Rockefeller University
Digital Commons @ RU

Student Theses and Dissertations

2011

Natural Product Biosynthesis In Uncultured Bacteria

Jeffrey Kim

Follow this and additional works at: http://digitalcommons.rockefeller.edu/student_theses_and_dissertations

 Part of the Life Sciences Commons

Recommended Citation
Kim, Jeffrey, "Natural Product Biosynthesis In Uncultured Bacteria" (2011). *Student Theses and Dissertations.* Paper 94.

This Thesis is brought to you for free and open access by Digital Commons @ RU. It has been accepted for inclusion in Student Theses and Dissertations by an authorized administrator of Digital Commons @ RU. For more information, please contact mcsweej@mail.rockefeller.edu.

NATURAL PRODUCT BIOSYNTHESIS IN UNCULTURED BACTERIA

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Jeffrey Kim

June 2011

© Copyright by Jeffrey Kim 2011

NATURAL PRODUCT BIOSYNTHESIS IN UNCULTURED BACTERIA

Jeffrey Kim, Ph.D.

The Rockefeller University 2011

A single gram of soil can contain thousands of unique bacterial species, only a small fraction of which is regularly cultured in the laboratory. Although the fermentation of cultured microorganisms has provided access to numerous bioactive secondary metabolites, with these same methods it is not possible to characterize the natural products encoded by the uncultured majority. The heterologous expression of biosynthetic gene clusters cloned from DNA extracted directly from environmental samples (eDNA) has begun to provide access to the chemical diversity encoded in the genomes of previously uncultured bacteria. The systematic exploration of natural product biosynthesis in uncultured bacteria, however, still faces several challenges that we sought to experimentally address. First, many natural product gene clusters cannot be detected in functional screens due to cloning and expression limitations of metagenomic library host strains. Second, the lack of robust and scalable gene cluster assembly methods precludes the functional characterization of a large number of natural product biosynthetic gene clusters from cosmid-based eDNA libraries. Third, the large-scale analysis of metagenomic natural product chemical diversity and the

phylogenetic context it is found in were previously unaddressed due to the complexities of microbial communities. To address these questions, we have:

1) Demonstrated that sequence-based screens can be used to systematically discover a diverse range of natural product gene clusters by screening two of the largest recombinant eDNA libraries reported to date. This approach circumvents many of the challenges of using functional screens to discover novel biosynthetic gene clusters. (Chapter 2)
2) Shown that transformation associated recombination in *S. cerevisiae* can be used to functionally reassemble large natural product gene clusters that exceed conventional eDNA cloning limits. This approach overcomes a significant barrier which prevented the functional characterization of many natural product gene clusters from eDNA libraries. (Chapter 3)
3) Developed a high throughput sequencing analysis framework to characterize environmental biosynthetic capacity. These results suggest that the continued construction and screening of soil-based eDNA libraries should provide access to additional novel pools of biosynthetic enzyme diversity. (Chapter 4)

This thesis is dedicated to

My family, friends, and colleagues.

In loving memory of William Morgan, Walter Carpenter and Tristan Campbell

ACKNOWLEDGMENTS

The majority of my childhood was spent trying to figure out how things worked by wantonly disassembling household appliances. This was an endless source of entertainment as a child but one that came at the risk of driving any normal parents mad. Sadly, I think this still provides a good description of my *modus operandi*, despite several decades of "maturation." I therefore want to thank my family, friends, and mentors for being understanding and patient with me. My early brushes with research were captivating and taught me the value of channeling my curiosity toward science. I still fondly remember my first day as a research assistant in Dr. Tarun Kapoor's laboratory. I wrestled with the FPLC in the cold room for hours and I was underdressed, soaked from a summer shower that morning, and shivering like a leaf. I could not have been happier. Since then, I have learned something new every day, and having the opportunity to work with Dr. Tarun Kapoor and Dr. Sean Brady was one of the most rewarding experiences of my life. I have had the fortune of working with talented, driven and innovative people during my studies and extracurricular pursuits. Dr. Kapoor and Dr. Brady combine these qualities with an enthusiasm for discovery that is unmatched and awe-inspiring. They introduced me to my lifelong passion and I cannot express how grateful I am for having had the opportunity to learn from them. I hope to make myself a better scientist and

person by following their example and by pursuing my future goals with equal intensity and creativity. I would like to thank Dr. Tom Muir, Dr. Sid Strickland, and Dr. Emily Harms for offering scientific and personal advice. I deeply appreciate the efforts made by my thesis committee to foster my scientific development. None of these explorations would have been possible without their guidance. I would like to thank Dr. Lars Dietrich for graciously serving on my committee and reviewing this thesis. I would like to thank the following people for being incredibly supportive during my studies: Drs. Howard Hang, Matthew Pratt, Benjamin Kwok, Ulf Peters, Jedidiah Gaetz, Sebastian Jayaraj, Stephen Lory, Christopher Paddon, Sarah Reisinger, Timothy Geistlinger, and Zach Serber. I would also like to thank Kristy Schopper and my LIC family, Bradley Kemp, my sisters Joan and Jeannie Kim, William Morgan, Herbert Moore, Christopher Hwang, Gary Leggett, and all of the members of the Brady Lab. I would not be here without you. Finally, I would like to thank my parents for teaching me that the most important and rewarding things in life are the people around us.

Table of Contents

Table of Contents ..... **vi**

List of Figures ..... **viii**

List of Tables ..... **xi**

List of Equations ..... **xii**

List of Abbreviations ..... **xiii**

CHAPTER 1 ..... **1**

1 Introduction and Background **1**

1.1 Natural Products and their Uses **1**

1.1.1 Natural Product Biosynthesis 4

1.1.2 Natural Products and Uncultured Bacteria 9

1.2 Metagenomic Strategies to Natural Product Discovery **13**

1.2.1 Environmental Culturing and Sequencing Efforts 13

1.2.2 Environmental Sample Selection 15

1.2.3 Environmental DNA Isolation Strategies 17

1.2.4 Environmental DNA Cloning Strategies 20

1.2.5 Screening Strategies 23

1.3 Molecules and their biosynthetic genes from metagenomic libraries **30**

1.3.1 Phenotypic Screening of Metagenomic Libraries 30

1.3.2 Sequence-Based Screening of Metagenomic Libraries 45

1.3.3 Tailoring Enzymes from eDNA 51

1.4 Future Challenges and Outlook **53**

CHAPTER 2 ..... **55**

2 Mining the Metagenome for Natural Product Diversity **55**

2.1 Introduction **55**

2.2 Results **56**

2.2.1 Library Cloning and Transduction 56

2.2.2 First Generation Sequence-Based Screens 62

2.2.3 Library Size Analysis 76

2.2.4 Second Generation Sequence-Based Screens 80

2.3 Discussion and Future Directions **96**

2.4 Materials and Methods **98**

2.4.1 eDNA Library Construction Details 98

2.4.2 General Screening Procedure for Library Size Analysis 100

2.4.3 Identification of Gene Clusters of Interest 101

2.4.4 General Procedure for Clone Recovery 104

CHAPTER 3 107

3 Assembly and Heterologous Expression of Large Natural Product Gene Clusters using TAR 107

3.1 Introduction 107

3.2 Results 112

3.2.1 Transformation Associated Recombination (TAR)-Mediated Assembly of Large Natural Product Gene Clusters 112

3.2.2 TAR Cloning for Culture-Based Natural Product Research 129

3.3 Discussion and Future Directions 133

3.4 Materials and Methods 135

3.4.1 pTARa Vector Construction 135

3.4.2 TAR Cloning 136

3.4.3 Pathway-Specific Capture Vector Construction 136

3.4.4 TAR Cloning and Pathway Assembly 140

3.4.5 Analysis of TAR Recombined Clones 142

3.4.6 Fluostatin Characterization 143

3.4.7 Defined Culture Medias 143

CHAPTER 4 144

4 Characterizing Metagenomic Chemical Diversity 144

4.1 Introduction 144

4.2 Results 147

4.2.1 Characterizing the Species Diversity of eDNA Libraries. 147

4.2.2 Characterizing Biosynthetic Diversity in Soil Microbiomes 157

4.2.3 Linking Phylogeny and Biosynthetic Function in Uncultured Bacteria 185

4.3 Discussion and Future Directions 191

4.4 Materials and Methods 193

4.4.1 eDNA Sample Preparation 193

4.4.2 Screening Procedure 194

4.4.3 Phylogeny Reconstruction and Biosynthetic Profiling 198

CHAPTER 5 206

5. Concluding Remarks 206

5.1 Future directions 209

5.1.1 Quantitatively Linking Phylogeny and Function 209

5.1.2 High-Throughput Sequencing and Gene Synthesis 211

5.1.3 Synthetic Biology and Natural Product Discovery 214

APPENDIX 219

GLOSSARY 225

REFERENCES 227

List of Figures

Figure 1: New chemical entities introduced over the past 25 years 2
Figure 2: Natural products as chemical biology tools 3
Figure 3: Condensations during templated biosynthesis 6
Figure 4: Central metabolism and building blocks for natural products 8
Figure 5: Bioassay guided fractionation 9
Figure 6: New classes of antibiotics introduced since 1925 10
Figure 7: Major bacterial divisions and their proportion of cultured isolates 11
Figure 8: A general overview of cultivation-independent natural product discovery. 15
Figure 9: eDNA library construction. 18
Figure 10: Alternative screening strategies. 28
Figure 11: Structures of terragines A-E (1-6) 31
Figure 12: Various N-acyl amino acids from eDNA screening 33
Figure 13: Proposed biosynthetic scheme for long chain fatty acid enol esters 34
Figure 14: Deoxyviolacein and violacein 36
Figure 15: Turbomycin A and B 37
Figure 16: Isonitrile-functionalized indole derivative 38
Figure 17: Antifungal PKS gene cluster 40
Figure 18: Indigo\indirubin 42
Figure 19: *Ralstonia metallidurans*-based libraries 43
Figure 20: Myristoylputrescine (26) and 1, 3-hydroxy myristoylputrescine (27) 45
Figure 21: PCR-based screening for novel natural products (type II PKS) 46
Figure 22: Dienic alcohols 49
Figure 23: Erdacin 51
Figure 24: Novel heptapeptide (teicoplanin) congeners from eDNA 53
Figure 25: Biosynthetic functional groups and modifications found in various secondary metabolites 62
Figure 26: Degenerate primer design strategy 64
Figure 27: Sequence-based screening of biosynthetic features 65
Figure 28: Various halometabolites 66
Figure 29: Proposed biosynthetic halogenation mechanisms 68
Figure 30: Screening and clone recovery outline 69
Figure 31: BB16, a gene cluster encoding a putative phosphinothricin tripeptide analog 73
Figure 32: First-generation type II PKS pathways 76
Figure 33: Library size analysis 79
Figure 34: Molecular phylogeny of α-KG (A) and FAD dependent (B) halogenases 80

Figure 35: α-KG Halometabolite gene clusters 82
Figure 36: NRPS gene cluster 86
Figure 37: PKS gene cluster 90
Figure 38: FRI gene cluster 94
Figure 39: Functional metagenomics with TAR-assembled gene clusters 112
Figure 40: TAR cloning and assembly of natural product gene clusters 114
Figure 41: pTARa *S. cerevisiae/E. coli/Streptomyces* shuttle capture vector 115
Figure 42: Pathway-specific capture vector assembly 117
Figure 43: Detailed analysis of TAR-reassembled gene clusters 121
Figure 44: TAR assembly of multi-clone eDNA-derived gene clusters 123
Figure 45: Sequencing-independent TAR assembly of gene clusters 124
Figure 46: Novel Fluostatins F-H from a TAR assembled biosynthetic pathway 128
Figure 47: Colibactin gene cluster 130
Figure 48: Direct TAR cloning of natural product gene clusters from sequenced organisms 132
Figure 49: Overview of high-throughput sequencing screens for biosynthetic diversity 146
Figure 50: β-diversity analysis of crude eDNA samples 151
Figure 51: Phylogenetic analysis of crude eDNA and associated library samples. 154
Figure 52: Phylogenetic analysis of crude eDNA samples 156
Figure 53: Barcoded clone recovery strategy 161
Figure 54: Adenylation domain analysis of crude eDNA samples 163
Figure 55: Adenylation and ketosynthase domain β-diversity in crude eDNA 166
Figure 56: Type I PKS relative sequence similarity 174
Figure 57: Type II PKS relative sequence similarity 175
Figure 58: Adenylation domain (NRPS) relative sequence similarity 176
Figure 59: Terpene cyclase relative sequence similarity 177
Figure 60: Halogenase relative sequence similarity 178
Figure 61: Indolocarbazole oxidase relative sequence similarity 179
Figure 62: Bacterial nitric oxide synthase relative sequence similarity 182
Figure 63: Major bacterial divisions detected with screening primers 187
Figure 64: Biosynthetic richness estimates 190
Figure 65: Linking phylogeny and function using microfluidic digital PCR 211
Figure 66: Theoretical sequencing-based workflow for natural product discovery 214
Figure 67: Heterologous production of artemisinin 217
Figure 68: Type II PKS domain rarefaction analysis 219
Figure 69: Adenylation domain (NRPS) rarefaction analysis 220
Figure 70: Flavin-depedent halogenase rarefaction analysis 221
Figure 71: Type I ketosynthase domain rarefaction analysis 222

Figure 72: Terpene cyclase rarefaction analysis 223
Figure 73: Bacterial nitric oxide synthase rarefaction analysis 224

List of Tables

Table 1: Various metagenomic libraries screened for natural products 22
Table 2: Primary eDNA libraries constructed and screened in these studies 60
Table 3: Cryptic NRPS gene cluster annotation 86
Table 4: Cryptic PKS gene cluster annotation 90
Table 5: FRI gene cluster annotation 94
Table 6: Primers used for pathway-specific capture vector construction 140
Table 7: Species evenness of eDNA extracts 152
Table 8: Species richness of eDNA libraries 155
Table 9: Sorenson similarity indices 166
Table 10: Biosynthetic enzyme diversity estimates 170
Table 11: High-throughput sequencing primers 195

List of Equations

Equation 1: Shannon evenness (E) 203
Equation 2: The general form of the Chao1 estimator..... 203
Equation 3: Sorenson's similarity index 203

List of Abbreviations

AD	Adenylation domain
AMP	Adenosine monophosphate
ATP	Adenosine triphosphate
ACP	Acyl carrier protein
BAC	Bacterial artificial chromosome
bla	Beta-lactamase (resistance gene)
BLAST	Basic local alignment search tool
CEN/ARS	Centromere/autonomous replicating sequence
CoA	Coenzyme A
CSM	Complete synthetic medium
DNA	Deoxyribonucleic acid (ds – double stranded)
DMSO	Dimethylsulfoxide
eDNA	Environmental DNA
ESI/MS	Electrospray ionization mass spectrometry
FAD	Flavin adenine dinucleotide
GFP	Green fluorescent protein
HPLC	High pressure liquid chromatography
HRMS	High resolution mass spectrometry
IPP	Isopentanyl pyrophosphate
Kan	Kanamycin (or associated resistance gene)
kb	Kilobase
KS	Ketosynthase domain

LC/MS	Liquid chromatography/Mass spectrometry
MB	Megabase
NCBI	National center for biotechnology information
NMR	Nuclear magnetic resonance spectroscopy
NRPS	Non ribosomal peptide synthetase
OTU	Operational taxonomic unit (i.e. unique sequence)
PCA	Polymerase cycling assembly
PCR	Polymerase chain reaction
PKS	Polyketide synthase
rRNA	Ribosomal ribonucleic acid
s.d.	Standard deviation
TAR	Transformation associated recombination
Tm	Melting temperature
TPS	Terpene synthase
YAC	Yeast artificial chromosome
YCP	Yeast centromeric plasmid

CHAPTER 1

1 Introduction and Background

1.1 Natural Products and their Uses

Natural products isolated from cultured organisms have provided many of the most important pharmacophores discovered to date (Figure 1). (Newman, Cragg et al. 2003; Newman and Cragg 2004; Newman and Cragg 2007) These small molecules display a range of activities including immunosuppressive, antibiotic, and antineoplastic properties, and novel functions are continually being characterized. (Howitz, Bitterman et al. 2003; Miao, Coëffet-Legal et al. 2005; Harrison, Strong et al. 2009) Their potent bioactivity is attributed in large part to selective pressures which have tuned the structural and chemical properties of natural products in order to effectively function in a biological setting. In their native microbial hosts, natural products have been shown to play critical roles in biological processes including nutrient scavenging, virulence, and self-defense. (Omura, Ikeda et al. 2001; Wolfgang, Kulasekara et al. 2003; Nougayrède, Homburg et al. 2006; Wyatt, Wang et al. 2010) It is also clear from the sequencing of bacterial genomes that natural products play important functions that are complementary to other biological substrates, as some bacteria dedicate up to 15% of their genomes to encoding canonical biosynthetic gene clusters.

(Wyatt, Wang et al. 2010) It is difficult to predict the types of synthetic chemical scaffolds that would be required to perturb a biological network *in vivo,* whereas natural products have been selected to function in this context. As a result, historically, most bioactive compounds have been developed from a much smaller subset of natural product scaffolds which have been synthetically tailored and derivitized to extend their utility.

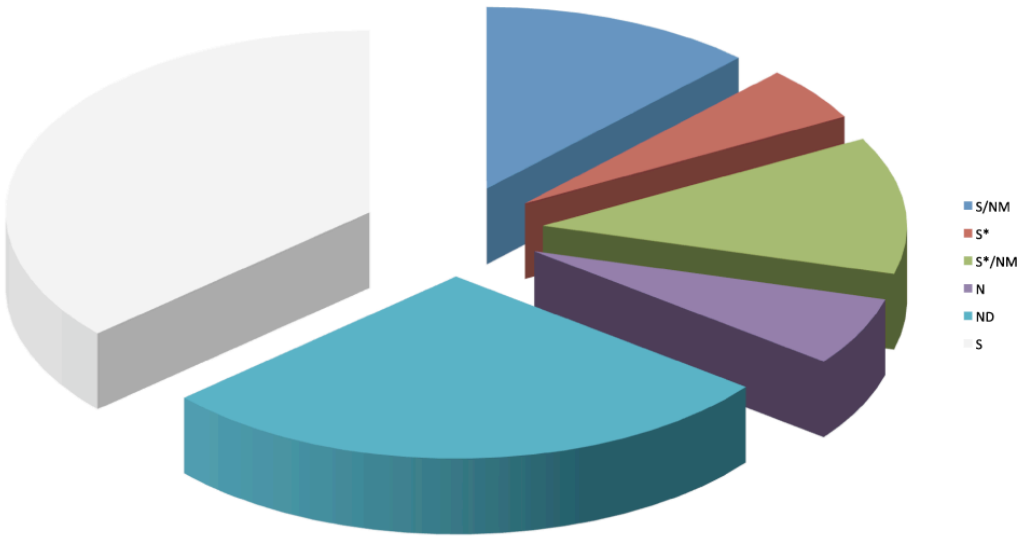


Figure 1: New chemical entities introduced over the past 25 years

Natural products comprise the majority of new chemical entities that have been released over the past 25 years (colored). N = natural product, ND = natural product derived, S*/NM = synthetic natural product mimic with a natural product pharmacophore, S* = synthetic natural product pharmacophore, S/NM = synthetic natural product mimic, S = fully synthetic. *Adapted from* (Newman, Cragg et al. 2003; Newman and Cragg 2004; Newman and Cragg 2007)

In addition to these functions, natural products have also provided useful tools to study other biological systems by offering a way to post-translationally perturb protein function for chemical biology applications (Figure 2). (Bertrand, Postle et al. 1983; Spencer, Belshaw et al. 1996; Stockwell and Schreiber 1998; Mootz and Muir 2002; Schneekloth, Fonseca et al. 2004; Pratt, Schwartz et al. 2007; Schwartz, Saez et al. 2007)

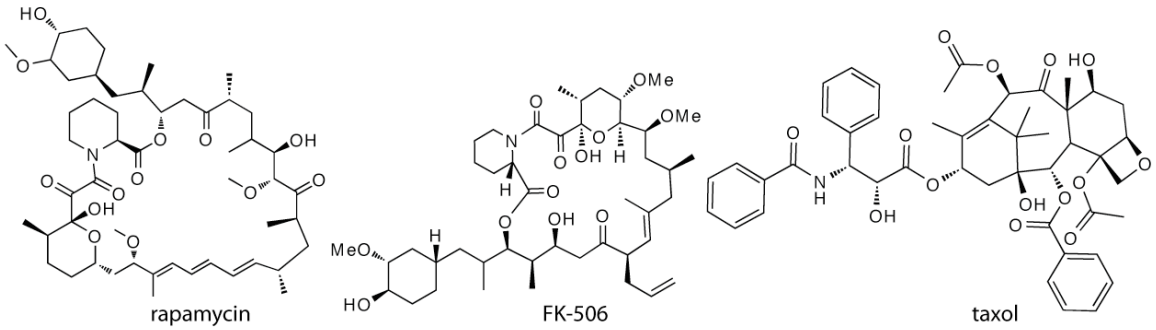


Figure 2: Natural products as chemical biology tools

In general, natural products represent a biologically-relevant source of chemical diversity that can contain *a priori* solutions to target complementarity, biological network specificity, solubility, and other features that often need to be optimized with synthetic chemical scaffolds. In addition, many of the features found in natural products are challenging or intractable to generate using traditional synthetic approaches to molecular design. Natural products are therefore an important raw source of chemical diversity. (Corey and Li 1999; Trost and Dong 2008; Nicolaou, Dalby et al. 2009) Many biosynthetically-derived chemical transformations achieve site-specificity, chirality, and reactivity that cannot be attained with synthetic

strategies, making biosynthetic catalysis an attractive complement to traditional chemical synthesis. (Savile, Janey et al. 2010) Natural products and their biosynthetic enzymes therefore play an important role in chemistry by providing bioactive pharmacophores, inspiring biologically relevant chemical design, complementing synthetic strategies, and providing chemical tools that can be used to study other biological processes.

1.1.1 Natural Product Biosynthesis

The chemical diversity found in natural products is generated in large part through the combinatorial shuffling of a much smaller set of conserved biosynthetic enzymes which catalyze the formation of natural products using basic chemical building blocks (Figure 3). For polyketide-like compounds, malonyl-CoA and methylmalonyl-CoA comprise the majority of monomer units incorporated during chain elongation. Starter units for this class of compounds can be comprised of thioesters of acetyl-, proprionyl-, benzoyl-CoAs, and variants of these basic structures. (Moore and Hertweck 2002) During polyketide biosynthesis, the decarboxylation of malonyl-CoA (or methylmalonyl-CoA), generates a nucleophilic enolate thioester which attacks an acyl thioester (or other starting unit) in a Claisen condensation (Figure 3a). For nonribosomal peptides, the elongation units consist of the 20 canonical amino acids in addition to a diverse range of non-proteinogenic amino and aryl acids which are often biosynthesized by genes found within a natural product gene cluster. (Schwarzer, Finking et al. 2003) Nonribosomal

peptide biosynthesis proceeds via the nucleophilic attack of an aminoacyl thioester on an upstream thioester to form an amide bond (Figure 3b). Isoprenoids are a diverse group of natural products (~40,000) encoded by non-templated biosynthetic systems. Chemically, they are generated via the condensation of five-carbon isoprene units (2-methyl-1,3-butadiene) which can be further modified via heterocyclizations and tailoring steps with various finishing enzymes. The universal five-carbon starting monomers for isoprenoid biosynthesis, IPP (isopentanyl pyrophosphate) or its isomer DMAPP (dimethylallylpyrophosphate), are generated using two primary constituent pathways, the mevalonate (MEV) and deoxyxylulose 5-pyrophosphate (DXP) pathway. The synthesis of all higher terpenoids (>10 carbon units) proceeds through polymers of IPP and DMAPP in the form of GPP (geranyl pyrophosphate), FPP (farnesyl pyrophosphate), and GGPP (geranylgeranyl pyrophosphate). (Lange, Rujan et al. 2000)

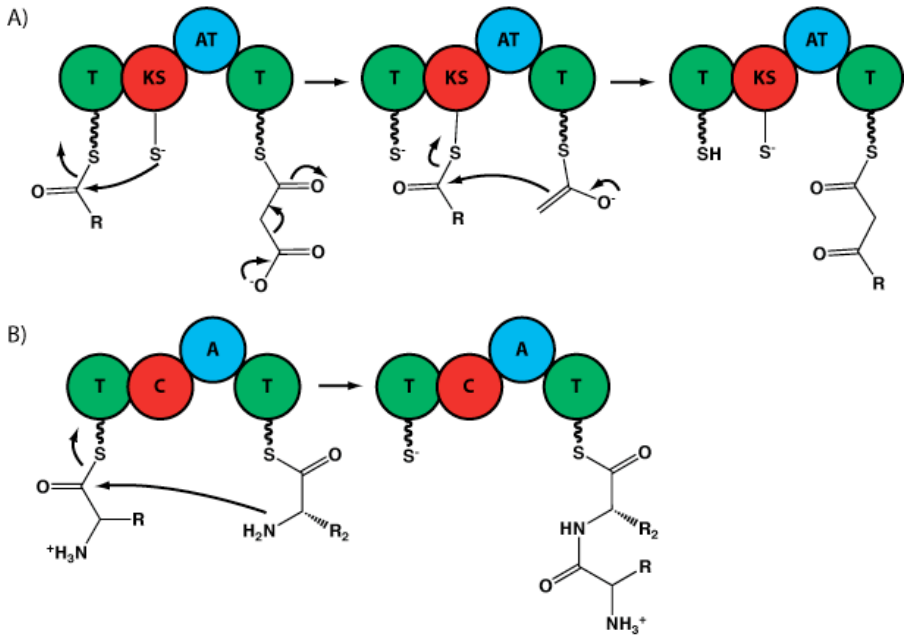


Figure 3: Condensations during templated biosynthesis

During polyketide biosynthesis, the decarboxylation of malonyl-CoA (or methylmalonyl-CoA), generates a nucleophilic enolate thioester which attacks an acyl thioester (or other starting unit) in a Claisen condensation (A). Nonribosomal peptide biosynthesis proceeds via the nucleophilic attack of an aminoacyl thioester on an upstream thioester to form an amide bond (B). T = thiolation domain, KS = ketosynthase, AT = acyltransferase, C = condensation, A = adenylation domain.

For templated biosynthetic systems, the number and arrangement of each enzyme in a gene cluster determines the final molecular structure as each chemical unit is typically added in a serial manner during chain elongation and modification. Templated biosynthetic systems can also be combined and formed from both PKS and NRPS modules to generate even more structural diversity. In addition, core natural product scaffolds can be further modified by tailoring enzymes which alter the final structure of a natural product via macrocyclization, halogenation, glycosylation,

phosphorylation, sulfation, methylation, and additional chemical transformations. Many of these tailoring steps are performed in a site-specific or chiral manner, which further expands the structural diversity found in natural products. (Savile, Janey et al. 2010) One advantage of the modularity of some biosynthetic systems is that a relatively small number of enzymes can be combined in different arrangements to produce a large number of possible output chemical structures. While this strategy is clearly utilized in nature, the rational rearrangement of biosynthetic proteins to produce novel natural products has met with only limited success. (Cane, Walsh et al. 1998; Palaniappan, Kim et al. 2003; Nguyen, Ritz et al. 2006; Fischbach, Lai et al. 2007) This is partially due to the complexity of these large, multi-component systems and the coevolution of allosteric domains that are required for successful intermediate catalysis. (Gokhale and Khosla 2000) Also, the manipulation and engineering of large biosynthetic gene clusters is non trivial as will be outlined in Chapter 3.

As a central tenet, the enzymes involved in core biosynthetic transformations and many tailoring steps are highly conserved as a result of their modular compatibility and the use of related starting monomers. The precursor molecules used for natural product biosynthesis are typically shunt products from central metabolic biochemical pathways (Figure 4). This feature lends itself to the possibility of heterologously producing natural products in non-native biological hosts as these pathways are generally

conserved. The central metabolism of model laboratory hosts can also be synthetically engineered to increase the levels of these precursors in an effort to produce higher quantities of natural products. (Martin, Pitera et al. 2003; Ro, Paradise et al. 2006; Chang, Eachus et al. 2007; Tsuruta, Paddon et al. 2009)

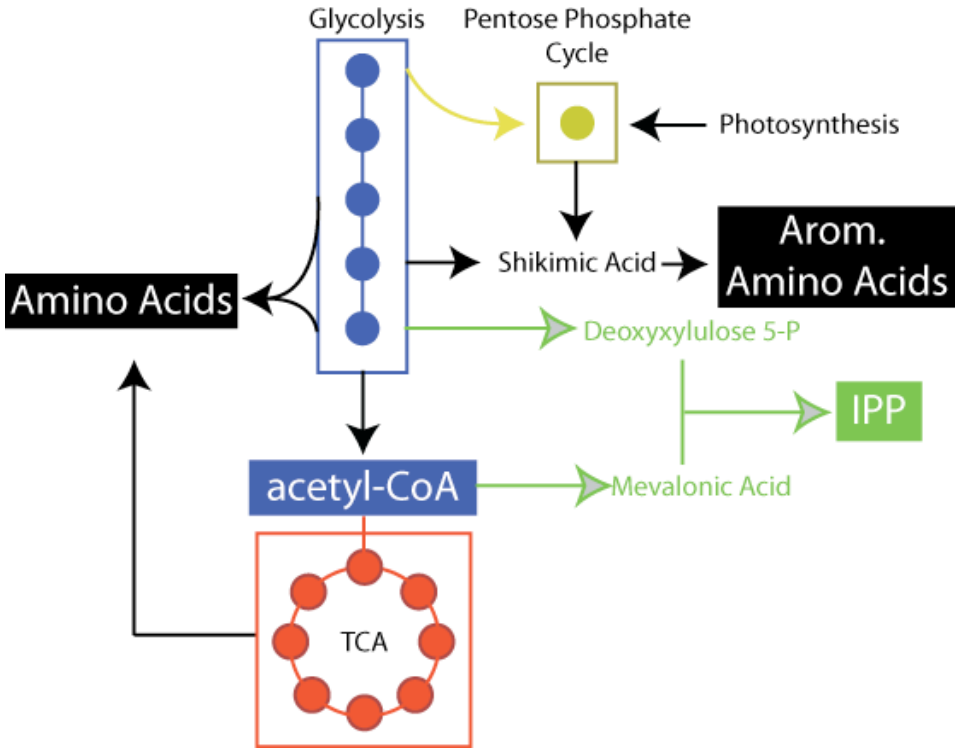


Figure 4: Central metabolism and building blocks for natural products

Central metabolism provides many of the core building blocks for natural product biosynthesis. The four primary sources of starting units for biosynthetic systems described in this thesis are shown (solid fill).

1.1.2 Natural Products and Uncultured Bacteria

In the mid 20th century, Selman Waksman helped initiate a productive era of antibiotic discovery by characterizing numerous antibiotics from cultured soil bacteria. In the decades since these early discoveries, the methods used to uncover novel natural products have remained largely unchanged (Figure 5). Culture-dependent natural product discovery efforts have identified many of the most important pharmacophores known to date. The rate of rediscovery of common natural product scaffolds from easily cultured bacteria is now, however, estimated to exceed 99%. (Zaehner and Fiedler 1999; Newman, Cragg et al. 2003; Newman and Cragg 2004; Newman and Cragg 2007) Novel antibiotics, of which natural products are a major source, have also become increasingly difficult to access due to the rediscovery of common antibiotic scaffolds from easily cultured sources (Figure 6).

