class 1 TSs (**Figure 1.3**).

1.7 Plants

Plants are by far the richest known source of terpenoid natural products, being the source of approximately 70% of all known structures (**Figure 1.2, Table 1.1**).[17] One reason for this molecular abundance could be seen in the high number of plant species (>300,000),[19] but also the extreme biosynthetic diversity within some of these species, some coding for more than 50 TS genes in their genomes.[20] Typical plant-type TSs form a monophyletic group (**Figure 1.3**) and within this clade show more clustering by phylum than catalytic activity, reflective of the whole phylogenetic analysis of TSs. Plant TSs have a modular architecture and are either two-or three-domain enzymes, consisting of an alpha-domain, which is responsible for class 1 TS activity, a beta domain, which shows the fold of archetypical squalene-hopene cyclases harboring an active site for class 2 cyclase activity, and the gamma domain, which plays a structural role but does not participate in catalysis.[16] Class 1 TS from plants can either have an alpha beta[21] or an alpha-beta-gamma structure (**Figure 1.4C**).[22] While didomain enzymes are exclusively class 1 TSs, three-domain enzymes can be bifunctional (class 1 and class 2 activity) or monofunctional, where either the alpha- or beta-domain has lost their respective class 1 or class 2 activity. This fact is illustrated by the production of *ent*-kaurene (**9**), the terpene precursor for different important plant hormones regulating growth and germination, like

the gibberellins in vascular plants (lycophytes, ferns, gymnosperms and angiosperms) and mosses.[23, 24] Compound **9** is synthesized in a two-step reaction (**Figure 1.4A**) in which first a class 2 cyclization takes place to form *ent*-copalyl diphosphate[10] and in the second step a class 1 cyclization converts **10** to *ent*-kaurene (**9**). While all plants seem to produce **9** the enzymes employed are different based in lineage. It is typical for vascular plants to employ two discrete, monofunctional three domain enzymes in which the **10**-synthase has a catalytically inactive alpha-domain and the **9**-synthase has an inactive beta-domain (**Figure 1.4B**).[25] In mosses, which diverged relatively early in the evolutionary history of plants, one enzyme with an active beta and alpha-domain is responsible for the direct conversion of GGPP to **9**.[26] The moss enzyme could be seen

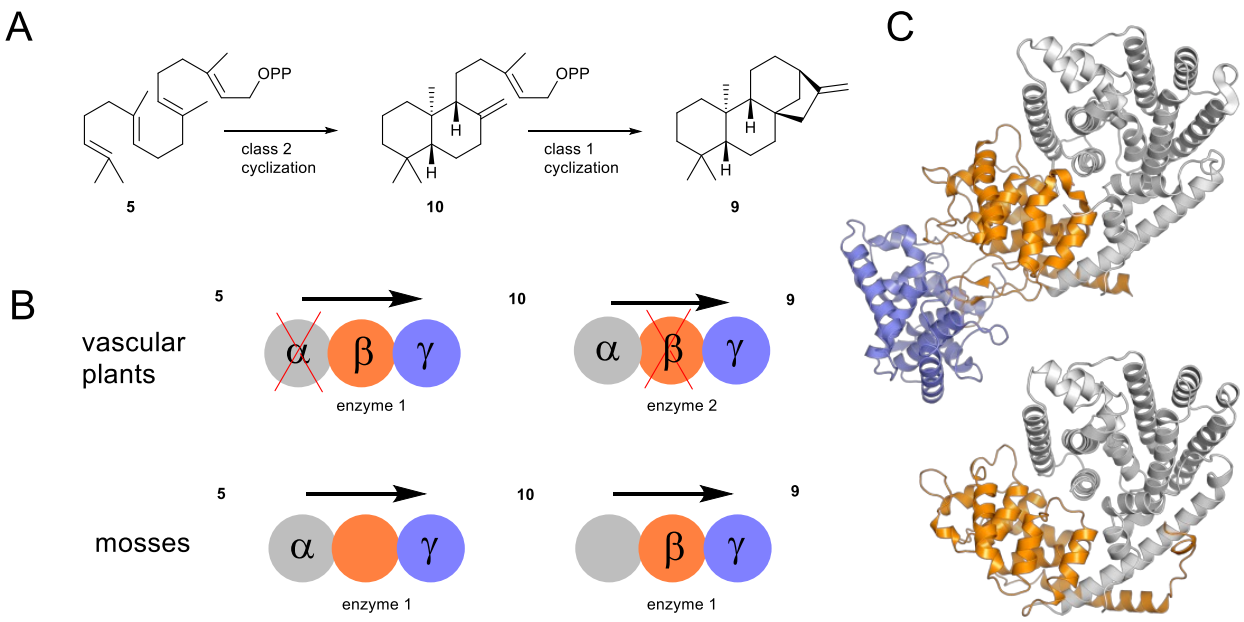


Figure 1.4 Function and architecture of plant TS enzymes. **A.** The two-step reaction of GGPP (**5**) to *ent*-kaurene (**9**), **B.** Different biosynthetic strategies for **9** production: vascular plants employ two monofunctional enzymes three domain enzymes that only contain one active domain each, while mosses use one bifunctional TS with two catalytically active domains that act consecutively. **C.** representative crystal structures of three- and di-domain plant TS, taxadiene synthase (up, PDB code 3p5r), epi-aristolochene synthase (down, PDB code: 5eas). Color code: gray = alpha-domain, orange = beta-domain, blue = gamma-domain.

as a preservation of the ancestral state of an *ent*-kaurene synthase, while vascular plants, which diverged later, evolved to distinct enzymes, most likely by duplication and loss of function of the unused domain respectively.[3] The characterization of several three-domain TSs from non-vascular plants (mosses, liverworts and hornworts) followed by a phylogenetic analysis of the whole plant TS clade showed two early gene duplication events in the last common plant ancestor resulting in three principle clades.[27] One of these contains three domain enzymes that are bifunctional *ent*-kaurene (**9**) synthases for non-vascular plants (mosses, hornworts, liverworts) and monofunctional class 2 *ent*-copalyl diphosphate (**10**) synthases in vascular plants, pointing to a bifunctional *ent*-aurene synthase as the ancestral plant TS.[27]

Duplication-neofunctionalization events similar to the diversification of *ent*-kaurene production have given rise to a large number of TSs that are involved in secondary metabolism and produce the vast number of structurally distinct terpene carbon scaffolds known from plants. One such example is represented in the momilactones from *Oryza sativa* (28). Momilactones are secreted by *O. sativa* into the soil, where they function as allelochemicals exhibiting phytotoxicity (29, 30). The carbon scaffold is assembled by a sequence like for the kaurene phytohormones, where the class 2 enzyme OsCPS4 produces *syn*-copalyl diphosphate (**14**),[31] which is converted by the class 1 cyclase OsKSL4 to form *syn*-pimaradiene (**15**).[32] *Syn*-pimaradiene is further functionalized by four cytochrome p450 monooxygenases (CYP) and one short chain dehydrogenase (SDH) resulting in the main phytotoxic compound momilactone B (**16**) (**Figure 1.5A**).[33] A part of the pathway is colocalized on *O. sativa* chromosome 4 as a biosynthetic gene cluster,[34]

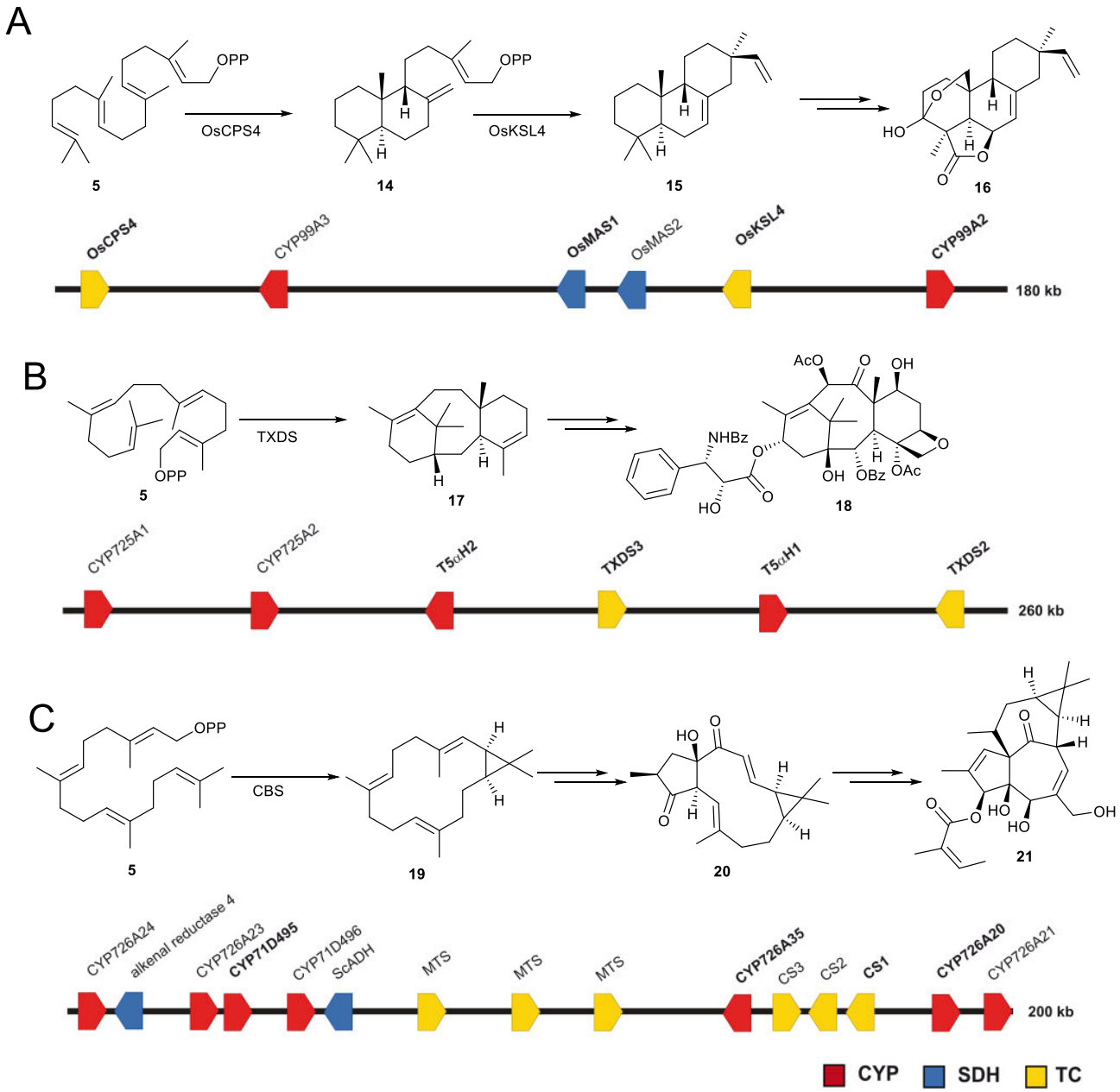


Figure 1.5 Diterpene-scaffold generation by tri-domain TSs and their connection with oxidized diterpenoids. **A.** syn-pimaradiene (**15**), the terpene precursor for momilactone B (**16**) is produced in two steps by a monofunctional class 2 and a monofunctional class 1 TS. All pathway enzymes are known, four of the seven genes are co-localized in a gene cluster in *O. sativa*. **B.** Taxadiene (**17**) is produced by taxadiene synthase, a monofunctional class 1 TS. Taxadiene synthase is colocalized with taxadiene-5-a-hydroxylase, a characterized CYP that oxidized **17**, on the genome of *Taxus chinensis*. **C.** A monofunctional class 1 TS converts GGPP to casbene (**19**), the precursor to a plethora of structurally diverse diterpenoids like ingenol (**21**). Casbene synthase is co-localized with enzymes responsible for the production of key intermediate jolkinol C (**20**) and further, uncharacterized CYP, SDH and TS genes on the genome of *Jatropha curcas*. Genes with names printed in bold have been biochemically verified.

Calohypnum plumiforme also produces momilactones via **15** and encodes for a gene cluster responsible for their production.[35] The clusters from *C. plumiforme* and *O. sativa* are not homologous and the *syn*-pimaradiene synthase in *C. plumiforme* is a bifunctional class 2 / class 1 enzyme,[36] which shows the same difference as *ent*-kaurene (**9**) production in mosses and vascular plants.

Monofunctional three-domain TSs also utilize linear precursors like GGPP to produce terpene carbon scaffold. Two examples are taxadiene (**17**) synthase from *Taxus spp*. and casbene (**19**) synthase, which is widespread in Euphorbiaceae.[37, 38]

Taxadiene (**17**) serves as the carbon scaffold in the taxane diterpenes including the anti- cancer medicine Taxol.[38] While not all biochemical conversions from taxadiene (**17**) to taxol (**18**) are known to this date, many enzymes have been identified by transcriptomic approaches, including the second committed step in taxol biosynthesis the hydroxylation of taxadiene by a taxadiene 5a-hydroxylase (T5aH).[39] Genome sequencing of two different species of *Taxus* showed that taxadiene synthase and T5aH are colocalized within larger gene clusters containing uncharacterized CYP genes (**Figure 1.5B**).[10, 40]

Casbene (**19**) is the precursor to a wide array of structurally complex diterpenoids in the plant family of Euphorbiaceae.[41] The biosynthetic pathways to higher functionalized diterpenoids like ingenol mebutate (**21**) are unknown, but the enzymes for the generation of the bicyclic jolkinol C (**20**) are characterized.[42] Compound **19** is oxidized by three CYPs and cyclized by a short chain dehydrogenase (SDH).[42] These genes were shown to be co-localized in larger gene clusters in several species from the Euphorbiaceae group including *Ricinus communis*,[43] *Jatropha curcas*,[44] and *Euphorbia peplus*.[45] The last study

identified two gene clusters containing casbene synthase and a multitude of CYPs and SDHs containing a total of 68 gene candidates. Gene silencing experiments determined jolkinol C as a branching point between the co-occurring groups of bicyclic jatrophane diterpenoids and tetracyclic ingenane diterpenoids like ingenol.[45]

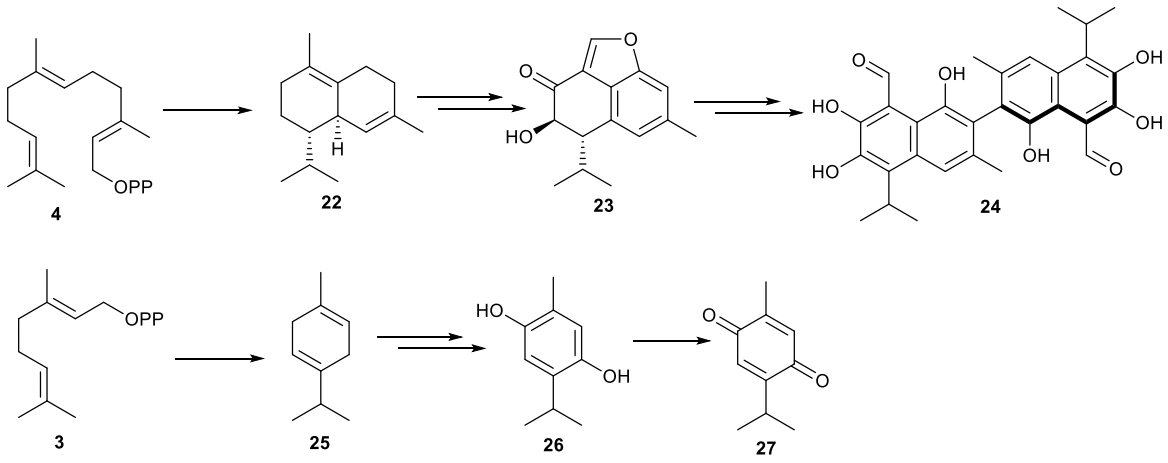


Figure 1.6 Sesqui- and monoterpenoids and their connected hydrocarbons produced by sesqui- and mono-TS.

All discussed examples so far are diterpenoids, but the evolution of the putative ancestral bifunctional diterpene synthase has also brought forth a huge repertoire of monofunctional sesqui- and mono-TS enzymes, that use farnesyl diphosphate (FPP, **4**) and geranyl diphosphate (GPP, **3**) as substrates respectively.[3] The dimeric sesquiterpenoid gossypol (**24**) is a defense compound produced by cotton plants (*Gossypium* spp.) in the presence of phytopathogenic fungi.[46] The first biosynthetic step is carried out by a didomain sesquiterpene synthase that converts **4** to delta cadinene (**22**).[47] Several biosynthetic steps towards the intermediate **23** are characterized and carried out by CYPs and SDHs.[48] The final step is facilitated by a dirigent protein, that orients two achiral monomeric compounds so the biaryl coupling produces atroposelectively (+)-**24**.[48]

Thymohydroquinone is a characteristic flavor compound in thyme and oregano and produced via the monoterpene gamma-terpineol (**25**) which is oxidized to **27** in three steps carried out by two CYP and one SDH in Thymus vulgaris.[49] Due to their hydrophobicity and relatively small molecular mass **27** and pathway intermediates are volatile, some of them contributing to the characteristic odor of their producing plants.[49] In fact, a big portion of fragrant plant volatiles consists of mono and sesquiterpenes, often as unmodified hydrocarbons produced by TS enzymes.[50] Volatile terpenes can act as direct or as indirect defense. The sesquiterpene hydrocarbon was shown to act as both, direct and indirect defense. Infestation of the maize by the larvae of the beetle *Diabrotica virgifera* leads to the belowground release of beta-caryophyllene through the roots, which attracts the entomopathogenic nematode *Heterorhabditis megidis*.[51] In *Arabidopsis thaliana* infected by *Pseudomonas syringae* beta-caryophyllene production is strongly increased and inhibits microbial growth.[52] Volatiles emitted by floral parts also act as attractants to pollinators.[53] *Nicotiana suaveolens* is a night-blooming plant mainly pollinated by hawkmoths.[54] Different studies found the concerted higher production of seven monoterpenes at night, the main compound being 1,8-cineole.[55, 56] Investigation of the origins of these terpenes showed a single monoterpene synthase to be responsible for the production of all of these compounds, highlighting the ability of class 1 TSs to produce a set of structurally different products, which simplifies regulation of a complex bouquet of volatiles together being responsible for a scent signal.[54]

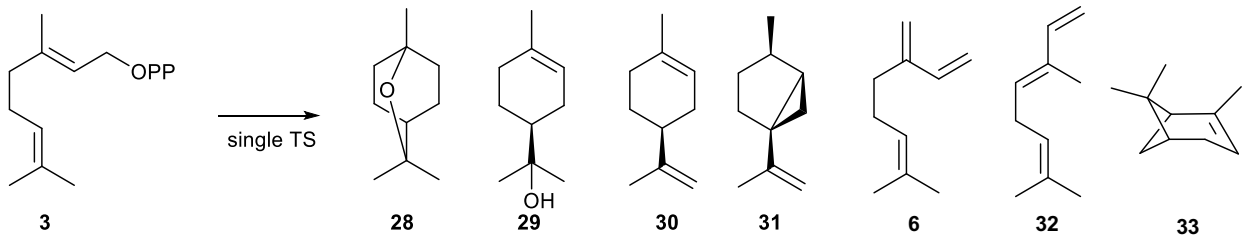


Figure 1.7 Several monoterpenes produced by a *Nicotiana suaveolens* mono-TS.

Besides the discussed two- and three-domain TSs, which were thought to exclusively responsible for all terpene TS activity in plants, several lineages of single-domain enzymes were discovered recently in the club moss *Selaginella moellendorffii*, which encodes 48 of these enzymes in its genome next to 18 classic two- and three-domain plant TSs.[57] Since these monodomain TSs are primarily known form microbes they were called microbial-type TSs and a later study found them to be abundant in almost all non-seed plants, but in no representative of a seed plant.[58]

Phylogenetic analyses revealed clustering with bacterial and fungal TS sequences, which, in addition to their domain structure points towards horizontal gene transfer from a microbe to non-seed plants. Since these microbial-type plant TSs are not forming a monophyletic clade, but rather occur as four monophyletic groups, evidence suggest multiple horizontal gene transfer events.[59] The ecological role of these synthases is not known, but all characterized members produce mono- and sesquiterpenes, often complex mixtures, which suggests involvement in volatile production.[58]

1.8 Fungi

Recent advances in fungal genome sequencing, metabolomics, and the development of synthetic biology approaches to characterize fungal natural products have accelerated our understanding of natural product biosynthesis in this kingdom.

Fungal terpenoids make up almost 8% of known terpenoid natural products, and 32% of non-plant derived terpenoid natural products, making them the third largest contributor to this class of compounds (**Figure 1.2, Table 1.1**).[17] Terpenoid natural products in fungi are primarily known from the two major fungal divisions. Ascomycota, or filamentous fungi, which account for ~60% of described species, and Basidiomycota, including mushroom forming fungi, which account for ~30% of described species.[60] Although there are many differences between the terpenomes of the two major fungal Phyla, class 1 fungal TSs have been found to be broadly distributed in both Phyla.[60] A comparative analysis of the two Phyla found terpenoids to be the major class of natural products in Basidiomycota, with approximately four times more TS sequences identified per genome than in publicly available Ascomycota genomes.[60] Identified TS sequences consist of typical class 1 architecture, consisting of an alpha-domain subunits responsible for class 1 TS activity, as well as unique chimeric fungal bifunctional TSs, which harbor a prenyl transferase (PT) domain and a class 1 TS domain, forming a bifunctional alpha-alpha-domain assembly.[20] Fungal terpene synthases are dispersed throughout the microbial-type TS clade (**Figure 1.3**) forming a polyphyletic group with clustering according to both function and taxonomy. Notably, many fungal TSs are associated with biosynthetic gene clusters which facilitates genome mining and pathway engineering efforts.[62] Fungal terpenoids derive from the mevalonate pathway and are typically sesquiterpenes (C15), diterpenes (C20), sesterterpenes (C25), and triterpenes (C30), with no fungal monoterpene synthases known to date.[60] Fungal terpenoids are functionally diverse and play numerous ecological roles ranging from defense and attraction, to mediating

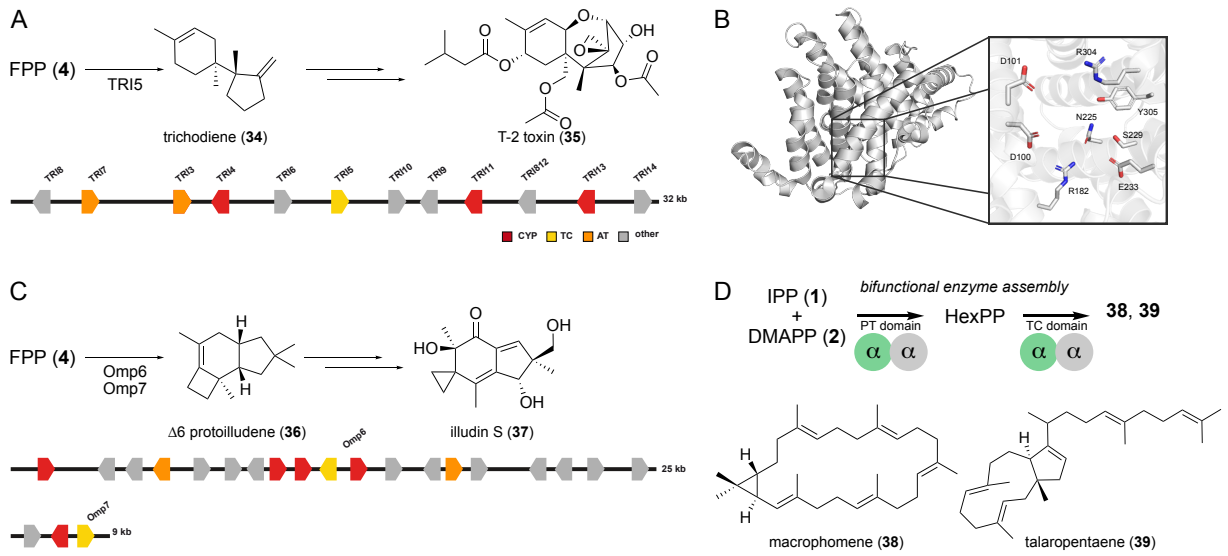


Figure 1.8 Fungal TS scaffold generation. **A.** Trichodiene (**34**), the terpene precursor for T-2 toxin (**35**), and the producing gene cluster from *Fusarium sporotrichioides*. **B.** Monomer of trichodiene synthase and highlighted active site residues. **C**. Δ6 protoilludene (**36**), the terpene precursor for the illuden family of sesquiterpenes, including the antitumor agent, illudin S (**37**). **D.** Bifunctional triterpenoid synthase with, macrophomene (**38**) and talaropentaene (**39**).

complex interkingdom chemical interactions, such as host plant and endophytic fungi signaling.[63] Additionally, many fungal terpenoids are highly bioactive with anticancer, antimicrobial, and antiviral activities. For example, the illudin family of sesquiterpenoids are effective against several tumor cell-lines, including metastatic prostate cancers[64] and the highly toxic trichothecene mycotoxins have a significant impact on global cereal crops and are known to contaminate livestock feed.[65]

The first characterized class 1 fungal TS comes from the Ascomycota plant pathogen *Fusarium sporotrichioides* and was found to catalyze the formation of trichodiene (**34**), which is the first committed step towards the trichothecene mycotoxins (**Figure 1.8A**). Detailed mechanistic characterization via site directed mutagenesis developed a cyclization cascade proposal which was later rationalized by a crystal structure[66] resulting in a structure-based model of the **34**-synthase reaction. One of the

most interesting observations from this early work is the startling conservation of the terpene synthase fold, which, much like in other organisms, consists of an alpha helical bundle with relatively low sequence similarity to other fungal TS sequences. Fungal TSs were found to maintain the two conserved aspartate-rich motifs, DDXXD/E and NSE/DTE (**Figure 1.8B**).[66] Trichodiene is then modified in *Fusarium* strains into the highly toxic trichothecenes, which share a common tetracyclic ring and includes T-2 toxin (**35**) produced by *F. langsethiae*, *F. poae*, and *F. sporotrichioides* (gene cluster AF359360.3).[67, 68]

Some of the most well-known terpenoid natural products in Basidiomycota belong to the illudin family of sesquiterpenoids, which derive from the ∆6 protoilludane (**36**) scaffold (**Figure 1.8C**).[69] Early isolation and structural characterization work identified *Omphalotus olearius*, or the Jack O'Latern mushroom, as a major producer of bioactive illudins.[64] Follow-up studies report the draft genome *of O. olearius* strain VT-653.13 and identified 11 TS sequences (Omp1-10, with 5a and 5b).[70] Biochemical characterization of each TS found Omp6 and 7 (62% AA % identity) to be highly specific **36**-synthases. Both TSs are located within biosynthetic gene clusters, with Omp6 located within a 25 kb, 18 gene cluster that is proposed to produce illudin compounds (**Figure 1.8C**). Omp7, is located within a smaller cluster, with only one P450 and a FAD binding protein. Using newly identified TS sequences as genetic hooks, genome mining of 40 publicly available Basidiomycota genomes revealed 531 putative TS sequences. Phylogenetic analysis highlights five putative functional clades, which were supported with biochemical validation of Fompi1 from *Fomitopsis pinicola*. This enzyme was validated as an α-

cuprenene sesquiterpene synthase which matched the predicted product according to the growing Basidiomycota TS phylogeny.

Fungal chimeric TSs were initially characterized in fusicoccin biosynthesis. The tricyclic hydrocarbon precursor for the fusicoccins is biosynthesized by a fusicoccadiene synthase, which contains both prenyltransferase and terpene cyclase activities. Notably, this enzyme is encoded as a part of a larger gene cluster.[71] Bifunctional TSs are also common in sestertepene biosynthesis in fungi, where catalytically independent domains work in tandem to generate C25 sestertepene products.[72] Recently, a new family of fungal chimeric class 1 triterpene synthases were discovered and functionally characterized.[73] Typically, triterpenoids are biosynthesized via the dimerization of FPP into squalene by class 2 TSs known as squalene synthases. Subsequent cyclizations are catalyzed by triterpene synthases, such as squalene hopene cyclases (SHC), resulting in products like pentacyclic hopene, the precursor of hopanoids. A recent study by Tao and coauthors, reports the discovery and biochemical characterization of two non-squalene-dependent triterpene synthases TvTS and MpMS from *Talaromyces verruculosus* and *Macrophomina phaseolina*, respectively. Much like chimeric fungal sestertepene synthases, this new family of class 1 triterpene synthases harbor both PT and TS domains (**Figure 1.8D**). Each PT domain was found to first convert **1** and **2** into hexaprenyl diphosphate or HexPP (C30), which is subsequently cyclized by the TS domain into the final triterpenoid product. Expression of full-length TvTS and MpMS in *Saccharomyces cerevisiae* resulted in two different triterpenoid products, talaropentaene (**38**) and macrophomene (**39**), respectively (**Figure 1.8D**). Notably, the 22-membered ring in **39** represents the largest macrocycle terpenoid known to-date and only structurally mirrored

by the 22-membered rings in nostocyclophanes, a class of oxidatively dimerized polyketides from cyanobacteria.[74] Additional protein structure-supported genome mining using AlphaFold2-predicted models and the crystal structure of TvTS-TC identified six new sequences with putative triterpenoid activity. Biochemical characterization of putative triterpenoid bifunctional synthases resulted in the discovery of CgCS from the host *Colletotrichum gloeosporioides*. CgCS is structurally similar to TvTS and makes a related product, colleterpenol. Overall, this work identifies a new enzymatic mechanism for the biosynthesis of triterpenenoids and expands the product boundaries for class 1 TSs.

1.9 Bacteria

While the group has expanded in recent years, bacterial terpenoids still make up less than 0.5% of all terpenoid natural products and approximately 2% of all non-plant terpenoid natural products (**Figure 1.2, Table 1.1**).[6] While many bacteria encode stand-alone terpene synthases, a growing number of bacterial terpene synthases are situated within biosynthetic gene clusters that encode functionalized molecules like albaflavenone,[75] pentalenolactone,[76] cyclooctatin (**41**),[77] and (+)-kolavenol.[78] While bacterial terpenoids like albaflavenone and *ent*-kaurene (**9**)[79] are the result of class 1 and class 2 cyclizations, these TSs are usually separate monofunctional enzymes which are clustered together, unlike the bifunctional alpha-beta-gamma domain plant TSs that produce similar terpenoids. Bacterial terpenoids, many of which are released as volatiles, are important ecological mediators of microbe-microbe interactions. These molecules play a particularly large role in soil ecosystems, where volatile terpenes have been shown to mediate interactions between bacteria, plants, fungi, and other microbes across long

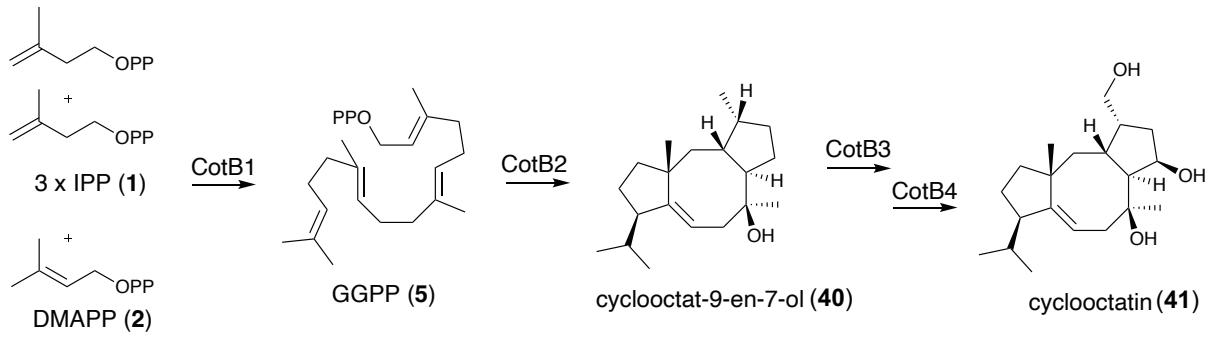


Figure 1.9 Cyclooctatin biosynthesis. GGPP synthase CotB1 catalyzes the production of GGPP (**5**) from IPP (**1**) and DMAPP (**2**). Terpene synthase CotB2 produces cyclooctat-9-en-7-ol (**40**), which is converted to cyclooctatin (**41**) by two CYP450s, CotB3 and CotB4.

distances.[80] Bacterial terpenes such as albaflavenone and β-pinene are known to inhibit other bacteria and fungi,[81-83] while bacterial D-cadinene and β-linalool can stimulate motility in protists.[83]

Several bacterial terpene synthases have been extensively characterized and a thorough summary of this work can be found elsewhere.[84] Most of the TSs characterized from bacteria are sesquiterpene synthases, though a handful of monoterpene and diterpene synthases have also been characterized. Perhaps the best studied bacterial TS is the diterpene synthase CotB2, which forms the characteristic 5-8-5 fused ring system of the potent lysophospholipase inhibitor cyclooctatin (**41**) (**Figure 1.9**). The cyclooctatin (**41**) gene cluster consists of a GGPP synthase (CotB1), the CotB2 terpene synthase, and two CYP450s (CotB3 and CotB4). CotB2 has been the subject of numerous structural and mechanistic studies, which has yielded unprecedented insight into the inner workings of this enzyme.[85-87] Several crystal structures of CotB2 have revealed that following the binding of GGPP and 3 Mg^{2+} ions, the C terminus “lid” of the enzyme closes on the active site, transitioning the enzyme from an open to a closed state. Several mutagenesis studies have shown that single point mutations to active site residues change the product

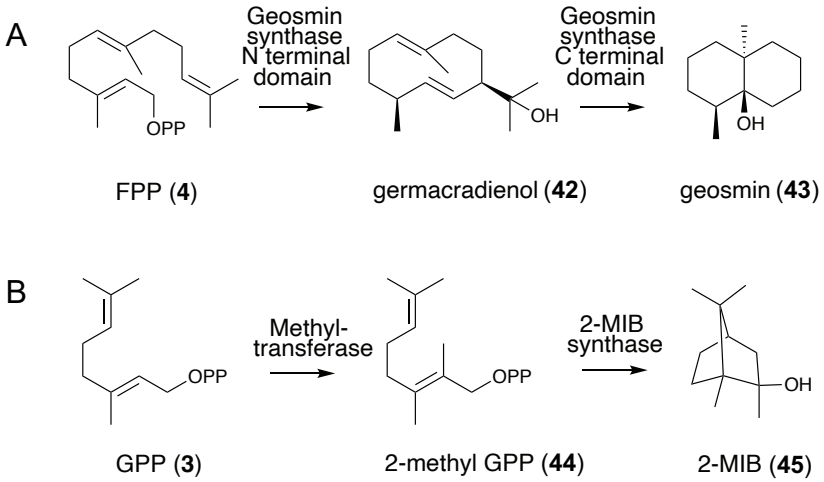


Figure 1.10 A. Geosmin (**44**) and **B.** 2-methylisoborneol (2-MIB) (**45**) biosynthesis.

scope drastically from fusicoccanes to dolabellanes or cembranoids. This makes CotB2 a promising candidate for biocatalysis since small mutations provide access to a wide range of scaffolds with antibiotic,[88] anticancer,[89] and antiparasitic activity.[90, 91]

Geosmin (**43**) and the related compound 2-methylisoborneol (2-MIB) (**45**) (**Figure 1.10**) are widespread bacterial terpenoids responsible for the earthy odor of soil. While humans and many animals can sense these compounds,[92] their ecological role is still poorly understood.[93] Compounds **43** and **45** were recently reported as predation deterrents for *Caenorhabditis elegans*.[94] Compound **43** also has a clear effect on certain insects including fruit flies, in which **43** impacts feeding, oviposition, and flight,[95, 96] and mosquitos, in which **43** acts as an attractant for oviposition sites.[97] Becher *et al.* also showed that **43** and **45** mediate symbiosis between *Streptomyces* species and springtails, by acting as feeding attractants for springtails, which facilitates Streptomyces spore dispersal through fecal pellets. It was recently reported that relatively low levels (50-5000ng/L) of **43** disrupted several processes involved in embryonic development in zebrafish.[98]

Geosmin (**43**) is biosynthesized by an unusual didomain enzyme. The N terminal domain of geosmin synthase converts FPP to germacradienol (**42**), while the C terminal domain converts **42** to geosmin (**43**). Both the N terminal and C terminal domains contain the usual type I TS active site moieties including the aspartate-rich DDXXD, the NSE triad, and the RY dimer. In some cases, the N terminal domain features an additional NSE triad approximately 38 amino acids downstream of the canonical NSE triad, but mutagenesis of the downstream NSE motif in *Streptomyces coelicolor* geosmin synthase did not significantly affect product composition or enzyme kinetics, suggesting this additional NSE may not play a role in catalysis.[99]

2-MIB (**45**) is the product of a two gene biosynthetic gene cluster consisting of a SAM-dependent methyl transferase and a monoterpene synthase. The methyl transferase starts by methylating GPP (**3**) at the C2 position to form (*E*)-2-methyl GPP (**44**), the substrate of the monoterpene synthase. Similar to many other monoterpene synthases, which isomerize GPP to linalyl diphosphate (LPP), 2-MIB synthase isomerizes **44** to 2-methyl LPP in the first step of catalysis, before catalyzing the rest of the cyclization to **45**. Many **45**-synthases found across several bacterial species are unusually large bacterial terpene synthases, with approximately 400-500 amino acid residues. Most 2-MIB synthases features a proline-rich (>25% proline) 13 kDa N-terminal domain which is not seen in other terpene synthases; however, this domain is entirely disordered and its function is unknown.[100] Crystal structures of 2MIB synthase from *Streptomyces coelicolor* reveal an “incompletely closed” active site and an unusual proline-rich domain of unknown function, which is only found in **45**-synthase homologues, but not other terpene synthases.[100]

1.10 Algae

Algae, including Chlorophytes (green algae), Rhodophytes (red algae), and for the purpose of this review, Chromista (brown algae, diatoms, etc), contribute ~2% of total known terpenoids and ~19% of non-plant derived terpenoid natural products (**Table 1.1**). Of the known terpenoids from algae, 48% are anticipated to be produced by class 1 TS enzymology (**Figure 1.2B, Table 1.1**). Notable features include oxidative scaffold tailoring with oxygen and halogens. The incorporation of halogens, like bromine and chlorine, is a hallmark of marine chemistry and rarely seen in terrestrial compounds. Red macroalgae, or seaweeds, are particularly gifted producers of halogenated terpenoids, or haloterpenoids. Algal haloterpenoids are hypothesized to play several ecological roles, such as, deterring herbivory[101] and preventing the formation of biofilms and biofouling via quorum sensing disruption.[102]

Some red algae are known to accumulate haloterpenoids in specialized refractile inclusion bodies, or *corps en cerise.* Additionally, there is evidence for the subcellular compartmentalization of terpenoid biosynthetic machinery in these unique cell structures. Paradas *et al.* describe a 'mevalonosome', which is a vacuolar organelle containing enzymes involved in the mevalonate pathway in *Plocamium brasiliense*.[103] Cytochemical analysis supports the presence of active 3-hydroxy-3-methylglutaryl-CoA synthase (HMGS) in the *P. brasiliense* mevalonosome, a key enzyme of isoprenoid biosynthesis in the mevalonate pathway. Further work to establish the subcellular localization of other terpenoid enzymes is needed.

The first biochemically characterized macroalgal sesquiterpene synthase comes from the red alga, *Laurencia pacifica*.[104] Kersten *et al.* identified and functionally characterized a bourbonane-producing sesquiterpene synthase, LphTS-A, using transcriptomics and heterologous expression in yeast (**Figure 1.11A**). LphTS-A is one of three terpene synthases identified in the transcriptome of *L. pacifica*. To resolve the absolute stereochemistry of prespatane (**46**), the major bourbonane sesquiterpene product of LphTS-A, Kersten *et al.* used a microgram-scale nuclear magnetic resonance-coupled crystalline sponge X-ray diffraction analysis. Three more TS transcripts were identified from *Laurencia dendroidea* from a publicly available dataset in the NCBI Sequence Read Archive (SRA). Phylogenetic analysis revealed a well-supported monophyletic clade comprising the six red algal sequences, separate from sesquiterpene synthases from other domains of life. Interestingly, this clade is nested between bacterial

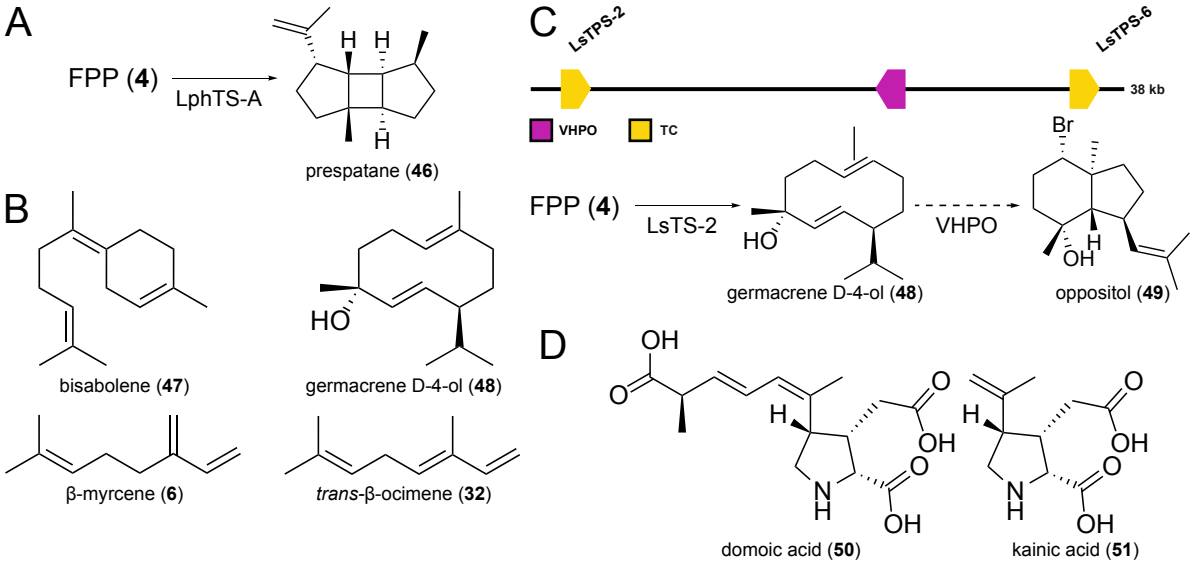


Figure 1.11 Algal TS scaffold generation. **A.** Prespatane (**46**), the product of LphTS-A **B.** Key haloterpenoid scaffolds, bisabolene (**47**), germacrene D-4-ol (**48**), β-myrcene (**6**), trans-β-ocimene (**32**), (**51**) **C**. Proposed pathway to oppositol (**49**) the putative product of the haloterpenoid associated gene cluster from *Laurencia subopposita*. **D.** Structures of domoic acid (**50**) and kainic acid (**51**)

and fungal TS sequences, suggesting a recent horizontal gene transfer event, possibly from a microbial source.

Systematic sequence analysis of publicly available red algal genomes and transcriptomes by Wei *et al.* further expands the small, but growing number of algal TSs.[7] In addition to the six already known sequences from *Laurencia* spp., three new TSs were identified from *Erythrolobus australicus* and *Porphyridium purpureum*. No typical plant TS sequences were found in any of the algal genomes analyzed. In-depth genomic analysis of *P. purpureum* validated red algal TS sequences are not from microbial contamination and are part of the nuclear genome. Upstream and downstream neighboring genes relative to the TS sequence were determined to be most closely related to eukaryotic genes. Notably, the assembled region containing the TS sequence had high contiguous coverage, which provides confidence in assembly accuracy and assignment of neighboring genes.

Work by Steele *et al.* reports ten new red algal TSs from the key haloterpenoid producing red algae *Portieria hornemannii*, *Plocamium pacificum*, and *Laurencia subopposita* (*in preparation*). This work supports the existence of microbial-type class 1 TSs, through whole genome sequencing and phylogenetic analysis. In all four species and the authors demonstrate a clear delineation between algal genes and those from associated microbes. *In vitro* reconstitution and biochemical characterization revealed the selective production of key haloterpenoid precursors, including the well-known plant terpenes β-myrcene (**6**), trans-β-ocimene (**32**), bisabolene (**47**), and germacrene D-4-ol (**48**) described for the first time from algal enzymes (**Figure 1.11B**). As sequencing data of algae improves, we are rapidly expanding our understanding of gene clustering in red

seaweeds. This is exemplified by the first known haloterpenoid associated gene cluster in *L. subopposita*, where the sequence lsTPS-2 was found to be co-clustered with a vanadium-dependent haloperoxidases, the predicted enzyme responsible for the halogenation of algal haloterpenoids (**Figure 1.11C**).[105] Haloperoxidases have been isolated from all classes of marine algae and are implicated in haloterpenoid biosynthesis.[105] The expansion of the red algal TS clade by Steele and Brunson *et al.* supports our understanding of terpenoid biosynthesis in red algae, however further studies of downstream red algal tailoring reactions and brown macroalgae are still required to fill gaps in algal terpenoid chemistry.

Kainoids are a family of potent marine neurotoxins produced by a variety of marine and terrestrial organisms. Oceanic kainoid production is primarily associated with the harmful algal bloom–forming diatom genus *Pseudonitzschia*, however, domoic acid (DA) (**50**) and kainic acid (KA) (**51**) are produced by both macro- and micro- algae. Compounds **50** and **51** both share a tri-substituted pyrrolidine ring, the hallmark of the kainoid family, and differ at the C4 substituent in prenyl chain length (**Figure 1.11D**). Early stable isotope experiments suggested marine kainoids are composed of L-glutamate and either DMAPP (**2**) or GPP (**3**). Work by Brunson *et al.* identified the four-gene cluster responsible for **50** production using transcriptome and genome sequencing.[106] Initially, one of the gene candidates, *dabA*, was annotated as a terpene cyclase, and structural prediction suggested a terpene cyclase-like fold. However, DabA was found to repurpose TS enzymology to catalyze an intermolecular N-prenylation reaction of a primary amine, rather than the canonical intramolecular cyclization reaction. Follow-up work by Chekan *et al.* identified homologous biosynthetic genes using whole genome sequencing of the

red seaweeds *Digenea simplex* and *Palmaria palmata*.[107] Here they identified the DabA homologs, dsKabA and ppKabA, respectively, which also catalyzes the N-prenylation of L-glutamate but accepts DMAPP (**2**) rather than GPP (**3**), to form the lower isoprene homolog of DA (**50**), KA (**51**). A later study by Chekan *et al.* reports the 2.1 Å crystal structure of DabA.[108] Like most terpene cyclases, DabA consists of a mostly alpha-helical core, however DabA has several key differences from canonical terpene cyclases. First, DabA contains N- and C- terminal extensions and an internal insertion, which connect to the main alpha-helical core via a β-hairpin extension. Next, DabA has an elongated hydrophobic tunnel which accommodates the unique binding of the pyrophosphate substrate. Similarly, DabA has distinct active site binding modifications that remodel the canonical magnesium (Mg^{2+})-binding motif found in terpene cyclases and enable the intermolecular prenylation reaction, where the canonical DDXXD motif is now NDXXA and NSE is NTE (**Figure 1.12**). The authors also identify key residues that dictate prenyl group chain length specificity, and thus, catalytically differentiates DabA and KabA.

Both DabA and KabA have remarkably low amino acid sequence similarity to any publicly available protein sequence and thus serve as a key genetic hook to query for kainoid biosynthetic genes. This is demonstrated in work by Steele and Brunson *et al.*

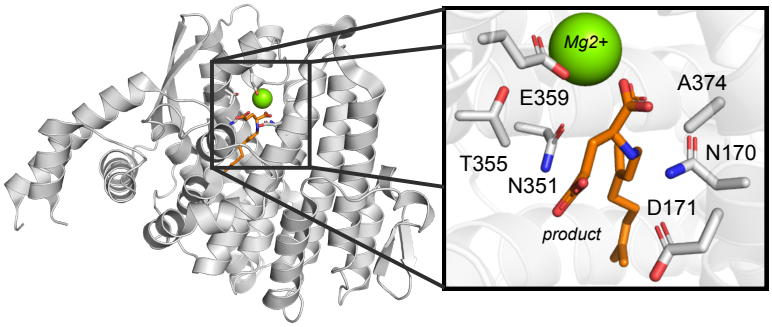


Figure 1.12 Crystal structure of DabA (PDB ID:6VL1). Panel highlights modified active site residues DDXXD is NDXXA and NSE is NTE. Orange is the DabA product *N*-geranyl-L-glutamate.

where the authors identify the *rad* biosynthetic gene cluster in the red macroalgae *Chondria amrata* and biochemically characterize the TS-like enzyme RadA, which despite having higher amino acid sequence similarity to the red algal kainic acid biosynthesis enzyme KabA, catalytically behaves more like the homologous diatom enzyme DabA.[109] The consistent monophyly of the A proteins from these distantly related taxa suggests a possible horizontal gene transfer event. Given the limited distribution of kainoid biosynthetic enzymes in modern diatoms and red macroalgae, as well as the lack of clear homology of the glutamate N-prenyltransferase (RadA, KabA, and DabA) to known algal

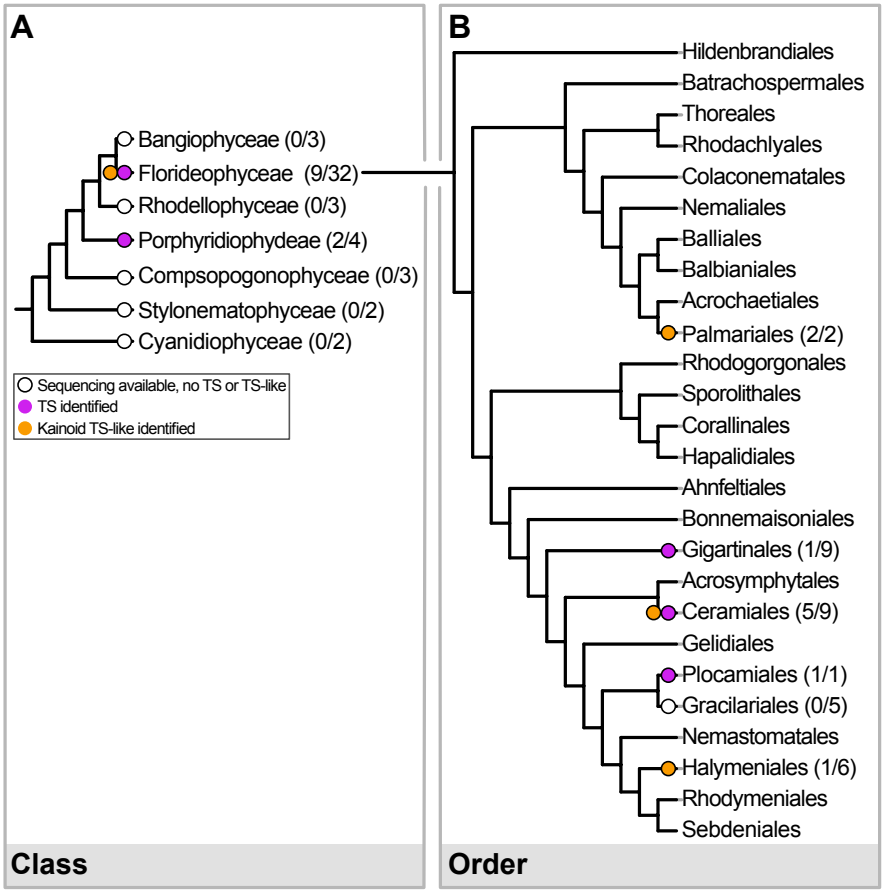


Figure 1.13: Cladogram of red algae by Order (**A**) by Class (**B**) within Florideophyceae. Blue circles denote sequencing data is available, but no TS or TS-like sequences have been identified. Yellow circles denote a typical microbial-type TS has been identified in the group. Red circles denote a TS-like 'A' protein, like DabA, has been identified.

proteins suggests HGT from a potential third host as a mechanism for the acquisition of these kainoid biosynthetic enzymes.

In summary, algal TS and repurposed TS-like sequences are not widespread across red algae (**Figure 1.13**). With the majority of known TS and TS-like sequences found within the class Florideophyceae. Both canonical terpenoid and kainoid producing sequences clade within the broader microbial-type class 1 family of TSs. This in addition to the sparse distribution of TS and TS-like sequences implies acquisition from an unknown microbial source, possibly via HGT. Algal terpenoids are typically highly functionalized and there remains a gap in knowledge of the downstream tailoring reactions, particularly in haloterpenoids. Similarly, no brown or green algal TS sequences are yet to be identified, despite many terpenoid compounds being isolate from these taxa (**Figure 1.2B, Table 1.1**). Interestingly, of the three algal lineages, green macroalgae are most closely related to land plants where terpenoids are common. However, terpenoids from green algae are rare and no TS genes have been identified, opening further questions around TS evolution and function in these distinct algal lineages.

1.11 Insects

In the chemical ecology of insects, terpenoids play important roles in defense and communication. For example (6R,7S)-himachala-9,11-diene (**54**) in the crucifer flea beetle *Phyllotreta striolata*, is a major male-produced aggregation pheromone.[110] Terpenoid biosynthesis in insects is an instance of striking chemical convergence. While insects contain typical IDS and mevalonate pathway enzymes, no homologs of plant or microbial-type TS genes have been found in the genomes of insects. Interestingly, many insect terpenoids are similar or even identical to plants, with many putative downstream

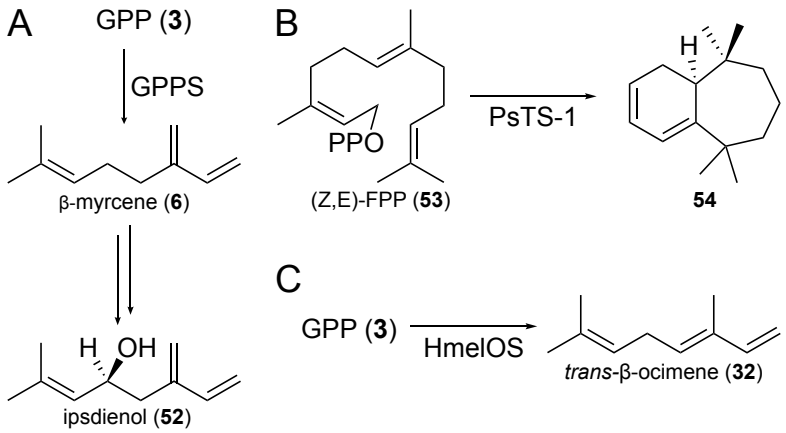


Figure 1.14 Insect terpenoid scaffolds **A.** β-myrcene (**6**) conversion to ipsdienol (**52**) in *Ips pini* **B.** Conversion of (*Z,E*)-FPP (**53**) to (6R,7S)-himachala-9,11-diene (**57**) by PsTPS1 in *Phyllotreta striolata* **C.** Production of (*E*)-β-ocimene (**32**) in *Heliconius melpomene* by HmelOS

oxidative tailoring reactions yet to be characterized. For example, (*E*)-β-caryophyllene is a sesquiterpene produced by plants such as maize, cannabis, and Arabidopsis[111] and in insects is produced by the Asian lady beetle.[112] Despite this shared chemistry, it is unlikely terpenoid biosynthesis arose in insects due to a common ancestral biosynthetic pathway with plants or other kingdoms, which highlights the fundamental question of whether insect terpenoids are synthesized de novo, derived from plant precursors, or produced by symbionts. Comprehensive analysis of insect genomes and transcriptomes suggests TS evolution occurred via gene duplication and neofunctionalization of insect IDS sequences. Interestingly, in several instances where high quality genome data is available, insect IDS-like TS sequences are found to be co-clustered with typical insect IDS sequences.[4] In the class of Insecta, all insect lineages sampled (Blattodea, Hemiptera, Phthiraptera, Hymenoptera, Coleoptera, Lepidoptera, and Diptera), except for the Phthiraptera (parasitic lice), contain characterized or predicted TS proteins.[113]

Early work by Martin *et al.* utilized cell-free assays of the terpenoid producing pine engraver beetle, *Ips pini*.[114] Cell-free assays using [1-3H] GPP (**3**) demonstrated **3** is converted to the acyclic monoterpene β-myrcene (**6**) which is a precursor to the *I. pini* aggregation pheromone ipsdienol (**52**) (**Figure 1.14A**). Historically, the biosynthesis of the aggregation pheromone components in *I. pini* had thought to occur only through hydroxylation of monoterpene olefins, such as β-myrcene (**6**) or α-pinene (**33**), derived from their conifer hosts.[115] However, the sex-specific and inducible enzymatic activity in extracts of male *I. pini* shifted the focus from **6** as a host-provided precursor to β-myrcene as an insect-derived biosynthetic intermediate.[116] Subsequent work successfully identified a geranyl diphosphate synthase (GPPS) with **6**-synthase activity in *I. pini*.[117] Recombinant enzyme produced both GPP (**3**) and **6**. Taken together this is the first evidence of genome encoded insect TS enzymatic machinery and the first example of a monoterpene synthase in the Metazoa.

More recent studies utilize transcriptome and genome sequencing to greatly expanded our understanding of TS enzymes in insects. Beran *et al.* investigate the biosynthesis of sesquiterpene aggregation pheromones in the flea beetle, *Phyllotreta striolata*, an important pest of crucifer crops in North America and Southeast Asia.[118] Here, the authors describe an emerging class of new, evolutionarily distinct, TS enzymes in *P. striolata*. Initially, no homologs to typical plant or microbial terpene synthases were identified, however, *P. striolata* was found to express a remarkably high number of trans-IDS sequences. From nine different IDS-type transcripts in the *P. striolata* transcriptome, two were found to encode bona fide *trans*-or *cis*-IDS enzymes, while four transcripts encode TSs, suggesting an evolutionary origin of these enzymes from IDS progenitors.

Using a combination of cell-free assays, transcriptomics, and heterologous expression, at least four different IDS-like TS enzymes were successfully identified and biochemically characterized, where the authors found PsTPS1 to convert (*Z,E*)-FPP to (6R,7S)-himachala-9,11-diene (**54**) (**Figure 1.14B**) and other sesquiterpenes observed in beetle extracts. Within the *P. striolata* transcriptome, two expressed enzymes had genuine IDS activity, with PsIDS1 synthesizing (*E,E*)-FPP (**4**), whereas PsIDS3 produced neryl diphosphate, (*Z,Z*)-FPP, and (*Z,E*)-FPP (**53**). Phylogenetic analysis of the new IDS-like TS enzymes, including several related species from the order Coleoptera, supports the evolutionary conversion of IDS to TS function. Interestingly, exon–intron structures of IDS and TS genes in *P. striolata* are conserved, further supporting this TS gene family evolved from IDS ancestors.

Works by Lancaster et al, identify and characterize IDS-like TS enzymes from the Southern green stink bug, *Nezara viridula* and the harlequin bug, *Murgantia histrionica*. Phylogenetic analysis suggests that insect TS activity might have emerged more recently from IDS proteins and this event has occurred independently in these and other insect taxa. Mining of brown marmorated stink bug, *Halyomorpha halys* genome and transcriptome revealed a family of six IDS-like TS genes.[4] This new family of IDS-like TS sequences was found to be organized in two separate clusters, each of which most likely emerged by gene duplication, where putative TS sequences were found to be co-clustered with typical insect IDS sequences. This model of identifying candidate TS enzymes by searching for IDS gene expansion was successful in the search for an (*E*)-β-ocimene (**32**) synthases in the butterfly genus *Heliconius*. Darragh *et al.* investigate the recent divergence of two butterfly species, *Heliconius melpomene* and *Heliconius cydno*,

which differ in the presence and absence of **32**.[119] Using linkage mapping, gene expression, and functional analyses, *H. melpomene* was found to have two new TS genes, one of which HmelOS, can produce **32** (**Figure 1.14C**) Furthermore, the *Heliconius* species TS enzymes are unrelated to previously described Coleopteran and Hemipteran sequences, supporting their independent evolutionary origins.

Although insect TS sequences are phylogenetically distant, the shared chemistry across kingdoms once again highlights the utility of terpenoids as an expansion of primary metabolic pathways across domains of life.

1.12 Amoeba

Amoebae are microeukaryotes that produce a variety of mono-, sesqui-, and diterpenoids.[120, 121] Dictyostelids, a subgroup of amoebae known for their unique life cycle consisting of both unicellular and multicellular phases, emit a complex mixture of terpenoids during the multicellular phases of their life cycles. The production of those dictyostelid terpenoids has been linked to several recently characterized terpene synthases, most of which are stand-alone terpene synthases, however a terpene

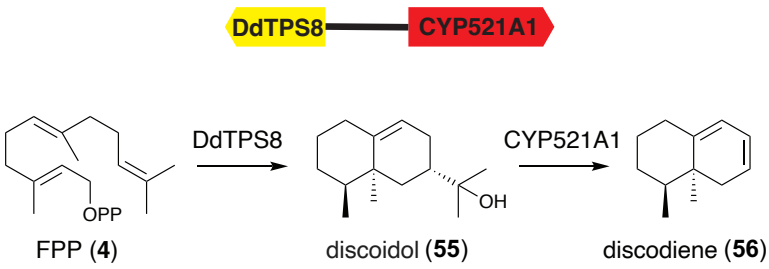


Figure 1.15 Discodiene biosynthesis in *Dictyostelium discoideum*. FPP (**4**) is converted to discoidol (**55**) by the class 1 TS DdTPS8, which is then converted to discoidiene (**56**) by the CYP450 CYP521A1.

synthase-CYP450 gene cluster was recently reported from an amoeba (**Figure 1.15**).[122] The first amoebal terpene synthases, DdTS1-9, were reported from *Dictyostelium discoideum* by Chen *et al.* in 2016.[123] Heterologously expressed DdTS1-9 produced a variety of sesquiterpenes, including those found in the headspace extracts of *D. discoideum*, such as (*E,E*)-a-farnesene, protoillud-7-ene, (3S)-(+)-asterisca-2(9),6-diene, and (–)-b-barbatene.[121, 123, 124] Interestingly, the expression of these terpene synthase genes in *D. discoideum* was linked to the amoeba's stages of development, with the highest expression during the latter half of the multicellular stage.[123] Further work revealed the presence of terpene synthases in other dictyostelid species, including *Dictyostelium purpureum*, *Polysphondylium pallidum*, and *Dictyostelium lacteum*.[121, 125]

While little is known about the ecological role of terpenes in amoeba, (*E,E*)-a-farnesene has been shown to have insect deterrent activity in plants,[126] which supports the hypothesis proposed by Chen *et al.* 2016 that (*E,E*)-a-farnesene could protect amoeba from predation by nematodes.[123] It's also hypothesized that the production of β-barbatene in the fruiting bodies stage of the amoebal life cycle could indicate its role in attracting insects for spore dispersal, an activity that has been documented for β-barbatene in fruiting body-producing fungi.[127] It was recently reported that some of these dictyostelid terpene synthases cluster with cytochrome P450 enzymes, the first reported case of biosynthetic gene clustering in amoeba.[122] DdTS8, which produces the sesquiterpene alcohol discoidol (**55**), clusters with CYP521A1, a cytochrome P450 that acts on **55** to produce discodiene (**56**), which is also found in the headspace of *D. discoideum* (**Figure 1.15**).[122] Similar TS-P450 clustering was also seen with 3 other *D. discoideum* TSs, as well as 3 TSs from *D. purpureum*, though the products of those

clusters were not determined. These amoebal clusters add to the growing list of terpenoid-encoding clusters from non-plant eukaryotes.