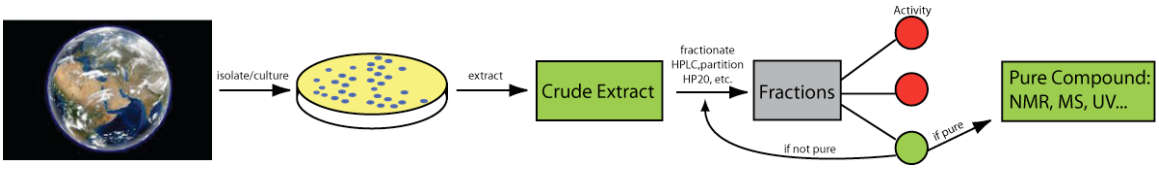


Figure 5: Bioassay guided fractionation

Cultured environmental samples (in this example, bacteria) are isolated and crude extracts are tested using bioactivity as a read out. Iterative rounds of fractionation and activity assays are used to deconvolve active compounds from crude extract mixtures until a pure compound is obtained for structural elucidation.

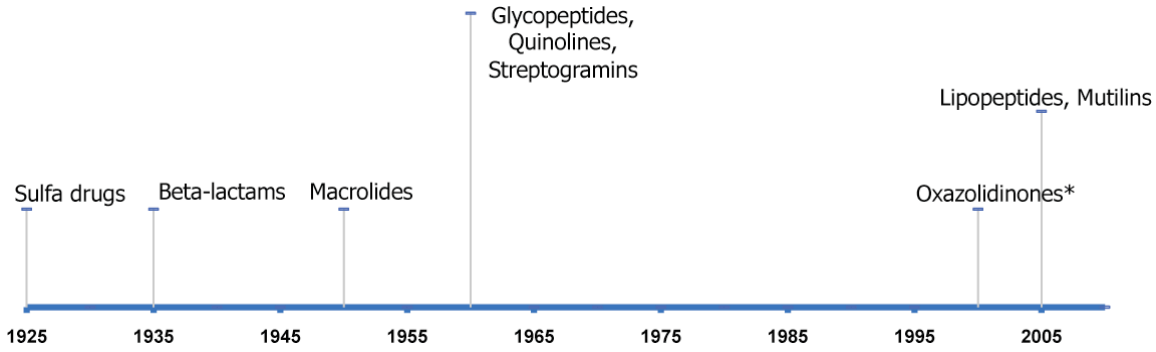


Figure 6: New classes of antibiotics introduced since 1925

Between 1962 and 2000 no new major classes of antibiotics were introduced. While antibiotics represent only one type of natural product, this trend extends to many other classes of natural products. * = synthetically derived. *Adapted from* (Fischbach and Walsh 2009)

The ability to grow an organism in pure culture is a prerequisite to producing and isolating bioactive compounds using traditional discovery strategies (Figure 5). For much of the last century, it has been known that there is a large discrepancy between the number of bacteria present in environmental samples and the number that could be easily cultured. (Jannasch and Jones 1959; Torsvik, Goksøyr et al. 1990) In general, the difficulty of culturing environmentally isolated bacteria has prevented the discovery of natural products from a significant portion of bacterial diversity. More recent molecular phylogenetic analyses have begun to quantitatively describe the severity of this barrier by analyzing single genes indirectly amplified from environmental samples in a cultivation-independent manner. These studies now indicate that traditional culturing methods provide access to less than 1% of the bacterial diversity in our environment.

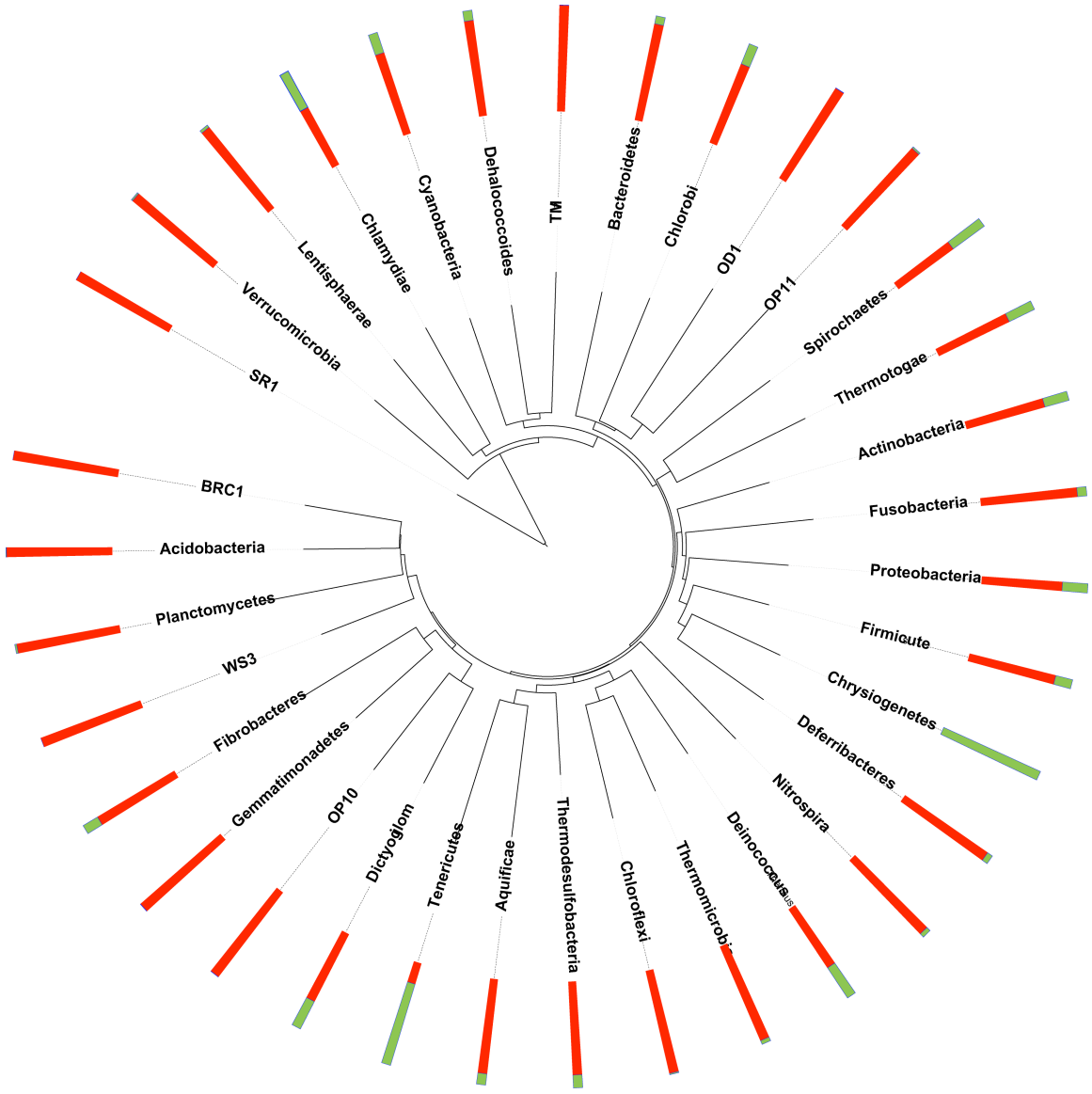


Figure 7: Major bacterial divisions and their proportion of cultured isolates

A phylogenetic tree showing the relationship between major bacterial divisions (not including candidate environmental bacterial divisions) highlights the proportion of cultivated versus presently uncultivated bacterial species. (Red = uncultivated, Green = cultured). Percentages are normalized per bacterial division for ease of visualization. Data sourced from the Ribosomal Database Project update 15. (Cole, Chai et al. 2007; Cole, Wang et al. 2009)

Microbial diversity analyses now suggest that more than 80 major bacterial divisions exist and less than half of these contain cultured isolates.

(Keller and Zengler 2004; Schloss and Handelsman 2004; DeSantis, Hugenholtz et al. 2006; DeSantis, Hugenholtz et al. 2006; Cole, Chai et al. 2007; Webster, Yarram et al. 2007; Cole, Wang et al. 2009) A single gram of soil is predicted to contain up to 10,000 unique bacterial species, and by most estimates, less than 1% of these have been grown in pure culture (Figure 7). (Torsvik, Goksøyr et al. 1990; Torsvik, Salte et al. 1990; Tankéré, Bourne et al. 2002; Torsvik, Øvreås et al. 2002; Rappé and Giovannoni 2003; Webster, Yarram et al. 2007) Additional studies have shown that uncultured bacteria from other microbiomes outnumber their cultured counterparts by several orders of magnitude. (Kaeberlein, Lewis et al. 2002; Qin, Li et al. 2010) Recent large-scale sequencing projects also indicate that the environment contains a large reservoir of novel genes that have not yet been studied due to cultivation barriers. (Rusch, Halpern et al. 2007) Taken together, these insights suggest that uncultured bacteria are likely the largest remaining pool of genetic diversity that has not been screened for the production of secondary metabolites.

To circumvent the challenges of culturing environmental bacteria, a novel discovery strategy has been developed which relies on the extraction of DNA from environmental samples (environmental DNA, eDNA) and the cloning of this DNA into easily cultured model bacterial hosts. The culture independent analysis of natural microbial communities is now known as metagenomics. (Handelsman, Rondon et al. 1998) Using this strategy,

environmentally-derived genes can be propagated, studied, and functionally examined in tractable laboratory hosts with established genetic tools. Functionally accessing the genomes of uncultured bacteria using this approach has been of particular interest to the natural products community because the genes required for the biosynthesis of a bacterial secondary metabolite, including regulatory, tailoring, and resistance (self immunity) enzymes, are often found clustered on bacterial chromosomes. Therefore, the direct isolation of high molecular weight DNA from the environment can, in theory, yield functionally intact natural product biosynthetic gene clusters that can confer the production of new metabolites in a heterologous host. (Brady, Simmons et al. 2009; Kim, Simmons et al. 2009; Craig, Chang et al. 2010) While many microbiomes harbor biosynthetic genes, including symbiotic bacteria, this chapter will focus on terrestrial microbiomes and will include a discussion of the metabolites which have been characterized to date using metagenomic techniques. (Schmidt, Nelson et al. 2005)

1.2 Metagenomic Strategies to Natural Product Discovery

1.2.1 Environmental Culturing and Sequencing Efforts

A number of novel culturing strategies, including consortia-culturing, single cell microdroplet encapsulation, environmental nutrient diffusion and very low nutrient growth, have been used to cultivate bacteria from environmental samples. (Kaeberlein, Lewis et al. 2002; Zengler, Toledo et al. 2002; Joseph, Hugenholtz et al. 2003; Zengler, Walcher et al. 2005; Green

and Keller 2006; Ingham, Sprenkels et al. 2007) While these methods represent technical advances, they do not provide general or scalable solutions for culturing the majority of bacteria present in our environment. Most environmental bacteria remain inaccessible using culture-based methods and therefore little is known about them beyond their 16s rRNA gene sequences. Due to the complexity of microbial communities only a handful of complete bacterial genomes have been successfully reconstructed from indirect environmental sequencing efforts. (Tyson, Chapman et al. 2004; Baker, Tyson et al. 2006; Dinsdale, Edwards et al. 2008) Complete sequencing of even dominant species in a phylogenetically diverse microbiome such as soil is predicted to require up to 5 billion base pairs of sequencing data. (Tringe, von Mering et al. 2005) Although there have been significant technical advances over the past decade, high throughput sequencing is still not a tractable approach to directly analyze large numbers of uncultured bacterial genomes from environmental microbiomes. (Margulies, Egholm et al. 2005; Shendure, Porreca et al. 2005; Eid, Fehr et al. 2009) Studying natural product gene clusters from environmental microbiomes therefore requires metagenomic cloning and sequence enrichment techniques which are described in the following sections.

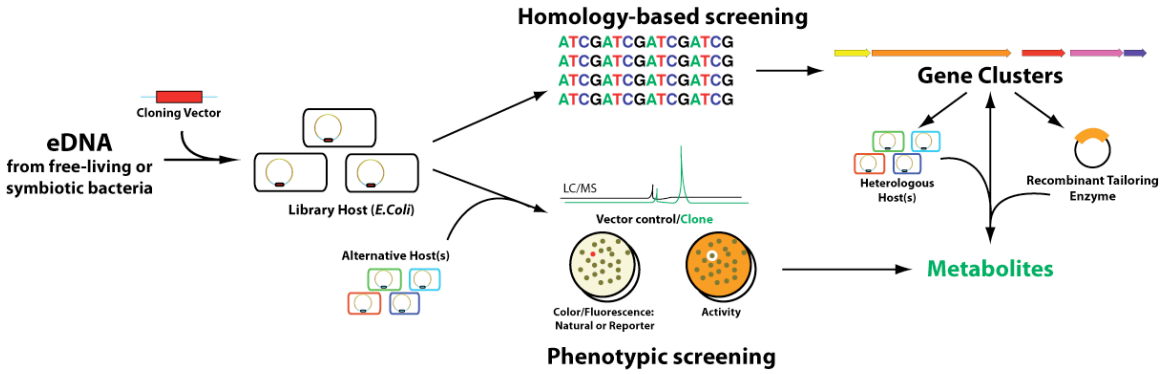


Figure 8: A general overview of cultivation-independent natural product discovery.

eDNA is purified and cloned into a vector that allows the high efficiency transformation of a tractable library host, typically *E. coli*. One of two primary screening methodologies are generally used (homology-based/phenotypic) in order to identify clone-specific molecules or biosynthetic gene clusters for downstream heterologous expression and characterization efforts. Metabolites identified in these screens are directly linked to the genes that encode their biosynthesis. *Adapted from* (Brady, Simmons et al. 2009)

1.2.2 Environmental Sample Selection

Most environments, including marine, freshwater, soil, and the human gut, have been shown to contain diverse microbiomes that could potentially serve as sources of metagenomic DNA. (Qin, Li et al. 2010) Soil contains the most diverse microbial population of these environments. For example, it is estimated that the bacterial diversity present in one ton of soil exceeds that of all marine environments combined. (Whitman, Coleman et al. 1998; Curtis, Sloan et al. 2002; Gans, Wolinsky et al. 2005) Recent analyses indicate that diverse terrestrial microbiomes can contain up to 10,000 unique bacterial species per gram of soil. (Whitman, Coleman et al. 1998; Curtis, Sloan et al. 2002; Gans, Wolinsky et al. 2005) Due to this extraordinary bacterial

diversity, soil has been an attractive yet challenging starting point for metegenomic studies designed to identify natural products. High-throughput sequencing indicates that there are often significant phylogenetic differences between soils at a species level. (Dinsdale, Edwards et al. 2008) Changes in temperature, water content, particle size, soil type, heavy metal contamination, and pH have all been shown to affect the microbial diversity present in a soil sample. (Courtois, Frostegård et al. 2001; Liles, Manske et al. 2003; Bertrand, Poly et al. 2005; Gans, Wolinsky et al. 2005; Liles, Williamson et al. 2008) A comparison of microbial diversity in different soils, using fluorescence microscopy and 16s rRNA analysis, a commonly used phylogenetic marker, found that the number of detectable species can vary by more than three orders of magnitude among different terrestrial sources. (Torsvik, Goksøyr et al. 1990) Although phylogenetic analysis of microbiomes using 16s rRNA sequencing has now become routine, there have been no systematic studies to identify which microbiomes contain a high diversity of biosynthetic enzymes. (DeSantis, Hugenholtz et al. 2006; DeSantis, Hugenholtz et al. 2006; Sogin, Morrison et al. 2006; Cole, Chai et al. 2007; Cole, Wang et al. 2009) Samples used to construct metagenomic libraries for secondary metabolite discovery have therefore been primarily selected based on the quality and quantity of high molecular weight eDNA that they yield. Future natural product discovery efforts will clearly benefit from comparative metagenomic analyses that shed light on which environments and

characteristics correlate with biosynthetic enzyme diversity. Chapter 4 will outline steps we took toward achieving this goal.

1.2.3 Environmental DNA Isolation Strategies

Two primary methods, direct DNA extraction and whole-cell purification, have been used to isolate DNA from environmental samples for metagenomic library construction. In direct DNA extraction protocols, bacteria present in an environmental sample are lysed *in situ* (using a combination of heat, detergents, enzymes, organic solvents, or physical perturbations), DNA is collected by alcohol precipitation from a centrifuge-clarified lysate, and pure eDNA is obtained from the crude precipitate using gel or affinity matrix purification (silica, Sephadex). (Miller, Bryant et al. 1999; Brady 2007; Sharma, Capalash et al. 2007; Sharma, Radl et al. 2007) Attempts to optimize direct DNA extraction methods indicate that chelating agents (CTAB (cetyl trimethylammonium bromide), ammonium acetate), which help remove humic acids, increase the likelihood of obtaining DNA that can be more efficiently cloned. It has also been demonstrated that the inhibition of contaminating nucleases with formamide or calcium carbonate increases the size of the recovered eDNA. (Zhou, Bruns et al. 1996; Liles, Manske et al. 2003; Sagova-Mareckova, Cermak et al. 2008) Direct DNA isolation strategies generally yield microgram quantities of 30-50 kb eDNA (~50 ug/gram) which is well suited for cosmid and fosmid based cloning systems. DNA isolated using direct methods, however, is typically not large

enough to construct insert libraries that exceed 50 kb in length. (Sagova-Mareckova, Cermak et al. 2008) Biosynthetic gene clusters are often longer than 50 kb which prevents many functionally intact gene clusters from being captured on single clones using direct eDNA isolation strategies. While this size limitation may prevent the detection of larger natural product gene clusters using functional screens, larger natural product gene clusters can, in theory, be recovered from collections of overlapping eDNA-derived cosmid clones. Chapter 2 outlines steps we took toward achieving this goal.

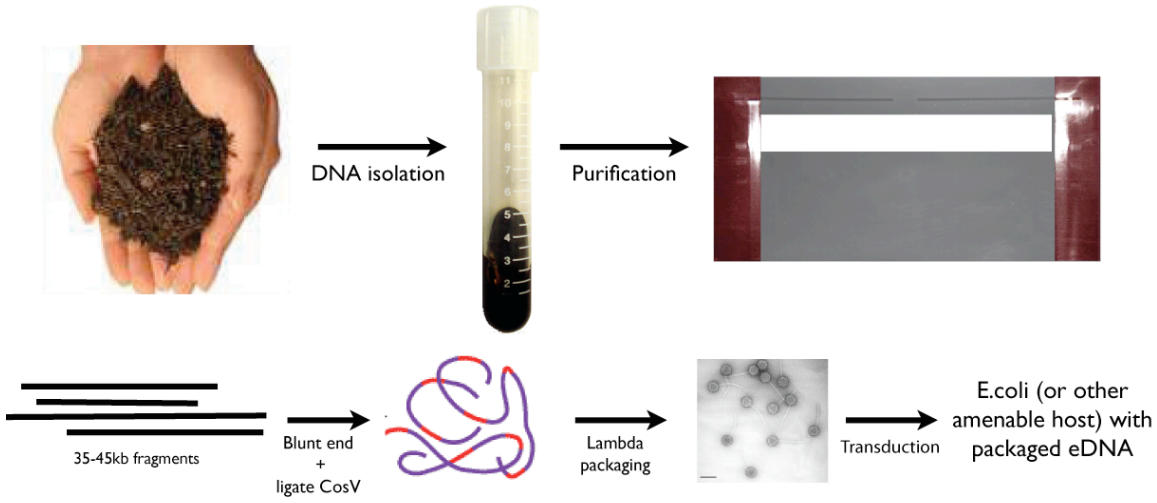


Figure 9: eDNA library construction.

An environmental sample (in this case soil) is processed with a combination of heat and chemical lysis to release DNA. The DNA is then precipitated, and crude eDNA is purified via gel electrophoresis to yield ~45 kb fragments of high molecular weight DNA. This DNA is blunt-end repaired, ligated to a cosmid or fosmid vector, packaged with lambda phage, and transduced into a library host (*E. coli*).

The enrichment of bacteria from environmental samples prior to cell lysis has been used to obtain higher molecular weight DNA. Bacterial

enrichment strategies are less efficient, however, so this approach often yields 10-100 fold less DNA than direct DNA isolation methods. (Courtois, Cappellano et al. 2003) In these experiments, cells are purified from environmental contaminants using differential centrifugation, filtration, or high-speed differential density centrifugation through a Nycodenz™ polymer. (Ford and Rickwood 1982; Rickwood, Ford et al. 1982; Gellert-Mortimer, Clarke et al. 1988; Courtois, Frostegård et al. 2001) The bacteria are then typically embedded in agarose and lysed using a combination of detergents and enzymes. Environmental DNA fragments in excess of 1Mb in size have been isolated using this approach, however low yields have made the construction of large eDNA libraries challenging. (Liles, Williamson et al. 2008) Direct extraction protocols, which yield higher overall quantities of eDNA, have therefore been the preferred method to create large metagenomic libraries from soil microbiomes.

One promising approach that could address the difficulty of isolating large quantities of higher molecular weight eDNA is multiple displacement amplification (MDA) with phi29 polymerase. Phi 29 is capable of efficiently amplifying ~70 kb lengths of DNA and can produce milligram quantities of amplified DNA from a single DNA fragment. (Abulencia, Wyborski et al. 2006; Yokouchi, Fukuoka et al. 2006; Zhang, Martiny et al. 2006; Marcy, Ishoey et al. 2007; Smolina, Lee et al. 2007; Stepanauskas and Sieracki 2007) While this method can yield chimeric amplification products, MDA-generated

DNA has been used to successfully construct metagenomic libraries from low abundance organisms. (Yokouchi, Fukuoka et al. 2006; Podar, Abulencia et al. 2007) Also, reductions in reaction volume have been shown to decrease the chances of generating chimeric artifacts using MDA. (Marcy, Ishoey et al. 2007) Recently, a novel electrophoretic technique based on the synchronous coefficient of drag alteration (SCODA) was used to efficiently extract and concentrate high molecular weight DNA (>100 kb) directly from soil samples with minimal preparation steps. (Marziali, Pel et al. 2005) This approach could be used to isolate higher molecular weight DNA from environmental samples for future library construction efforts. The application and refinement of new techniques such as MDA and SCODA will prove useful for the construction of metagenomic libraries for future natural product gene cluster discovery efforts.

1.2.4 Environmental DNA Cloning Strategies

Natural product biosynthetic gene clusters can range from a few kilobases to hundreds of kilobases in length. Efforts to clone functionally intact gene clusters for natural product discovery have therefore relied on cosmid, fosmid, or bacterial artificial chromosome (BAC) vectors that are capable of cloning large fragments of DNA. Cosmid and fosmid-based systems, which rely on a lamda phage packaging step, can accept 35-40 kb DNA inserts. (Bierman, Logan et al. 1992; Kim, Shizuya et al. 1992) While intact gene clusters can be captured on cosmid-sized clones, many large gene

clusters exceed this cloning limit precluding their functional detection. Although BACs are capable of cloning larger inserts, the isolation of very high molecular weight DNA from environmental samples is challenging, and the average insert size for most metagenomic BAC libraries has rarely exceeded 50 kb. (Liles, Williamson et al. 2008) Additionally, cosmid/fosmid based approaches are three to four orders of magnitude more efficient than BACs at cloning metagenomic DNA, and libraries in excess of $1x10^7$ clones can be routinely constructed. (Banik and Brady 2008; King, Bauer et al. 2009; Kim, Feng et al. 2010) Soil environments are among the most diverse microbiomes that have ever been analyzed. Efficient cosmid-based transduction systems in *E. coli* have therefore been the preferred method for creating large soil-based metagenomic libraries.

Table 1: Various metagenomic libraries screened for natural products

Metagenomic libraries constructed from soils have been used for the discovery of natural products. The majority of these libraries have been constructed using cosmid-based vectors due to their high cloning efficiency. *broad host range cosmid vector. 1: (Rondon, August et al. 2000; Gillespie, Brady et al. 2002; Liles, Manske et al. 2003; Riesenfeld, Goodman et al. 2004) 2: (Wang, Graziani et al. 2000) 3: (Brady and Clardy 2000) 4: (Brady, Chao et al. 2001) 5: (MacNeil, Tiong et al. 2001) 6: (Brady, Chao et al. 2002) 7: (Courtois, Cappellano et al. 2003) 8: (Ginolhac, Jarrin et al. 2004) 9: (Brady, Chao et al. 2004)

Origin	Vector Type	No. of Clones	Insert Size (kb)	Total DNA (Gb)	Genes	Date
Uncultivated soil	BAC	3,648: 24,576	27:44.5	1,19	Antimicrobials, resistance genes, 16s rRNA, var. biocatalysts	2000[1]
NA	NA	NA	NA	NA	Antimicrobials	2000[2]
NA	Cosmid	700,000	NA	24.5	Antimicrobials	2000[3]
NA	Cosmid	NA	NA	NA	Pigments	2001[4]
Uncultivated soil	BAC	12,000	37	0.42	Antimicrobials	2001[5]
NA	Cosmid	NA	NA	NA	Fatty acid enol esters	2002[6]
Arable field	Cosmid	5000	NA	0.18	Polyketide synthases, various other activities	2003[7]
Clay loam sandy type	Fosmid	100,000	30-40	3.5	Polyketide synthases, various other activities	2004[8]
NA	Cosmid	NA	NA	NA	Long-chain N-acyltyrosines	2004[9]
Various desert soils	Cosmid	10-15,000,000	35	350-525	This study*	2009-2010
Various topsoils	Cosmid*	100-500,000	35	17.5	This study*	2009

1.2.5 Screening Strategies

1.2.5.1 Functional Screening and Host Selection

Functional and sequence-based screens have been used to identify metagenomic clones that contain natural product biosynthetic gene clusters. (Brady, Simmons et al. 2009; Kim, Simmons et al. 2009; Craig, Chang et al. 2010) Although several novel metabolites have been uncovered using functional screens, these approaches require the coordinated expression of multiple foreign biosynthetic genes in a library host bacterium under the culture conditions used for a particular assay or screen. Also, the recombinant clone must generate sufficient quantities of a metabolite to be detected in the selected assay. Because the gene clusters isolated from a metagenomic sample are of diverse phylogenetic origin, the likelihood that a gene cluster meets all of these requirements is very low, and hit rates for functional screens of metagenomic libraries are generally around 0.01%. (Courtois, Cappellano et al. 2003; Williamson, Borlee et al. 2005; Guan, Ju et al. 2007; Brady, Simmons et al. 2009) Due to these limitations, functional screens of metagenomic libraries have been designed so that they can be easily carried out on a large number of clones. Antimicrobial activity, color production, LC/MS analysis of culture extracts, and reporter gene activation have all been successfully used as readouts in high-throughput assays designed to find small molecule producing clones. (Courtois, Cappellano et al. 2003; Williamson, Borlee et al. 2005; Guan, Ju et al. 2007; Banik and Brady

2008; Brady, Simmons et al. 2009) In general, *E. coli* remains the most efficient cloning host for creating large genomic DNA libraries from metagenomic samples. Libraries have therefore initially been constructed in *E. coli* and later shuttled into different hosts including *S. lividans, Ralstonia metallodurans, Rhizobium leguminosraum, Agrobacterium tumefaciens, Burkholderia graminis, Caulobacter vibrioides*, and *Pseudomonas putida.* (Wang, Graziani et al. 2000; Martinez, Kolvek et al. 2004; Li, Wexler et al. 2005; Craig, Chang et al. 2009; Craig, Chang et al. 2010) The phylogenetic diversity present in soil microbiomes makes the selection of a heterologous host for natural product discovery efforts challenging. (Torsvik, Øvreås et al. 2002) Heterologous expression hosts are inherently limited in their capacity to functionally process foreign DNA. The expanded use of phylogenetically diverse heterologous expression hosts will, therefore, most likely continue to benefit functional screening efforts.

One primary advantage of a functional assay is that no *a priori* knowledge of a biosynthetic enzyme is required to discover novel secondary metabolites. In functional screens, a secondary metabolite is directly linked to its biosynthetic genes allowing the unbiased discovery of novel biosynthetic sequences which have not been characterized before. A complementary sequence-based screen, based on genes discovered using functional assays, can then be used to find additional examples of gene clusters and secondary

metabolites, as demonstrated by Brady et al. in 2002 and 2007. (Brady, Chao et al. 2002; Brady, Bauer et al. 2007)

1.2.5.2 Sequenced-based Screens

As described earlier, metagenomic cloning strategies rely primarily on cosmid-based vectors due to their high efficiency and the ease of isolating eDNA that can be directly utilized in transduction reactions. Many canonical biosynthetic gene clusters captured using this approach, however, are truncated because they are too large to be cloned on single cosmids. Aside from the challenge of selecting an optimal heterologous screening host, this technical barrier highlights a major limitation of using functional assays to discover natural products from cosmid-based eDNA libraries. Sequence-based screening strategies circumvent many of the limitations of functional screens and have the potential to provide access to a more diverse collection of gene clusters. In sequence-based screens, degenerate primers based on conserved regions in natural product biosynthetic genes are used to PCR amplify novel homologues from eDNA libraries. These sequences are then used to recover large insert clones from a metagenomic library typically through dilution fractionation or sequence hybridization. In contrast to functional screens, sequence-based strategies are expression-independent and can be used to recover complete gene clusters from overlapping cosmid clones containing truncated portions of a pathway. Although the assembly of natural DNA constructs of this size is challenging, Red/ET recombineering has been

utilized to reconstruct gene clusters carried on overlapping metagenomic clones. (Vetcher, Tian et al. 2005; Wenzel, Gross et al. 2005) Recombineering approaches can be prohibitively challenging for natural product gene clusters carried on more than two clones, however, so we developed a more robust recombination method that allows us to assemble multi-clone gene clusters in *S. cerevisiae*. Chapter 3 describes the development and application of this technique toward assembling large natural product gene clusters. Some estimates indicate that more than $1x10^{11}$ clones with an average insert size of 100 kb would be required to capture the total bacterial diversity present in one gram of soil. (Handelsman, Rondon et al. 1998; Daniel 2005) Despite the challenge of building libraries of this scale, a wide range of gene clusters have be recovered using DNA based screening of libraries that are $1x10^{7}$ (40 kb inserts) clones in size. (Brady 2007; Banik and Brady 2008; Kim, Feng et al. 2010)

Although gene clusters can now routinely be isolated from eDNA libraries using sequenced-based screening methods, heterologous expression still poses a significant challenge as described for functional screens. Promoter activation, ribosome binding site recognition, product toxicity, and differences in primary metabolite diversity are just some of the factors that can impede heterologous expression. In metagenomic studies, this difficulty is compounded by the fact that neither the molecule produced by the gene cluster nor the source organism is generally known. Despite this limitation,

both polyketide and nonribosomal peptide gene clusters have been successfully heterologously expressed in *E. coli* and various *Streptomycetes*. (Brady, Simmons et al. 2009) As with phenotypic screening, the development of phylogenetically diverse model bacterial systems that can be used for heterologous expression studies should significantly increase the number and diversity of new metabolites that are discovered from eDNA-derived gene clusters using this approach.

1.2.5.3 Other screens

Due to the microbial diversity found in environmental samples, natural product biosynthetic genes represent only a small fraction of the total DNA in a metagenomic library. To overcome this, a number of methods have been successfully used to proportionally enrich biosynthetic sequences, including stable isotope labeling, subtractive hybridization, fluorescence in-situ hybridization (FISH) combined with cell sorting (FACS), affinity capturing, and phage display. (Demidov, Bukanov et al. 2000; Galbraith, Antonopoulos et al. 2004; Kalyuzhnaya, Nercessian et al. 2005; Yin, Straight et al. 2005; Kalyuzhnaya, Zabinsky et al. 2006; Yin, Straight et al. 2007; Kalyuzhnaya, Lapidus et al. 2008; Kalyuzhnaya, Lidstrom et al. 2008) A number of additional enrichment strategies could also be adapted for metagenomic studies, including biotinylated nucleotide hybridization and magnetic capture, gene targeting with selectable markers, and any combination of fluorogenic gene-specific reporters in conjunction with FACS

sorting. Recently, two intracellular reporter assays were developed to identify clone-specific small molecules from metagenomic libraries. METREX (metabolite regulated expression) leverages the conserved architecture of quorum sensing systems to screen for quorum sensing mimics from metagenomic libraries. In this assay, activated LuxR (bound to N-acyl homoserine lactone (AHL) or a mimic from a metagenomic library) induces the expression of green fluorescent protein (GFP), which can be detected using microscopy or FACS.

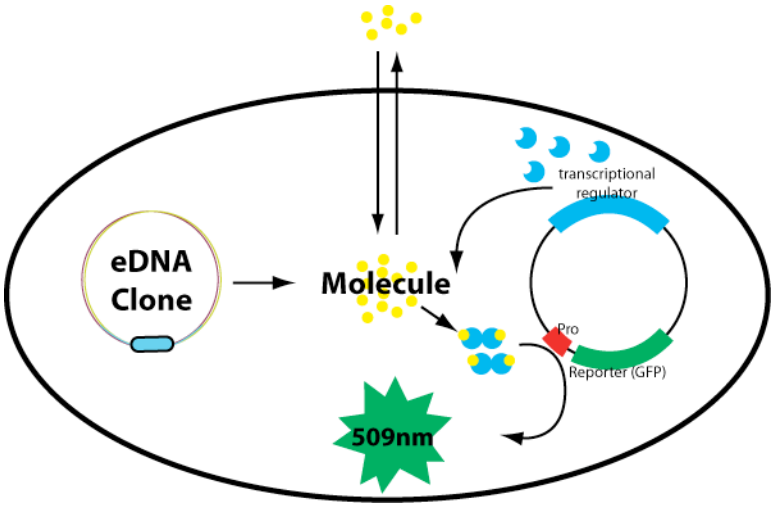


Figure 10: Alternative screening strategies.
Substrate-induced reporter assays such as METREX can be used to detect clone specific molecules from metagenomic libraries. In these approaches, a genetic reporter system is activated by the presence of a secondary metabolite. One potential advantage of this approach is the increased sensitivity of the assay as any secondary metabolites produced by library clones do not have to exit the cell in order to be functionally detected.

This technique was successfully applied to a gypsy moth midgut metagenome and resulted in the identification of indigo and indirubin producing clones.

(Guan, Ju et al. 2007) Small molecules detected in a METREX screen do not have to exit the cell to be detected and therefore provide a potential sensitivity advantage over other functional assays. As a demonstration, the authors of a related study showed that 10 out of 11 clones that were active in the intracellular METREX screen did not activate GFP expression when the sensor was used in an overlay. (Williamson, Borlee et al. 2005) The second related screen, SIGEX (substrate induced gene expression) utilizes operon-trap GFP expression and FACS sorting to identify catabolic gene clusters. Many gene clusters are organized in operons that are induced by the substrate of the gene cluster. By adding exogenous substrate, related gene clusters can be activated and detected using an integrated GFP reporter and FACS sorting. Novel clones containing biosynthetic gene clusters have been identified in metagenomic libraries using this approach. (Uchiyama, Abe et al. 2005; Yin, Straight et al. 2005; Uchiyama and Watanabe 2007; Yin, Straight et al. 2007; Uchiyama and Watanabe 2008) These represent only a few examples of techniques that would prove useful for screening and enriching metagenomic libraries in future studies.

1.3 Molecules and their biosynthetic genes from metagenomic libraries

1.3.1 Phenotypic Screening of Metagenomic Libraries

1.3.1.1 Terragines

A study performed in 2000, which resulted in the isolation of Terragines A-E, represented the first time that a secondary metabolite was functionally identified, heterologously expressed, and structurally characterized from a metagenomic clone. (Wang, Graziani et al. 2000) Earlier reports of metagenomic library construction and phenotypic screening had appeared but none of these studies identified a small molecule and most relied on small insert plasmid libraries which are incapable of functionally capturing the majority of natural product biosynthetic gene clusters. Using DNA isolation methods developed earlier by Davies and co-workers (Yap, Li et al. 1996), cosmid libraries in *E. coli* were constructed from soils collected in British Columbia and Canada. The library cloning vector contained shuttle elements that allowed the transfer of clones from *E. coli* into *S. lividans* via intergeneric conjugation. (Tobias Kieser 2000) HPLC-ESIMS analysis of culture broth extracts from the resulting *S. lividans* transformants revealed two clones that produced clone-specific metabolites. The structural characterization of metabolites found in these extracts resulted in the identification of terragines A-E (**1-5**) in addition to the known microbial

siderophore norcardamine (**6**) (Figure 11). Continued screening of the recombinant library revealed that 18 of approximately 1,000 unique *S. lividans* recombinants were found to produce members of the terragine/norcardamine families. None of the clones were sequenced in this study so the biosynthetic origin of the terragines remains unknown. Although none of the compounds identified in this initial screen displayed any notable biological activity, this work was the first demonstration of accessing the chemistry of uncultured bacteria using metagenomics.

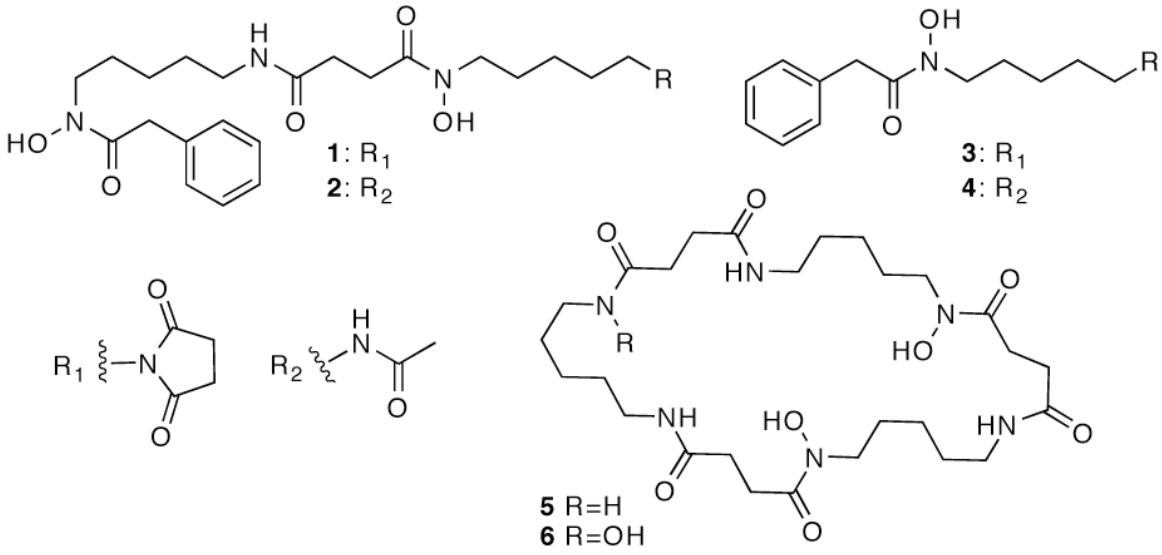


Figure 11: Structures of terragines A-E (1-6)

A cosmid library built in E. coli that was transferred to S. lividans and two recombinants were found to produce these metabolites.

1.3.1.2 N-acyl amino acids

The first novel bioactive small molecules derived from the heterologous expression of an eDNA-derived natural product gene cluster were a family of long chain N-acyl amino acids with antibacterial activity. (Brady, Chao et al.

2001) In this study, approximately 700,000 soil-derived eDNA cosmid clones were functionally screened and of these, 65 exhibited antibacterial properties. In this particular antibiosis assay, the metagenomic library was cultured on agar and a zone of inhibition in a *Bacillus subtilis* overlay indicated the production of a clone-specific metabolite with antibiotic activity. The activity of one of these clones was traced to the production of a family of N-acyltyrosine derivatives with saturated and monounsaturated fatty acid side chains between 8 and 18 carbons in length. Sequencing of the gene cluster and transposon mutagenesis showed that a single open reading frame (ORF) was necessary for the production of the N-acyltyrosines (Genbank No. AF324335). Later subcloning of this ORF indicated that it was also sufficient for the production of this family of secondary metabolites in *E. coli*. This study marked the first systematic exploration of the biosynthesis of a metagenomically-derived small molecule and highlighted the advantage of coupling a natural product with its cloned biosynthetic genes. Neither of these long chain N-acylated antibiotics nor their biosynthetic enzymes (N-acyl amino synthase (NAS)) had been previously reported from cultured bacteria.

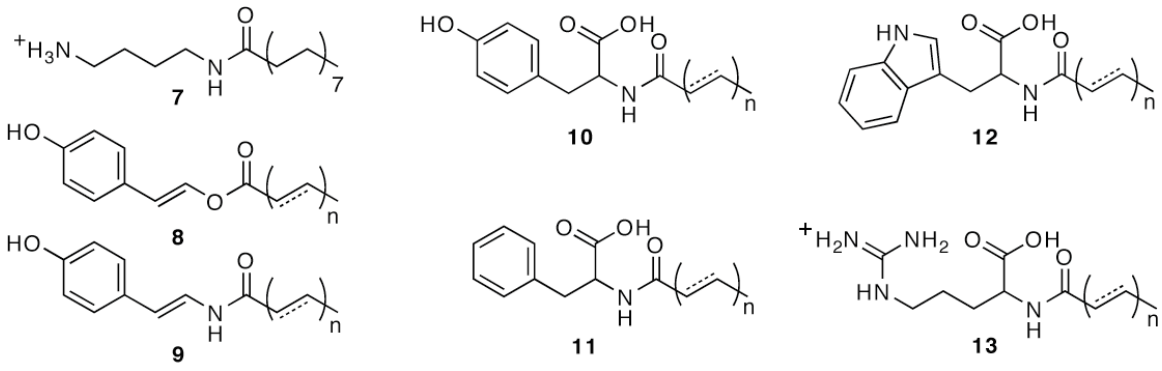


Figure 12: Various N-acyl amino acids from eDNA screening
(Brady and Clardy 2000; Brady, Chao et al. 2004; Brady and Clardy 2005)

Further analysis of the extracts obtained from one of the clones that produced long chain N-acyltyrosines indicated the presence of two additional families of compounds: long chain N-acyl eneamides and long chain N-acyl enol esters (**8-10**). (Brady, Chao et al. 2002) Sequencing of the clone insert revealed that the NAS responsible for the formation of the long chain N-acyltyrosine antibiotics was part of a 13 open reading frame gene cluster (FeeA-M, Genbank No. AY128669) that encodes the biosynthesis of all three families of natural products. The characterization of transposon mutants of this gene cluster that accumulated intermediates allowed the authors to propose the biosynthetic scheme outlined in Figure 13. In the proposed biosynthetic scheme, an acyl carrier protein (either the ACP (FeeL), or the native *E. coli* ACP) is charged with a fatty acid that is transferred to tyrosine by the NAS (FeeM) to generate long chain N-acyl tyrosine intermediates. (Van Wagoner and Clardy 2006) The N-acyltyrosines are then oxidatively decarboxylated by FeeG and the long chain eneamides then undergo an N-O acyl transfer catalyzed by FeeH to yield the corresponding enol esters. Since

the initial characterization of long chain N-acyltyrosine antibiotics, a number of additional antibacterially active structures have been identified from eDNA-derived gene clusters. (Brady, Chao et al. 2004; Brady and Clardy 2004; Brady and Clardy 2005)

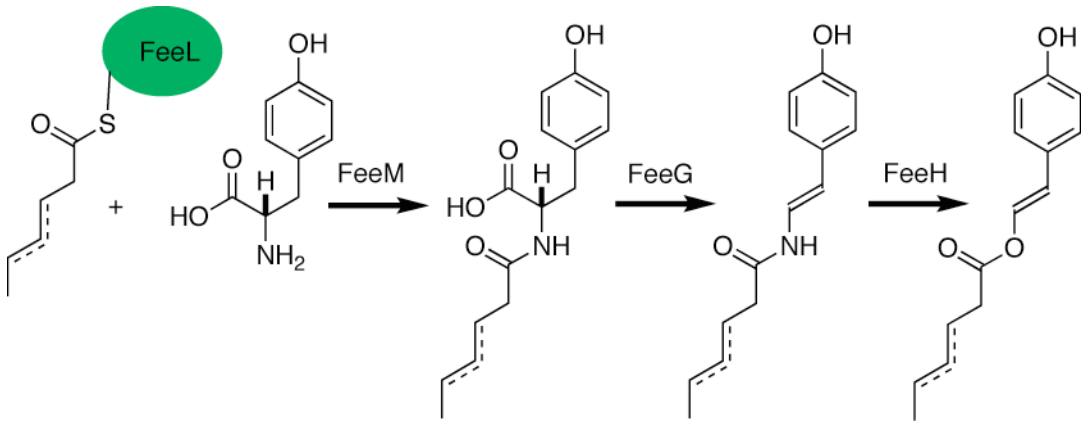


Figure 13: Proposed biosynthetic scheme for long chain fatty acid enol esters

These include clones that produce long chain N-acylphenylalanines (**11**), long chain N-acyltryptophans (**12**) long chain N-acylarginines (**13**), and long chain N-acylputrescines (**7**). (Brady, Chao et al. 2004; Brady and Clardy 2004; Brady and Clardy 2005; Clardy and Brady 2007) In each case, a single ORF was reported to be responsible for the biosynthesis of the metabolites (GenBank Nos. AY214919, AY214920, DQ224236, AY632377). A full-length 16s rRNA gene sequence was found in a gene cluster that conferred the production of long chain N-acylphenylalanines which allowed the authors to determine the phylogenetic origin of the metagenomic insert. While most DNA cloned from the environment cannot be easily taxonomically assigned due to the lack of reference genes, this fortuitous discovery indicated that the

gene cluster was most likely derived from a previously uncultured β-Proteobacterium.

1.3.1.3 Violacein

The appearance of color in a bacterial colony is often an indication of small molecule biosynthesis. Color can therefore be used as a simple indicator for eDNA-derived metagenomic clones that contain natural product biosynthetic gene clusters. In 2001, the visual inspection of a cosmid library derived from soil eDNA revealed several blue-colored *E. coli* colonies. (Brady, Chao et al. 2001) HRMS and NMR analysis of the colored metabolites produced by these colonies identified violacein (**15**) and deoxyviolacein (**14**). These small molecules are tryptophan dimers that were originally isolated from the cultured bacterium *Chromobacterium violaceum*. (August, Grossman et al. 2000) Transposon mutagenesis of the metagenomic clone inserts identified a group of four genes that were responsible for the production of the secondary metabolites (VioA-D, GenBank No. AF367409). The organization of the eDNA-derived violacein biosynthetic gene cluster is identical to that of the violacein gene cluster from the cultured bacterium *Chromobacterium violaceum*. Interestingly, despite conserved gene organization, the genes themselves only show 48, 62, 71, and 69% amino acid identity respectively.

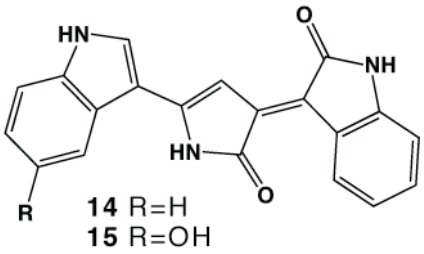
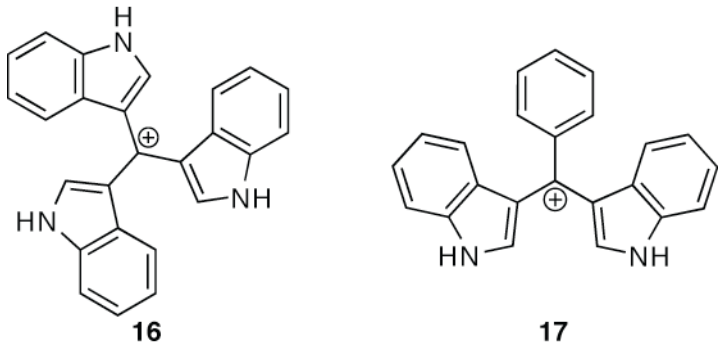


Figure 14: Deoxyviolacein and violacein

1.3.1.4 Turbomycins

In 2002, the visual screening of eDNA libraries constructed in *E. coli* revealed 3 pigmented colonies. (Rondon, August et al. 2000; Gillespie, Brady et al. 2002) The brown pigment produced by each of the clones suggested the presence of melanin, a common secondary metabolite identified in many metagenomic screens. During the characterization of the melanin-like material, two colored metabolites consistently appeared in the acid precipitate at elevated levels compared to the acid precipitates from identically treated vector control cultures. HRMS and NMR analysis of the colored material indicated that the metabolites were a previously reported triaryl cation named turbomycin A (**16**) and a new analog named turbomycin B (**17**). (Rondon, August et al. 2000; Gillespie, Brady et al. 2002) Subsequent transposon mutagenesis and subcloning demonstrated that a single ORF (GenBank No. AF511570), which showed high sequence similarity to members of the 4-hydroxyphenylpyruvate dioxygenase (4HPPD) family of enzymes, was necessary and sufficient for the increased production of the turbomycins. While 4HPPD enzymes are known to catalyze the formation of

homogentisic acid (HGA), which can undergo spontaneous polymerization to HGA-melanin, the biosynthetic origins of the turbomycins in this system remains unclear.



16 17

Figure 15: Turbomycin A and B

1.3.1.5 Isocyanide functionalized antibiotics

The functional screening of a cosmid library derived from soil collected in Massachusetts revealed a clone that produced a novel isocyanide-functionalized indole (**18**). (Brady and Clardy 2005) Saturating transposon mutagenesis of an antibacterially active subclone of the original cosmid indicated that two ORFs, IsnA and IsnB (GenBank No. DQ084328), were required for the production of the compound in *E. coli*. Although the isocyanide functional group had been previously characterized in other bacterial secondary metabolites, its biosynthesis and the source of the constituent atoms were unknown. The cloning and heterologous expression of an isonitrile biosynthetic enzyme in *E. coli* made it possible to perform controlled feeding experiments to uncover the origin of both the nitrogen and

carbon in the isonitrile functional group. (Brady and Clardy 2005) Using "inverse labeling," researchers were able to show that the isocyanide carbon was derived from the C2 carbon of five carbon sugar isomers found in the pentose phosphate pathway. Additional feeding studies suggested tryptophan as the source of the nitrogen. *In vitro* reconstitution experiments using purified IsnA, IsnB, tryptophan, and ribulose-5-phosphate confirmed the putative origin of the isocyanide atoms.

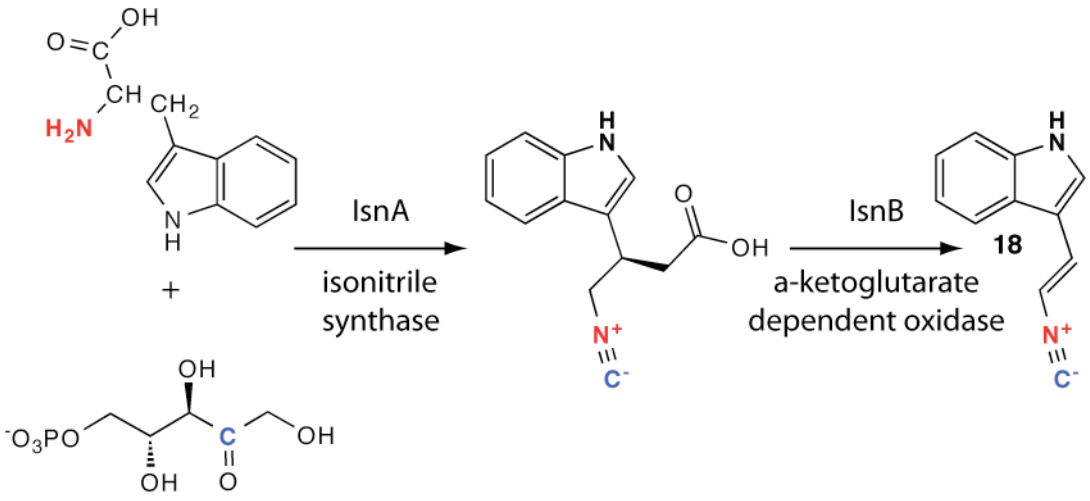


Figure 16: Isonitrile-functionalized indole derivative

Detailed transposon mutagenesis and inverse labeling experiments allowed the elucidation of the biosynthetic scheme leading to the production of (**18**) in *E. coli*. (Brady and Clardy 2005)

In this study, a well executed functional screen allowed for the unbiased identification of novel biosynthetic sequence motifs from a metagenomic library. In a subsequent DNA-based screening study, degenerate primers based on the eDNA-derived IsnA sequence and predicted homologs found in a BLAST search were used to PCR-amplify IsnA-related

sequences from the DNA cloned in four environmental DNA cosmid libraries. (Brady, Bauer et al. 2007) Twelve clones with predicted IsnA homologs were isolated and sequenced to reveal several unique IsnA-containing biosynthetic operons. Each of the IsnA-containing operons was subsequently PCR amplified, cloned into bacterial protein expression vectors and transformed into either *E. coli* or *Pseudomonas aeruginosa* for culturing and expression studies. The analysis of extracts derived from these cultures led to the characterization of 9 clone-specific metabolites. (Brady, Bauer et al. 2007)

1.3.1.6 An antifungal active PKS gene cluster

In 2008, Korean scientists reported the construction of two metagenomic libraries containing approximately 100,000 total clones. (Chung, Lim et al. 2008) Overlay bioassays using *Saccharomyces cerevisiae* provided evidence for clone-specific growth inhibition for one clone. Sequence analysis of the 40 kb insert revealed a biosynthetic gene cluster that appeared to encode a bacterial type II polyketide. Transposon mutants that disrupted the production of antifungal activity fell in the core PKS components (Figure 17). Unfortunately, despite exhaustive isolation and characterization efforts, no small molecules were reported in this work. There have been several reports of functionally active biosynthetic gene clusters that have not yielded a molecular structure. (Nougayrède, Homburg et al. 2006) This highlights just one of the many challenges facing the characterization of heterologously expressed natural products. Based on the

transposon mutagenesis results and the sequence of the gene cluster, the product is presumed to be a polyketide containing at least one nitrogen atom. While this work did not yield a characterized small molecule, the study demonstrated the successful application of a functional screen using a novel assay strain, *S. cerevisiae*.

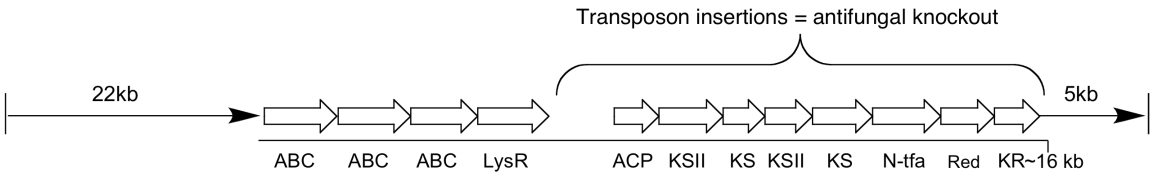


Figure 17: Antifungal PKS gene cluster

Transposon mutagenesis of the antifungal gene cluster indicated that PKS-type biosynthetic genes were required for the observed activity (GenBank No. DQ000460). The site of the transposon insertions and the predicted function of each ORF are shown. *Adapted from* (Kim, Simmons et al. 2009)

1.3.1.7 Indigo/indirubin

Functional screens of metagenomic libraries for small molecule production typically uncover 4 common classes of biosynthetic genes and compounds. These include long chain N-acyl amino acid-producing clones, red-pigmented clones that express aminolevulinic acid synthases (HemA), brown clones that produce melanin-like compounds, and blue clones that produce mixtures of indigo (**16**) and indirubin (**17**). (Gillespie, Brady et al. 2002; Wilkinson, Jeanicke et al. 2002; Brady and Clardy 2005; Huang, Lai et al. 2009) Three separate studies have reported the discovery of either indigo or indirubin-producing clones from metagenomic libraries (Figure 18). A BAC

clone found in a 12,000-membered eDNA library generated from New England soil (GenBank No. DQ000460) contained two predicted indole dioxygenases, each capable of producing the observed pigments. (MacNeil, Tiong et al. 2001) A fosmid clone isolated from a 110,000-membered library constructed from Korean soil (GenBank No. EF569599) contained a single predicted monooxygenase that was responsible for indigo and indirubin production. (Lim, Chung et al. 2005) The third example was a plasmid clone found in a 800,000-membered library generated from the midgut of gypsy moths. (Guan, Ju et al. 2007) This clone contained a two-component flavin dependent monooxygenase system (MoxZ/Y, GenBank No. AR053980). In this case, MoxY was sufficient to produce the small molecules, but MoxZ, a predicted NADH:flavin oxidoreductase, enhanced the color production. Two of these indigo-producing clones were identified in simple colorimetric screens while the third example was discovered using a METREX reporter gene assay designed to detect quorum-sensing mimics produced by clones in metagenomic libraries (Please see section 1.2.5.3). (Guan, Ju et al. 2007) In this study, the clone that produced indigo and indirubin also appeared to produce an additional uncharacterized metabolite that was recalcitrant to isolation and structural elucidation. METREX was also used to discover an AHL quorum sensing inducer that most closely resembles N-(3-oxohexanoyl)-L-HSL. (Williamson, Borlee et al. 2005).

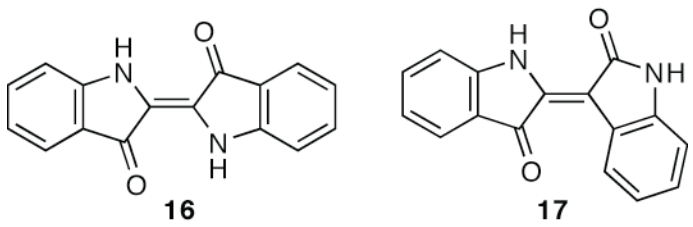


Figure 18: Indigo\indirubin

1.3.1.8 Molecules from Ralstonia metallidurans

The majority of functional metagenomic library screens have utilized either *E. coli* or *Streptomyces* as heterologous expression hosts. This is primarily due to the ease of genetic manipulation and relatively fast growth rate of these bacteria. Recent work in the β-proteobacterium *Ralstonia metallidurans* indicates that expanding the host repertoire for metagenomic natural product discovery will likely increase the number and types of small molecules discovered from eDNA libraries. (Craig, Chang et al. 2009) In this study, phenotypic screens of an eDNA library hosted in *R. metallidurans* yielded 2 recombinants which produced clone-specific metabolites. One clone (RM3) was identified with simple visual inspection as it produced a yellow pigment, and the other (RM57) was identified based on antibiotic activity. Upon sequencing, clone RM3 (GenBank No. FJ151553) was found to contain a 6-gene operon with homology to carotenoid gene clusters. The yellow metabolite from this clone was spectroscopically identical to β-carotene. Sequencing of clone RM57 (GenBank No. FJ151552) did not reveal an obvious source of antibiotic activity. Transposon mutagenesis of cosmid RM57 indicated that antibiotic activity was dependent on a type III polyketide

synthase. A total of 7 small molecules were subsequently isolated from cultures of clone RM57. Compounds (**20-25**) are long- and short-chain substituted resourcinol-like compounds while compound (**24**) is a pyrone heterodimer with both saturated and monounsaturated side chains. Compound (**25**) was a novel tricyclic isocoumarin-based scaffold. Both compounds (**24**) and (**25**) displayed antibacterial activity against *Staphylococcus aureus* and *Bacillus subtilis* (Figure 19).

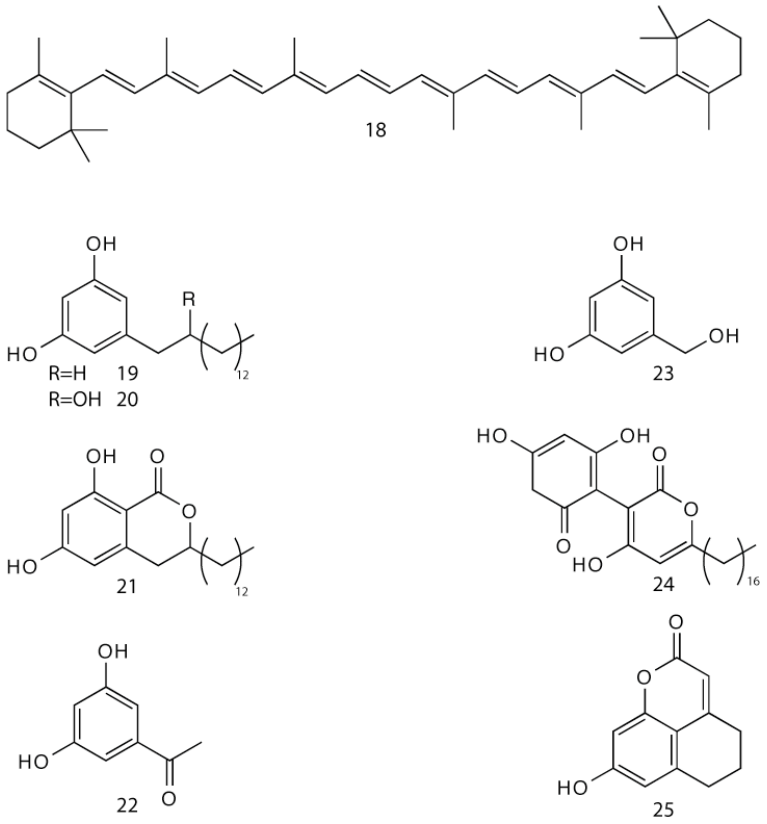


Figure 19: *Ralstonia metallidurans*-based libraries

Both known (18-23) and novel (24-25) secondary metabolites were characterized from clones found in phenotypic screens of eDNA libraries hosted in *R. metallidurans*. (Craig, Chang et al. 2009)

1.3.1.9 Molecules from diverse proteobacteria

Additional natural products from metagenomic libraries were recently uncovered using a diverse panel of genetically tractable proteobacterial hosts. (Craig, Chang et al. 2010) The authors of this study selected a panel of expression hosts based on phylogenetic analyses which indicate that proteobacteria are often well represented in soil microbiomes. After isolating eDNA using established methods, the authors first generated three libraries between 130,000-450,000 clones in size using a cosmid vector with broad host range elements (pJWC1) in *E. coli*. This vector allowed the stable propagation of eDNA inserts in a wide range of bacterial hosts. After isolating metagenomic library DNA from *E. coli*, the libraries were transferred into *Agrobacterium tumefaciens, Blumeria graminis, Caulobacter vibrioides*, and *Ralstonia metallidurans* by electroporation, and into *Pseudomonas putida* by biparental conjugation. (Tobias Kieser 2000) Simple colorimetric, morphological, and antibiosis overlay assays were then used to identify clone-specific features. The detailed analysis of culture broth extracts from a panel of candidate clones identified 2 novel metabolites (**26-27**), myristoylputrescine and an N-acylated derivative of 4-amino-2-hydroxybutamine (Figure 20). To demonstrate the necessity of expanding the host range for metagenomic natural product discovery efforts, the authors isolated cosmids from the small molecule producing hosts and transferred them to other hosts used in the study. Screening these shuffled host-cosmid

pairs revealed that only certain combinations of hosts and gene clusters could be detected in functional assays. Organic extracts of seemingly inactive and non-producing host-cosmid pairs did reveal, however, that low levels of the predicted small molecules were being generated in many cases. This indicates that there is potentially a large number of small molecule-producing eDNA clones which fall below the threshold of detection in commonly used hosts, and that developing novel phylogentically diverse heterologous expression hosts will most likely lead to the continued discovery of novel small molecules.

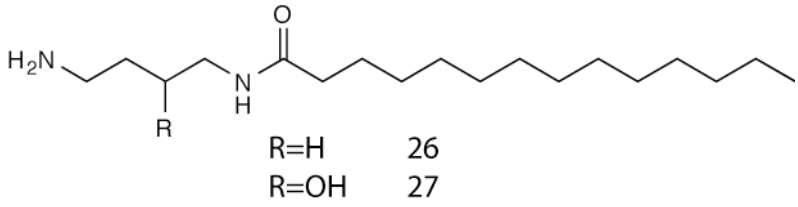


Figure 20: Myristoylputrescine (26) and 1, 3-hydroxy myristoylputrescine (27)
Compounds 26 and 27 were discovered in a phenotypic screen of diverse proteobacteria utilizing a novel broad host-range cosmid vector.