1.13 Sponges

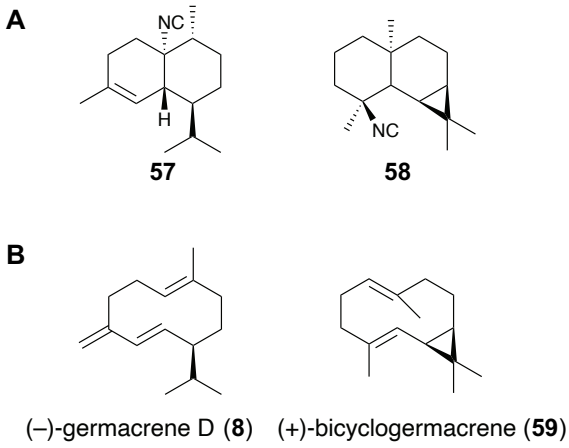


Figure 1.16 A. Sponge nitrogenous terpenes. **B.** Products of sponge terpene synthases.

Marine sponges are the most prolific source of marine natural products to date, with more than 9400 reported sponge molecules on CMNPD.[128] Sponges contain a rich array of terpenoids, including bioactive terpenoids such as the anti-malarial molecule kalihinol A, the anti-tumor compound avarol, the anti-inflammatory agent manoalide. Approximately 4300 sponge natural products are isoprenoids, comprising nearly half of all known sponge derived molecules. The majority of sponge isoprenoids likely originate from class 2 TSs, such as steroids and meroterpenoids. Scaffolds that originate from class 1 TSs comprise a smaller portion of sponge terpenoids, with approximately 1450 reported class 1 sponge terpenoids (**Figure 1.2, Table 1.1**). Many of the sponge class 1 terpenoids feature nitrogen-containing functional groups like isonitriles, isocyanates, and

isothiocyanates. These molecules, also known as nitrogenous terpenes, are only found in marine sponges and the nudibranchs that feed on them.

Very little is known about the biosynthesis of sponge derived terpenoids, however recently, a novel group of TSs were reported for the first time from marine sponges (**Figure 1.16**).[9] These TSs were reported from multiple species of marine sponges, several of which are reported to produce nitrogenous terpenes. They contain the usual DDXXD and RY motifs, as well as a slightly unusual NGD moiety and they clade distinctly from other known TSs from other domains of life. The genes that encode these TSs do not appear to cluster with other biosynthetic genes. Interestingly, these TSs were encoded by the sponge animal itself, which helped to shed light on a question that had long puzzled marine natural products scientists: Are sponge derived terpenoids produced by the sponge animal or its microbiome? Marine sponges feature a rich microbiome, which has made it difficult to determine whether a molecule derived from a sponge originates from the sponge animal host or a microbial symbiont. It has been well established that sponge associated microbes are responsible for the production of sponge derived molecules from other natural product classes, like PKSs, NRPSs, and PDBEs. Early cell sorting experiments, however, have found that terpenoids are often stored in the sponge animal host cells, which further supports the importance of the sponge host animal in the production of terpenoids.[129–132]

1.14 Discussion

Terpenoids persist across taxonomic kingdoms, geographic locations, and environmental niches because life is uniquely primed to evolve and maintain specialized terpenoid biosynthesis. Primary isoprenoid biosynthesis pathways are ubiquitous

throughout nature, providing the necessary precursors for secondary terpenoid biosynthesis. Similarly, the plasticity of terpene synthases raises the possibility that an inherent lack of substrate specificity in some TSs may play a role in their evolution. Additionally, enzymes, like CYP450s, often implicated in terpenoid tailoring, are widely distributed and often maintain enough substrate promiscuity to act on a variety of terpenoid products and facilitate the evolution of new pathways. Together, this suggests the cost of evolving a new terpenoid biosynthesis pathway is relatively low as the precursor production and tailoring machinery is already included. All that is required is to acquire a terpene synthase, either through horizontal gene transfer or neofunctionalization of an existing IDS, to produce new specialized terpenoids.

The small size (300-500 amino acids) and lack of specialized redox partners or complex cofactors make TSs excellent candidates for horizontal gene transfer and easy incorporation into an organism's metabolism, driving the emergence of new terpenoid pathways. As a group, TSs have a rugged evolutionary landscape and high plasticity, meaning small mutations often affect large changes to product promiscuity. Aside from a few conserved active site motifs, TSs typically have a high degree of sequence variability and small mutations can greatly affect product production. Coupled with the potent bioactivity of terpenoid products, positive evolutionary pressure to maintain and expand bioactive and ecologically relevant terpenoids can push a TS towards new product specificity. This high plasticity of TSs allows for many different sequence possibilities to produce a functional TS.

Terpenoids play a crucial role in mediating interspecies interactions. For example, terpenoid chemoattractants facilitate spore and gamete dispersal and antibacterial

terpenoids that protect plants from pathogens.[133–135] These ecological responses are triggered by terpenoids activating receptors, which in turn activate various signaling pathways. Terpenoids are well documented to activate olfactory receptors in insects, mammals, and other animals, which may orchestrate processes like development, growth, and reproduction. Recent research suggests that nuclear receptors, a group of ligand-activated transcription factors involved in reproduction and development across all metazoan groups, evolved from ancient terpene synthases and retained the same ligand-binding domain core of an ancestral terpene synthase shared with a metazoan ancestor.[136] These nuclear receptors are basal for metazoans, indicating that the evolutionary history of metazoans is closely linked to that of isoprenoids. Isoprenoid precursors are ubiquitous across all animals, as are the receptors that sense isoprenoids.

Terpene synthases can be acquired by organisms through two routes: neofunctionalization and horizontal gene transfer. In insects, neofunctionalization is the dominant route, as their terpene synthases closely resemble their own native IDSs, more than those of other bacterial or fungal terpene synthases. IDSs and TSs share an alpha helical fold and similar mechanisms, indicating that they may have a shared common ancestor. Horizontal gene transfer is currently the most common mode of TS evolution, as seen in fungi,[137] red algae,[108, 109] and plants.[59] The more distinct groups of two- and three-domain plant terpene synthases, coral,[8]and sponge[9] terpene synthases are less clear, but due to the structural similarity to microbial TS enzymes an ancient horizontal gene transfer event in both lineages is a possibility. The polyphyletic nature of the microbial-type TSs clade also supports the hypothesis of dynamic horizontal gene transfer events throughout these Phyla. Similarly, in red algae, phylogenetic analysis of

the TS-like prenyltransferases (A proteins) suggests a horizontal gene transfer event happened recently.

As we sequence a greater number and diversity of species, we continue to discover new specialized terpene biosynthesis genes across distinct evolutionary lineages. Many recent terpene synthase discoveries have come from non-model organisms like corals, sponges, algae, and insects. As such, there is an ongoing need to sequence unusual organisms and to deposit the resulting sequences into publicly available databases, such as NCBI and ENA. Additionally, bioinformatic toolsets need to be developed with terpene synthase discovery in mind. Unlike many other natural product biosynthesis genes such as PKSs or NRPSs, which have high sequence homology and are highly predictable by homology-based analysis tools, terpene synthases are often missed by standard genome mining programs. It is quite likely novel terpene synthases are waiting to be discovered in organisms that have yet to be sequenced. As shown in the discoveries of insect, coral, and sponge TSs, all of which were initially missed by traditional enzyme discovery tools. To continue to discover and expand our understanding of TS enzymology and evolution, a combinatorial approach of sequencing, homology-independent search tools, and biochemical validation is needed to discover truly novel terpene synthases.

1.15 References

1. M. G. Chevrette, K. Gutiérrez-García, N. Selem-Mojica, C. Aguilar-Martínez, A. Yañez-Olvera, H. E. Ramos-Aboites, P. A. Hoskisson, F. Barona-Gómez, Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat. Prod. Rep.* **37**, 566–599 (2020).

2. E. Pichersky, R. A. Raguso, Why do plants produce so many terpenoid compounds? *New Phytol.* **220**, 692–702 (2018).

3. F. Chen, D. Tholl, J. Bohlmann, E. Pichersky, The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J* **66**, 212–229 (2011).

4. D. Tholl, Z. Rebholz, A. V. Morozov, P. E. O'Maille, Terpene synthases and pathways in animals: enzymology and structural evolution in the biosynthesis of volatile infochemicals. *Nat. Prod. Rep.* **40**, 766-793 (2023).

5. M. B. Quin, C. M. Flynn, C. Schmidt-Dannert, Traversing the fungal terpenome. *Nat. Prod. Rep.* **31**, 1449–1473 (2014).

6. J. D. Rudolf, T. A. Alsup, B. Xu, Z. Li, Bacterial Terpenome. *Nat. Prod. Rep.* **23**, 9–10 (2006).

7. G. Wei, Q. Jia, X. Chen, T. G. Köllner, D. Bhattacharya, G. K.-S. Wong, J. Gershenzon, F. Chen, Terpene biosynthesis in red algae is catalyzed by microbial type but not typical plant terpene synthases. *Plant Physiol.* **179**, 382–390 (2019).

8. I. Burkhardt, T. de Rond, P. Y.-T. Chen, B. S. Moore, Ancient plant-like terpene biosynthesis in corals. *Nat. Chem. Biol.* **18**, 664–669 (2022).

9. K. Wilson, T. de Rond, I. Burkhardt, T. S. Steele, R. J. B. Schäfer, S. Podell, E. E. Allen, B. S. Moore, Terpene biosynthesis in marine sponge animals. Proc. Natl. Acad. Sci*. U.S.A.* **120**, e2220934120 (2023).

10. X. Xiong, J. Gou, Q. Liao, Y. Li, Q. Zhou, G. Bi, C. Li, R. Du, X. Wang, T. Sun, L. Guo, H. Liang, P. Lu, Y. Wu, Z. Zhang, D.-K. Ro, Y. Shang, S. Huang, J. Yan, The *Taxus* genome provides insights into paclitaxel biosynthesis. *Nat. Plants* **7**, 1026–1036 (2021).

11. C. J. Paddon, J. D. Keasling, Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat. Rev. Microbiol*. **12**, 355–367 (2014).

12. J. Lombard, D. Moreira, Origins and early evolution of the mevalonate pathway of isoprenoid biosynthesis in the three domains of life. *Mol. Biol. Evol.* **28**, 87–99 (2011).

13. G. Flesch, M. Rohmer, Prokaryotic hopanoids: the biosynthesis of the bacteriohopane skeleton. *Eur. J. Biochem.* **175**, 405–411 (1988).

14. A. L. Jaffe, C. J. Castelle, P. B. Matheus Carnevali, S. Gribaldo, J. F. Banfield, The rise of diversity in metabolic platforms across the Candidate Phyla Radiation. *BMC Biology* **18**, 69 (2020).

15. Y. Hoshino, E. A. Gaucher, On the origin of isoprenoid biosynthesis. *Mol. Biol. Evol.* **35**, 2185–2197 (2018).

16. R. Cao, Y. Zhang, F. M. Mann, C. Huang, D. Mukkamala, M. P. Hudock, M. E. Mead, S. Prisic, K. Wang, F.-Y. Lin, T.-K. Chang, R. J. Peters, E. Oldfield, Diterpene cyclases and the nature of the isoprene fold. *Proteins: Struct. Funct.* **78**, 2417–2432 (2010).

17. Dictionary of Natural Products (April 13, 2023).

18. B.-J. Yoon, Hidden Markov Models and their applications in biological sequence analysis. *Curr. Genomics* **10**, 402–415 (2009).

19. M. J. M. Christenhusz, J. W. Byng, The number of known plants species in the world and its annual increase. *Phytotaxa* **261**, 201–217 (2016).

20. K. D. Allen, K. McKernan, C. Pauli, J. Roe, A. Torres, R. Gaudino, Genomic characterization of the complete terpene synthase gene family from *Cannabis sativa*. *PLOS ONE* **14**, e0222363 (2019).

21. C. M. Starks, K. Back, J. Chappell, J. P. Noel, Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase. *Science* **277**, 1815–1820 (1997).

22. M. Köksal, Y. Jin, R. M. Coates, R. Croteau, D. W. Christianson, Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature* **469**, 116–120 (2011).

23. P. Hedden, The current status of research on gibberellin biosynthesis. *Plant Cell Physiol.* **61**, 1832–1849 (2020).

24. C. Schwechheimer, Gibberellin signaling in plants – the extended version. *Front. Plant. Sci*. **2** (2012).

25. C. I. Keeling, H. K. Dullat, M. Yuen, S. G. Ralph, S. Jancsik, J. Bohlmann, Identification and functional characterization of monofunctional ent-copalyl diphosphate and ent-kaurene synthases in white spruce reveal different patterns for diterpene synthase evolution for primary and secondary metabolism in gymnosperms. *Plant Physiol.* **152**, 1197–1208 (2010).

26. K. Hayashi, H. Kawaide, M. Notomi, Y. Sakigi, A. Matsuo, H. Nozaki, Identification and functional analysis of bifunctional ent-kaurene synthase from the moss *Physcomitrella patens*. *FEBS Letters* **580**, 6175–6181 (2006).

27. Q. Jia, R. Brown, T. G. Köllner, J. Fu, X. Chen, G. K.-S. Wong, J. Gershenzon, R. J. Peters, F. Chen, Origin and early evolution of the plant terpene synthase family. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2100361119 (2022).

28. T. Kato, C. Kabuto, N. Sasaki, M. Tsunagawa, H. Aizawa, K. Fujita, Y. Kato, Y. Kitahara, N. Takahashi, Momilactones, growth inhibitors from rice, *Oryza sativa*, *Tetrahedron Lett.* **14**, 3861–3864 (1973).

29. T. Toyomasu, M. Usui, C. Sugawara, K. Otomo, Y. Hirose, A. Miyao, H. Hirochika, K. Okada, T. Shimizu, J. Koga, M. Hasegawa, M. Chuba, Y. Kawana, M. Kuroda, E. Minami, W. Mitsuhashi, H. Yamane, Reverse-genetic approach to verify physiological roles of rice phytoalexins: characterization of a knockdown mutant of OsCPS4 phytoalexin biosynthetic gene in rice. *Physiologia Plantarum* **150**, 55–62 (2014).

30. H. Kato-Noguchi, M. Hasegawa, T. Ino, K. Ota, H. Kujime, Contribution of momilactone A and B to rice allelopathy. *J. Plant Physiol*. **167**, 787–791 (2010).

31. K. Otomo, Y. Kanno, A. Motegi, H. Kenmoku, H. Yamane, W. Mitsuhashi, H. Oikawa, H. Toshima, H. Itoh, M. Matsuoka, T. Sassa, T. Toyomasu, Diterpene cyclases responsible for the biosynthesis of phytoalexins, momilactones a, b, and oryzalexins a–f in rice. *Biosci. Biotechnol. Biochem.* **68**, 2001–2006 (2004).

32. P. R. Wilderman, M. Xu, Y. Jin, R. M. Coates, R. J. Peters, Identification of syn-pimara-7,15-diene synthase reveals functional clustering of terpene synthases involved in rice phytoalexin/allelochemical biosynthesis. *Plant Physiol.* **135**, 2098–2105 (2004).

33. R. De La Peña, E. S. Sattely, Rerouting plant terpene biosynthesis enables momilactone pathway elucidation. *Nat. Chem. Biol.* **17**, 205–212 (2021).

34. K. Shimura, A. Okada, K. Okada, Y. Jikumaru, K.-W. Ko, T. Toyomasu, T. Sassa, M. Hasegawa, O. Kodama, N. Shibuya, J. Koga, H. Nojiri, H. Yamane, Identification of a biosynthetic gene cluster in rice for momilactones. *J. Biol. Chem.* **282**, 34013–34018 (2007).

35. L. Mao, H. Kawaide, T. Higuchi, M. Chen, K. Miyamoto, Y. Hirata, H. Kimura, S. Miyazaki, M. Teruya, K. Fujiwara, K. Tomita, H. Yamane, K. Hayashi, H. Nojiri, L. Jia, J. Qiu, C. Ye, M. P. Timko, L. Fan, K. Okada, Genomic evidence for convergent evolution of gene clusters for momilactone biosynthesis in land plants. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 12472–12480 (2020).

36. K. Okada, H. Kawaide, K. Miyamoto, S. Miyazaki, R. Kainuma, H. Kimura, K. Fujiwara, M. Natsume, H. Nojiri, M. Nakajima, H. Yamane, Y. Hatano, H. Nozaki,

K. Hayashi, HpDTC1, a stress-inducible bifunctional diterpene cyclase involved in momilactone biosynthesis, functions in chemical defense in the moss *Hypnum plumaeforme*. *Sci. Rep.* **6**, 25316 (2016).

37. M. R. Wildung, R. Croteau, A cDNA clone for taxadiene synthase, the diterpene cyclase that catalyzes the committed step of taxol biosynthesis. *J. Biol. Chem.* **271**, 9201–9204 (1996).

38. M. T. Dueber, W. Adolf, C. A. West, Biosynthesis of the diterpene phytoalexin casbene: partial purification and characterization of casbene synthetase from *Ricinis communis*. *Plant Physiol*. **62**, 598–603 (1978).

39. S. Jennewein, R. M. Long, R. M. Williams, R. Croteau, Cytochrome P450 taxadiene 5α-hydroxylase, a mechanistically unusual monooxygenase catalyzing the first oxygenation step of taxol biosynthesis. *Chem. Biol.* **11**, 379–387 (2004).

40. J. Cheng, X. Wang, X. Liu, X. Zhu, Z. Li, H. Chu, Q. Wang, Q. Lou, B. Cai, Y. Yang, X. Lu, K. Peng, D. Liu, Y. Liu, L. Lu, H. Liu, T. Yang, Q. Ge, C. Shi, G. Liu, Z. Dong, X. Xu, W. Wang, H. Jiang, Y. Ma, Chromosome-level genome of Himalayan yew provides insights into the origin and evolution of the paclitaxel biosynthetic pathway. *Mol. Plant* **14**, 1199–1209 (2021).

41. M. Fattahian, M. Ghanadian, Z. Ali, I. A. Khan, Jatrophane and rearranged jatrophane-type diterpenes: biogenesis, structure, isolation, biological activity and SARs (1984–2019). *Phytochem. Rev.* **19**, 265–336 (2020).

42. D. Luo, R. Callari, B. Hamberger, S. G. Wubshet, M. T. Nielsen, J. Andersen-Ranberg, B. M. Hallström, F. Cozzi, H. Heider, B. Lindberg Møller, D. Staerk, B. Hamberger, Oxidation and cyclization of casbene in the biosynthesis of Euphorbia factors from mature seeds of *Euphorbia lathyrism*. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E5082–E5089 (2016).

43. A. J. King, G. D. Brown, A. D. Gilday, T. R. Larson, I. A. Graham, production of bioactive diterpenoids in the Euphorbiaceae depends on evolutionarily conserved gene clusters. *The Plant Cell* **26**, 3286–3298 (2014).

44. A. J. King, G. D. Brown, A. D. Gilday, E. Forestier, T. R. Larson, I. A. Graham, A cytochrome p450-mediated intramolecular carbon–carbon ring closure in the biosynthesis of multidrug-resistance-reversing lathyrane diterpenoids. *ChemBioChem* **17**, 1593–1597 (2016).

45. T. Czechowski, E. Forestier, S. H. Swamidatta, A. D. Gilday, A. Cording, T. R. Larson, D. Harvey, Y. Li, Z. He, A. J. King, G. D. Brown, I. A. Graham, Gene discovery and virus-induced gene silencing reveal branched pathways to major classes of bioactive diterpenoids in *Euphorbia peplus*. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2203890119 (2022).

46. P. Heinstein, Stimulation of sesquiterpene aldehyde formation in *Gossypium arboreum* cell suspension cultures by conidia of *Verticillium dahliae*. *J. Nat. Prod.* **48**, 907–915 (1985).

47. A. A. Bell, R. D. Stipanovic, C. R. Howell, P. A. Fryxell, Antimicrobial terpenoids of Gossypium: Hemigossypol, 6-methoxyhemigossypol and 6-deoxyhemigossypol. *Phytochem.* **14**, 225–231 (1975).

48. X. Tian, J.-X. Ruan, J.-Q. Huang, C.-Q. Yang, X. Fang, Z.-W. Chen, H. Hong, L.-J. Wang, Y.-B. Mao, S. Lu, T.-Z. Zhang, X.-Y. Chen, Characterization of gossypol biosynthetic pathway. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E5410–E5418 (2018).

49. S. T. Krause, P. Liao, C. Crocoll, B. Boachon, C. Förster, F. Leidecker, N. Wiese, D. Zhao, J. C. Wood, C. R. Buell, J. Gershenzon, N. Dudareva, J. Degenhardt, The biosynthesis of thymol, carvacrol, and thymohydroquinone in Lamiaceae proceeds via cytochrome P450s and a short-chain dehydrogenase. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2110092118 (2021).

50. M. Rosenkranz, Y. Chen, P. Zhu, A. C. Vlot, Volatile terpenes – mediators of plant-to-plant communication. *Plant J.* **108**, 617–631 (2021).

51. S. Rasmann, T. G. Köllner, J. Degenhardt, I. Hiltpold, S. Toepfer, U. Kuhlmann, J. Gershenzon, T. C. J. Turlings, Recruitment of entomopathogenic nematodes by insect-damaged maize roots. *Nature* **434**, 732–737 (2005).

52. M. Huang, A. M. Sanchez-Moreiras, C. Abel, R. Sohrabi, S. Lee, J. Gershenzon, D. Tholl, The major volatile organic compound emitted from *Arabidopsis thaliana* flowers, the sesquiterpene (*E*)-β-caryophyllene, is a defense against a bacterial pathogen. *New Phytol.* **193**, 997–1008 (2012).

53. J. T. Knudsen, R. Eriksson, J. Gershenzon, B. Ståhl, Diversity and distribution of floral scent. *Bot. Rev* **72**, 1 (2006).

54. S. Roeder, A.-M. Hartmann, U. Effmert, B. Piechulla, Regulation of simultaneous synthesis of floral scent terpenoids by the 1,8-cineole synthase of *Nicotiana suaveolens. Plant Mol. Biol.* **65**, 107–124 (2007).

55. J. H. Loughrin, T. R. Hamilton-Kemp, H. R. Burton, R. A. Andersen, Effect of diurnal sampling on the headspace composition of detached *Nicotiana suaveolens* flowers. *Phytochem.* **32**, 1417–1419 (1993).

56. R. A. Raguso, R. A. Levin, S. E. Foose, M. W. Holmberg, L. A. McDade, Fragrance chemistry, nocturnal rhythms and pollination “syndromes” in *Nicotiana*. *Phytochem.* **63**, 265–284 (2003).

57. G. Li, T. G. Köllner, Y. Yin, Y. Jiang, H. Chen, Y. Xu, J. Gershenzon, E. Pichersky, F. Chen, Nonseed plant *Selaginella moellendorffii* has both seed plant and

microbial types of terpene synthases. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14711–14715 (2012).

58. Q. Jia, G. Li, T. G. Köllner, J. Fu, X. Chen, W. Xiong, B. J. Crandall-Stotler, J. L. Bowman, D. J. Weston, Y. Zhang, L. Chen, Y. Xie, F.-W. Li, C. J. Rothfels, A. Larsson, S. W. Graham, D. W. Stevenson, G. K.-S. Wong, J. Gershenzon, F. Chen, Microbial-type terpene synthase genes occur widely in nonseed land plants, but not in seed plants. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12328–12333 (2016).

59. Q. Jia, T. G. Köllner, J. Gershenzon, F. Chen, MTPSLs: New terpene synthases in nonseed plants. *Trends Plant Sci.* **23**, 121–128 (2018).

60. C. Schmidt-Dannert, Biosynthesis of terpenoid natural products in fungi. *Adv. Biochem. Eng. Biotechnol*. **148**, 19–61 (2015).

61. P. Zhang, G. Wu, S. C. Heard, C. Niu, S. A. Bell, F. Li, Y. Ye, Y. Zhang, J. M. Winter, Identification and characterization of a cryptic bifunctional type i diterpene synthase involved in talaronoid biosynthesis from a marine-derived fungus. *Org. Lett*. **24**, 7037–7041 (2022).

62. Y. Yuan, S. Cheng, G. Bian, P. Yan, Z. Ma, W. Dai, R. Chen, S. Fu, H. Huang, H. Chi, Y. Cai, Z. Deng, T. Liu, Efficient exploration of terpenoid biosynthetic gene clusters in filamentous fungi. *Nat. Catal.* **5**, 277–287 (2022).

63. J. M. Galindo-Solís, F. J. Fernández, Endophytic fungal terpenoids: Natural role and bioactivities. *Microorganisms* **10**, 339 (2022).

64. R. Schobert, S. Knauer, S. Seibt, B. Biersack, Anticancer active illudins: recent developments of a potent alkylating compound class. *Curr. Med. Chem.* **18**, 790–807 (2011).

65. M. Kimura, T. Tokai, N. Takahashi-Ando, S. Ohsato, M. Fujimura, Molecular and genetic studies of fusarium trichothecene biosynthesis: Pathways, genes, and evolution. *Biosci. Biotechnol. Biochem*. **71**, 2105–2123 (2007).

66. M. J. Rynkiewicz, D. E. Cane, D. W. Christianson, Structure of trichodiene synthase from *Fusarium sporotrichioides* provides mechanistic inferences on the terpene cyclization cascade. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 13543–13548 (2001).

67. A. V. Nathanail, E. Varga, J. Meng-Reiterer, C. Bueschl, H. Michlmayr, A. Malachova, P. Fruhmann, M. Jestoi, K. Peltonen, G. Adam, M. Lemmens, R. Schuhmacher, F. Berthiller, Metabolism of the fusarium mycotoxins T-2 toxin and HT-2 toxin in wheat. J. Agric. *Food Chem.* **63**, 7862–7872 (2015).

68. Tri6 encodes an unusual zinc finger protein involved in regulation of trichothecene biosynthesis in *Fusarium sporotrichioides*. *Appl. Environ. Microbiol.* **61**, 1923-1930 (1995).

69. M. M. Cadelis, B. R. Copp, S. Wiles, A review of fungal protoilludane sesquiterpenoid natural products. *Antibiotics* **9**, 928 (2020).

70. G. T. Wawrzyn, M. B. Quin, S. Choudhary, F. López-Gallego, C. Schmidt-Dannert, Draft genome of *Omphalotus olearius* provides a predictive framework for sesquiterpenoid natural product biosynthesis in Basidiomycota. *Chem. Biol.* **19**, 772–783 (2012).

71. M. Chen, W. K. W. Chou, T. Toyomasu, D. E. Cane, D. W. Christianson, Structure and function of fusicoccadiene synthase, a hexameric bifunctional diterpene synthase*. ACS Chem. Biol*. **11**, 889–899 (2016).

72. R. Chiba, A. Minami, K. Gomi, H. Oikawa, identification of ophiobolin f synthase by a genome mining approach: A sesterterpene synthase from *Aspergillus clavatus*. *Org. Lett*. **15**, 594–597 (2013).

73. H. Tao, L. Lauterbach, G. Bian, R. Chen, A. Hou, T. Mori, S. Cheng, B. Hu, L. Lu, X. Mu, M. Li, N. Adachi, M. Kawasaki, T. Moriya, T. Senda, X. Wang, Z. Deng, I. Abe, J. S. Dickschat, T. Liu, Discovery of non-squalene triterpenes. *Nature* **606**, 414–419 (2022).

74. J. L. Chen, R. E. Moore, G. M. L. Patterson, Structures of nostocyclophanes A-D. *J. Org. Chem.* **56**, 4360–4364 (1991).

75. B. Zhao, X. Lin, L. Lei, D. C. Lamb, S. L. Kelly, M. R. Waterman, D. E. Cane, Biosynthesis of the sesquiterpene antibiotic albaflavenone in *Streptomyces coelicolor* a3(2). *J. Biol. Chem.* **283**, 8183–8189 (2008).

76. C. N. Tetzlaff, Z. You, D. E. Cane, S. Takamatsu, S. Omura, H. Ikeda, a gene cluster for biosynthesis of the sesquiterpenoid antibiotic pentalenolactone in *Streptomyces avermitilis*. *Biochem.* **45**, 6179–6186 (2006).

77. S.-Y. Kim, P. Zhao, M. Igarashi, R. Sawa, T. Tomita, M. Nishiyama, T. Kuzuyama, Cloning and heterologous expression of the cyclooctatin biosynthetic gene cluster afford a diterpene cyclase and two p450 hydroxylases. *Chem. Biol.* **16**, 736–743 (2009).

78. C. Nakano, M. Oshima, N. Kurashima, T. Hoshino, Identification of a new diterpene biosynthetic gene cluster that produces O-methylkolavelool in *Herpetosiphon aurantiacus*. *ChemBioChem* **16**, 772–781 (2015).

79. W. Liu, X. Feng, Y. Zheng, C.-H. Huang, C. Nakano, T. Hoshino, S. Bogue, T.-P. Ko, C.-C. Chen, Y. Cui, J. Li, I. Wang, S.-T. D. Hsu, E. Oldfield, R.-T. Guo, Structure, function and inhibition of ent-kaurene synthase from *Bradyrhizobium japonicum*. *Sci. Rep.* **4**, 6214 (2014).

80. K. Schulz-Bohm, L. Martín-Sánchez, P. Garbeva, Microbial volatiles: Small molecules with an important role in intra- and inter-kingdom interactions. *Front. Microbiol*. **8**, 2484 (2017).

81. H. Gürtler, R. Pedersen, U. Anthoni, C. Christophersen, P. H. Nielsen, E. M. H. Wellington, C. Pedersen, K. Bock, Albaflavenone, a sesquiterpene ketone with a zizaene skeleton produced by a Streptomycete with a new rope morphology. *J. Antibiot.* **47**, 434–439 (1994).

82. P. Garbeva, C. Hordijk, S. Gerards, W. de Boer, Volatile-mediated interactions between phylogenetically different soil bacteria. *Front. Microbiol.* **5**, 289 (2014).

83. C. Song, R. Schmidt, V. de Jager, D. Krzyzanowska, E. Jongedijk, K. Cankar, J. Beekwilder, A. van Veen, W. de Boer, J. A. van Veen, P. Garbeva, Exploring the genomic traits of fungus-feeding bacterial genus Collimonas. *BMC Genomics* **16**, 1103 (2015).

84. J. S. Dickschat, Bacterial terpene cyclases. Nat. Prod. Rep. **33**, 87–110 (2015).

85. T. Tomita, S.-Y. Kim, K. Teramoto, A. Meguro, T. Ozaki, A. Yoshida, Y. Motoyoshi, N. Mori, K. Ishigami, H. Watanabe, M. Nishiyama, T. Kuzuyama, Structural insights into the CotB2-catalyzed cyclization of geranylgeranyl diphosphate to the diterpene cyclooctat-9-en-7-ol. *ACS Chem. Biol.* **12**, 1621–1628 (2017).

86. R. Driller, S. Janke, M. Fuchs, E. Warner, A. R. Mhashal, D. T. Major, M. Christmann, T. Brück, B. Loll, Towards a comprehensive understanding of the structural dynamics of a bacterial diterpene synthase during catalysis. *Nat. Commun.* **9**, 3971 (2018).

87. R. Janke, C. Görner, M. Hirte, T. Brück, B. Loll, The first structure of a bacterial diterpene cyclase: CotB2. *Acta. Cryst. D.* **70**, 1528–1537 (2014).

88. F. Aqil, M. Zahin, K. A. El Sayed, I. Ahmad, K. Y. Orabi, J. M. Arif, Antimicrobial, antioxidant, and antimutagenic activities of selected marine natural products and tobacco cembranoids. *Drug Chem. Toxicol.* **34**, 167–179 (2011).

89. A. H. de Boer, I. J. de V. Leeuwen, Fusicoccanes: diterpenes with surprising biological functions. *Trends Plant Sci.* **17**, 360–368 (2012).

90. N. P. Thao, B. T. T. Luyen, R. Brun, M. Kaiser, P. Van Kiem, C. Van Minh, T. J. Schmidt, J. S. Kang, Y. H. Kim, Anti-protozoal activities of cembrane-type diterpenes from vietnamese soft corals. *Molecules* **20**, 12459–12468 (2015).

91. R. Driller, D. Garbe, N. Mehlmer, M. Fuchs, K. Raz, D. T. Major, T. Brück, B. Loll, Current understanding and biotechnological application of the bacterial diterpene synthase CotB2. *Beilstein J. Org. Chem*. **15**, 2355–2368 (2019).

92. V. Liato, M. Aïder, Geosmin as a source of the earthy-musty smell in fruits, vegetables and water: Origins, impact on foods and water, and review of the removing techniques. *Chemosphere* **181**, 9–18 (2017).

93. P. Garbeva, M. Avalos, D. Ulanova, G. P. van Wezel, J. S. Dickschat, Volatile sensation: The chemical ecology of the earthy odorant geosmin. *Environ. Microbiol.*, 1–10 (2023).

94. L. Zaroubi, I. Ozugergin, K. Mastronardi, A. Imfeld, C. Law, Y. Gélinas, A. Piekny, B. L. Findlay, The ubiquitous soil terpene geosmin acts as a warning chemical. *Appl. Environ. Microbiol*. **88**, e00093-22 (2022).

95. M. C. Stensmyr, H. K. M. Dweck, A. Farhan, I. Ibba, A. Strutz, L. Mukunda, J. Linz, V. Grabe, K. Steck, S. Lavista-Llanos, D. Wicher, S. Sachse, M. Knaden, P. G. Becher, Y. Seki, B. S. Hansson, A conserved dedicated olfactory circuit for detecting harmful microbes in drosophila. *Cell* **151**, 1345–1357 (2012).

96. P. G. Becher, M. Bengtsson, B. S. Hansson, P. Witzgall, Flying the fly: Long-range flight behavior of *Drosophila melanogaster* to attractive odors. *J. Chem. Ecol.* **36**, 599–607 (2010).

97. N. Melo, G. H. Wolff, A. L. Costa-da-Silva, R. Arribas, M. F. Triana, M. Gugger, J. A. Riffell, M. DeGennaro, M. C. Stensmyr, Geosmin attracts *Aedes aegypti* mosquitoes to oviposition sites. *Curr. Biol.* **30**, 127–134 (2020).

98. W. Zhou, Y. Wang, J. Wang, C. Peng, Z. Wang, H. Qin, G. Li, D. Li, Geosmin disrupts energy metabolism and locomotor behavior of zebrafish in early life stages. *Sci. Total Environ.* **859**, 160222 (2023).

99. J. Jiang, X. He, D. E. Cane, Biosynthesis of the earthy odorant geosmin by a bifunctional *Streptomyces coelicolor* enzyme. *Nat. Chem. Biol.* **3**, 711–715 (2007).

100. M. Köksal, W. K. W. Chou, D. E. Cane, D. W. Christianson, Structure of 2-methylisoborneol synthase from streptomyces coelicolor and implications for the cyclization of a noncanonical c-methylated monoterpenoid substrate. *Biochemistry* **51**, 3011–3020 (2012).

101. V. J. Paul, M. E. Hay, J. E. Duffy, W. Fenical, K. Gustafson, Chemical defense in the seaweed *Ochtodes secundiramea* (Montagne) Howe (Rhodophyta): effects of its monoterpenoid components upon diverse coral-reef herbivores. *J. Exp. Mar. Biol. Ecol.* **114**, 249–260 (1988).

102. M. Sandy, J. N. Carter-Franklin, J. D. Martin, A. Butler, Vanadium bromoperoxidase from *Delisea pulchra*: Enzyme-catalyzed formation of bromofuranone and attendant disruption of quorum sensing. *Chem. Commun.* **47**, 12086–12088 (2011).

103. W. C. Paradas, T. M. Crespo, L. T. Salgado, L. R. de Andrade, A. R. Soares, C. Hellio, R. R. Paranhos, L. J. Hill, G. M. de Souza, A. G. A. C. Kelecom, B. A. P. Da Gama, R. C. Pereira, G. M. Amado-Filho, Mevalonosomes: Specific vacuoles containing the mevalonate pathway in *Plocamium brasiliense* cortical cells (Rhodophyta). *J. Phycol.* **51**, 225–235 (2015).

104. R. D. Kersten, S. Lee, D. Fujita, T. Pluskal, S. Kram, J. E. Smith, T. Iwai, J. P. Noel, M. Fujita, J.-K. Weng, A red algal bourbonane sesquiterpene synthase defined by microgram-scale NMR-coupled crystalline sponge X-ray diffraction analysis. *J. Am. Chem. Soc.* **139**, 16838–16844 (2017).

105. J. N. Carter-Franklin, A. Butler, Vanadium bromoperoxidase-catalyzed biosynthesis of halogenated marine natural products. *J. Am. Chem. Soc.* **126**, 15060–15066 (2004).

106. J. K. Brunson, S. M. K. McKinnie, J. R. Chekan, J. P. McCrow, Z. D. Miles, E. M. Bertrand, V. A. Bielinski, H. Luhavaya, M. Oborník, G. J. Smith, D. A. Hutchins, A. E. Allen, B. S. Moore, Biosynthesis of the neurotoxin domoic acid in a bloom-forming diatom. *Science* **361**, 1356–1358 (2018).

107. J. R. Chekan, S. M. K. McKinnie, M. L. Moore, S. G. Poplawski, T. P. Michael, B. S. Moore, Scalable biosynthesis of the seaweed neurochemical, kainic acid. *Angew. Chem., Int. Ed.* **58**, 8454–8457 (2019).

108. J. R. Chekan, S. M. K. McKinnie, J. P. Noel, B. S. Moore, Algal neurotoxin biosynthesis repurposes the terpene cyclase structural fold into an N-prenyltransferase. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 12799–12805 (2020).

109. T. S. Steele, J. K. Brunson, Y. Maeno, R. Terada, A. E. Allen, M. Yotsu-Yamashita, J. R. Chekan, B. S. Moore, Domoic acid biosynthesis in the red alga *Chondria armata* suggests a complex evolutionary history for toxin production. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2117407119 (2022).

110. R. J. Bartelt, B. W. Zilkowski, A. A. Cossé, U. Schnupf, K. Vermillion, F. A. Momany, male-specific sesquiterpenes from Phyllotreta flea beetles. *J. Nat. Prod.* **74**, 585–595 (2011).

111. F. Beran, T. G. Köllner, J. Gershenzon, D. Tholl, Chemical convergence between plants and insects: biosynthetic origins and functions of common secondary metabolites. *New Phytol.* **223**, 52–67 (2019).

112. A. E. Brown, E. W. Riddick, J. R. Aldrich, W. E. Holmes, Identification of (−)-β-caryophyllene as a gender-specific terpene produced by the multicolored asian lady beetle. *J. Chem. Ecol.* **32**, 2489–2499 (2006).

113. Z. Rebholz, L. Shewade, K. Kaler, H. Larose, F. Schubot, D. Tholl, A. V. Morozov, P. E. O'Maille, Emergence of terpene chemical communication in insects: Evolutionary recruitment of isoprenoid metabolism. *Protein Sci.,* e4634 (2023).

114. D. Martin, J. Bohlmann, J. Gershenzon, W. Francke, S. J. Seybold, A novel sex-specific and inducible monoterpene synthase activity associated with a pine bark beetle, the pine engraver, *Ips pini*. *Naturwissenschaften* **90**, 173–179 (2003).

115. P. R. Hughes, Myrcene: A precursor of pheromones in Ips beetles. *J. Insect Physiol.* **20**, 1271–1275 (1974).

116. S. J. Seybold, D. R. Quilici, J. A. Tillman, D. Vanderwel, D. L. Wood, G. J. Blomquist, *De novo* biosynthesis of the aggregation pheromone components ipsenol and ipsdienol by the pine bark *beetles Ips paraconfusus* (Lanier) and *Ips pini* (Say) (Coleoptera: Scolytidae). *Proc. Natl. Acad. Sci. U.S.A.* **92**, 8393–8397 (1995).

117. A. B. Gilg, C. Tittiger, G. J. Blomquist, Unique animal prenyltransferase with monoterpene synthase activity. *Naturwissenschaften* **96**, 731–735 (2009).

118. F. Beran, P. Rahfeld, K. Luck, R. Nagel, H. Vogel, N. Wielsch, S. Irmisch, S. Ramasamy, J. Gershenzon, D. G. Heckel, T. G. Köllner, Novel family of terpene synthases evolved from trans-isoprenyl diphosphate synthases in a flea beetle. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2922–2927 (2016).

119. K. Darragh, A. Orteu, D. Black, K. J. R. P. Byers, D. Szczerbowski, I. A. Warren, P. Rastas, A. Pinharanda, J. W. Davey, S. F. Garza, D. A. Almeida, R. M. Merrill, W. O. McMillan, S. Schulz, C. D. Jiggins, A novel terpene synthase controls differences in anti-aphrodisiac pheromone production between closely related *Heliconius* butterflies. *PLOS Biology* **19**, e3001022 (2021).

120. X. Chen, T. G. Köllner, W. Xiong, G. Wei, F. Chen, Emission and biosynthesis of volatile terpenoids from the plasmodial slime mold *Physarum polycephalum*. *J. Org. Chem*. **15**, 2872–2880 (2019).

121. J. Rinkel, T. G. Köllner, F. Chen, J. S. Dickschat, Characterisation of three terpene synthases for β-barbatene, β-araneosene and nephthenol from social amoebae. *Chem. Commun.* **55**, 13255–13258 (2019).

122. X. Chen, K. Luck, P. Rabe, C. Q. Dinh, G. Shaulsky, D. R. Nelson, J. Gershenzon, J. S. Dickschat, T. G. Köllner, F. Chen, A terpene synthase-cytochrome P450 cluster in *Dictyostelium discoideum* produces a novel trisnorsesquiterpene. *eLife* **8**, e44352 (2019).

123. X. Chen, T. G. Köllner, Q. Jia, A. Norris, B. Santhanam, P. Rabe, J. S. Dickschat, G. Shaulsky, J. Gershenzon, F. Chen, Terpene synthase genes in eukaryotes beyond plants and fungi: Occurrence in social amoebae. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12132–12137 (2016).

124. P. Rabe, J. Rinkel, B. Nubbemeyer, T. G. Köllner, F. Chen, J. S. Dickschat, Terpene cyclases from social amoebae. *Angew. Chem., Int. Ed.* **55**, 15420–15423 (2016).

125. X. Chen, T. G. Köllner, G. Shaulsky, Q. Jia, J. S. Dickschat, J. Gershenzon, F. Chen, Diversity and functional evolution of terpene synthases in Dictyostelid social amoebae. *Sci. Rep.* **8**, 14361 (2018).

126. M. Huang, C. Abel, R. Sohrabi, J. Petri, I. Haupt, J. Cosimano, J. Gershenzon, D. Tholl, Variation of herbivore-induced volatile terpenes among *Arabidopsis* ecotypes depends on allelic differences and subcellular targeting of two terpene synthases, TPS02 and TPS03. *Plant Physiol.* **153**, 1293–1310 (2010).

127. J. Fäldt, M. Jonsell, G. Nordlander, A.-K. Borg-Karlson, Volatiles of bracket fungi *Fomitopsis pinicola* and *Fomes fomentarius* and their functions as insect attractants. *J. Chem. Ecol.* **25**, 567–590 (1999).

128. C. Lyu, T. Chen, B. Qiang, N. Liu, H. Wang, L. Zhang, Z. Liu, CMNPD: a comprehensive marine natural products database towards facilitating drug discovery from the ocean. *Nucleic Acids Res.* **49**, D509–D515 (2021).

129. M. D. Unson, D. J. Faulkner, Cyanobacterial symbiont biosynthesis of chlorinated metabolites from *Dysidea herbacea* (Porifera). *Experientia* **49**, 349–353 (1993).

130. A. E. Flowers, M. J. Garson, R. I. Webb, E. J. Dumdei, R. D. Charan, Cellular origin of chlorinated diketopiperazines in the dictyoceratid sponge *Dysidea herbacea* (Keller). *Cell Tissue Res.* **292**, 597–607 (1998).

131. M. J. Uriz, X. Turon, J. Galera, J. M. Tur, New light on the cell location of avarol within the sponge *Dysidea avara* (Dendroceratida). *Cell Tissue Res.* **285**, 519–527 (1996).

132. M. J. Garson, J. E. Thompson, R. M. Larsen, C. N. Battershill, P. T. Murphy, P. R. Bergquist, Terpenes in sponge cell membranes: Cell separation and membrane fractionation studies with the tropical marine sponge *Amphimedon sp. Lipids* **27**, 378–388 (1992).

133. P. Becher, V. Verschut, M. Bibb, M. Bush, B. Molnar, E. Barane, M. Al-Bassam, G. Chandra, L. Song, G. Challis, M. Buttner, K. Flärdh, Developmentally regulated volatiles geosmin and 2-methylisoborneol attract a soil arthropod to *Streptomyces* bacteria promoting spore dispersal. *Nat. Microbiol.* **5**, 1–9 (2020).

134. D. A. T. Boncan, S. S. K. Tsang, C. Li, I. H. T. Lee, H.-M. Lam, T.-F. Chan, J. H. L. Hui, Terpenes and terpenoids in plants: Interactions with environment and insects. *Int. J. Mol. Sci.* **21**, 7382 (2020).

135. B. Singh, R. A. Sharma, Plant terpenes: Defense responses, phylogenetic analysis, regulation and clinical applications. *Biotech.* **5**, 129–151 (2015).

136. D. R. Houston, J. G. Hanna, J. C. Lathe, S. G. Hillier, R. Lathe, Evidence that nuclear receptors are related to terpene synthases. *J. Mol. Endocrinol.* **68**, 153–166 (2022).

137. Q. Jia, X. Chen, T. G. Köllner, J. Rinkel, J. Fu, J. Labbé, W. Xiong, J. S. Dickschat, J. Gershenzon, F. Chen, Terpene synthase genes originated from bacteria through horizontal gene transfer contribute to terpenoid diversity in fungi. *Sci. Rep*. **9**, 9223 (2019).

1.16 Acknowledgements

Chapter 1 contains material currently being prepared for publication, coauthored with Wilson, K.A., Moore, B.S., and Burkhardt, I. The dissertation author was the primary co-author of this chapter with Wilson, K.A.

CHAPTER 2. Domoic acid biosynthesis in the red alga Chondria armata suggests a complex evolutionary history for toxin production

2.1 Introduction and Context for Chapter 2

The kainoid family of natural products includes kainic acid, domoic acid, and acromelic acid, all of which act as potent ionotropic glutamate receptor antagonists.[1] In Japan, historically, kainoid-producing red macroalgae, like *Digena simplex* and *Chondria armata* have been used as anthelmintic agents, or deworming agents, to treat parasitic roundworm infections.[2] Kainic acid is still used in a laboratory setting to generate mouse model systems to study neurological diseases, such as temporal lobe epilepsy.[3] Notably, the name domoic acid originates from the Japanese word for *C. armata*, 'domoi', from which domoic acid was originally isolated in the 1950s.[3\4] The biosynthesis of acromelic acid, produced by the fungi *Clitocybe acromelalga*, is currently unknown, presenting an opportunity for future studies of this family of natural products. However, the biosynthetic pathways for domoic acid and kainic acid have been recently solved in two distantly related algal lineages: planktonic diatoms[5] and red macroalgae,[6] respectively.

Oceanic domoic acid production is mainly associated with the harmful algal bloom–forming diatom genus *Pseudo-nitzschia*. The biosynthetic pathway to isodomoic acid A was recently elucidated in the species *Pseudo-nitzschia multiseries* using differential transcriptomics, establishing the genetic basis marine kainoid production.[5] Follow-up studies characterized kainic acid biosynthesis pathways from the red macroalgae *D. simplex and Palmaria palmata*.[6] The successful sequencing and discovery of kainic acid biosynthesis gene clusters in *D*. *simplex and P*. *palmata* led to follow-up sequencing efforts with the red macroalgae, *C. armata*, which is described in the following chapter.

The work described in this chapter was published in the Proceedings of the National Academy of Sciences in 2022 and describes the sequencing and assembly of

the draft genome of *C. armata*, and the identification of several copies of the red algal domoic acid (*rad*) biosynthetic gene cluster. The *rad* gene clusters are organized similarly to the previously known diatom domoic acid biosynthesis clusters in terms of gene organization and synteny. Unlike the kainic acid producing clusters identified in the related red macroalgae *D. simplex and P. palmata*, the *C. armata* clusters include a class of enzyme known as a cytochrome P450 (CYP450), which is critical to domoic acid production. Biochemical characterization of core biosynthetic enzymes supports a slightly altered domoic acid biosynthesis model in *C. armata* via the related compound isodomoic acid B. Furthermore, phylogenetic analysis of the *rad* genes supports two modes of evolution driving kainoid biosynthesis in marine systems: native eukaryotic CYP450 neofunctionalization combined with horizontal gene transfer of core biosynthesis enzymes within each respective algal lineage.

The work presented in this chapter supports the feasibility of sequencing and characterizing red algal genomes for the discovery of natural product biosynthesis genes, as well as informs the distribution and evolutionary history of domoic acid biosynthesis from two distinct taxonomic algal lineages. This work is in many aspects a foundational study for chapters 3 and 4 of this dissertation, because of the optimization of high-molecular weight DNA extraction protocols for red macroalgae and the successful implementation of long-read, third generation sequencing for the discovery of biosynthetic gene clusters. Moreover, this chapter describes the biochemical validation of biosynthesis genes from a red macroalgae, one of only a handful of studies to date.

2.2 References for Chapter 2 Introduction

1. P. Werner, M. Voigt, K. Keinänen, W. Wisden, P. H. Seeburg, Cloning of a putative high-affinity kainate receptor expressed predominantly in hippocampal CA3 cells. *Nature* **351**, 742–744 (1991).

2. T. Higa, M. Kuniyoshi, Toxins associated with medicinal and edible seaweeds. *J. Toxicol. Toxin Rev.* **19**, 119–137 (2000).

3. X. Y. Zheng, H. L. Zhang, Q. Luo, J. Zhu, Kainic acid-induced neurodegenerative model: potentials and limitations. *J. Biomed Biotechnol.* **2011**, 457079 (2011).

4. W. H. Gerwick, Plant sources of drugs and chemicals. *Encyclopedia of Biodiversity (Second Edition*). 129–139 (2013).

5. J. K. Brunson, S. M. K. McKinnie, J. R. Chekan, J. P. McCrow, Z. D. Miles, E. M. Bertrand, V. A. Bielinski, H. Luhavaya, M. Oborník, G. J. Smith, D. A. Hutchins, A. E. Allen, B. S. Moore, Biosynthesis of the neurotoxin domoic acid in a bloom-forming diatom. *Science* **361**, 1356–1358 (2018).

6. J. R. Chekan, S. M. K. McKinnie, M. L. Moore, S. G. Poplawski, T. P. Michael, B. S. Moore, Scalable biosynthesis of the seaweed neurochemical, kainic acid. *Angew. Chem., Int. Ed.* **58**, 8454–8457 (2019).

7. M. Harizani, E. Ioannou, V. Roussis, The *Laurencia* Paradox: An endless source of chemodiversity, *Progress in the Chemistry of Organic Natural Products* **102,** 91–252 (2016).

2.3 Reprint of “Domoic acid biosynthesis in the red alga *Chondria armata* suggests a complex evolutionary history for toxin production”

Domoic acid biosynthesis in the red alga *Chondria armata* suggests a complex evolutionary history for toxin production

Taylor S. Steele[a,b,1], John K. Brunson[a,c,1], Yukari Maeno[d], Ryuta Terada[e], Andrew E. Allen[c,f], Mari Yotsu-Yamashita[d], Jonathan R. Chekan[g,2], and Bradley S. Moore[a,h,2]

[a]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92093; [b]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, CA 92093; [c]Microbial and Environmental Genomics Group, J. Craig Venter Institute, La Jolla, CA 92037; [d]Graduate School of Agricultural Science, Tohoku University, Sendai 980-8572, Japan; [e]United Graduate School of Agricultural Sciences, Kagoshima University, Kagoshima 890-0065, Japan; [f]Integrative Oceanography Division, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92037; [g]Department of Chemistry and Biochemistry, University of North Carolina Greensboro, Greensboro, NC 27412; and [h]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA 92093

Edited by Rodney Croteau, Institute of Biological Chemistry, Washington State University, Pullman, WA; received September 22, 2021; accepted December 10, 2021

Domoic acid (DA), the causative agent of amnesic shellfish poisoning, is produced by select organisms within two distantly related algal clades: planktonic diatoms and red macroalgae. The biosynthetic pathway to isodomoic acid A was recently solved in the harmful algal bloom–forming diatom *Pseudonitzschia multiseries*, establishing the genetic basis for the global production of this potent neurotoxin. Herein, we sequenced the 507-Mb genome of *Chondria armata*, the red macroalgal seaweed from which DA was first isolated in the 1950s, identifying several copies of the red algal DA (*rad*) biosynthetic gene cluster. The *rad* genes are organized similarly to the diatom DA biosynthesis cluster in terms of gene synteny, including a cytochrome P450 (CYP450) enzyme critical to DA production that is notably absent in red algae that produce the simpler kainoid neurochemical, kainic acid. The biochemical characterization of the *N*-prenyltransferase (RadA) and kainoid synthase (RadC) enzymes support a slightly altered DA biosynthetic model in *C. armata* via the congener isodomoic acid B, with RadC behaving more like the homologous diatom enzyme despite higher amino acid similarity to red algal kainic acid synthesis enzymes. A phylogenetic analysis of the *rad* genes suggests unique origins for the red macroalgal and diatom genes in their respective hosts, with native eukaryotic CYP450 neofunctionalization combining with the horizontal gene transfer of *N*-prenyltransferases and kainoid synthases to establish DA production within the algal lineages.

natural products | neurotoxin | genomics | seaweed | biosynthetic gene cluster

Harmful algal blooms of the diatom genus *Pseudonitzschia* produce high levels of domoic acid (DA), a neurotoxic glutamate receptor agonist with far-reaching food web implications because of its bioaccumulation in shellfish (1). The consumption of seafood contaminated with DA can cause acute amnesic shellfish poisoning (ASP) in humans, a malady characterized by seizures, short-term memory loss, and even death (2, 3). In 1987, the first recorded outbreak of ASP occurred on Prince Edward Island, Canada, wherein 107 people contracted the illness from eating mussels containing high levels of DA (4, 5). While oceanic DA production is primarily linked to the harmful algal bloom–forming diatom genus *Pseudonitzschia*, the compound was originally discovered in the 1950s by Daigo and coauthors from the red macroalga *Chondria armata* (6). The chemical constituents of *C. armata*, along with several other seaweeds, were of particular interest due to the historical usage of algae as anthelmintic agents in Japan, with the name "domoic acid" deriving from the word "domoi," the Japanese name for *C. armata* (6, 7). Indeed, both DA and the related red algal metabolite kainic acid (KA) display anthelmintic activity, with the latter being used to treat roundworm infections until the 1990s (8, 9). These two compounds are commonly referred to as kainoids, sharing a homologous glutamate and isoprenoid-derived pyrrolidine scaffold.

The evolutionary origin of kainoid biosynthesis in marine algae remains unclear. The production of KA is limited to red algae (Rhodophyta), and while multiple orders within Rhodophyta contain species capable of making KA, the distribution of known kainoid-producing genera is sparse (*SI Appendix*, Table S1 and Fig. S1). Meanwhile, DA production in red algae is further constrained to just one family, *Rhodomelaceae*. After the initial discovery of DA in *C. armata,* further chemical studies identified potential kainoid production in other red algae, including *Alsidium corallinum* (10, 11), *Digenea simplex* (12), *Amansia glomerata* (12), *Chondria baileyana* (13), and *Vidalia obtusiloba* (12). The ability of diatoms (Bacillariophyta) to produce DA is surprising considering no other taxa has been described to produce the neurotoxin besides Rhodophyta and Bacillariophyta, two significantly divergent clades of algae (Fig. 1*A*). Among diatoms, only

Significance

Originally isolated from the red alga *Chondria armata*, domoic acid (DA) is best known as a potent marine neurotoxin produced by oceanic harmful algal blooms of planktonic diatoms. Sequencing efforts to date of kainoid-producing red algae have focused exclusively on a closely related molecule, kainic acid, leaving a gap in the understanding of DA biosynthesis in red algae and its evolutionary linkage to diatoms. Here, we present the phylogenetic and biochemical investigation of DA biosynthesis in *C. armata*. This work demonstrates the high synteny of DA biosynthetic genes between relatively distant taxonomic groups of algae and suggests a complex evolutionary history for DA biosynthesis involving gene transfer and neofunctionalization.

Author contributions: T.S.S., J.K.B., J.R.C., and B.S.M. designed research; T.S.S., J.K.B., Y.M., M.Y.-Y., and J.R.C. performed research; R.T. contributed new reagents/analytic tools; T.S.S., J.K.B., Y.M., M.Y.-Y., J.R.C., and B.S.M. analyzed data; T.S.S., J.K.B., J.R.C., and B.S.M. wrote the paper; and A.E.A. provided helpful discussions.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]T.S.S. and J.K.B. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: jrchekan@uncg.edu or bsmoore@ucsd.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2117407119/-/DCSupplemental.

Published February 2, 2022.

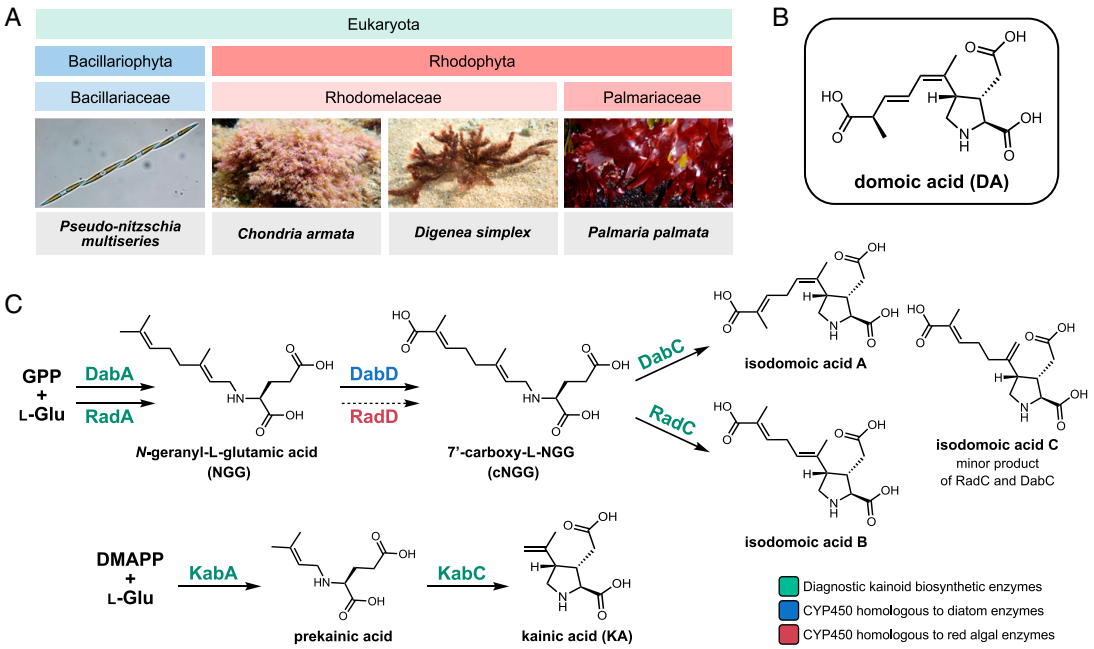


Fig. 1. Kainoid biosynthetic genes and producing organisms. (*A*) DA-producing diatom *P. multiseries* (Bacillariophyta) and KA-producing red algae *D. simplex* and *P. palmata* with DA-producing red alga *C. armata* (Rhodophyta), two significantly divergent phyla of kainoid-producing algae. (*B*) Structure of DA. (*C*) Proposed DA and KA biosynthetic pathways showing diagnostic kainoid biosynthetic enzymes in teal performing the two key enzymatic transformations forming the core kainoid pyrrolidine scaffold. Verified activities and major products from this work and previous studies (17, 18) shown with solid arrows.

the genera *Pseudonitzschia* and *Nitzschia* contain species that have been demonstrated to produce DA, although it is possible that other understudied diatoms, such as species belonging to the genus *Amphora*, may also produce the toxin (14–16). Understanding the distribution and evolutionary history of DA biosynthesis may help identify understudied emerging sources of toxicity in the marine environment.

The genes encoding DA (Fig. 1*B*) biosynthesis (*dab*) and KA biosynthesis (*kab*) were recently elucidated (Fig. 1*C*) (17, 18). Both pathways begin with the *N*-prenylation of L-glutamate by either geranyl diphosphate (GPP) or dimethylallyl diphosphate (DMAPP) to establish the linear precursors to DA and KA, respectively (19). Subsequent oxidative cyclization to build the characteristic pyrrolidine ring of either molecule is catalyzed by an α-ketoglutarate (αKG)-dependent Fe^{2+} oxidase, also known as a kainoid synthase enzyme (DabC/KabC). These two unusual enzymatic transformations are diagnostic for kainoid biosynthesis in algae, forming a core biosynthetic gene cluster and serving as bioinformatic hooks to probe sequencing data for kainoid biosynthetic pathways (Fig. 1*C*). The biosynthesis of DA, unlike that of KA, requires further oxidation to install the carboxylic acid functionality on the monoterpene-derived alkyl side chain via a cytochrome P450-catalyzed (DabD) reaction occurring prior to oxidative cyclization of the pyrrolidine. Additionally, the biosynthesis of DA also involves an uncharacterized isomerization transformation catalyzed by a yet to be identified enzyme. The discovery of the various kainoid pathways was facilitated by transcriptomic and genomic sequencing of the DA-producing diatom *Pseudonitzschia multiseries* (17) and of the KA-producing red algae *Digenea simplex* and *Palmaria palmata* (18). Notably, the kainoid biosynthetic genes from all organisms sequenced thus far appear to cocluster within each respective genome. Follow-up sequencing efforts of red algae to date have focused exclusively on organisms producing KA, leaving a gap in our understanding of the enzymology underlying DA biosynthesis in red algae.

To address this fundamental question, we sequenced the genome of *Chondria armata* and identified unique copies of a biosynthetic gene cluster suggestive of DA biosynthesis. Subsequent in vitro characterization of the *C. armata* pyrrolidine-forming enzymes further implicates their role in DA biosynthesis, leading to the designation of these sequences as red algal domoic acid (*rad*) biosynthesis genes. Comparative enzyme substrate assays, syntenic comparisons of gene cluster organization, and phylogenetic analysis provide key insights into the evolution of DA biosynthesis in diatoms and red algae. Our findings suggest a combination of horizontal gene transfer (HGT) and neofunctionalization of native enzymology served to establish neurotoxin production in the divergent algal lineages.

Results and Discussion

Genome Sequencing and Assembly. We collected *C. armata* from Kyushu Island, Japan, and sequenced its genomic DNA with a combination of Oxford Nanopore Technologies (ONT) and Illumina platforms. Genome size was estimated to be 480 Mb using the 17-mer histogram generated by Platanus v1.2.4 (20), repeat content was estimated to be 45.49%, and kmer distributions were consistent with tetraploidy (*SI Appendix*, Fig. S2). Hybrid assembly using MaSuRCA v3.4.2 (21) yielded a genome with a total size, N50, and GC-content of 507 Mb, 643 kb, and 45.34%, respectively (*SI Appendix*, Table S3). Notably, the longest assembled contig was 3.3 Mb in length. Genome contiguity and completeness was assessed with the eukaryota benchmarking universal

single-copy orthologs (BUSCO v4.0.5) database (22). Because of a lack of rhodophyta-specific datasets, we assessed the 10 publicly available red macroalgal genomes with the eukaryota_odb10 BUSCO database (23–26). We found the *C. armata* assembly to be relatively complete, containing 69% of complete, single-copy eukaryotic gene orthologs, where current published red macroalgal genomes contain on average 65% of complete, single-copy eukaryotic gene orthologs (*SI Appendix*, Table S4 and Fig. S3). To further assess overall genome completeness, we mapped ONT reads to the assembled genome using minimap2 (27), in which 88% of ONT reads were present in the assembled genome, a clearer indicator of genome completeness. Genome sequences and assembly are deposited in the National Center for Biotechnology Information (NCBI), BioProject identification PRJNA762367.