1.3.2 Sequence-Based Screening of Metagenomic Libraries

1.3.2.1 Type II PKS KSβ from soil DNA

The first functional exploration of natural product biosynthesis in uncultured bacteria involved the PCR amplification and heterologous expression of soil-derived type II polyketide synthase (PKS) genes. (Seow, Meurer et al. 1997) Type II PKS gene clusters are responsible for the

biosynthesis of a large number of structurally diverse aromatic polyketides and are found in phylogenetically diverse bacteria. Minimal type II PKS systems are comprised of three core proteins: two β-ketoacyl synthases (KSα and KSβ) that catalyze sequential Claisen-condensation reactions and control the polyketide chain length, and one acyl carrier protein (ACP) that provides a covalent anchor for the nascent polyketide during the chain elongation (Figure 21).

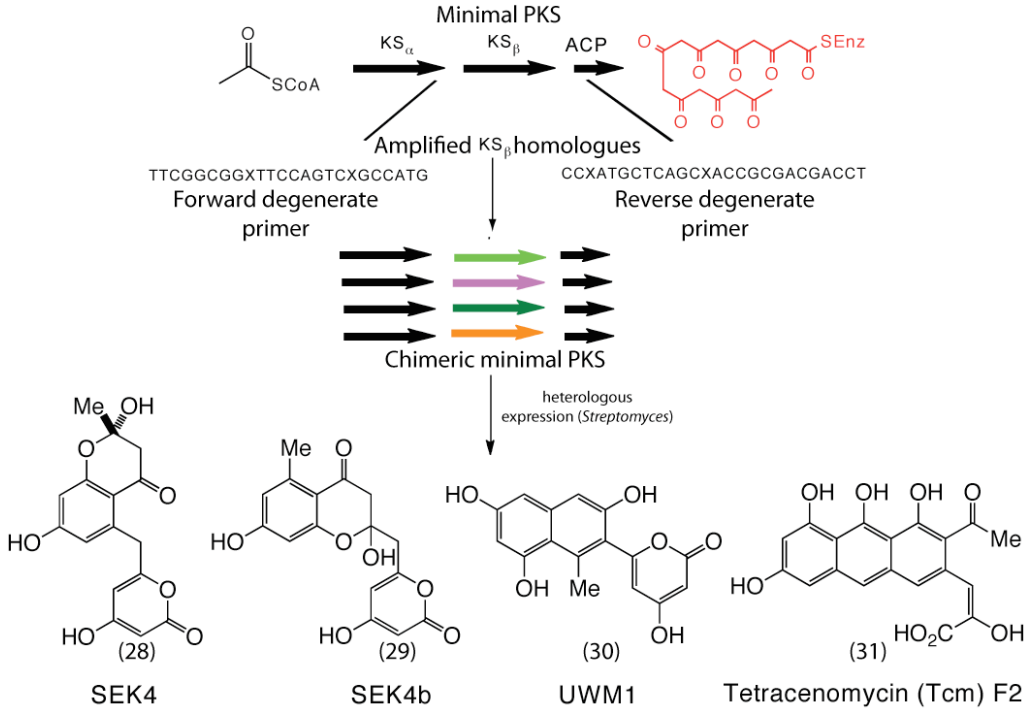


Figure 21: PCR-based screening for novel natural products (type II PKS)

Degenerate PCR primers designed to recognize a conserved region in KS_{α} sequences and a conserved sequence in the downstream ACP were used to amplify full-length KS_{β} sequences from environmental DNA. The amplified sequences were cloned into minimal PKS systems and introduced into *streptomyces* for heterologous expression studies. Four metabolites (**28-31**) were uncovered using this strategy. (Seow, Meurer et al. 1997) *Adapted from* (Kim, Simmons et al. 2009)

In many actinomycete-derived type II PKS gene clusters, the minimal PKS genes are organized with the KSα and ACP genes on either side of the KSβ gene (KSα-KSβ-ACP). With this knowledge, researchers designed a strategy to access novel type II polyketide biosynthetic enzymes from environmental samples using degenerate PCR.

Using KSα and ACP-specific primer sets, PCR amplicons of the correct predicted size were obtained from DNA isolated directly from soil. Two unique sequences that showed similarity to known KSβ genes were cloned from the eDNA-derived amplicon. To analyze the function of these genes in detail, hybrid PKS cassettes were constructed by replacing the KSβ gene from an existing minimal PKS with one of the two eDNA-derived sequences. The heterologous expression of the hybrid constructs in two *streptomycetes* yielded several octa- and decaketide metabolites (**28-31**). While compounds **28-31** had been previously described (Fu, Hopwood et al. 1994), this study also revealed two compounds that appeared novel; however, their molecular structures were not determined due to insufficient material. Based on chromatographic and spectroscopic analyses, one of the novel metabolites appeared to be related to SEK4 (**28**) and SEK4b (**29**). The other uncharacterized metabolite did not appear to resemble any of the observed compounds and was presumed by the authors to be novel. While no new metabolites were fully characterized in this study, it does represent the first example of the heterologous

expression of novel biosynthetic genes from metagenomic DNA to yield small molecules.

PCR-based discovery strategies are generally limited to single genes or small fragments of gene clusters which are typically not sufficient to produce novel metabolites as biosynthetic pathways often contain multiple enzymes. Davies and colleagues did perceptively note, however, that this approach could be extended and that sequences amplified using degenerate PCR could be used as tags to recover the rest of a gene cluster from cosmid or BAC libraries generated from the same metagenomic sample. (Seow, Meurer et al. 1997) All subsequent sequence-based metagenomic screens for natural product gene clusters would follow this model exactly.

1.3.2.2 Type I PKS clones

In 2003 a research group screened a 5,000-member soil-derived eDNA cosmid library using functional and sequenced-based techniques. (Courtois, Cappellano et al. 2003) Clones from this library were assayed for antibiotic activity and the DNA isolated from these clones was screened with PCR primers targeting both 16s rRNA and type I PKS genes. 47 unique 16s sequences, many of which represented new candidate species, were identified in this relatively small library. Using degenerate PCR primers designed against conserved regions flanking PKS type I ketoacyl synthetase (KS) domains, researchers were able to identify 11 unique sequences resembling known ketosynthases in the library. Three of the KS sequences were found to

arise from a single cosmid insert. This cosmid was fully sequenced revealing 6 ORFs. While the gene cluster appeared to be truncated, the information suggested that other clones identified in the ketosynthase PCR screen could contain natural product gene clusters that may yield novel secondary metabolites.

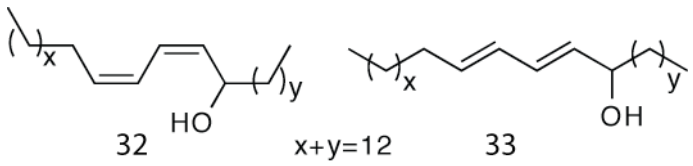


Figure 22: Dienic alcohols

The initial library was constructed using a shuttle vector (pOS700I), which allowed clones to be mobilized from *E. coli* into *S. lividans* for phenotypic screening. Each clone that was found to contain a KS sequence in the degenerate PCR screen was introduced by conjugation into *S. lividans* TK24. (Tobias Kieser 2000) No antibacterial activity was detected in a screen of the *S. lividans* clones. Extracts derived from culture broths of the recombinant *S. lividans* were subsequently analyzed by HPLC for the presence of clone specific small molecules. Two recombinants that appeared to produce the same mixtures of metabolites were identified and selected for further examination. Novel aliphatic dienic alcohol isomers **(32)** and **(33)** were subsequently isolated and characterized from culture extracts of these recombinant *S. lividans* clones (Figure 22). (Courtois, Cappellano et al. 2003)

1.3.2.3 Type II PKS gene clusters

In 2009 and 2010, researchers extended the scope of Seow, et al.'s seminal type II PKS study by screening cosmid libraries using degenerate primers which target the same minimal PKS modules. (King, Bauer et al. 2009; Bauer, King et al. 2010) Although the diversity of soil microbiomes was appreciated in the earlier type II PKS study, later experiments would reveal a staggering number of bacterial species in terrestrial samples that far exceeded any initial estimates. With this in mind, the authors of this study constructed a cosmid library of ~$1x10^7$ clones, approximately two orders of magnitude larger than any other cosmid library generated from a metagenomic sample. The authors amplified and sequenced 21 unique KSβ sequences from this cosmid library. Only one of these showed greater than 80% identity to a previously reported KSβ gene. From here, a panel of ten cosmid clones containing type II PKS gene clusters were isolated from the library using the PCR amplified sequences as recovery probes. The cosmid clones containing the minimal PKS genes were purified from *E. coli*, retrofitted with conjugative mating elements for *Sreptomyces*, and mobilized into *S. lividans* and *S. albus* for heterologous expression studies. (Tobias Kieser 2000) Several recombinants appeared to generate clone-specific metabolites as they produced color and one high-producing clone was selected for detailed investigation. The major component of an ethyl acetate extract of

this clone (V167) was isolated, structurally characterized and reported as Erdacin (**34**) (Figure 23).

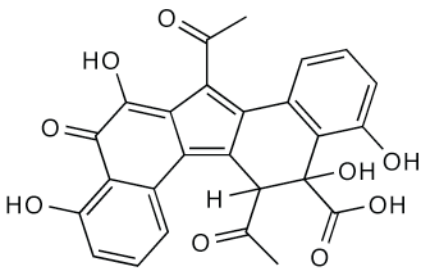


Figure 23: Erdacin

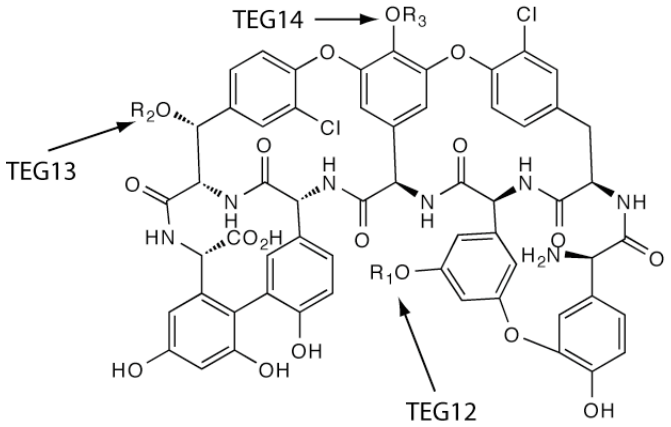
Sequenced-based screens were used to isolate compound 34 from a 1x10^7 member cosmid library constructed from soil collected in Utah.

Detailed transposon mutagenesis of the gene cluster and labeled precursor studies helped the authors propose a biosynthetic scheme which involves a common octaketide precursor that undergoes two distinct cyclizations to form the two halves of Erdacin. The continued characterization of the Erdacin gene cluster (ERD) from this primary screen also allowed the discovery of Utahmycins A and B, two azaquinones produced by the environmental DNA clone. Erdacin represents the first example of heterologously expressing a relatively complex secondary metabolite from a single eDNA-derived clone.

1.3.3 Tailoring Enzymes from eDNA

In addition to core biosynthetic enzymes, bacterially derived natural product gene clusters often contain a range of resistance and tailoring enzymes that can modify a core natural product chemical scaffold. With this

in mind, researchers set out in 2008 to isolate gene clusters that contain enzymes which could confer the production of novel congeners of the well characterized glycopeptides vancomycin and teicoplanin. (Banik and Brady 2008) Using an oxidative coupling enzyme (OxyC) involved in the cyclization of the heptapeptide backbone of these non ribosomal peptides, the authors were able to identify two novel homologues in a $1x10^7$ member cosmid library. They subsequently isolated portions of gene clusters encoding what appeared to be novel congeners of this important class of antibiotic. Sequencing of the two pathways TEG and VEG (teicoplanin-like and vancomycin-like eDNA derived gene cluster: GenBank Nos. EU874253, EU874252) revealed a group of novel sulfotransferases in the TEG pathway. Although this pathway was incomplete and no overlapping cosmids could be identified in the metagenomic library, the TEG pathway sulfotransferases provided an opportunity to modify the known teicplanin scaffold in a unique way. To achieve this, the enzymes were subcloned, recombinantly expressed and purified, and used *in vitro* to sulfate the teicoplanin aglycone in three distinct regions (**36-42**). Although the authors were unable to isolate clone-specific molecules from either pathway, this work demonstrated the potential for the combinatorial biosynthetic diversification of existing natural product scaffolds using enzymes found in metagenomic libraries. (Banik and Brady 2008)



	R_1	R_2	R_3
Aglycone (35)	H	H	H
A (36)	SO_3H	H	H
B (37)	H	SO_3H	H
C (38)	H	H	SO_3H
D (39)	SO_3H	SO_3H	H
E (40)	SO_3H	H	SO_3H
F (41)	H	SO_3H	SO_3H
G (42)	SO_3H	SO_3H	SO_3H

Figure 24: Novel heptapeptide (teicoplanin) congeners from eDNA

The sulfo-teicoplanins A-G (**36-42**) were produced *in vitro* from the aglycone (**35**) using three recombinantly expressed and purified sulfotransferases from an eDNA-derived biosynthetic pathway (TEG). *Adapted from* (Banik and Brady 2008)

1.4 Future Challenges and Outlook

The efforts described in this chapter clearly demonstrate that uncultured bacteria are likely to be a rich source of novel biologically active small molecules. However, the systematic exploration of natural product diversity in uncultured bacteria still faces significant challenges. These include 1) the need for more robust strategies that can be used to efficiently clone and stably maintain very large insert libraries (>50 kb) in a more diverse collection of bacterial hosts; 2) the development of novel genetically tractable bacterial hosts from diverse phylogenetic origins; 3) innovative high throughput assays to enrich biosynthetic sequences from complex metagenomic libraries; 4) novel assays to detect the production of small quantities of molecules; 5) scalable solutions to produce large quantities of heterologously expressed secondary metabolites; 6) novel methods to

manipulate and engineer biosynthetic gene clusters; and 7) a more fundamental understanding of where biosynthetic gene clusters are phylogenetically derived from and how they are regulated. The subsequent results in this thesis begin to address many of these hurdles, but a complete solution will require the coordinated efforts of multiple disciplines.

CHAPTER 2

2 Mining the Metagenome for Natural Product Diversity

2.1 Introduction

The experiments outlined in Chapter 1 clearly demonstrate the potential of uncultured bacteria as a rich source of chemical diversity. Many of these experiments utilized functional screens of soil-based eDNA libraries to discover novel metabolites from uncultured bacteria (Table 1). Of the advances that have been made since the earliest experiments outlined in Chapter 1, the most important was arguably the refinement of high molecular weight eDNA isolation and cloning methods from soil metagenomes by Brady et al.. (Brady 2007; Banik and Brady 2008; King, Bauer et al. 2009; Kim, Feng et al. 2010) With simple functional screens, Brady and coworkers demonstrated that additional new classes of secondary metabolites can be discovered from eDNA libraries of sufficient size. (Brady and Clardy 2000; Brady, Chao et al. 2002; Brady, Chao et al. 2004; Brady and Clardy 2005; Brady, Bauer et al. 2007; Craig, Chang et al. 2009; Craig, Chang et al. 2010) As described earlier, however, functional screens can be hindered by the challenges of heterologous expression and cloning limits which often yield truncated natural product gene clusters. This prevents the

discovery of many biosynthetic pathways using functional assays regardless of how large a library is. We hypothesized that sequence-based methods would provide a general platform for discovering a more diverse range of biosynthetic clones as this approach does not rely on functional expression and can be used to recover larger gene clusters from groups of overlapping clones. In the course of early experiments, we discovered that eDNA libraries needed to be much larger in size to recover complete natural product gene clusters using sequence-based screens. We subsequently designed a framework that would allow us to systematically access overlapping clones that constitute complete biosynthetic gene clusters from soil metagenomes. In order to accomplish this we needed to 1) determine the size a library needed to be in order to capture overlapping cosmids comprising a complete biosynthetic pathway, 2) construct libraries that were large enough to ensure the recovery of complete gene clusters, and 3) demonstrate access to a diverse range of large biosynthetic genes clusters using simple sequence-based screens. The results presented in this chapter describe successes along these aims.

2.2 Results

2.2.1 Library Cloning and Transduction

The direct extraction of high molecular weight DNA (40-50 kb) from soil microbiomes has recently become established. (Brady 2007) Despite general correlations of bacterial diversity with factors such as pH, salinity,

heavy metal content, and temperature, there have been no systematic studies of natural product chemical diversity in environmental samples. (Please see Chapter 4 for experiments which begin to address this) Based on general knowledge of the phylogenetic diversity of terrestrial microbiomes, we set out to create a panel of crude eDNA extracts from a variety of soils, including multiple desert and arid environments, topsoils, and cryptobiotic crusts. It is difficult to access the genomes of rare members of bacterial communities using metagenomic approaches due to the uneven phylogenetic distribution of most microbiomes. (Daniel 2005) One potential solution to this problem is the pre-normalization of the population prior to library construction. We attempted basic experiments along this aim using two differential DNA stains (bis-(6-chloro-2-methoxy-9-acridinyl)spermine (AT-rich), and actinomycin D (GC-rich)) and whole-cell bacterial populations isolated using density centrifugation through a Nycodenz™ polymer matrix.(Rickwood, Ford et al. 1982; Gellert-Mortimer, Clarke et al. 1988; Courtois, Frostegård et al. 2001) From here, we used FACS sorting and 16s rRNA restriction-fragment length polymorphism (RFPL) profiling to show that we could enrich populations based on differential DNA staining (data not shown). Other studies have used a combination of 16s rRNA-specific FISH probes and FACS to successfully isolate bacterial subgroups from complex microbiomes.(Kalyuzhnaya, Zabinsky et al. 2006) Unfortunately, neither of these methods yields sufficient material for the construction of large eDNA

libraries due to the limitations of FACS throughput. The continued refinement of approaches like these could aid in accessing rare members of bacterial communities in future experiments.

Several methods for eDNA extraction were attempted during preliminary studies including combinations of heat, enzymatic treatments, high-speed grinding, freezing and thawing, chemical lysis, and whole bacterial cell purification. Of these methods, the strategy outlined below consistently yielded the highest quantity and quality of environmental DNA that was suitable for the efficient construction of large cosmid libraries. (See Materials and Methods) First, approximately 250-500 grams of soil were heated in lysis buffer containing EDTA (ethylenediaminetetraacetic acid), detergents, and CTAB (cetyl trimethylammonium bromide). Large soil particulates were removed via centrifugation, and DNA was precipitated in crude form with isopropanol. Crude eDNA was then further purified using gel electrophoresis and electroelution. In general, this strategy typically yielded crude eDNA that was approximately 45-50 kb in length. This made further processing of the eDNA unnecessary as it is the ideal size for cosmid transduction based on phage packaging limitations. (Brady 2007) Other eDNA extraction methods generally yielded either insufficient quantities of crude eDNA or lower molecular weight eDNA that was unsuitable for cosmid cloning. After assessing the quality of eDNA extraction via gel electrophoresis, eDNA samples were end repaired using the concerted

activities of T4 polynucleotide kinase (PNK) and T4 DNA polymerase, ligated to a cosmid cloning vector, and transduced into *E. coli* with lambda phage (Figure 9). Transduction reactions were titered to determine the packaging efficiency and libraries were subarrayed into aliquots to facilitate downstream screening and clone isolation efforts from the mixed pools of cosmids. Each unique pool in a subarray was subsequently stored as glycerol stocks, and DNA was purified for screening purposes. With this method, we were able to routinely generate recombinant eDNA libraries containing millions of clones and >100,000 genome equivalents of DNA (considering an average bacterial genome of approximately 4MB in size). Several smaller eDNA libraries were constructed using slight variations of this method (lysozyme, proteinase k, SDS (sodium dodecyl sulfate), and grinding for example) and were used for early sequence-based screens. (As described in section 2.2.3) Three soil samples collected in Utah, Anza Borego California (AB), and Arizona (ARZ) were utilized for the majority of subsequent screens because eDNA extracted from these samples allowed the efficient construction of libraries containing 1-1.5x10^7 clones. These represent three of the largest recombinant eDNA libraries reported to date and were a critical development in natural product gene cluster discovery efforts as demonstrated in later sections.

Table 2: Primary eDNA libraries constructed and screened in these studies

1: (Kim, Feng et al. 2010), 3: (Banik and Brady 2008; King, Bauer et al. 2009; Bauer, King et al. 2010; Kim, Feng et al. 2010)

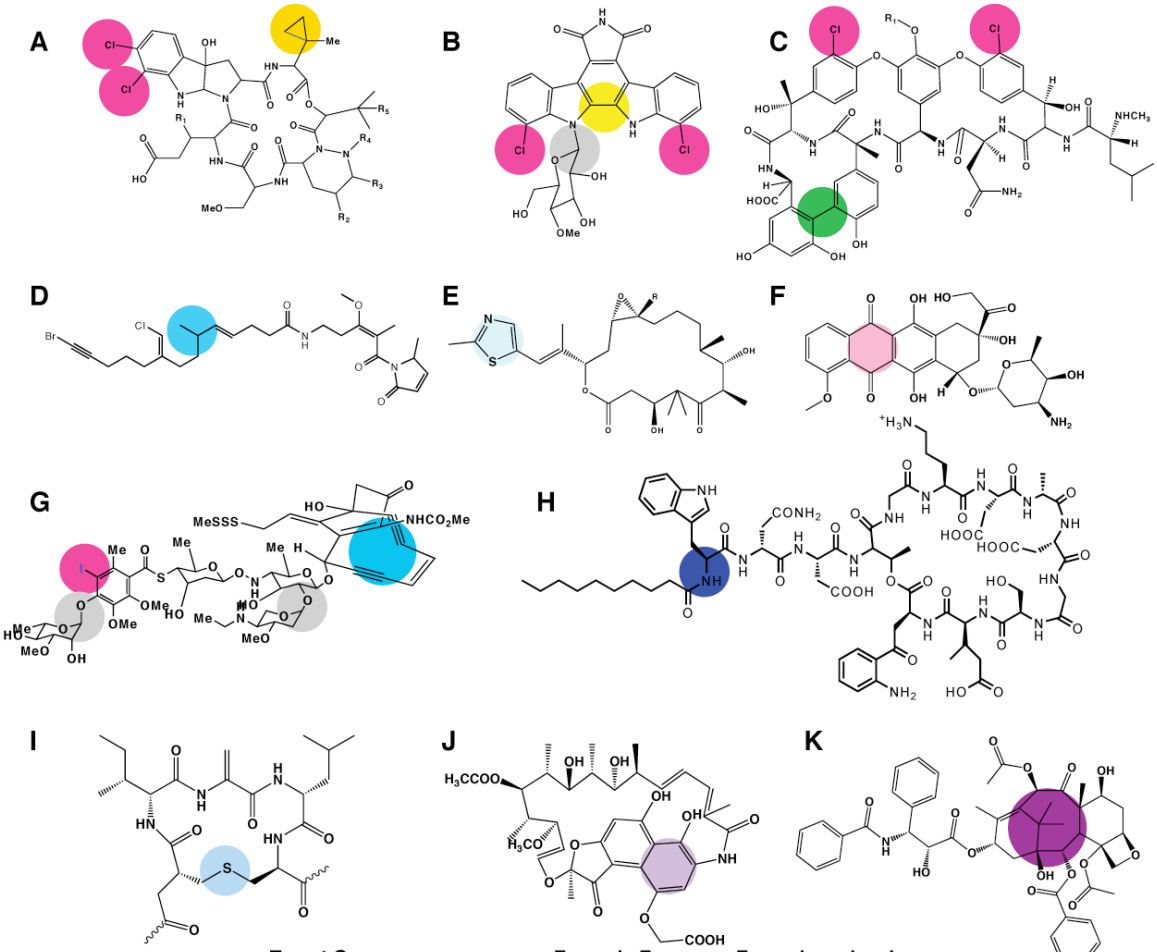
Library	Vector	Insert Size (kb)	Total Size (clones)	Sequence size (Gb)	References
Anza Borego	Cosmid	35-40	1.5x10^7	525	1
Arizona	Cosmid	35-40	1.5x10^7	525	This work
Utah	Cosmid	35-40	1.0x10^7	350	3

Based on the amount of crude eDNA isolated from the Arizona, Anza Borego, and Utah samples the resulting libraries could, in fact, be expanded further but practical storage and handling limitations impede efforts to build libraries much larger in size. While eDNA libraries containing ~1-1.5x10^7 clones still do not exhaustively sample the bacterial diversity present in the soil, we hypothesized that libraries of this size should provide access to a diverse range of biosynthetic gene clusters.

Many natural product gene clusters are too large to be routinely captured on individual cosmid clones. With metagenomic libraries of sufficient size and sequence coverage, large gene clusters that cannot be captured on a single cosmid insert could be accessed by recovering collections of overlapping eDNA clones. Although larger-insert cloning systems do exist (BAC), they are typically several orders of magnitude less efficient than cosmid/fosmid based systems and the routine isolation of higher molecular weight (>50 kb) eDNA from soils has been historically challenging. (Liles,

Williamson et al. 2008) Cosmids, in conjunction with direct eDNA extraction methods outlined here, therefore currently represent the most efficient method of capturing genomic DNA from soil microbiomes for natural product discovery.

2.2.2 First Generation Sequence-Based Screens



Target Group	Example Enzyme	Example molecule
Terpene Cyclase	TxnC	Taxol (K)
Ketosynthase (PKS)	KS	Jamaicamide (D)
Adenylation Domain (NRPS)	AD	Daptomycin (H)
FAD Halogenase	ComH, PrnA, RebH	Complestatin (C)
bNOS	Nos	Resistance Mechanism
indolocarbazole cyclase	RebD, RebC	Staurosporine (B)
TII PKS	DpsA	Doxorubicin (F)
thiazoles	EpoA,B	Epothilone (E)
lantibiotic cyclases	SpaC	Subtilin (I)
AHBA synthases	RifK	Rifamycin (J)
Glycosyltransferase	various	Calicheamycin (G)
Acyl-ACP ligase	various	Daptomycin(H)

Figure 25: Biosynthetic functional groups and modifications found in various secondary metabolites.

Chemical features generated by biosynthetic enzymes are colored according to the legend. Example molecules include A) kutznerides, B) staurosporine, C) complestatin, D) jamaicamide A, E) epothilone B, F) doxyrubicin, G) calicheamycin, H) daptomycin, I) calicheamycin, J) rifamycin, and K) taxol.

In excess of 35,000 unique microbial natural products have been characterized using culture-based discovery methods. (Laatsch 2009) Although microbial genomic sequencing efforts are becoming routine, only a small percentage of these secondary metabolites have been directly associated with the biosynthetic genes which catalyze their formation. Those gene clusters that have been sequenced, however, indicate that the structural diversity seen in natural products appears to arise in large part from the natural combinatorial shuffling of a much smaller set of conserved biosynthetic enzymes (Figure 25). Sequencing has also revealed that biosynthetic genes are often found clustered on bacterial genomes along with regulatory, tailoring, resistance, and export proteins. Degenerate PCR primers designed to recognize conserved regions in natural product biosynthetic genes are therefore useful for identifying novel biosynthetic gene clusters that catalyze the formation of a diverse collection of small molecules (Figures 26-27). In this basic strategy, a sequence alignment of known biosynthetic homologues is used to find regions of nucleotide conservation that can be used to design degenerate PCR primers. These primers can subsequently be used to amplify and detect novel sequence variants contained within cosmids from eDNA libraries. From here, these sequences can be used as tags to isolate individual cosmids containing a biosynthetic gene cluster from eDNA libraries (Figure 26).

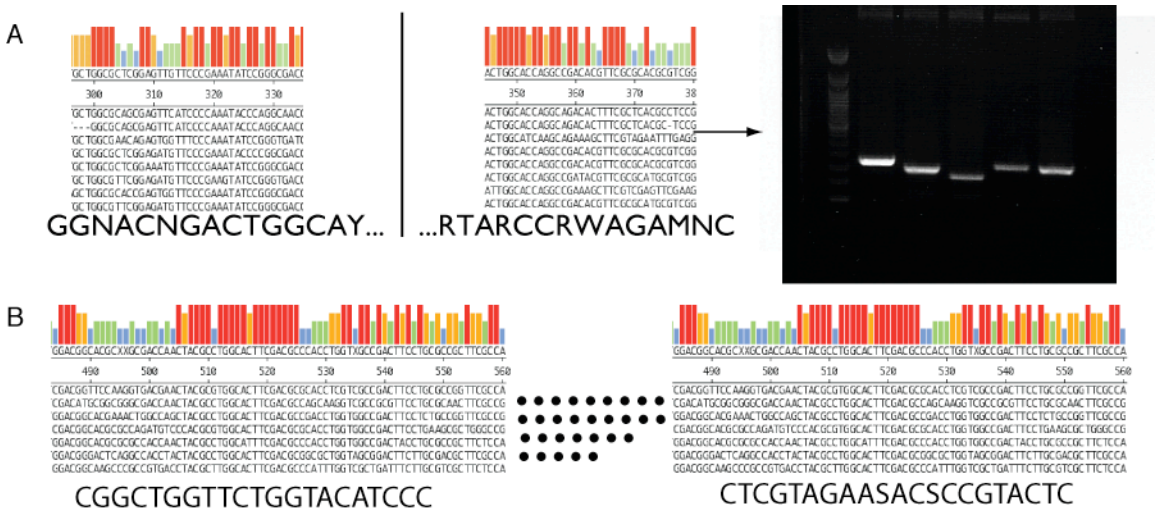


Figure 26: Degenerate primer design strategy

A multiple sequence alignment of known homologues (α-KG halogenases (A) and FAD dependent aromatic halogenases (B)) was used to design degenerate PCR primers that target conserved sequences. The degenerate primers can then be used to amplify novel sequences from environmental samples or eDNA libraries for gene cluster recovery efforts.

In addition to core biosynthetic enzymes (Figure 6), biosynthetic gene clusters often contain tailoring enzymes which modify natural product scaffolds (Figure 25). By targeting either core biosynthetic enzymes found in many pathways (PKS/NRPS) or tailoring enzymes involved in the biosynthesis of specific classes of metabolites, it is possible to screen for a diverse range of gene clusters (Figure 27). To demonstrate the versatility of sequenced-based screening methods, we identified and recovered numerous biosynthetic gene clusters encoding core biosynthetic systems, a broad subgroup of halometabolites, and class-specific lipopeptides using PKS, halogenase, and Acyl-ACP ligase degenerate primers respectively.