Kainoid Cluster Identification and Comparison. To identify kainoid biosynthesis genes in the *C. armata* genome, we used *dab* and *kab* genes as genetic hooks (17, 18). Previous work demonstrated the remarkably low sequence similarity of *dabA* and *kabA* genes to any publicly available protein sequence (19). tBLASTn searches with the highly conserved *N*-prenyltransferase genes initially revealed three copies of a kainoid biosynthetic gene cluster. Notably, unlike previously characterized red algal kainoid gene clusters, two of the *C. armata* clusters included a coclustered cytochrome P450 (CYP450), which is suggestive of DA biosynthesis. We refer to these clusters as red algal domoic acid (*rad*) biosynthetic gene clusters. The first *rad* cluster (*rad1*, GenBank accession: OK169902) is contained within a 7.9-kb region, and complete coding regions for all three key DA biosynthetic genes were clearly identified (Fig. 2). The second copy of the *rad* cluster (*rad2*, GenBank accession: OK169903) is contained within a shorter 7.5-kb region, and complete coding regions of an *N*-prenyltransferase and kainoid synthase were identified. However, this copy of the *rad* cluster appears to contain a degraded CYP450 with premature STOP codons. The third copy of the *rad* cluster (*rad3*, GenBank accession: OK169904) is contained within a 5.8-kb region and completely lacks a CYP450. This cluster is located 271 kb into an assembled contig, 443 kb in length. Further analysis did not reveal any

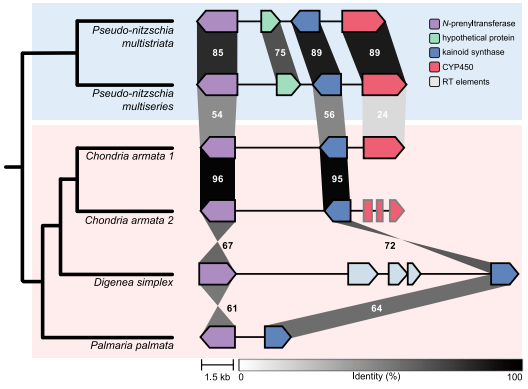


Fig. 2. Visualization of syntenic comparisons between *dab*, *rad*, and *kab* gene clusters. Blue background highlights diatom sequences, and red background highlights red algal sequences. Cladogram connects clusters based on taxonomic relationships of organisms, not to scale with respect to evolutionary time. Intensity of shading between genes is relative to the similarity of gene pairs by amino acid sequence percent identity in agreement with discrete measures of sequence percent identity as labeled. Retro transposable elements are abbreviated as RT elements.

DA CYP450 gene homologs or general CYP450 enzymes within this assembled sequence space (*SI Appendix*, Fig. S4). Curiously, despite *C. armata* appearing to be tetraploid, only three copies of the *rad* cluster were initially identified. To determine if one copy of the *rad* clusters was a collapsed paralog, the coverage of each *rad* cluster copy was calculated using mapped ONT reads. The mean coverage of *rad1* was almost double (39×) that of *rad2* (21×) and *rad3* (21×) and is indicative of a collapsed paralog, supporting the presence of all four expected genomic copies of the *rad* cluster.

Publicly available *dab* and *kab* gene clusters were used to construct a comparative analysis with new *rad* gene clusters from *C. armata* (Fig. 2) (28). Despite core kainoid biosynthetic genes being more similar by amino acid percent identity to the KA-producing red alga *D. simplex*, a relative of *C. armata* also belonging to the family *Rhodomelaceae*, all-versus-all global amino acid alignments, revealed the overall structure and organization of the *rad* gene cluster to be more like *dab* clusters from diatoms. Aside from lacking the hypothetical "DabB" protein found in the *dab* clusters, the *rad* and *dab* clusters surprisingly display nearly identical gene organization and orientation (Fig. 2). Amino acid percent identity remains consistent across the *N*-prenyltransferase (*radA*/*dabA*) and kainoid synthase (*radC*/*dabC*) enzymes, at 54 and 56%, respectively. However, despite the conservation of gene location and orientation, the amino acid percent identity between the coclustered CYP450s *radD* and *dabD* decreased to 24%.

In Vitro Validation of Biosynthetic Gene Function. In addition to DA, several additional isomers have also been isolated from *C. armata* and other sources (29–31). Although production of isodomoic acid A, B, and C have all been described in *Pseudo-nitzschia*, these isomers are found at substantially lower levels than DA in the diatom species and isolates studied thus far (14, 32). A similar trend has also been observed with *C. armata* (33). However, high-resolution liquid chromatography–mass spectrometry (LC-MS) analysis of methanolic algal extracts suggests that isodomoic acid B is an especially abundant DA isomer in our independently collected *C. armata* isolate (Fig. 3*A*) (*SI Appendix*, Fig. S5). We hypothesized that the differential abundance of DA isomers in *C. armata* extracts is linked to the activities of one or more kainoid synthase enzymes expressed by the seaweed, a hypothesis consistent with our previous observations that the kainoid synthase DabC can generate structural diversity (17).

To determine the basis for differences in isomer production, we examined the newly identified *rad1* cluster from *C. armata*, beginning with kainoid synthase. Full-length RadC1 from the intact *rad1* cluster was expressed as an N-terminal hexahistidine (N-His_6)–soluble construct in *Escherichia coli* (*SI Appendix*, Fig. S6). After affinity chromatography purification, RadC1 was observed to perform oxidative cyclization on the linear precursor cNGG, similar to the previously described activity for DabC (17). Overnight RadC1 and DabC reactions both proceed to near-complete substrate consumption when cNGG was added. However, RadC1 makes isodomoic acid B as its major cyclization product, whereas DabC produces isodomoic acid A (Fig. 3). While isodomoic acid B production by RadC1 is consistent with the relative isomer abundance observed in our *C. armata* extracts, the presence of isodomoic acid A and C in the algae is yet to be explained biochemically. Although trace quantities of isodomoic acid C are observed from the RadC1 reaction, it remains to be seen whether stereoisomer diversity is also enriched upstream of cyclization through incorporation of neryl diphosphate. Initial expression, purification, and enzymatic assay of RadC2, encoded in the *rad2* cluster, did not reveal substantial activity toward cNGG, although small quantities of isodomoic acid B are produced. As the *rad2* cluster also contains a degraded

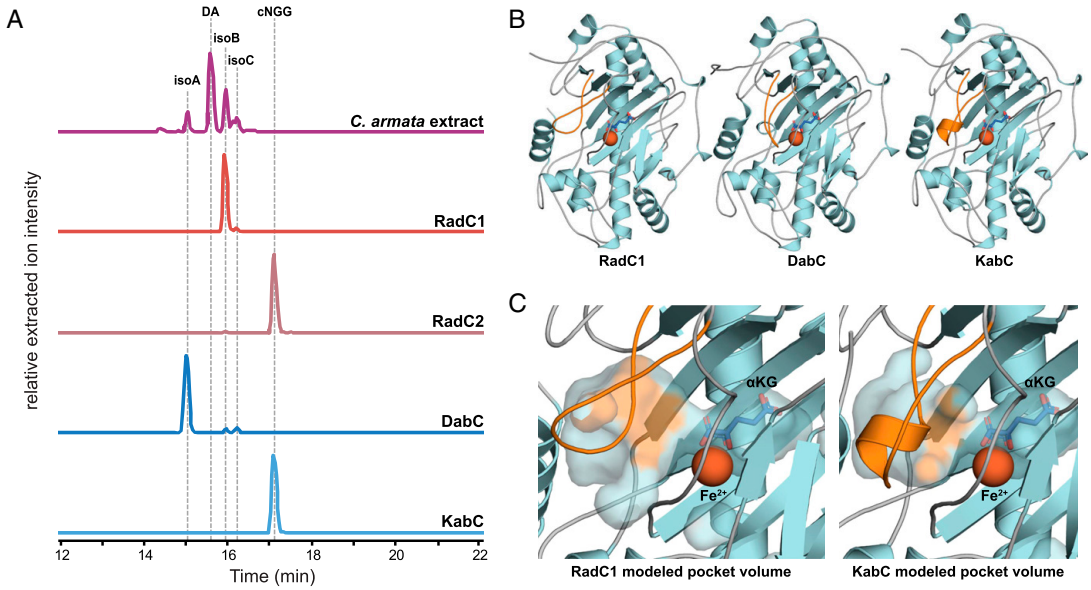


Fig. 3. Activities and predicted structures of the known kainoid synthase enzymes. (*A*) Comparison of combined extracted ion chromatogram profiles for cNGG substrate (*m/z* 312.1 ± 0.2) and isodomoic acid (*m/z* 310.1 ± 0.2) from RadC1, RadC2, DabC, and KabC assays and *C. armata* extract. Relative intensity of extracted ions is shown, and all substrates and observed DA isomers are compared to purified standards (*SI Appendix*, Fig. S5). (*B*) AlphaFold2 predicted models for RadC1, DabC, and KabC. (*C*) Modeled pocket volume for RadC1 and KabC.

copy of the coclustered RadD CYP450 with several premature STOP codons, we suspect that this second cluster may be undergoing pseudogenization and that the encoded enzymes may not function in vivo. The third copy of the kainoid synthase enzyme RadC3 displays 100% amino acid sequence identity to RadC2 and is assumed to be similarly inactive.

Both RadC1 and DabC generate a similar series of dainic acid molecules, a set of higher isoprene homologs of kainic acid, from the less-favored NGG substrate. RadC1 appears to approach complete substrate turnover in our preliminary, qualitative in vitro analysis, whereas DabC does not (*SI Appendix*, Fig. S7) (17). Because RadC1 is more closely related in amino acid sequence to the KA-forming KabC enzyme than the diatom DabC enzyme, the substrate specificity of KabC from *D. simplex* (DsKabC, GenBank accession: QCC62383.1) was also further investigated. Intriguingly, KabC does not catalyze cyclization on either the cNGG or NGG precursor (Fig. 3 and *SI Appendix*, Fig. S7). As demonstrated previously, KabC efficiently cyclizes the shorter prekainic acid substrate to KA, and this activity can also be seen in this study for RadC and DabC, although neither of the latter reactions proceed to completion (*SI Appendix*, Fig. S8) (18). This inability of KabC to catalyze cyclization on the longer cNGG and NGG substrates represents a marked difference between the DA isomer–forming RadC and DabC versus the KA-forming KabC. While no obvious single residue can be attributed to these differences in substrate specificity, structural prediction via AlphaFold2 (34, 35) reveals KabC enzyme models to have an extended loop that pushes further into the active site than in RadC or DabC models (Fig. 3*B*). The quantification of active site area and volume solvent–accessible surface area using CASTp (36) identified an overall decrease in KabC modeled pocket volume (115 Å^3) when compared to RadC (208 Å^3) and DabC (262 Å^3) models (Fig. 3*C* and *SI Appendix*, Fig. S9). This follows an expected trend according to each enzyme's observed substrate specificities (Fig. 3*A*).

The glutamate *N*-prenyltransferase enzyme RadA1 was also expressed as a soluble construct following a seven amino acid N-terminal truncation and fusion to maltose-binding protein (MBP) containing an N-His$_6$- tag (*SI Appendix*, Fig. S10). Initial overnight assays with purified MBP-Δ7-RadA1 suggest both GPP and DMAPP can be accepted as substrates to make NGG and prekainic acid, respectively. This observation contrasts the selectivity previously shown by DabA and, to a lesser extent, KabA (17, 18). Ongoing exhaustive experiments on RadA kinetics and structure will provide additional information regarding potential prenyl group promiscuity to compare directly with previously described DabA and KabA activity (19). Overall, the ability of RadA1 and RadC1 to produce DA intermediates and isomers, similar to what was observed previously in diatoms, supports their role in overall biosynthesis of DA in *C. armata* (17).

Phylogenetic Analysis. Previous efforts have attempted to describe the evolutionary history of the glutamate *N*-prenyltransferase enzyme present in both DA and KA biosynthesis (19). Despite structural homology to known bacterial terpene cyclases, the enzyme does not clade well with any extant terpene cyclase because of its unique amino acid sequence (*SI Appendix*, Fig. S11). To further explore the evolutionary history of the remaining DA and KA biosynthetic genes, maximum likelihood (ML) phylogenetic trees were constructed for both the αKG-dependent kainoid synthases (DabC/KabC/RadC) and the CYP450s (DabD/RadD). Beginning with RadC, initial BLAST searches revealed few closely related homologs outside of known KA or DA biosynthetic genes in the NCBI protein database. The closest homologs were primarily from bacteria, including the well-studied αKG-dependent oxidase deacetoxycephalosporin C synthase (*cefE*, UniProtKB/Swiss-Prot identification: P18548.1) in *Streptomyces clavuligerus*, an enzyme that performs the penicillin to cephalosporin ring expansion. To ensure a broad understanding of kainoid synthase enzyme phylogeny in relation to other

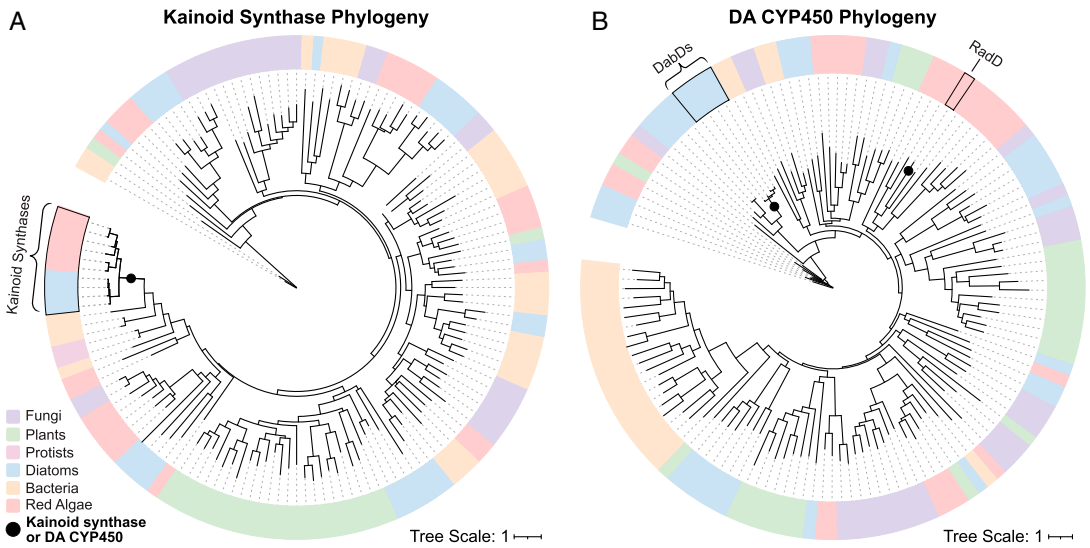


Fig. 4. Phylogenetic analysis of (*A*) kainoid synthase and (*B*) coclustered DA CYP450 enzymes (circle denotes key branch points). The maximum likelihood trees were built using representative sequences from key taxonomic groups (*SI Appendix*, Tables S5 and S6). Kainoid synthase enzymes form their own distinct branch, independent of taxonomic origin, while the DA CYP450s are nested within their respective taxonomic groups.

αKG-dependent Fe^{2+}-containing oxidases, representative sequences of key taxonomic groups were generated from the isopenicillin N synthase–like InterPro family (IPR027443). These representative sequences, in addition to top BLAST hits and other known kainoid synthases, were used to construct a ML phylogenetic tree, with all sequence accession numbers listed in *SI Appendix*, Table S5. This process of using BLAST hits and representative InterPro family sequences to construct a ML tree was repeated with RadD. Unlike RadC, BLAST searches using the hypothetically assigned RadD revealed several significant alignments with other red algal CYP450s. To ensure a well-populated phylogeny, representative sequences from the cytochrome P450 InterPro family (IPR001128) were generated from the same taxonomic groups as before (*SI Appendix*, Table S6). These sequences, in addition to representative diatom P450 sequences from publicly available transcriptomics and top BLASTp hits for RadD and DabD, were used to construct a ML phylogenetic tree (37).

Based on our phylogenetic analysis, kainoid synthases form their own distinct branch independent of taxonomic origin (Fig. 4*A*), a finding that is consistent with previous phylogenetic studies on *N*-prenyltransferases (19). Notably, the bacterial enzyme cefE is the only well-characterized close relative to RadC. The monophyletic clading of both the *N*-prenyltransferase enzymes and kainoid synthase enzymes from these distantly related taxa suggests a possible horizontal gene transfer event. This observation is further supported by high levels of overall sequence identity between the two enzymatic groups. On the other hand, the CYP450 enzymes are independently nested within their respective taxonomic groups, an observation that supports convergent evolution toward the installation of the carboxylic acid on the prenyl sidechain (Fig. 4*B*). Based on this taxonomic clustering, the presence of a CYP450 enzyme in DA biosynthesis may be a result of gene duplication and neofunctionalization of P450 enzymes, a phenomenon that has been previously described in plants (38, 39). This hypothesis is further strengthened by the presence of additional CYP450s closely related to RadD or DabD found within the genome of *C. armata* or *P. multiseries*, respectively (*SI Appendix*, Figs. S12 and S13). The evolutionary history of CYP450 families in diatoms has been the subject of previous research, with diatoms displaying substantial P450 gene family acquisition and loss despite encoding relatively few CYP450 genes compared to land plants (40). The diatom DabD protein bears the most homology to other CYP450 enzymes found in diatoms, including *Thalassiosira pseudonana* and *Fragilariopsis cylindrus* (bootstrap > 99%), suggestive of an ancient acquisition of this CYP450 family in the diatom lineage (*SI Appendix*, Fig. S12). The biochemical function of these related enzymes is unknown, and the specific CYP450 family classification of DabD-like enzymes remains to be determined.

Conclusions and Discussion

Our discovery of the DA biosynthetic machinery within the genome of the red alga *C. armata*, the first-described DA producer, provides crucial insight into the evolutionary history of kainoid biosynthesis in diatoms and red algae. The overall sparse distribution of kainoid biosynthetic genes in these two distantly related taxa and conservation of gene synteny in *C. armata* and *Pseudonitzschia* suggests a possible HGT event. However, direct HGT of the entire gene cluster from red algae to diatoms, or vice versa, is not likely in modern ecological contexts. Alternatively, many genes in diatoms and other microalgae were indeed acquired from red algae in distant evolutionary history through endosymbiotic gene transfer (EGT). Endosymbiosis of unicellular red algae played a major role in the evolution of the diatom clade, and several red algal genes were transferred to the host nuclear genome via EGT to contribute to the diverse origins of diatom genomes (41, 42). However, an ancient EGT event moving the DA biosynthetic gene cluster from red algae to diatoms is unlikely given the limited distribution of kainoid biosynthetic enzymes in modern diatom genera, whereas genes acquired via EGT are typically widespread in the diatom lineage. Nevertheless, the lack of clear homology of the glutamate

N-prenyltransferase (RadA, KabA, and DabA) and kainoid synthase (RadC, KabC, and DabC) enzymes to known algal proteins suggests HGT from a potential third host as a mechanism for the acquisition of these kainoid biosynthetic enzymes. Furthermore, the observation that the enzymes RadD and DabD are most like native CYP450s within the red algal and diatom genomes, respectively, supports a history invoking two modes of evolution toward DA biosynthesis: HGT combined with gene duplication and neofunctionalization of native biochemistry.

From our biochemically supported phylogenetic analysis, it appears that the glutamate *N*-prenyltransferase and kainoid synthase genes form a core biosynthetic gene cluster and were acquired via HGT from an unknown source. These specific chemical transformations are unprecedented in biochemistry outside of diatom and red algal kainoid biosynthesis and do not seem to be repurposed enzymology from within the respective algal lineages as evidenced by their low similarity to terpene cyclases and αKG-dependent Fe^{2+} oxidases from public algal genomes and transcriptomes. These enzymes may be bacterial in origin due to the structural homology of the *N*-prenyltransferase to bacterial terpene cyclases and the modest amino acid sequence homology of the kainoid synthase to known bacterial enzymes, although no strong candidates for bacterial kainoid biosynthetic enzymes have been found to date (19). Notably, HGT from bacteria is well described in diatoms and red algae. Diatoms especially have been the subject of substantial genome sequencing efforts, revealing that ∼5% of their functional genes have been acquired from bacteria in relatively recent evolutionary history (43, 44). The acquisition of HGT-derived bacterial genes, such as urea cycle enzymes and proteorhodopsins, has contributed to the unique biochemistry and subsequent ecological success of the diatom lineage (45, 46). Similarly, HGT is an important mechanism for the expansion of genes related to both primary and secondary metabolism in red algae. The extremophilic red alga *Galdieria sulphuraria* is estimated to have acquired at least 5% of protein-coding genes through HGT (47). These HGT-derived genes typically confer an adaptive advantage to the organism such as in the genus *Pyropia*, wherein an HGT-derived carbonic anhydrase gene acts as a key component of the carbon-concentrating mechanism and aids in overcoming carbon limitations (48). Red algae also seem to have acquired other terpene cyclases via HGT from bacteria, although these enzymes appear unrelated to the *N*-prenyltransferase enzymes in DA and KA biosynthesis (49). HGT has also been suggested to play a role in the evolution of saxitoxin biosynthesis. Saxitoxin, another notable HAB neurotoxin and the causative agent of paralytic shellfish poisoning, is produced by prokaryotic freshwater cyanobacteria and eukaryotic marine dinoflagellates (50). Genes encoding saxitoxin biosynthesis were first identified in the cyanobacterium *Cylindrospermopsis raciborskii* and later in dinoflagellates belonging to the genus *Alexandrium*, suggesting a potentially ancient HGT event from cyanobacteria to dinoflagellates based on conservation of diagnostic domain architectures between key biosynthetic enzymes (51–53).

Following acquisition of the core kainoid biosynthetic genes, both *C. armata* and *Pseudonitzschia* spp. appear to have repurposed CYP450 enzymology to install the characteristic carboxylic acid on the terpene-derived DA side chain important for toxicity (54). Enzymes within the CYP450 family catalyze a diversity of chemical transformations, and conversion of a methyl group to a carboxylic acid is a well-established biotransformation with precedence in diverse eukaryotic and prokaryotic organisms (55). Although the activity of RadD1 remains to be confirmed in vitro, the high degree of synteny between the *dab* and *rad1* clusters suggests that the red algal RadD1 fulfills a similar role to DabD in diatoms. The observed activity toward the cNGG substrate in the RadC1 and DabC systems also supports the hypothesis that a common sequence of DA biosynthetic reactions is shared between diatoms and red algae, with CYP450-catalyzed oxidation to a carboxylic acid occurring prior to cyclization in both organisms (17). However, until RadD is experimentally characterized, we recognize that this reaction series may differ in red algae given the apparent activity of RadC1 toward NGG. Regardless of reaction order, we hypothesize that a similar biosynthetic reaction toward installing a carboxylic acid on the DA scaffold has evolved independently in diatoms and red algae. An analogous independent neofunctionalization of CYP450 activities has been previously implicated in the evolution of the plant hormone diosgenin, a plant steroid saponin with a distinctive CYP450-installed spiroketal motif that is not limited to a specific taxon of plants but is instead broadly and sparsely distributed (38). An additional example of CYP450-driven convergent evolution is seen in the furanocoumarin pathway, also in plants (39). The independent selection for these enzyme activities in unrelated plants provides precedence for the neofunctionalization of CYP450s to install similar chemical features in distantly related organisms, a phenomenon that may have occurred in the evolution toward DA biosynthesis as reflected by the red algal–derived RadD and diatom-derived DabD.

Evolution toward DA production is further supported by the apparent substrate scope of the different kainoid synthases. Although RadC1 is more like the red algal KabC in its amino acid sequence, RadC1 displays an activity more similar to the diatom enzyme DabC in its ability to convert both NGG and cNGG to DA-like molecules. These substrates are not similarly cyclized by the KA-forming DsKabC. Despite the similar activities exhibited by the *C. armata* and *Pseudonitzschia* kainoid synthases, the difference in DA isomers generated by RadC1 (isodomoic acid B) and DabC (isodomoic acid A) raises additional biosynthetic questions. Notably, the 1,3-olefin isomerization step required to make DA itself continues to remain elusive (17, 29). Because isodomoic acid B would require both a 1,3-isomerization and a separate *trans* to *cis* isomerization of the second olefin closest to the kainoid ring, it is worth reconsidering whether the key 1,3-isomerization occurs before or after kainoid synthase cyclization to create the most abundant, canonical DA isomer in diatoms and red algae. A similar 1,3-isomerization step occurs immediately following the triple hydroxylation of a methyl group to a carboxylic acid in lysergic acid biosynthesis, a process thought to be carried out by one or more CYP450 enzymes in the ergot fungus *Claviceps purpurea* (56). However, such an olefin isomerization was not revealed by previous studies on DabD expressed in *Saccharomyces cerevisiae* (17). Meanwhile, hypothetical protein DabB identified from the diatom biosynthetic cluster is missing from the respective *C. armata* gene cluster and does not appear to be encoded within the sequenced *C. armata* genome. In any case, the responsible isomerase remains unidentified. Ongoing investigations of candidate genes present in these two distantly related species, including further study of the activities exhibited by the coclustered CYP450 enzymes DabD and hypothetically assigned RadD, will shed additional light on this remaining piece of the DA biosynthetic puzzle.

The apparent evolution of kainoid synthase substrate selectivity in combination with evidence for CYP450 neofunctionalization and coclustering suggests a potential advantage afforded by DA biosynthesis over KA biosynthesis. While DA is chemically distinct from KA and has enhanced bioactivity against ionotropic glutamate receptors, the true ecological function of DA has remained elusive (57). Previous studies have demonstrated limited trace metal chelation by DA, affording a potential selective advantage for diatoms thriving in iron-limited coastal regimes (58, 59). Other work has suggested a role for DA in grazer defense, particularly with respect to copepod grazing in the diatom system (60). However, no studies to date have provided a concrete ecological mechanism beyond reasonable doubt in both diatoms and red algae. While DA could conceivably

serve different functions in the red alga *C. armata* and *Pseudonitzschia* diatoms, the evolutionary history described here provides insight to the ecological importance of this marine biotoxin.

Materials and Methods

SI Appendix includes methods for expression, purification, and mutagenesis of all enzymes used in this study. Methodologies for activity assays, sequencing, and phylogenetics are also detailed. Finally, general chemical procedures and small molecule characterization are available in *SI Appendix*. The remaining data and methods are present in *SI Appendix*. The NCBI accession numbers of sequences deposited and used in this study are listed in *SI Appendix*, Table S7.

Data Availability. BioProject, BioSample, Genome, Sequence Read Archive (SRA), and GenBank data have been deposited in NCBI (PRJNA762367, SAMN21392177, JAIWHZ000000000, SRR15927350, SRR15927349, SRR15927348, OK169902, OK169903, OK169904). All other study data are included in the article and/or *SI Appendix*. Previously published data were used for this work (DOI: 10.1126/science.aau0382, https://doi.org/10.1002/anie.201902910).

ACKNOWLEDGMENTS. We thank T. Michael (Salk Institute for Biological Studies) for assistance with DNA sequencing, G. J. Smith (Moss Landing Marine Laboratories) for providing the diatom image, T. Teruya (University of Ryukyus) and G. Saunders (University of New Brunswick) for providing seaweed images, and G. Rouse and T. Fallon (Scripps Institution of Oceanography) for helpful discussions. We acknowledge R. Botts (Point Loma Nazarene University) for his valuable mentorship and guidance. This research was supported by the National Oceanic and Atmospheric Administration (NA19NOS4780181 to B.S.M. and A.E.A.), the NSF through the Graduate Fellowship Research Program to T.S.S., the NIH (F31ES030613 to J.K.B.), the Simons Foundation Fellowship of the Life Sciences Research Foundation to J.R.C., and the University of North Carolina at Greensboro research startup funds to J.R.C.

1. V. L. Trainer *et al.*, *Pseudo-nitzschia* physiological ecology, phylogeny, toxicity, monitoring and impacts on ecosystem health. *Harmful Algae* **14**, 271–300 (2012).
2. T. M. Perl *et al.*, Amnesic shellfish poisoning: A new clinical syndrome due to domoic acid. *Can. Dis. Wkly. Rep.* **16**, 7–8 (1990).
3. O. M. Pulido, Phycotoxins by harmful algal blooms (HABs) and human poisoning: An overview. *Int. Clin. Pathol. J.* **2**, 145–152 (2016).
4. J. L. C. Wright *et al.*, Identification of domoic acid, a neuroexcitatory amino acid, in toxic mussels from eastern Prince Edward Island. *Can. J. Chem.* **67**, 481–490 (1989).
5. S. S. Bates *et al.*, Pennate diatom *Nitzschia pungens* as the primary source of domoic acid, a toxin in shellfish from eastern Prince Edward Island, Canada. *Can. J. Fish. Aquat. Sci.* **46**, 1203–1215 (1989).
6. K. Daigo, Studies on the constituents of *Chondria armata*. II: Isolation of an anthelmintical constituent. *Yakugaku Zasshi* **79**, 353–356 (1959).
7. W. H. Gerwick, "Plant sources of drugs and chemicals" in *Encyclopedia of Biodiversity*, S. A. Levin, Ed. (Elsevier, vol. **2**, 2013), pp. 129–139.
8. Y. Komiya, A. Kobayashi, Techniques applied in Japan for the control of *Ascaris* and hookworm infections – A review. *Jpn. J. Med. Sci. Biol.* **18**, 1–17 (1965).
9. J. F. Tremblay, Shortage of kainic acid hampers neuroscience research. *Chem. Eng. News Arch.* **78**, 14–15 (2000).
10. G. Impellizzeri *et al.*, Amino acids and low-molecular-weight carbohydrates of some marine red algae. *Phytochemistry* **14**, 1549–1557 (1975).
11. Y. Shimizu *et al.*, Dinoflagellate and other microalgal toxins: Chemistry and biochemistry. *Pure Appl. Chem.* **61**, 513–516 (1989).
12. M. Sato *et al.*, Distribution of neuroexcitatory amino acids in marine algae. *Phytochemistry* **42**, 1595–1597 (1996).
13. M. V. Laycock, A. S. W. de Freitas, J. L. C. Wright, Glutamate agonists from marine algae. *J. Appl. Phycol.* **1**, 113–122 (1989).
14. S. S. Bates, K. A. Hubbard, N. Lundholm, M. Montresor, C. P. Leaw, *Pseudo-nitzschia*, *Nitzschia*, and domoic acid: New research since 2011. *Harmful Algae* **79**, 3–43 (2018).
15. D. G. Mann *et al.*, Ripe for reassessment: A synthesis of available molecular data for the speciose diatom family Bacillariaceae. *Mol. Phylogenet. Evol.* **158**, 106985 (2021).
16. B. C. Dhar *et al.*, Molecular detection of a potentially toxic diatom species. *Int. J. Environ. Res. Public Health* **12**, 4921–4941 (2015).
17. J. K. Brunson *et al.*, Biosynthesis of the neurotoxin domoic acid in a bloom-forming diatom. *Science* **361**, 1356–1358 (2018).
18. J. R. Chekan *et al.*, Scalable biosynthesis of the seaweed neurochemical, kainic acid. *Angew. Chem. Int. Ed. Engl.* **58**, 8454–8457 (2019).
19. J. R. Chekan, S. M. K. McKinnie, J. P. Noel, B. S. Moore, Algal neurotoxin biosynthesis repurposes the terpene cyclase structural fold into an *N*-prenyltransferase. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 12799–12805 (2020).
20. R. Kajitani *et al.*, Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nat. Commun.* **10**, 1702 (2019).
21. A. V. Zimin *et al.*, Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27**, 787–792 (2017).
22. M. Seppey, M. Manni, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).
23. J. Collén *et al.*, Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5247–5252 (2013).
24. J. Lee *et al.*, Analysis of the draft genome of the red seaweed *Gracilariopsis chorda* provides insights into genome size evolution in Rhodophyta. *Mol. Biol. Evol.* **35**, 1869–1886 (2018).
25. S. Jia *et al.*, High-quality de novo genome assembly of *Kappaphycus alvarezii* based on both PacBio and HiSeq sequencing. bioRxiv [Preprint] (2020). https://doi.org/10.1101/2020.02.15.950402. Accessed 16 June 2021.
26. S. H. Brawley *et al.*, Insights into the red algae and eukaryotic evolution from the genome of *Porphyra umbilicalis* (Bangiophyceae, Rhodophyta). *Proc. Natl. Acad. Sci. U.S.A.* **114**, E6361–E6370 (2017).
27. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
28. C. L. M. Gilchrist, Y.-H. Chooi, Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).
29. J. Clayden, B. Read, K. R. Hebditch, Chemistry of domoic acid, isodomoic acids, and their analogues. *Tetrahedron* **61**, 5713–5724 (2005).
30. M. Meda *et al.*, Structures of isodomoic acids A, B and C, novel insecticidal amino acids from the red alga *Chondria armata*. *Chem. Pharm. Bull. (Tokyo)* **34**, 4892–4895 (1986).
31. L. Zaman *et al.*, Two new isomers of domoic acid from a red alga, *Chondria armata*. *Toxicon* **35**, 205–212 (1997).
32. A. Lelong, H. Hégaret, P. Soudant, S. S. Bates, *Pseudo-nitzschia* (Bacillariophyceae) species, domoic acid and amnesic shellfish poisoning: Revisiting previous paradigms. *Phycologia* **51**, 168–216 (2012).
33. Y. Maeno *et al.*, Six domoic acid related compounds from the red alga, *Chondria armata*, and domoic acid biosynthesis by the diatom, *Pseudo-nitzschia multiseries*. *Sci. Rep.* **8**, 356 (2018).
34. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
35. M. Mirdita, S. Ovchinnikov, M. Steinegger, ColabFold – Making protein folding accessible to all. bioRxiv [Preprint] (2021). https://doi.org/10.21203/rs.3.rs-1032816/v1. Accessed 8 September 2021.
36. W. Tian, C. Chen, X. Lei, J. Zhao, J. Liang, CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res.* **46**, W363–W367 (2018).
37. P. J. Keeling *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.* **12**, e1001889 (2014).
38. B. Christ *et al.*, Repeated evolution of cytochrome P450-mediated spiroketal steroid biosynthesis in plants. *Nat. Commun.* **10**, 3206 (2019).
39. C. Villard *et al.*, A new P450 involved in the furanocoumarin pathway underlies a recent case of convergent evolution. *New Phytol.* **231**, 1923–1939 (2021).
40. L. Teng *et al.*, Diversity and evolution of cytochromes P450 in stramenopiles. *Planta* **249**, 647–661 (2019).
41. H. Qiu, H. S. Yoon, D. Bhattacharya, Algal endosymbionts as vectors of horizontal gene transfer in photosynthetic eukaryotes. *Front Plant Sci* **4**, 366 (2013).
42. D. Moreira, P. Deschamps, What was the real contribution of endosymbionts to the eukaryotic nucleus? Insights from photosynthetic eukaryotes. *Cold Spring Harb. Perspect. Biol.* **6**, a016014 (2014).
43. E. Vancaester, T. Depuydt, C. M. Osuna-Cruz, K. Vandepoele, Comprehensive and functional analysis of horizontal gene transfer events in diatoms. *Mol. Biol. Evol.* **37**, 3243–3257 (2020).
44. C. Bowler *et al.*, The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239–244 (2008).
45. A. E. Allen *et al.*, Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**, 203–207 (2011).
46. A. Marchetti, D. Catlett, B. M. Hopkinson, K. Ellis, N. Cassar, Marine diatom proteorhodopsins and their potential role in coping with low iron availability. *ISME J.* **9**, 2745–2748 (2015).
47. G. Schönknecht *et al.*, Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* **339**, 1207–1210 (2013).
48. D. Wang *et al.*, *Pyropia yezoensis* genome reveals diverse mechanisms of carbon acquisition in the intertidal environment. *Nat. Commun.* **11**, 4028 (2020).
49. G. Wei *et al.*, Terpene biosynthesis in red algae is catalyzed by microbial type but not typical plant terpene synthases. *Plant Physiol.* **179**, 382–390 (2019).
50. K. D. Cusick, G. S. Sayler, An overview on the marine neurotoxin, saxitoxin: Genetics, molecular targets, methods of detection and ecological functions. *Mar. Drugs* **11**, 991–1018 (2013).
51. R. Kellmann *et al.*, Biosynthetic intermediate analysis and functional homology reveal a saxitoxin gene cluster in cyanobacteria. *Appl. Environ. Microbiol.* **74**, 4044–4053 (2008).

52. A. Stüken *et al.*, Discovery of nuclear-encoded genes for the neurotoxin saxitoxin in dinoflagellates. *PLoS One* **6**, e20096 (2011).
53. R. J. S. Orr, A. Stüken, S. A. Murray, K. S. Jakobsen, Evolution and distribution of saxitoxin biosynthesis in dinoflagellates. *Mar. Drugs* **11**, 2814–2828 (2013).
54. Y. Maeno *et al.*, Preparation of domoic acid analogues using a bioconversion system, and their toxicity in mice. *Org. Biomol. Chem.* **19**, 7894–7902 (2021).
55. X. Zhang, S. Li, Expansion of chemical space for natural products by uncommon P450 reactions. *Nat. Prod. Rep.* **34**, 1061–1089 (2017).
56. T. Haarmann, I. Ortel, P. Tudzynski, U. Keller, Identification of the cytochrome P450 monooxygenase that bridges the clavine and ergoline alkaloid pathways. *ChemBioChem* **7**, 645–652 (2006).
57. J. A. Larm, P. M. Beart, N. S. Cheung, Neurotoxin domoic acid produces cytotoxicity via kainate- and AMPA-sensitive receptors in cultured cortical neurones. *Neurochem. Int.* **31**, 677–682 (1997).
58. C. G. Trick *et al.*, Iron enrichment stimulates toxic diatom production in high-nitrate, low-chlorophyll areas. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 5887–5892 (2010).
59. E. Rue, K. Bruland, Domoic acid binds iron and copper: A possible role for the toxin produced by the marine diatom *Pseudo-nitzschia*. *Mar. Chem.* **76**, 127–134 (2001).
60. S. Harðardóttir *et al.*, Transcriptomic responses to grazing reveal the metabolic pathway leading to the biosynthesis of domoic acid and highlight different defense strategies in diatoms. *BMC Mol. Biol.* **20**, 7 (2019).

Supplementary Information for

Domoic acid biosynthesis in the red alga *Chondria armata* suggests a complex evolutionary history for toxin production

Taylor S. Steele[a,b,1], John K. Brunson[a,c,1], Yukari Maeno[d], Ryuta Terada[e], Andrew E. Allen[c,f], Mari Yotsu-Yamashita[d], Jonathan R. Chekan[g,2], Bradley S. Moore[a,h,2]

[1]These authors contributed equally

[2]To whom correspondence may be addressed. Email: jrchekan@uncg.edu and bsmoore@ucsd.edu

This PDF file includes:

Supplementary text
Figures S1 to S14
Tables S1 to S7
SI References

Table of Contents

1. **General Methods/Chemical Methods**

2. **Molecular Biology/Biochemical Methods**

3. **Supplementary Tables**
   - Table S1. Known KA and DA producers
   - Table S2. Primers used in this study
   - Table S3. QUAST assessment of *Chondria armata* assembly
   - Table S4. BUSCO assessment of red macroalgal genomes
   - Table S5. Kainoid synthase representative sequences
   - Table S6. CYP450 representative sequences
   - Table S7. Accession numbers of sequences deposited and used in this study

4. **Supplementary Figures**
   - Figure S1. Cladogram of DA and KA producers
   - Figure S2. 17-mer histogram
   - Figure S3. Graphic of BUSCO assessment of red macroalgal genomes
   - Figure S4. Visual representation of *rad* clusters and coverage
   - Figure S5. Domoic acid isomers in *C. armata*
   - Figure S6. 10% SDS-PAGE gel
   - Figure S7. NGG substrate screen
   - Figure S8. prekainic acid substrate screen
   - Figure S9. CASTp output of modeled pocket volumes
   - Figure S10. RadA1 substrate screen
   - Figure S11. N-prenyltransferase enzyme maximum-likelihood phylogenetic tree
   - Figure S12. Expanded diatom CYP450 maximum-likelihood phylogenetic tree
   - Figure S13. Expanded *C. armata* CYP450 maximum-likelihood phylogenetic tree
   - Figure S14. ^{1}H NMR spectrum of isodomoic acid C

5. **Supplementary References**

Methods

General Methods/Chemical Methods

General

All chemicals and solvents were used as received from the commercial supplier (SigmaAldrich or Fisher). Domoic acid and kainic acid were purchased from commercial suppliers (National Research Council of Canada or Chem-Impex, respectively). All protein purification from *Escherichia coli* was performed on an ÄKTA Pure 25 L1 instrument (Cytiva) with a fraction collector F9-C and sample pump S9 with all solvents filtered through a nylon 0.2 µm GDWP membrane (Merck) prior to use. FPLC data was analyzed with UNICORN version 7 software. All protein quantification was done by method of Bradford using the Protein Assay Dye Reagent Concentrate (Bio-Rad) on protein sample dilutions in MilliQ water.

High resolution liquid chromatography mass spectrometry (HRMS) measurements were carried out on an Agilent Technologies 1200 Series system with a diode-array detector coupled to an Agilent Technologies 6530 accurate-mass Q-TOF LCMS run in negative ionization mode. Compounds were separated by reversed-phase chromatography on a Phenomenex Kinetex 5 mm C18 100 Å 150 x 4.6 mm LC column with water + 0.1% formic acid (solvent A) and acetonitrile + 0.1% formic acid (solvent B) as eluents. Two LC methods were used for separation of all compounds for HRMS. For LC Method A, the following gradient was applied at a flow rate of 0.75 mL/min: hold at 5% B for 1 minute, 5% to 35% B over 30 min, 35 to 100% B over 1 minute, hold at 100% B for 1.5 min, 100% to 5% B over 2.5 min, hold at 5% B for 2 min. For LC Method B, the following gradient was applied at a flow rate of 0.75 mL/min: hold at 5% B for 1 minute, 5% to 35% B over 15 min, 35 to 100% B over 1 minute, hold at 100% B for 1.5 min, 100% to 5% B over 2.5 min, hold at 5% B for 2 min.

Preparation of substrates and standards

Preparation of all non-commercial substrates and standards has been described elsewhere in the chemical literature (1–3). Briefly, the chemical syntheses of *N*-geranyl glutamate (NGG), 7'-carboxy-*N*-gernayl glutamate (cNGG), prekainic acid, dimethyl-allyl diphosphate (DMAPP), and geranyl diphosphate (GPP) were performed as previously described (1, 2). Enzyme biocatalytic synthesis of dainic acid A, dainic acid C, and isodomoic acid A was performed as previously described (1). Purification of isodomoic acid B from *Chondria armata* was performed as previously described (3). Isodomoic acid C was co-purified with isodomoic acid B using the previously described methodology for isodomoic acid B. The two compounds were separated using a Mightysil RP-18 GP Aqua column (4.6 × 250 mm, 5 µm, Kanto Chemical, Tokyo, Japan) with water–acetonitrile–formic acid (90 : 10 : 0.1, v/v/v). Proton nuclear magnetic resonance (^{1}H-NMR) measurement of isodomoic acid C was performed in D_2O at 600 MHz (Figure S14). The signal of the residual MeOD was adjusted at δ 3.30 ppm as the internal reference in D_2O. The identity of isodomoic acid B was similarly confirmed by ^{1}H-NMR by comparison to literature spectra (3, 4).

Molecular Biology/Biochemical Methods

DNA Extraction

C. armata was collected from Hanazezaki (31˚11'40" N 130˚30'30" E), Ibusuki City, Kagoshima Prefecture, Kyushu Island, Japan. The sample was dried by lyophilization and finely ground. 1 g of this *C. armata* powder was used to extract DNA according to the previously described protocol (2). Extracted DNA was further purified using a Short Read Eliminator Kit, XSRE (Circulomics) and size selected using the BluePippin system with a High Pass Plus 15 kb cassette (Sage Science).

Oxford Nanopore Sequencing

Size-selected, high-molecular weight (HMW) DNA was sequenced using the Oxford Nanopore MinION and PromethION platforms (Oxford Nanopore Technologies (ONT), Oxford, UK). Two one-dimensional (1d) libraries were prepared with 900 ng HMW DNA using the Ligation Sequencing Kit (SQK-LSK109) (ONT, Oxford, UK). An initial library was loaded onto a R9.4 flowcell and run for 48 hr, resulting in 6.9 Gb of sequence with a read length N50 of 15 kb. A second library was generated using the same method as before and loaded onto a R9.4.1 flowcell and run using the PromethION platform 24 (P24), yielding 46 Gb of sequence with a read length N50 of 12.5 kb. The data from these two runs were pooled and used for assembly. NCBI Sequence Read Archive accession: SRR15927349 and SRR15927348.

Illumina Sequencing

NEBnext sequencing libraries were generated to polish genome assemblies (New England Biolabs, Beverly, MA, USA). NEBnext sequencing libraries were created with 100 ng of DNA and quality-controlled on a bioanalyzer. Resulting libraries were sequenced on an Illumina MiSeq 2x150 bp to check quality and quantity (Illumina, San Diego, CA). The libraries were then sequenced on an Illumina NovaSeq S4 200 (PE100, 10x) run that resulted in 74 Gb of sequence. NCBI Sequence Read Archive accession: SRR15927350.

Genome assembly

Illumina reads were trimmed with Trimmomatic v0.3 (5) and used for genome size estimation and initial assembly by Platanus v1.2.4 (6) using the following parameters: -k 17 -s 10 -u 0.2 -t 24 -m 500. Using the 17-mer histogram, *C. armata* genome size was estimated to be 480 Mb (Figure S2).

Raw pooled ONT and illumina reads were assembled using the MaSuRCA pipeline v3.4.2 (7). Configuration file parameter FLYE_ASSEMBLY=1 was used for final assembly of corrected mega-reads. All other parameters were set to default settings.

Genome quality assessment

C. armata genome assembly contiguity and completeness were assessed with BUSCO (v4.0.5) (8) using the eukaryota benchmarking universal single-copy orthologs eukaryota_odb10 dataset. This method was applied to the ten publicly available red macroalgal genomes – *Agarophyton vermiculophyllum* (AgarVerm_1.0), *Asparagopsis taxiformis* (ASM1839795v1), *Chondrus crispus* (ASM35022v2), *Digenea simplex* (ASM479842v1), *Gracilariopsis chorda* (GraCho1.0), *Gracilariopsis lemaneiformis* (Glem_v01), *Kappaphycus alvarezii* (ASM220596v3), *Neoporphyra haitanensis* (ASM982973v1), Neopyropia yezoensis (ASM982973v1) and *Porphyra umbilicalis* (P_umbilicalis_v1) – to establish a relative comparison of *C. armata* eukaryotic gene orthologs. Raw ONT reads were mapped to the final assembled genome using minimap2 (9). Mapped reads were processed, sorted, and indexed using SAMtools (10) with the sorted bam file reporting 88.32% of reads mapped to the assembled genome. The average depth of read coverage for *rad1, rad2*, and *rad3* was calculated to be 39x, 21x, and 21x, respectively.

PCR and cloning

Polymerase chain reaction (PCR) was carried out using standard thermocycling protocols using Prime Star MAX (TaKaRa) using the templates and primers as specified for each reaction. All reactions using plasmids as template were treated with DpnI (New England Biolabs) following PCR amplification to remove contaminating template DNA. Assembly of all PCR fragments was performed using the NEBuilder HiFi DNA Assembly Mix (New England Biosciences) following manufacturer protocols. Initial amplification of *radA1*, *radC1*, and *radC2* was performed using the primer sets RadA_p28_N_001-002 and RadC_p28_N_001-

002 with *C. armata* gDNA as template. The pET28a vector was linearized using the primer set p28-N_BB_001-002. Assembly of *radA* and *radC* amplicons into the linearized pET28a vector incorporated the N-terminal His_6 affinity tag, whereupon the assembled constructs were transformed into *E. coli* DH5a chemically competent cells (ThermoFisher) and plated on kanamycin (50 mg/mL) LB plates. Purification of plasmid DNA and subsequent Sanger sequencing (GeneWiz) confirmed the presence of *radA1*, *radC1*, and *radC2* inserts. All constructs were transformed into chemically competent *E. coli* BL21(DE3) (New England Biolabs).

Following expression testing of constructs, further truncations and tags were screened for *radA1* and *radC2* to yield soluble protein expression. A 7-amino acid N-terminal truncation was generated for *radA1* using the primers RadA_Δ7_MBP and p28_3prime_Amp using the construct generated above as a template. A modified pET28 vector containing an N-terminal maltose binding protein (MBP) with N-terminal His_6 tag was linearized using the primer set MBP_BB_001-002 (11). Assembly of the truncated *radA1* into the linearized MBP vector yielded the construct N-His_6-MBP-D7-RadA1, which was transformed and purified, as described above, and verified by Sanger sequencing.

An 11-amino acid N-terminal truncation was generated for *radC2* using the primers RadC2_Δ11_p28-C and RadC2_CtermOH_Amp with the original RadC2 construct generated above as template. The pET28a vector was linearized using the primer set p28-C_BB_001-002 to allow the addition of a C-terminal His_6 tag to the insert. Assembly of the truncated *radC2* into the linearized pET28a vector yielded the construct D11-RadC2-His_6-C, which was transformed, purified, and sequenced as described above to verify the construct.

Extraction of domoic acid isomers from *C. armata*

To extract domoic acid isomers from *C. armata,* 50 mg of dried algae was pulverized and extracted with 500 μL of 50% methanol. This mixture was mashed and vortexed every 15 mins, for 90 mins at 60˚C. The mixture was then centrifuged for 15 mins at (21000 xg, 20 min, 20 ˚C) to pellet debris. The supernatant was gently removed and filtered prior to injection (10 μL) on LC-HRMS.

Protein Purification

Both DabC and DsKabC were purified as described previously (1, 2). For the newly generated N-His_6-MBP-Δ7-RadA1 (RadA1, hereafter), N-His_6-RadC1 (RadC1, hereafter) and Δ11-RadC2-His_6-C (RadC2, hereafter) protein expression constructs, culturing and protein purification was performed using standard methodology. Cultures of *E. coli* BL21 (DE3) chemically competent cells co-transformed with RadA1, RadC1, and RadC2 protein expression plasmids were shaken at 37 ˚C in 1 L of Terrific Broth (Fisher Bioreagents) to an OD_{600} of ~0.6. Then, cultures were chilled to 18 ˚C and induced with 1 mM of isopropylthio-β-galactoside (IPTG). Flasks were shaken overnight (~16 hr) and cells were harvested by centrifugation (8000 xg, 10 min). Cell pellets were resuspended in 25 mL of lysis buffer (20 mM Tris pH 8, 500 mM NaCl, 10 mM imidazole, 10% glycerol) and stored at -80 ˚C for future purification.

All frozen cell pellets were defrosted, and cells were lysed by sonication using a Qsonica 6 mm tip at 40% amplitude for 20 cycles of 15 seconds on and 45 seconds off, gently mixing after 15 cycles. Cell lysate was then centrifuged for 30 mins at 20,000xg to pellet cell debris. Initial immobilized metal-affinity chromatography (IMAC) purification of RadA1, RadC1, and RadC2 was performed similarly for all soluble protein constructs. Briefly: clarified lysate was loaded at 2 mL/min onto a 5 mL HisTrap FF column (Cytiva) using a pre-equilibrated with wash buffer (20 mM Tris, 500 mM NaCl, 10 mM imidazole, pH 8). The loaded column was washed with 10 column volumes of 8% elution buffer (20 mM Tris, 500 mM NaCl, 250 mM imidazole, pH 8). Protein was eluted using a linear gradient of 8-100% elution buffer over 15 column volumes in 4 mL fractions. Fractions were assessed using SDS-PAGE, and target protein containing

fractions were pooled and concentrated using Amicon Ultra-15 50 kDa-cutoff (RadA1) and 30 kDa-cutoff (RadC1 and RadC2) centrifugal filters (Millipore Sigma).

RadC1 and RadC2 were further purified by size exclusion chromatography. Briefly: protein was concentrated to 1 mL using an Amicon Ultra-15 30 kDa-cutoff concentrator and further purified at a flow rate of 1 mL/min with a HiLoad 16/60 Superdex 75 prep grade column (GE Healthcare Life Sciences) pre-equilibrated with 50 mM HEPES pH 8.0, 250 mM NaCl, and 10% glycerol. Purity was checked using SDS-PAGE and relevant fractions were pooled, concentrated, and stored immediately at -80 ˚C.

Following IMAC purification, concentrated RadA1 protein was desalted and buffer-exchanged using PD-10 columns (Sephadex G-25 M, Cytiva) pre-equilibrated with storage buffer (50 mM HEPES pH 8, 250 mM NaCl, 10% glycerol) also containing 5 mM $MgCl_2$ to help with RadA1 stability. The buffer-exchanged RadA1 was further concentrated, aliquoted, and stored immediately at -80 ˚C. RadA1 was not stable upon attempted size exclusion chromatography using 50 mM HEPES pH 8.0, 250 mM NaCl, 5 mM $MgCl_2$, and 10% glycerol.

Enzymatic activity assays

Enzyme assays to demonstrate kainoid synthase (RadC1, RadC2, DabC, KabC) function were carried out as previously described (1). Assays were conducted in 100 mM HEPES (pH 8.0), 100 mM KCl, 10% glycerol buffer with 1 mM ascorbate and 6.25 mM alpha-ketoglutarate (αKG). Prenylated glutamate substrate (NGG, cNGG, prekainic acid) was added to 1 mM, followed by 50 mM of enzyme and 50 mM of $FeSO_4$. Total reaction volume was 100 μL. Reactions were allowed to incubate at room temperature (~25 ˚C) overnight (~18 hr) and were then quenched with 100 μL of ice-cold methanol. Quenched reactions were centrifuged (21000 xg, 20 min, 4 ˚C) in a tabletop centrifuge and subsequently filtered with a nylon 0.22 μm pore CA membrane (Costar Spin-X) prior to injection (10 μL) on LC-HRMS. LC Method A was used for all kainoid synthase assays using the geranylated substrates (NGG and cNGG) and associated standards, whereas LC Method B was used for kainoid synthase assays on pre-kainic acid and associated standards. Substrates and products were identified using retention time and mass [M-1] in comparison to injections of prepared chemical standards. A similar 10 μL injection of *C. armata* DA extract, prepared as described above, was also analyzed by LC-HRMS in comparison with cNGG enzyme assays and standards to ascertain relative DA isomer abundance in our *C. armata* isolate.

Enzyme assays to demonstrate *N*-prenyltransferase function were carried out as previously described, with modifications (1, 2, 12). Assays were conducted in 100 mM HEPES (pH 8.0), 100 mM KCl, 10% glycerol buffer with 5 mM $MgCl_2$. Isoprene diphosphate (DMAPP or GPP) was added to 1 mM together with 20 mM of L-glutamate. Reactions were allowed to incubate at room temperature (~25 ˚C) overnight (~18 hr) and were then quenched with 100 μL of ice-cold methanol. Quenched reactions were centrifuged, filtered, and injected (10 μL) onto LC-HRMS. LC Method B was used for all *N*-prenyltransferase assays described here. Substrates and products were identified using retention time and mass [M-1] in comparison to injections of prepared chemical standards.

Phylogenetic analysis

Kainoid synthase tree: Initial BLAST searches revealed few closely related homologs outside of known KA or DA biosynthetic genes. These top tBLASTn hits were pooled with representative sequences from the isopenicillin N synthase-like (IPNS-like) InterPro family (IPR027443). Representative sequences were selected from UniProt using the UniRef50 function to provide a clustered set of sequences. Select sequences were further clustered using the CD-HIT suite, and a representative sequence from the largest 150 clusters was selected for further analysis, final selection of sequences listed in Table S5. A multiple

sequence alignment was generated using kalign (v2.04) (13). A maximum-likelihood (ML) phylogenetic tree was built using IQ-TREE (14), and the best-fit substitution model, VT+R5, was automatically selected. This process of using top BLAST hits and representative InterPro family sequences to construct a ML tree was repeated with RadD using the cytochrome P450 InterPro family (IPR001128). In addition to top BLAST hits and representative sequences, diatom P450 sequences were mined from publicly available transcriptomic datasets and a subset of related P450s was seeded into the tree, final selection of sequences listed in Table S6. The substitution model LG+R6 was automatically selected. All trees were visualized in iTOL (15).

Kainic Acid					
order	**family**	**genus**	**species**	**extract**	**gene**
Ceramiales	Rhodomelaceae	Digenea	simplex	yes	yes
Ceramiales	Rhodomelaceae	Laurencia	papiillosa	yes	no
Ceramiales	Rhodomelaceae	Vidalia	obtusiloba	yes	no
Ceramiales	Ceramiaceae	Centroceras	clavulatum	yes	no
Palmariales	Palmariaceae	Palmaria	palmata	yes	yes
Palmariales	Palmariaceae	Palmaria	hecatensis	yes	no
Palmariales	Rhodophysemataceae	Rhodophysema	elegans	yes	yes
Halymeniales	Halymeniaceae	Grateloupia	filicina	yes	yes
Domoic Acid					
order	**family**	**genus**	**species**	**extract**	**gene**
Ceramiales	Rhodomelaceae	Digenea	simplex	yes	no*
Ceramiales	Rhodomelaceae	Chondria	armata	yes	yes
Ceramiales	Rhodomelaceae	Osmundaria	obtusiloba	yes	no
Ceramiales	Rhodomelaceae	Amansia	glomerata	yes	no
Ceramiales	Rhodomelaceae	Alsidium	helminthochorton	yes	no

Table S1. Summary of known KA and DA red algal producers from the literature. Right columns indicate if the study included chemical characterization of extracts and/or validation of biosynthetic genes (16–19). *A homolog for RadD was not found in the sequenced *D. simplex* sample.

Primer Name	Sequence
RadA_p28_N_001	GGTGCCGCGCGGCAGCCATATGAAGGTACTTGCAGAAGACGACCC
RadA_p28_N_002	GGTGGTGGTGGTGCTCGAGTCAAGTCGCTGTACTAATTATTTTAACGAGCTCG
RadC_p28_N_001	GGTGCCGCGCGGCAGCCATATGTTTACGATCAAAGGAACGGAACTGAAC
RadC_p28_N_002	GGTGGTGGTGGTGCTCGAGCTAGTAGTAGCCATGAAGAAACTTGTATTTTACG
p28_3prime_Amp	TGGTGGTGGTGGTGCTCGAG
RadA_d7_MBP	CTGTACTTCCAATCCGGATCCGACCCAAATGATGCACTAGCCCGTATCAAAACC
RadC2_d11_p28-C	CTTTAAGAAGGAGATATACCATGGATTTTAATCCCTTGGAAGTAGAGAAGCTCAATTG
RadC2_CtermOH_Amp	GCTCGAGTGCGGCCGCAAGCTTGTAGTACCCATGAGTAAACTTGTATTTTACGTAAGTGA
MBP_BB_001	GGATCCGGATTGGAAGTACAGGTTCTCAGATCC
MBP_BB_002	CTCGAGCACCACCACCACCACCACTGAG
p28-N_BB_001	CATATGGCTGCCGCGCGGCACC
p28-N_BB_002	CTCGAGCACCACCACCACCACCACTGAG
p28-C_BB_001	CATGGTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGGGGAATTG
p28-C_BB_002	AAGCTTGCGGCCGCACTCGAGC

Table S2. Primers used in this study

Statistics without reference	Assembly
# contigs	2991
# contigs (>= 0 bp)	2991
# contigs (>= 1000 bp)	2818
# contigs (>= 5000 bp)	2271
# contigs (>= 10000 bp)	2035
# contigs (>= 25000 bp)	1769
# contigs (>= 50000 bp)	1490
Largest contig	3292845
Total length	507692717
Total length (>= 0 bp)	507692717
Total length (>= 1000 bp)	507578445
Total length (>= 5000 bp)	506026847
Total length (>= 10000 bp)	504384905
Total length (>= 25000 bp)	499967534
Total length (>= 50000 bp)	489680988
N50	643002
N75	256360
L50	226
L75	532
GC (%)	45.34
Mismatches	
# N's	19500
# N's per 100 kb	3.84

Table S3. Quast statistics of the final assembled *C. armata* genome.

Scientific name	Assembly	Complete	Single -copy	Duplicated	Fragmented	Missing	n=255
Chondria armata	SIO_Carma_v2.2	74.5	70.2	4.3	7.5	18.0	255
Agarophyton vermiculophyllum	AgarVerm_1.0	59.6	58.8	0.8	6.3	34.1	255
Asparagopsis taxiformis	ASM1839795v1	78.0	69.4	8.6	5.5	16.5	255
Chondrus crispus	ASM35022v2	72.2	71.0	1.2	8.2	19.6	255
Digenea simplex	ASM479842v1	71.4	69.8	1.6	7.5	21.1	255
Gracilariopsis chorda	GraCho1.0	76.9	74.9	2.0	5.5	17.6	255
Gracilariopsis lemaneiformis	Glem_v01	75.3	74.1	1.2	6.7	18.0	255
Kappaphycus alvarezii	ASM220596v3	59.6	41.6	18.0	11.8	28.6	255
Neoporphyra haitanensis	OUC_PyHait	52.6	51.4	1.2	13.3	34.1	255
Neopyropia yezoensis	ASM982973v1	54.9	54.5	0.4	13.7	31.4	255
Porphyra umbilicalis	P_umbilicalis_v1	42.4	42.0	0.4	14.1	43.5	255

Table S4. BUSCO assessment of publicly available red macroalgal genomes

D4E802	A0A2V3J2T9	A0A6S7ZS46	A0A6P4B8E0	A0A0W0G3P1
A0A485B8P9	A0A6V0PIX2	A0A6V1FBN5	A0A371EF89	A0A094A4Z7
A0A1I2FBH3	M2WQW5	A0A6V1VUR4	A0A4S4F1G2	A0A093Z4A8
A0A0Q0DE95	R7QT26	A0A6S8Y206	A0A5N5L5D4	A0A7D7ZSX3
A0A2X4YWX6	A0A5J4YHC7	A0A6U4KKR1	B9RI27	WP_075477798
A0A6M4A7N8	A0A6T5W5I9	K0RCE9	A0A2U1NEA5	WP_112713920
W6RJX4	A0A6T6N7U3	A0A6V3GZM2	A0A2G3D1B6	A0A1R2C824
A0A653YTM6	A0A5J4Z228	A0A6U1BI44	A0A5N5P4G6	A0A1R2CN02
A0A1H4S654	A0A6T9YZV4	A0A6U3NAE4	A0A5J9TM82	P18548
A0A3N1V457	A0A6T6NYH4	A0A6V0NRI9	A0A094H7E5	
A0A562MRU1	A0A6V0MET1	A0A6U0GHJ2	A0A135S7A4	
A2WDT8	A0A5J4Z328	A0A6U6F6P8	A0A139IR59	
M5D0U2	A0A6T6LKS5	A0A6V3DN33	A0A1S9RV91	
A0A663ASE4	A0A2V3J2G5	A0A6U2NKB2	A0A5N6J8W1	
A0A4R3ZW85	A0A6T6KDA2	A0A124SC96	A0A2G8SRB9	
A0A2N3PW13	M1VKP3	A0A438J382	A0A3M7LXI8	
A0A0P9P600	A0A6V0QHG1	A0A2U1NH25	A0A1L9TCD4	
A0A387HMF6	A0A1X6PK80	A0A4V3WLV0	A0A1X6MM30	
A0A286HD09	A0A5J4Z952	A0A6A2YJD0	A0A444RNV4	
A0A021VNH7	A0A6V1LTB9	A0A5B6VHC0	A0A0F2M517	
Q98N28	A0A6V1I903	A0A3S3QTW4	A0A4U0WWH5	
I3UAA2	A0A6S8QUQ0	W9QYP6	A0A395SJD4	
Q3JLE9	A0A6U3NP22	A0A498I7M4	A0A0G2FG99	
A0A6P2CA42	A0A6U1FFP2	M4CSL2	A0A2P2HFL0	
C4WFH3	A0A6T0NQR4	A0A3Q7I749	A0A364MXH3	
A0A6T6Q0U9	A0A6T0LLC5	A0A540MTJ5	A0A135SK80	
A0A6T6NV76	A0A6V1Y066	A0A314KJM9	A0A4P7NST2	
R7QC14	A0A6S8G773	A0A5J5BV07	B8MRF9	
A0A6T6CML0	A0A6T2XKK6	A0A0B0MG73	W3WSK3	
A0A5J4Z4R2	A0A6T0K2C4	A0A1J6JT00	A0A1S9D541	

Table S5. Kainoid synthase representative sequences. UniProt accession numbers of representative sequences of key taxonomic groups from the isopenicillin N synthase-like (IPNS-like) InterPro family (IPR027443). Sequences highlighted in grey represent NCBI accession numbers of top BLAST hits.

A0A6V1GDW8	A0A4R2CXH8	A0A094AXQ3	A0A498I3I0	A0A6T6AFA8
A0A6V1V780	A0A2Z4V4B8	A0A2N1NSW2	A0A3Q7IZP9	M2VTT7
A0A3S7L8P2	A0A1H5YLD1	A0A094JJ65	A0A5N6QLC6	M1VHJ8
A0A6U5MHZ5	A0A4R4VQ17	A0A2G7FJ09	A0A200RBS5	A0A2V3IHK9
A0A6V3BWI7	A0A249PFW7	A0A094HR63	A0A2G3C5P5	247725
K0RTJ9	A0A0M4Q8A0	A0A135TVV1	A0A6A1UQ32	202814
A0A6S9PJ72	A0A290XS85	A0A2G7FQT3	A0A6L2M770	130274
A0A6U3SSW4	A0A498PLX0	A0A0G4KS67	A0A5J4ZT80	32491
A0A6U0ZEH0	X8CRZ0	W9CUQ1	A0A371HFR0	
A0A6U1BAX8	A0A2G6CH09	A0A2P4QZ99	A0A1R3I738	
A0A6U2T721	A0A4D4L8P2	A0A165K1U7	A0A6T5VV58	
A0A6T0HIW2	A0A3N6E6F4	A0A0G4NKU9	A0A1X6P1P4	
A0A6U0GZL9	A0A0P0R9V1	A0A175W394	A0A6V0RYK1	
A0A6V0Y2P1	A0A1I4AUB2	A0A4Y9ZEV9	A0A6T6MKE5	
A0A6V0Q6S4	A0A2I7WCV3	A0A4U9EEN7	A0A5J4YL64	
A0A6S8EW11	I7FQV2	A0A6N2LYY4	A0A5J4YNM0	
A0A6T9B9Y7	A0PL28	A0A5D2RYA7	A0A2V3IWP3	
A0A6V2X6A8	A0A6H9IY50	A0A2N9HUK6	A0A6T6AE48	
A0A6V2E486	A0A3G2J7I5	A0A2N9H9F4	A0A6T6Q6L2	
A0A1E7EPM2	A0A1Q5HD23	A0A5N5LFD9	A0A1X6PC05	
A0A6S8LNV9	A0A2U3E828	A0A200Q069	A0A5J4YWA9	
A0A6V2JND1	A0A421J420	A0A6A2XWI3	R7QH57	
A0A6V4QEM2	A0A1S9DKV5	A0A498J2J3	A0A6V0RHZ8	
A0A6U1UVS9	A0A484GA20	A0A6N2LLD8	A0A6T6K777	
A0A6U3S297	A0A094D2J3	A0A166FQ11	A0A1X6P2T0	
A0A6N0WX86	A0A2U3DT16	A0A1Q3CNA7	A0A6V0RF11	
U5ELP1	A0A2N1NTF4	A0A6A3B2M2	R7QBA0	
A0A3L8C4K3	J4KML6	A0A5B6WX83	A0A6T6B6C5	
C2W9K6	A0A4S9AE14	A0A498I2M7	A0A6T6ATK7	
A0A3Q8XTT3	A0A1S9RMP3	A0A498KMK2	A0A6V0MZH5	

Table S6. CYP450 representative sequences. UniProt accession numbers of representative sequences of key taxonomic groups from the cytochrome P450 InterPro family (IPR001128). Sequences highlighted in grey represent the Protein ID of sequences from the JGI PhycoCosm.

Repository	Accession	Sequence
BioProject	PRJNA762367	*Chondria armata* BioProject
BioSample	SAMN21392177	*Chondria armata* BioSample
Genome	JAIWHZ000000000	*Chondria armata* Assembly (SIO_Carma_v2.2)
SRA	SRR15927350	Illumina Sequencing of *Chondria armata*
SRA	SRR15927349	Nanopore Sequencing of *Chondria armata*
SRA	SRR15927348	Nanopore Sequencing of *Chondria armata*
GenBank	OK169902	*rad1* gene cluster
GenBank	OK169903	*rad2* gene cluster
GenBank	OK169904	*rad3* gene cluster
GenBank	QCC62383.1	*Digena simplex* KabC
GenBank	AYD91075.1	*Pseudo-nitzschia multiseries* DabC

Table S7. Accession numbers of sequences deposited and used in this study

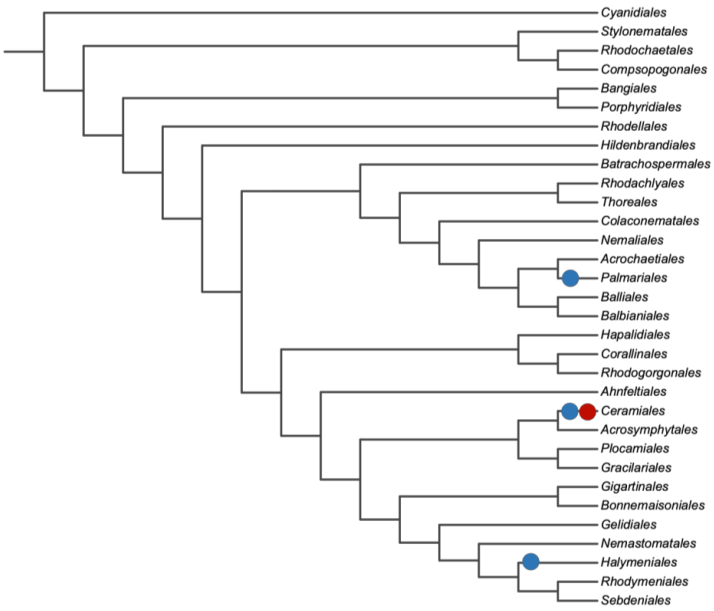


Figure S1. Cladogram of DA and KA producers by order in Rhodophyta. Blue and red dots indicate KA and DA producing orders, respectively. See Table S1 for a list of producing organisms.

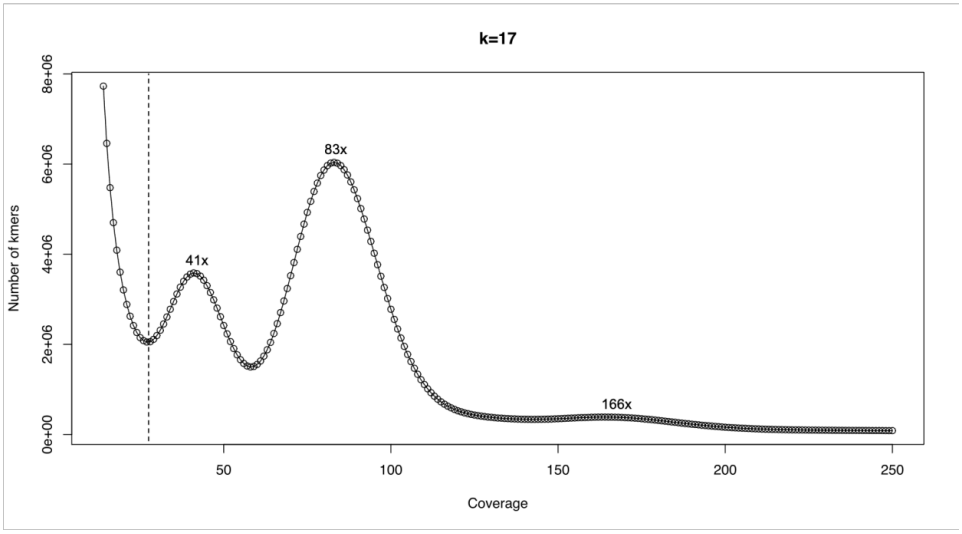


Figure S2. K-mer analysis of the *C. armata* genome. Tetraploid (41x), diploid (83x), and haploid (166x) peaks are labeled. Sequences to the left of the dashed line were omitted. Haploid genome size was estimated to be 480 Mb.

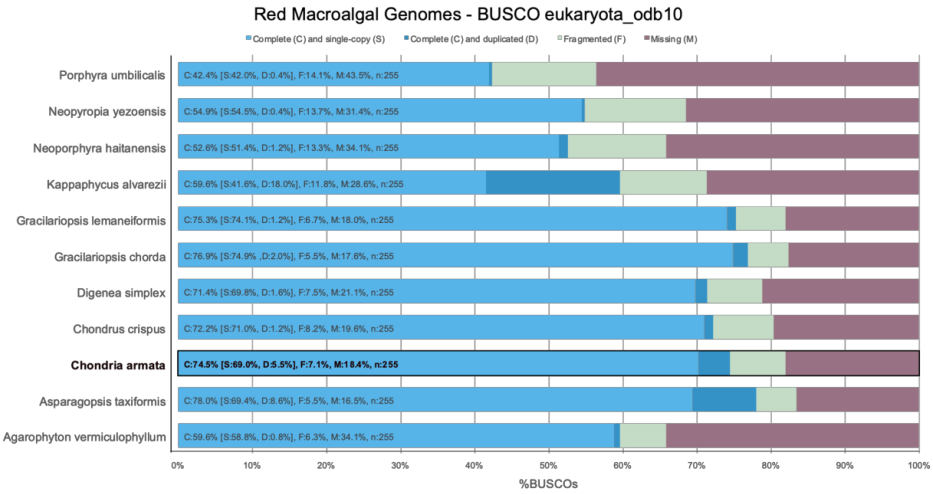


Figure S3. BUSCO assessment of red macroalgal genomes. *C. armata* BUSCO assessment highlighted in black box.

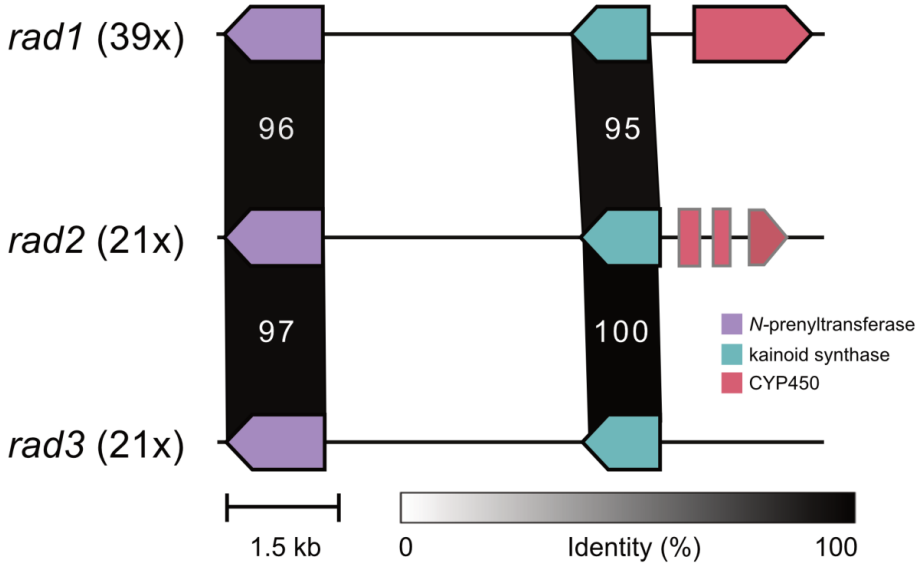


Figure S4. The three copies of the *rad* cluster within the *C. armata* genome. Mean coverage across each copy in parentheses. The sequences of *rad1*, *rad2*, and *rad3* are deposited in GenBank, accession numbers OK169902, OK169903, and OK169904, respectively.

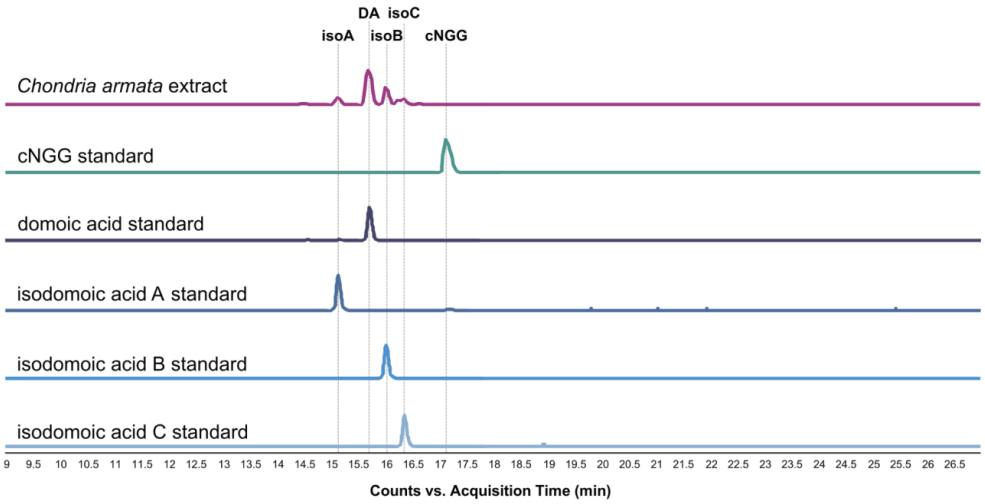


Figure S5. Domoic acid isomers in *C. armata*. Negative mode LC-HRMS chromatograms using LC Method A of domoic acid (DA), isodomoic acids (isoA, isoB, isoC), and cNGG standards compared with *C. armata* extract, showing extracted ion chromatograms (EIC) for anticipated DA isomer and cNGG masses (*m/z* 310.1 ± 0.2 and *m/z* 312.1 ± 0.2, respectively).

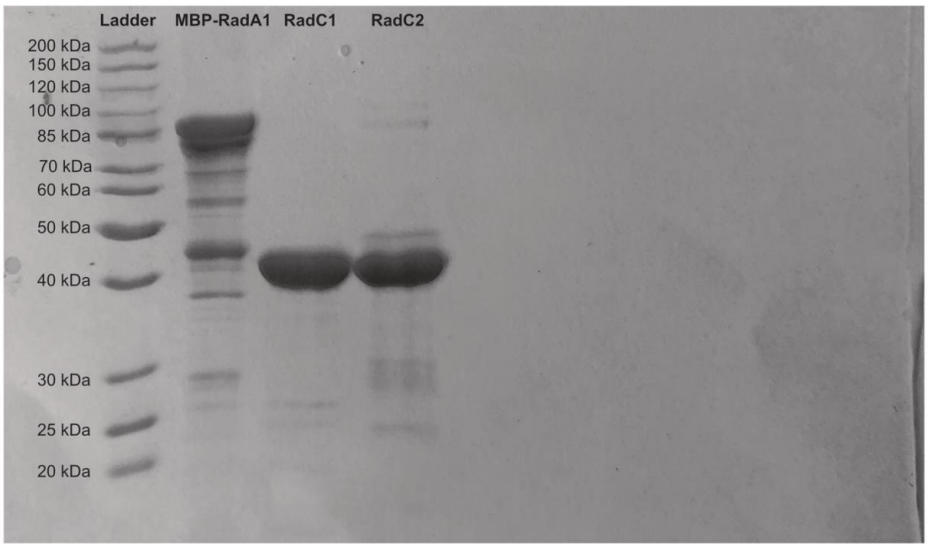


Figure S6. 10% SDS-PAGE gel. Sample order is as follows: Ladder, MBP-RadA1, RadC1, RadC2. Note the additional band in the RadA lane is the MBP tag without RadA.

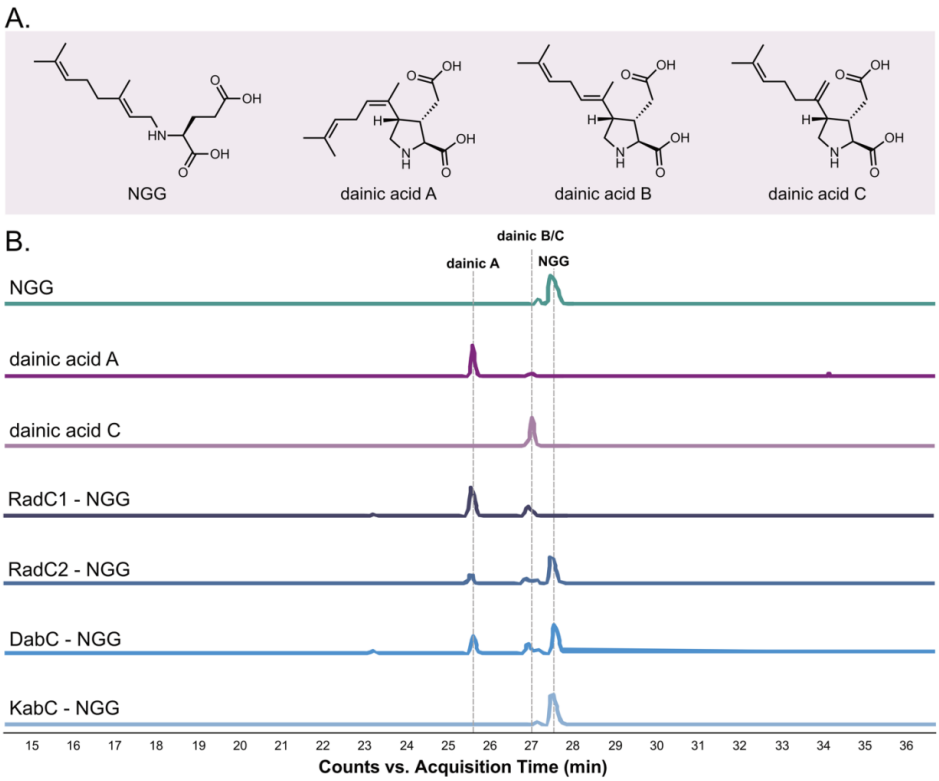


Figure S7. Kainoid synthase reactions were set up as previously described using NGG (1 mM). (*A*) NGG and dainic acid isomers. (*B*) Negative mode LC-HRMS chromatograms using LC Method A of kainoid synthase reactions are shown together with dainic acid standards. Dainic B and C have been previously shown to co-elute using a similar acidified normal phase chromatography method (1). Each trace is the combined extracted ion chromatograms for anticipated products and substrates (*m/z* 280.2 ± 0.2 and *m/z* 282.2 ± 0.2, respectively).

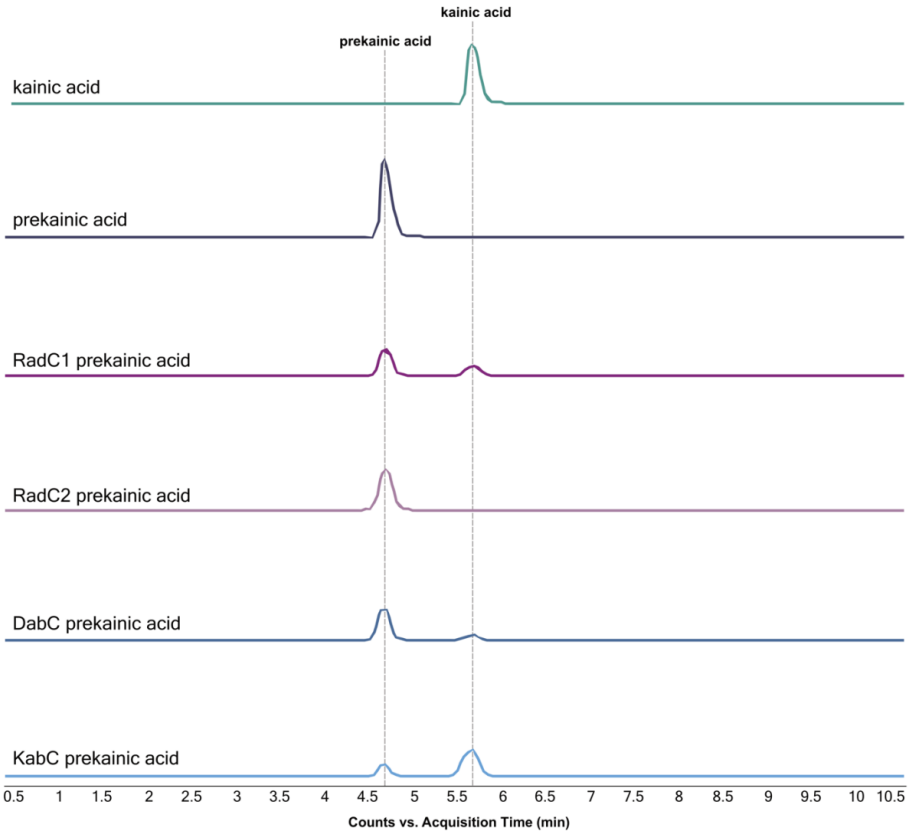


Figure S8. Prekainic acid substrate screen. Kainoid synthase reactions were set up as previously described using prekainic acid (1 mM). Negative mode LC-HRMS chromatograms using LC Method B of kainoid synthase reactions are shown together with prekainic acid and kainic acid standards. Each trace represents the combined extracted ion chromatograms for anticipated products and substrates (*m/z* 212.0 ± 0.2 and *m/z* 214.1 ± 0.2, respectively).

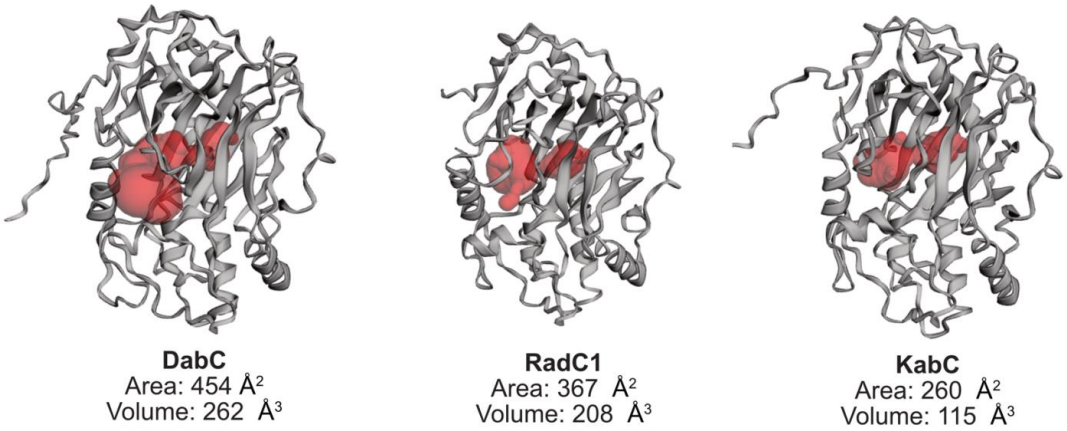


Figure S9. CASTp output of modeled pocket volumes for *Pseudo-nitzschia multiseries* DabC, *C. armata* RadC1, and *Digenea simplex* KabC. Enzyme binding pockets highlighted in red, measurements of pocket area and volume listed below each model label (20).

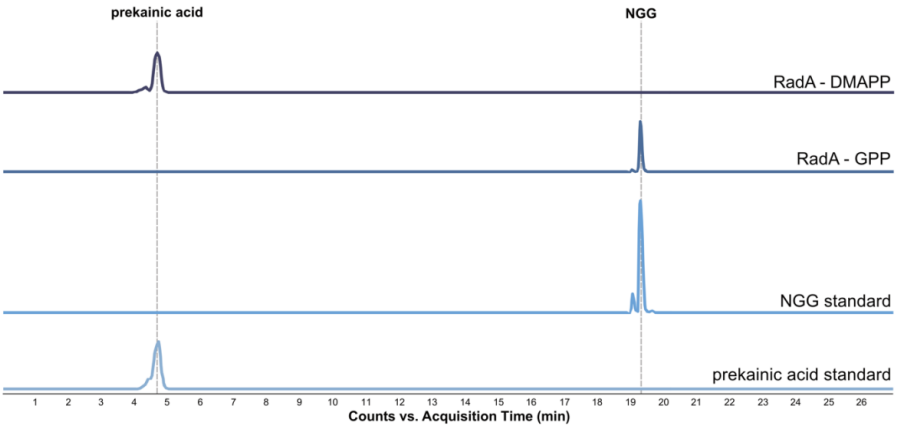


Figure S10. Overnight assays with purified MBP-Δ7-RadA1. Substrate screening suggest both GPP and DMAPP can be accepted as substrates to make NGG and prekainic acid, respectively. RadA1 *N*-prenyltransferase reactions were set up as previously described using L-glutamate acid (20 mM) and DMAPP or GPP prenyl donors (1 mM). Negative mode HRMS chromatograms using LC Method B of RadA1 reactions are shown together with prekainic acid and NGG standards. Each trace represents the combined extracted ion chromatograms for anticipated NGG and prekainic acid products (*m/z* 282.2 ± 0.2 and *m/z* 214.1 ± 0.2, respectively).

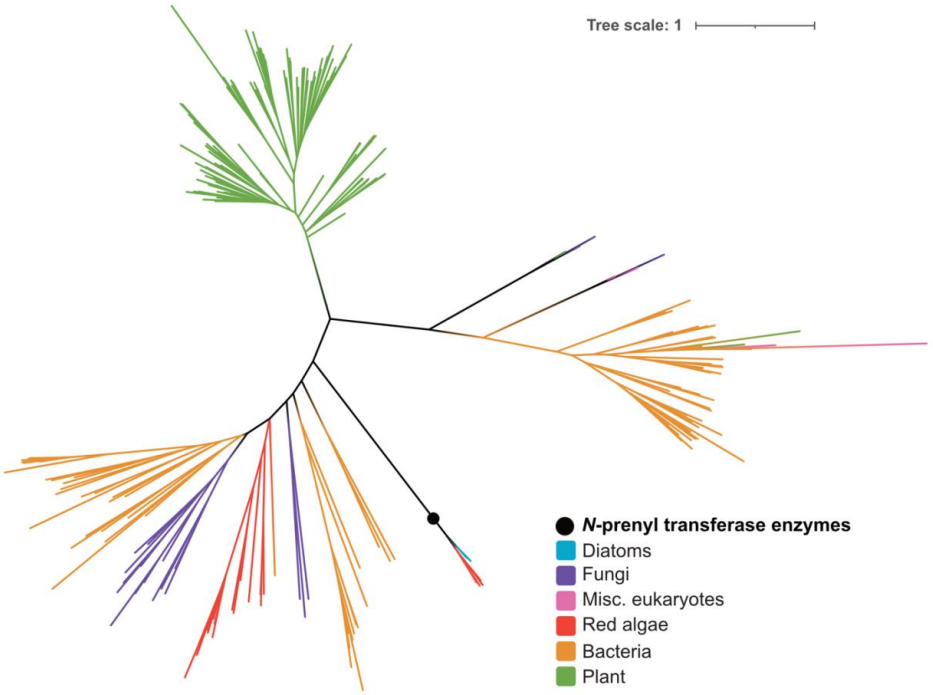


Figure S11. *N*-prenyltransferase enzyme maximum-likelihood phylogenetic tree. RadA sequences mapped to previously reported tree (12). They clade with the other *N*-prenyltransferases and are very distantly related to other terpene cyclases.

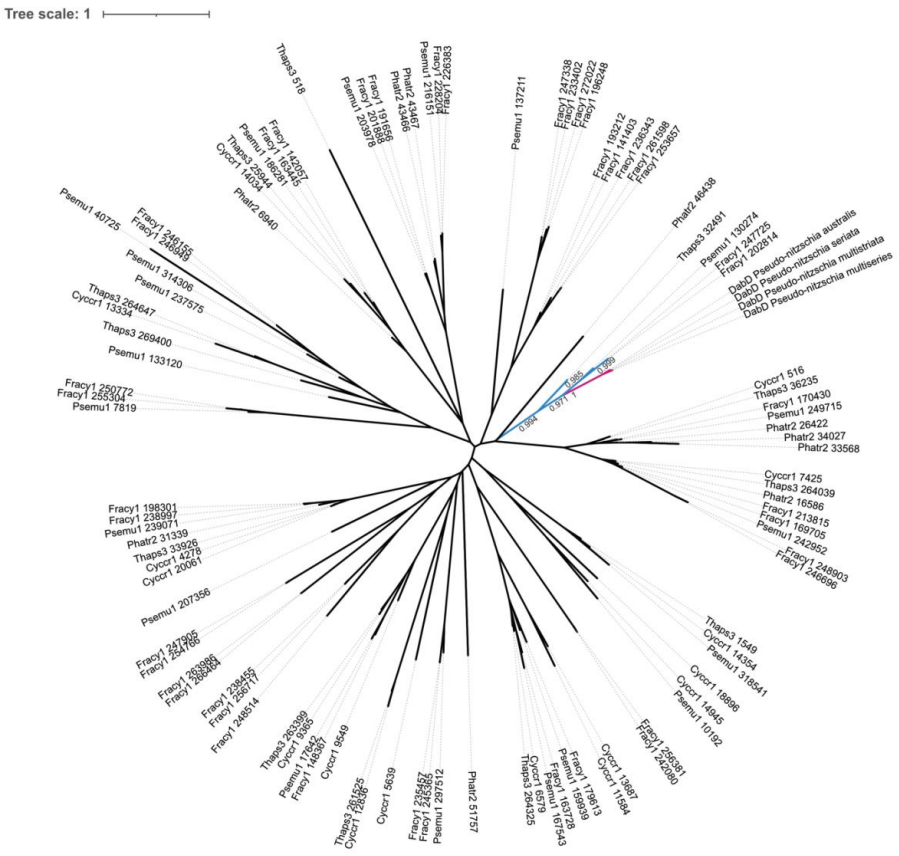


Figure S12. Expanded diatom CYP450 maximum-likelihood phylogenetic tree. Diatom sequences used in Figure 4B highlighted in blue and red. Numeric identifiers are Joint Genome Institute Protein IDs (JGI PIDs) from the respective publicly available diatom genomes through JGI. Psemu1 is the identifier for the *Pseudo-nitzschia multiseries* CLN-47 sequencing dataset.

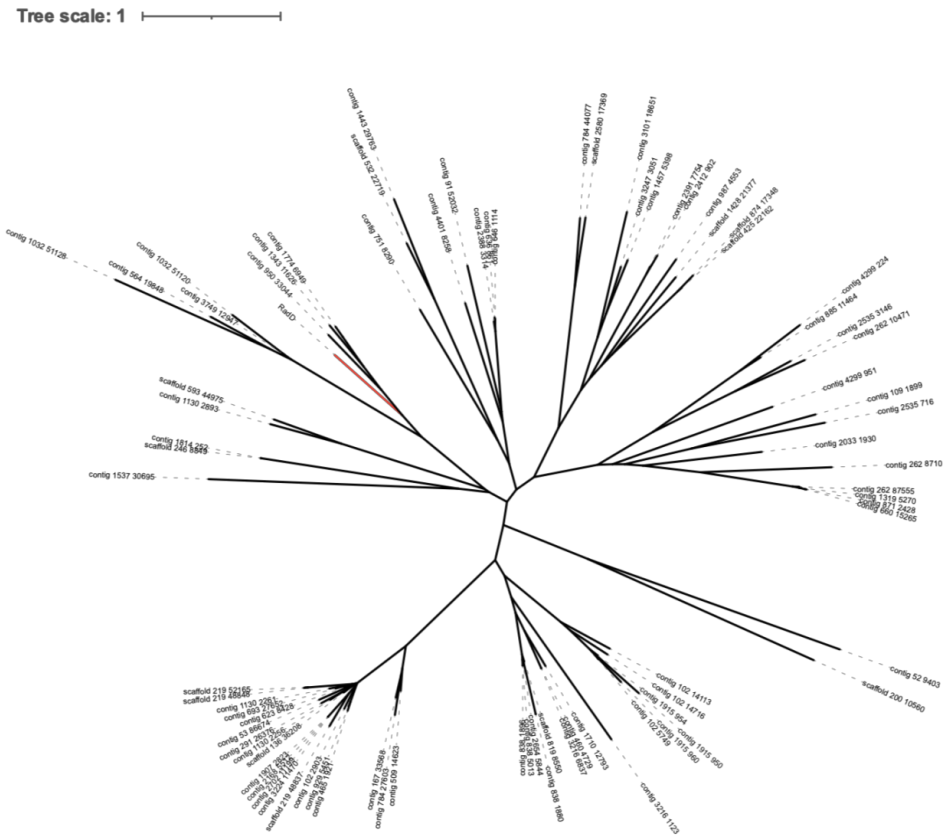


Figure S13. Expanded *C. armata* CYP450 maximum-likelihood phylogenetic tree. RadD highlighted in red.

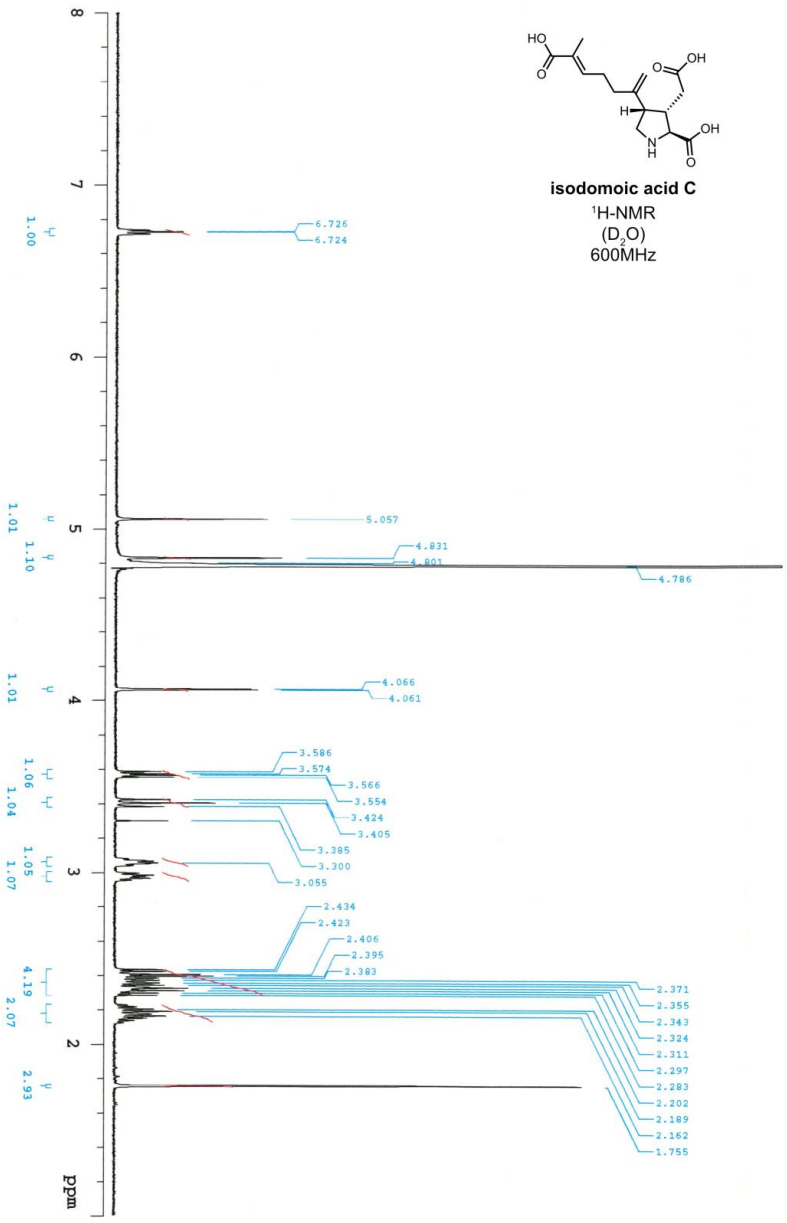


Figure S14. ^{1}H-NMR spectrum of isodomoic acid C. Analysis performed in D_2O/600 MHz. The signal of the residual MeOD was adjusted at δ 3.30 ppm as the internal reference in D_2O.

Supplementary References

1. J. K. Brunson, *et al.*, Biosynthesis of the neurotoxin domoic acid in a bloom-forming diatom. *Science* **361**, 1356–1358 (2018).

2. J. R. Chekan, *et al.*, Scalable biosynthesis of the seaweed neurochemical, kainic acid. *Angew. Chem. Int. Ed.* **58**, 8454–8457 (2019).

3. Y. Maeno, *et al.*, Six domoic acid related compounds from the red alga, *Chondria armata*, and domoic acid biosynthesis by the diatom, *Pseudo-nitzschia multiseries*. *Sci. Rep.* **8**, 356 (2018).

4. M. Meda, *et al.*, Structures of isodomoic acids A, B and C, novel insecticidal amino acids from the red alga *Chondria armata*. *Chem. Pharm. Bull.* **34**, 4892–4895 (1986).

5. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

6. R. Kajitani, *et al.*, Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nat. Commun.* **10** (2019).

7. A. V. Zimin, *et al.*, Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res*. **5**, 787-792 (2017).

8. M. Seppey, M. Manni, E. M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).

9. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**, 3094-3100 (2018).

10. H. Li, *et* al., 1000 genome project data processing subgroup, the sequence alignment/map format and SAMtools. *Bioinformatics*. **25**, 2078-2079 (2009).

11. S. W. Lee, *et* al., Discovery of a widely distributed toxin biosynthetic gene cluster. *Proc. Natl. Acad. Sci.* **15**, 5879-5884 (2008).

12. J. R. Chekan, S. M. K. McKinnie, J. P. Noel, B. S. Moore, Algal neurotoxin biosynthesis repurposes the terpene cyclase structural fold into an N-prenyltransferase. *Proc. Natl. Acad. Sci.* **23**, 12799-12805 (2020).

13. T. Lassmann, Kalign 3: multiple sequence alignment of large datasets. *Bioinformatics* **36**, 1928–1929 (2020).

14. B. Q. Minh, *et al.*, IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

15. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. **49**, W293-W296 (2021).

16. G. Impellizzeri, *et al.*, Amino acids and low-molecular-weight carbohydrates of some marine red algae. *Phytochemistry* **14**, 1549–1557 (1975).

17. Y. Shimizu, *et al.*, Dinoflagellate and other microalgal toxins: chemistry and biochemistry. *Pure Appl. Chem.* **61**, 513–516 (1989).

18. M. Sato, *et al.*, Distribution of neuroexcitatory amino acids in marine algae. *Phytochemistry* **42**, 1595–1597 (1996).

19. M. V. Laycock, A. S. W. de Freitas, J. L. C. Wright, Glutamate agonists from marine algae. *J. Appl. Phycol.* **1**, 113–122 (1989).

20. W. Tian, C. Chen, X. Lei, J. Zhao, J. Liang, CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res.* **46**, W363–W367 (2018).

2.4 Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in *The Proceedings of the National Academy of Sciences*, Steele, T.S., Brunson, J.K., Maeno, Y., Terada, R., Allen, A.E., Yotsu-Yamashita, M., Chekan, J.R., and Moore, B.S., 2022. The dissertation author was the primary co-researcher and co-author of this chapter with Brunson, J.K.

CHAPTER 3. Biosynthesis of haloterpenoids in red algae via microbial-like type I terpene synthases

3.1 Introduction and Context for Chapter 3

Terpenoids are a rich source of bioactive seaweed chemistry.[1] Red algae are well-known for producing halogenated terpenes, or haloterpenoids, with some of the most prolific producers from the genera *Laurencia*,[2,3] *Plocamium,*[4,5] and *Portieria*.[6,7] The work presented in this chapter has a long history at the Scripps Institution of Oceanography, with much of our understanding of seaweed chemistry established by my committee member Professor William Fenical,[3,8,9] and his peers Professors D. John Faulkner (SIO),[4,10] and Richard E. Moore (Univ. of Hawaii).[11,12] Their early research is foundational to our understanding of seaweed chemistry and the inspiration for uncovering the genetic basis of haloterpenoid biosynthesis in red seaweeds. There are many ways to connect genes and chemistry, here we analyze sequencing datasets through the lens of the early chemical literature for the discovery of natural product biosynthesis genes.

The first efforts to elucidate haloterpenoid biosynthesis genes examined transcriptomic data from the genus *Laurencia* and established the existence of a new family of red algal terpene synthases.[13] Follow-up studies determined red algal terpene synthases are phylogenetically unrelated to typical plant terpene synthases[14] and most likely originated in red algae via horizontal gene transfer from an unknown microbial source.[15] However, to-date no algal terpene synthases have been characterized that produce hypothesized on-pathway haloterpenoid scaffolds. Similarly, very little sequencing data for haloterpenoid producing red algae has been generated, hindering further research of haloterpenoid biosynthesis.

The work in this chapter describes the sequencing of notable haloterpenoid producing red seaweeds, including: *Plocamium pacificum, Portieria hornemannii,*

Laurencia pacifica, and *Laurencia subopposita.* These sequencing datasets were generated in partnership with my co-author Malia Moore (*P.* hornemannii) and the U.S. Department of Energy Joint Genome Institute (*P. pacificum, L. pacifica,* and *L. subopposita*). In-depth discussion of DNA extraction and sequencing protocols can be found in Chapter 4 of this dissertation. In Chapter 3, we identify and biochemically characterize new red algal terpene synthases and validate the selective production of key haloterpenoid backbones. Additionally, we identify the first haloterpenoid associated gene pair. This discovery sets the stage for future work on haloterpenoid biosynthesis projects and clarifies long-standing questions around the biosynthesis of haloterpenoids in marine systems.

3.2 References for Chapter 3 introduction

1. Dictionary of Natural Products (April 13, 2023).

2. M. Harizani, E. Ioannou, V. Roussis, The *Laurencia* Paradox: An endless source of chemodiversity, *Prog. Chem. Org. Nat.* **102,** 91–252 (2016).

3. J. J. Sims, William. Fenical, R. M. Wing, Phillip. Radlick, Marine natural products. IV. Prepacifenol, a halogenated epoxy sesquiterpene and precursor to pacifenol from the red alga, *Laurencia filiformis*. *J. Am. Chem. Soc.* **95**, 972–972 (1973).

4. J. S. Mynderse, D. J. Faulkner, Polyhalogenated monoterpenes from the red alga *Plocamium cartilagineum*. *Tetrahedron* **31**, 1963–1967 (1975).

5. M. G. Knott, A review of secondary metabolites isolated from *Plocamium* species worldwide. *Int. Sci. Technol. J. Namibia* **6**, 75-93 (2015).

6. R. W. Fuller, J. H. Cardellina, Y. Kato, M. R. Boyd, L. S. Brinen, J. Clardy, K. M. Snader, A pentahalogenated monoterpene from the red alga *Portieria hornemannii* produces a novel cytotoxicity profile against a diverse panel of human tumor cell lines. *J. Med. Chem.* **35**, 3007–3011 (1992).

7. D. A. Payo, J. Colo, H. Calumpong, O. de Clerck, Variability of non-polar secondary metabolites in the red alga *Portieria*. *Marine Drugs* **9**, 2438–2468 (2011).

8. W. Fenical, Halogenation in the Rhodophyta, *J. Phycol.* **11**, 245–259 (1975).

9. B. M. Howard, A. M. Nonomura, W. Fenical, Chemotaxonomy in marine algae: Secondary metabolite synthesis by *Laurencia* in unialgal culture. *Biochem. Syst. Ecol.* **8**, 329–336 (1980).

10. D. J. Faulkner, Interesting aspects of marine natural products chemistry. *Tetrahedron* **33**, 1421–1443 (1977).

11. R. E. Moore, E. Moore, J. A. Pettus, M. S. Doty, A. Abbott, E. H. Williamson, A. Volatile compounds from marine algae. *Ethnobotanical* (1976).

12. B. J. Burreson, F. X. Woolard, R. E. Moore, Evidence for the biogenesis of halogenated myrcenes from the red alga *Chondrococcus Hornemanni. Chem. Lett.* **4**, 1111–1114 (1975).

13. R. D. Kersten, S. Lee, D. Fujita, T. Pluskal, S. Kram, J. E. Smith, T. Iwai, J. P. Noel, M. Fujita, J.-K. Weng, A red algal bourbonane sesquiterpene synthase defined by microgram-scale NMR-coupled crystalline sponge X-ray diffraction analysis. *J. Am. Chem. Soc.* **139**, 16838–16844 (2017).

14. G. Li, T. G. Köllner, Y. Yin, Y. Jiang, H. Chen, Y. Xu, J. Gershenzon, E. Pichersky, F. Chen, Nonseed plant *Selaginella moellendorffii* has both seed plant and microbial types of terpene synthases. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14711–14715 (2012).

15. G. Wei, Q. Jia, X. Chen, T. G. Köllner, D. Bhattacharya, G. K.-S. Wong, J. Gershenzon, F. Chen, Terpene biosynthesis in red algae is catalyzed by microbial type but not typical plant terpene synthases. *Plant Physiol.* **179**, 382–390 (2019).

3.3 Abstract

Red algae, or seaweeds, produce highly distinctive halogenated terpenoid compounds, including the bromochlorinated monoterpene halomon that was once heralded as a promising anticancer agent. The first dedicated step in the biosynthesis of these natural product molecules is expected to be catalyzed by a terpene synthase. Recent work has demonstrated an emerging class of type-I microbial-type terpene synthases in red algal terpene biosynthesis. However, only one terpene cyclase from a notoriously haloterpenoid-producing red alga (*Laurencia pacifica*) has been investigated functionally. Here we report ten new microbial-type terpene synthases (TSs) from the red algae *Portieria hornemannii, Plocamium pacificum, Laurencia pacifica*, and *Laurencia subopposita.* Whole genome sequencing and phylogenetic analysis confirmed the existence of TSs of clear eukaryotic origin in all four species. *In vitro* reconstitution and biochemical characterization revealed selective production of key halogenated terpene precursors, including the well-known plant terpenes myrcene, *trans*-β-ocimene, bisabolene, and germacrene D-4-ol described for the first time from algal enzymes. Finally, gene co-occurrence analysis revealed the first haloterpenoid associated biosynthetic gene pair linking a TS and an algal vanadium-dependent haloperoxidase. These results expand on a small, but growing number of characterized red algal terpene synthases and offer insight into the biosynthesis of iconic halogenated algal natural products.

3.4 Introduction

Terpenoids are a class of natural products ubiquitous in both marine and terrestrial organisms.[1] Specialized terpenoids have been characterized in plants,[2] fungi,[3] bacteria,[4] algae,[5] as well as insects[6] and many marine animals, like octocorals,[7] sponges,[8] and mollusks.[9] The pervasiveness of specialized terpenoids can be attributed to several factors, including the versatility of terpenoids as a source of chemical diversity and their numerous biological functions, such as signaling, defense, and attraction.[10] Terpenoids or isoprenoids are both synonymous for natural products constructed from two basic 5-

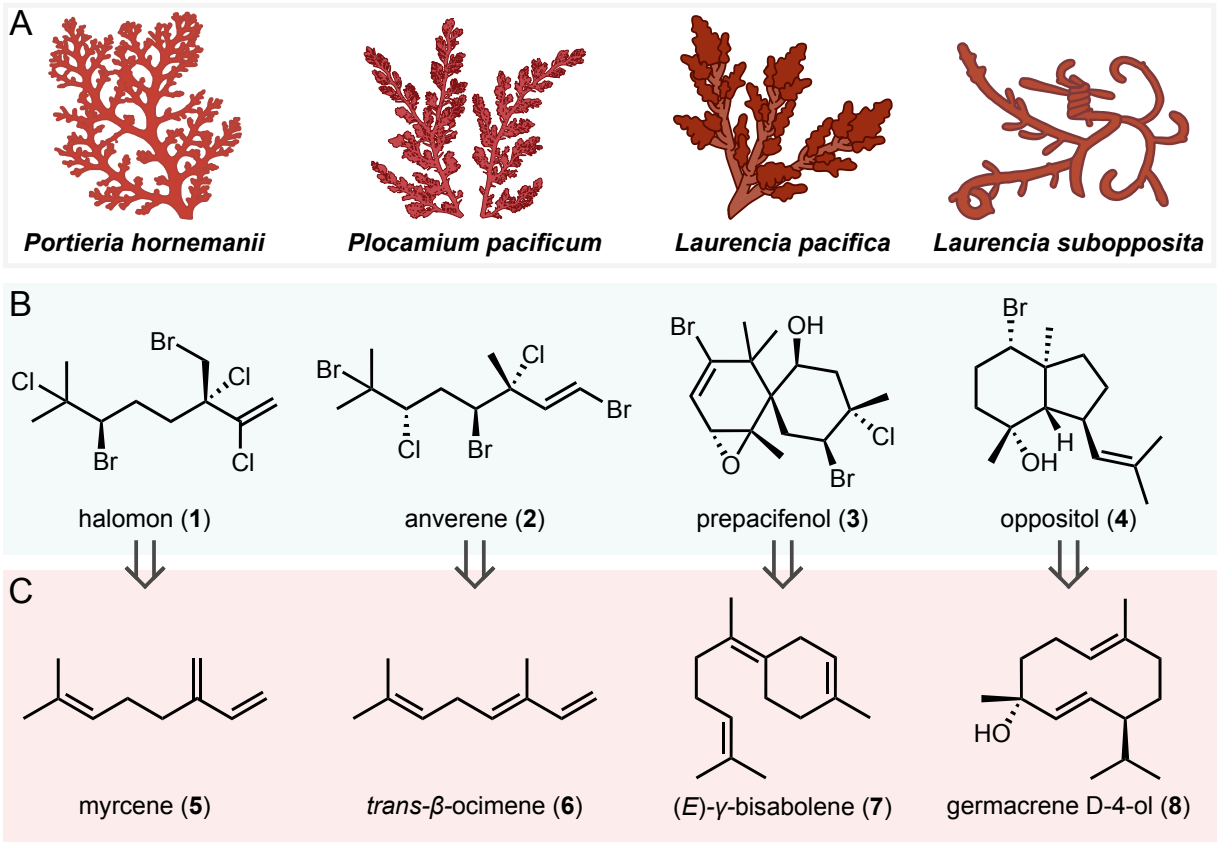


Figure 3.1 Haloterpenoid producing red macroalgae and representative associated metabolites. **A.** Red macroalgae included in this study **B.** Representative haloterpenoid secondary metabolites associated with each red macroalga in panel '**A**', samples selected includes two primarily monoterpene producing species (*P. hornemannii and P.* pacificum) and two primarily sesquiterpene producing species (*L. pacifica* and *L.* subopposita) **C.** Proposed scaffolds for each haloterpenoid metabolite shown in panel '**B**'.

carbon (C5) building blocks: dimethylallyl diphosphate (DMAPP) and isopentenyl diphosphate (IPP). DMAPP and IPP derive from two possible biosynthetic pathways, namely the mevalonate pathway (MEV) or the methylerythritol 4-phosphate (MEP) pathway.[11,12] The production of terpene scaffolds in terrestrial plants is known to be catalyzed by two groups of evolutionarily distinct type I of terpene synthases (TS), either typical plant-type TSs or a newly emerging class of microbial-type TSs (MTS).[13] Type I TSs utilize prenyl diphosphate substrates to generate an enormous diversity of carbon scaffolds, which in turn may be subject to downstream tailoring reactions, such as halogenation or oxidation.[14]

The biogenesis of halogenated terpenes has been of interest to the natural products community for decades since their initial discovery in the 1960s from the cosmopolitan genus *Laurencia* (**Figure 3.1A**).[15] Marine red algae, or seaweeds, are prominent producers of haloterpenoids, many of which exhibit potent differential cytotoxicity (**Figure 3.1B**). The first dedicated step in the biosynthesis of haloterpenoids is expected to be catalyzed by a terpene synthase, and retrobiosynthetic analysis predicts key terpenoid scaffolds for anticipated downstream tailoring reactions (**Figure 3.1C**). The production of haloterpenoids is a distinctive trait of red algae, but it is not uniformly distributed throughout Rhodophyta. Instead, haloterpenoid biosynthesis is mainly concentrated in just a few genera, namely *Laurencia*, *Plocamium*, and *Portieria*.[16]

Recent work demonstrated the sporadic occurrence of type I terpene synthases in red algae.[17] Red algal TSs are evolutionarily more closely related to microbial-type and microbial terpene synthases than to typical plant terpene synthase genes.[18,19] Transcriptomes of the genus *Laurencia,* in particular *Laurencia pacifica* and *Laurencia*

dendroidea, were each found to contain genes for three different type I MTSs, but only one was biochemically characterized.[20] The red algae *Porphyridium purpureum* and *Erythrolobus australicus* were each found to have one and two MTS, respectively, however these species are not known to produce haloterpenoids.[17] Overall, the distribution of these genes appears to not be widespread, with only a few, sparsely distributed species of the classes Porphyridiophyceae and Florideophyceae known to harbor MTSs or 7 out of 49 investigated species to-date (**Figure 3.S1**, **Table 3.S2**).

Here, we report ten new MTSs from *Plocamium pacificum, Portieria hornemannii,* and *Laurencia subopposita*. We also biochemically characterized the remaining MTSs from *L. pacifica.* Phylogenetic analysis demonstrated a clear delineation between algal and microbial genes and suggests a recent acquisition from bacteria via horizontal gene transfer (HGT).[17] Notably, in one case a terpene synthase gene was found to be physically co-localized with a vanadium-dependent haloperoxidase (VHPO) gene, the enzyme predicted to be responsible for the halogenation of algal haloterpenoids.[21,22] *In vitro* reconstitution and biochemical characterization of algal TS sequences revealed the selective production of both mono- and sesquiterpene scaffolds, including key halogenated terpene precursors. This work continues to expand the red algal terpene synthase family and provides insight into the biosynthetic pathways of algal halogenated natural products.

3.5 Results and Discussion

Sequencing and Genome Mining

We collected *L. pacifica, L. subopposita,* and *P. pacificum* from La Jolla, CA, USA and *P. hornemannii* from Oahu, HI, USA. Organisms were selected to capture a broad range of red algal haloterpenoid diversity regarding both structure and activity, with two predominantly sesquiterpene producers, *L. pacifica and L. subopposita,* and two predominately monoterpene producers, *P. pacificum* and *P. hornemannii* (**Figure 3.1**). We analyzed each algal sample by GC-MS and confirmed production of algal haloterpenoids with authentic standard molecules and found consistent production of haloterpenoids matching those previously reported in the literature.[1,5]

Following validation of haloterpenoid production, we sequenced the genome and transcriptome of each sample (**Table 3.S3, Figure 3.S2**). To identify red algal haloterpenoid biosynthesis genes, we queried algal sequencing data using an in-house Hidden Markov model (HMM) generated from a larger diversity of characterized TS sequences from octocorals, bacteria, and fungi.[7] To complement HMM results, we performed tBLASTn searches with known red algal microbial-type class I TS sequences as genetic hooks.[20] Querying the genome and transcriptome of each respective organism yielded a total of ten previously unknown well-scoring hits, all with canonical class I TS motifs (**Figure 3.S3**).[14] As expected, the three previously published MTSs were identified from new *L. pacifica* genomic data. Six new MTSs were identified from *L. subopposita*, three new MTS sequences from *P. hornemannii* and notably, only a single MTS was identified from *P. pacificum.* Phylogenetic analysis revealed a well-supported (bootstrap

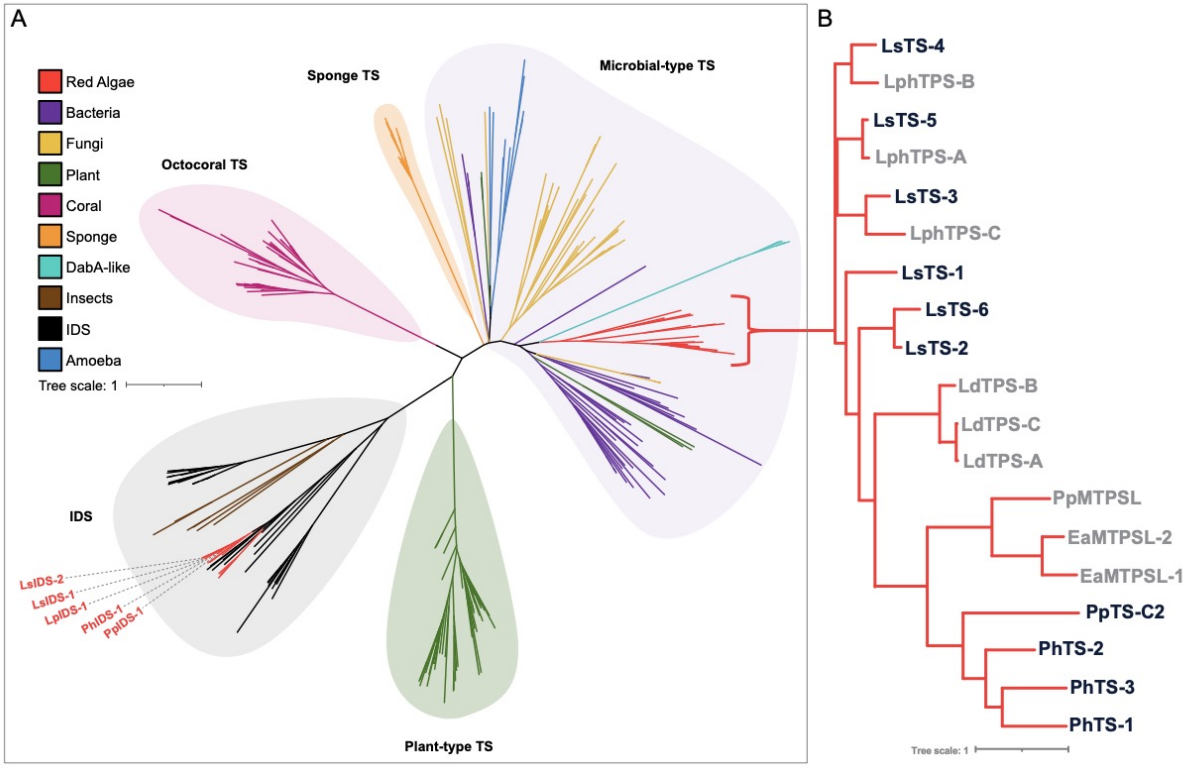


Figure 3.2 Phylogenetic analysis of red algal TS sequences. **A.** Phylogeny of TS sequences, including representative bacterial, fungal, amoebal, coral, sponge, and plant TS sequences. Putative red algal IDS sequences highlighted in red, sequences from this study are labeled. The scale measures evolutionary distances in substitutions per amino acid. **B.** Subclade of red algal sequences, the first two or three letters of each TS name represent the taxonomy of the macroalgae from which it originated. Names of new sequences are in black and previously identified sequences are in grey.

>99%) monophyletic clade comprised exclusively of red algal MTSs (**Figure 3.2A**); nomenclature of red algal TS sequences from the literature is maintained for continuity.[18,20] (**Figure 3.2B**). As expected, the expanded red algal clade nests within the broader microbial-type terpene synthase clade alongside bacterial and fungal TS sequences. This contrasts with recent work exploring sponge and coral MTSs, which form distinct monophyletic clades separate from bacterial and fungal MTSs, while still maintaining the overall shared structure and conserved key active site residues with bacterial sequences, suggesting an ancient HGT event. Similarly, growing evidence does

not support red algal type I TS evolution via a recent gene duplication and neofunctionalization of a red algal IDS like in insect TS evolution (**Figure 3.2A**).[23] Red algal IDS sequences from both publicly available genomes (**Table 3.S4**) and sequences from this study appear to be more closely related to other IDSs than to newly identified TS sequences. Overall, the sparse taxonomic distribution, nested monophyly within the broader microbial clade, and conserved active site residues supports a recent HGT event as the mode of evolution for red algal MTSs.

Red macroalgae are prolific producers of terpenoids, with roughly half (1081 of 2266) of known red algal compounds considered to be terpenoid in part.[1] The first dedicated step in the biosynthesis of non-triterpenoid based haloterpenoids is likely catalyzed by a type I terpene synthase. While three red algal TSs have been biochemically characterized, their reported products to-date do not appear to be biosynthetic intermediates of target haloterpenoids, such as halomon (**1**), anverene (**2**), prepacifenol (**3**), and oppositol (**4**) (**Figure 3.1**). We hypothesized the newly identified MTS sequences may produce on-pathway terpene scaffolds for these well-known halogenated products as they are phylogenetically unique from previously characterized sequences (**Figure 3.2A**).