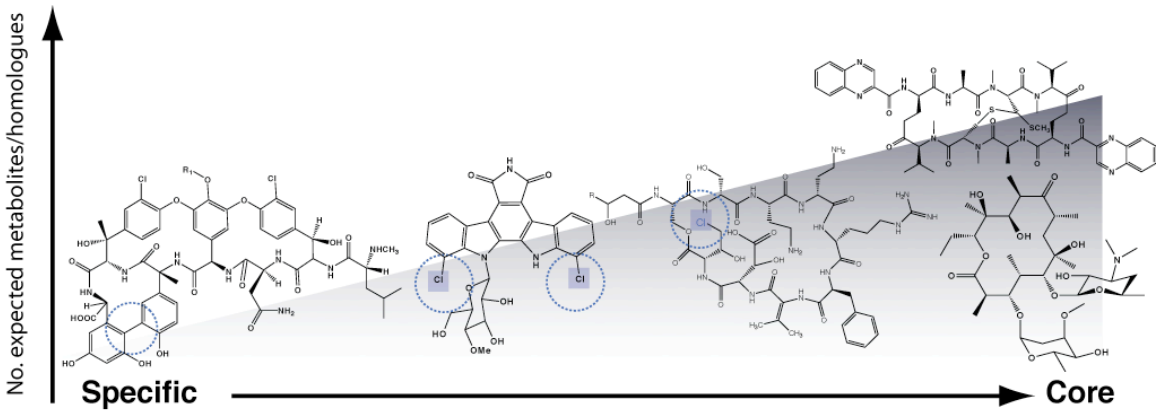


Figure 27: Sequence-based screening of biosynthetic features

Class-specific or core biosynthetic enzymes can be targeted to recover a range of biosynthetic gene clusters using sequenced-based methods. Examples of enzymes and modifications discussed in this work are circled in blue. (OxyC oxidative coupling enzymes, FAD-mediated halogenases, aKG-halogenases)

2.2.2.1 Halogenases

There are currently more than 5,000 structurally distinct halometabolites which have been isolated from phylogeneticially diverse bacteria of both marine and terrestrial origin. Halogenations are found in more than 50% of high-throughput screening compounds, more than 70% of bioactive natural products, and they can play a significant role in conferring bioactivity by increasing lipophilicity and cell permeability (Figure 28). (Hornung, Bertazzo et al. 2007) Halogenases are also easily distinguished from primary metabolic enzymes, making them ideal screening targets to isolate novel secondary metabolite gene clusters from eDNA libraries. (Hornung, Bertazzo et al. 2007)

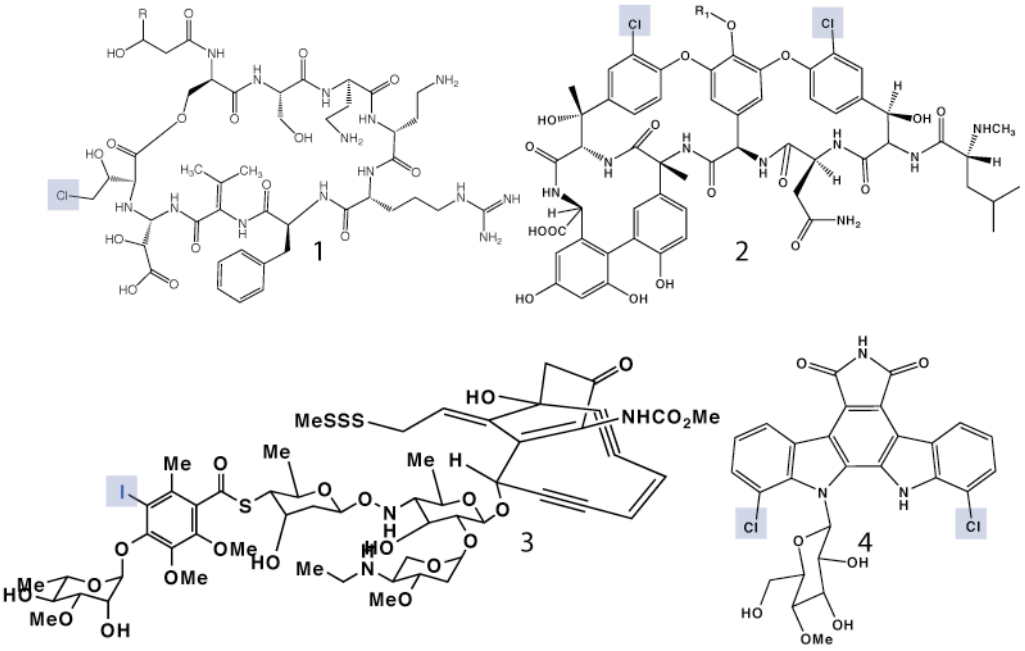


Figure 28: Various halometabolites

Example halometabolites with their respective sites of halogenation (blue) are shown. Syringomycin utilizes an α-KG dependent aliphatic halogenase (1) while complestatin (2), calicheamycin (3), and rebecamycin (4) utilize FAD-mediated aromatic halogenases.

From a functional standpoint, these enzymes can provide insight into the mechanisms nature has evolved to incorporate halogens into often unreactive substrates. (Neumann, Fujimori et al. 2008) The factors that this transformation normally requires using synthetic strategies include high heat, a catalyst, and tight spatial regulation of the reactive site making it difficult to rationalize *in vivo*. However, recent studies have proposed at least three enzymatic mechanisms. (Blasiak, Vaillancourt et al. 2006) The first class of halogenases to be characterized were the haloperoxidases. Haloperoxidases generate a free hypohalous acid in solution that targets

aromatic carbons on a substrate with the aid of hydrogen peroxide and a heme or vanadium cofactor. Many secondary metabolites, however, display regioselective or aliphatic halogenations. These modifications are carried out by α-ketoglutarate (α-KG) and FAD-mediated halogenases (Figure 30). (Blasiak, Vaillancourt et al. 2006) α-KG halogenases use α-KG and iron to form a Cl-Fe(IV)-oxo species which abstracts a hydrogen from the substrate, forming a radical and Cl-Fe(III) intermediate. Oxidative transfer of Cl- to the substrate radical catalytically reforms Fe(II) for the next round of halogenation (Figure 29a). FAD halogenases use flavin as a cofactor to site-specifically halogenate aromatic carbons on tyrosyl or tryptophan moieties. FAD-dependent halogenases generate a flavin hydroperoxide (-OOH) intermediate that is used to form hypohalous acid when a halide attacks the distal oxygen. These enzymes covalently trap the reactive hypohalous acid on a lysine in the active site, however, in order to sterically position the reaction (Figure 29b). Recombinantly expressed halogenases have been used to diversify natural product scaffolds *in vitro* and novel homologues could provide general solutions to overcome site-specificity and reactivity in synthetic halogenation reactions. (Weissman and Leadlay 2005; Savile, Janey et al. 2010)

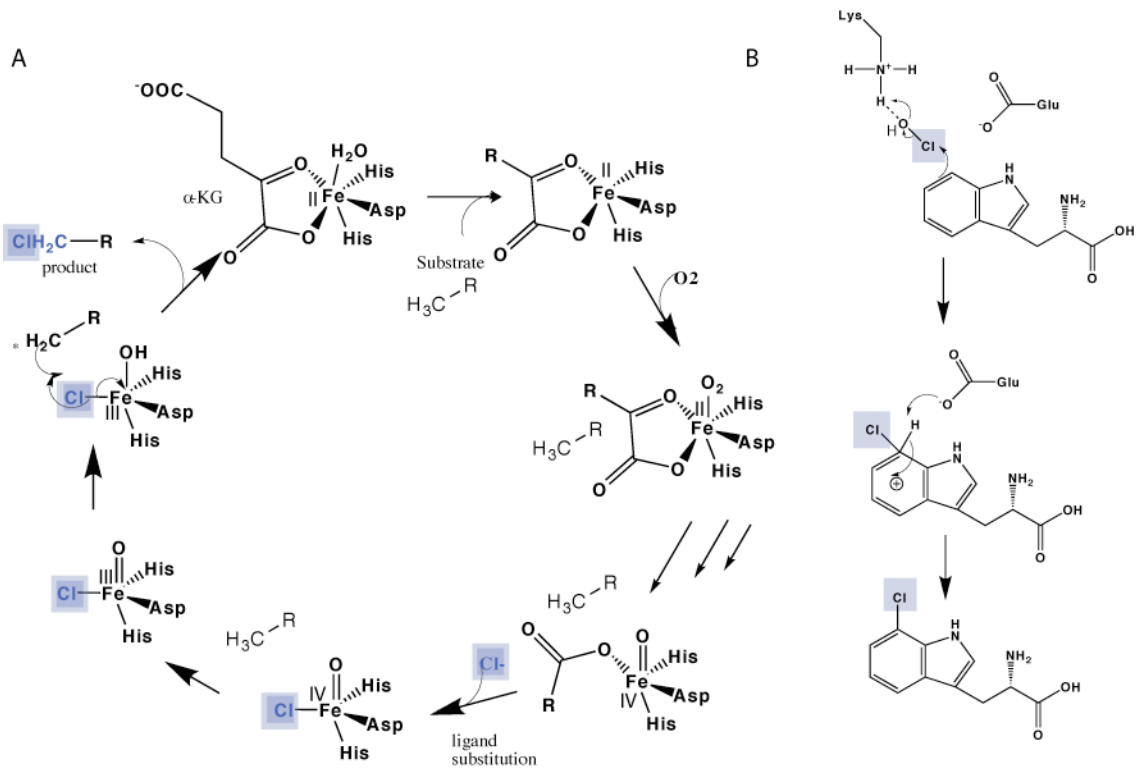


Figure 29: Proposed biosynthetic halogenation mechanisms

α-Ketoglutarate catalyzed halogenases were recently discovered and the proposed mechanism is shown (A). FAD dependent halogenases utilize a hypohalous acid intermediate that is trapped on a lysine in the active site which affords the subsequent site specific halogenation of aromatic substrates (B).

In preliminary screening experiments, we targeted α-KG mediated halogenases in order to recover novel natural product gene clusters. Using a sequence alignment of known α-KG halogenases including CmaB, SyrB2, and BarB2, (a coronatine, syringomycin, and barbamide halogenase respectively), we designed 9 sets of degenerate PCR primers to detect novel homologues from our recombinant eDNA libraries (Figure 26). Initially, crude eDNA extracts from multiple soil samples (BB1-2, LPSG\RW, TZ5\9, CLBR, and OL1\2), were screened using these degenerate primers in all combinations

with a range of annealing temperatures and buffer conditions to optimize degenerate PCR conditions. This optimization procedure is normally required as these degenerate primers are used to screen complex mixtures of genomic DNA as opposed to traditional PCR strategies. One set of primers was found to yield amplicons consistently and of the appropriate predicted size from all of the crude samples tested (CmaBF3: 5'-GARGGNACNGACTGGCAYCAG3', CmaBR3RC -5'-TAGTCRTARCCRWAGAMNCC-3') Cloning and sequencing of PCR products from this preliminary screen confirmed that the amplicons contained novel halogenase homologues. (See Materials and Methods)

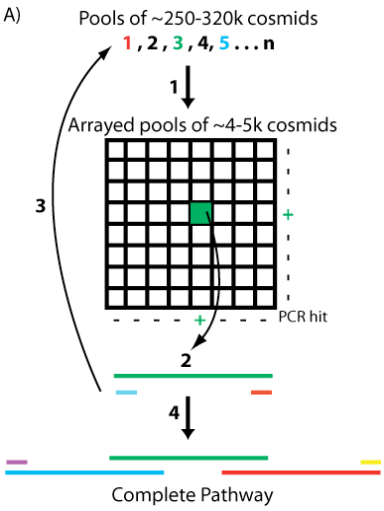


Figure 30: Screening and clone recovery outline

Degenerate PCR is used to identify library subarrays which contain a biosynthetic sequence of interest (1). Second generation libraries utilized pooled rows and columns to facilitate fractionation efforts (2). Dilution fractionation of PCR-positive pools is then used to isolate a cosmid containing a biosynthetic homologue. If necessary, sequence information from the ends of a biosynthetic cosmid is used to iteratively recover overlapping cosmids using the same strategy.

Subarrays from first-generation libraries derived from these eDNA samples were screened with optimized degenerate PCR conditions to yield 15 unique groups of α-KG halogenase homologues. The sequences derived from subarray screens were used to design specific primers to isolate cosmids containing halometabolite biosynthetic gene clusters. Cosmids were recovered from library subarrays using either colony hybridization with homologous DIG™ (digoxigenin, Roche) labeled sequence probes or whole-cell PCR and dilution fractionation (Figure 30). Three cosmids containing halogenase sequences were recovered from first-generation eDNA libraries (BB16, OlI4, BB48) using this strategy. These cosmids were end sequenced using vector-specific primers, and two of the pathways were clearly truncated based on the presence of partial biosynthetic genes (OLI4, BB48). We subsequently designed primers based on end sequencing in an effort to recover overlapping clones carrying the remaining portions of these two gene clusters. Despite exhaustive screening efforts, overlapping clones could not be detected for these pathways in any of the first generation libraries which contained between 20,000-200,000 clones each.

We opted to focus on the remaining gene cluster, BB16, which appeared complete based on comparisons to a related sequenced pathway encoding the phytotoxin phosphinothricin tripeptide (PTT). PTT (Figure. 31) is a glutamine tRNA synthetase inhibitor and is the most widely used broad-spectrum herbicide and selection agent for genetically modified (GM) crops

worldwide (Bayer Liberty Link™). An acetyltransferase from the producing species *Streptomyces viridochromogenes* confers resistance to PTT by acetylating and sterically preventing PTT from binding glutamine tRNA synthetase. GM crops are stably transfected with a recombinant form of the acetyltransferase, allowing them to grow under PTT selection. The phosphinothricin moiety, which is thought to chemically mimic phosphorylation without being subjected to native dephosphorylation mechanisms, is rare, and only a handful of other natural products which contain this functional group have been reported to date. *In vitro* studies of phosphinothricin acetyltransferase have also shown that it displays high substrate specificity indicating that this gene cluster could provide a novel herbicide and resistance gene combination. (Davies, Tata et al. 2005; Herouet, Esdaile et al. 2005; Davies, Tata et al. 2007)

Full sequencing revealed that cosmid BB16 appeared to contain the biosynthetic modules which encode the tripeptide in addition to an α-KG halogenase which was used as the initial screening target. No halogenated phosphinothricin tripeptide derivatives have been isolated or generated synthetically to date. In the interest of analyzing additional genes outside of the canonical gene cluster, we attempted to recover overlapping cosmids but were unable to detect any candidates in first-generation libraries. Cosmid BB16 was retrofitted with a *streptomyces* integrative expression cassette using reported lambda-mediated recombination protocols. (Court, Sawitzke et

al. 2002; Sawitzke, Thomason et al. 2007; Sharan, Thomason et al. 2009) The pathway was then mobilized into a panel of *streptomyces* including *S. lividans, S. albus, S. viridochromogenes, S. toyocaensis,* and *S. lavendulae* using biparental conjugation. (Tobias Kieser 2000) From here, reported culture conditions for the original isolation of PTT from *S. viridochromogenes* and three standard *streptomyces* fermentation medias (R5, R2YE, M9) were used to grow the recombinant strains in 25mL shake flasks over a 7 day timecourse. (Tobias Kieser 2000) Organic and water soluble extracts of the culture timecourses were analyzed by LC/MS (a hydrophilic HILIC column was used for water soluble fractions as reported) and compared to synthetic PTT standards. Unfortunately, we were unable to detect any clone-specific metabolites in both organic and water soluble extracts via LC/MS using multiple culture conditions and heterologous hosts. Overlay assays with *Bacillus subtilis* employed for the functional detection of PTT in the native host also did not reveal any clone-specific bioactivity in heterologous *streptomyces* hosts compared to *S. viridochromogenes* controls. (Blodgett, Thomas et al. 2007)

Although the BB16 pathway contained a homologous canonical NRPS motif required for generating the phosphinothricin tripeptide backbone, the organization of the ORFs was entirely distinct from the sequenced PTT gene cluster. This raised the possibility that additional biosynthetic modules required for the production of the natural product may exist outside of the

original cosmid. Because we were unable to recover any overlapping clones in first-generation eDNA libraries, we ceased further examination of this particular biosynthetic system after initial heterologous expression trials were unsuccessful. Although no clone-specific metabolite was uncovered, this preliminary study was a clear demonstration that sequence-based screens could be used to discover classically rare natural product biosynthetic motifs from eDNA libraries.

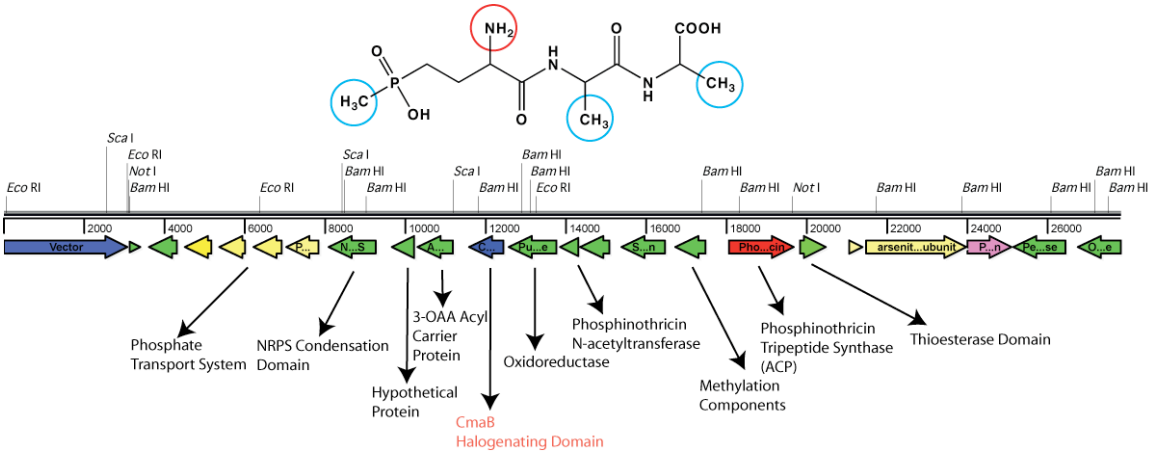


Figure 31: BB16, a gene cluster encoding a putative phosphinothricin tripeptide analog

Cosmid BB16 was recovered using α-KG-specific degenerate primers in a screen of ~200,000 cosmid clones. Potential sites of halogenation are circled in blue while the site of acetylation by the resistance gene, *bar,* is circled in red.

2.2.2.2 Type II polyketides

Type II polyketide biosynthetic gene clusters are found in phylogenetically diverse bacteria and have been used in several studies in search of novel natural products from environmental samples. (Seow, Meurer et al. 1997; King, Bauer et al. 2009; Kim, Feng et al. 2010) The domain

organization of type II minimal PKS genes allows KSβ sequences to be amplified using conserved nucleotide sequences in flanking ACP and KSα genes. The details of this domain organization, primer design, and general screening strategy are discussed in Chapter 1 (Sequence-based screens). Based on the knowledge that type II PKS systems are sometimes encoded by smaller gene clusters, we set out to recover cosmid clones containing these biosynthetic systems from our first generation eDNA libraries. Initial screens of two first generation soil-based eDNA libraries (BB, LPSG) yielded four clones which contained type II PKS genes (LPSG47, LPSG48, LPSG72, BB76). End-sequencing of clone BB76 revealed that the pathway was incomplete as two truncated biosynthetic genes were found on the ends of the cosmid insert. We were unable to detect overlapping clones in the BB library using PCR primers designed against the ends of clone BB16. More comprehensive sequencing of the remaining PKS-like clones also revealed that these gene clusters (LPSG 47. 48. 72) were truncated and no additional overlapping clones could be detected in this first-generation eDNA library. After retrofitting each cosmid with a *streptomyces* integration cassette we conjugatively introduced the clones into *S. lividans* for expression studies. (Tobias Kieser 2000) We were unable to detect any clone-specific metabolites from these recombinants but this result was not entirely unexpected as the gene clusters were clearly truncated based on sequencing efforts. Although no clone specific metabolites could be detected in the culture broths of these

incomplete pathways, this study demonstrated that a diverse range of type II PKS systems could be recovered from eDNA libraries using sequence-based screens. We would subsequently use this same strategy in later studies to recover several complete type II PKS gene clusters from larger eDNA libraries (section 2.2.4).

Together, these preliminary results demonstrated that a diverse range of biosynthetic cosmids can be recovered from soil-derived eDNA libraries using sequence based screens. Although each of the pathways recovered in preliminary screens were truncated and no overlapping clones could be detected in first-generation libraries, these results set a clear precedent for the discovery of a diverse range of complete gene clusters from larger eDNA libraries using sequenced-based methods. We therefore designed a strategy to create eDNA libraries of sufficient size to allow the recovery of groups of overlapping clones comprising complete biosynthetic pathways. The following sections describe successes along this aim.

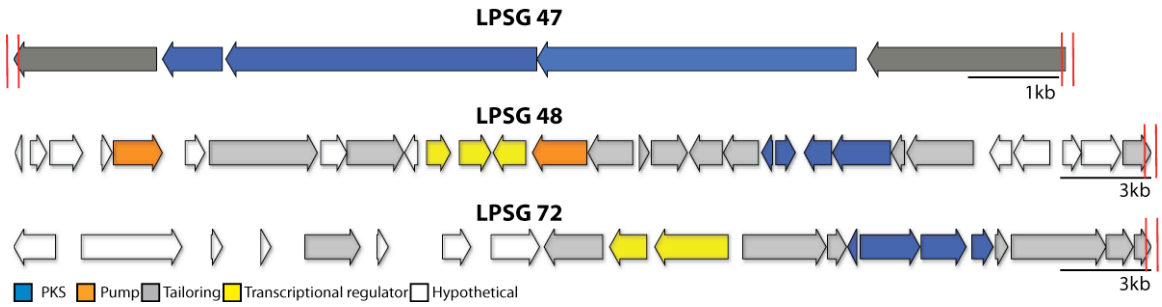


Figure 32: First-generation type II PKS pathways

LPSG 47, 48, and 72 were recovered from a first generation eDNA library (LPSG) using type II PKS specific primers. (Seow, Meurer et al. 1997; King, Bauer et al. 2009; Kim, Feng et al. 2010) The pathways were determined to be truncated after end sequencing (red hashes indicate incomplete biosynthetic ORFs) and no additional overlapping clones could be detected in smaller first-generation libraries.

2.2.3 Library Size Analysis

The rich microbial diversity present in soils makes them attractive but challenging starting points for the culture-independent discovery of new natural product biosynthetic gene clusters. Much of the difficulty in working with soil microbiomes stems from their complexity, which requires the construction of large clone libraries in order to ensure that complete biosynthetic pathways can be recovered from randomly cloned fragments of metagenomic DNA. Sequence-based screens of first-generation libraries (20,000-200,000 clones in size) led to the discovery of a range of type II polyketide and halometabolite gene clusters (section 2.2.2). All of these biosynthetic pathways were incomplete, however, and overlapping cosmids comprising the remaining portions of the pathways could not be detected in first-generation libraries. This prevented the further examination of these pathways as they did not contain all of the genes required for the

biosynthesis of a natural product. In order to address this, we designed a simple screening strategy which would allow us to determine the size a soil-based eDNA library needed to be in order to recover overlapping clones comprising complete biosynthetic gene clusters.

For this study, DNA isolated from soil collected in Utah and Anza Borego (California) was used to construct a series of independent 750,000- and 320,000-membered eDNA cosmid libraries respectively. Using these unique aliquots of library samples, we set out to determine the point at which redundant biosynthetic sequences would begin to appear. We hypothesized that this would serve as a general indicator of when two distinct library clones contained overlapping fragments of eDNA, most likely from the same biosynthetic pathway. Earlier studies suggest that type II (aromatic, iterative) PKS biosynthetic systems are highly conserved and are found in phylogenetically diverse bacteria. (Seow, Meurer et al. 1997) We therefore chose type II PKS pathways as a model system for this analysis. Both the Anza Borego and Utah libraries were screened by John Bauer and Zhiyang Feng for the presence of beta-ketoacyl synthase (KSβ) gene sequences using degenerate PCR primers designed to recognize type II PKS systems. (Seow, Meurer et al. 1997; King, Bauer et al. 2009) In total, 19 and 73 distinct KSβ gene sequences were amplified from the Utah and Anza Borego libraries respectively (Figure 33). Redundant KSβ sequences began to regularly appear once ~$3x10^6$ and ~$2.25x10^6$ clones had been examined in the Utah and

California libraries. Additional screens using primers designed to recognize other conserved natural product biosynthetic gene sequences have also shown that redundant sequences begin to appear once libraries exceed 1-3x10^6 clones in size. (Banik and Brady 2008) The libraries used in our efforts to recover complete natural product gene clusters were therefore expanded until they contained 1-1.5x10^7 unique clones, approximately 5-10 times the number of clones needed to identify the first redundant biosynthetic sequences. While even eDNA libraries containing 1-1.5x10^7 clones are unlikely to permit the recovery of rare biosynthetic gene clusters, this analysis suggested that they should contain collections of clones encoding complete PKS gene clusters and, by extension, clones comprising many other types of complete biosynthetic gene clusters found in the genomes of uncultured bacteria.

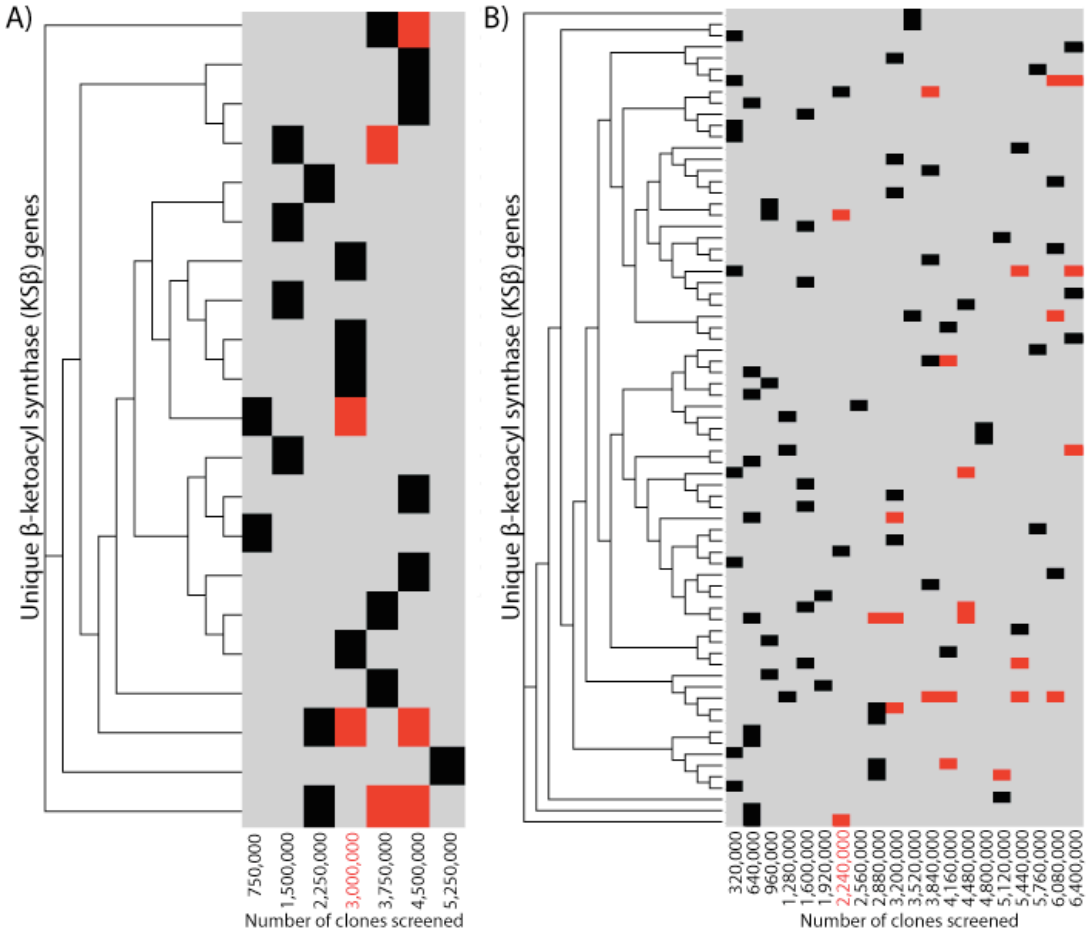


Figure 33: Library size analysis

The Utah (A) and Anza Borego (B) eDNA libraries were used to determine the point at which redundant KSβ sequences (overlapping type II PKS clones) could be detected in unique library aliquots (red). *Adapted from* (Kim, Feng et al. 2010).

2.2.4 Second Generation Sequence-Based Screens

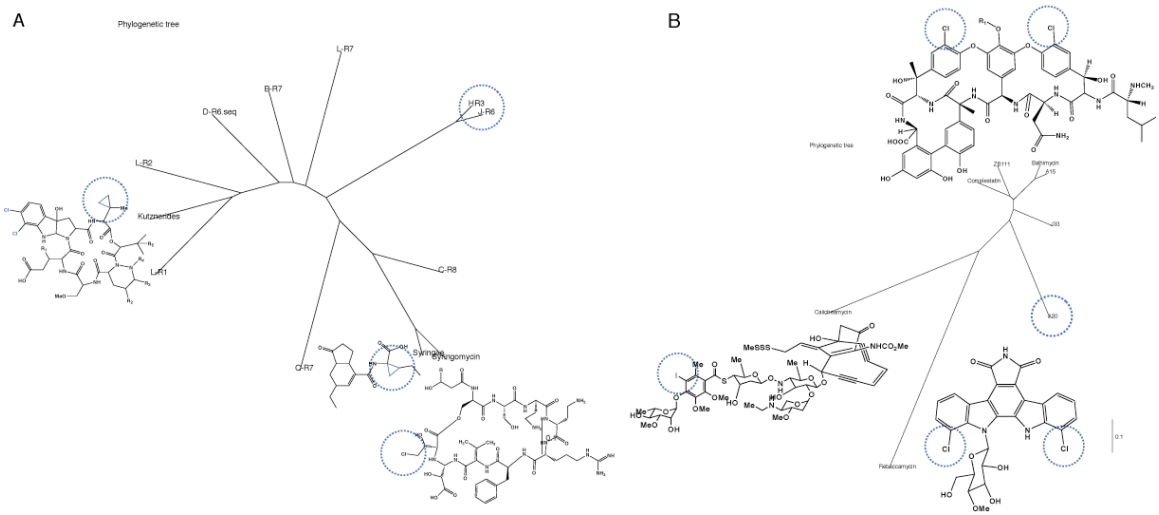


Figure 34: Molecular phylogeny of α-KG (A) and FAD dependent (B) halogenases
This figure shows the sequence relationships between known halogenases with their associated structures and novel homologues identified in a screen of a second generation 1x10^7 member cosmid library (Utah). Cosmids from all branches were recovered. Halogenations are shown circled (blue).

2.2.4.1 Halogenases

To recover additional halometabolite biosynthetic pathways, we focused all subsequent screening efforts on larger second-generation cosmid libraries containing 1-1.5x10^7 clones. Our analysis of these libraries with minimal type II PKS specific primers indicated that they would most likely contain overlapping clones for a range of biosynthetic pathways as opposed to smaller first-generation libraries. Preliminary screens of a 1x10^7 member cosmid library derived from soil collected in Utah revealed 8 unique groups of α-KG halogenase homologues (Figure 34a). Cosmids representing all of these groups of sequences were recovered from the Utah eDNA library using

specific primers derived from sequences identified in preliminary screens. End sequencing revealed that, in each case, the pathways were most likely truncated as biosynthetic genes were found at the ends of the cosmid inserts. Additional PCR primers were therefore designed against the ends of initial cosmid isolates to recover additional clones containing the remaining portions of each gene cluster (Figure 30). Overlapping clones for all of the initial cosmids were successfully isolated demonstrating that the $1x10^7$ membered Utah library was large enough to recover a diverse range of halometabolite enoding biosynthetic pathways carried on multiple cosmids. The process of end sequencing and clone recovery was iterated until non-biosynthetic genes were identified in the proximal and distal ends of a putative biosynthetic gene cluster. Full sequencing of the cosmids comprising potentially complete pathways revealed that most encoded unknown metabolites, as they showed no similarities in enzyme organization or gene homology to characterized gene clusters (Figure 35). Two of the pathways appear to encode the previously reported phytotoxin syringomycin as they contained similar gene organization to the sequenced biosynthetic gene cluster (GenBank No. ADGB01000000). (Fukuchi, Furihata et al. 1992) One of the cryptic clusters (J48) spanned more than 120 kb and was recovered on 5 overlapping cosmids. This was the first demonstration that eDNA-derived pathways of this size could be recovered using sequence-based screens of large soil libraries.