Biochemical Characterization

Genes for heterologous expression were purchased from TWIST Biosciences as codon-optimized (IDT DNA algorithm) sequences cloned into pET 28a(+) vectors. To determine each enzyme's product, recombinant proteins were tested *in vitro* and screened for activity with common terpene diphosphate precursors (**Figure 3.S4**). The

resulting reaction products were analyzed with gas chromatography-mass spectrometry (GC-MS) and NMR spectroscopy (**Figure 3.S5-S16**, **Figure 3.S17-S24**). All recombinant proteins, except LsTS-6, showed TS activity (**Figure 3.3A, Figure 3.S14**). This is unsurprising as LsTS-6 was not present in the transcriptome of *L. subopposita* and appears to be truncated in the genome is *L. subopposita* (**Figure 3.S3**). Of the ten new

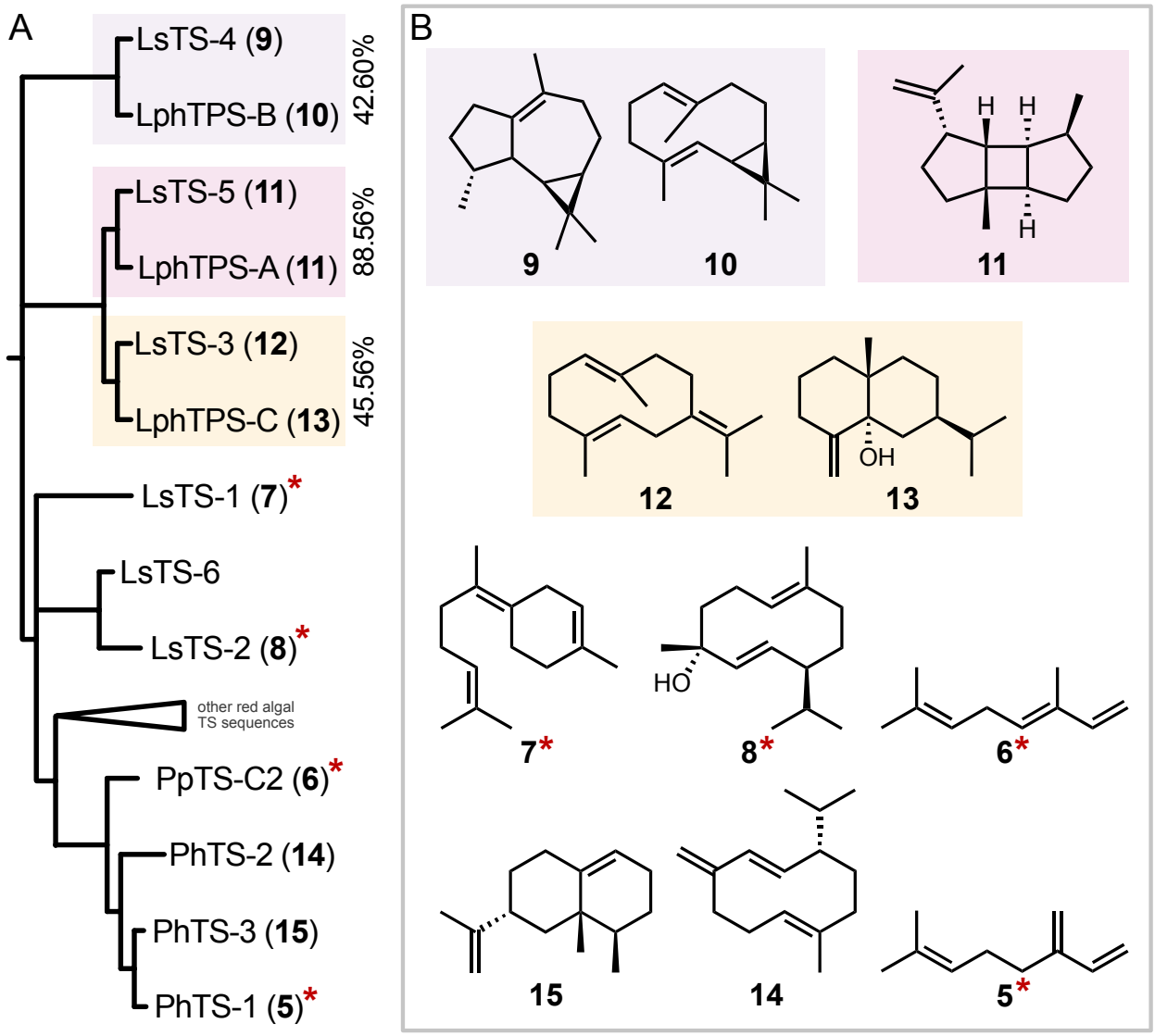


Figure 3.3 Algal TS biochemical characterization. **A.** Phylogeny of newly characterized red algal terpene synthases and **B.** their respective products. Algal sequences not included in this study are collapsed. Sequences highlighted with a red asterisk produce products predicted to be on pathway towards algal haloterpenoid products. Percent similarity by amino acid sequence is shown next to each *Laurencia* TS pair.

sequences and two uncharacterized *L. pacifica* sequences, nine acted as selective sesquiterpene synthases, converting farnesyl diphosphate (FPP) to cyclic terpene products (**Figure 3.3B**). The selective production of sesquiterpene products corresponds with the known metabolome of the genus *Laurencia*, which is predominately comprised of sesquiterpene terpene scaffolds.[1] Interestingly, the structural similarity of sesquiterpene products produced by the paired homologs between *L. subopposita* and *L. pacifica* reflects the percent similarity by amino acid sequence (**Figure 3.3A**). For example, LsTS-5 and LphTS-A share 88.56% identity by amino acid sequence and make the same bourbonane scaffold (**11**) (**Figure 3.S13**), whereas the homolog pairs LsTS-4/LphTPS-B and LsTS-3/LphTPS-C share 42.60% and 45.56% identity, respectively and make the related sesquiterpene scaffolds **9**/**10** and **12**/**13**, respectively (**Figure 3.S12/3.S15, Figure 3.S11/3.S16**). Despite *L. subopposita* not being a major producer of chamigrane haloterpenoids, LsTS-1 was found to make (*E*)-γ-bisabolene (**7**), the postulated precursor to the chamigrane class of haloterpenoids (**3**) abundant in the genus *Laurencia* (**Figure 3.S9**).[24] Notably, none of the TSs found in *L. pacifica* make a **7**-type scaffold, however, **7** is detected in the crude extract, pointing to either production by a microbial symbiont or involvement of a different enzyme class in terpene hydrocarbon production. LsTS-2 was found to selectively make germacrene D-4-ol (**8**), the hypothesized biosynthetic precursor to the haloterpenoid oppositol (**4**) (**Figure 3.S10**). Two of the enzymes, PpTS-C2 and PhTS-1, acted as selective monoterpene synthases, converting geranyl diphosphate (GPP) to acyclic terpene products (**Figure 3.S5, Figure 3.S6**), both of which are the principal scaffolds of haloterpenoid products in *P. pacificum* and *P. hornemannii*. The single TS identified from *P. pacificum* showed the selective

formation of *trans*-β-ocimene (**6**), the terpene proposed precursor to anverene (**2**), as well as all known terpenoids isolated from *Plocamium pacificum* and the closely related *Plocamium cartilagineum* from which anverene was isolated. The monoterpene producing enzyme obtained from *P. hornemannii* selectively forms the terpene scaffold β-myrcene (**5**), which is the hypothesized biosynthetic precursor of the notable red algal natural product, halomon (**1**), and other known haloterpenoids from *P. hornemannii*. Two additional TSs identified in both the genome and transcriptome of *P. hornemannii*, PhTS-2 and PhTS-3, showed the selective formation of germacrene D (**14**) and valencene (**15**), respectively, despite no sesquiterpenoids being reported from this organism (**Figure 3.S7, Figure 3.S8**).[1] Collectively, our experiments establish the functions of new red algal MTS genes and reveal the selective formation of key terpenoid scaffolds towards the biosynthesis of haloterpenoids in red macroalgae.

Haloterpenoid Gene Clustering

We next examined the genomic context of each terpene synthase. Annotating neighboring genes serves two primary purposes: to validate their origin as algal and to explore the physical colocalization, or clustering, of biosynthetic genes. Eukaryotic biosynthetic gene clusters (BGCs) have several distinguishing features from their microbial counterparts. Eukaryotic BGCs are typically spaced over several kilobases, have large intergenic regions, and are often flanked, or contain, viral retrotransposable elements.[25] Because of these features, identifying BGCs in eukaryotes is historically challenging. However, rapid innovations in long-read sequencing have greatly improved genome assembly and contiguity. In plants, certain natural product pathways, including

many terpenoids, are now understood to cluster biosynthetic genes.[26] Recently, the biosynthetic genes of numerous marine eukaryotic secondary metabolites have also been shown to cluster in the nuclear genome. For example, in octocorals, TS genes were found to be physically colocalized with genes encoding putative terpenoid tailoring enzymes.[7,27], In red algae, there are a handful of examples of gene clustering, however, algal clusters known to-date either contain VHPOs [28] or MTS genes.[29,30] It is important to note, the only MTS containing gene clusters that have been identified and biochemically validated are found in kainoid producing red macroalgae. These sequences are MTS-like, meaning they maintain MTS motifs and structure, but catalyze an *N*-prenylation reaction. Therefore, a haloterpenoid producing gene cluster from red macroalgae has yet to be identified.

To identify possible haloterpenoid BGCs in red algae, we used a targeted genome mining approach to screen the draft genome assemblies of *L. pacifica*, *L. subopposita*, *P. hornemannii* and *P. pacificum* with the Pfams for terpene synthases (PF19086), VHPOs (PF01569), and polyprenyl synthetases (PF00348). We complemented this dataset with the previously generated results from screening each genome with an in-house generated HMM targeting MTSs.[7] Within the *P. hornemannii* and *P. pacificum* genomes we did not identify any instances of a red algal TS co-clustered with any family of halogenating enzyme. Excitingly, we identified two sets of gene pairs in *L. pacifica* and *L. subopposita* (**Figure 3.4A and B**)**.** In *L. pacifica*, two of the biochemically characterized red algal TSs were found to cluster together 152kb apart on a single contiguous sequence 484kb in length (**Figure 3.4A**). In *L. subopposita,* one of the biochemically characterized red algal TSs, LsTS-2, was found to cluster with a VHPO and the fragmented sequence of LsTS-

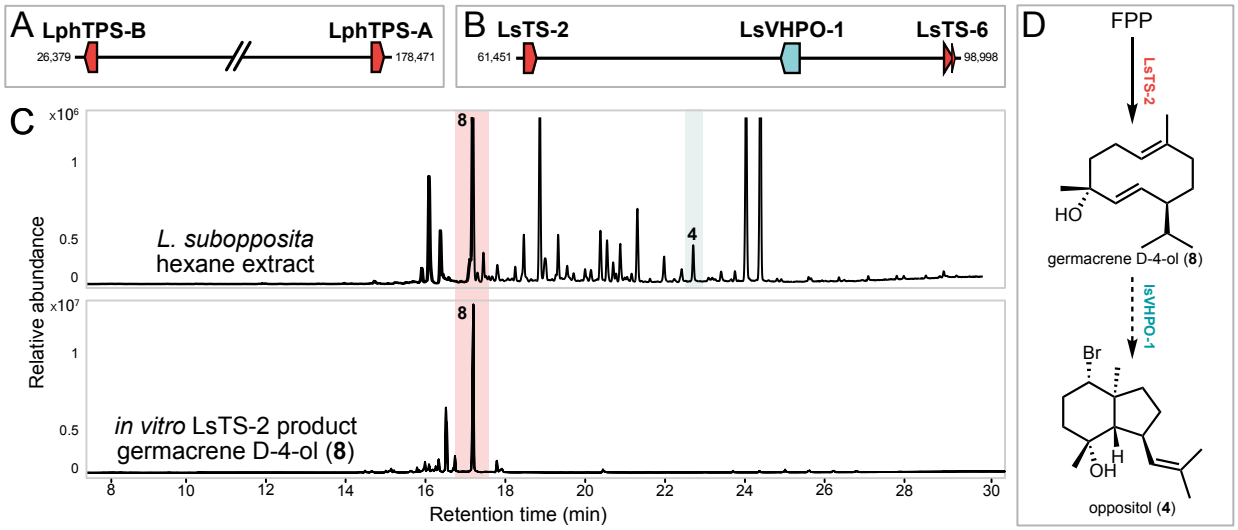


Figure 3.4 Assembled contigs from in *L. pacifica* and *L subopposita* show the physical co-localization of TSs with genes encoding for other TSs (red) or a VHPO (blue) **A.** Gene pair comprised of LphTPS-B and LphTPS-A **B.** Gene cluster comprised of LsTS-2, LsVHPO-1, and LsTS-6 **C.** GC-MS detection of gene pair associated sesquiterpenes in *L. subopposita* hexane extract and reconstitution of the production of germacrene D-4-ol (**8**) in *E. coli* by recombinant expression of LsTS-2. The red bar highlights sesquiterpene **8**; the blue bar highlights putative brominated sesquiterpene alcohol like **4**. **D.** Proposed retrobiosynthesis of **4**, the putative product of *L. subopposita* TS-VHPO gene cluster in panel 'B'.

6 (**Figure 3.4B**). This cluster was about 25kb upstream of one a coding sequence annotated as a glycosyltransferase using a Conserved Domain Database search; closest BLASTp hits support this sequence as algal in origin (**Table 3.S5**)**.** Both clusters are interspersed with retrotransposable (RT) elements and other mobile genetic elements common in eukaryotic genomes.[31] Thus, transposable element-mediated recombination may contribute to cluster formation in red macroalgae. The TS-VHPO gene cluster in *L. subopposita* was found to be transcriptionally active, and germacrene D-4-ol (**8**), the *in vitro* product of LsTS-2, is present in the crude hexane extract of *L. subopposita* (**Figure 3.4C**).

Retrobiosynthetic analysis of oppositol (**4**) supports a two-enzyme biosynthesis in agreement with the TS-VHPO gene pair, where **4** could be generated from **8** in one

halogenation step (**Figure 3.4D**). VHPOs are known to catalyze electrophilic halogenation reactions in many different organisms, including red algae, with several examples of C-C bond formation via a bromonium-induced halocyclization cascade similar to the hypothesized germacrene D-4-ol (**8**) reaction (**Figure 3.S25**).[22,32,33] Further, the

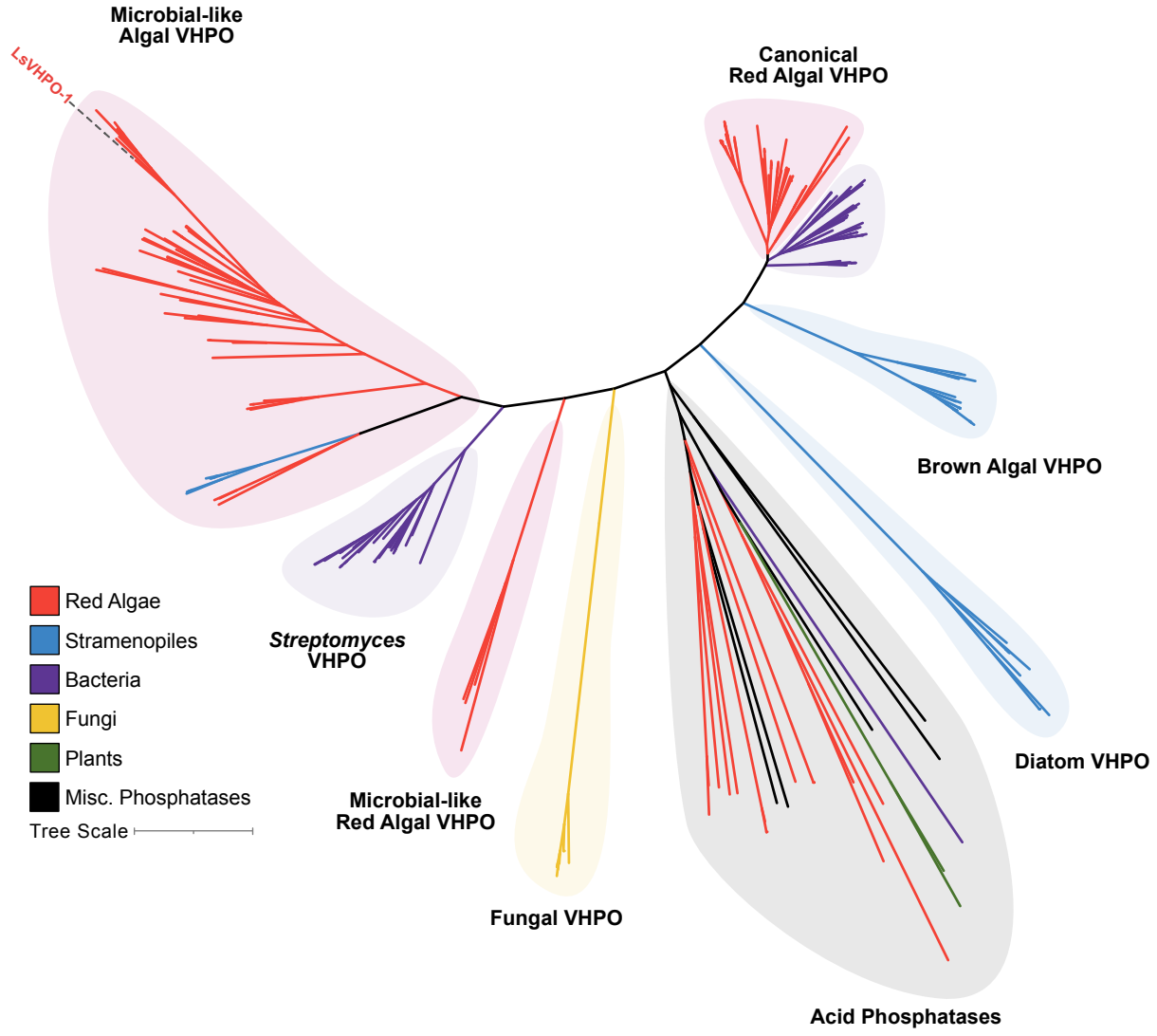


Figure 3.5. Vanadium-dependent haloperoxidase phylogeny including representative bacterial, fungal, algal, and plant VHPO sequences. Putative red algal VHPO sequences highlighted in red, LsVHPO-1 is labeled. The scale measures evolutionary distances in substitutions per amino acid.

stereocenters identified in germacrene D-4-ol produced by LsTS-2 show the same configuration as in the putative biosynthetic product oppositol (**4**) (**Figure 3.S25**).[1]

To generate a broader comparison of LsVHPO-1 relative to other VHPO enzymes, we constructed a maximum-likelihood phylogenetic tree using newly generated haloterpenoid producing red algal sequencing datasets and select publicly available VHPO sequences (**Table 3.S4**), acid phosphatase sequences were included as an outgroup. LsVHPO-1 was found to clade tightly with sequences from the genus *Laurencia* and other red algal sequences from new and public datasets. Interestingly, this group of red algal VHPOs appears to be distinct, but more closely related to microbial VHPOs from the genus *Streptomyces*. VHPOs characterized from streptomyces are involved in the biosynthetic pathways of meroterpenoids like the merochlorins[2] and the napyradiomycins.[3] In the biosynthesis of the latter, VHPOs catalyze bromonium induced cyclization reactions. Previously reported VHPOs from red algae, like the biochemically validated VHPO from *Corallina officinalis*[34], appear to form a second clade of VHPOs adjacent to brown algal sequences. This discovery sheds first light on haloterpenoid biosynthesis in red algae, and ongoing investigations of the new microbial-like VHPO clade, including further study of the co-clustered VHPO enzyme LsVHPO-1 will shed additional light on the biosynthesis of red algal haloterpenoids.

3.6 Conclusions

In summary, we report the discovery of the first terpene synthases from the haloterpenoid producing red macroalgae *L. subopposita, P. pacificum, and P. hornemannii*, and resolve previously uncharacterized sesquiterpene-synthases from *L.*

pacifica by the combination of transcriptomics, genomics, and *in vitro* biochemical characterization. Here, we showed red macroalgae can produce both mono- and sesquiterpenes via class I microbial-type terpene synthases and that these terpene scaffolds are likely on pathway towards haloterpenoids of interest. Additionally, this study shows the occurrence of a gene pair in red macroalgae, which potentially encodes enzymes for haloterpenoid biosynthesis. This discovery adds to a growing number of eukaryotic BGCs, and future studies will focus on the systematic characterization of the remaining haloterpenoid associated biosynthesis enzymes. Further sequencing efforts of haloterpenoid producing red macroalgae could expedite the discovery of haloterpenoid chemistry and biochemistry in source-limited non-model organisms and expand the evolutionary history of terpene synthases in Rhodophyta and enrich our understanding of gene clustering in higher organisms for specialized pathways.

3.7 Main Text References

1. Dictionary of Natural Products (April 13, 2023).

2. F. Chen, D. Tholl, J. Bohlmann, E. Pichersky, The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* **66**, 212–229 (2011).

3. M. B. Quin, C. M. Flynn, C. Schmidt-Dannert, Traversing the fungal terpenome. *Nat. Prod. Rep.* **31**, 1449–1473 (2014).

4. J. S. Dickschat, Bacterial terpene cyclases. *Nat. Prod. Rep.* **33**, 87–110 (2015).

5. R. E. Moore, E. Moore, J. A. Pettus, M. S. Doty, A. Abbott, E. H. Williamson, A. Volatile compounds from marine algae. *Ethnobotanical* (1976).

6. F. Beran, T. G. Köllner, J. Gershenzon, D. Tholl, Chemical convergence between plants and insects: Biosynthetic origins and functions of common secondary metabolites. *New Phytol.* **223**, 52–67 (2019).

7. I. Burkhardt, T. de Rond, P. Y.-T. Chen, B. S. Moore, Ancient plant-like terpene biosynthesis in corals. *Nat. Chem. Biol.* **18**, 664–669 (2022).

8. K. Wilson, T. de Rond, I. Burkhardt, T. S. Steele, R. J. B. Schäfer, S. Podell, E. E. Allen, B. S. Moore, Terpene biosynthesis in marine sponge animals. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2220934120 (2023).

9. N. Nocchi, A. R. Soares, M. L. Souto, J. J. Fernández, M. N. Martin, R. C. Pereira, Detection of a chemical cue from the host seaweed *Laurencia dendroidea* by the associated mollusc *Aplysia brasiliana*. *PLOS ONE* **12**, e0187126 (2017).

10. M. Avalos, P. Garbeva, L. Vader, G. P. van Wezel, J. S. Dickschat, D. Ulanova, Biosynthesis, evolution and ecology of microbial terpenoids. *Nat. Prod. Rep.* **39**, 249–272 (2022).

11. G. Flesch, M. Rohmer, Prokaryotic hopanoids: the biosynthesis of the bacteriohopane skeleton. *Eur. J. Biochem.* **175**, 405–411 (1988).

12. J. Lombard, D. Moreira, Origins and early evolution of the mevalonate pathway of isoprenoid biosynthesis in the three domains of life. *Mol. Biol. Evol.* **28**, 87–99 (2011).

13. G. Li, T. G. Köllner, Y. Yin, Y. Jiang, H. Chen, Y. Xu, J. Gershenzon, E. Pichersky, F. Chen, Nonseed plant *Selaginella moellendorffii* has both seed plant and microbial types of terpene synthases. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14711–14715 (2012).

14. D. W. Christianson, Structural and chemical biology of terpenoid cyclases. *Chem. Rev.* **117**, 11570–11648 (2017).

15. Irie, T. , M. Suzuki. E. Kurosawa, and T. Masamune. Laurinterol and debromolaurinterol, constituents from *Laurencia intermedia*. *Tetrahedron Lett*. 1837-1840 (1966).

16. C. Lyu, T. Chen, B. Qiang, N. Liu, H. Wang, L. Zhang, Z. Liu, CMNPD: a comprehensive marine natural products database towards facilitating drug discovery from the ocean. *Nucleic Acids Res* **49**, D509–D515 (2021).

17. G. Wei, Q. Jia, X. Chen, T. G. Köllner, D. Bhattacharya, G. K.-S. Wong, J. Gershenzon, F. Chen, Terpene biosynthesis in red algae is catalyzed by microbial type but not typical plant terpene synthases. *Plant Physiol.* **179**, 382–390 (2019).

18. Q. Jia, G. Li, T. G. Köllner, J. Fu, X. Chen, W. Xiong, B. J. Crandall-Stotler, J. L. Bowman, D. J. Weston, Y. Zhang, L. Chen, Y. Xie, F.-W. Li, C. J. Rothfels, A. Larsson, S. W. Graham, D. W. Stevenson, G. K.-S. Wong, J. Gershenzon, F. Chen, Microbial-type terpene synthase genes occur widely in nonseed land plants, but not in seed plants. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12328–12333 (2016).

19. Q. Jia, T. G. Köllner, J. Gershenzon, F. Chen, MTPSLs: New terpene synthases in nonseed plants. *Trends Plant Sci* **23**, 121–128 (2018).

20. R. D. Kersten, S. Lee, D. Fujita, T. Pluskal, S. Kram, J. E. Smith, T. Iwai, J. P. Noel, M. Fujita, J.-K. Weng, A red algal bourbonane sesquiterpene synthase defined by microgram-scale NMR-coupled crystalline sponge X-ray diffraction analysis. *J. Am. Chem. Soc.* **139**, 16838–16844 (2017).

21. A. Butler, J. N. Carter-Franklin, The role of vanadium bromoperoxidase in the biosynthesis of halogenated marine natural products. *Nat. Prod.* **21**, 180–188 (2004).

22. J. N. Carter-Franklin, A. Butler, Vanadium bromoperoxidase-catalyzed biosynthesis of halogenated marine natural products. *J. Am. Chem. Soc.* **126**, 15060–15066 (2004).

23. F. Beran, P. Rahfeld, K. Luck, R. Nagel, H. Vogel, N. Wielsch, S. Irmisch, S. Ramasamy, J. Gershenzon, D. G. Heckel, T. G. Köllner, Novel family of terpene synthases evolved from trans-isoprenyl diphosphate synthases in a flea beetle. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2922–2927 (2016).

24. J. J. Sims, William. Fenical, R. M. Wing, Phillip. Radlick, Marine natural products. IV. Prepacifenol, a halogenated epoxy sesquiterpene and precursor to pacifenol from the red alga, *Laurencia filiformis*. *J. Am. Chem. Soc.* **95**, 972–972 (1973).

25. B. Field, A. E. Osbourn, Metabolic diversification—independent assembly of operon-like gene clusters in different plants. *Science* **320**, 543–547 (2008).

26. H.-W. Nützmann, A. Huang, A. Osbourn, Plant metabolic clusters – from genetics to genomics. *New Phytol.* **211**, 771–789 (2016).

27. P. D. Scesa, Z. Lin, E. W. Schmidt, Ancient defensive terpene biosynthetic gene clusters in the soft corals. *Nat. Chem. Biol.* **18**, 659–663 (2022).

28. H. R. Thapa, Z. Lin, D. Yi, J. E. Smith, E. W. Schmidt, V. Agarwal, Genetic and biochemical reconstitution of bromoform biosynthesis in *Asparagopsis* lends insights into seaweed reactive oxygen species enzymology. *ACS Chem Biol* **15**, 1662–1670 (2020).

29. T. S. Steele, J. K. Brunson, Y. Maeno, R. Terada, A. E. Allen, M. Yotsu-Yamashita, J. R. Chekan, B. S. Moore, Domoic acid biosynthesis in the red alga *Chondria armata* suggests a complex evolutionary history for toxin production. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2117407119 (2022).

30. J. R. Chekan, S. M. K. McKinnie, M. L. Moore, S. G. Poplawski, T. P. Michael, B. S. Moore, Scalable biosynthesis of the seaweed neurochemical, kainic acid. *Angew. Chem., Int. Ed.* **58**, 8454–8457 (2019).

31. J. N. Wells, C. Feschotte, A field guide to eukaryotic transposable elements. *Annu. Rev. Genet.* **54**, 539–561 (2020).

32. P. Bernhardt, T. Okino, J. M. Winter, A. Miyanaga, B. S. Moore, A stereoselective vanadium-dependent chloroperoxidase in bacterial antibiotic biosynthesis. *J. Am. Chem. Soc.* **133**, 4268–4270 (2011).

33. D. Mondal, B. F. Fisher, Y. Jiang, J. C. Lewis, Flavin-dependent halogenases catalyze enantioselective olefin halocyclization. *Nat. Commun.* **12**, 3268 (2021).

34. J. N. Carter, K. E. Beatty, M. T. Simpson, A. Butler, Reactivity of recombinant and mutant vanadium bromoperoxidase from the red alga *Corallina officinalis*. *J. Inorg. Biochem.* **91**, 59–69 (2002).

3.8 Supplementary Information

General Methods and Chemical Methods

All chemicals and solvents were used as received from the commercial supplier (SigmaAldrich, Fisher, or TCI). The substrates GPP, FPP, and GGPP were synthesized as reported previously.[1,2] Thin layer chromatography was performed with thin layer chromatography plates purchased from Merck (silica gel 60 F254, glass plates). Column chromatography was performed using a Flash EZ prep combiflash system (Teledyne ISCO). GC–MS analysis was performed on an Agilent 7890A gas chromatograph with an Agilent 5975C mass spectrometer using an Rtx-5Sil 30 m × 0.25 mm, 0.25-μm column. The oven program was as follows: hold 70 °C for 3 min, 10 °C min^{-1} to 325 °C and hold at 325 °C for 3 min. The flow rate was 1 mL min^{-1}, and the injection volume was 1 μl. The splitless injection inlet temperature was 270 °C, the MS transfer line temperature was 280 °C, the MS source temperature was 230 °C and the MS quad temperature was 150 °C. Data were analyzed using a Masshunter B06.00. For the detection of bicyclogermacrene (**10**) the same temperature program was used but the inlet temperature was lowered to 200 °C and the transfer line temperature was lowered to 250 °C. NMR spectra were recorded on a Bruker Avance III spectrometer (600 MHz) using a 1.7-mm inverse detection triple resonance (H-C/N/D) cryoprobe or a JEOL ECZ spectrometer (500 MHz) and were referenced against deuterated benzene (δ = 7.16 ppm) or deuterated chloroform (δ = 7.26 ppm) for 1H-NMR and deuterated benzene (δ = 128.06 ppm) or deuterated chloroform (δ = 77.16 ppm) for 13C-NMR. Data for NMR spectra are reported as follows: shift (δ) in ppm; s, singlet; d, doublet; t, triplet; q, quartet; m, multiplet or unresolved; br, broad signal; *J*, coupling constant(s) in Hz. Data were

analyzed using MestreNova 12.0.0. Optical rotation measurements were performed using a Jasco P-2000 polarimeter.

All protein purification from *Escherichia coli* was performed on an ÄKTA Pure 25 L1 instrument (Cytiva) with a fraction collector F9-C and sample pump S9 with all solvents filtered through a nylon 0.2 µm GDWP membrane (Merck) prior to use. FPLC data was analyzed with UNICORN version 7 software. All protein quantification was done by method of Bradford using the Protein Assay Dye Reagent Concentrate (Bio-Rad) on protein sample dilutions in MilliQ water.

Molecular Biology and Biochemical Methods

Nucleotide Extraction

Portieria hornemannii was collected from Honolulu, HI, USA (21.256947, -157.79708) and *Plocamium pacificum, Portieria hornemannii, Laurencia pacifica*, and *Laurencia subopposita* were collected from La Jolla, CA, USA (32.839909, -117.28243) at low tide from the intertidal zone. The samples were snap-frozen in liquid nitrogen and stored at -80 ˚C. High-molecular weight (HMW) DNA sample preparation protocols for *Portieria hornemannii, Plocamium pacificum, Laurencia subopposita,* and *Laurencia pacifica* HMW DNA was extracted according to the protocol reported in full in Chapter 4 and published here: dx.doi.org/10.17504/protocols.io.14egn2dnpg5d/v1 Extracted DNA was further purified, and size selected using the BluePippin system with a High Pass Plus 15 kb cassette (Sage Science).

Total RNA was extracted with the QIAGEN RNeasy Plant isolation kit according to the manufacturer's instructions. RNA quality was assessed by Agilent Bioanalyzer.

Illumina Sequencing

P. pacificum, L. pacifica, and *L. subopposita* sequencing libraries were constructed using an Illumina TruSeq DNA PCR-free library kit using standard protocols. DNA was sheared to 350 base-pairs using a Covaris Focused-Ultrasonicator (LE220) and indexed adaptors ligated to the sheared fragments. Libraries were sequenced on a NovaSeq 6000 using paired ends and a read length of 150 base pairs.

P. hornemannii NEBnext sequencing libraries were generated to polish genome assemblies (New England Biolabs, Beverly, MA, USA). NEBnext sequencing libraries were created with 100 ng of DNA and quality-controlled on a bioanalyzer. Resulting libraries were sequenced on an Illumina MiSeq 2x150 bp to check quality and quantity (Illumina, San Diego, CA). The libraries were then sequenced on an Illumina NovaSeq S4 200 (PE100, 10x) run that resulted in 64 Gb of sequence. NCBI Sequence Read

For *P. pacificum, P. hornemannii,* and *L. subopposita* RNA-seq, a strand-specific mRNA library was prepared and sequenced on an Illumina NovaSeq S4 200 (PE100, 10x) run resulting in approximately 200 million reads per sample.

Oxford Nanopore Sequencing

Size-selected, high-molecular weight (HMW) DNA was sequenced using the Oxford Nanopore PromethION platform (Oxford Nanopore Technologies (ONT), Oxford, UK). A library was prepared using the Ligation Sequencing Kit V14 (SQK-LSK114) (ONT, Oxford, UK) and loaded onto an R10.4.1 flowcell. The sample was run with MinKnow version 22.10.7, High Accuracy (HAC) Basecalling mode, and 260 bps pore speed for 48 hr, resulting in 7.4 Gb of sequence with a read length N50 of 37 kb.

PacBio Sequencing

P. pacificum, L. pacifica, and *L. subopposita* Long-read sequencing was performed using Circular Consensus Sequencing (CCS) mode on a PacBio Sequel II instrument. DNA was sheared on a Diagenode Megaruptor and libraries were constructed using a SMRTbell Template Prep Kit 3.0 followed by sizing (10-50kb) on a SAGE Blue Pippin instrument. Sequencing was performed using a 30 hour movie time with 2 hour pre-extension and the resulting raw data was processed using the CCS6 algorithm.

L. subopposita was also sequenced with a modified low input PacBio protocol. The library was started with 500ng of DNA, sheared on a Diagenode Megaruptor and libraries were constructed using a SMRTbell Template Prep Kit 2.0 with extended damage repair time. After bead sizing to remove small fragments, sequencing was performed using a 30 hour movie time with 2 hour pre-extension and the resulting raw data was processed using the CCS6 algorithm.

Genome and Trascriptome Assembly

For transcriptomic analysis, all samples were processed as follows: RNA-seq read quality was assessed with FastQC[3] and trimmed with Trimmomatic v0.3[4]. The resulting trimmed and filtered reads were assembled by Trinity v2.9.1[5] using the following parameters: --seqType fq --max_memory 500G --CPU 12.

For genome assembly, Illumina reads were processed as above, and an initial short read assembly was generated by SPAdes v3.14.0[6] using the following parameters: metaspades.py -k 21,33,55,77,99,127 --memory 650 --threads 32. ONT and PacBio

reads were assembled by Flye[7] using the following parameters: --nano-raw or--pacbio-raw --threads 24 –meta. The resulting assemblies were polished three times with Racon v1.5.0[8] and twice with Pilon v1.23[9].

Genome completeness was assessed with BUSCO v4.0.5[10] using the eukaryota benchmarking universal single-copy orthologs eukaryota_odb10 dataset. This method was applied to the following publicly available red macroalgal genomes: Agarophyton vermiculophyllum (AgarVerm_1.0), Asparagopsis taxiformis (ASM1839795v1), Chondrus crispus (ASM35022v2), Digenea simplex (ASM479842v1), Gracilariopsis chorda (GraCho1.0), Gracilariopsis lemaneiformis (Glem_v01), Kappaphycus alvarezii (ASM220596v3), Neoporphyra haitanensis (ASM982973v1), Neopyropia yezoensis (ASM982973v1), *Chondria armata* (SIO_Carma_v2.2), and Porphyra umbilicalis (P_umbilicalis_v1) – to establish a relative comparison of *P. pacificum, P. hornemannii, L. pacifica,* and *L. subopposita* genome assemblies.

Phylogenetic Analysis

All sequence alignments were generated with Kalign v.2.04[11] and phylogenetic analysis was perform using IQ-TREE multicore v1.6.12[12] with the following parameters: -bb 1000 -nt AUTO -m 'LG+R6'. Sequences included in TS analysis are previously reported and additional red algal IDS sequences are summarized in Table 3.S4.[13,14] Sequences included in the VHPO analysis are selected from InterPro familes IPR016119 (Bromoperoxidase/chloroperoxidase C-terminal) and IPR036938 (Phosphatidic acid phosphatase type 2/haloperoxidase superfamily). VHPO sequences are summarized in Table 3.S4. All trees were visualized in iTOL.[15]

Protein Expression and Purification

Genes for heterologous expression were purchased from TWIST Biosciences as codon-optimized (Integrated DNA Technologies) sequences cloned into pET 28a(+) vectors (cut sites XhoI and NdeI). Protein expression plasmids were chemically transformed into E. coli BL21 (DE3) and selected on LB agar plates with 50 μg mL^{-1} kanamycin. A single colony was inoculated into a 5 mL LB starter culture and grown at 37 °C with 220 rpm shaking overnight. 1 L of Terrific Broth (Fisher Bioreagents) was then inoculated with 1 vol% of *E. coli* overnight culture for each respective TS. Cultures were shaken (180 rpm) at 37 ˚C to an optical density at 600 nm (OD600) of ~0.6-0.8. Then, cultures were chilled to 18 ˚C and induced with 300 μM of isopropylthio-β-galactoside (IPTG). Flasks were shaken overnight (180 rpm, ~18 hr) and cells were harvested by centrifugation (8000 xg, 15 min). Cell pellets were resuspended in 25 mL of lysis buffer (20 mM Tris, 300 mM NaCl, 10 mM imidazole, 10% glycerol, pH 8) and stored at -80 ˚C for future purification or processed immediately.

All frozen cell pellets were defrosted, and cells were lysed by sonication using a Qsonica 6 mm tip at 50% amplitude for 20 cycles of 15 seconds on and 45 seconds off, gently mixing after 15 cycles. Cell lysate was then centrifuged for 45 mins at 20,000 xg to pellet cell debris. Initial immobilized metal-affinity chromatography (IMAC) purification of TS enzymes was performed similarly for all soluble protein constructs. Briefly: clarified lysate was loaded at 2 mL/min onto a 5 mL HisTrap FF column (Cytiva) using a pre-equilibrated with wash buffer (20 mM Tris, 300 mM NaCl, 30 mM imidazole, pH 8). The loaded column was washed with 10 column volumes of 8% elution buffer (20 mM Tris,

300 mM NaCl, 250 mM imidazole, pH 8). Protein was eluted using a linear gradient of 8-100% elution buffer over 15 column volumes in 4 mL fractions. Fractions were assessed using SDS-PAGE, and target protein containing fractions were pooled and concentrated using Amicon Ultra-15 30 kDa-cutoff centrifugal filters (Millipore Sigma). TS enzymes were further purified, and buffer exchanged using PD-10 Desalting Columns (Sephadex G-25 M, Cytvia) pre-equilibrated with storage buffer (50 mM HEPES, 250 mM NaCl, 5 mM $MgCl_2$,10% glycerol, pH 7.8) according to the manufacturer's instructions. Purity was checked using SDS-PAGE and relevant fractions were pooled, concentrated, and stored immediately at -80 ˚C.

Enzymatic Assays

Analytical scale assays were performed in 1 mL of TS reaction buffer (50 mM HEPES, 5 mM $MgCl_2$, 250 mM NaCl, 10% glycerol, pH 7.8) with an enzyme concentration of 3 μM and 0.25 mg mL^{-1} GPP, FPP or GGPP to test mono-, sesqui-, and diterpene synthase activity. The assays were run for 18 h at room temperature and extracted with hexanes (200 μl). The extracts were dried over $MgSO_4$ and analyzed by GC–MS.

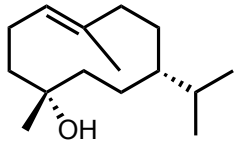
For product isolation, GPP and FPP (60 mg) were dissolved in ammonium bicarbonate solution (50 mL, 50 mM) and added dropwise over 8 h to 300 mL of TS reaction buffer containing 1–6 μM enzyme at room temperature. The assay was run for 18 h at room temperature and extracted with pentane (3 × 200 mL). The organic layer was dried over $MgSO_4$ and filtered and concentrated in vacuo to 0.5 mL. The crude product was purified by silica gel column chromatography. Fractions were checked by thin layer

chromatography (stained with phosphomolybdic acid stain), and product-containing fractions were pooled, concentrated to dryness, and analyzed by NMR spectroscopy.

Structure elucidation

Germacrene D-4-ol

Following the general procedure for product isolation the crude product obtained from the extracted assay of LsTS-2 was purified by column chromatography using hexane/EtOAc 5:1 to give germacrene D-4-ol (**8**, 4 mg). ^{1}H- and ^{13}C- NMR spectra were recorded and compared to literature data, showing identical spectra.[16] The absolute configuration was assigned as (+)-germacrene D-4-ol by optical rotation and comparison to literature data.[16]



Column chromatography: Hexanes/ EtOAc (5:1), **yield**: 4.0 mg (**13**%).
Optical rotatory power: $[\alpha]_D^{25.0}$ = +30.5 (*c* 0.18, $CHCl_3$). **^{1}H-NMR** (500 MHz, C_6D_6): δ = 5.25 (dd, $^{3}J_{H,H}$ = 15.7, 10.0, 1 H), 5.01 (d, $^{3}J_{H,H}$ = 15.7, 1 H), 4.92 (d, $^{3}J_{H,H}$ = 11.6, 1H), 2.62 (dddd, $^{3}J_{H,H}$ = 14.6, 12.8, 11.6, 3.2, 1 H), 2.27-2.21 (m, 1 H), 1.95-1.86 (m, 2 H), 1.54 (s, 3 H), 1.44 (dt, $^{3}J_{H,H}$ = 13.8, 3.6, 1 H), 1.41-1.35 (m, 1 H), 1.33-1.23 (m, 3 H), 1.07 (s, 3 H), 0.90 (d, $^{3}J_{H,H}$ = 6.7, 3 H), 0.86 (d, $^{3}J_{H,H}$ = 6.8, 3 H). **^{13}C-NMR** (125 MHz, C_6D_6): δ = 140.5, 132.4, 129.3, 126.0, 72.7, 55.1, 41.7, 40.0, 33.4, 31.3. 26.5, 24.2, 21.0, 19.2, 16.8. **EI-MS** (70 eV): *m*/*z* (%) = 222 (1), 207 (12), 205 (3), 204 (16), 189 (5), 162 (9), 161 (50), 149 (3), 147 (5), 137 (4), 135 (7), 134 (7), 133 (10), 131 (3), 124 (6), 123 (23), 122 (5), 121 (16), 120 (6), 119 (21), 117 (4), 115 (3), 111 (3), 109 (14), 108 (5), 107 (11), 106 (5), 105 (30), 97 (7), 95 (17), 94 (5), 93 (20), 92 (5), 91 (20), 83 (5), 82 (10), 81 (100), 80 (16), 79 (19), 77 (11), 71 (11), 69 (11), 67 (12), 65 (4), 55 (12), 53 (7).

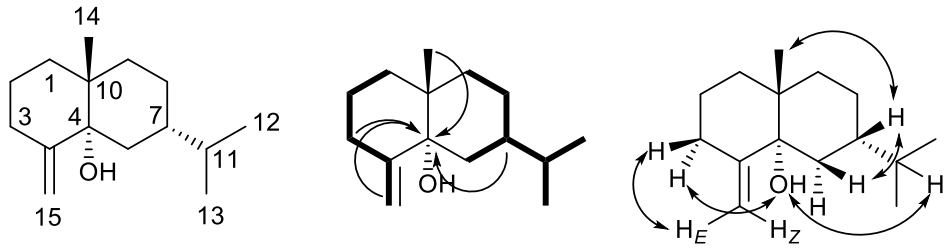
(5S,7S,10R)-Selin-4(14)-en-5-ol

Following the general procedure for product isolation the crude product obtained from the extracted assay of LphTS-B was purified by column chromatography using hexane/EtOAc 9:1 to give (5S,7S,10R)-Selin-4(14)-en-5-ol (**13**, 4 mg). The structure including relative configuration was assigned using one- and two-dimensional NMR data (see **Table 3.S1**). The obtained NMR data are identical with a report in the literature,[17] but the stereocenter at carbon 7 was inverted based upon NOE correlations to the shown structure.

Table 3.S1 NMR data for LpTPS-C product (500MHz, $CDCl_3$), in line with literature

C	^{13}C	^{1}H
1	34.9	1.78-1.73 (m); 1.03-0.98 (m)
2	22.3	1.71-1.61 (m); 1.60-1.53 (m)
3	31.6	2.60 (dtt, ^{2}J = 13.4, ^{3}J = 6.2, ^{4}J = 1.9) 2.07 (ddt, ^{2}J = 13.5, ^{3}J = 5.1, ^{4}J = 1.8)
4	152.6	-
5	76.6	-
6	30.6	1.85 (dd, ^{2}J = 14.7, ^{3}J = 6.3) 1.78-1.76 (m)
7	40.9	1.37-1.31 (m)
8	22.4	1.73-1.68 (2H, m)
9	30.9	1.73-1.73 (m)
10	38.3	-
11	29.1	2.12 (dsept, ^{3}J = 10.6, ^{3}J = 6.6)
12*	22.9	0.90 (d, ^{3}J = 6.5)
13*	22.2	0.92 (d, ^{3}J = 6.6)
14	20.3	0.90 (s)
15	107.2	4.76 (H_E, t, ^{4}J = 1.9) 4.67 (H_Z, t, ^{4}J = 1.7)

*: Carbons with designated numbers are interchangeable because very small differences in ^{1}H and ^{13}C shifts.



Column chromatography: Hexanes/ EtOAc (9:1), **yield**: 4.0 mg (10%).
Optical rotary power: $[\alpha]_D^{25.0}$ = +497.8 (*c* 0.45, $CHCl_3$). **TLC** (hexanes / EtOAc 9:1): R_f = 0.5. **NMR**: See **Table 3.S1**. **EI-MS** (70 eV): *m*/*z* (%) = 222 (14), 207 (3), 205 (7), 204 (33), 191 (9), 189 (17), 179 (7), 175 (4), 164 (4), 163 (4), 162 (15), 161 (100), 159 (4), 153 (9), 152 (5), 149 (4), 147 (9), 145 (5), 139 (4), 137 (5), 135 (11), 134 (6), 133 (50), 131 (8), 129 (5), 128 (7), 125 (6), 124 (11), 123 (10), 122 (7), 121 (14), 120 (6), 119 (34), 118 (3), 117 (14), 115 (10), 109 (20), 108 (6), 107 (18), 106 (8), 105 (69), 104 (4), 103 (6), 97 (7), 96 (10), 95 (50), 94 (8), 93 (29), 92 (8), 91 (60), 83 (10), 82 (13), 81 (50), 80 (4), 79 (29), 78 (5), 77 (22), 71 (4), 70 (3), 69 (19), 68 (5), 67 (28), 65 (9), 55 (26), 53 (12).

Table 3.S2 Distribution of terpene synthase sequences in Rhodophyta

Organism	Class	Order	TS
Porphyra umbilicalis	Bangiophyceae	Bangiales	
Porphyra yezoensis	Bangiophyceae	Bangiales	
Pyropia yezoensis	Bangiophyceae	Bangiales	
Compsopogon caeruleus	Compsopogonophyceae	Compsopogonales	
Madagascaria erythrocladioides	Compsopogonophyceae	Erythropeltidales	
Rhodochaete parvula	Compsopogonophyceae	Rhodochaetales	
Cyanidio schyzonmerolae	Cyanidiophyceae	Cyanidiales	
Galdieria sulphuraria	Cyanidiophyceae	Galdieriales	
Ceramium kondoi	Florideophyceae	Ceramiales	
Heterosiphonia pulchra	Florideophyceae	Ceramiales	
Laurencia dendroida	Florideophyceae	Ceramiales	Y
Laurencia pacifica	Florideophyceae	Ceramiales	Y
Laurencia subopposita	Florideophyceae	Ceramiales	Y
Polysiphonia japonica	Florideophyceae	Ceramiales	
Symphyocladia latiuscula	Florideophyceae	Ceramiales	
Chondria armata	Florideophyceae	Ceramiales	
Digenea simplex	Florideophyceae	Ceramiales	
Betaphycus gelatinae	Florideophyceae	Gigartinales	
Chondrus crispus	Florideophyceae	Gigartinales	
Dumontia simplex	Florideophyceae	Gigartinales	
Eucheuma denticulatum	Florideophyceae	Gigartinales	
Gloeopeltis furcata	Florideophyceae	Gigartinales	
Gymnogongrus ftabelliformis	Florideophyceae	Gigartinales	
Kappaphycus alvarezii	Florideophyceae	Gigartinales	
Mazzaella japonica	Florideophyceae	Gigartinales	
Portieria hornemannii	Florideophyceae	Gigartinales	Y
Gracilaria asiatica	Florideophyceae	Gracilares	
Gracilaria blodgettii	Florideophyceae	Gracilares	
Gracilaria chouae	Florideophyceae	Gracilares	
Gracilaria lemaneiformi	Florideophyceae	Gracilares	
Gracilariopsis chorda	Florideophyceae	Gracilares	
Grateloupia licina	Florideophyceae	Halymeniales	
Grateloupia livida	Florideophyceae	Halymeniales	
Grateloupia turuturu	Florideophyceae	Halymeniales	
Prionitis divaricata	Florideophyceae	Halymeniales	
Sinotubimorpha guangdongensis	Florideophyceae	Halymeniales	
Grateloupia filcina	Florideophyceae	Halymeniales	
Palmaria palmata	Florideophyceae	Palmariales	
Rhodophysema elegans	Florideophyceae	Palmariales	
Plocamium pacificum	Florideophyceae	Plocamiales	Y
Erythrolobus australicus	Porphyridiophyceae	Porphyridiales	
Erythrolobus australicus	Porphyridiophyceae	Porphyridiales	Y
Porphyridium purpureum	Porphyridiophyceae	Porphyridiales	Y

Table 3.S2 Cont. Distribution of terpene synthase sequences in Rhodophyta

ORGANISM	CLASS	ORDER	
Timspurckia oligopyrenoides	Porphyridiophyceae	Porphyridiales	
Glaucosphaera vacuolata	Rhodellophyceae	Glaucosphaerales	
Rhodella maculata	Rhodellophyceae	Rhodellales	
Rhodella violacea	Rhodellophyceae	Rhodellales	
Chroodactylon ornatum	Stylonematophyceae	Porphyridiales	
Rhodosorus marinus	Stylonematophyceae	Stylonematales	
		Total	**7**

Table 3.S3 Genome sequencing summary statistics.

Red algal species	Illumina (DNA) yield	PacBio yield	ONT yield	Total assembly length	# of TS genes
Plocamium pacificum	68 Gb	26 Gb	N/a	900 mb	1
Portieria hornemannii	64 Gb	n/a	7.4 Gb	100 Mb	3
Laurencia pacifica	65 Gb	6 Gb, 8 Gb, 33 Gb	n/a	805 Mb	3
Laurencia subopposita	64 Gb	29 Gb, 23 Gb	n/a	444 Mb	6

Table 3.S4 UniProt accession numbers of red algal (RA) IDS sequences supplementing previously published TS list and VHPO sequences for constructing TS and VHPO phylogenies.

IDS Seqs	VHPO Seqs					
Q9TLS1	A0A077JFN1	A7KH32	A0A7S1XAN2	A0A955QNU1	B7ZGM2	A0A0M8V3S5
P51268	A0A084SSH7	A0A940Z5B9	A0A7S1XAQ5	A0A846EKF2	B7ZGM4	A0A285DN85
Q1XDL8	A0A1G8YKF6	A0A1G8VJJ3	A0A2V3J298	A0A0A8J7X1	B7ZGM5	
A0A0X8D962	A0A357I714	A0A1M7Q8B0	A0A7S0T7I2	A0A1X6P568	A0A6H5JGA0	
A0A2I6DH45	A0A963BFY3	A0A8T6H5C3	A0A7S0T7V1	A0A6P0NBQ5	A7KH04	
A0A6G5XNW3	B0C4R0	A0A2V3IDV4	A0A5J4YQJ6	A0A6P0NVN6	K7ZUA3	
A0A7J7IJS0	T2JIS8	A0A2V3IPI8	A0A7S0G5P5	A0A6P0P0Q8	O81959	
	A0A0C1V4J5	R7QFJ4	M2VZ03	A0A6P0P0D2	Q8LLW7	
	A0A4V2E3J3	R7QGE4	M1UVK2	R7Q9K8	O81960	
	A0A6I5NV21	R7QGE8	Q5TZ07	A0A2V3IBA2	A0A1I4BJF3	
	A0A6I5RGX7	R7QHE1	Q6NLA5	A0A2V3ILN7	A0A358JL35	
	A0A6M0F707	A0A2V3IQV0	Q86IX2	A0A2V3ILP4	A0A838SD74	
	A0A6M0F8H9	A0A2V3IUT4	P0A924	A0A2V3ILT2	A0A8T3PKH0	
	A0A6P0YXA1	A0A2V3J3Z6	O59747	L7YCT6	A0A955QAD8	
	A0A0P7YQX6	A0A2V3IUW3	Q8BJ52	A0A6M8Q0S7	M2RSX2	
	A0A2W4ZUK7	R7QBK3	A0A1Z5JQS5	A0A6M8Q894	M2UAK1	
	A0A928WQH8	R7QM91	A0A1Z5KQ12	A0A6M8Q8W3	N4XFG5	
	B4WRG3	R7QMJ0	A0A1Z5K0G0	A0A6M8Q946	A0A7U2FD96	
	A0A1Z3HMR0	R7QTE4	A0A6T6HJX8	R7QIP4	Q0UE79	
	A0A1H9G5V1	A0A7S1TBR6	A0A448ZHQ4	R7QNK3	A0A6A6U3N1	
	A0A1V6LWA6	A0A7S1TEU3	A0A448ZHR5	A0A6M8PSH6	A0A7S0BI13	
	A0A2N0FRM7	A0A7S1TF10	A0A448ZND3	A0A6M8PXA4	A0A7S0BIB9	
	A0A355A8C0	A0A2V3J2U3	A0A448ZPU2	R7Q9E9	A0A7S3E6M0	
	A0A357JIJ9	A0A6H5JTZ2	A0A3G3BM28	R7QUC0	Q9V576	
	A0A239CNJ0	D8LQU5	B7ZGM6	R7QTB2	Q86AF0	
	A0A963AUP5	D7G662	Q4LDE6	A0A7G1K7W0	A0A2V3J6U4	
	K9PWV9	A0A6H5K7C3	O82433	A0A7G1K986	A0A5J4Z065	
	A0A7X1B886	D8LQU9	P81701	A0A7G1K9C3	A0A7J7INR0	
	B5JIY3	D8LQR8	A0A3G3BM34	A0A140JTC4	A7KH27	
	A0A0F7N9T7	D8LSP1	A0A411NJS5	A0A140JTC8	A0A4R4MN06	
	A0A2K9FC11	R7QSC8	A0A411NJY0	A0A140JTC9	A0A7W7IDA2	
	A0A4R9EYF0	A0A1X6P8T2	A0A411NJW8	A0A140JTC5	A0A4R5E3U0	
	A0A941BT52	P49053	Q7X9V0	A0A140JTC6	A0A940YWQ5	
	M4T7F4	A0A364MZW7	Q7X9V1	A0A2K9F6Y0	A0A3S8VXG6	
	A7KH05	R0IP96	B7ZGM1	A0A4V3J8P4	A0A6V8L3P7	
	A7KGZ9	A0A8H5Z974	B7ZGM3	A0A931GJ53	A0A0F7NFF4	

Table 3.S5 Top 30 BlastP hits of VHPO-TS co-localized glycosyltransferase

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. ident	Accession
Starch phosphorylase	Gracilaria domingensis	Gracilaria domingensis	397	397	76%	4.00E-128	KAI0558613.1
Glycogen phosphorylase 1	Gracilariopsis chorda	Gracilariopsis chorda	396	396	76%	7.00E-128	PXF45736.1
Starch phosphorylase	Chondrus crispus	Chondrus crispus	390	390	76%	2.00E-125	XP_005717739.1
hypothetical protein BU14_0266s0010	Porphyra umbilicalis	Porphyra umbilicalis	349	349	76%	1.00E-109	OSX74821.1
hypothetical protein BU14_0266s0019	Porphyra umbilicalis	Porphyra umbilicalis	349	349	76%	1.00E-109	OSX74830.1
Glycogen phosphorylase 1	Porphyridium purpureum	Porphyridium purpureum	333	333	76%	2.00E-103	KAA8495610.1
Non-essential glycogen phosphorylase	Cyanidiococcus yangmingshanensis	Cyanidiococcus yangmingshanensis	318	318	76%	2.00E-97	KAF6000966.1
glycogen phosphorylase	Cyanidioschyzon merolae strain 10D	Cyanidioschyzon merolae strain 10D	317	317	76%	3.00E-97	XP_005535523.1
starch phosphorylase isoform 1	Galdieria sulphuraria	Galdieria sulphuraria	303	303	76%	6.00E-92	XP_005703547.1
starch phosphorylase isoform 2	Galdieria sulphuraria	Galdieria sulphuraria	303	303	76%	8.00E-92	XP_005703546.1
hypothetical protein GpartN1_g7297.t1	Galdieria partita	Galdieria partita	300	300	76%	6.00E-91	GJQ15506.1
hypothetical protein AXG93_1335s1250	Marchantia polymorpha subsp. ruderalis	Marchantia polymorpha subsp. ruderalis	290	290	76%	3.00E-87	OAE28633.1
Alpha-glucan phosphorylase 2	Nymphaea thermarum	Nymphaea thermarum	278	278	76%	6.00E-87	KAF3775924.1
hypothetical protein Mapa_007697	Marchantia paleacea	Marchantia paleacea	289	289	76%	8.00E-87	KAG6550793.1
Phosphorylase isoform 1	Carex littledalei	Carex littledalei	279	279	76%	1.00E-86	KAF3334102.1
Glycogen phosphorylase 1	Galdieria sulphuraria	Galdieria sulphuraria	287	287	76%	4.00E-86	GJD11123.1
Phosphorylase isoform 2	Theobroma cacao	Theobroma cacao	279	279	76%	8.00E-86	EOY33812.1
hypothetical protein KP509_31G037900	Ceratopteris richardii	Ceratopteris richardii	275	275	76%	3.00E-85	KAH7288711.1
hypothetical protein KP509_31G037900	Ceratopteris richardii	Ceratopteris richardii	275	275	76%	8.00E-85	KAH7288709.1
Alpha-glucan phosphorylase 2, cytosolic	Ananas comosus	Ananas comosus	274	274	76%	9.00E-85	OAY77179.1
alpha-glucan phosphorylase, H isozyme isoform	Nymphaea colorata	Nymphaea colorata	280	280	76%	9.00E-85	XP_031488934.1
Putative Phosphorylase	Rhizopus microsporus	Rhizopus microsporus	273	273	76%	1.00E-84	CEI94481.1
glycogen/starch/alpha-glucan phosphorylase	Tetrasphaera jenkinsii	Tetrasphaera jenkinsii	270	270	76%	1.00E-84	WP_162200002.1
alpha-glucan phosphorylase, H isozyme isoform	Nymphaea colorata	Nymphaea colorata	281	281	76%	3.00E-84	XP_031488932.1
alpha-glucan phosphorylase, H isozyme isoform	Nymphaea colorata	Nymphaea colorata	280	280	76%	3.00E-84	XP_031488933.1
alpha-glucan phosphorylase 2, cytosolic isoform	Eutrema salsugineum	Eutrema salsugineum	278	278	76%	3.00E-84	XP_024013464.1
hypothetical protein CY35_10G106300	Sphagnum magellanicum	Sphagnum magellanicum	279	279	76%	4.00E-84	KAH9551112.1
Glycogen/starch/alpha-glucan phosphorylase	Arabidopsis thaliana x Arabidopsis arenosa	Arabidopsis thaliana x Arabidopsis arenosa	281	281	76%	8.00E-84	KAG7559934.1
unnamed protein product	Arabidopsis arenosa	Arabidopsis arenosa	281	281	76%	9.00E-84	CAE6075344.1

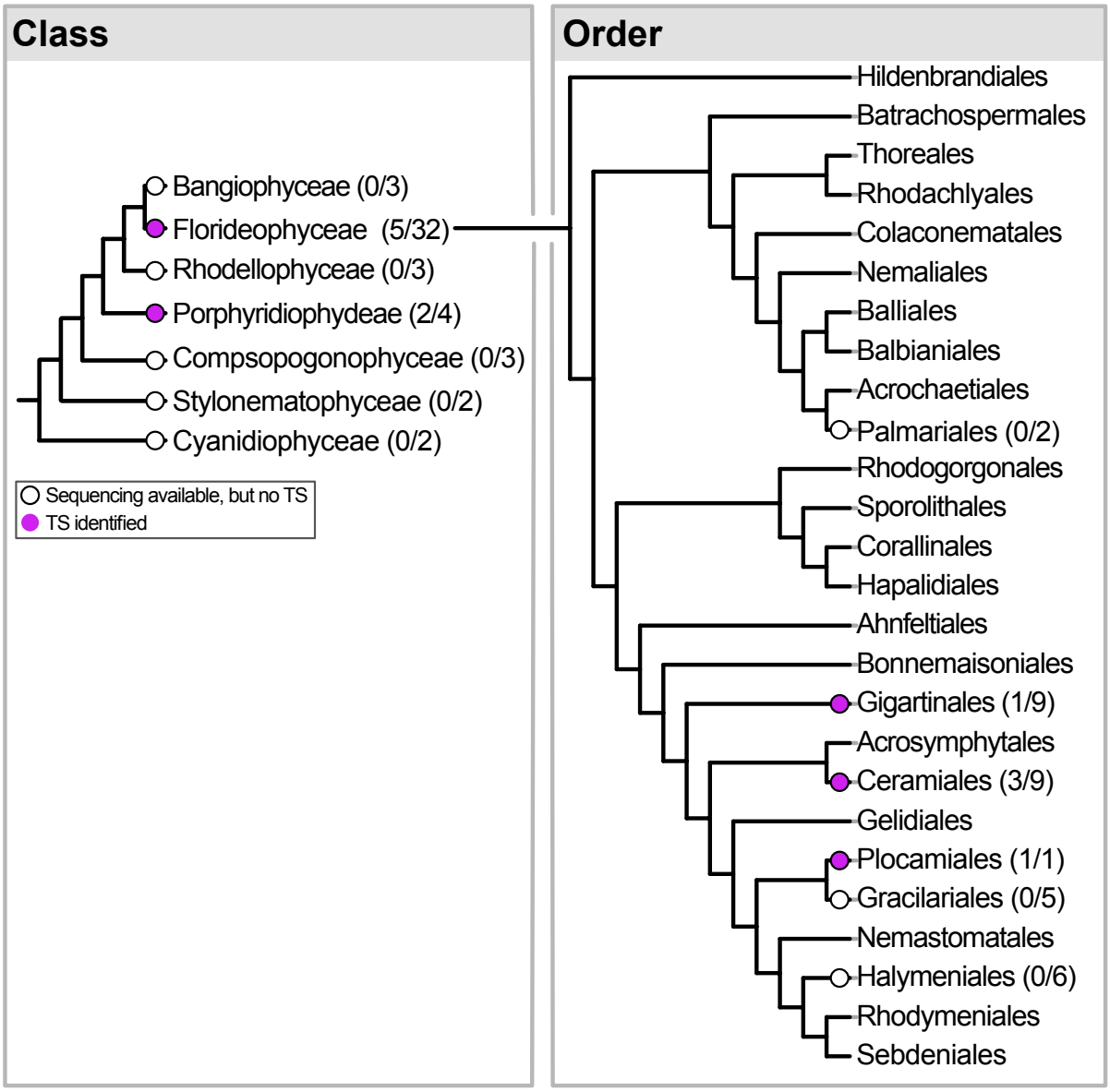


Figure 3.S1 The distribution of TS sequences by class and order. Purple circle indicates the presence of a TS and a white circle indicates sequencing data is available, but no TS sequence has been identified. The classes Porphyridiophyceae and Florideophyceae harbor MTSs for 7 out of 49 investigated red algal species to-date.

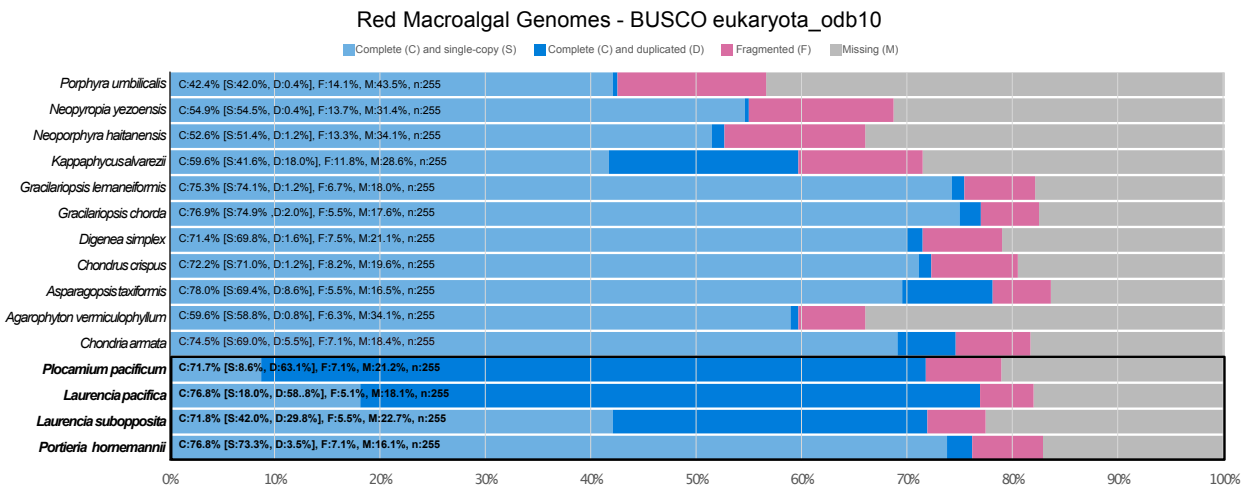


Figure 3.S2 BUSCO assessment of red macroalgal genomes. *P. pacificum, L. subopposita L. pacifica, and P. hornemannii* BUSCO assessment highlighted in black box.

Figure 3.S3 Sequence alignment of ten new red algal TS sequences, key active site residues are highlighted in yellow and summarized in blue in sequence logo (bottom).

```
lsTS-2       SGDPNDVLSSPYIQ-MMTRLAVYHISYSNDIIGFHKERKETSLNNLIKVMAKNNQQSFED   272
lsTS-6       SGDSSDVLSSPYIQ-MMTRLAVHHVSYCNDIIGFHKECKETSLNNLVKVMAKDNQQSFGD   272
lsTS-1       KGSPEDLPFMSYME-IMKHLAIKHIAFVNDIVSFNKERYEAVNNNIIIVLANDRDHTCKD   280
lsTS-3       AGSSYEVMLPPYTS-ILLRMATNHIHWVNDIISLNKEAKEAISGNMVEIISNRDQYTVTR   275
lsTS-4       TGDPADVLSTPHIL-IMTRLVVNHVSWFNDIIGVNKEKKEAVNNNIVFVMESKKGQTWEG   275
lsTS-5       TGCPSSVLSMPFLQ-IMTNLAVNHVAWVNDIVGANKERKEAVNNNIVFVMANDRGLTMSG   276
pcTS-C2      SRGKC-EYASCFLLQQMVSYADHQVIWINDIIGVNTDCQDGVEGNIVVALKNSRGLTWDE   284
phTS-1       Q-APQRVKKSCYVD-ELFRSGSSVVAWVNDVYTVWKDIRDKTSDNVVLVLVNSRELSWKE   355
phTS-2       KDSPENITNNAYLI-ELLRMATIHVQSVNDFFSMAKDIRDCTGDNIVVVLAKERELSWPE   271
phTS-3       KERALDVHGSCFLS-ELMRAGTLHVAWLNDVFSVCKDLRDGTGDNVVLVLSRENGLSWEA   274
                          .    :       :   **.   .:  :   .*::  :      :

lsTS-2       ALEGALKSTNQLVDAFLNLEKMVHIHGLSLHVGRDRLNDDILKYIEVLKYWMRGSLDWHF   332
lsTS-6       ALKGALKTNNQLVDAFLNVEEMVSIHGLTLLKDRKR------------------------   308
lsTS-1       AIGDAIKLSNQNLETFLKLESALITKLDP------TIHGDSLAFIDVLKNCMRGNIDWHF   334
lsTS-3       AVEDALKRVNEQCEAFLDLESKLHSSGLL------EGNEDTIIFIAVLKYWMRGSLDWHF   329
lsTS-4       AVQDSVKRVNEECETFLELEAELHESGLL------EDNEDALNYIEVLKYWIRGSMDWHF   329
lsTS-5       AVKEAVKRTNQECEVFLNFEHKLHAGGVV------VDADDLLNYIEVLKYWMRGSLDWHF   330
pcTS-C2      AVDCAINMVNDEIDGFIALEKEVEY-----LRGQGELSNGTAEFINCLKSYMRGSMEWSA   339
phTS-1       ALNSSIQLVDEEMENLLALERNVAG-----MIG--GEDPGIVAIIKAVNDCVRGNYDWSI   408
phTS-2       AIECSLDLINKEMKAFLSLESKLED-----MMG--PMDRDVVKVVESMKHLMRGNRDWCG   324
phTS-3       AVEHVMGLVDDETENILSLEENLED-----LMG--SLDYETKTFVEGLNYIVRGSYDWSM   327
              *:   :   :.  . :: .*  :

lsTS-2       ESARYKDYTPISS-------------------------------- 345
lsTS-6       -----------------------------------------------  308
lsTS-1       ESVRYRTSQ------------------------------------- 343
lsTS-3       ESKRYAVDN------------------------------------- 338
lsTS-4       ESGRYRVNTEK----------------------------------- 340
lsTS-5       ESKRYKVKACK----------------------------------- 341
pcTS-C2      ETDRYKEWRMMAEAEV------------------------------ 355
phTS-1       ETPRYRDVKIQNGTGKDK--LVV----EQRDAKLSSK*------- 439
phTS-2       SSGRYRDINSYVESSA*----------------------------- 340
phTS-3       ESTRYAGINGLEEEQPVSGRTISKKQAGERSTYMSDSGLASCEE* 371
```

Figure 3.S3 Cont. Sequence alignment of ten new red algal TS sequences, key active site residues are highlighted in yellow and summarized in blue in sequence logo (bottom).

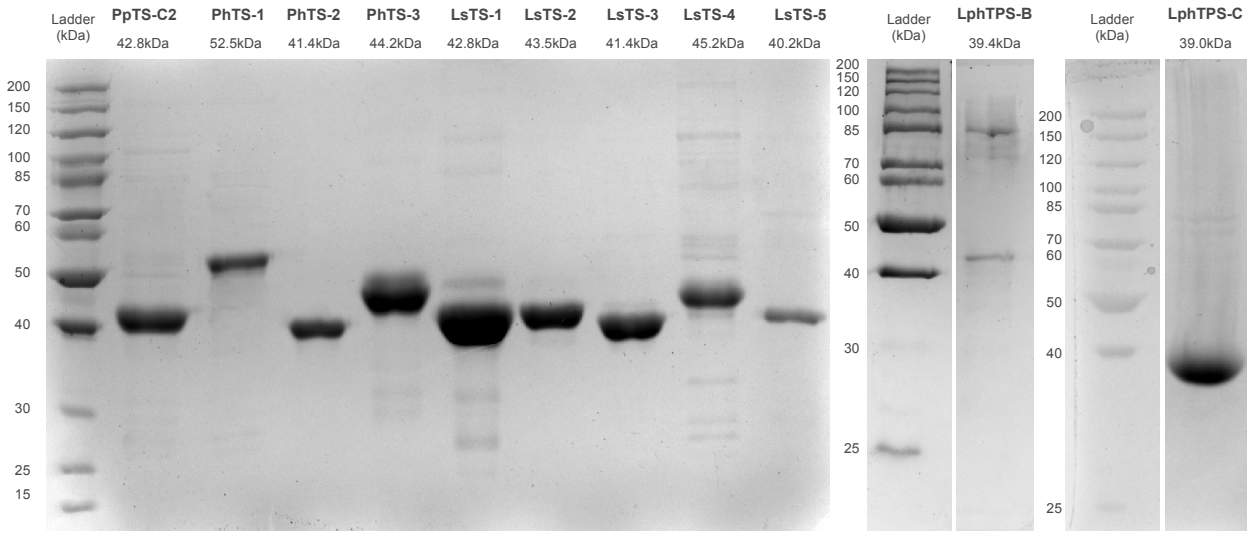


Figure 3.S4 10% SDS-PAGE gel. Sample order is as follows: Ladder, PpTS-CA, PhTS-1, PhTS-2, PhTS-3, LsTS-1, LsTS-2, LsTS-3, LsTS-4, LsTS-5, LsTS-6, Ladder, LpTS-B, Ladder, LpTS-C

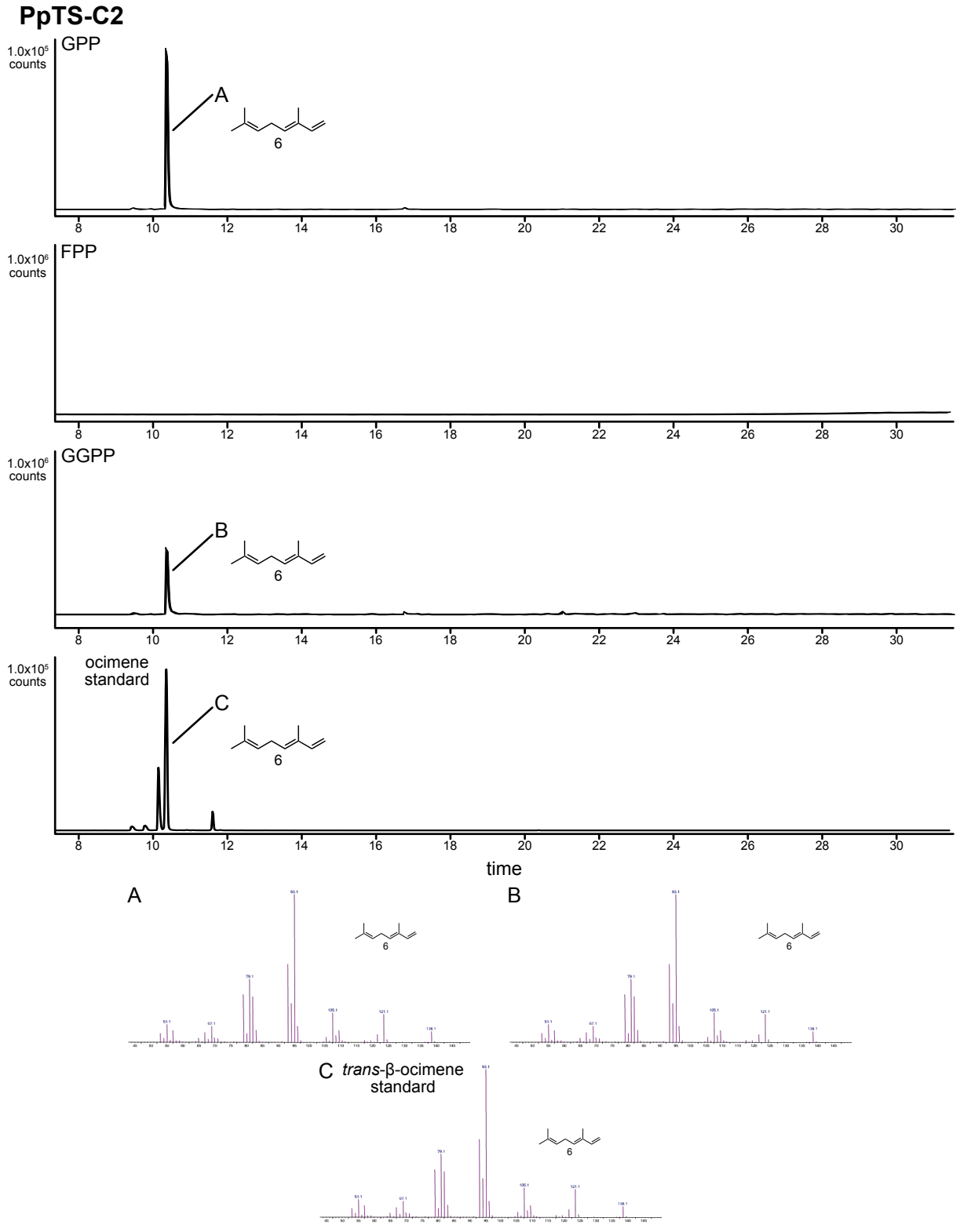


Figure 3.S5 GCMS analysis of PpTS-C2 with the substrates GPP (top) FPP (middle top) GGPP (middle bottom) and standard of ocimene isomers (bottom). Mass spectra for major peaks are shown.

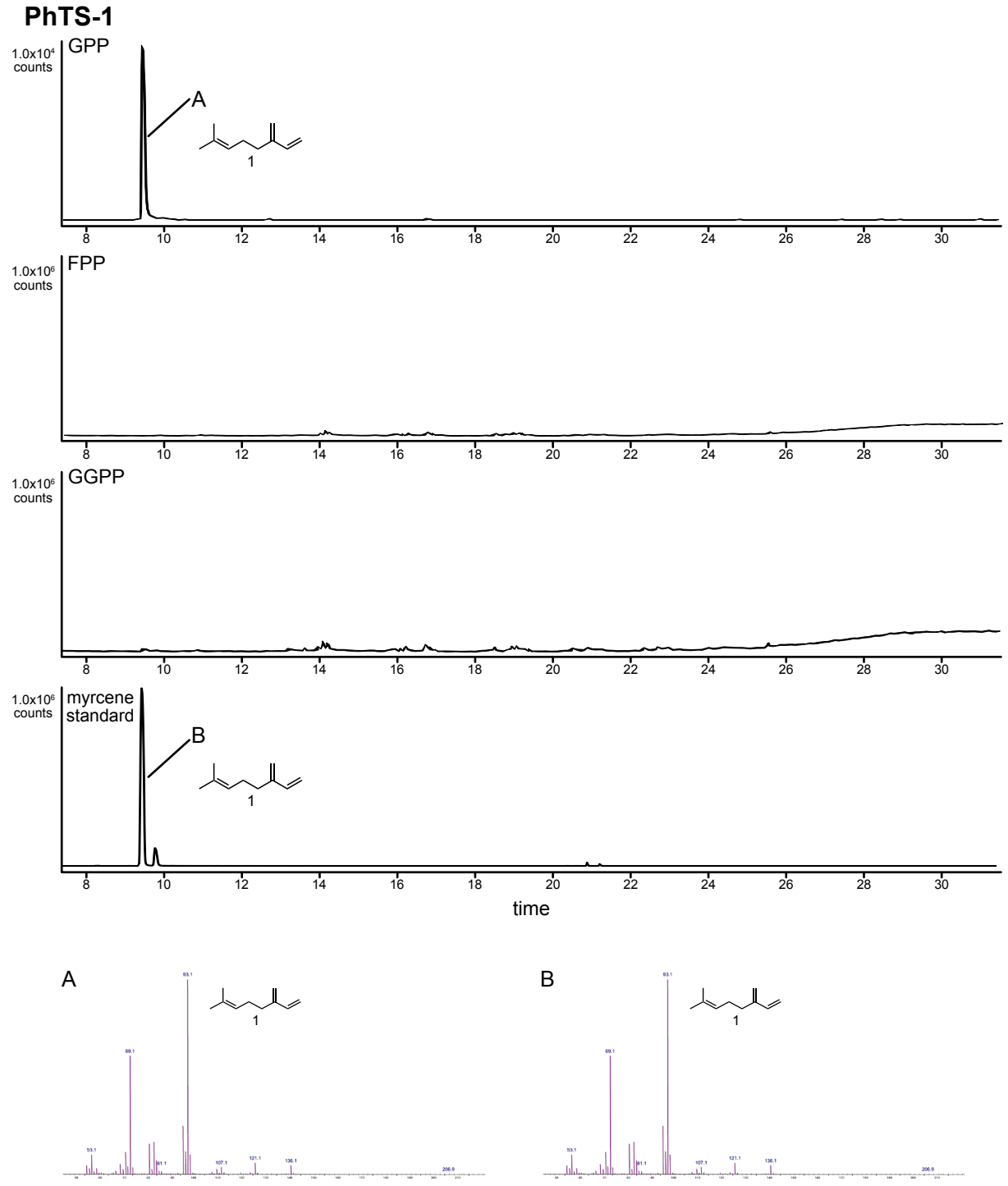


Figure 3.S6 GCMS analysis of PhTS-1 with the substrates GPP (top) FPP (middle top) GGPP (middle bottom) and standard of myrcene isomers (bottom). Mass spectra for major peaks are shown.

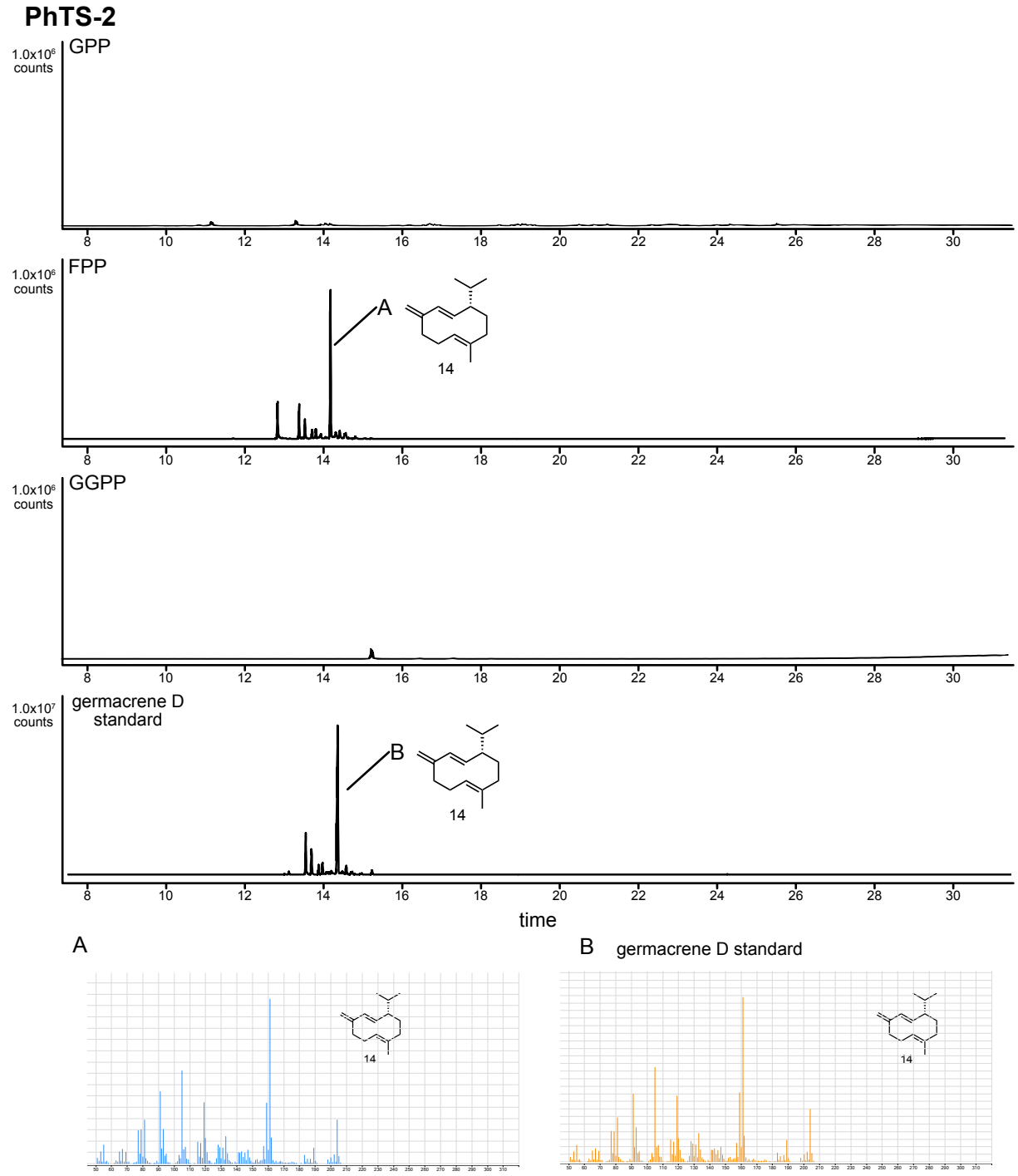


Figure 3.S7 GCMS analysis of PhTS-2 with the substrates GPP (top) FPP (middle top) and GGPP (middle bottom) and germacrene D (**14**) standard (bottom). Mass spectra for major peaks are shown.

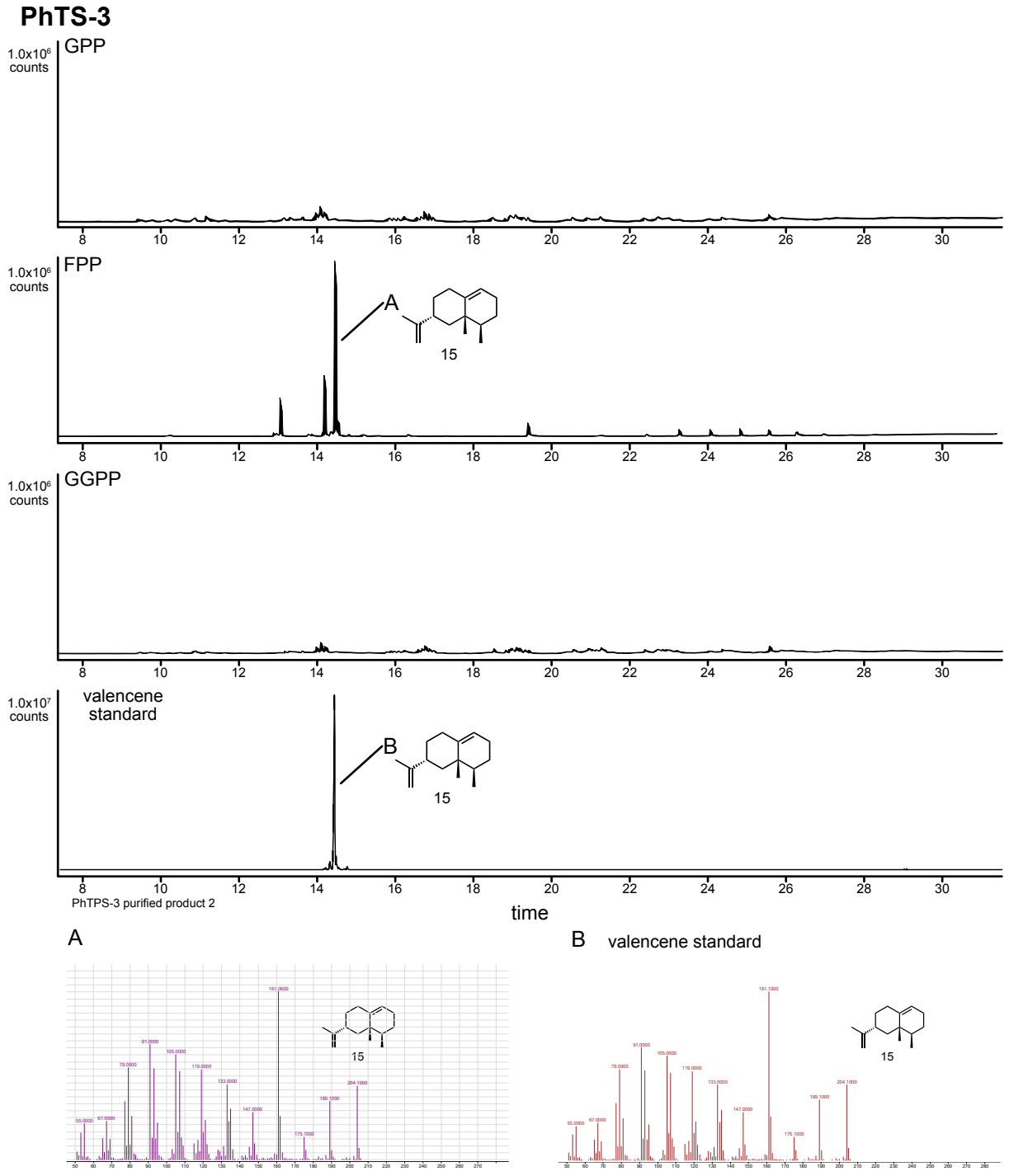


Figure 3.S8 GCMS analysis of PhTS-3 with the substrates GPP (top) FPP (middle top) and GGPP (middle bottom) and valencene (**15**) (bottom). Mass spectra for major peaks are shown.

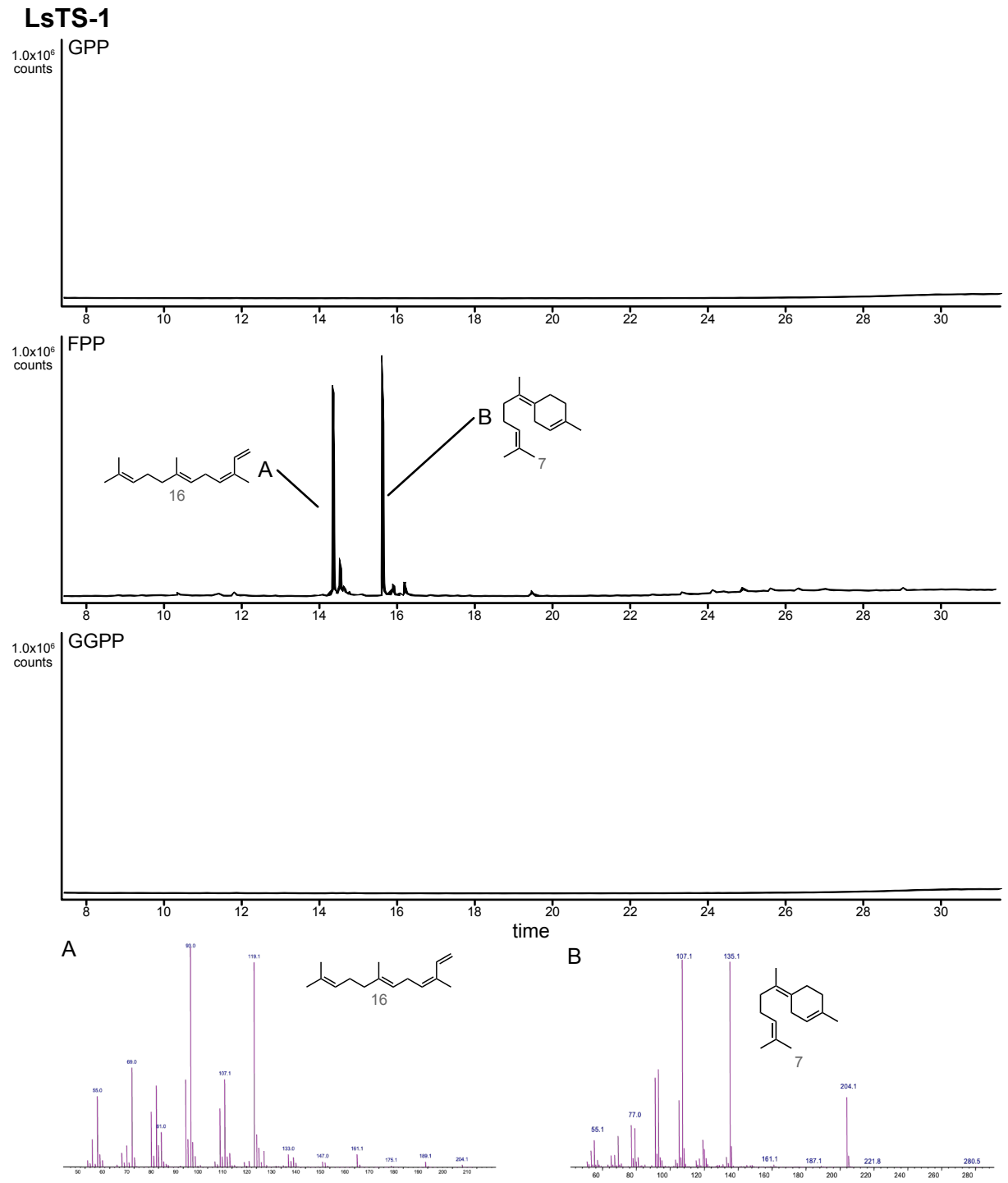


Figure 3.S9 GCMS analysis of LsTS-1 with the substrates GPP (top) FPP (middle) and GGPP (bottom). Mass spectra for major peaks are shown.

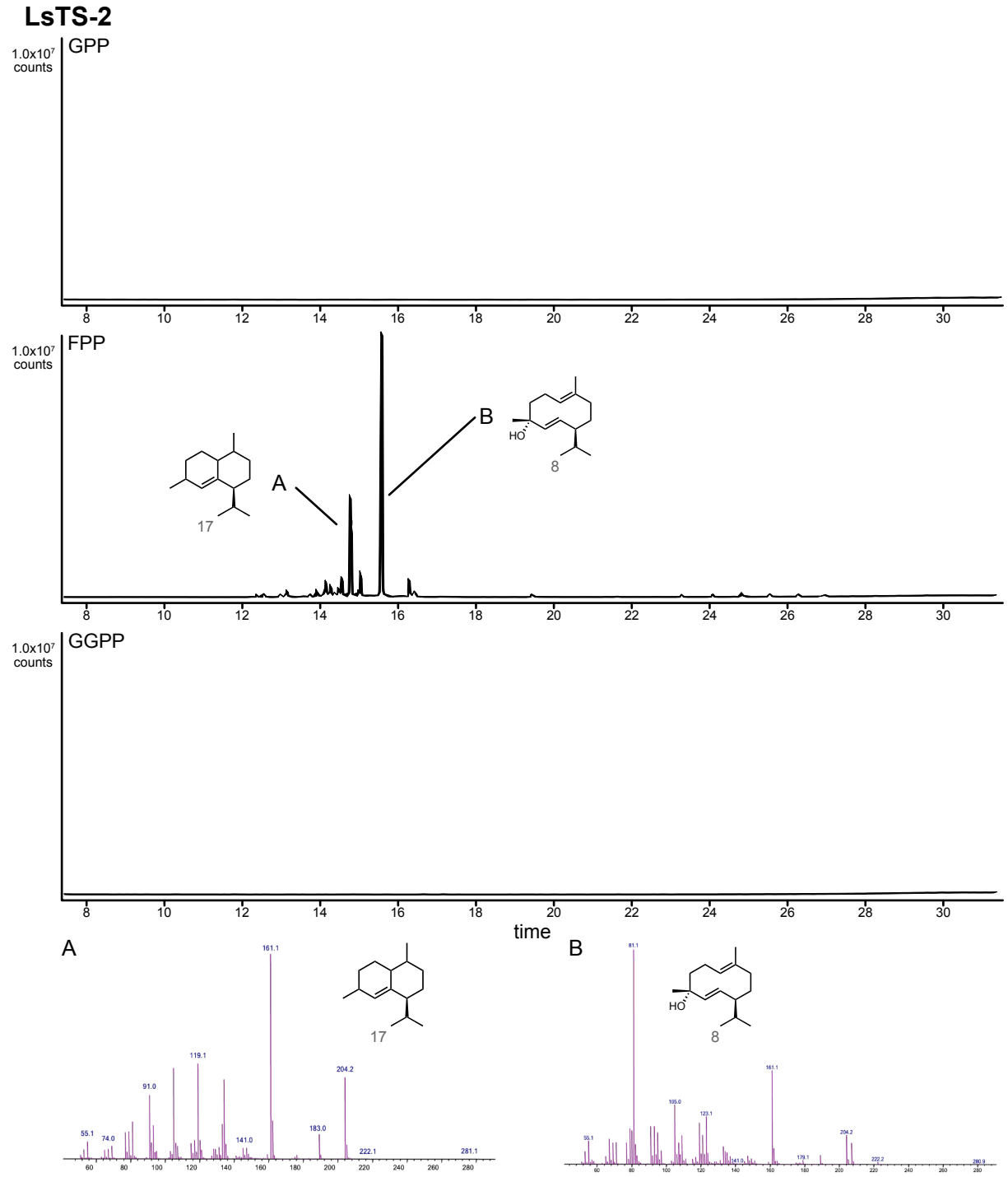


Figure 3.S10 GCMS analysis of LsTS-2 with the substrates GPP (top) FPP (middle) and GGPP (bottom). Mass spectra for major peaks are shown.

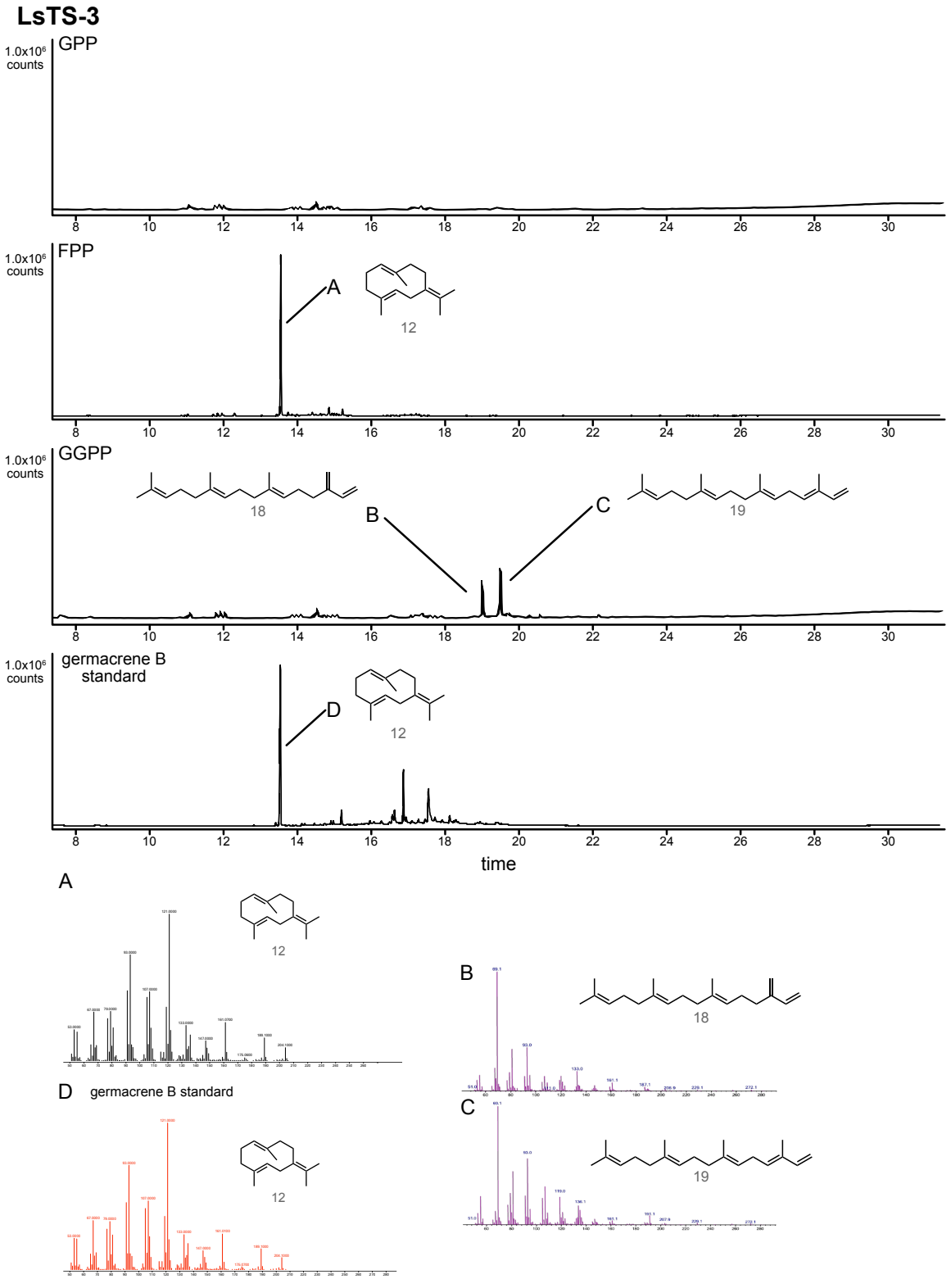


Figure 3.S11 GCMS analysis of LsTS-3 with the substrates GPP (top) FPP (middle top) and GGPP (middle bottom) and germacrene B (**12**) (bottom). Mass spectra for major peaks are shown.

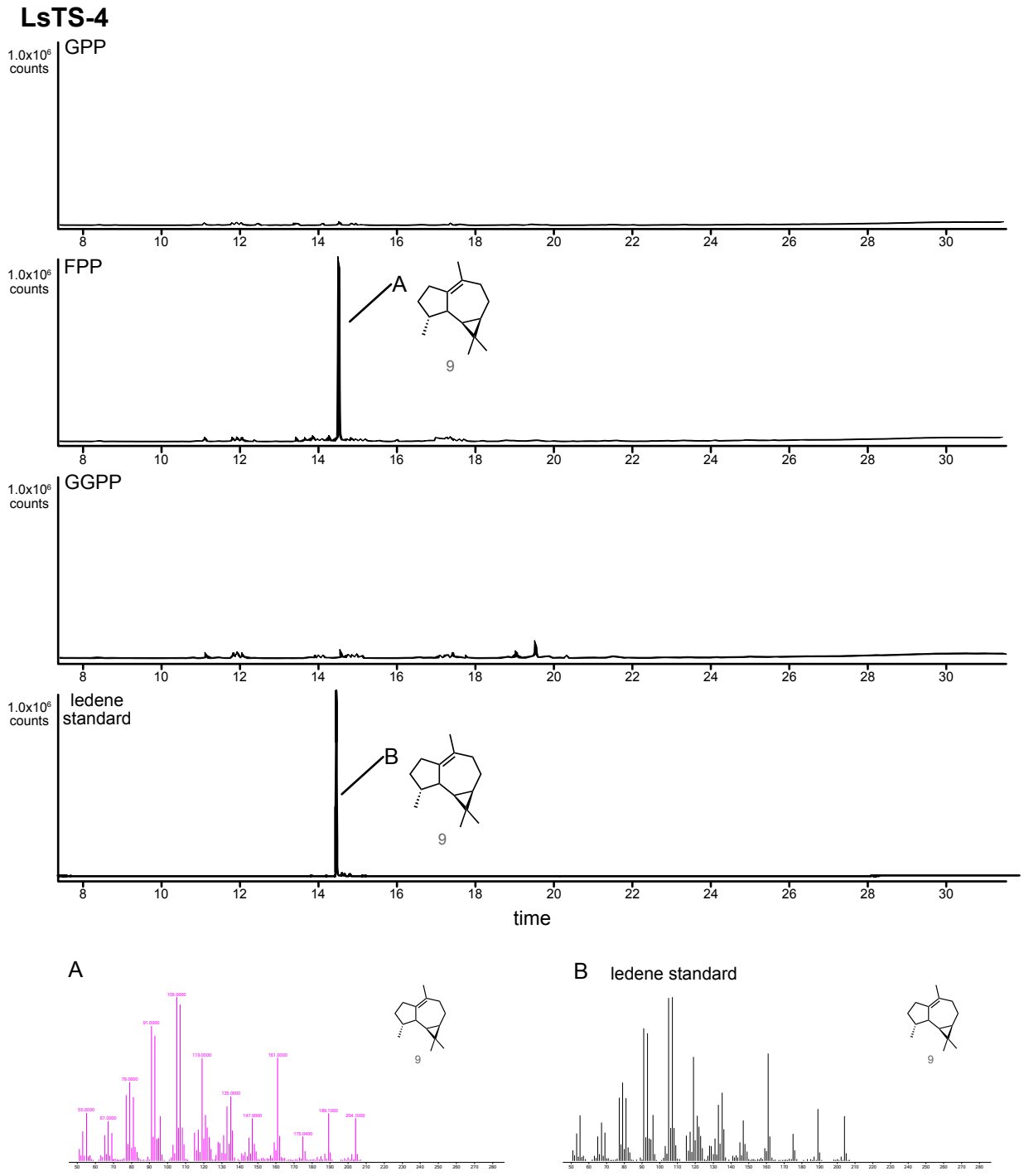


Figure 3.S12 GCMS analysis of LsTS-4 with the substrates GPP (top) FPP (middle top) GGPP (middle bottom) and ledene (**9**) standard (bottom). Mass spectra for major peaks are shown.

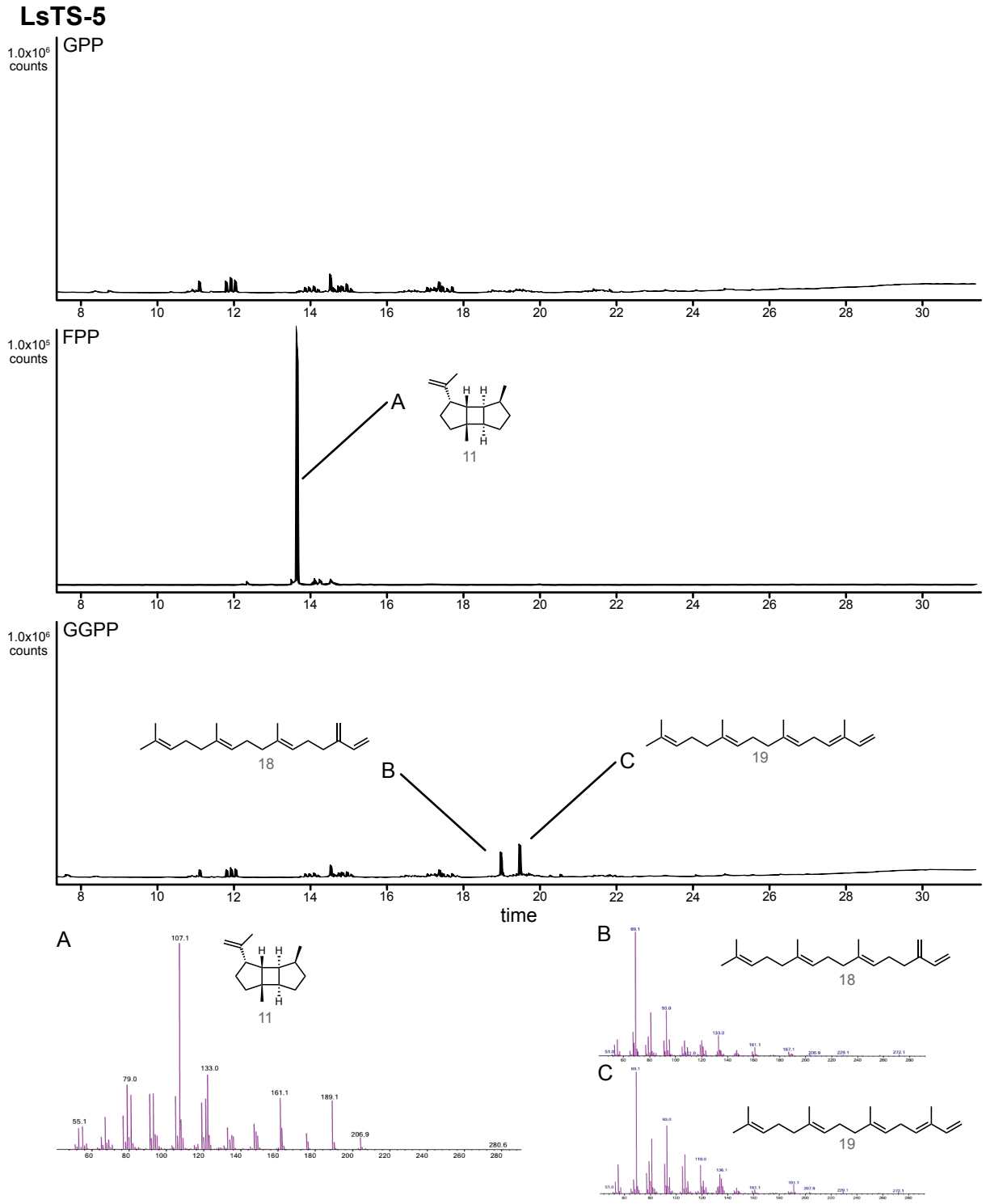


Figure 3.S13 GCMS analysis of LsTS-5 with the substrates GPP (top) FPP (middle) and GGPP (bottom). Mass spectra for major peaks are shown.

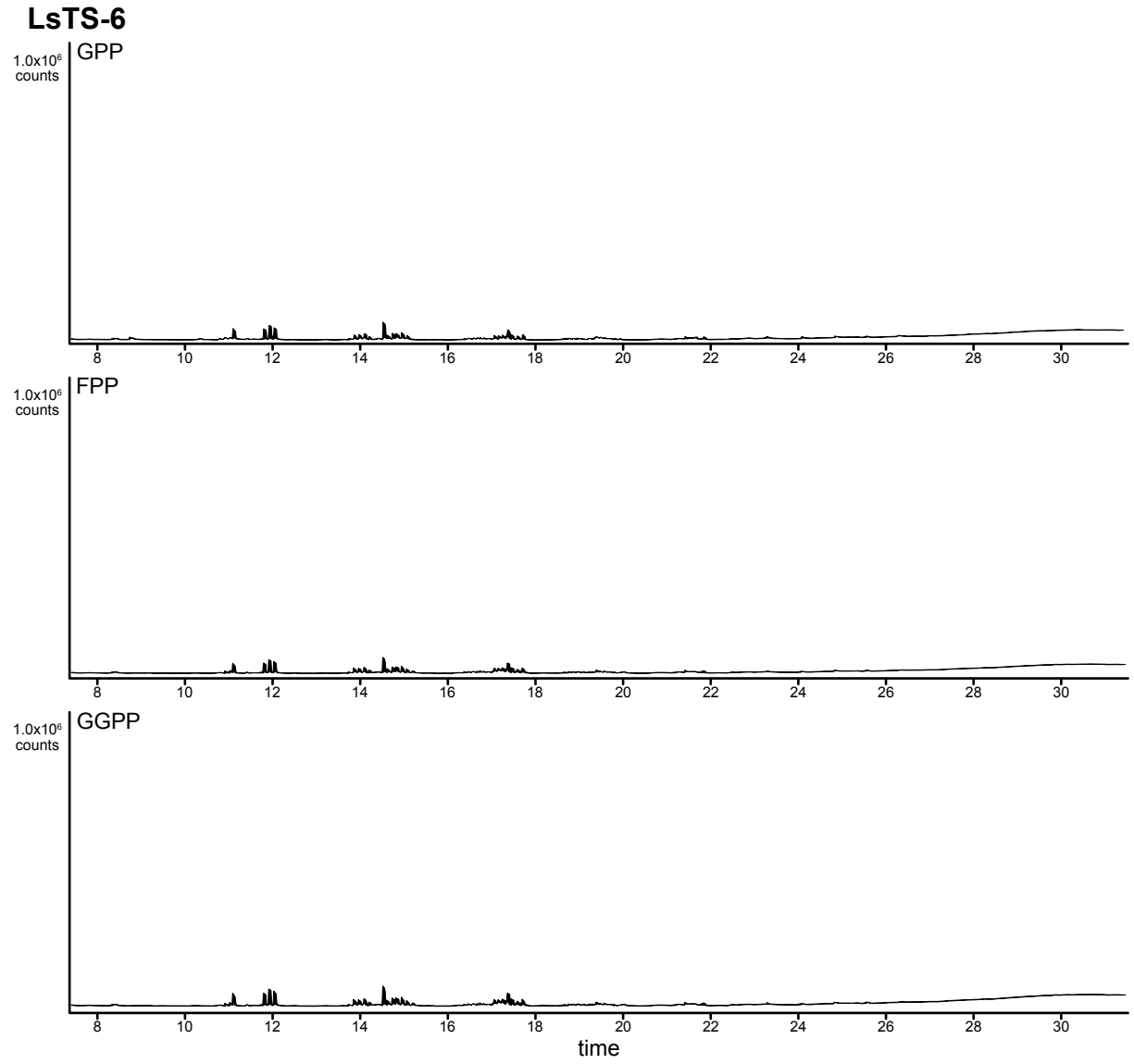


Figure 3.S14 GCMS analysis of LsTS-6 with the substrates GPP (top) FPP (middle) and GGPP (bottom).

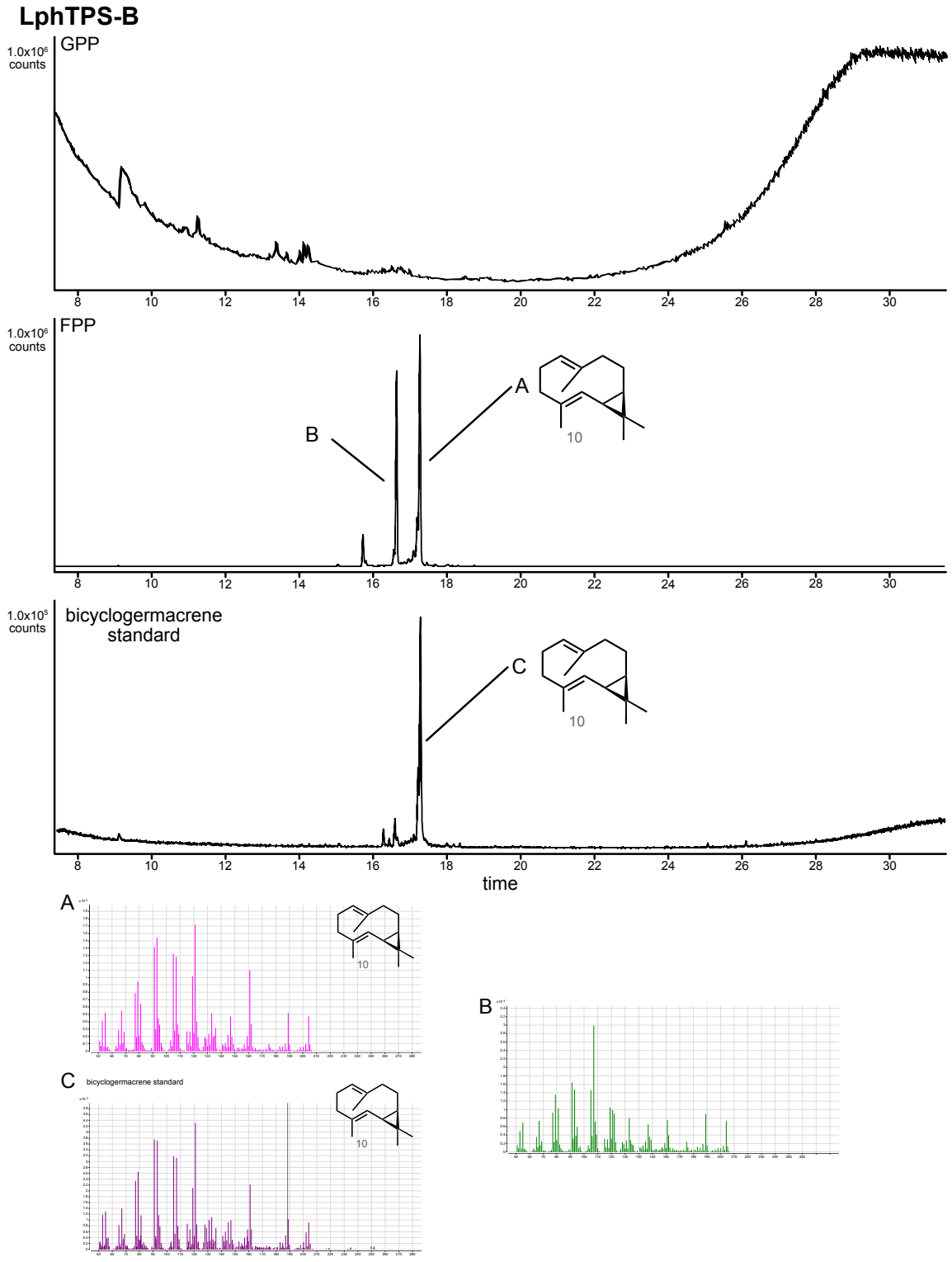


Figure 3.S15 GCMS analysis of LphTPS-B with the substrates GPP (top) FPP (middle) and bicylogermacrene standard (**10**) (bottom). Mass spectra for major peaks are shown.

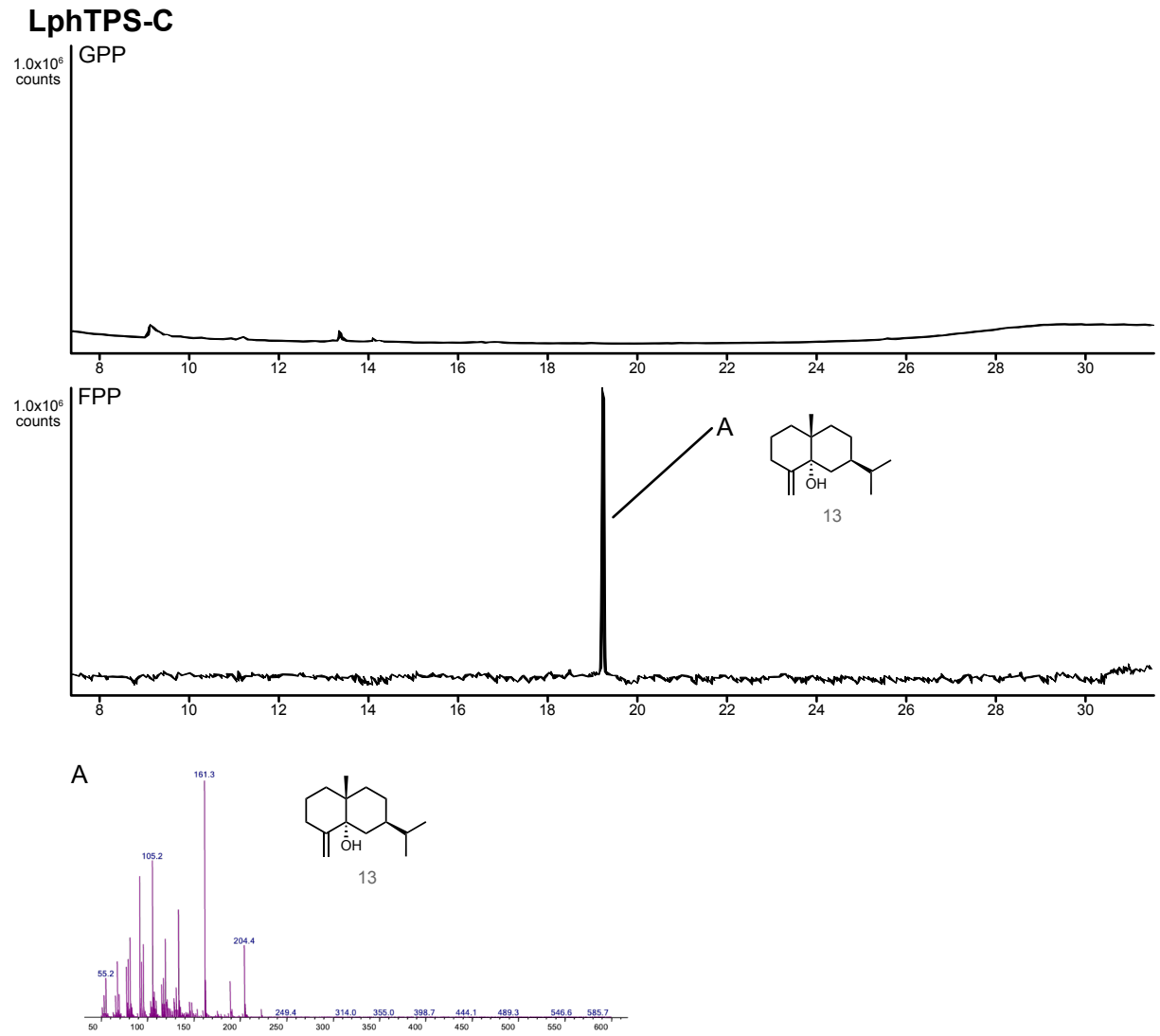


Figure 3.S16 GCMS analysis of LphTPS-C with the substrates GPP (top) and FPP (bottom). Mass spectra for major peaks are shown.

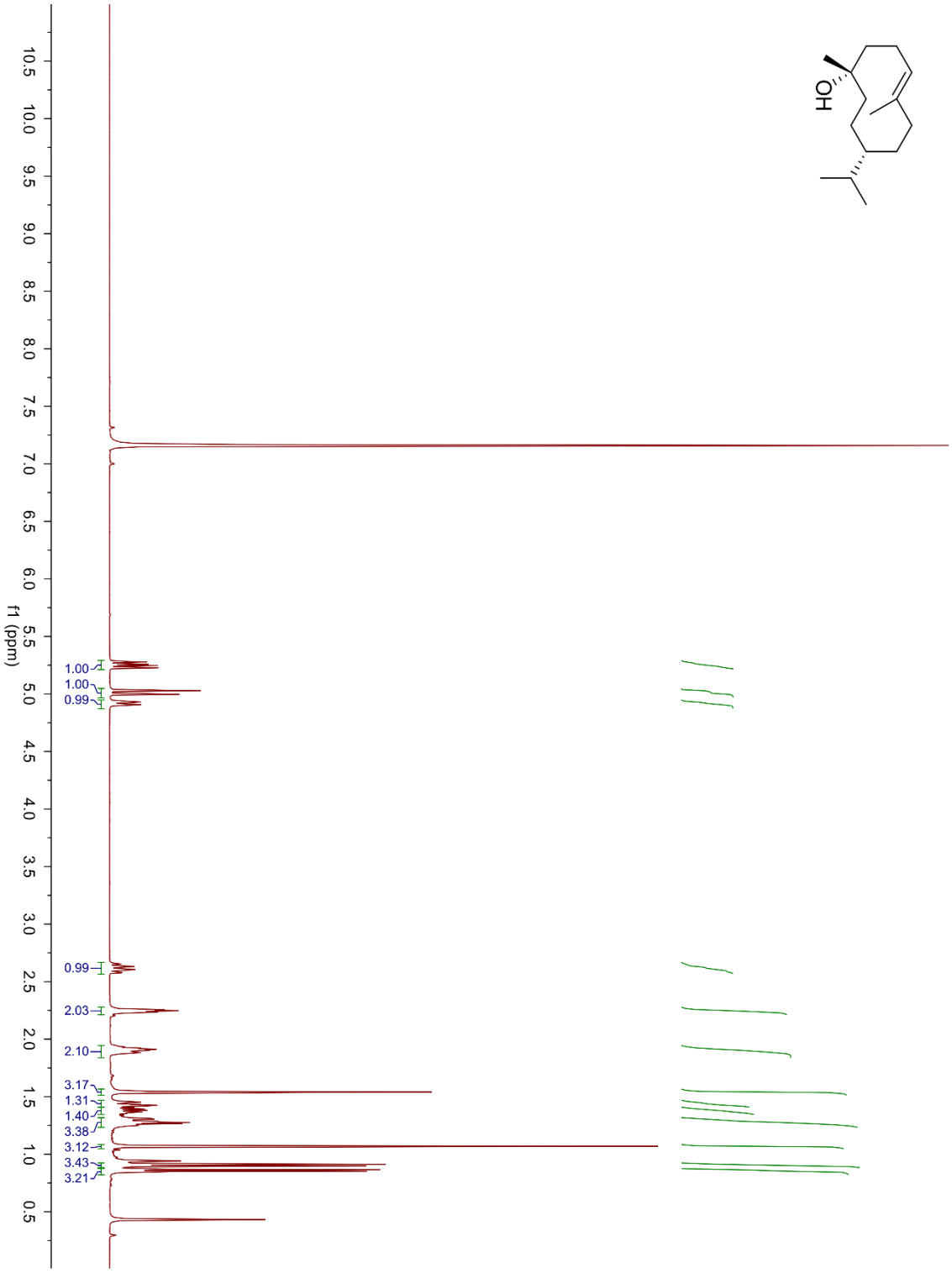


Figure 3.S17 ^{1}H-NMR spectrum of **8** in C_6D_6.

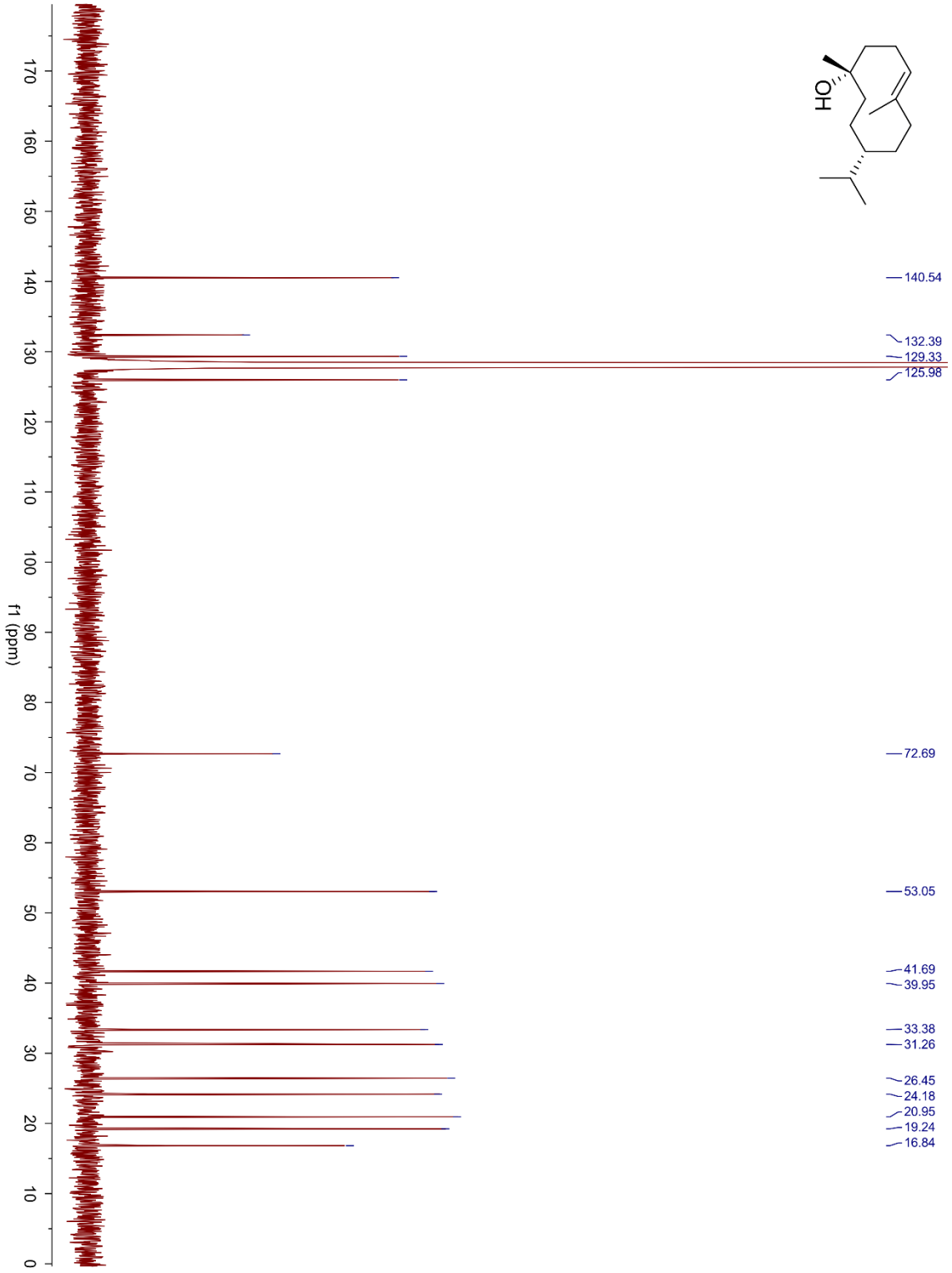


Figure 3.S18 ^{13}C-NMR spectrum of **8** in C_6D_6.

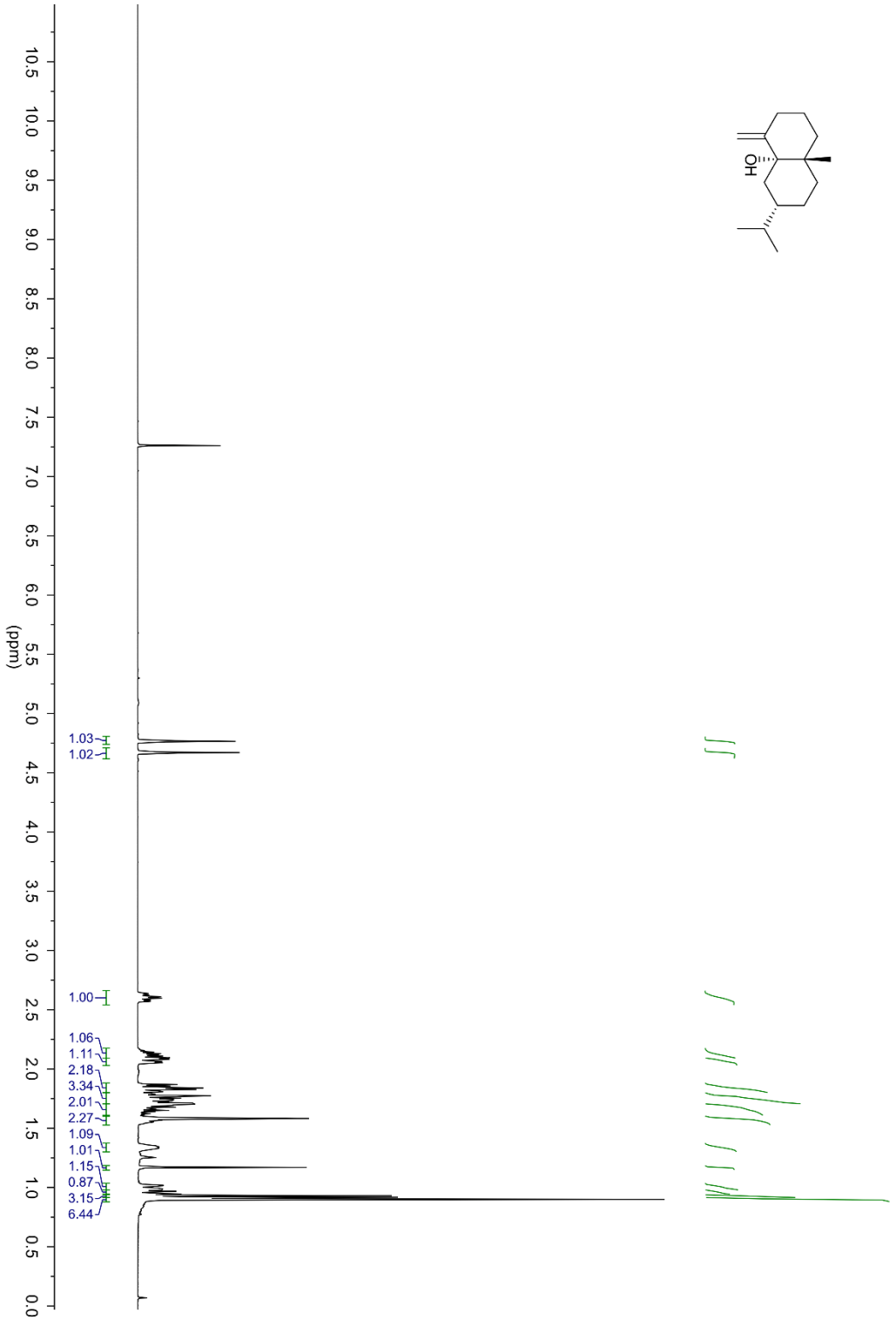


Figure 3.S19 ^{1}H-NMR spectrum of **13** in $CDCl_3$.