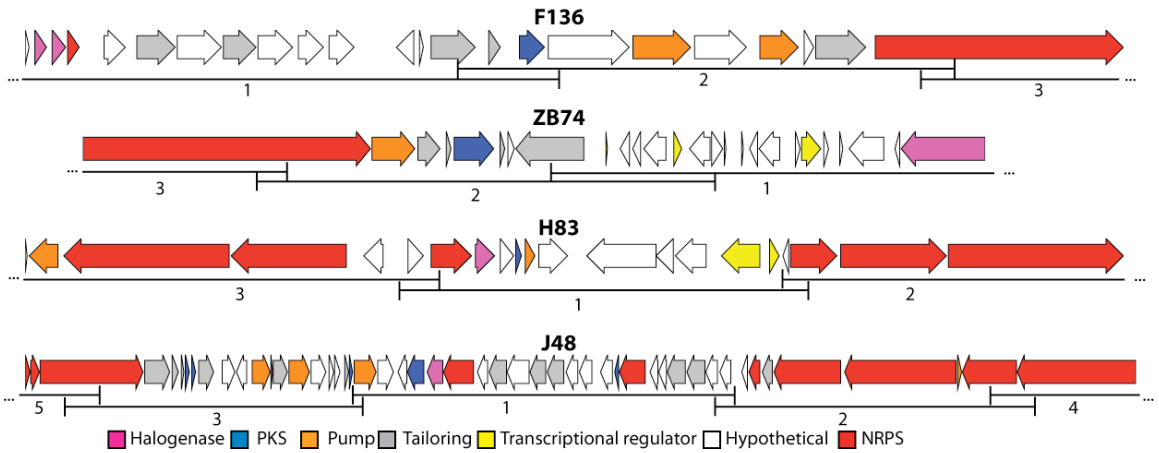


Figure 35: α-KG Halometabolite gene clusters

Large halometabolite gene clusters were recovered from groups of overlapping clones using α-KG halogenase specific primers (shown in purple) and a second generation 1x10^7 member cosmid library. The cosmid inserts (~40 kb each) and number of rounds of clone recovery required to isolate the cosmids are shown in black. Hashes (…) indicate sequence regions where all ORFs were classified as hypothetical or unknown proteins.

The functional boundaries of a metagenomically derived gene cluster are typically determined by comparing it to related gene clusters or by identifying clearly non-biosynthetic genes in proximal and distal ends of a putative pathway. All of the α-KG halogenase containing gene clusters contained large regions (>20 kb) of uncharacterized and hypothetical proteins making the sequence-based determination of functional boundaries and putative structural assignments challenging. Also, there are very few well characterized gene clusters that contain α-KG halogenases making homology-based comparisons difficult. We isolated additional overlapping clones for all of the pathways in an attempt to identify genes which would indicate whether the cluster was fully recovered, but these cosmids were exclusively comprised of hypothetical and unknown proteins (Figure 35). Each of the α-

KG halogenase gene clusters spanned at least 100 kb in length and multiple overlapping cosmids, making downstream heterologous expression efforts very challenging in the absence of robust reassembly solutions. Many of the ORFs contained within these pathways were also only partially cloned on each fragment of DNA making functional studies impossible for multi-plasmid expression systems. Because of the cryptic nature of these gene clusters, the difficulty of assigning functional boundaries, and the challenge of systematically approaching multi-plasmid heterologous expression experiments, additional halometabolite gene cluster screening efforts were focused on FAD-mediated halogenases.

There are numerous sequenced and well characterized FAD-halogenase containing biosynthetic pathways which encode a diverse array of secondary metabolites. (Hornung, Bertazzo et al. 2007) The homology-based determination of the functional boundaries of an eDNA derived gene cluster is therefore more straightforward in comparison to α-KG halogenase pathways. Many compounds containing FAD-mediated halogenations have well-defined bioactivities and the discovery of novel congeners would be of general interest (Figure 34b). Also, FAD-halogenases have been used for the *in vitro* modification of related chemical scaffolds. (Yeh, Garneau et al. 2005) Novel homologues identified in these screens could therefore, in theory, be used to enzymatically complement synthetic halogenation strategies. (Banik and Brady 2008) Most importantly, FAD-halogenases contain high levels of

sequence similarity among structurally related halometabolites. This allows novel gene clusters containing FAD halogenases to be grouped with related chemical scaffolds using the sequence of this single enzyme. (Hornung, Bertazzo et al. 2007) In general, the ability to determine both the functional boundaries of a novel gene cluster and potential structural features of the encoded secondary metabolite prior to full pathway recovery represents a major advantage.

When screening for FAD-dependent aromatic halogenases, we focused exclusively on a second-generation eDNA library containing $1x10^7$ clones (Utah), as it was now clear that most canonical biosynthetic pathways would need to be recovered from overlapping groups of clones. FAD-mediated halogenases contain a highly conserved FAD binding site and a group of tryptophan residues ~400-800 nucleotides 3' to the active site. These two regions were trimmed from sequenced FAD halogenase homologues including ComH, RebH, and PrnA, (a glycopeptide, indolocarbazole, and pyrrolnitrin halogenase respectively), and used for two independent multiple sequence alignments. A crude eDNA extract, Pennsylvania 1 (courtesy of Jeffrey Craig), was used to optimize six pairs of compatible degenerate PCR primers with various annealing temperatures and buffer conditions (See Materials and Methods). One set of primers (THalF3: 5'-CGGCTGGTTCTGGTACATCCC-3', ThalR2: 5'-GAACTCGTAGAASACSCCGTACTC-3') consistently yielded an appropriately

sized amplicon. Cloning and sequencing of the amplicons generated in this preliminary screen confirmed that novel halogenase homologues could be detected in crude eDNA samples using these primers. Degenerate PCR screening of the Utah library yielded four groups of novel FAD-halogenase sequences (Figure 34b). Specific primers were designed using these sequences, and a cosmid containing what appeared to be a glycopeptide antibiotic halogenase was recovered from the Utah cosmid library (ZA41). End sequencing of this cosmid revealed that the pathway was truncated so additional primers were designed to recover overlapping cosmids comprising a complete gene cluster. Overlapping cosmids Q87 and J2 were recovered in two additional rounds of clone recovery and screening. The cosmids comprising a complete pathway were fully sequenced revealing that the NRPS gene cluster most likely encodes a novel halogenated enneapeptide with tailoring and non-canonical amino acid backbone similarities (including dihydroxyphenylglycine (DHPG) and hydroxyphenylglycine (HPG)) to glycopeptide antibiotics such as vancomycin and teicoplanin. The NRPS pathway spans approximately 90 kb in length and was recovered on three overlapping cosmids (Figure 36). A fourth overlapping cosmid (ZB118), which was recovered from the same library, was determined to be unnecessary as

sequencing later revealed that it contained no additional biosynthetic genes.

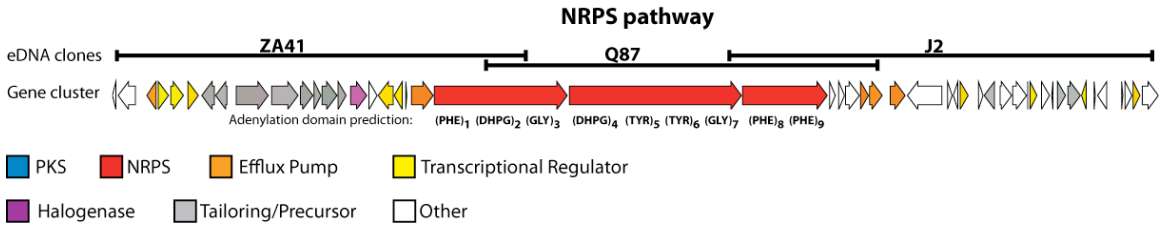


Figure 36: NRPS gene cluster

FAD halogenase specific primers were used to recover a gene cluster encoding a novel enneapeptide spanning approximately 90 kb. This gene cluster appears to encode a novel halogenated enneapeptide. The halogenase domain used for initial isolation is shown (purple). Additional annotations are colored according to the legend. NRPS modules were predicted using NRPSpredictor (Rausch, Weber et al. 2005) (DHPG – dihydroxyphenylglycine).

Table 3: Cryptic NRPS gene cluster annotation

Detailed ORF annotations of the cryptic NRPS gene cluster are shown with the closest homologues. Overlapping cosmids are listed (a=Za41, b=Q87, c=J2) for reference.

ORF	Size	Source [a]	Protein(s), organism, accession number of closest homologue	e value
NRPS1	114	a	SAV6581, *S. avermitilis*, NP827757.1	2e-155
NRPS2	517	a	Amidase, *Rubrobacter xylanophilus*, YP643254.1	3e-05
NRPS3	279	a	Transporter, *S. sviceus* ATCC29083, YP002204564.1	5e-90
NRPS4	280	a	Regulator, *S. sviceus* ATCC29083, YP00224563.1	9e-143
NRPS5	353	a	phospho-2-dehydro-3-deoxyheptonate aldolase, *Actinoplanes Teichomyceticus,* CAE53368.1	4e-154
NRPS6	340	a	StrR Treg, *Actinoplanes Teichomyceticus*, CAE53369.1	4e-99
NRPS7	372	a	Hmo, *S. toyocaensis*, AAM80552.1	2e-156
NRPS8	367	a	HmaS, *S.toyocaensis*, AAM80551.1	1e-147
NRPS9	926	a	LuxR, *Actinoplanes Teichomyceticus*, CAG15028.1	5e-144
NRPS10	751	a	HpgT, *S. fungicidus*, ABD65949.1	0
NRPS11	372	a	DpgA, *S.toyocaensis*, AAM80548.1	3e-167
NRPS12	221	a	DpgB, *S.toyocaensis*, AAM80547.1	6e-61
NRPS13	439	a	DpgC, *Amycolytopsis balhimycina*, CAC48380.1	1e-161

NRPS14	270	a	DpgD, *S.toyocaensis*, AAM80545.1	6e-121
NRPS15	498	a	Halogenase, *Streptomyces sp.* CB2664, AB82426.1	0
NRPS16	270	a	A/B Hydrolase fold, *Franka sp.* CcI3, YP481545.1	1e-99
NRPS17	425	a	Histidine kinase, *Catenulispora acidiphila*, ZP04370999.1	2e-84
NRPS18	252	a	Response regulator, *Catenulispora acidiphila*, ZP04370998.1	1e-65
NRPS19	75	a	NA, *Gordonia bronchialis*, ZP03883065	9e-25
NRPS20	686	a	ABC transporter, *Thermomonospora curvata*, ZP04032681.1	0
NRPS21	3672	a	Putative NRPS, *S. griseus*, YP001824777.1	0
NRPS22	4745	a-b	NRPS, *S. ghanensis sp.*, ZP04685019.1	0
NRPS23	2360	b-c	CDA peptide synthase III, *S.coelicolor* A3(2), NP733597.1	0
NRPS24	205	b-c	Hypothetical protein, *Kribella flavida*, ZP03860135.1	5e-08
NRPS25	203	b-c	Hypothetical protein, *Catenulispora acidiphila*, ZP04376206.1	2e-07
NRPS26	384	b-c	Peptidoglycan binding protein, *Streptosporangium roseum*, ZP04470394.1	8e-35
NRPS27	255	b-c	ABC-type antimicrobial peptide transport system, *Streptosporangium roseum*, ZP04470395.1	2e-63
NRPS28	399	c	Hypothetical protein, *Micromonospora sp.* ZP04604412.1	1e-81
NRPS29	427	c	Antiporter, *A. balhimycina*, CAC48373.1	2e-93
NRPS30	976	c	Hypothetical, *S. clavuligerus*, ZP05003411.1	0
NRPS31	170	c	Hypothetical protein SGR6859, *S. griseus*, YP001828371.1	3e-59
NRPS32	136	c	Hypothetical protein, *Corynebacterium efficiens*, NP737671.1	9e-08
NRPS33	251	c	LysR, *S. avermitilis*, NP822360.1	5e-82
NRPS34	167	c	Secreted protein, *S. pristinaespiralis*, ZP05014329.1	1e-42
NRPS35	319	c	Prephenate dehydrogenase, *S. coelicolor* A3(2), NP733544.1	6e-100
NRPS36	337	c	Hypothetical protein, *S. roseosporus*, ZP04711333.1	5e-85

NRPS37	393	c	Hypothetical protein, *S. roseosporus*, ZP04696149.1	6e-100
NRPS38	46	c	Two component regulator, *S. griseus,* YP001823365.1	0.059
NRPS39	224	c	Two component regulator, *S. clavuligerus*, ZP05008347.1	1e-94
NRPS40	275	c	Hypothetical protein, *Actinoplanes friuliensis,* CAD32905.1	5e-146
NRPS41	43	c	NA	
NRPS42	269	c	Methyltransferase, S.albus, ZP04705340.1	4e-78
NRPS43	386	c	Glycosyltransferase*, S. albus,* ZP0405339.1	1e-146
NRPS44	147	c	Transcriptional regulator, *S. coelicolor*, NP630524.1	1e-36
NRPS45	71	c	MbtH-like protein, *S. fungicidus*, ABD65966.1	1e-27
NRPS46	287	c	Transcriptional regulator, *S. fugicidus*, ABD65942.1	1e-76
NRPS47	84	c	Proline Racemase, *S. sviceus*, ZP05016839.1	0.10
NRPS48	222	c	dihydrodipicolinate synthase, *S. erythraea*, YP001105451.1	2e-39
NRPS49	237	c	Transcriptional regulator, *Kribbela flavida*, ZP03860465.1	5e-38
NRPS50	464	c	Magnesium dependent protein, *S. sp* Mg1, ZP04997562.1	0.16

2.2.4.2 Type II PKS

To demonstrate that additional types of biosynthetic gene clusters could be discovered using sequence-based screens of large eDNA libraries, John Bauer, Michael Clark Pearson and I recovered a panel of 10 type II PKS cosmids from a $1x10^7$ member eDNA library derived from soil collected in Utah using the same KSβ specific primers discussed in previous sections. (Seow, Meurer et al. 1997; King, Bauer et al. 2009) After retrofitting these cosmids with integrative expression cassettes, they were conjugatively shuttled into *S. lividans* and *S. albus* for heterologous expression studies. (Tobias Kieser 2000) One of the recombinant *S. albus* clones (V167) led to the discovery of the polyketide Erdacin (**34**) outlined in Chapter 1. Many of the remaining type II PKS pathways appeared to be truncated after sequencing and required the recovery of overlapping clones. In contrast to first-generation library screens which did not yield complete type II PKS pathways, we were able to recover overlapping clones comprising a complete gene cluster (hereon referred to as the PKS pathway) from the larger second generation Utah eDNA library. The overlapping clone, V48, was recovered from the library using PCR primers designed against the end sequences of the original cosmid isolate X16. While the eDNA-derived PKS gene cluster does not closely resemble any known pathway, the appearance of primary metabolic enzymes in the sequence surrounding the conserved natural

product biosynthetic genes suggests that it was likely recovered in its entirety (Figure 37).

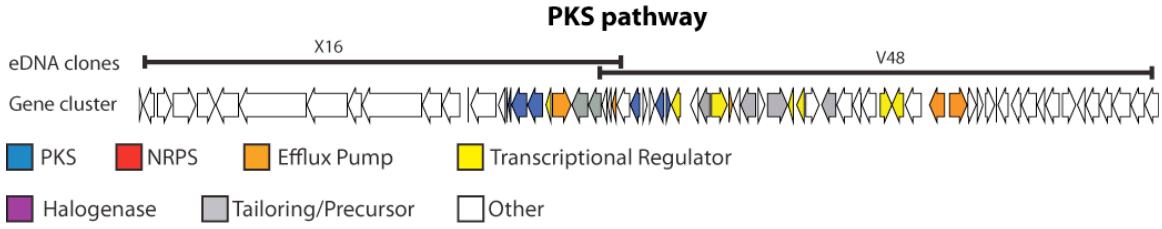


Figure 37: PKS gene cluster

The PKS cluster was recovered from a 1x10^7 member soil eDNA library (Utah) using type II minimal PKS degenerate primers.(Seow, Meurer et al. 1997; King, Bauer et al. 2009; Kim, Feng et al. 2010)

Table 4: Cryptic PKS gene cluster annotation

Detailed annotation of the cryptic PKS pathway recovered from a 1x10^7 member cosmid library constructed from soil collected in Utah. Overlapping cosmids are listed (a=X16, b=V48) for reference.

ORF	Size	Source [a]	Protein(s), organism, accession number of closest homologue	e value
PKS1	678	a	Hypothetical protein, *S. pristinaespiralis,* ZP05013621.1	6e-16
PKS2	207	a	Hypothetical protein, *Kribbella flavida*, ZP03865820.1	5e-92
PKS3	53	a	Hypothetical protein, *Janibacter sp.,* ZP00994426.1	0.007
PKS4	80	a	Acyl carrier protein, *S. arenae*, AAD20269.1	3e-16
PKS5	406	a	Polyketide Synthase, *Saccharopolyspora hirsute*, AAA26489.1	1e-134
PKS6	423	a	Ketosynthase, *S. antibioticus*, CAC05671.1	3e-178
PKS7	144	a	MerR transcriptional regulator, *Janibacter sp.,* ZP0996281.1	2e-30
PKS8	496	a	Efflux pump, *Streptomyces sp. Mg1*, YP002181945.1	4e-92
PKS9	414	a	Cytochrome P450, *S. avermitilis* MA4680, NP823553.1	0.005
PKS10	355	a-b	NapB (Me-transferase), *S. aculeolatus*, ABS50465.1	3e-127
PKS11	139	a-b	Hypothetical protein, *Frankia alni* ACN14A,	2e-04

			YP714822.1	
PKS12	111	a-b	No Similarity (NS)	NS
PKS13	128	a-b	Hypothetical protein, *Ricinus communis*	1.1
PKS14	333	b	OxyF*, S. rimosus*, AAZ78330.1	7e-80
PKS15	273	b	Cyclase, *Streptomyces sp.,* AAG30197.1	2e-38
PKS16	137	b	Hypothetical protein, *Fulvimarina pelagi*, ZP01439050.1	3e-28
PKS17	145	b	NS	NS
PKS18	250	b	Cyclase, *Streptomyces sp*. R1228, AAG30196.1	6e-82
PKS19	133	b	Aln5, *Streptomyces sp*. CM020, ACI88872.1	1e-16
PKS20	250	b	Regulatory protein, *S. avermitilis*, NP823545.1	4e-44
PKS21	212	b	Hypothetical protein, *S. erythraea* NRRL2338, YP001102450.1	3e-28
PKS22	295	b	Carboxylesterase, *M.smegmatis* MC2 155, YP885948.1	1e-64
PKS23	427	b	TetR, *S. erythraea* NRRL2338, YP001109142.1	2e-28
PKS24	111	b	ABC transporter, *Arthrobacter sp*. FB24, YP82923.1	3e-15
PKS25	204	b	Hypothetical protein, *S. roseum*, ZP04469915.1	4e-33
PKS26	394	b	Acyl-CoA dehydrogenase, *Micromonospora sp*. ZP04609304.1	8e-47
PKS27	190	b	Holo acyl carrier protein synthase, *Leptospirillum ferrodiazotrophum*, EES52839.1	2e-13
PKS28	533	b	NS	NS
PKS29	158	b	LuxR, *Acidothermus cellulyticus*, YP873148.1	9e-08
PKS30	200	b	TetR, *Methylobacterium nodulans*, YP002496490.1	3e-09
PKS31	361	b	Non-heme bromoperoxidase, *Frankia alni*, YP710982.1	6e-19
PKS32	385	b	Acyltransferase, *Brevibacterium linens* BL2, ZP00377888.1	4e-94
PKS33	405	b	Hypothetical protein, *Brevibacterium linens* BL2, ZP00377876.1	2e-132
PKS34	233	b	Hypothetical protein, *Trichodesmium erythraeum*, YP724068.1	3e-48
PKS35	393	b	Hypothetical protein, *Corynbacterium glutamicum*, YP001137300.1	6e-25
PKS36	313	b	XRE Treg, *Arthrobacter chlorophenolicus*,	1e-89

			YP002489420.1	
PKS37	322	b	Anchor protein, *Streptococcus pneumoniae,* YP002741036.1	3e-07
PKS38	434	b	FAD dehydrogenase, *Brevibacterium linens,* ZP00377889.1	4e-147
PKS39	248	b	NS	
PKS40	149	b	Hypothetical protein, *Streptosporangium roseum*, ZP04473826.1	1e-07

2.2.4.3 Acyl-ACP ligase

To demonstrate that specific classes of natural product gene clusters can be recovered using sequence-based screens, we isolated a biosynthetic pathway encoding a lipopeptide natural product. Lipopeptide antibiotics are characterized by a canonical acyl chain that is attached to a non ribosomal peptide backbone through the coordinated actions of an acyl-ACP ligase and downstream NRPS modules. Fang Chang designed degenerate primers that recognize acyl-ACP ligases found in lipopeptide antibiotic gene clusters (dptE (daptomycin), lipA (friulimicin)). These primers were then used to identify and recover a cosmid (1679) from a $1.5x10^7$ member eDNA library constructed from soil collected in Anza Borego (California). Comprehensive sequencing revealed that cosmid 1679 contained the same ORF organization as a portion of the sequenced friulimicin gene cluster (GenBank No. 126635107). PCR primers designed against the end sequences of cosmid 1679 were subsequently used to identify and recover overlapping clones (1451, 201) in two rounds of iterative clone recovery from the same library. These additional clones were also fully sequenced. The eDNA-derived FRI gene cluster and the friulimicin gene cluster from *Actinoplanes friuliensis* have the same gene organization and are 89% identical over the 68 kb region that is predicted to comprise the functional biosynthetic pathway. A comparison of these two gene clusters suggests that the entire FRI gene cluster was likely captured on the three overlapping eDNA cosmids that were recovered. The

high degree of homology to the sequenced friulimicin gene cluster and the lack of any significantly divergent open reading frames in the eDNA-derived FRI pathway suggest that this gene cluster most likely also encodes friulimicin.

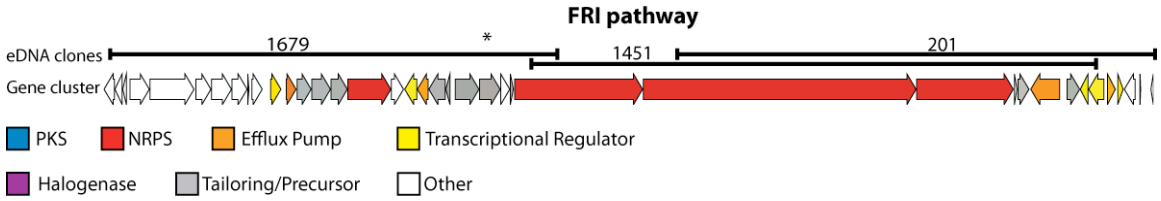


Figure 38: FRI gene cluster

The overall ORF organization and high degree of homology between the FRI gene cluster and the sequenced friulimicin gene cluster suggest that the eDNA-derived FRI pathway also encodes friulimicin. The Acyl-ACP ligase used for initial screening and clone recovery is shown (*).

Table 5: FRI gene cluster annotation

Detailed annotation of the FRI pathway recovered from a 1x10^7 member cosmid library constructed from soil collected in California (AB). Overlapping cosmids are listed (a=1679, b=1451, c=201) for reference

ORF	Size	Source[a]	Protein(s), organism, accession number of closest homologue	e value
FRI1	260	a	Hypothetical protein, *Aspergillus Niger*, XP658143.1	1e-04
FRI2	196	a	Predicted protein, *S. pristinaespiralis*, ZP05014280.1	2e-04
FRI3	102	a	Transferase, *S. pristinaespiralis*, ZP05009388.1	0.077
FRI4	490	a	CRISPR, *Clostridium thermocellum*, ZP03149922.1	3e-16
FRI5	1083	a	Hypothetical protein, *Roseobacter sp.*, ZP01058183.1	5e-26
FRI6	419	a	Hypothetical protein, *Acidovorax avenae sub. Citrulli*, YP971303.1	2.6
FRI7	505	a	Phospholipase D, *Azorhizobium caulinodans*, YP001525272.1	0.39
FRI8	374	a	TPR repeat, *Micromonospora sp.* ZP04603960.1	8e-17
FRI9	56	a	Alkaline phosphatase, *A. friuliensis*, CAM56764.1	1e-126

FRI10	273	a	RegA, *A. friuliensis*, CAM56765.1	8e-115
FRI11	266	a	ExpB, *A. friuliensis*, CAM56766.1	1e-93
FRI12	358	a	DabA, *A. friuliensis*, CAM56767.1	0
FRI13	481	a	DabB, *A. friuliensis*, CAM56768.1	0
FRI14	408	a	DabC, *A. friuliensis*, CAM56769.1	6e-144
FRI15	1064	a	NRPS, *A. friuliensis*, CAD32904.2	0
FRI16	303	a	Hypothetical protein, *A. friuliensis*, CAD32905.1	3e-149
FRI17	305	a	regB, *A. friuliensis*, CAD32906.1	1e-06
FRI18	280	a	ABC transporter, *A. friuliensis*, CAD32907.1	1e-123
FRI19	416	a	Glutamase mutase sub B, *A. friuliensis*, CAD32908.1	9e-160
FRI20	115	a	Glutamase mutase sub A, *A. friuliensis*, CAD32909.1	9e-72
FRI21	589	a	Acyl CoA synthase, *A. friuliensis*, CAD32910.1	0
FRI22	515	a	Acyl CoA dehydrogenase, *A. friuliensis*, CAJ18234.2	0
FRI23	236	a	Hypothetical protein, *A. friuliensis*, CAJ18235.1	1e-89
FRI24	89	a	Acyl carrier protein, *A. friuliensis*, CAJ18236.1	9e-40
FRI25	3139	a,b	NRPS B, *A. friuliensis*, CAJ18237.2	0
FRI26	6670	a,b,c	NRPS C, *A. friuliensis*, CAM56770.1	0
FRI27	2376	b,c	NRPS D, *A. friuliensis*, CAM56771.1	0
FRI28	82	b,c	NS	NS
FRI29	269	b,c	NS	NS
FRI30	727	b,c	Mem, *A. friuliensis*, CAM56774.1	0
FRI31	324	b,c	Lysine cyclodeaminase, *A. friuliensis,* CAM56775.1	2e-166
FRI32	215	b,c	NS	NS
FRI33	357	b,c	RegD, *A. friuliensis,* CAM56777.1	3e-139
FRI34	186	c	Hypothetical protein, *A. friuliensis,* CAN56778.1	5e-85
FRI35	159	c	Hypothetical protein, *Nocardioides sp.*, YP922990.1	6e-29
FRI36	284	c	Hypothetical protein, *Sphaerobacter thermophilus*, ZP04494592.1	6e-12
FRI37	32	c	NS	NS
FRI38	125	c	Transposase, *S. griseus*, YP001821630.1	9e-26
FRI39	82	c	NS	NS

2.3 Discussion and Future Directions

Many biosynthetic gene clusters are too large to be captured on a single insert using cosmid-based cloning strategies. This precludes their discovery using functional screens of eDNA libraries because all of the genes required for the biosynthesis of the natural product are not found on a single clone. To discover a broader range of biosynthetic pathways, an expression-independent sequence-based strategy was applied to recover cosmids containing biosynthetic gene clusters. Preliminary sequence-based screens demonstrated the successful recovery of biosynthetic cosmids encoding a diverse range of halometabolite and type II polyketides. Many of these pathways were truncated, however, and overlapping clones could not be detected in smaller first-generation libraries (<$1x10^6$ clones). We therefore empirically investigated the size an eDNA library needed to be in order to recover complete natural product gene clusters. Using a core biosynthetic motif as a readout (type II PKS: KSβ), we showed that redundant biosynthetic sequences, and by extension, overlapping fragments of gene clusters, begin to appear when libraries exceed ~2-$3x10^6$ clones in size. Other studies with different biosynthetic enzymes have yielded similar results. (Banik and Brady 2008) We subsequently constructed two of the largest eDNA-derived cosmid libraries reported to date, each of which contains more

than $1x10^7$ clones and over 100,000 bacterial genome equivalents[1]. Based on our KSβ sequencing results, we hypothesized that libraries of this size should permit the recovery of overlapping cosmids containing complete biosynthetic gene clusters. We next demonstrated systematic access to overlapping cosmids comprising complete gene clusters using a range of canonical biosynthetic motifs including core (PKS/NRPS), tailoring (Halogenase), and class specific (Acyl-ACP ligase) enzymes. Using sequence-based screens, we recovered several examples of what are classically considered rare gene clusters (PTT, FRI) in addition to several cryptic biosynthetic pathways which appear to encode novel secondary metabolites.

As suggested by these results, cosmid libraries containing in excess of 10 million clones appear to provide sufficient coverage of soil metagenomes to allow access to a diverse range of complete natural product biosynthetic gene clusters. Additional studies performed in our lab using other biosynthetic enzymes have now demonstrated that sequence-based screens of libraries of this size can be used to consistently recover a diverse range of large natural product gene clusters in their entirety. This general platform overcomes a major limitation that prevented the detection and isolation of many natural product gene clusters from cosmid-based eDNA libraries as truncated pathways cannot typically be detected in functional screens. In order to

[1] For an average bacterial genome of approximately 4Mb

functionally characterize larger eDNA-derived biosynthetic gene clusters captured on multiple cosmid clones, it is first necessary to assemble the pathway into a single construct. Pathway reassembly makes downstream heterologous expression efforts tractable, as multi-plasmid expression systems can become exceedingly complex. Also, many natural product gene clusters contain large ORFs, such as NRPS's which can individually span >30 kb, that are often only partially cloned on each individual cosmid comprising a complete pathway. This prevents their functional production during multi-plasmid heterologous expression efforts. In general, the lack of robust clone assembly strategies poses a major barrier to functionally characterizing larger gene clusters recovered on multiple eDNA-derived cosmid clones. To overcome this challenge, I developed and applied a novel recombination strategy in *Saccharomyces cerevisiae* which allows the rapid, single-reaction reassembly of multiple cosmids comprising complete biosynthetic gene clusters. (The details of these experiments are outlined in Chapter 3.)

2.4 Materials and Methods

2.4.1 eDNA Library Construction Details

Although a variety of different methods were initially tested, including combinations of grinding/pulverization, enzymatic lysis, and freeze/thawing, the three primary eDNA libraries used in these studies (Anza Borego, Arizona, Utah) were constructed using the following protocol: Between 250-500 grams of soil was incubated at 70°C in lysis buffer (2% sodium dodecyl

sulfate (w/v), 100 mM Tris-HCl, 100 mM ethylenediaminetetraacetic acid (EDTA), 1.5 M NaCl, 1% cetyl trimethylammonium bromide (w/v)) for two hours. Large particulates were then removed by centrifugation (4,000 x g, 30 min). DNA was precipitated from the resulting supernatant with the addition of 0.6 volumes of isopropyl alcohol, pelleted by centrifugation (4,000 x g, 30 min), washed with 70% ethanol and resuspended in a minimum volume of TE (10 mM Tris, 1 mM EDTA, pH 8).

High molecular weight DNA that was purified from the crude extract by gel electrophoresis (1% agarose, 0.5x Tris/Borate/EDTA, 16 hours, 20 V) was blunt-ended (End-It, Epicentre Biotechnologies), ligated into precut pWEB or pWEB-TNC (Epicentre Biotechnologies), packaged into lambda phage and transduced into *Escherichia coli* (EC100, Epicentre Biotechnologies). Individual library aliquots equivalent to approximately 4,000-5,000 colony forming units (CFU) were either plated on LB agar plates or inoculated into 5 mL of liquid LB and then allowed to incubate overnight at 37°C with the appropriate antibiotic selection. Once colonies formed, the plate-grown aliquots were resuspended in 5 mL of LB. Matching glycerol stocks (15% glycerol) and DNA miniprep pairs were created from each unique library aliquot. The minipreps were arrayed in 8 x 8 grids corresponding to 250,000-320,000 total cosmids and DNA from the rows and columns of each grid was pooled. To facilitate library screening, pooled rows and columns were further combined to yield master aliquots, each representing a single 8

x 8 grid of minipreps. Each unique *E. coli* transduction yielded three master aliquots (~750,000 clones) of the Utah library and one master aliquot (~320,000 clones) of the Anza Borego library. In total, the Utah soil library contains ~10 million unique cosmid clones and the Anza Borego soil library contains ~15 million unique cosmid clones.