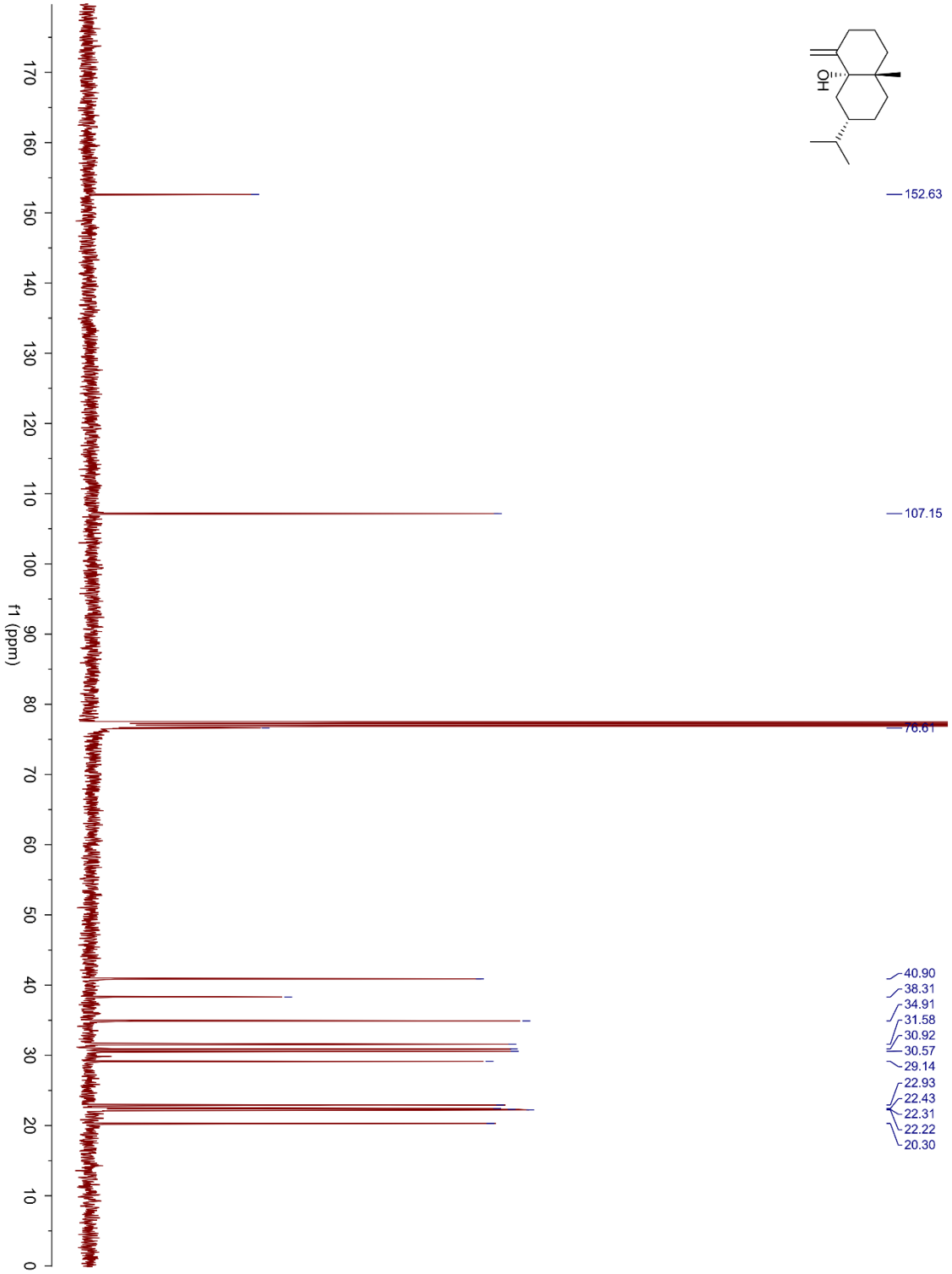


Figure 3.S20 ^{13}C-NMR spectrum of **13** in $CDCl_3$.

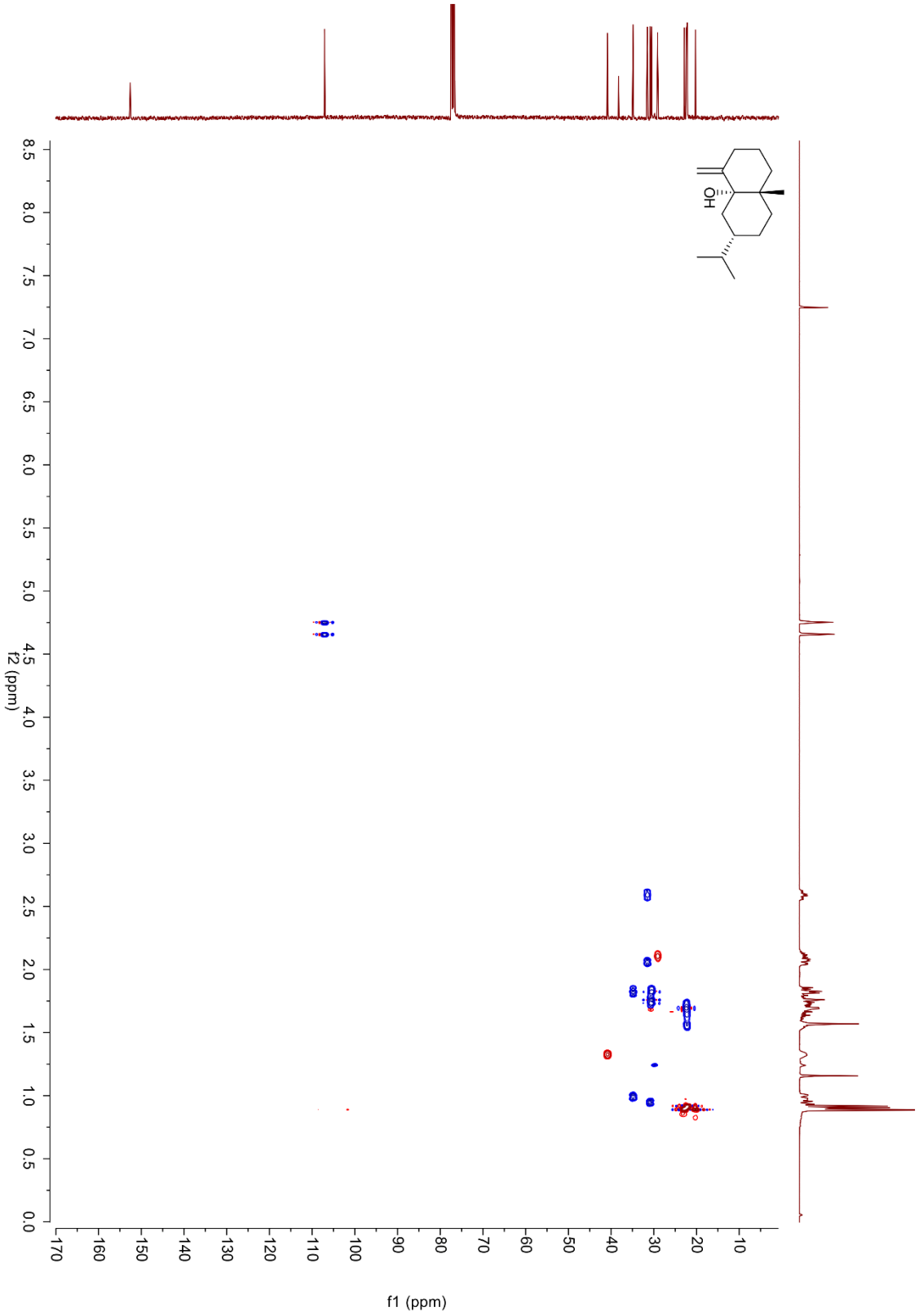


Figure 3.S21 ^{1}H- ^{13}C-HSQC spectrum of **13** in $CDCl_3$.

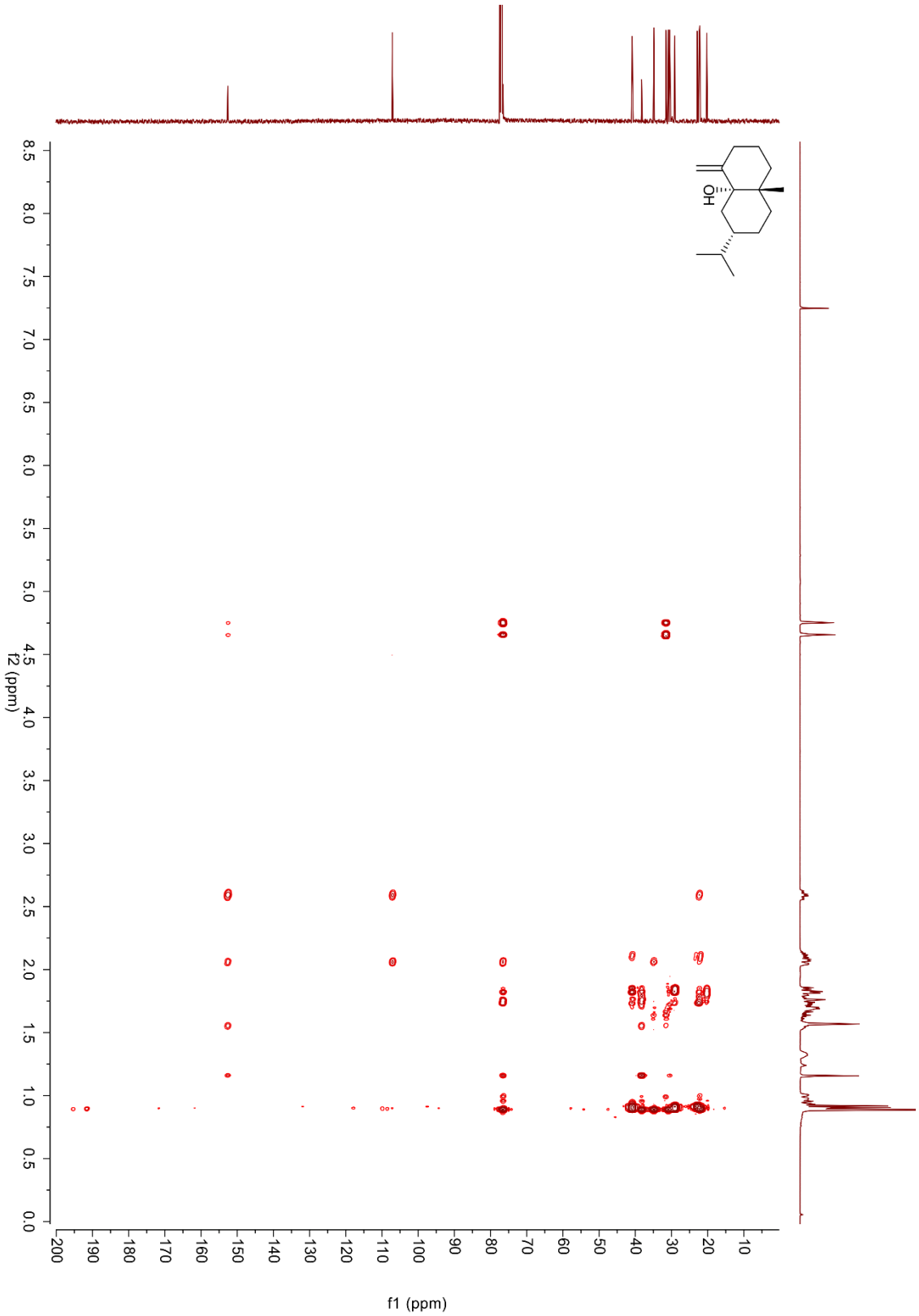


Figure 3.S22 ^{1}H- ^{13}C-HMBC spectrum of **13** in $CDCl_3$.

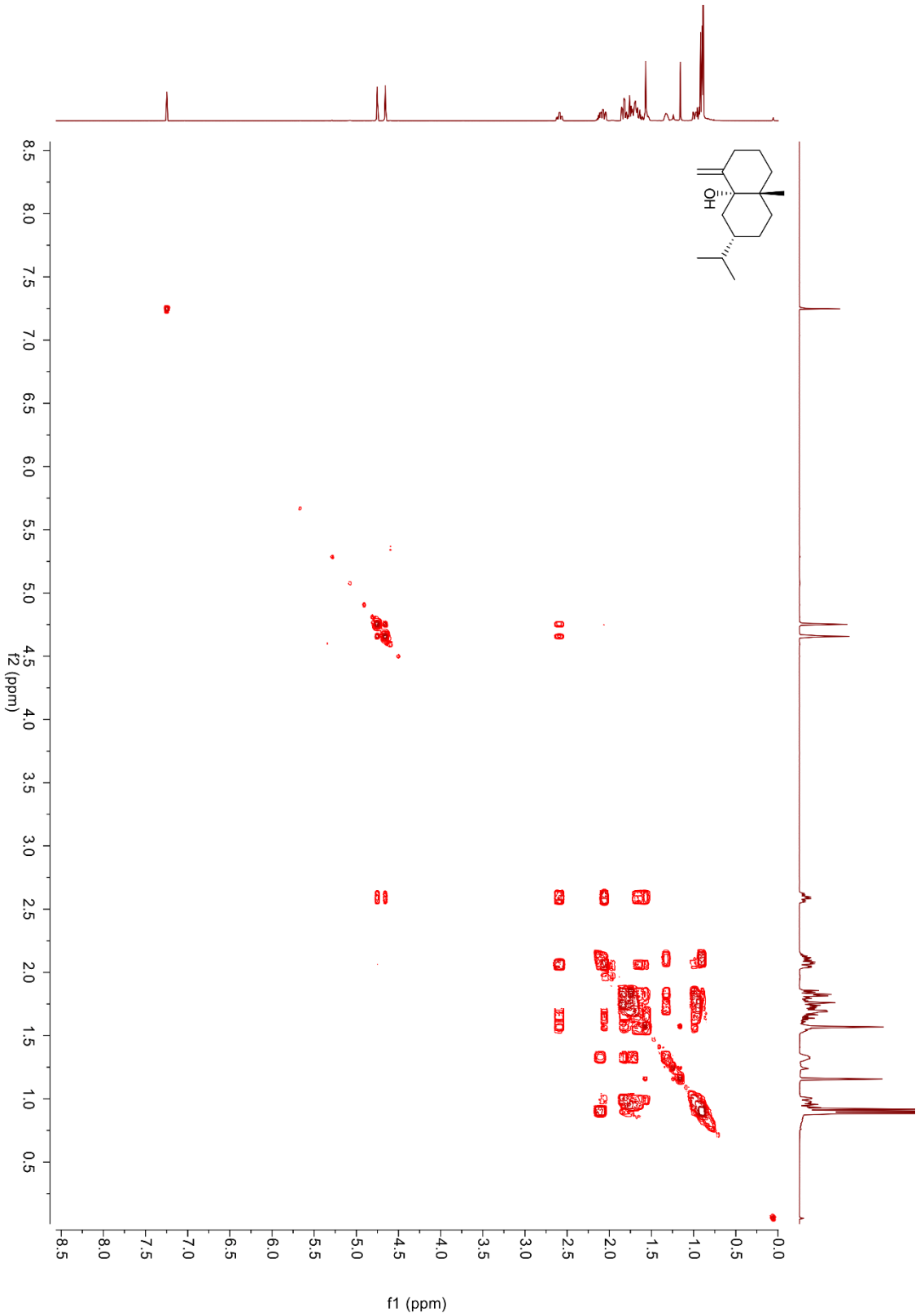


Figure 3.S23 ^{1}H- ^{1}H--COSY spectrum of **13** in $CDCl_3$.

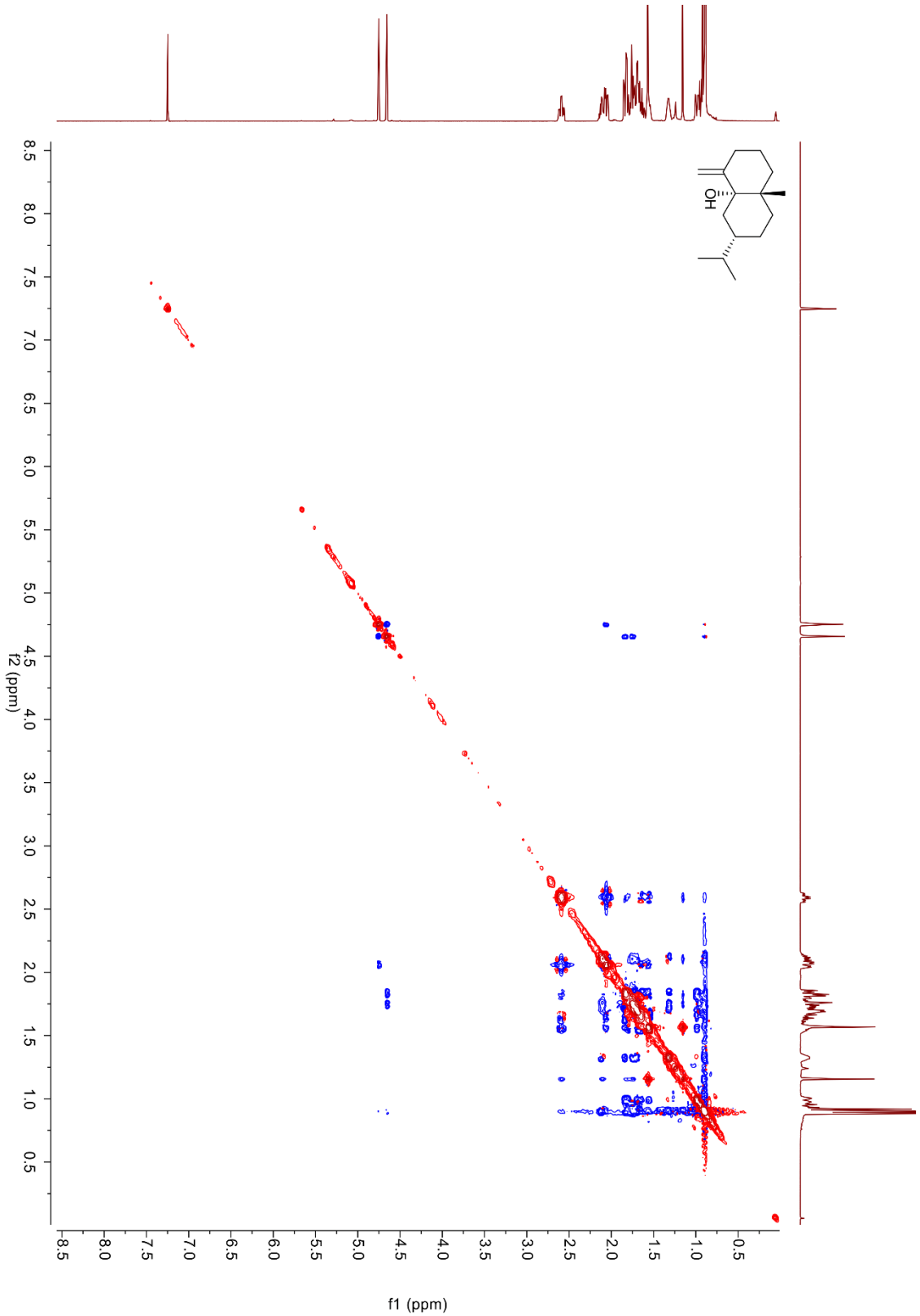


Figure 3.S24 ^{1}H- ^{1}H-NOESY spectrum of **13** in $CDCl_3$

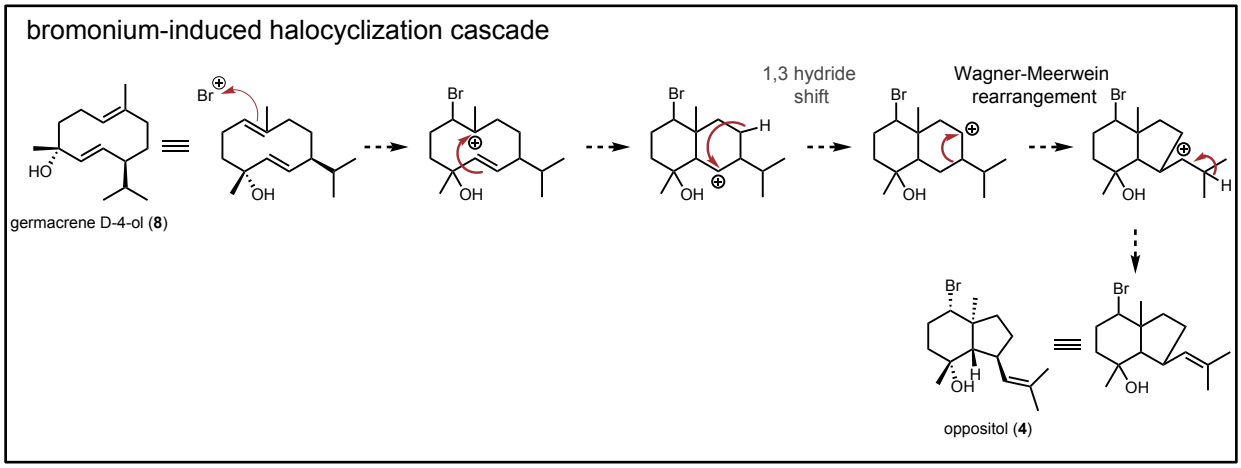


Figure 3.S25 Proposed bromonium-induced halocyclization cascade by LsVHPO-1 in oppositol (**4**) biosynthesis. LsVHPO-1 is co-clustered with the germacrene D-4-ol (**8**) producing TS, LsTS-2**.**

3.9 Supplementary Information References

1. V. J. Davisson, A. B. Woodside, C. D. Poulter, Synthesis of allylic and homoallylic isoprenoid pyrophosphates, *Meth. Enzymol.* 130–144 (1985)

2. A. B. Woodside, Z. Huang, and C. D. Poulter, Trisammonium geranyl diphosphate. *Org. Synth.* **66**, 211 (1988).

3. S. Andrews. FastQC: A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010)

4. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

5. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

6. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

7. M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

8. R. Vaser, I. Sović, N. Nagarajan, M. Šikić, Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737-746 (2017).

9. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, A. M. Earl, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLOS ONE, **9**, (11) e112963 (2014).

10. M. Seppey, M. Manni, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).

11. T. Lassmann, Kalign 3: multiple sequence alignment of large datasets. *Bioinformatics* **36**, 1928–1929 (2020).

12. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

13. P. D. Scesa, Z. Lin, E. W. Schmidt, Ancient defensive terpene biosynthetic gene clusters in the soft corals. *Nat. Chem. Biol.* **18**, 659–663 (2022).

14. Wilson, K.A., de Rond, T., Burkhardt, I., Steele, T.S**.,** Schäfer, R., Podell, S., Allen, E.E., Moore, B.S. Terpene biosynthesis in marine sponge animals. *Proc. Natl. Acad. Sci. U.S.A*. **120,** (9) e2220934120 (2023).

15. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49**, W293–W296 (2021).

16. I. Burkhardt, N. B. Kreuzenbeck, C. Beemelmanns, J. S. Dickschat, Mechanistic characterization of three sesquiterpene synthases from the termite-associated fungus *Termitomyces*. *Org. Biomol. Chem.* **17**, 3348–3355 (2019).

17. M. Suzuki, M. Segawa, H. Kikuchi, T. Suzuki, E. Kurosawa, (5S,7R,10R)-selin-4(14)-en-5α-ol, a sesquiterpene alcohol from the red alga *Laurencia nipponica*. *Phytochemistry* **24**, 2011–2012 (1985).

3.10 Acknowledgements

Chapter 3 contains material currently being prepared for publication, coauthored with Burkhardt, I., Moore, M., de Rond, T., Michael, T., and Moore, B.S. We are grateful to the Joint Genome Institute, Department of Energy for their ongoing support and sequencing data generated for this project (CSP-504335). The dissertation author was the primary researcher and author of this material.

CHAPTER 4. Ongoing investigations of red algal genomics and haloterpenoid biosynthesis downstream tailoring reactions

4.1 Introduction and Context for Chapter 4

Chemistry is the language of the natural world, and marine organisms are gifted producers of bioactive small molecules. Unlike in terrestrial plants and animals, which are primarily known to incorporate chlorine and iodine into small molecules, marine organisms are well understood to also utilize bromine.[1] Bromination is a hallmark of marine chemistry, and made possible by the composition of ocean water, which is approximately 0.5 M chloride, 0.9 mM bromide, and 0.4 μM iodide.[1] Marine red macroalgae are highly productive producers of halogenated terpenoids, or haloterpenoids, with some species capable of producing up to 5% of their dry weight.[2] Early research of algal haloterpenoids conducted in the 1960s and 1970s catalogued the immense structural diversity and rich bioactivities of algal haloterpenoids and the discovery and characterization of algal haloterpenoids marked a significant paradigm shift during this period. As science-policy makers were grappling with the use and management of commercial halogenated pesticides and other anthropogenic sources of halogenated small molecules, it was remarkable to discover naturally halogenated molecules.[3] Now, we endeavor to understand the genetic basis for haloterpenoid biosynthesis and apply new sequencing technologies for the discovery of biosynthesis genes in red macroalgae.

Vanadium-dependent haloperoxidase (VHPO) have been isolated from all classes of marine algae and widely implicated in the biosynthesis of halogenated marine natural products.[1,4] Foundational work from Professor Alison Butler's group established the role of VHPOs in red algal haloterpenoid biosynthesis, [4,5,6] yet, a significant challenge remains to understand the underlying genetic basis of halogenation biochemistry in red algal

natural products. Further research has been hindered by a dearth of genetic information from haloterpenoid producing red algae and limited access to key haloterpenoid producing organisms. Similarly, marine red algal genomes are inherently difficult to sequence due high levels of microbial contamination from marine environments and large quantities of polysaccharides and polyphenolics inhibiting sufficient, clean, DNA extractions.

In Chapter 3, I presented work outlining initial sequencing and genome mining efforts in haloterpenoid producing red macroalgae and characterized the first step in haloterpenoid biosynthesis. Chapter 4 complements this work, with an in-depth discussion of whole genome sequencing of red macroalgae and attempts to answer outstanding questions around gene clustering through the development of gene co-occurrence networks to identify new haloterpenoid associated gene clusters. Finally, Chapter 4 explores the current status of continued work on the biosynthesis of red algal haloterpenoids.

4.2 References for Chapter 4 introduction

1. V. Agarwal, Z. D. Miles, J. M. Winter, A. S. Eustáquio, A. A. El Gamal, B. S. Moore, Enzymatic halogenation and dehalogenation reactions: Pervasive and mechanistically diverse. *Chem. Rev.* **117**, 5619–5674 (2017).

2. W. Fenical, Halogenation in the Rhodophyta, *J. Phycol.* **11**, 245–259 (1975).

3. W. Fenical, D. John Faulkner (1942–2002): Marine natural products chemistry and marine chemical ecology. *Angew. Chem., Int. Ed.* **42**, 1438–1439 (2003).

4. Alison. Butler, J. V. Walker, Marine haloperoxidases. *Chem. Rev.* **93**, 1937–1944 (1993).

5. A. Butler, J. N. Carter-Franklin, The role of vanadium bromoperoxidase in the biosynthesis of halogenated marine natural products. *Nat. Prod. Rep.* **21**, 180–188 (2004).

6. J. N. Carter-Franklin, A. Butler, Vanadium bromoperoxidase-catalyzed biosynthesis of halogenated marine natural products. *J. Am. Chem. Soc.* **126**, 15060–15066 (2004).

4.3 Introduction

The study of the chemistry of the natural world is an ever-expanding frontier. The foundations of natural products research were laid in the latter half of the twentieth century, and the introduction of technologies like nuclear magnetic resonance (NMR) instruments and mass spectrometers profoundly transformed natural product isolation and structural elucidation.[1] Similarly, SCUBA, in addition to remotely operated and autonomous underwater vehicles (ROV/AUV) have expanded access to the marine world beyond nearshore and intertidal zones.[2] The wealth of extremely bioactive marine natural products isolated from charismatic creatures like red macroalgae,[3] sponges,[4] and octocorals[5] have often proven to be valuable drug targets (**Figure 4.1**). However, supply

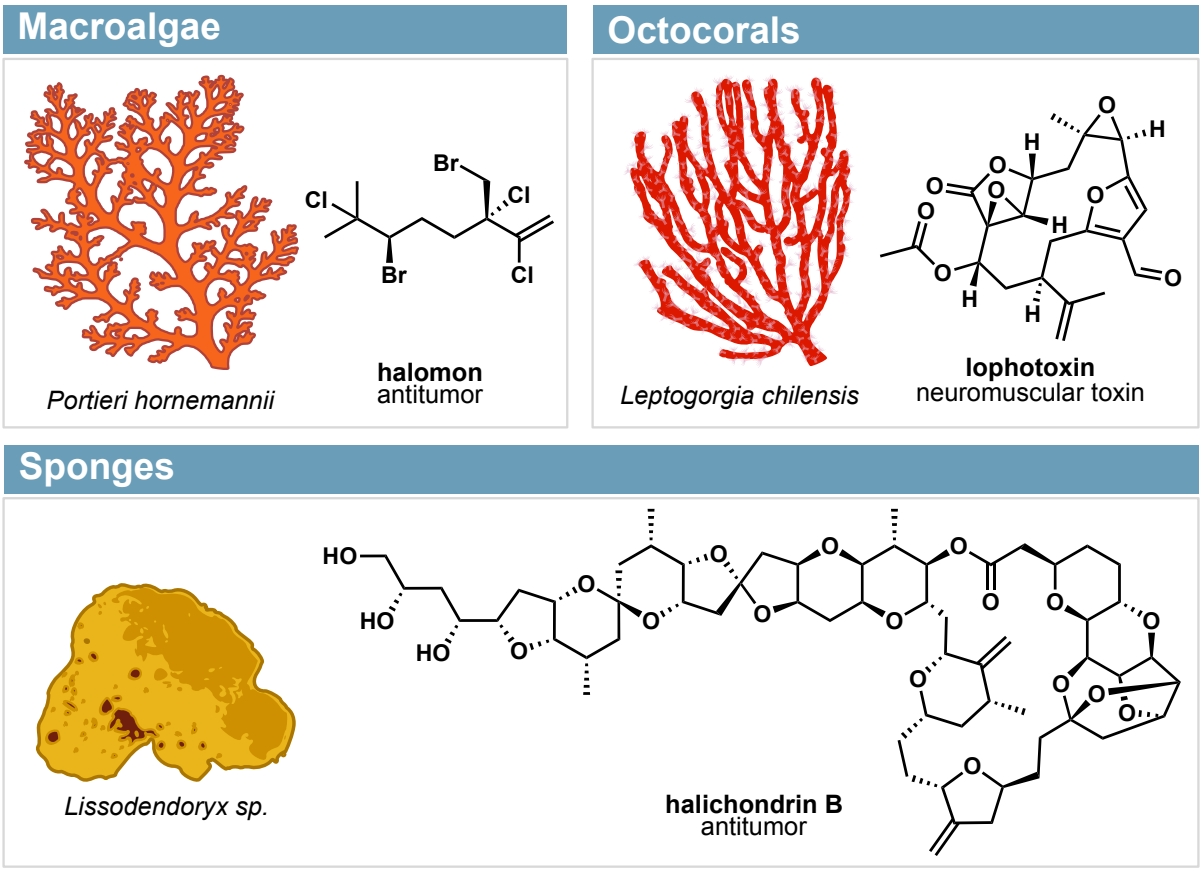


Figure 4.1 Supply limited bioactive marine natural products and their host organisms.

challenges due to the availability and scalability of sensitive marine organisms,[6] which typically remain unculturable, has limited efforts to bring promising compounds to the clinic.[7] Thus, the complementary approaches of synthetic and biosynthetic research in natural products are invaluable strategies to preserve sensitive marine organisms while gaining access to their unique and potent chemistry. These branches of research have emerged and grown in the last several decades, with natural products biosynthesis research becoming intimately tied to the growth of sequencing technologies.

The field of natural products biosynthesis has been fundamentally transformed by the rapid accumulation of genomic and transcriptomic information in what is now referred to as the "post-genomics era". Once sequencing data is available for an organism, we can access the genetic blueprints for a given small molecule and utilize synthetic and biosynthetic approaches to circumvent source organism supply issues. Advancements in computing resources and computational tools, as well as the emergence of technologies like long-read, single-molecule DNA sequencing from Pacific Bioscience (PacBio) and Oxford Nanopore Technologies (ONT) have made it cheaper, faster, and more efficient to connect genes to chemistry.[8,9] As a result, as marine natural products biosynthesis research shifts into non-model eukaryotic organisms, like red macroalgae, a new sequencing frontier has emerged.

Unlike the relatively simple genomes of microbes, eukaryotes typically have large, highly repetitive genomes, which are less amenable to conventional sequencing and genome mining workflows established in microbial systems. For example, biosynthetic genes cassettes in red macroalgae are often in tandem duplications and nested in repetitive sequences. This makes identifying biosynthetic gene clusters in red macroalgal

genomes almost impossible without long-read sequencing data. Similarly, contamination in marine sequencing datasets can cause significant issues when constructing quality genome assemblies. The associated microbiomes of many marine organisms are dense and diverse in comparison to their terrestrial counterparts, with microscopic organisms representing more than two-thirds of the biomass in the ocean.[10] Red macroalgal genome sequencing has proven particularly challenging due to the close association of seaweeds with a diverse array of marine microbes and crustaceans, which means sequencing efforts often result in a highly complex hologenome.[11–13]

The generation of single reads in the hundreds of kilobases (kb) with PacBio HiFi and/or Oxford Nanopore long-read data is vital to differentiating between host and microbial sequences. Large high-quality scaffolds showcase classic eukaryotic features, such as long intergenic regions, repetitive elements and genes containing introns, additionally a larger genomic context of target genes/regions enables more accurate taxonomic classification. While non-host contamination does cause issues with downstream assembly and analysis, it is worth noting the value of associated microbial sequences in efforts to differentiate between host and microbial sequences. Investigations into the biosynthetic origins of many marine-derived molecules have traditionally focused on the associated microbiome of an organism and not the host itself. Until recently, natural products from sessile marine animals and plants were hypothesized to be primarily biosynthesized by symbiotic microbes rather than by the host organism. However, this long-standing paradigm has been repeatedly challenged in recent years, with biosynthetic genes and gene clusters now known from macroalgae,[14,15,15] corals,[16,17] and sponges.[18] Additionally, these discoveries highlight the importance of long-read

sequencing data for the discovery of biosynthesis genes in non-model eukaryotic organisms.

Previous work by the Moore lab and others has demonstrated the feasibility of red macroalgal genome sequencing for natural products discovery,[14,15,19] where the goal is often identifying a single gene or genomic region. This unique application typically does not require a complete, chromosome-scale, genome assembly,[20] but often necessitates a high-quality and contiguous genome assembly. For example, Chekan *et al.* performed whole genome sequencing on the well-studied kainic acid producer *Digenea simplex* to identify kainic acid biosynthetic genes and determine their genomic context. This study utilized single-molecule, long-read sequencing technologies from Oxford Nanopore Technologies (ONT) to generate a 299 Mb uncorrected genome assembly with 8246 contigs and a contig N50 of 57 kb. Notably, the longest raw read was 1.2 Mb in length. Although by no means a reference quality, chromosome-scale, genome assembly, this draft assembly enabled the discovery of one of the first known biosynthesis gene clusters in red macroalgae outside of mycosporine-like amino acid biosynthesis.[11,12] This work is also the basis for optimized nucleotide isolation strategies reported here.

Beyond managing microbial contamination, it is vital to obtain high-quality nucleic acids from target organisms. Long-read sequencing platforms like ONT and PacBio require sufficient, high molecular weight (HMW) DNA and both platforms can be sensitive to physical contamination co-purified with DNA samples. In particular, nucleotide extraction from marine plants, like algae and seagrass, is hampered by the large quantity of polysaccharides and polyphenolics present in sample tissues. Contamination of nucleic acids with problematic biomolecules can lead to reduced library prep and sequencing

outcomes. Similarly, clean-up strategies like column-based clean-up kits result in fragmented DNA samples not suitable for long-read sequencing applications. At present, several protocols are available for extracting HMW DNA from red algae,[12,14] however inconsistent results and large biomass requirements have hindered further development. Here we report experimental protocols tailored for red algal systems, including high-molecular weight DNA extraction procedures and genome mining strategies.

Finally, this work reports progress towards the reconstitution of haloterpenoid biosynthesis pathways and efforts to biochemically characterize vanadium-dependent haloperoxidases (VHPOs) in red macroalgal systems. VHPOs are the predicted enzyme responsible for the halogenation of algal haloterpenoids and in Chapter 3 of this dissertation I report the first haloterpenoid associated biosynthetic gene pair in red macroalgae, where a terpene synthase (TS) and VHPO were co-clustered in the genome of the red macroalga *Laurencia subopposita*. Here I describe *in vitro* efforts achieve soluble expression of red algal VHPOs from *Plocamium pacificum*, *Portieria hornemannii, Laurencia pacifica,* and *L. subopposita*. I also report *in vivo* efforts to biochemically characterize the co-clustered VHPO found in the *L. subopposita* VHPO-TS gene pair. VHPOs are promising biocatalysts not only for the production of native algal haloterpenoid products, which are often inherently bioactive and valuable drug targets, but also for more general enzymatic halogenation reactions. Many pharmaceuticals are halogenated, and the addition of a halogen is a well-proven strategy to increase the efficacy of pharmaceuticals.[21] Overall, this section is intended as a progress report on the current status of these ongoing projects and a roadmap for future lab members interested in this work.

4.4 Red Algal DNA Barcoding

Ahead of committing whole genome sequencing resources, we have found value in utilizing red algal DNA barcoding techniques to validate species identifications. Briefly, algal genomic DNA (gDNA) is extracted with the GeneJET Genomic DNA Purification Kit (ThermoFisher) according to the manufacturer's instructions. If the GeneJET DNA kit or other column-based kits are utilized, it may be necessary to remove PCR inhibiting contaminates common in red algal samples. Red algal gDNA samples are amenable to clean-up with the Genomic DNA Clean & Concentrator kit (Zymo Research). Although resulting DNA samples are not sufficient for long read sequencing applications, samples are adequate for PCR amplification and short-read sequencing.

Following gDNA extraction, we employ PCR primers widely used in Florideophyte phylogenetic and barcoding studies developed by the Saunders group.[22] Although there is no perfect "universal" primer, we have found using a combination of reported primers and target genes is sufficient to confirm taxonomic assignment. Here we apply the ribulose-1,5-bisphosphate carboxylase large subunit (rbcL) primer set, and the cytochrome c oxidase subunit 1 extended fragment (COI) primer set (**Table 4.1 and 4.2**) with the following amplification profile: 94 °C for 2 min; 35 cycles of 94 °C for 20 s, 55 °C annealing for 30 s, 72 °C extension for 30 s; followed by 72 °C final extension for 10 min. Both 5 PRIME HotMasterMix (QuantaBio) or LongAmp *Taq* 2x Master Mix (NEB) polymerase mixes performed well with algal samples. Amplified PCR fragments were submitted for Sanger sequencing and sequencing results were compared to top BLASTn hits. Querying NCBI reference databases usually resulted in good quality BLASTn results with >95% identity matches by nucleotide sequence. For example, the sample visually

identified as *Laurencia pacifica* had a >99% match by nucleotide sequence with reference database samples submitted from reliable phycologists and algal taxonomists.

Table 4.1 DNA barcoding primers used in this study: RuBisCo: Ribulose-1,5-bisphosphate carboxylase large subunit (rbcL, ~1,350 bp)

Name	**Primer**
F57	GTAATTCCATATGCTAAAATGGG
rbcLrevNEW	AACATTTGATGTTGGAGTYTC

Table 4.2 DNA barcoding primers used in this study: CO1: Cytochrome c oxidase subunit 1 extended fragment (COI, ~1,232 bp)

Name	**Primer**
M13LF2	TGTAAAACGACGGCCAGTCAAAYCAAYAARGATATHGGNAC
M13Ri	CAGGAAACAGCTATGACGGRTGICCRAARAAYCARAA

In the case of *Plocamium pacificum*, further species validation was necessary. This species has been called *P. cartilagineum* and *P. cartilagineum* subspp. *pacificum*, but *P. cartilagineum* was described by Carl Linnaeus and no doubt did not come from the Pacific Ocean. *Plocamium pacificum* was described by Harald Kylin from Friday Harbor, WA in 1925.[23] In collaboration with Dr. Kathy Ann Miller (UCB), we confirmed *Plocamium pacificum* is the current name, and a better bet for this widespread west coast species. In conclusion, assigning taxonomy in macroalgal systems is convoluted,[24] and in this work, we utilize a combination of traditional and molecular techniques to confidently assign species identifications.

4.5 High Molecular Weight DNA Extraction for Marine Macroalgal Tissue

Developed in collaboration with Malia Moore and the Michael lab at the Salk Institute for Biological Studies. The following section is a reprint in full of a protocol published on protocols.io (dx.doi.org/10.17504/protocols.io.14egn2dnpg5d/v1)

Introduction

Marine macroalgae contain a variety of unique cell wall components including sulfated polysaccharides and polyphenolics. These components often co-elute with high molecular weight (HMW) DNA and lead to reduced library prep and sequencing outcomes. This protocol incorporates polyvinylpolypyrrolidone (PVPP) and β-mercaptoethanol (BME) to reduce polyphenolic contamination, and an early salting out step with potassium acetate (KOAc) to address polysaccharides.[14] This protocol is largely adapted from an Oxford Nanopore HMW DNA extraction from Arabidopsis leaves, which incorporates the QIAGEN Blood and Cell Culture DNA Midi Kit for column cleanup.[25] The DNA product often requires additional cleanup after elution, and we suggest the BluePippin 15kb size selection for all HMW applications.

Equipment

- Lyophilizer
- Stir plate
- Heat block or water bath
- Vortex
- Mortar and pestle
- Refrigerated centrifuge for spins up to 3,500 x g with 50 mL
- Suggested Sage Science BluePippin

Consumables

- Stock solution: 1M Tris-HCl, pH 9.5
- Stock solution: 5M sodium chloride (NaCl)

- Stock solution: 500mM ethylenediaminetetraacetic acid (EDTA)
- Stock solution: 5M potassium acetate (KOAc)
- Cetyltrimethylammonium bromide (CTAB)
- Polyethylene glycol (PEG) 8000
- β-mercaptoethanol (BME)
- Polyvinylpolypyrrolidone (PVPP)
- RNase A, 100mg/mL (eg. QIAGEN Mat. #1007885)
- 100% isopropanol
- 95-100% ethanol
- Nuclease-free water
- QIAGEN Blood and Cell Culture DNA Midi Kit (Cat. #13343)
- Tris-EDTA (TE) buffer
- 50 mL Falcon Tubes
- 1.5 mL collection tubes (eg. Eppendorf DNA LoBind tubes Cat. #022431021)
- Suggested: Sage Science High Pass Plus Cassette (BPLUS10 or BPLUS03) for BluePippin

Lyophilizing algal tissue

1. Flash-freeze algal tissue in liquid nitrogen (target ≥5g wet tissue).
2. Quickly transfer sample to lyophilization container and freeze dry for 36-48 hours.
3. Macerate the tissue with a clean spatula to increase surface area and put on the lyophilizer for another 24 hours.
4. Remove and refrigerate with desiccant for immediate use, or store at -80C for longer periods.

Setting up the DNA extraction

1. Prepare desired volume of Carlson lysis buffer (100 mM Tris-HCl, pH 9.5, 2% CTAB, 1.4 M NaCl, 1% PEG 8000, 20 mM EDTA) and mix overnight on a magnetic stirrer. The stock solutions suggested under consumables will yield a homogenous buffer with no precipitate.
2. Pre-heat a heat block or water bath to 65 °C and place in a fume hood.
3. For each extraction, transfer 20 mL of Carlson lysis buffer to a 50 mL Falcon tube. In a fume hood, add 400 µl BME (originally 50uL) and mix by vortexing. Pre-warm the solution to 65 °C for 30 minutes before starting the extraction.
5. Scoop 0.5 teaspoons lyophilized plant tissue into a clean mortar and add 50-100mg powdered PVPP. Grind with pestle for ~30 seconds, until tissue is powdered and combined, but not long enough to introduce significant moisture. Move immediately into DNA extraction.

Lysis and first precipitation

1. Pour tissue into the warm buffer. Invert 5 times.
2. Add 40 µl of RNase A and vortex for 5 seconds.
3. Optional: If using a heat block with mixing, set the block (still at 65 °C) to mixing at 300rpm for 5 minutes.

4. Incubate for 1 hour at 65 °C. Invert 10 times every 15 minutes. At 30 minutes, add another 40 uL of RNase A, inverting 10 times to combine.
5. Allow the tubes to cool down to room temperature for 10 minutes.
6. Add 20 mL chloroform and vortex for two pulses of 5 seconds each.
7. Centrifuge the tubes at 3500 x g for 15 minutes at 4 °C.
8. In a fume hood, transfer the top layer of lysate from each tube to a new 50 mL Falcon tube, without disturbing the interphase.
   Tip: The lysate layer should be 14-18 mL of solution, but it is recommended to use wide-bore tips, transferring 1 mL at a time. Tips can also be widened by cutting standard P1000 tips.
9. Mix supernatant with 0.4X 5M potassium acetate (KOAc) at room temperature, inverting at least 10 times to combine, then incubate on ice for 20 min.
10. Centrifuge the tubes at 3500 x g for 45 minutes at 4 °C.
11. Remove and retain the supernatant.
    Tip: This may best be done by pouring slowly and observing the polysaccharide-salt pellet, which may be mobile. Leave some liquid behind in favor of avoiding the pellet.
12. Add 0.7X volumes of isopropanol. Invert 10 times. Incubate at -80 °C for 15 minutes. Do not extend this incubation.
13. Centrifuge the sample at 3500 x g for 45 minutes at 4 °C.
    Tip: If available, a fixed-angle centrifuge will create a pellet on the wall of the tube that has greater surface area for dissolution in step 15 (as compared to a conical pellet at the base of a falcon tube from a swinging bucket).
14. Discard the supernatant without disturbing the pellet. Use sterile wipes to absorb the liquid on the tube walls, being careful not to disturb the pellet.
15. To each pellet, add 10 mL G2 buffer, from the QIAGEN kit. Incubate at 50 °C for 30-60 minutes, or until the pellet is dissolved. Swirl the pellet to mix but do not try to pipette or vortex.

<u>Column cleanup</u>

1. Equilibrate a QIAGEN Genomic-tip 100/G column with 4 mL of Buffer QBT.
2. Pour the DNA in G2 buffer through the equilibrated column and allow it to flow through with just gravity.
3. Once all the lysate has passed through, wash the column with 8 mL of Buffer QC.
4. Repeat the wash with another 8 mL of Buffer QC.
5. Place the column over a clean 50 mL Falcon tube, and elute the genomic DNA with 5 mL of Buffer QF, pre-warmed to 55 °C.
6. Allow the eluate to cool down to room temperature.
7. Add 3.5 mL of isopropanol to the eluted DNA and mix by inverting the tube 10 times.
8. Incubate the tube at -20 °C for at least 3 hours, or overnight.

<u>Final precipitation</u>

1. Centrifuge the tube at 3500 x g for 45 minutes at 4 °C.
2. Discard the supernatant without disturbing the pellet.

3. Add 4 mL of ice-cold 70% ethanol to the pelleted DNA and invert the tube 10 times.
4. Centrifuge at 3500 x g for 10 minutes at 4 °C.
   Tip: If using a swinging bucket centrifuge the DNA will pellet at the base of the tube and be easy to locate and resuspend. If using a fixed angle, mark the side of the tube that faces outwards in order to locate the pellet for washes and elution.
5. Discard the supernatant without disturbing the pellet. Use sterile wipes to dry the tube walls, being careful not to disturb the pellet.
6. Resuspend the DNA in 100 µL of TE buffer and incubate at room temperature, typically overnight.
7. Transfer the DNA into a nuclease-free 1.5 mL tube (DNA LoBind tube preferred) using a wide-bore tip, and store at 4 °C.
   Tip: Often, waiting a further 48 hours before quantifying on Nanodrop and Qubit will allow the DNA to further relax and yield the most accurate results.
8. Carry samples forward to BluePippin size selection, if available. This gel separation will retain DNA fragments greater than 15 kb and discard any residual contamination still evident on a Nanodrop trace. For these benefits, expect 50-70% loss of DNA.

Additional tips

- In the field or in lab, it is vital to scrape off all surface epiphytes and wash the sample in clean water before flash freezing to reduce contaminants common in the marine environment that confound genome assembly.
- Marine macroalgae are incredibly diverse in biochemical content, so individual
- seaweeds may require troubleshooting. Suggested alterations include varying input tissue type or quantity, increasing CTAB or BME percent, or adding a second chloroform separation.
- It may be necessary to carry out extractions of the same tissue in parallel to yield sufficient DNA, especially when large losses from BluePippin are expected. It is not suggested to combine multiple extractions onto the same column, as this may lead to overloading and a dirty sample. This protocol as written, paired with BluePippin, has produced sequencing-quality DNA for Nanopore from a red alga *Porteria hornemanii* and a brown alga *Macrocystis pyrifera*. For *P. hornemanii*, a single 20 mL extraction produced sufficient DNA for sequencing, but for *M. pyrifera*, three parallel extractions of 20 mL were necessary.

4.6 Genome Mining and Co-occurrence Networks

Historically, genome mining for natural products biosynthesis genes has focused on microbial systems. Microbes typically cluster biosynthetic genes associated with the production of a specialized metabolite into tightly organized, co-regulated, gene cassettes. Gene clustering in conjunction with the relatively small size of microbial genomes, has made genome mining and discovery efforts in microbial systems much simpler. In recent years, technological advancements in long-read sequencing have facilitated the generation of more contiguous assemblies capable of addressing clustering questions in eukaryotic genomes, which are characterized by long intergenic regions, repetitive elements, and intron-containing genes. For example, the biosynthetic genes of some specialized plant metabolites appear to be clustered in the genomes of higher plants.[26] There are now a handful of examples of biosynthetic gene clustering in marine plants (macroalgae), which includes gene clusters associated with kainoid biosynthesis,[14,15] bromoform biosynthesis,[19] and mycosporine-like amino acid biosynthesis.[11,12] With this in mind, I set about developing an unbiased genome mining strategy to identify physically co-localized, or co-occurring, genes of interest in complex metagenomic red algal datasets.

To capture genes encoding target enzymes in a biosynthetic pathway of interest, I utilized profile hidden Markov models (profile-HMMs), which are statistical models and search methods that ultimately rely on conserved residues or domains in a biological sequence to classify and annotate gene sequences. Because of their reliance on conserved motifs and not overall homology, Profile-HMMs are more flexible than BLAST-based searches for genome mining and well-suited for eukaryotic systems. To identify

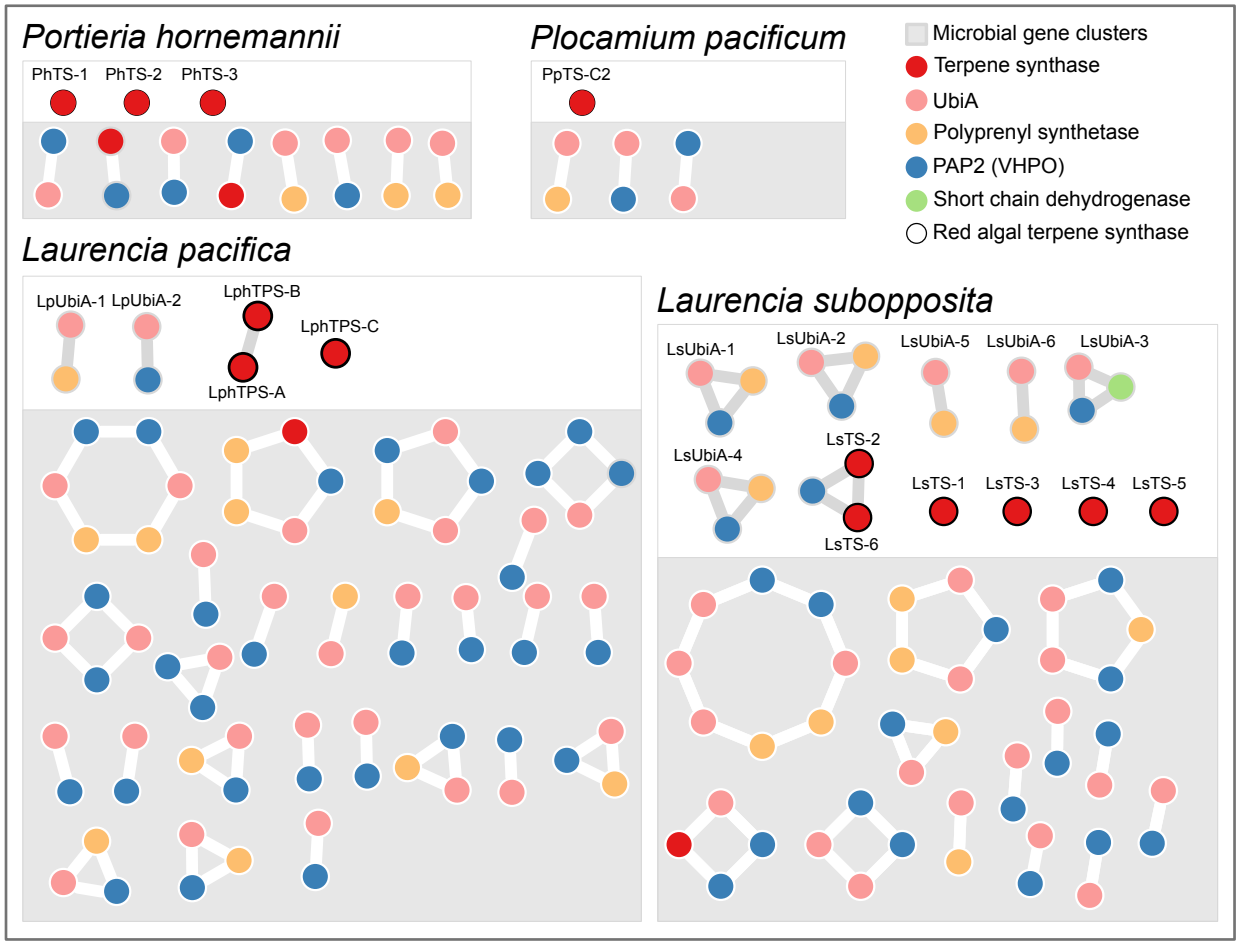


Figure 4.2 Gene co-occurrence network. Each dot represents a gene, and each connection indicates physical co-occurrence on a contiguous sequence without respect to order or distance. Pink – UbiA, orange – IDS, blue – VHPO, green – short chain dehydrogenase.

haloterpenoid associated gene clusters in red macroalgae, I screened the draft genome assemblies of *L. pacifica*, *L. subopposita*, *P. hornemannii* and *P. pacificum* with the following Pfams: Terpene synthase family 2, C-terminal metal binding (PF19086), UbiA prenyltransferase superfamily (PF01040), PAP2 superfamily (VHPO) (PF01569), and Polyprenyl synthetases (PF00348). Seeing a variety of well scoring hits for both terpene synthases and UbiA's we chose to include both classes of terpenoid associated enzymes in our analysis to capture a broader understanding of terpenoid biochemistry in red macroalgae.

Next, the results from each Pfam search were queried for two hits occurring on the same contiguous sequence (contig). All contigs with more than two hits from different Pfam searches were extracted from their larger genome files and annotated with the Conserved Domain Database to quickly scan all known protein domains in addition to the initial Pfam search results. The resulting list of annotated sequences were then sorted into putative 'host' and 'microbial' bins based on overall gene cluster structure and annotation results. This step is necessary because the sequencing datasets utilized in this study constitute a hologenome of each organism, which can be defined as the total

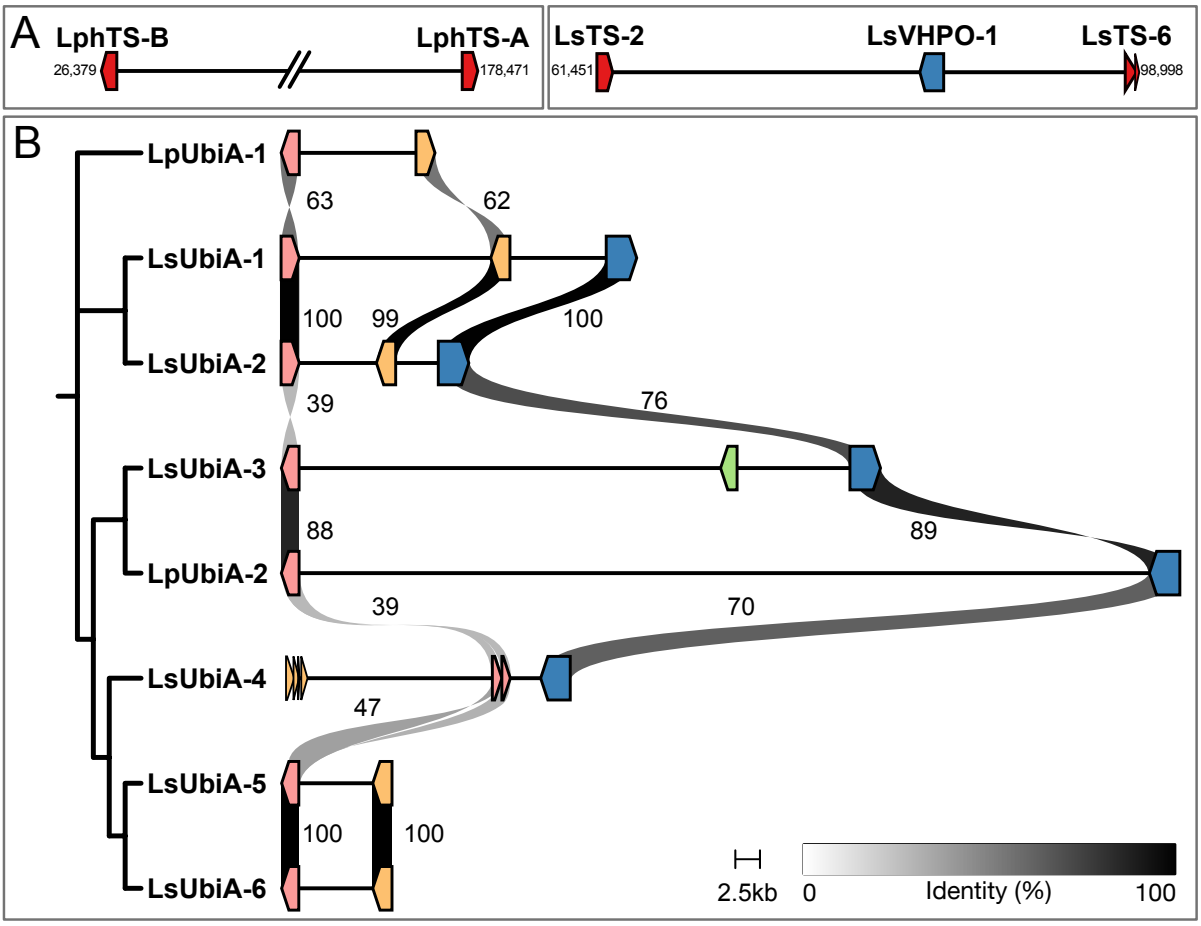


Figure 4.3 Gene clusters in *Laurencia subopposita* and *Laurencia pacifica*. **A.** TS gene clusters, red – algal TS, blue – VHPO. **B.** Syntenic comparison of UbiA gene clusters. Pink – UbiA, orange – IDS, blue – VHPO, green – short chain dehydrogenase. Short broken up CDS regions indicate stop codons/indels. shading between genes is relative to the similarity of gene pairs by amino acid sequence percent identity in agreement with discrete measures of sequence percent identity as labeled.

genomic components of the host (nuclear, mitochondrial, and plastid genomes) and associated microbiota (bacterial, viral, archaeal, and fungal).[27] Finally, the annotated and binned list of sequences was used to generate a gene co-occurrence network where each node represents a gene and each edge (connection) indicates physical co-occurrence on a contig without respect to gene order or distance (**Figure 4.2**).

Within the *P. hornemannii* and *P. pacificum* genomes we did not identify any instances of a red algal terpenoid genes co-clustered with any family of halogenating enzyme. Excitingly, in addition to previously identified TS-containing gene clusters in both *L. pacifica* and *L. subopposita* (**Figure 4.3A**), there were numerous instances of UbiA-containing gene clusters. UbiA sequences were found on eukaryotic contigs and generally co-occur with polyprenyl synthetases (IDS), which could provide the necessary diphosphate precursors for terpenoid production. Additionally, most clusters contained a VHPO sequence, and in one instance, a short chain dehydrogenase (SDH) (**Figure 4.3B**). Both VHPO and SDH enzymes are commonly used in the oxidative functionalization of terpenoid natural products.[17,28] Phylogenetic analysis of the UbiA sequences revealed gene conservation between *L. pacifica* and *L. subopposita* and clarified the presence of two sets of alleles (100% amino acid sequence identity) in *L.* subopposita, LsUbia-1 & -2 and LsUbiA-5 & -6. Syntenic comparison of the UbiA-containing gene clusters supports the emergence of two UbiA-containing gene cluster structures conserved between *L. pacifica* and *L. subopposita*. The first consists of a core gene pair with a UbiA and an IDS, and in some cases a VHPO, like in the LsUbiA-1 cluster (**Figure 4.3B**). The second is spread over a relatively larger genetic space, includes a UbiA and a VHPO, and in the LsUbiA-3 cluster, also contains an SDH. The discovery of

this new family of UbiA-containing gene clusters unique to the genus *Laurencia* could explain halogenated terpenoid chemistry not accounted for by the MTPS family of terpene synthases. However, it is clear the diversity of the known metabolome correlates to the diversity of sequence as the genus *Laurencia* accounts for almost half of know natural products in red seaweeds. Ongoing investigations of candidate genes present in these two related species, including further study of the newly identified UbiA-containing gene clusters, will shed additional light on this remaining piece of the algal haloterpenoid biosynthetic puzzle. Additionally, this method of identifying physically co-occurring biosynthetic genes in eukaryotes could be modified into an automated pipeline to rapidly analyze sequencing datasets.