2.4.2 General Screening Procedure for Library Size Analysis

Unique aliquots of ~750,000 eDNA clones (Utah library) were constructed independently from a single soil sample and then arrayed into the ~4,000-membered pools described above. These ~750,000-membered sublibraries (equal to three 8 x 8 grids described above) were then screened to determine the number of unique clones required to identify redundant examples of β-ketoacyl synthase (KS_b) genes in an eDNA mega-library. All together, six unique 750,000-membered library aliquots (~4,500,000 unique clones in total) were examined in detail. DNA from the ~4,000 membered pools was used as a template in PCR reactions with degenerate primers designed to amplify β-Ketoacyl synthase gene sequences. β-Ketoacyl synthase genes were amplified using the following pair of degenerate PCR primers: dp:KS_α, 5' TTC GGS GGI TTC CAG WSI GCS ATG and dp:ACP, 5' TCS AKS AGS GCS AIS GAS TCG TAI CC (I=deoxyinosine).(Seow, Meurer et al. 1997; King, Bauer et al. 2009; Kim, Feng et al. 2010) Each 25 µl PCR reaction contained 50 ng eDNA template, 2.5 µM each primer, 2 mM dNTPs, 1X ThermoPol Reaction Buffer (New England Biolabs, Ipswich MA), 0.5 U *Taq*

DNA polymerase (New England Biolabs) and 5% DMSO. A touchdown amplification protocol was used in the initial PCR screen of the library: denaturation (95°C; 45 sec), annealing (touchdown of 65°C to 58°C over 8 cycles dT -1°C, then 58°C for 35 cycles; 1 min), and extension (72°C; 2 min). (Seow, Meurer et al. 1997; King, Bauer et al. 2009; Kim, Feng et al. 2010) Amplicons of the correct predicted size (~1.5 kb) were identified by gel electrophoresis, gel purified (Qiagen MinElute™ gel purification kit) and directly sequenced. In total, 19 unique β-ketoacyl synthase sequences were identified in the seven 750,000-membered library aliquots that were examined in detail from the Utah library. 73 unique β-ketoacyl synthase sequences were identified in the twenty 320,000-membered library aliquots that were examined for the Anza Borego library. Library aliquots that failed to yield any PCR products were not included in our analysis. The identification of redundant sequences in a screen of unique library aliquots is a prerequisite to identifying overlapping clones that could be used to reconstruct large natural product gene clusters. The libraries used in these studies were therefore expanded to contain 1.0-1.5x10^7 clones, 4-5 times the number of clones needed to detect the first redundant β-ketoacyl synthase sequences.

2.4.3 Identification of Gene Clusters of Interest

In general, for all degenerate PCR primer designs estimated Tm, G/C content, and secondary structures were taken into consideration to yield

primers that would be suitable for standard PCR conditions and annealing temperatures (55-75°C) during PCR. Multiple sets of suitable primers were typically designed for each biosynthetic target. These primer sets were tested in all combinations across a gradient of annealing temperatures (55-75°C), with touchdown PCR protocols, with and without DMSO, and in conjunction with FailSafe™ buffers A-K (Epicentre Biotechnologies; Δ(betaine\$MgCl_2$)) to determine which conditions and degenerate primer pairs consistently yielded amplicons of correct size and sequence. From here, an optimal set of degenerate primers and conditions was generally selected for all downstream library screening applications. After sequence verifying the degenerate PCR amplicons, via TOPO™ cloning (Invitrogen) or direct sequencing, specific primers were used to recover the cosmids containing the sequences of interest using the clone recovery procedure described in section 2.4.4.

PCR reactions with degenerate primers designed to amplify β-Ketoacyl synthase gene sequences were used to detect type II polyketide synthase (PKS) sequences.(Seow, Meurer et al. 1997; King, Bauer et al. 2009; Kim, Feng et al. 2010) Degenerate primers designed to detect flavin-dependent halogenases (TyrohalF3: 5'-CGGCTGGTTCTGGTACATCCC-3', TyrohalR2: 5'-GAACTCGTAGAASACSCCGTACTC-3') were used to identify the nonribosmal peptide synthetase (NRPS) gene cluster. The FRI gene cluster was identified using primers that recognize conserved sequences in acyl-ACP ligases found in lipopeptide antibiotic gene clusters (DpFrEFWD1: 5'-

TSMTSCAGTACACSTCSGG-3' and DpFrEREV1: 5'-WDGTCGTASGCGAAGTCSG-3'). Type II PKS sequences were amplified using the same PCR conditions outlined for the library size analysis. α–KG mediated halogenases were detected using primers (CmaBF3: 5'-GARGGNACNGACTGGCAYCAG-3' and CmaBR3RC: 5'-TAGTCRTARCCRWAGAMNCC-3'). α–KG mediated halogenase homologue PCR reactions contained the following components: Each 25 μL reaction contained primer added to a final concentration of 2.5 μM, ~1 μL of eDNA template (~100 ng), 1x FailSafe Buffer G (Epicentre Biotechnologies), and 1 U of *Taq* DNA polymerase. Reactions were cycled using the following touchdown protocol: initial denaturation (95°C, 3 min); 25 standard cycles (95°C, 1 min; 65°C, 1 min; 72°C, 30 sec) and a final extension step (72°C, 4 min). FAD-dependent halogenases were amplified using the following PCR conditions: Each 20 μL reaction contained primer added to a final concentration of 2.5 μM, 0.5 μL of eDNA template (~100 ng), 1x FailSafe Buffer G (Epicentre Biotechnologies), and 1 U of *Taq* DNA polymerase. Reactions were cycled using the following touchdown protocol: initial denaturation (95°C, 2 min); 9 touchdown cycles (95°C, 30 sec; 70°C (dt -1°C/cycle), 30 sec; 72°C, 30 sec), 30 standard cycles (95°C, 30 sec; 60°C, 30 sec; 72°C, 30 sec) and a final extension step (72°C, 5 min). The acyl-CoA ligase homologues were identified using the following reaction conditions: 25 μL reactions contained primer added to a final concentration of 2.5 μM, 0.5

μL of eDNA template (~100 ng), 1x ThermoPol Buffer, 2 mM dNTPs, and 0.5 U of *Taq* DNA polymerase. Reactions were cycled using the following touchdown protocol: initial denaturation (95°C, 2 min); 6 touchdown cycles (95°C, 30 sec; 65°C (dt -1°C/cycle), 30 sec; 72°C, 30 sec), 30 standard cycles (95°C, 30 sec; 58°C, 30 sec; 72°C, 30 sec) and a final extension step (72°C, 2 min). Amplicons of the correct predicted size were typically gel purified (Qiagen MinElute™) prior to sequencing.

2.4.4 General Procedure for Clone Recovery

Clones of interest were recovered from the eDNA library using successive rounds of library dilution and PCR screening. As described in the text, sub-libraries with PKS-containing clones were initially identified by PCR screening of the pooled 8 x 8 miniprep grids. Individual clones were recovered from a sub-library of interest by serial dilution of the corresponding glycerol stock and subsequent plating of 20 μL aliquots of the dilute sample into individual wells of two 96-well microtiter plates prefilled with LB-agar (50 μg/mL ampicillin). A 10^{-4} dilution of a cell suspension with an $OD_{600} = 0.2$ was used in the initial round of PCR screening. These dilutions were generally titered with a parallel agar 100 mm^2 petri dish to determine the approximate number of colonies in an individual well of a 96-well plate. Following incubation overnight at 37°C, 25 μL of LB was added to each well and the plates were vortexed for 1-3 min to resuspend the bacterial colonies. After vortexing, an additional 50 μL of LB was added to each well, and 1 μL of

the cell suspension from each well was used as a template in whole cell PCR reactions. Pooled rows (8) were typically screened followed by individual wells in a PCR-positive row (12) to reduce the amount of PCR screening. A single PCR positive well was then subjected to an additional round of dilution and PCR screening. Either a 10^{-5} or 10^{-6} dilution of a cell suspension with an OD_{600} = 0.2 was used for this second round of PCR screening. Individual colonies from PCR positive wells identified in the second round of dilution PCR screening were then screened by colony PCR after plating on LB agar (50 μg/mL ampicillin). A similar dilution strategy using liquid LB-filled 96-well microtiter plates was used to recover some clones.

Individual recovered clones were inoculated into 100-250 mL LB (50 μg/mL ampicillin) for overnight growth at 37 °C (225RPM) and midiprep DNA isolation (Qiagen HiSpeed Midi Kit™). Each cosmid was then end sequenced using vector-specific universal primers (M13-40 and the T7 promoter). We generally determined if a pathway was complete based on homology comparisons between end sequencing data and known biosynthetic sequences. PCR primer sets were designed to these end sequences and were used to rescreen the library if the pathway appeared to be truncated. Overlapping clones identified with the end primer set were then recovered using the same strategy described above. This process was repeated iteratively until the end sequencing of distal and proximal cosmids no longer revealed enzymes suggestive of secondary metabolism. Subsequent full

sequencing of each recovered cosmid clone using transposon-based methods (BB16, LPSG47, LPSG72, LPSG48, BB72, OLI4, ZA41) or 454 GS FLX pyrosequencing (all other cosmids) aided in determining when pathways were fully recovered. The cryptic NRPS gene cluster was recovered on three cosmid isolates (Za41, Q87, J2), the friulimicin-like gene cluster (FRI) was recovered on three cosmid clones (1697, 1451, 201) and the PKS cluster was recovered on two cosmid isolates (X16, V48), the fluostatin gene cluster was recovered on two overlapping cosmids (AB649/1850). All sequence information was assembled using Newbler (Roche), or VELVET (Abulencia, Wyborski et al. 2006), and annotated using Genemark (Lukashin and Borodovsky 1998; Borodovsky, Mills et al. 2003), and homologues were identified using BLASTx.(Altschul, Gish et al. 1990) (http://blast.ncbi.nlm.nih.gov/Blast.cgi) Gene cluster images were generated using MacVector. Sequences for each pathway have been deposited with GenBank sequence deposition info: PKS: GQ452294.2, NRPS: GQ452294.3, FRI: GQ452294.4. AB649/1850: HM193369. The amino acid substrate specificity for each adenylation domain found in the cryptic NRPS gene cluster was predicted using NRPSpredictor. (Rausch, Weber et al. 2005)

CHAPTER 3

3 Assembly and Heterologous Expression of Large Natural Product Gene Clusters using TAR

3.1 Introduction

The heterologous expression of biosynthetic gene clusters cloned directly from environmental DNA can now provide access to the chemical diversity encoded in the genomes of uncultured bacteria. A major challenge facing this approach is that many natural product biosynthetic gene clusters are too large to be readily captured on a single fragment of cloned eDNA. Although larger insert cloning vectors do exist (BAC), soil-based metagenomic libraries are generally not constructed using these systems due to the challenge of efficiently cloning and isolating large quantities of high molecular weight eDNA that exceeds 50 kb in length. (Liles, Williamson et al. 2008) The recovery of large eDNA-derived natural product gene clusters from collections of overlapping cosmid clones represents one potential solution to this problem. (Chapter 2) To discover novel metabolites from eDNA-derived gene clusters isolated using sequence-based strategies, the pathway must be heterologously expressed. Expression systems that depend on the cotransformation of multiple plasmids, however, become prohibitively complex for many heterologous hosts, especially when a gene cluster is

composed of more than two clones. Also, natural product gene clusters often contain large ORFs that are only partially cloned on individual cosmids comprising a complete biosynthetic pathway. These truncated genes cannot be functionally expressed in a multi-plasmid expression strategy. It is therefore necessary to reassemble large natural product gene clusters captured on multiple cosmids into single constructs for downstream heterologous expression studies. Unfortunately, traditional methods for the assembly of large plasmids from constituent fragments of DNA can be technically challenging. Naturally-derived DNA sequences of this length scale often do not contain unique recognition sites, making restriction-mediated cloning strategies impossible. Also, large DNA constructs can suffer from instability and inefficient transformation in *E. coli* which has presented a major hurdle to assembling larger gene systems for other areas of research as well. (Gibson, Benders et al. 2008; Gibson, Glass et al. 2010) In general, the lack of robust and scalable DNA assembly strategies poses a major barrier to eDNA-based natural product discovery efforts. For this reason, all of the secondary metabolites that have been characterized to date from soil-based eDNA libraries have been derived from the heterologous expression of single clones that encode relatively simple chemical structures. (Chapter 1) Successful functional metagenomic natural product discovery studies carried out on marine samples and other microbiomes have also primarily been restricted to single clones. (Schmidt, Nelson et al. 2005)

Lambda-mediated recombination (recombineering) presents one potential solution to engineering natural DNA constructs by overcoming the requirement for unique restriction sites. (Court, Sawitzke et al. 2002; Wenzel, Gross et al. 2005) In these studies, lambda phage recombinase genes (α/β/γ), under the control of an inducible promoter, are co-expressed in the presence of fragments of DNA containing overlapping regions of homology. These regions subsequently undergo homologous recombination to yield an assembled DNA construct. Lambda-mediated recombination in *E. coli* represents a significant technical advance as it does not depend on the availability of unique restriction sites. It was recently applied to functionally reassemble a 43 kb biosynthetic gene cluster from a cultured bacterium (*Stigmatella aurantiaca*) that was heterologously expressed in *E. coli* and *Pseudomonas*. (Wenzel, Gross et al. 2005) We have used this system to retrofit cosmids with integrative elements for *Streptomyces* conjugation and heterologous expression as well. (Chapters 1 and 2) (Tobias Kieser 2000) Unfortunately, recombineering is inefficient and technically challenging when working with large biosynthetic pathways cloned on more than two cosmids. This strategy requires the stepwise introduction of a unique selectable marker for each component of a gene cluster, and also relies on the transformation of large linear DNA constructs into *E. coli,* a highly inefficient process. In our attempts to utilize this system to assemble the NRPS pathway, we have also noticed clone instability and pathway truncations

most likely caused by the overexpression of recombinase genes (data not shown). In general, due to these technical restrictions, lambda recombineering is not a robust or scalable method to rapidly assemble large numbers of multi-clone biosynthetic gene clusters isolated from eDNA libraries.

One alternative to manipulating natural DNA constructs is to engineer a synthetic gene cluster which maintains codon fidelity while offering convenient cloning features. An elegant example of this approach was the total synthesis and assembly of a synthetic 32 kb erythromycin gene cluster. (Kodumal, Patel et al. 2004; Reisinger, Patel et al. 2006) In this work, the sequence of the biosynthetic pathway was determined and computationally redesigned to introduce convenient restriction sites for multi-part assembly while maintaining codon fidelity. (Jayaraj, Reid et al. 2005) The gene cluster was then synthesized in 1 kb fragments which were shuttled on carrier plasmids containing one of two selectable markers. Using a ligation by selection approach, industrial researchers were able construct a functionally intact, codon optimized version of this large biosynthetic gene cluster in a few weeks. (Comeron and Aguade 1998; Menzella, Reid et al. 2005; Menzella, Reisinger et al. 2006) From here, they were able to demonstrate the successful heterologous expression of 6-dEB (deoxyerythronolide), the erythromycin aglycone, in *E. coli*. By using synthetic DNA constructs and a novel assembly method, the authors overcame cloning barriers and

functionally reconstituted the 32 kb biosynthetic pathway in a heterologous host. The restriction sites they engineered into the gene cluster also allowed the later incorporation of alternate biosynthetic modules at defined sites. While this approach offers an elegant solution to gene cluster engineering and heterologous expression, the total synthesis of multiple DNA constructs of this length scale is unfortunately cost prohibitive outside of a commercial context. It is therefore not applicable to large numbers of eDNA-derived biosynthetic pathways.

Many of the pathways recovered in our preliminary sequence-based screens presented in Chapter 2 were found on multiple clones, making heterologous expression studies challenging. To overcome this difficulty, we developed a novel experimental framework that permits the reassembly of large natural product biosynthetic gene clusters on overlapping soil-derived eDNA cosmids using transformation-associated recombination (TAR) in *Saccharomyces cerevisiae*. In this chapter, we demonstrate the utility of this method by rapidly assembling a diverse range of eDNA-derived biosynthetic gene clusters contained within multiple overlapping cosmids. We then demonstrate the TAR-mediated functional reassembly of a large natural product gene cluster which led to the discovery of a novel group of polyketides (Fluostatins F-H). In addition, we also demonstrate the targeted cloning of a sequenced biosynthetic gene cluster from a cultured organism without the need to construct and screen a genomic DNA library. This approach provides

a robust platform to study a more diverse range of large natural product gene clusters which exceed conventional eDNA cloning limits. In doing so, it overcomes a major barrier that prevented the functional characterization of large eDNA-derived natural product gene clusters.

3.2 Results

3.2.1 Transformation Associated Recombination (TAR)-Mediated Assembly of Large Natural Product Gene Clusters

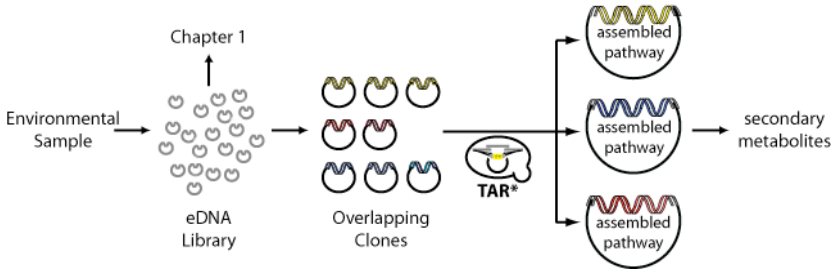


Figure 39: Functional metagenomics with TAR-assembled gene clusters
All metagenomically derived molecules that have been characterized to date have been expressed from single eDNA clones because of the difficulties associated with engineering large biosynthetic pathways. (Chapter 1) TAR now provides access to a much more diverse range of large gene clusters and overcomes conventional eDNA cloning limitations.

Homologous recombination in yeast has been widely used for conventional, restriction-independent cloning. (Ma, Kunes et al. 1987) *S. cerevisiae* offers several advantages as a cloning host when compared to *E. coli,* including the ability to stably maintain DNA constructs that are

hundreds of kilobases to megabases in length. (Murray and Szostak 1983) It is also possible use the native homologous recombination machinery of *S. cerevisiae* to assemble multiple homologous fragments of DNA into a stable plasmid without the need for restriction-mediated ligation and cloning. (Ma, Kunes et al. 1987) In addition, the efficient homologous recombination machinery present in *S. cerevisiae* can be used to modify DNA constructs by simply transforming PCR products or oligonucleotides. (Oldenburg, Vo et al. 1997; Gibson, Benders et al. 2008; Gibson 2009) Transformation associated recombination (TAR), a natural extension of conventional *S. cerevisiae* gap-repair methods, has been recently used to selectively clone a known sequence from a mixture of genomic DNA. In TAR cloning, genomic DNA and a "capture" vector with short homology arms corresponding to sequences flanking the region of interest are co-transformed into *S. cerevisiae*. The capture vector arms and homologous target DNA undergo recombination to yield a stable plasmid containing the targeted genomic region. TAR was originally developed to allow the targeted cloning of large genomic fragments without the need to construct and screen a genomic DNA library. (Larionov, Kouprina et al. 1994; Larionov, Kouprina et al. 1996; Larionov, Kouprina et al. 1996; Kouprina, Graves et al. 1997; Kouprina and Larionov 2006; Kouprina, Noskov et al. 2006; Kouprina and Larionov 2008) Recent studies extended the scope of this methodology by showing that it could be used to assemble 25 co-transformed overlapping DNA fragments into a complete 592

kb synthetic genome in a single reaction, and that multiple PCR products could be assembled into small biochemical pathways. (Gibson, Benders et al. 2008; Gibson, Benders et al. 2008; Shao, Zhao et al. 2009) Based on these experiments, I hypothesized that TAR could provide a rapid gene cluster assembly strategy and could also provide a powerful extension to culture-based natural product gene cluster cloning efforts.

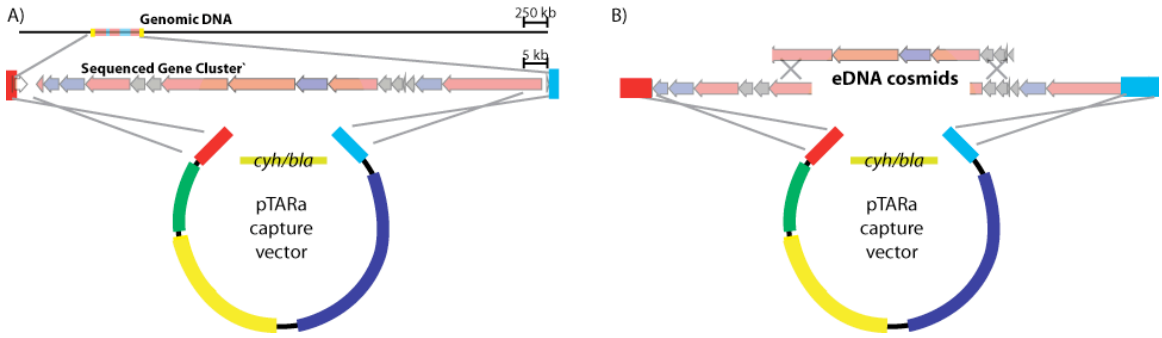


Figure 40: TAR cloning and assembly of natural product gene clusters

TAR can be used to directly clone a natural product gene cluster without constructing and screening a genomic DNA library (A). TAR can also be used to reassemble biosynthetic pathways from multiple eDNA-derived cosmids to functionally reconstitute large natural product gene clusters.

To facilitate TAR reassembly of large natural product gene clusters as well as subsequent heterologous expression studies with reassembled pathways, pTARa, a BAC-based *S. cerevisiae/E. coli/Streptomyces* shuttle capture vector was created (Figure 40). This vector contains elements that allow pathways to be assembled in *S. cerevisiae*, characterized in *E. coli*, and conjugatively transferred into a wide range of Streptomycetes for heterologous expression studies. (Tobias Kieser 2000) We included these elements to facilitate Streptomycete-based heterologous expression studies,

but any number of species-specific genetic elements can be incorporated into pTARa to allow the rapid transfer of pathways into a variety of tractable bacterial hosts. (Wolfgang, Kulasekara et al. 2003; Mathee, Narasimhan et al. 2008) As a demonstration of the utility of pTARa as a shuttle vector, we propagated the vector in *S. cerevisiae* (CRY1-2*::ura-*) selecting for uracil auxotrophy, transformed and isolated the vector from *E. coli,* and successfully conjugated into a number of different Streptomycetes, including *S. toyocaensis, S. lividans,* and *S. albus.*

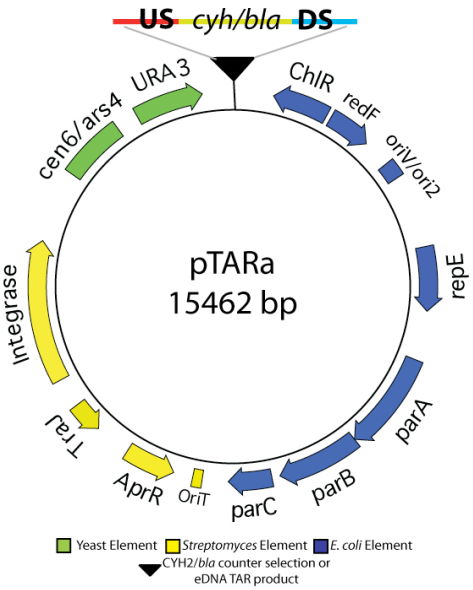


Figure 41: pTARa *S. cerevisiae/E. coli/Streptomyces* shuttle capture vector
pTARa was created to facilitate cloning large natural product gene clusters from sequenced organisms and to reassemble overlapping eDNA clones into single pathways. The capture vector contains elements that allow for the reassembly of the gene clusters in *S. cerevisiae* (green), the stable propagation of the assembled construct in *E. coli*. (blue) and the integration and heterologous expression of the constructs in various *Streptomycetes* (yellow). *Adapted from* (Kim, Feng et al. 2010).

For TAR-mediated assembly of multi-clone gene clusters, we created a unique pathway-specific capture vector containing homology arms corresponding to sequences at the proximal (US) and distal ends (DS) of the gene cluster to be reassembled (Figure 41). The homology arms for reassembly experiments were generated by PCR amplifying approximately 600-900bp of sequence from the proximal and distal cosmids for the pathway to be assembled. It has been shown that as little as 40bp of homology can be used to create capture vector homology arms, but we opted for longer sequence regions to avoid spurious recombination events between conserved sites within biosynthetic gene clusters. (Larionov, Kouprina et al. 1994; Larionov, Kouprina et al. 1996; Larionov, Kouprina et al. 1996; Kouprina, Graves et al. 1997; Kouprina and Larionov 2006; Kouprina, Noskov et al. 2006; Kouprina and Larionov 2008) To increase the proportion of appropriately recombined DNA constructs, we incorporated a counter selection cassette, *cyh2*, amplified from pLLX8. (See Materials and Methods) *cyh2* renders yeast sensitive to cycloheximide so recombination-dependent pathway reassembly events can be counterselected using this compound (Figure 40b). Capture vectors were initially created in *S. cerevisiae* using gap-repair recombination between linearized pTARa, the homology arms for a specific pathway, and a counter selection cassette, *cyh2*, with an ampicillin resistance gene for selection (Figure 42). Each of these components contained 40bp of corresponding homology to allow for gap repair assembly *in vivo*. (See

Materials and Methods) This procedure provided a rapid and highly efficient way to generate pathway-specific capture vectors for downstream gene cluster assembly experiments. (Oldenburg, Vo et al. 1997) In later experiments, we opted to PCR-amplify homology arms containing a unique intermediate restriction site (HpaI) in order to directly ligate the amplified homology arms into pTARa (Figure 42b). Although this approach does not allow for cycloheximide counter selection, we have optimized the assembly procedure to a point where it is not necessary for most experiments.

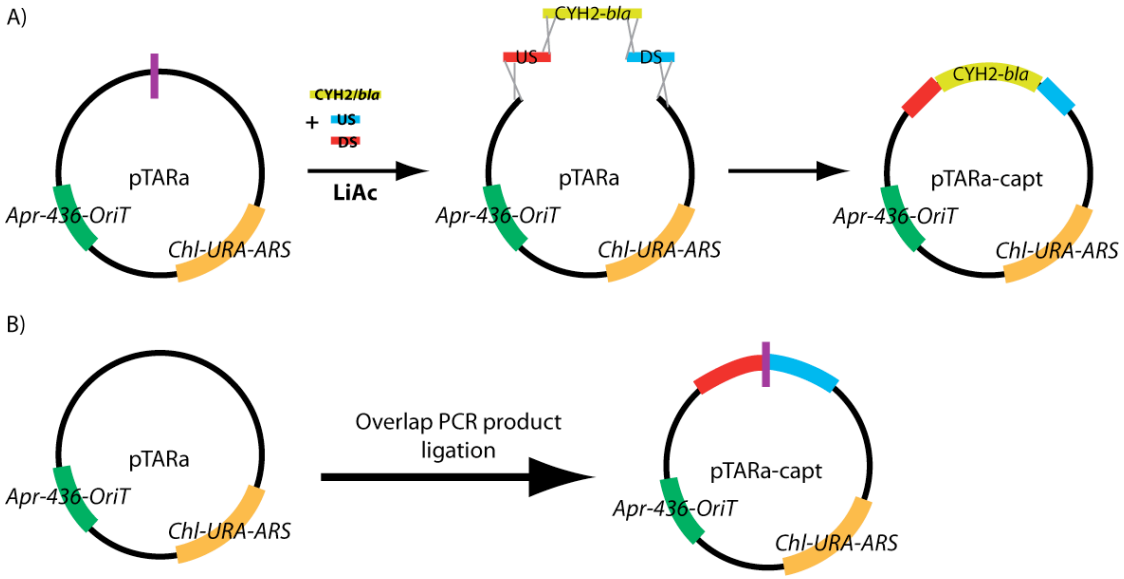


Figure 42: Pathway-specific capture vector assembly

Pathway-specific capture vectors were assembled *in vivo* by transforming chemically competent (LiAc) *S. cerevisiae* with linearized pTARa (purple), up, and downstream homology arms and a counter selection cassette all containing 40bp regions of homology to overlapping components (A). Some capture vectors were also assembled using overlap PCR and direct ligation of homology arms with pTARa (B).

Homologous recombination in *S. cerevisiae* is stimulated by the presence of double stranded breaks adjacent to recombination sites. (Larionov, Kouprina et al. 1994) The individual cosmids to be used in the reassembly of a gene cluster were therefore linearized by restriction digestion with DraI and then co-transformed with a linearized pathway-specific capture vector into competent *S. cerevisiae (ura-)*. DraI, which recognizes the AT-rich hexamer, TTTAAA, digests the cosmid backbone, yet rarely cuts in sequences found in biosynthetic gene clusters. TAR reassembly experiments were also successfully performed with intact circular cosmids, although the percentage of recombination events is significantly higher with linearized DNA (data not shown). We selected the NRPS, PKS, and FRI pathways as model systems (Chapter 2) to use in proof-of-principle TAR reassembly experiments. Each of these pathways was recovered on multiple overlapping cosmid clones from assorted 1-1.5x10^7 member soil eDNA libraries. The concentration of the components used in the co-transformation step was empirically selected to yield, on average, one assembled construct per *S. cerevisiae* spheroplast. This is important, as downstream analysis steps can become complex if more than one assembled construct is present in a single spheroplast. The capture vector and cosmids comprising a complete gene cluster were co-transformed into *S. cerevisiae* spheroplasts along with ssDNA carrier DNA which has been shown to increase transformation efficiency. (Woods and Gietz 2001; Gietz and Schiestl 2007; Gietz and Schiestl 2007)

Each spheroplast recombination reaction was plated in Complete Synthetic Media (CSM-Ura) top agar lacking uracil supplemented with 1M sorbitol and equilibrated to 50°C. For experiments with *in vivo* assembled capture vectors, cycloheximide was added to bottom agar to counterselect for recombined constructs. After 3-5 days of incubation at 30°C on CSM-Uracil dropout agar, recovered spheroplasts were restruck on new CSM-Uracil dropout agar plates (with or without cycloheximide) and grown for 24 hours at 30°C. This step was necessary to reduce background contamination caused by DNA from the original TAR reaction during the downstream PCR analysis. Yeast colonies were screened using multiplex PCR with primers specific to each unique cosmid fragment predicted to be present in a re-assembled gene cluster. Using this approach, *S. cerevisiae* colonies that contained intact biosynthetic gene clusters could be easily identified. In the course of our experiments, between 30-70% of the yeast colonies were found to be PCR-positive for all fragments predicted to be present in a pathway. Reassembly experiments using large PCR amplicons comprising a gene cluster yielded even higher percentages of successfully recombined gene clusters (90%). This is most likely due to the ability to control the amount of overlap between each gene cluster fragment using PCR and the absence of background library vector DNA in these cases.

Assembled gene clusters were isolated from PCR-positive yeast clones and electroporated into *E. coli* for detailed restriction analysis. In each case,

the large construct obtained from a TAR reassembly reaction produced a restriction map that was identical to predicted maps based on sequencing of the individual clones used for reassembly. The 39 kb PKS gene cluster was successfully subcloned from the central region of cosmids X16 and V48, two cosmids that contain 2.1 kb of overlap. The entire 89 kb cryptic NRPS gene cluster described in Chapter 2 was successfully reconstructed in a single *S. cerevisiae* spheroplast transformation reaction from three overlapping eDNA cosmid clones. The 90 kb eDNA-derived FRI gene cluster was also assembled using a single *S. cerevisiae* spheroplast transformation reaction and three overlapping eDNA-derived cosmid clones (Figures 43 and 44).