4.7 Vanadium-Dependent Haloperoxidase Expression Efforts Introduction

Halogenated natural product compounds represent a diverse class of anthropologically relevant organic molecules, with many applications in human health and the environment. However, supply challenges often constrain the potential of promising bioactive natural products from rare or environmentally sensitive marine organisms. Our discovery of the halogenated terpenoid biosynthesis genes in marine algae and progress towards flagship bioactive molecules, like the antitumor agent halomon, offers a path to bioengineering sustainable laboratory microbial fermentation.

Halomon was selected for preclinical drug development in the 90s for its selective activity against aggressive tumor cell lines, yet stalled due to insufficient natural material.[3,6] A scalable, enantioselective synthesis of halomon was recently developed to address product supply issues and introduce a new synthetic enantioselective

dihalogenation strategy.[7] Despite elegant synthetic methods developed for the assembly of these complex halogenated compounds, current syntheses rely on directing groups and high loadings of a chiral catalyst to obtain selective dihalogenation.[29–31] We seek to develop a complementary biocatalytic method that circumvents those synthetic challenges. In this section I describe efforts to characterize alkene halofunctionalization biochemistry in the production of halogenated natural products from red algae.

The biogenesis of bromochlorinated terpenes has been of interest for decades, with early hypotheses suggesting the coordinated enzymatic addition of bromonium and chloride to an alkene by a dihalogenase.[32–34] Despite promising biocatalytic potential,[35,36] no such enzyme has ever been discovered nor engineered. With several red algal genomes and transcriptomes now available, we are poised to solve this lingering

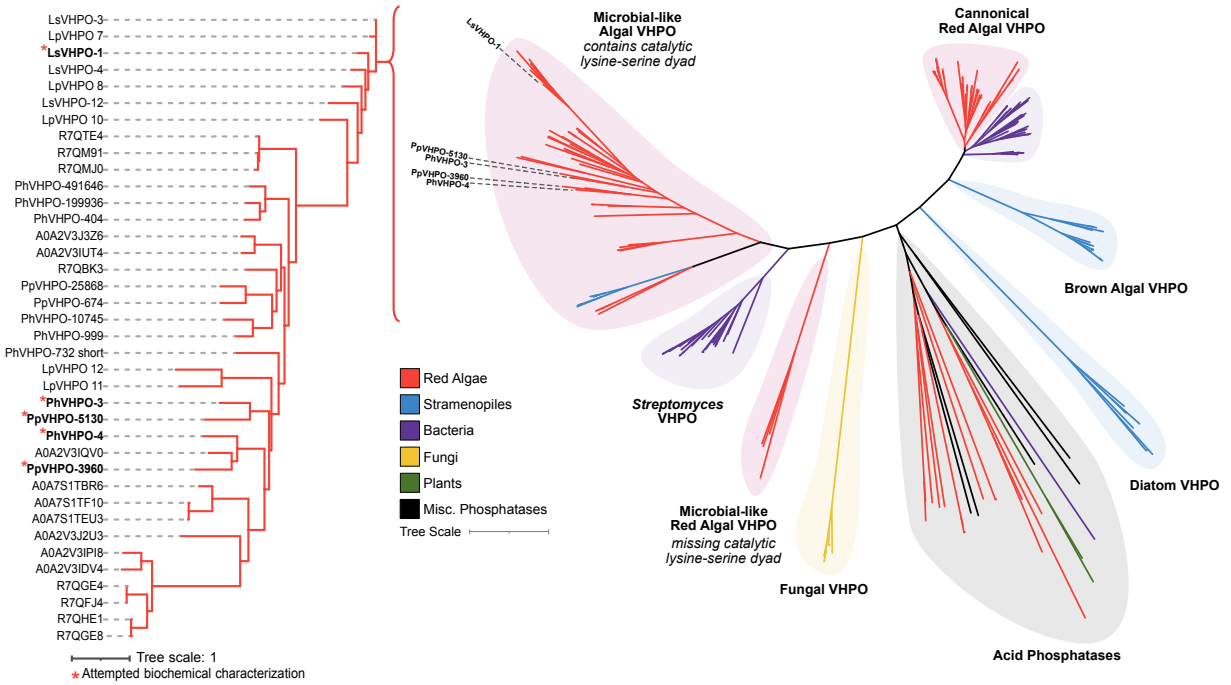
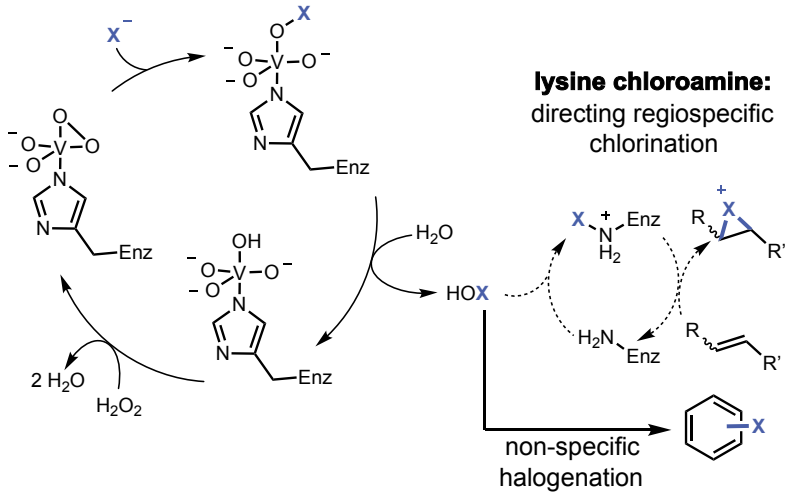


Figure 4.4 Representative phylogeny of vanadium-dependent haloperoxidases (VHPO) across various taxa and detailed view of new 'microbial-like' algal VHPO clade. Sequences attempted to be biochemically characterized are bolded and highlighted with an asterisk and labeled in the context of the larger tree.

biochemical challenge and develop new biocatalytic strategies in chemo-, regio- and enantioselective dihalogenation reactions. Based on preliminary genomic and biochemical data, we discovered a new family of vanadium-dependent haloperoxidase (VHPO) enzymes proposed to catalyze the long-sought 'dihalogenation' reaction. (**Figure 4.4**).

Preliminary sequencing data has resulted in draft genomes and transcriptomes for the notable red algal haloterpenoid producers *Portieria hornemannii, Plocamium pacificum, Laurencia pacifica*, and *Laurencia subopposita*. Genome mining and phylogenetic analysis of these sequencing datasets have enabled us to discover a new clade of VHPOs unique to red algae (**Figure 4.4**). This new red algal clade is closely related to Streptomyces spp. VHPOs that we first discovered over 10 years ago, including those involved in napyradiomycin (NapH1 and NapH3)[37,38] and merochlorin (Mcl24 and Mcl40)[39] biosynthesis. Recent work from the Moore lab describes a lysine-serine dyad near the vanadate binding site in the Streptomyces VHPOs, including NapH1 (PDB:



Scheme 4.1 Vanadium-dependent haloperoxidase (VHPO) mechanism. Specific microbial VHPOs are proposed to form a lysine chloroamine intermediate, which directs a regiospecific chlorination reaction. All other known VHPOs lack an additional lysine in the active site and perform non-specific halogenation reactions.

3W36), which is critical for stereoselective alkene halofunctionalization via formation of a chloramine intermediate (**Figure 4.5**).[38] While this lysine-serine dyad is absent in all previously characterized, non-selective algal VHPOs that generate diffusible hypohalous acids,[40,41] the dyad is present in the newly identified red algal haloperoxidase family (**Figure 4.6**). Notably, identified sequences from producers of bromochlorinated compound producers *P. hornemannii* and *P. pacificum* group together, possibly representing a clade of new dihalogenating enzymes. All candidate sequences are actively transcribed under the same conditions as the TSs that produce haloterpenoid

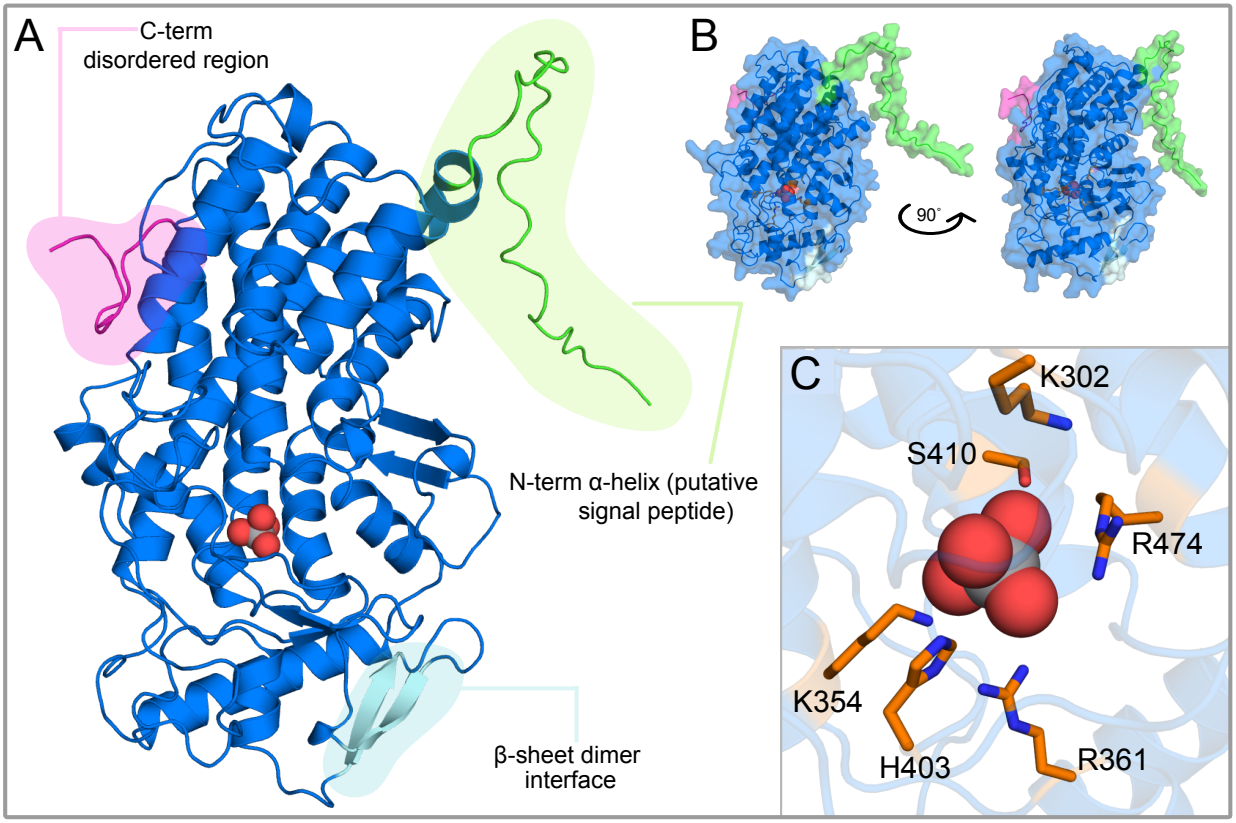


Figure 4.5 Model of LsVHPO-1 the co-clustered VHPO with a terpene synthase in the *Laurencia subopposita* genome **A.** Overall structure of LsVHPO-1 model, C-terminus disordered region highlighted in pink, N-terminus putative signaling peptide highlighted in green, and dimer interface highlighted in teal. **B**. Overall structure of LsVHPO-1 model with surface shown. Active site has limited solvent accessibility, perhaps contributing to the specificity. **C.** Active site of LsVHPO-1 with key residues labeled and shown in orange.

precursors, and one microbial-type algal VHPO was found to be co-clustered with an algal terpene synthase (Chapter 3). However, while seemingly promising, using VHPOs identified in this new clade as a biocatalyst remains problematic. To-date, no VHPO from this new clade has been successfully biochemically characterized. Although soluble protein was obtained from yeast (*Saccharomyces cerevisiae*) and actinomycete (*Streptomyces coelicolor* CH999) expression systems, none of the constructs expressed were biochemically active when tested with either the classical monochlorodimedone (MCD) assay[38,42] or with putative substrates in a variety of assay conditions (**Table 4.3** and **Table 4.4**). This has prevented further characterization and engineering efforts. No construct expressed in *Escherichia coli* has been soluble, which we hypothesize is due to structural features unique to microbial-type VHPOs. This includes an N-terminal α-helix, typically annotated as a chloroplast signaling peptide,[43] next there is a large, disordered region of varying length on the C-terminus (**Figure 4.6**). These features, as well as what can only be described as an unknown X-factor, has made the *in vitro* and *in vivo* reconstitution of this new clade of algal halogenases challenging. However, the promise of a new selective biocatalytic dihalogenation strategy to address haloterpenoid product supply issues makes further study of this system worthwhile.

4.8 General Strain Methodology

All *E. coli* strains were grown on LB agar or in LB broth with appropriate antibiotics for selection at 37 °C. For conjugation purposes, *E. coli* was grown using 2TY medium (1.6% (wt/vol) tryptone, 1% (wt/vol) yeast extract and 0.5% (wt/vol) NaCl) with appropriate antibiotic selection. For protein production, LB or TB media (1.2% w/v tryptone, 2.4% w/v

yeast extract, 0.4% (v/v) glycerol, 2.31% w/v KH_2PO_4, 12.54% w/v K_2HPO) was used, specific methods screened recorded below.

All *S. coelicolor* strains were grown on SFM agar plates (2% (wt/vol) D-mannitol, 2% (wt/vol) soya flour and 2% (wt/vol) agar) for conjugation and strain maintenance at 30 °C or in TSBY liquid medium (3% (wt/vol) tryptic soy broth, 10.3% (wt/vol) sucrose and 0.5% (wt/vol) yeast extract) at 30 °C with shaking at 220 r.p.m.[44] For protein production in *S. coelicolor* CH999, a TSBY preculture (100 µl) was used to inoculate the 30-mL 'primary culture' of Super YEME medium (0.3% (wt/vol) yeast extract, 0.5% (wt/vol) peptone, 1% (wt/vol) glucose, 0.3% (wt/vol) malt extract, 34% (wt/vol) sucrose, 0.5% (wt/vol) glycine, 0.235% (vol/vol) $MgCl_2 \cdot 6H_2O$ (2.5 M), 7.5×10^{-3}% (wt/vol) L-proline, 7.5×10^{-3}% (wt/vol) L-arginine, 7.5×10^{-3}% (wt/vol) L-cysteine, 0.01% L-histidine and 1.5×10^{-3}% (wt/vol) uracil, pH 7.2) with 50 µg/mL apramycin in a 250-mL flask with bottom spring. This culture was grown at 30 °C with shaking at 220 r.p.m. After 4 days, 4 mL of primary culture was used to inoculate 40 mL of Super YEME supplemented with apramycin to generate the 'secondary culture'. This culture was grown at 30 °C for 2 days. Finally, 600 mL of 'expression culture' was inoculated using 30 mL of secondary culture supplemented with apramycin (50 µg/mL) in 2.8-liter flasks with metal springs. After 2 days, thiostrepton was added to a final concentration of 10 µg/mL to induce protein expression.[44]

All *S. cerevisiae* strains were grown on YPD media[45] at 30 °C for maintenance. For transformation and selection, all *S. cerevisiae* strains were grown on Yeast Selection media (0.67% wt/vol) BD Difco Yeast Nitrogen Base without Amino Acids, (19.2×10^{-2}% wt/vol) Yeast Synthetic Drop-out Medium Supplements without uracil, (0.5% wt/vol)

glucose, for selection plates add (0.6% wt/vol) agar. For protein production a Yeast Selection media preculture (1 mL) was used to inoculate the 10 mL 'expression culture' of Yeast Selection media without glucose supplemented with (2% wt/vol) galactose and (0.2% wt/vol) glucose, for *in vivo* studies the 'expression culture' also contained (5 mM) KCl (5 mM) KBr, (0.5 mM) terpene substrate, and 10 μM Na_3VO_4.

Antibiotics used for selection: kanamycin 50 μg/mL, apramycin 50 μg/mL, tetracycline 8 μg/mL, ampicillin 100 μg/mL, chloramphenicol 25 μg/mL, nalidixic acid 25 μg/mL.

4.9 Attempted Expression of VHPOs in *Escherichia coli*

Thus far, no soluble expression conditions have been identified in *E. coli*. All efforts in this section were attempted with the most highly expressed microbial-type VHPOs from *Portieria hornemannii* (PhVHPO-3 and PhVHPO-4) and their respective homologs from *Plocamium pacificum* (PpVHPO-5130 and PpVHPO-3960) (**Figure 4.4**). In addition to standard BL21(DE3) *E. coli* competent cells, we screened OverExpress C41(DE3) (Sigma) and Lemo21(DE3) (NEB) *E. coli* strains which are both expression systems suitable for membrane and toxic proteins prone to insoluble expression, however neither system improved target protein solubility. All other optimization conditions were attempted in BL21(DE3) *E. coli* competent cells with an N-terminal hexahistidine (N-His6) tag in a pET28a vector. In addition to screening various *E. coli* expression strains, I also screened co-expression conditions with the Chaperone Plasmid Set (Takara) which did not yield any success. Efforts to optimize expression conditions include varying temperature (16 ˚C, 18 ˚C, and 30 ˚C), optical density upon induction, isopropyl-β-D-thiogalactopyranoside

(IPTG) concentration ranging from 10 μM - 500 μM, and media type (TB vs LB). These conditions yielded either no expression of target proteins or a small amount of insoluble protein. If some overexpression was detectable in the cell pellet fraction attempts to screen purification buffer conditions did not result in any solubility improvements. After identifying a signaling peptide on the N-terminus, I made several truncations based on homology models and signal peptide prediction software (SignalP[43] and Plant-mPLoc[46]) to remove the α-helix, which did not improve expression in any of the above conditions. I also noted a disordered region of varying length on the C-terminus and made the appropriate truncations based on homology models with similar poor outcomes. The use of the maltose binding protein (MBP) tag, a common protein expression tag, did not improve solubility, even when screened with various purification buffers and expression conditions. Finally, an attempt with the pET22b expression system, which targets proteins of interest to the periplasmic region of the *E. coli* cell with the pelB signal sequence, produced no detectable overexpression of protein. Although disheartening, this screening process made me a better biochemist, and I am grateful for the skills and confidence gained. When it was clear *E. coli* was not a suitable host for this system, we moved to two other expression systems, *S. cerevisiae* and *S. coelicolor* CH999. Perhaps in another set of hands, expression in *E. coli* may be successful as many conditions did yield a small amount of insoluble protein which could be a starting point for future efforts.

4.10 Attempted Characterization of VHPOs in *Streptomyces coelicolor* CH999

Although initially promising, the expression results in *S. coelicolor* CH999 are puzzling. All efforts in this section were attempted with the most highly expressed

microbial-type VHPO from *Portieria hornemannii,* PhVHPO-3. After expression cultures reached an appropriate density, cells were harvested by centrifugation (8000 xg, 30 min), this cycle was repeated (~3 times) until media was mostly cleared. Cell pellets were resuspended in 25 mL of lysis buffer (20 mM Tris, 300 mM NaCl, 10 mM imidazole, 10% glycerol, pH 8) and stored at -80 ˚C for future purification. All frozen cell pellets were defrosted, and cells were lysed by sonication using a Qsonica 6 mm tip at 70% amplitude for 20 cycles of 15 seconds on and 45 seconds off, gently mixing after 15 cycles. Cell lysate was then centrifuged for 45 mins at 20,000 xg to pellet cell debris. Protein purification was performed on an ÄTKApurifier FPLC (GE Healthcare Life Sciences) at 4

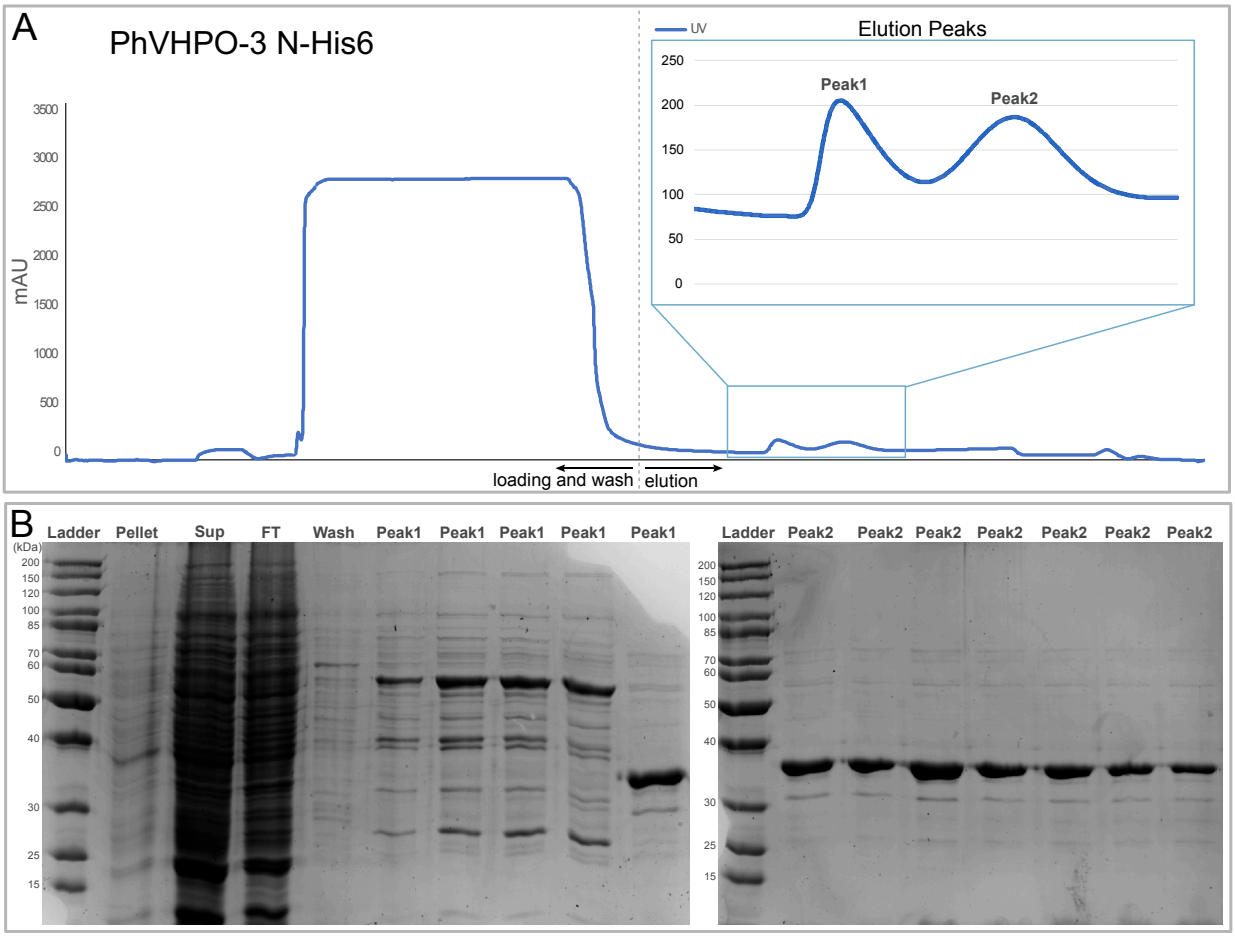


Figure 4.6 *Streptomyces coelicolor* CH999 expression results of PhVHPO-3. **A.** FPLC trace of Ni-NTA affinity chromatography run **B.** 10% SDS-PAGE gel of PhVHPO-3 purification.

°C. Proteins were purified by Ni^{2+}-affinity chromatography using 5 mL HisTrap FF (GE Healthcare) columns pre-equilibrated in wash buffer (500 mM NaCl, 20 mM Tris, 30 mM imidazole, 10 μM Na_3VO_4, pH 8.0). Lysate was then loaded onto the column at 1.5 mL/min after which the column was washed with 8 column volumes (40 mL) of wash buffer. Protein was eluted by a linear gradient of 0 – 100% elution buffer (500 mM NaCl, 20 mM Tris, 250 mM imidazole, 10 μM Na_3VO_4, pH 8.0) over 30 min at a flow rate of 2.0 mL/min. Protein-containing fractions were identified by SDS-PAGE gel (10% acrylamide) and then combined, concentrated using Amicon Ultra centrifugal filters with a 50 kDa molecular weight cut-off (EMD Millipore) and buffered exchanged using PD-10 Desalting Columns (Sephadex G-25 M, Cytvia) pre-equilibrated with storage buffer (50 mM HEPES, 250 mM NaCl, 10 μM Na_3VO_4,10% glycerol, pH 6 and pH 8). Concentrated and buffer exchanged samples were snap-frozen in liquid nitrogen and

stored immediately at -80 ˚C or used in enzymatic assays immediately.

PhVHPO-3 was found to be soluble, however, a second protein band consistently appeared, even after the addition of protease inhibitors (Complete EDTA-free, protease inhibitor cocktail tablets, Roche) (**Figure 4.7A**). Peak 1 was found to be the correct

Table 4.3 *Streptomyces coelicolor* CH999 *in vitro* PhVHPO-3 assay conditions.

Variable					
Vanadium in culture and/or assay	10 μM	25 μM	50 μM	100 μM	
Substrate co-solvent	DMSO	IPA			
pH	6	8			
H2O2	1 mM	2 mM	5 mM	10 mM	25 mM
Br/Cl equivalent	30 mM	300 mM			
Substrate concentration	1 μM	5 μM	10 μM	25 μM	A lot
Temperature	30˚C	RT			
Shaking	180 r.p.m				

molecular weight for PhVHPO-3 of 57.4 kDa. Both bands were tested for biochemical activity with standard MCD assay conditions and with the putative monoterpene substrates, β-myrcene and *trans*-β-ocimene. When enzyme assays were analyzed by GC-MS after 3 hrs and 12 hrs, there appeared to be no consumption of substrate or appearance of new peaks. Further optimization of assay conditions was attempted; however, no successful combination of conditions could be identified (**Table 4.3**).

4.11 Attempted Characterization of VHPOs in *Saccharomyces cerevisiae*

This system appears to be the most promising due to literature precedent for success with difficult heterologous protein expression in yeast expression platforms.[47] All efforts in this section were attempted with the co-clustered VHPO in the VHPO-TS gene pair described in Chapter 3, LsVHPO-1. LsVHPO-1 was heterologously expressed as a GFP fusion protein, and production of LsVHPO-1 was validated via Western Blot (**Figure 4.8A**) and GFP fluorescence was monitored in induced vs not induced cultures (**Figure 4.8B**). For *in vivo* assays, 10 mL expression cultures were grown overnight with the putative terpene substrate germacrene D-4-ol at varying concentrations (**Table 4.4**). After < 18 hrs, cells were harvested by centrifugation (5000 xg, 10 min) and the supernatant was quickly removed and saved. The cell pellet was resuspended in acetone and moved to a glass vial. Two volumes of hexanes, relative to the amount of acetone, were added and allowed to extract overnight with mixing. Similarly, the collected supernatant was extracted with 2:1 hexanes:acetone and allowed to extract overnight with mixing.

Figure 4.7 *Saccharomyces cerevisiae* expression of LsVHPO-1. **A.** Western blot of LsVHPO-1, from left to right: ladder, supernatant, and pellet. **B.** Fluorescence of GFP in induced and uninduced cultures.

Samples were analyzed by GC-MS, however, no discernable halogenation of the terpene substrate was detected. For crude lysate assays: After < 18 hrs, 10 mL expression cultures grown at 30 ˚C, were harvested by centrifugation (5000 xg, 10 min) in tared 50 mL falcon tube. Supernatant was removed and cell pellets were quickly resuspended according to wet mass in YeastBuster Protein Extraction Reagent (5 mL per 1 g) and TCEP solution (50 μL of 100 mM solution per 1 g). Resuspended pellets were shaken gently at RT for 15 – 20 mins and then bead beat with 0.5 mm glass beads for 45 s at 4.5 intensity x3 with 1 min on ice in between each round of bead beating. The resulting solutions were centrifuged (15,000 xg, 10 min) and both cell debris and supernatant were resuspended in either pH 6 or pH 8 storage buffer (50 mM HEPES, 250 mM NaCl, 10 μM Na_3VO_4,10% glycerol, pH 6 and pH 8) for immediate use in enzyme assay screening. Enzyme assays were analyzed by GC-MS; however, no discernable halogenation of the terpene substrate was detected.

Table 4.4 *Saccharomyces cerevisiae* assay conditions for both *in vivo* and crude lysate assays. The addition of 100 μM did appear to improve intensity of GFP fluorescence.

Variable					
Vanadium in culture and/or assay	10 μM	25 μM	50 μM	100 μM	
Substrate co-solvent	DMSO	IPA			
pH	6	8			
H2O2	1 mM	2 mM	5 mM	10 mM	25 mM
Br/Cl equivalent	30 mM	300 mM			
Substrate concentration	1 μM	5 μM	10 μM	25 μM	A lot
Temperature	30 ˚C	RT			
Shaking	180 r.p.m				

4.12 Discussion

Whole genome sequencing has revolutionized the field of natural products discovery and biosynthesis. Traditional methods of natural product discovery relied on the isolation of compounds from natural sources, followed by elucidation of their structures. Likewise, biosynthetic pathways could be characterized through isotopically labeled feeding studies of putative precursors, which often provided key insight into biosynthesis pathways and ‘order of operations questions’ but was synthetically labor intensive and limited by source organism requirements. However, with the advent of whole genome sequencing, we can identify and study entire biosynthetic pathways of natural products in a systematic and comprehensive manner.

Now, with substantial improvements in nucleotide isolation strategies and sequencing accuracy, read length, and throughput,[48] we can revisit these complex marine systems to connect biosynthetic genes to chemistry. This motto of genes to chemistry has been a huge driver of biosynthesis work in the Moore lab, as well as the work described in this dissertation. Red algal natural products, particularly haloterpenoids, are

arguably some of the simplest, yet structurally interesting compound in natural products research. These small terpene scaffolds are bursting with bromines and chlorines, and the underlying enzymatic machinery responsible for this overeager halogenation has been a source of fascination since their discovery. After the discovery of VHPOs from red algal systems, several studies have attempted to answer this longstanding question around haloterpenoid biosynthesis, yet little progress has been made to identify a dihalogenase capable of the long-sought bromochlorination reaction. The discovery of this reaction would not only be a huge leap towards understanding haloterpenoid biosynthesis, but it would also open doors to development of algal VHPOs as biocatalysts.

Biocatalysis is a major area of research sparked by the immense growth of sequencing data and rapid discovery of novel enzymes in biosynthesis research. By understanding the mechanisms of biosynthesis, researchers can engineer these enzymes to produce new chemical entities or to modify existing ones, leading to the production of high-value drug targets. Biocatalytic enzymes can be utilized to synthesize high-value drug targets by performing complex reactions in relatively mild reaction conditions, and once optimized can be more efficient, selective, and environmentally friendly than chemical synthesis. Algal VHPOs are incredibly exciting prospective biocatalysts due to their hypothesized role in haloterpenoid biosynthesis and possible dihalogenation capabilities. However, the biocatalytic potential of these enzymes is hindered by difficult heterologous expression and lack of clear biochemical data. Therefore, our goal is to optimize heterologous expression conditions towards soluble and active protein to support rapid biochemical characterization of microbial-like algal

VHPOs. This work would not only ensure an efficient and scalable route to high-value drug candidates like the antitumor molecule, halomon, but could be more broadly applied for halogenation chemistry in biocatalysis.

One intriguing application of a dihalogenase is in fluorination. In the hypothesized dihalogenation reaction, the first halogen atom acts as an electrophile and the second halogen atom acts as a nucleophile, something fluoride does exceptionally well (**Figure 4.9**). Fluorinated drugs are gaining increasing importance, and currently about 20% of all pharmaceuticals on the world market contain fluorine substituents.[21] Fluorination is well known to enhance bioavailability and increase receptor selectivity, key issues in pharmaceutical development.[49] Taking this into account, it is clear that fluorine incorporation could be a wildly useful biocatalytic reaction. On the basis of these conclusions, we hypothesis this newly discovered clade of red algal VHPOs could be capable of generating new-to-nature haloterpenoids as well as engineered to perform key pharmaceutically relevant fluorination reactions.

This chapter describes work towards the reconstitution of algal halogenases and ultimately towards the reconstitution of haloterpenoids pathways. Additionally, this

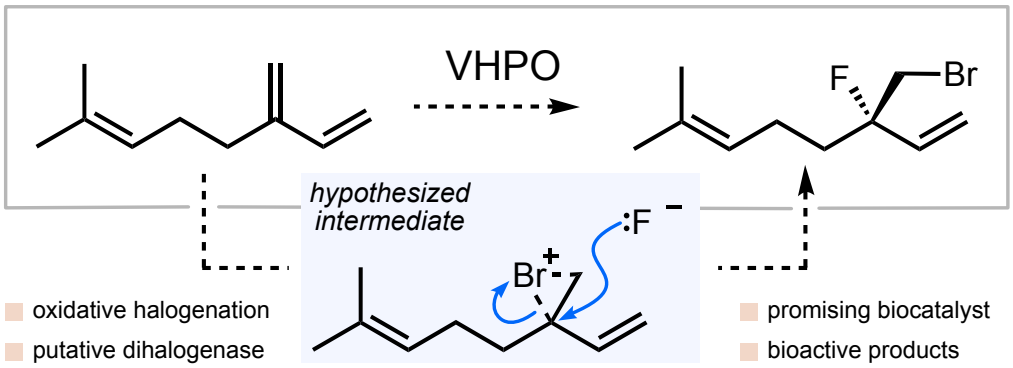


Figure 4.8 Proposed dihalogenation reaction to incorporate fluorine hypothesized could be catalyzed by the newly identified class of red algal vanadium-dependent haloperoxidases.

chapter describes recently optimized HMW DNA extraction protocols and genome mining strategies to support natural product biosynthesis research in red algal systems. Beyond the work described here, there are additional unexplored classes of red algal natural products. Both the families of halogenated fatty acids known as acetogenins[50] and halogenated triterpenoids like armatol[51] are red algal natural products with promising bioactivities and sourced from organisms with genomes reported for the first time in this dissertation. No biosynthetic genes are currently attributed to these systems, leaving their biosynthesis open to investigation. The ongoing work described in this chapter aims to lay the groundwork for future investigations of halogenated red algal natural products biosynthesis and biocatalysis.

4.13 References

1. W. H. Gerwick, B. S. Moore, Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. *Chem Biol* **19**, 85–98 (2012).

2. P. J. Scheuer, Marine natural products research: A look into the dive bag. *J. Nat. Prod.* **58**, 335–343 (1995).

3. R. W. Fuller, J. H. Cardellina, Y. Kato, M. R. Boyd, L. S. Brinen, J. Clardy, K. M. Snader, A pentahalogenated monoterpene from the red alga *Portieria hornemannii* produces a novel cytotoxicity profile against a diverse panel of human tumor cell lines. *J. Med. Chem.* **35**, 3007–3011 (1992).

4. D. Uemura, K. Takahashi, T. Yamamoto, C. Katayama, J. Tanaka, Y. Okumura, Y. Hirata, Norhalichondrin A: an antitumor polyether macrolide from a marine sponge. *J. Am. Chem. Soc.* **107**, 4796–4798 (1985).

5. W. Fenical, R. K. Okuda, M. M. Bandurraga, P. Culver, R. S. Jacobs, Lophotoxin: a novel neuromuscular toxin from Pacific Sea whips of the genus *Lophogorgia*. *Science* **212**, 1512–1514 (1981).

6. R. W. Fuller, J. H. Cardellina, J. Jurek, P. J. Scheuer, B. Alvarado-Lindner, M. Mcguire, G. N. Gray, J. Rios Steiner, J. Clardy, E. Menez, R. H. Shoemaker, D. J. Newman, K. M. Snader, M. R. Boyd, Isolation and structure/activity features of halomon-related antitumor monoterpenes from the red alga *Portieria homemannii J. Med. Chem.* **37**, 4407–4411 (1994).

7. C. Bucher, R. M. Deans, N. Z. Burns, Highly selective synthesis of halomon, plocamenone, and isoplocamenone. *J. Am. Chem. Soc.* **137**, 12784–7 (2015).

8. M. H. Medema, T. de Rond, B. S. Moore, Mining genomes to illuminate the specialized chemistry of life. *Nat. Rev. Genet.* **22**, 553–571 (2021).

9. K. D. Bauman, K. S. Butler, B. S. Moore, J. R. Chekan, Genome mining methods to discover bioactive natural products. *Nat. Prod. Rep.* **38**, 2100–2129 (2021).

10. A. Abreu, E. Bourgois, A. Gristwood, R. Troublé, S. G. Acinas, P. Bork, E. Boss, C. Bowler, M. Budinich, S. Chaffron, C. de Vargas, T. O. Delmont, D. Eveillard, L. Guidi, D. Iudicone, S. Kandels, H. Morlon, F. Lombard, R. Pepperkok, J. J. P. Karlusich, G. Piganeau, A. Régimbeau, G. Sommeria-Klein, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, O. Zablocki, D. Arendt, J. Bilic, R. Finn, E. Heard, B. Rouse, J. Vamathevan, R. Casotti, I. Cancio, M. Cunliffe, A. E. Kervella, W. H. C. F. Kooistra, M. Obst, N. Pade, D. M. Power, I. Santi, T. M. Tsagaraki, J. Vanaverbeke, Tara Ocean Foundation, Tara Oceans, European Molecular Biology Laboratory (EMBL), European Marine Biological Resource Centre - European Research Infrastructure Consortium (EMBRC-ERIC), Priorities for ocean microbiome research. *Nat. Microbiol.* **7**, 937–947 (2022).

11. J. Collén, B. Porcel, W. Carré, S. G. Ball, C. Chaparro, T. Tonon, T. Barbeyron, G. Michel, B. Noel, K. Valentin, M. Elias, F. Artiguenave, A. Arun, J.-M. Aury, J. F. Barbosa-Neto, J. H. Bothwell, F.-Y. Bouget, L. Brillet, F. Cabello-Hurtado, S. Capella-Gutiérrez, B. Charrier, L. Cladière, J. M. Cock, S. M. Coelho, C. Colleoni, M. Czjzek, C. D. Silva, L. Delage, F. Denoeud, P. Deschamps, S. M. Dittami, T. Gabaldón, C. M. M. Gachon, A. Groisillier, C. Hervé, K. Jabbari, M. Katinka, B. Kloareg, N. Kowalczyk, K. Labadie, C. Leblanc, P. J. Lopez, D. H. McLachlan, L. Meslet-Cladiere, A. Moustafa, Z. Nehr, P. N. Collén, O. Panaud, F. Partensky, J. Poulain, S. A. Rensing, S. Rousvoal, G. Samson, A. Symeonidi, J. Weissenbach, A. Zambounis, P. Wincker, C. Boyen, Genome structure and metabolic features in the red seaweed Chondrus crispus shed light on evolution of the Archaeplastida. *Proc. Natl. Acad. Sci. U.S.A*. **110**, 5247–5252 (2013).

12. S. H. Brawley, N. A. Blouin, E. Ficko-Blean, G. L. Wheeler, M. Lohr, H. V. Goodson, J. W. Jenkins, C. E. Blaby-Haas, K. E. Helliwell, C. X. Chan, T. N. Marriage, D. Bhattacharya, A. S. Klein, Y. Badis, J. Brodie, Y. Cao, J. Collén, S. M. Dittami, C. M. M. Gachon, B. R. Green, S. J. Karpowicz, J. W. Kim, U. J. Kudahl, S. Lin, G. Michel, M. Mittag, B. J. S. C. Olson, J. L. Pangilinan, Y. Peng, H. Qiu, S. Shu, J. T. Singer, A. G. Smith, B. N. Sprecher, V. Wagner, W. Wang, Z.-Y. Wang, J. Yan, C. Yarish, S. Zäuner-Riek, Y. Zhuang, Y. Zou, E. A. Lindquist, J. Grimwood, K. W. Barry, D. S. Rokhsar, J. Schmutz, J. W. Stiller, A. R. Grossman, S. E. Prochnik, Insights into the red algae and eukaryotic evolution from the genome of Porphyra umbilicalis (Bangiophyceae, Rhodophyta). *Proc. Natl. Acad. Sci. U.S.A*. **114**, E6361–E6370 (2017).

13. A. Collens, E. Kelley, L. A. Katz, The concept of the hologenome, an epigenetic phenomenon, challenges aspects of the modern evolutionary synthesis. *J. Exp. Zool. B. Mol. Dev. Evol.* **332**, 349–355 (2019).

14. J. R. Chekan, S. M. K. McKinnie, M. L. Moore, S. G. Poplawski, T. P. Michael, B. S. Moore, Scalable biosynthesis of the seaweed neurochemical, kainic acid. *Angew. Chem., Int. Ed.* **58**, 8454–8457 (2019).

15. T. S. Steele, J. K. Brunson, Y. Maeno, R. Terada, A. E. Allen, M. Yotsu-Yamashita, J. R. Chekan, B. S. Moore, Domoic acid biosynthesis in the red alga *Chondria armata* suggests a complex evolutionary history for toxin production. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2117407119 (2022).

16. P. D. Scesa, Z. Lin, E. W. Schmidt, Ancient defensive terpene biosynthetic gene clusters in the soft corals. *Nat. Chem. Biol.* **18**, 659–663 (2022).

17. I. Burkhardt, T. de Rond, P. Y.-T. Chen, B. S. Moore, Ancient plant-like terpene biosynthesis in corals. *Nat. Chem. Biol.* **18**, 664–669 (2022).

18. Wilson, K.A., de Rond, T., Burkhardt, I., Steele, T.S., Schäfer, R., Podell, S., Allen, E.E., Moore, B.S. Terpene biosynthesis in marine sponge animals. *Proc. Natl. Acad. Sci. U.S.A*. **120,** (9) e2220934120 (2023).

19. H. R. Thapa, Z. Lin, D. Yi, J. E. Smith, E. W. Schmidt, V. Agarwal, Genetic and biochemical reconstitution of bromoform biosynthesis in *Asparagopsis* lends insights into seaweed reactive oxygen species enzymology. *ACS Chem. Biol.* **15**, 1662–1670 (2020).

20. T. P. Michael, F. Jupe, F. Bemm, S. T. Motley, J. P. Sandoval, C. Lanz, O. Loudet, D. Weigel, J. R. Ecker, High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nat. Commun.* **9**, 541 (2018).

21. J. Wang, M. Sánchez-Roselló, J. L. Aceña, C. del Pozo, A. E. Sorochinsky, S. Fustero, V. A. Soloshonok, H. Liu, Fluorine in pharmaceutical industry: fluorine-containing drugs introduced to the market in the last decade (2001-2011). *Chem. Rev.*, **114***,* 2432–2506 (2014).

22. G. W. Saunders, T. E. Moore, Refinements for the amplification and sequencing of red algal DNA barcode and RedToL phylogenetic markers: a summary of current primers, profiles and strategies. *Algae* **28**, 31–43 (2013).

23. H. Kylin, The Marine Red Algae in the Vicinity of the Biological Station at Friday Harbor*,* Washington, *C.W.K. Gleerup*, (1925).

24. G. W. Saunders, Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **360**, 1879–1888 (2005).

25. Nanopore, Arabidopsis Leaf gDNA. https://community.nanoporetech.com/extraction_method_groups/plant-leaf-gDNA

26. H.-W. Nützmann, A. Huang, A. Osbourn, Plant metabolic clusters – from genetics to genomics. *New Phytol.* **211**, 771–789 (2016).

27. T. J. Carrier, A. M. Reitzel, The hologenome across environments and the implications of a host-associated microbial repertoire. *Front. Microbiol.* **8**, 802 (2017).

28. P. D. Sonawane, U. Heinig, S. Panda, N. S. Gilboa, M. Yona, S. P. Kumar, N. Alkan, T. Unger, S. Bocobza, M. Pliner, S. Malitsky, M. Tkachev, S. Meir, I. Rogachev, A. Aharoni, Short-chain dehydrogenase/reductase governs steroidal specialized metabolites structural diversity and toxicity in the genus Solanum. *Proc. Natl. Acad. Sci. U.S.A*. **115**, E5419–E5428 (2018).

29. M. L. Landry, N. Z. Burns, Catalytic enantioselective dihalogenation in total synthesis. *Acc. Chem. Res.* **51**, 1260–1271 (2018).

30. A. J. Cresswell, S. T.-C. Eey, S. E. Denmark, Catalytic, stereospecific syn-dichlorination of alkenes. *Nature Chem.* **7**, 146–152 (2015).

31. B. B. Gilbert, S. T.-C. Eey, P. Ryabchuk, O. Garry, S. E. Denmark, Organoselenium-catalyzed enantioselective syn-dichlorination of unbiased alkenes. *Tetrahedron* **75**, 4086–4098 (2019).

32. B. J. Burreson, F. X. Woolard, R. E. Moore, Evidence for the biogenesis of halogenated myrcenes from the red alga *Chondrococcus hornemanni*. *Chem. Lett.* **4**, 1111–1114 (1975).

33. M. D. Higgs, D. J. Vanderah, D. J. Faulkner, Polyhalogenated monoterpenes from *Plocamium cartilagineum* from the British coast. *Tetrahedron* **33**, 2775–2780 (1977).

34. V. Agarwal, Z. D. Miles, J. M. Winter, A. S. Eustáquio, A. A. El Gamal, B. S. Moore, Enzymatic halogenation and dehalogenation reactions: Pervasive and mechanistically diverse. *Chem. Rev.* **117**, 5619–5674 (2017).

35. E. de Boer, H. Plat, M. G. M. Tromp, R. Wever, M. C. R. Franssen, H. C. van der Plas, E. M. Meijer, H. E. Schoemaker, Vanadium containing bromoperoxidase: An example of an oxidoreductase with high operational stability in aqueous and organic media. *Biotechnol. Bioeng.* **30**, 607–610 (1987).

36. J. Latham, E. Brandenburger, S. A. Shepherd, B. R. K. Menon, J. Micklefield, Development of halogenase enzymes for use in synthesis. *Chem. Rev.* **118**, 232–269 (2018).

37. P. Bernhardt, T. Okino, J. M. Winter, A. Miyanaga, B. S. Moore, A Stereoselective vanadium-dependent chloroperoxidase in bacterial antibiotic biosynthesis. *J. Am. Chem. Soc.* **133**, 4268–4270 (2011).

38. P. Y.-T. Chen, S. Adak, J. R. Chekan, D. K. Liscombe, A. Miyanaga, P. Bernhardt, S. Diethelm, E. N. Fielding, J. H. George, Z. D. Miles, L. A. M. Murray, T. S. Steele, J. M. Winter, J. P. Noel, B. S. Moore, Structural basis of stereospecific vanadium-dependent haloperoxidase family enzymes in napyradiomycin biosynthesis. *Biochemistry* **61**, 1844–1852 (2022).

39. Z. D. Miles, S. Diethelm, H. P. Pepper, D. M. Huang, J. H. George, B. S. Moore, A unifying paradigm for naphthoquinone-based meroterpenoid (bio)synthesis. *Nat. Chem.* **9**, 1235–1242 (2017).

40. A. Butler, J. N. Carter-Franklin, The role of vanadium bromoperoxidase in the biosynthesis of halogenated marine natural products. *Nat. Prod. Rep.* **21**, 180–188 (2004).

41. J. M. Winter, B. S. Moore, Exploring the chemistry and biology of vanadium-dependent haloperoxidases. *J. Biol. Chem.* **284**, 18577–18581 (2009).

42. S. M. K. McKinnie, Z. D. Miles, B. S. Moore, Characterization and biochemical assays of streptomyces vanadium-dependent chloroperoxidases. *Meth. Enzymol*. **604**, 405–424 (2018).

43. J. J. Almagro Armenteros, K. D. Tsirigos, C. K. Sønderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne, H. Nielsen, SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).

44. K. D. Bauman, V. V. Shende, P. Y.-T. Chen, D. B. B. Trivella, T. A. M. Gulder, S. Vellalath, D. Romo, B. S. Moore, Enzymatic assembly of the salinosporamide γ-lactam-β-lactone anticancer warhead. *Nat. Chem. Biol.* **18**, 538–546 (2022).

45. YPD media. *Cold Spring Harb. Protoc*. pdb.rec12315 (2010).

46. K. C. Chou, H. B. Shen, Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PLOS ONE*. **5**, 6 (2010)

47. R. C. Boswell-Casteel, J. M. Johnson, R. M. Stroud, F. A. Hays, Integral membrane protein expression in *Saccharomyces cerevisiae, Methods Bol. Biol* **1432**, 163–186 (2016).

48. Y. Wang, Y. Zhao, A. Bollas, Y. Wang, K. F. Au, Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).

49. C. Weiß, T. Bogner, B. Sammet, N. Sewald, Total synthesis and biological evaluation of fluorinated cryptophycins. *Beilstein J. Org. Chem.* **8**, 2060–2066 (2012).

50. G. M. König, A. D. Wright, New C15 acetogenins and sesquiterpenes from the red alga *Laurencia gracilis. J. Nat. Prod.* **57**, 477–485 (1994).

51. B. S. Underwood, J. Tanuwidjaja, S.-S. Ng, T. F. Jamison, Total syntheses of the squalene-derived halogenated polyethers ent-dioxepandehydrothyrsiferol and armatol A via bromonium- and Lewis acid-initiated epoxide-opening cascades. *Tetrahedron* **69**, 5205–5220 (2013).

4.14 Acknowledgements

Chapter 4 contains unpublished material coauthored with Moore, M., Burkhardt, I., and Moore, B.S. The dissertation author was the primary researcher and author of this chapter.

CHAPTER 5. Perspectives and outlook

Human evolution throughout history has been shaped by our ability to harness the chemistry of the natural world to adapt and thrive. One of the most well-known examples of these genetically encoded small molecules, or natural products, is the early use of willow bark to treat pain. The first written evidence comes from an Egyptian scroll, which details the use of *tjeret* or *salix* (willow) for the treatment of non-specific pains.[1] We now know willow contains the bioactive analgesic natural product, salicin.[2] In the 1800s, salicin was chemically modified to develop the more potent and less irritating natural product derivative, aspirin. Aspirin is now one of the most commonly used drugs in the world and has proved lifesaving in the prevention of cardiovascular disease and pain management.[1] As of 2020, almost two-thirds of approved therapeutic agents are natural products or natural product derivatives[3] and to-date, 24 marine natural products have been approved to treat various cancers, hypertriglyceridemia, chronic pain, and viral infections.[4]

These ‘natural products’ are generally derived from an organism’s secondary metabolism, which is defined as dynamic, undergoing selection to offer a competitive evolutionary advantage, and may have a profound effect on survival.[5] I want to revisit this definition of secondary metabolism, because while stories like the above discussion of aspirin are compelling and set precedence for funding and further discovery work, organic molecules in the environment also serve innumerable ecological roles. For example, many natural products mediate inter- and intra-species communication by chemical means or improve the fitness of an organism by acting as a feeding deterrent, and simply render an organism unpalatable. Advances in genomic sequencing methods have revolutionized the way we study natural product biosynthesis, as well as how we

understand evolutionary patterns and drivers of biosynthesis gene diversification. In this dissertation, I examine how marine red algae evolve and maintain biosynthesis genes and apply long-read sequencing to explore questions centered around gene clustering in eukaryotes.

One of the largest and most diverse families of natural products are the terpenoids, with over 100,000 characterized compounds to-date. In the chemical ecology of many organisms, terpenoid natural products play a key role in defense and communication. In Chapter 1 of this dissertation, we explore the question:

Why are specialized terpenoid natural products so pervasive throughout all branches of the tree of life?

Here, we argue life is uniquely primed to evolve and maintain specialized terpenoid biosynthesis. This is in part due to the widespread prevalence of the underlying primary metabolic machinery. Therefore, we propose there is a relatively low evolutionary barrier to evolving new secondary terpenoid biosynthesis pathways – 'just add a terpene synthase'. With examples of terpene synthase evolution through both horizontal gene transfer[6,7] and via gene duplication and neofunctionalization.[8] Next, because an abundance of organisms already maintain receptors for terpene natural products, terpenoids could be considered a universal chemical language throughout nature.[9–11] Similarly, recent research suggests that nuclear receptors, which are involved in reproduction and development across all metazoan groups, evolved from ancient terpene synthases and retain the same ligand-binding domain core of an ancestral terpene synthase.[12] Finally, terpene synthase evolution is incredibly dynamic. Terpene synthases that make the same product can have widely divergent amino acid sequences,

confounding many standard genome mining efforts. This requires homology-independent search tools like Hidden Markov Models and broad reference terpene synthase databases. As such, there is an ongoing need to expand publicly available sequencing datasets.

In Chapter 2 of this dissertation, I use a sequencing forward approach to identify the biosynthesis genes responsible for domoic acid production in the red macroalga, *Chondria armata*. Here, we ask the key question:

Is domoic acid biosynthesis the same across two distinct algal lineages: red macroalgae and diatoms?

Until the publication of this chapter, there remained a gap in knowledge of domoic acid producing red algae, as only kainic acid producing red algae had been studied to-date. Upon sequencing *C. armata,* we quickly identified a domoic acid gene cluster using the terpene synthase-like gene *dabA* as a genetic hook. This cluster retains the three-gene structure (A – terpene synthase-like, C – kainoid synthase, and D – CYP450) observed in previous diatom domoic acid gene clusters. However, an all-versus-all comparison revealed an unexpected finding.

One of the first observations made when comparing this new set of clusters to previous clusters was the uneven sequence homology across the gene cluster. Irrespective of taxonomy, there was a remarkable level of sequence similarity between the 'A' and 'C' genes across different organisms. Conversely, the 'D' genes exhibited relatively lower homology between diatoms and red macroalgae. This result invokes two modes of evolution toward domoic acid biosynthesis: HGT combined with gene duplication and neofunctionalization of native CYP450 biochemistry. Based on these

results, I propose that there is a founder core gene-pair (A & C genes) accompanied by a more recently evolved CYP450 (D gene), which is one of the most common tailoring enzymes in terpenoid biosynthesis. Formation of new metabolic clusters from a founder gene pair template is common in plant biosynthesis.[13,14] In this scenario, an ancestral gene pair may give rise to new metabolic clusters as a result of independent events involving gene rearrangement and recruitment of new genes. Previous investigations of kainoid biosynthesis primarily concentrated on the lower isoprene kainoid homolog, kainic acid, which lacks an oxidized side-chain. Consequently, the analysis of CYP450s had primarily focused on diatoms, and we were not aware of the dynamic nature of the evolution of domoic acid associated CYP450s. CYP450s represent one of the largest enzyme families and are well understood to play a significant role in the structural diversification of terpene scaffolds. Interestingly, the kainoid clusters in all organisms were significantly enriched in transposable elements. Thus, transposable element-mediated recombination may contribute to cluster formation in both diatoms and red macroalgae.

Current work on kainoid biosynthesis centers on the uncharacterized isomerization transformation. Notably, the 1,3-olefin isomerization step required to make domoic acid continues to remain elusive. With the support of additional sequencing data, we now aim to clarify this outstanding question in domoic acid biosynthesis. Ongoing investigations of candidate genes present in these two distantly related species, including further study of the activities exhibited by the co-clustered CYP450 enzymes DabD and hypothetically assigned RadD, will shed additional light on this remaining piece of the DA biosynthetic puzzle.

Chapters 3 and 4 of this dissertation are a two-part exploration of halogenated terpene (haloterpenoid) biosynthesis in red macroalgae. Here we Identify the terpene synthases responsible for producing key terpene backbones and begin to explore downstream tailoring steps in haloterpenoid biosynthesis and ask the key question:

How are algal haloterpenoids biosynthesized in a regio-, chemo-, and enatio-selective manner?

The haloterpenoid projects are ones I was most excited to explore with my research. The depth of supporting haloterpenoid chemical literature is daunting and these projects are built on the shoulders of marine natural products giants like Professors William Fenical (SIO), D. John Faulkner (SIO), Richard E. Moore (Univ. of Hawaii), and Alison Butler (UCSB). Their early research is foundational to our understanding of seaweed chemistry, and I am humbled to continue this work. My main goal in continuing this work was to apply new long-read sequencing technologies to uncover the genetic basis of haloterpenoid biosynthesis in red macroalgae and unravel haloterpenoid biosynthesis gene clustering questions. Early sequencing efforts utilized a combination of Oxford Nanopore Technologies (ONT) and Illumina datasets. Later efforts, in collaboration with the Joint Genome Institute, Department of Energy (CSP-504335), also included Pacific Biosciences (PacBio) datasets. If possible, I would encourage future projects to include other sequencing experiments, like Hi-C and Bionano genome imaging to scaffold and improve the assemblies reported here. I targeted three local red algae, *Plocamium pacificum, Laurencia subopposita,* and *Laurencia pacifica*, as well as *Portieria hornemannii*, the notable producer of the haloterpenoid antitumor agent, halomon.

In Chapter 3, I identified and biochemically characterized ten new red algal terpene synthases, four of which produce proposed on-pathway precursors to target haloterpenoids. When I first identified these sequences, I was initially skeptical of their taxonomic origin. At the time, there was sparse literature data to help validate our sequences as *bona fide* red algal terpene synthases and not from a contaminate. However, the serendipitous publication of work from the Feng Chen lab supported our growing suspicion terpene synthases in red algae are phylogenetically distinct from typical plant terpene synthases and most likely originated in red macroalgae via horizontal gene transfer from an unknown microbial source. Our work expands the red algal terpene synthase clade and continues to demonstrate a clear delineation between algal and microbial genes. One of the noticeable gaps in Chapter 3 is the lack of a bisabolene synthase in *Laurencia pacifica*. We postulate bisabolene is the main terpene scaffold of the chamigrane family of haloterpenoids in the genus *Laurencia*. While we did find a bisabolene synthase in *Laurencia subopposita*, this species of *Laurencia* is not a major producer of chamigranes. Algal chamigranes are characterized by a densely functionalized spirocyclic core and would be an excellent target for future biosynthetic endeavors. Additionally, there are several other unaccounted terpenoid scaffolds from the genus *Laurencia*, primarily the laurene-type sesquiterpenoids. We hypothesize the new UbiA-containing gene clusters reported in Chapter 4 from *L. pacifica* and *L. subopposita* could be associated with these molecules and are the subject of ongoing research in the Moore lab.

In Chapters 3 and (primarily) 4, I describe attempts to elucidate downstream tailoring steps in haloterpenoid biosynthesis. Although there are still many outstanding

questions around the role of vanadium dependent halogenases in haloterpenoid biosynthesis, one of the main successes from this work was implementing long-read sequencing and developing co-occurrence networks to identify the first putative haloterpenoid associated gene clusters in red macroalgae. Beyond what I presented in this work – gene co-occurrence networks could be a way to assess gene clustering in other eukaryotic systems and a clear future direction for this work is to improve and automate this pipeline. To quote my co-author Malia Moore, ‘genomes are something we all benefit from’, and the algal genome assemblies and gene clusters presented in Chapters 3 and 4 are the first steps toward answering long-standing questions surrounding the biosynthesis of algal haloterpenoids.

As I consider the broader impacts of the work presented in this dissertation, I want to share a quote from Bill Fenical from the Natural Prodcast, “I never thought that I could combine a training in chemistry with a hobby of putting your head underwater”. What a wonderful, interdisciplinary field. Marine natural products has its origins in marine toxins, with work from Professors Paul Scheuer and Dick Moore, and like any scientific field, is fueled by technological advancements. One of the early examples of the impacts of technology on marine natural products research is the invention of the first successful safe open-circuit SCUBA system by Jacques Cousteau and Émile Gagnan, which expanded our capacity for diving and specimen collection.[15] This early work aimed to get a sense of ‘what was there', focusing on isolation and structure elucidation, systematically cataloging the overwhelmingly distinct chemistry of marine organisms. Natural products research next shifted to study the impacts of those molecules on fitness, and their role in the chemical ecology of an organism. As it became apparent the molecules that were

being isolated were extremely bioactive, research questions expanded to include drug discovery. This transition shifted marine natural products research to focus on microbial systems, which could be cultured at scale, a strategy to overcome source organism supply issues that limited previous efforts to advance molecules from eukaryotic systems. However, now that natural products biosynthesis is well supported by advancements in sequencing and computing technologies, we can assemble the biological blueprints for promising drug targets for both prokaryotic and eukaryotic systems. This growth in sequencing capabilities, along with growing interest in biocatalysis to complement synthetic methods,[16] has fundamentally transformed natural products biosynthesis research. Now, we can develop biotransformation processes to access promising drug targets in a sustainable, source-independent manner. Biosynthesis research, hastened by a growing need to depart from our dependence on fossil fuels, is an avenue towards a more sustainable future. As Sir Alister Hardy wrote in his 1965 book, The Open Sea: Its Natural History, “All the discoveries of the marine biologists, however remote they may now seem from practical issues, will be required in their different ways for a full understanding of the life in the sea and its more rational exploitation.”

1. M. J. R. Desborough, D. M. Keeling, The aspirin story – from willow to wonder drug. *Br. J. Haematol.* **177**, 674–683 (2017).

2. P. E. Schindler, Aspirin therapy *Biblio.com* (1978).

3. D. J. Newman, G. M. Cragg, natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770–803 (2020).

4. Oceans and Human Health - 2nd Edition (May 15, 2023).

5. M. G. Chevrette, K. Gutiérrez-García, N. Selem-Mojica, C. Aguilar-Martínez, A. Yañez-Olvera, H. E. Ramos-Aboites, P. A. Hoskisson, F. Barona-Gómez, Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat. Prod. Rep.* **37**, 566–599 (2020).

6. I. Burkhardt, T. de Rond, P. Y.-T. Chen, B. S. Moore, Ancient plant-like terpene biosynthesis in corals. *Nat. Chem. Biol.* **18**, 664–669 (2022).

7. K. Wilson, T. de Rond, I. Burkhardt, T. S. Steele, R. J. B. Schäfer, S. Podell, E. E. Allen, B. S. Moore, Terpene biosynthesis in marine sponge animals. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2220934120 (2023).

8. F. Beran, P. Rahfeld, K. Luck, R. Nagel, H. Vogel, N. Wielsch, S. Irmisch, S. Ramasamy, J. Gershenzon, D. G. Heckel, T. G. Köllner, Novel family of terpene synthases evolved from trans-isoprenyl diphosphate synthases in a flea beetle. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2922–2927 (2016).

9. R. Kramer, W. R. Abraham, Volatile sesquiterpenes from fungi: what are they good for? *Phytochem Rev.* **11**, 15–37 (2012).

10. A. Das, S. H. Lee, T. K. Hyun, S. W. Kim, J. Y. Kim, Plant volatiles as method of communication. *Plant Biotechnol. Rep.* **7**, 9–26 (2013).

11. K. Schulz-Bohm, L. Martín-Sánchez, P. Garbeva, Microbial volatiles: Small molecules with an important role in intra- and inter-kingdom interactions. *Front. Microbiol.* **8**, 2484 (2017).

12. D. R. Houston, J. G. Hanna, J. C. Lathe, S. G. Hillier, R. Lathe, Evidence that nuclear receptors are related to terpene synthases. *J. Mol. Endocrinol.* **68**, 153–166 (2022).

13. H.-W. Nützmann, A. Huang, A. Osbourn, Plant metabolic clusters – from genetics to genomics. *New Phytol.* **211**, 771–789 (2016).

14. R. Bharadwaj, S. R. Kumar, A. Sharma, R. Sathishkumar, Plant metabolic gene clusters: Evolution, organization, and their applications in synthetic biology. *Front. Plant Sci.* **12** (2021).

15. J.Y. Cousteau, F. Dumas, The silent world. *MW Books Ltd.* 5th Edition (1953)

16. E. L. Bell, W. Finnigan, S. P. France, A. P. Green, M. A. Hayes, L. J. Hepworth, S. L. Lovelock, H. Niikura, S. Osuna, E. Romero, K. S. Ryan, N. J. Turner, S. L. Flitsch, Biocatalysis. *Nat. Rev. Methods Primers* **1**, 1–21 (2021).

17. P. V. Stefanoudis, W. Y. Licuanan, T. H. Morrison, S. Talma, J. Veitayaki, L. C. Woodall, Turning the tide of parachute science. *Curr.* **31**, R184–R185 (2021).