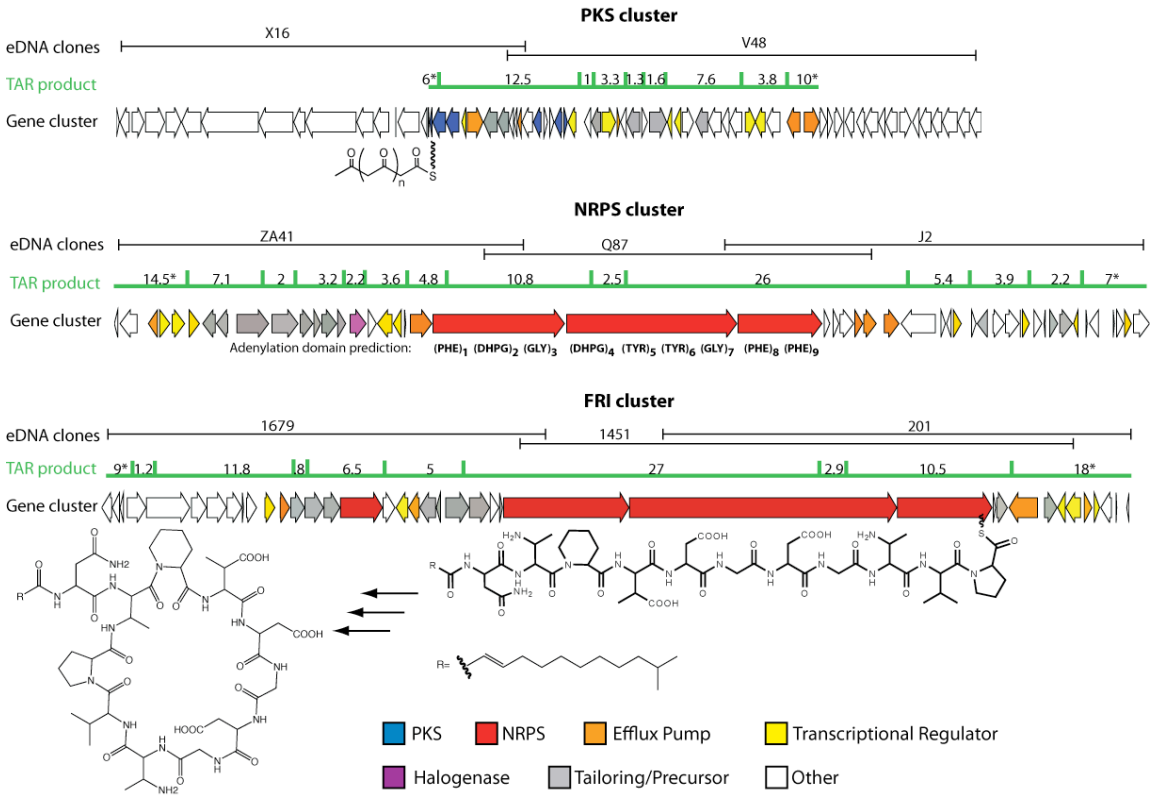
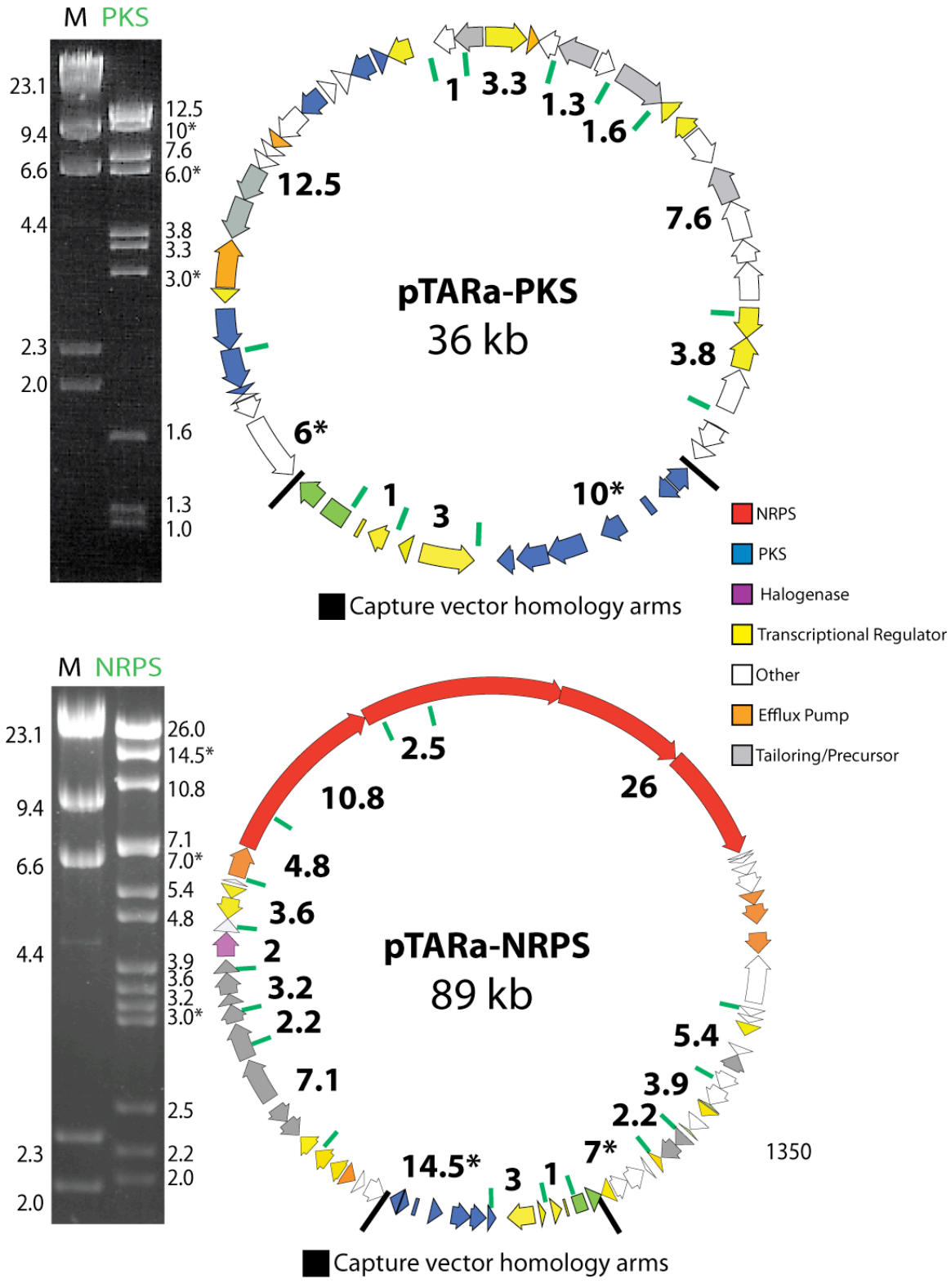


Figure 43: Detailed analysis of TAR-reassembled gene clusters

Gene clusters and annotations of predicted function are shown for three model biosynthetic systems used for TAR reassembly experiments (PKS, NRPS, FRI). The regions that were TAR-cloned are shown in green along with their predicted restriction maps based on sequence information. The cosmids comprising the complete gene cluster are shown in black. Substrate predictions for the NRPS gene cluster were calculated with NRPSPredictor. (Rausch, Weber et al. 2005) The reported structure for friulimicin is shown below the FRI gene cluster. *Adapted from* (Kim, Feng et al. 2010)


M PKS
23.1
9.4
6.6
4.4
2.3
2.0
12.5
10*
7.6
6.0*
3.8
3.3
3.0*
1.6
1.3
1.0
pTARa-PKS
36 kb
1
3.3
1.3
1.6
7.6
3.8
10*
3
1
6*
12.5
Capture vector homology arms
NRPS
PKS
Halogenase
Transcriptional Regulator
Other
Efflux Pump
Tailoring/Precursor
M NRPS
23.1
9.4
6.6
4.4
2.3
2.0
26.0
14.5*
10.8
7.1
7.0*
5.4
4.8
3.9
3.6
3.2
3.0*
2.5
2.2
2.0
pTARa-NRPS
89 kb
2.5
26
10.8
4.8
3.6
2
3.2
2.2
7.1
14.5*
3
1
7*
2.2
3.9
5.4
1350
Capture vector homology arms

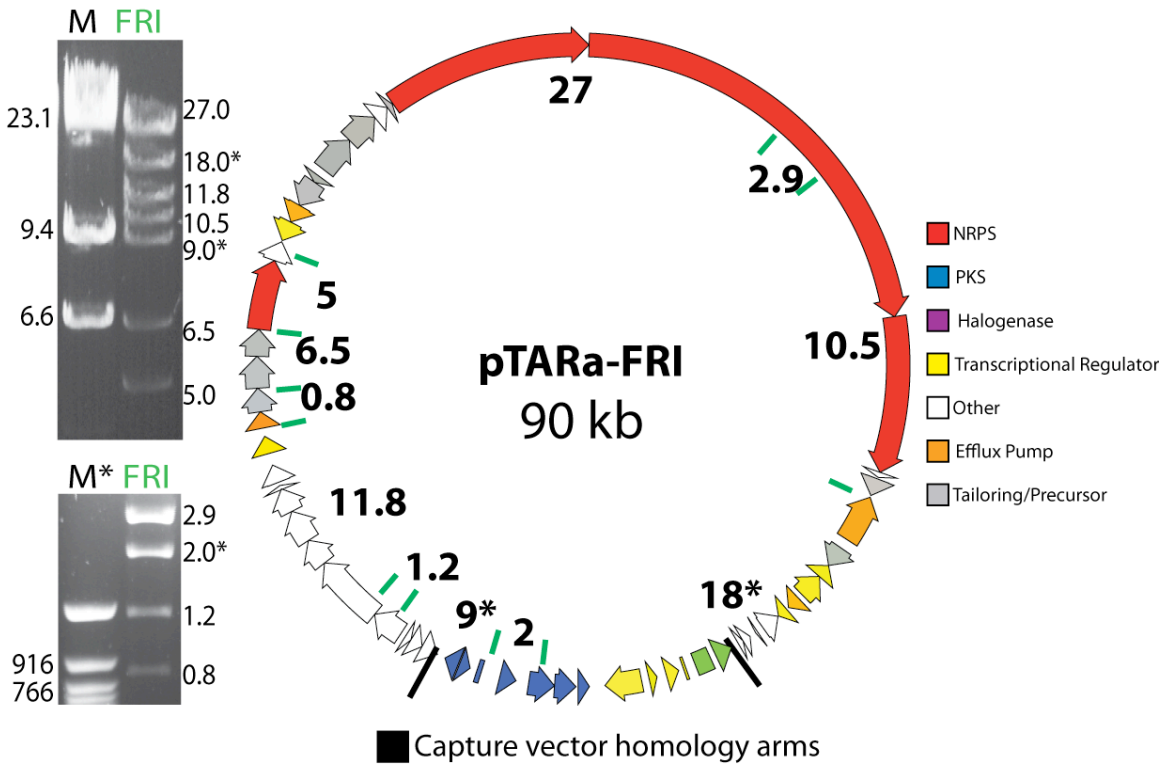


Figure 44: TAR assembly of multi-clone eDNA-derived gene clusters

Detailed restriction maps of TAR reassembled gene clusters are shown with representative annotations colored according to predicted function (Figure 43, Chapter 2). Gel images of the same digest were taken at two different timepoints from the same digest to highlight critical bands for the FRI pathway. Vector specific bands are highlighted (*). *Adapted from* (Kim, Feng et al. 2010)

While the PKS and NRPS gene clusters were initially assembled from fully sequenced sets of cosmids, TAR reassembly experiments can also be performed in the absence of comprehensive sequencing. This represents an additional advantage over other methods, such as lambda-mediated recombination, which typically rely on sequencing before engineering strategies can be designed. The FRI pathway was originally reassembled with only end-sequencing data for each cosmid clone predicted to comprise the complete gene cluster. A capture vector based on end-sequencing data

from the proximal and distal ends of the two outermost clones, cosmids 1679 and 201, was used to reassemble the gene cluster in a single TAR reaction. The successful reassembly of the gene cluster was first confirmed by multiplex PCR and then by comparing detailed restriction maps of the reassembled construct with those produced by the cosmids used in the reassembly experiment (Figure 45). Subsequent full sequencing of the clones comprising the FRI gene cluster confirmed the restriction mapping and successful sequencing-independent TAR assembly experiment.

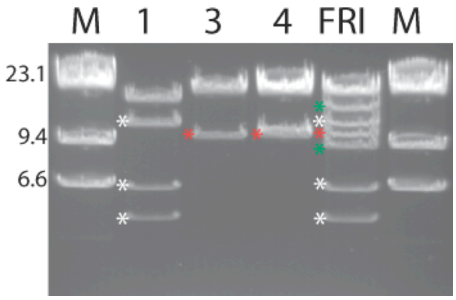


Figure 45: Sequencing-independent TAR assembly of gene clusters

A high molecular weight restriction map (EcoRI) of the three cosmids (1=1679, 3=201, 4=1451) comprising a complete FRI gene cluster suggested that it was successfully reassembled prior to comprehensive sequencing. Conserved fragments are highlighted (*) for each cosmid and the fully assembled FRI construct, and pTARa-specific fragments are shown in green (*). Subsequent full sequencing of the gene cluster and detailed mapping (Figures 43-44) confirmed the successful assembly.

The TAR reassembled constructs were all successfully shuttled into *S. lividans, S. albus, S. toyocaensis, S. lavendulae, S. roseosporus, S. griseus,* and *S. coelicolor* for heterologous expression experiments. A timecourse fermentation in shake flasks using 25 mL of standard defined vegetative and

expression media (R5, R2YE, SAM, SVM) for the individual portions of the NRPS, PKS, and FRI pathways and the fully recombined constructs unfortunately did not reveal any novel clone-specific metabolites. (Tobias Kieser 2000) We also attempted reported fermentation and isolation strategies (Diaion™ resin enrichment and HILIC-LC/MS) for related natural products but were unable to detect any novel clone specific metabolites in the components of and fully reassembled pathways. (Mchenney, Hosted et al. 1998; Miao, Coëffet-Legal et al. 2005; Nguyen, Ritz et al. 2006; Müller, Nolden et al. 2007). Although related natural products have been isolated from cultured bacteria (Daptomycin, CDA, Friulimicin), there has been only one reported case of successfully heterologously expressing a gene cluster of this size and complexity from a cultured bacterium. (Mchenney, Hosted et al. 1998)

Using these three model systems, we demonstrated that TAR can be used to rapidly reassemble collections of clones comprising large natural product gene clusters that exceed conventional eDNA cloning limits. In doing so, this strategy overcomes a major technical barrier which prevented the functional characterization of large biosynthetic pathways isolated from soil eDNA cosmid libraries. To further demonstrate the utility of this platform we pursued additional TAR reassembly experiments on large biosynthetic pathways which contained individual cosmids that produced clone-specific secondary metabolites. Using this strategy, we were able to demonstrate the

functional reassembly and expression of a large biosynthetic gene cluster to discover a novel collection of natural products that could not be derived from the individual cosmids comprising this pathway.

3.2.2.1 Novel fluostatin analogs from TAR assembled gene clusters

To approach the functional reassembly and heterologous expression of large natural product gene clusters using TAR, Zhiyang Feng and I assembled two type II polyketide gene clusters hereon referred to as the 649/1850 and 1105 pathways. A $1.5x10^7$ membered eDNA cosmid library constructed from soil collected in the Anza Borrego desert (AB) was screened for clones containing Type II PKS biosynthetic machinery using reported KSβ-specific primers. (Seow, Meurer et al. 1997; King, Bauer et al. 2009; Kim, Feng et al. 2010) Each minimal PKS containing clone recovered from the library was retrofitted with an integrative expression cassette and conjugatively introduced into *Streptomyces albus* for heterologous expression studies using reported protocols. (Tobias Kieser 2000) Extracts from cultures of *S. albus* transformed with the individual cosmids AB1105 and AB649 were found to contain clone-specific metabolites. AB649 appeared to encode the tetracyclic polyketides rabelomycin (**1**) and dehydrorabelomycin (**2**) (6-hydroxytetrangulol) (Figure 46). (Liu, Parker et al. 1970; Imamura, Kakinuma et al. 1982) Full sequencing of cosmids AB1105 and AB649 suggested that additional biosynthetic ORFs responsible for the production of a secondary metabolite may reside outside of these individual clones. PCR

primers targeting the ends of these cosmids were therefore designed and used to recover clones containing overlapping portions of the biosynthetic gene clusters from the Anza Borego (California) eDNA-library. Clones AB273 and AB949 were recovered using primers designed to detect overlapping fragments of the AB1105 pathway. Clone AB1850, which was recovered using primers designed to detect overlapping fragments of clone AB649, appeared to contain additional biosynthetic machinery that we hypothesized could further modify the metabolites produced by AB649.

To reassemble these gene clusters for heterologous expression studies, the collections of clones (AB273/1105/949 or AB649/AB1850), and a pathway-specific capture vector were co-transformed into *S. cerevisiae* and allowed to recombine in a single TAR reaction using methods described in this chapter. The assembled AB649/1850 gene cluster (GenBank No. HM193369), a >70 kb type II polyketide biosynthetic pathway, was completely sequenced and found to be a faithful reconstruction of the overlapping eDNA sequences captured on AB649 and AB1850. The fully assembled 1105 pathway was conjugated into *S. albus* and *S. lividans* but did not appear to produce any novel pathway-specific metabolites compared to the single AB1105 cosmid control. Further examination of the 1105 pathway and the set of cosmids comprising this gene cluster is currently being pursued by Dimitris Kallifidas. The fully assembled AB649/1850 pathway, which was reconstructed using TAR, was conjugated into *S. albus* and found to confer the production of at least four

additional clone specific metabolites not seen in cultures transformed with AB649 alone (Figure 46).

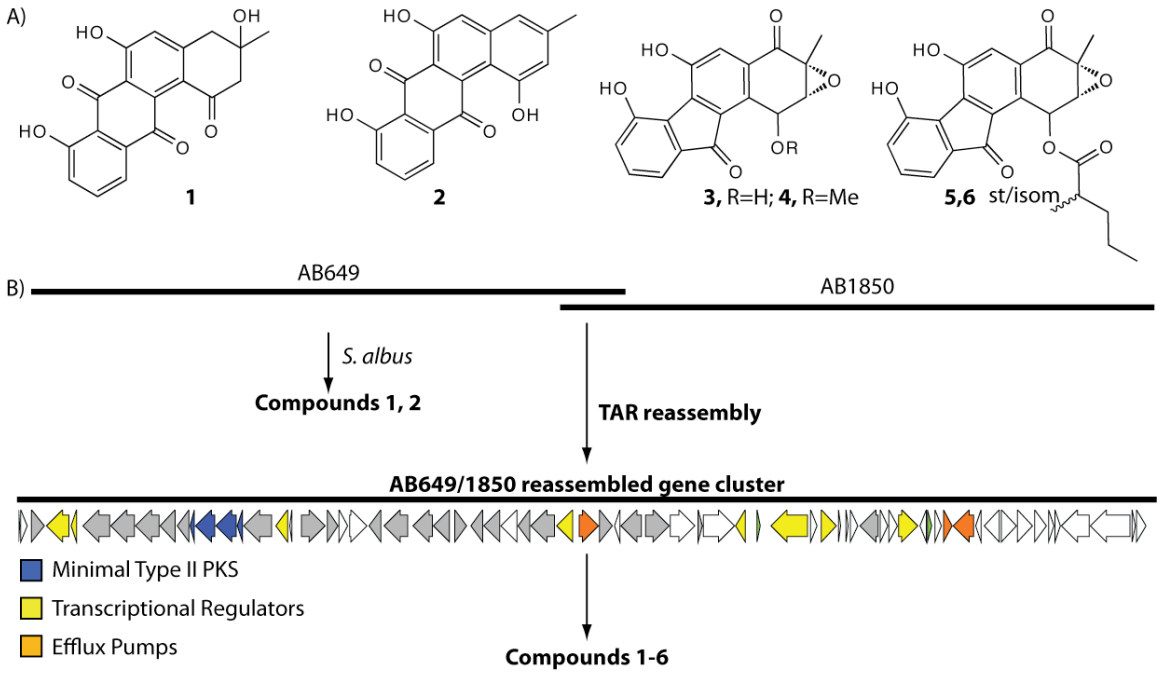


Figure 46: Novel Fluostatins F-H from a TAR assembled biosynthetic pathway
Novel Fluostatins F-H (compounds 4-6) were found in extracts of a TAR reassembled type II PKS gene cluster from a soil eDNA library AB649/1850. These compounds were not found in extracts of individual cosmids. *Adapted from* (Feng, Kim et al. 2010)

Compounds 4, 5, and 6 are novel natural products that have been named Fluostatins F, G, and H (Figure 46). These compounds show moderate antibacterial activity against *Bacillus subtilis* (37.5, 37.5, 21.2 μg/mL MIC respectively). The expression, purification, and characterization of this novel set of compounds from the TAR reassembled AB649/1850 pathway was performed by Zhiyang Feng and Sean Brady. (Feng, Kim et al. 2010)

The reassembly and heterologous expression of the AB649/1850 pathway suggests that TAR will be a generally useful tool for reconstructing large functionally intact biosynthetic gene clusters that exceed conventional eDNA cloning limits. In doing so, TAR significantly increases the value of cosmid eDNA libraries for future natural product discovery efforts. Several additional large gene clusters comprised of overlapping clones have been recovered from screening additional eDNA libraries in the time since these studies. The results presented here suggest that the TAR assembly of these gene clusters could reveal additional novel metabolites that were previously inaccessible using single-clone heterologous expression approaches.

3.2.2 TAR Cloning for Culture-Based Natural Product Research

Recent sequencing projects have revealed that bacterial genomes are rich in natural product gene clusters that are not associated with a reported secondary metabolite. These "cryptic" gene clusters are difficult to manipulate in their nascent genomic setting, making the cloning and engineering of biosynthetic pathways in a heterologous host an attractive strategy. In contrast to cultivation-independent strategies, large quantities of genomic DNA in excess of 1Mb can be isolated from cultured organisms. The cloning of large natural product gene clusters into BAC vectors still remains challenging, however, because the likelihood of capturing an intact biosynthetic pathway on a single clone from a genomic DNA library decreases as the size and complexity of the cluster increases. Therefore, many rounds of

repeated screening and clone isolation are often required before an intact biosynthetic gene cluster can be isolated from a cultured organism. (Mchenney, Hosted et al. 1998; Miao, Coëffet-Legal et al. 2005) In some instances, despite extensive genomic coverage (>15x) with BACs (80-180 kb), larger gene clusters from cultured organisms still cannot be found on single clones. (Wenzel, Gross et al. 2005) To demonstrate the utility of pTARa for culture-based natural products research, we set out to directly clone a large biosynthetic gene cluster without the need to construct and screen a genomic DNA library.

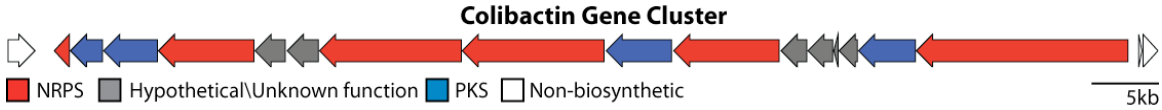


Figure 47: Colibactin gene cluster

The colibactin pathway appears to encode a hybrid PKS/NRPS, which causes mammalian cell cycle arrest upon transient exposure to bacterium harboring this gene cluster. ORFs were annotated based published predictions. (Nougayrède, Homburg et al. 2006; Putze, Hennequin et al. 2009)

In 2006, Nougayrede, et al. reported that a 56 kb NRPS/PKS gene cluster was necessary and sufficient to cause double-stranded DNA breaks and cell cycle arrest in mammalian cells infected with bacteria harboring the pathway. (Nougayrède, Homburg et al. 2006; Putze, Hennequin et al. 2009) Although the authors were able to functionally reconstitute the cell cycle arrest phenotype in a heterologous host, they were unable to characterize a natural product despite comprehensive fermentation efforts including transcriptional monitoring and proteomic analysis. (Homburg, Oswald et al.

2007) The colibactin gene cluster presented a model system to test TAR cloning capabilities as it cannot be captured in its entirety using cosmid/fosmid based approaches due to its size. The sequence of the gene cluster was published in this study (GenBank No. AM229678), and transposon mutagenesis in conjunction with a phenotypic screen identified the functional boundaries of the biosynthetic pathway (Figure 47). (Nougayrède, Homburg et al. 2006; Putze, Hennequin et al. 2009) Using this information, we designed a capture vector that targeted the proximal and distal regions of the gene cluster. We isolated genomic DNA from *Citrobacter koseri*, a sequenced pathogenic bacterium that contains the genomic island, using standard alkaline lysis and alcohol precipitation from 5 mL of overnight culture. 5 μg of the purified genomic DNA was then added to competent yeast spheroplasts along with 1 μg of the purified capture vector. After 72 hours of growth at 30°C on CSM-uracil top agar supplemented with 1 M sorbitol, the TAR captured gene cluster was identified using 8 sets of reported PCR primers. (Nougayrède, Homburg et al. 2006; Putze, Hennequin et al. 2009) From here, yeast spheroplasts containing the intact gene cluster were grown in CSM-uracil dropout media for 24 hours at 30°C (225RPM). The captured gene cluster was then isolated from the yeast spheroplasts, and transformed into *E. coli* for characterization.

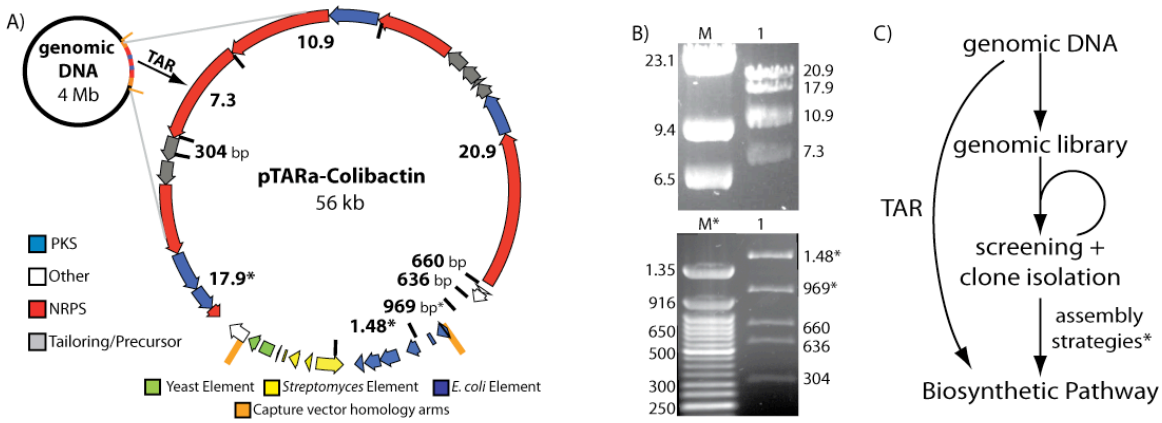


Figure 48: Direct TAR cloning of natural product gene clusters from sequenced organisms

The colibactin gene cluster was directly TAR cloned from *C. koseri* genomic DNA without the need to construct and screen a genomic library (A). Predicted fragment sizes based on the sequenced gene cluster are shown along with annotations of predicted ORF function (A). Detailed restriction mapping of the TAR captured gene cluster in conjunction with multiplex PCR confirmed the successful cloning of the gene cluster (B). TAR provides direct access to sequenced natural product gene clusters and should facilitate heterologous expression experiments for a wide range of biosynthetic systems(C).

Detailed restriction mapping showed that the 56 kb colibactin gene cluster was successfully directly cloned from genomic DNA isolated from the cultured bacterium, *C. koseri* (Figure 46). (Nougayrède, Homburg et al. 2006; Putze, Hennequin et al. 2009) The gene cluster was captured without the need to construct and screen a genomic DNA library, and TAR can be easily adapted to selectively clone any natural product gene cluster from a sequenced bacterium. As an added demonstration of the utility of TAR, we also PCR-amplified large regions of the sequenced colibactin gene cluster and assembled the amplicons into a full pathway using this same capture vector. As demonstrated by these experiments, TAR cloning should provide a general

and rapid means to directly access natural product biosynthetic gene clusters from the genomes of sequenced organisms without traditional cloning and library-based screening limitations.

3.3 Discussion and Future Directions

Previous studies have shown that novel natural products can be heterologously expressed from gene clusters captured on individual soil-derived eDNA clones (Chapter 1). (Brady, Simmons et al. 2009) The metabolites described in these early experiments were derived from single clones primarily due to the challenge of efficiently isolating large gene clusters. We demonstrated in Chapter 2 that sequence-based screens can be used to recover a diverse range of larger natural product gene clusters carried on multiple cosmid clones. This strategy overcomes conventional eDNA cloning and functional screening limitations and provides access to a more diverse range of natural product gene clusters from uncultured bacteria. The lack of robust reassembly methods for DNA constructs of this size, however, limited downstream efforts to heterologously express these large biosynthetic gene clusters. The only novel congeners identified from larger eDNA-derived natural product gene clusters have therefore been generated using the *in vitro* activity of single ORFs or partial pathways that contained enzymes capable of modifying native metabolites produced by a heterologous host. (Banik and Brady 2008)

Here we have shown that TAR can be used to rapidly reassemble multiple overlapping eDNA-derived clones into a single construct containing a large natural product gene cluster. We were also able to demonstrate that TAR can be used to functionally reassemble large biosynthetic pathways which exceed conventional eDNA cloning limitations to uncover novel secondary metabolites (Fluostatins F-H). The maximum number of DNA fragments that can be simultaneously assembled in TAR experiments has yet to be determined, but even the largest gene clusters are unlikely to require more than 3 or 4 overlapping cosmids which is well within the established limits of TAR. (Larionov, Kouprina et al. 1994; Larionov, Kouprina et al. 1996; Larionov, Kouprina et al. 1996; Gibson, Benders et al. 2008; Gibson, Benders et al. 2008; Gibson 2009; Gibson, Glass et al. 2010) We have also shown that TAR can be used to directly and specifically clone large natural product gene clusters from sequenced organisms without the need to construct and screen a genomic library. For culture-dependent natural product discovery efforts, this provides a rapid, single-step method to clone a biosynthetic gene cluster from any sequenced organism for heterologous expression and engineering efforts. The TAR-dependent assembly of biosynthetic pathways from overlapping clones found in eDNA libraries provides a robust platform for accessing functionally intact natural product gene clusters that exceed conventional eDNA cloning limits. In doing so, it eliminates a major barrier which prevented the widespread functional

characterization of large natural product gene clusters found in eDNA libraries.

Many of the cosmids and assembled gene clusters described in previous chapters did not yield novel secondary metabolites. This is due in large part to heterologous expression requirements which represent the most significant challenge preventing the discovery of novel natural products from uncultured bacteria. Sequenced bacterial genomes have revealed a large number of biosynthetic gene clusters in cultured model organisms that are not associated with the production of any small molecules despite extensive fermentation, expression, and engineering efforts. Similarly, only a small percentage of gene clusters isolated from eDNA libraries are predicted to be successfully heterologously expressed in their native form. The reassembly of biosynthetic pathways from cosmid eDNA libraries using TAR represents a critical step toward being able to systematically approach multiple heterologous expression and engineering avenues to access a more diverse range of secondary metabolites from uncultured bacterium.

3.4 Materials and Methods

3.4.1 pTARa Vector Construction

The yeast autonomous replicating sequence (ARSH4), centromeric maintenance element (CEN6), and uracil (URA3) markers were obtained

from pLLX13 by digestion with EcoRI and HindIII. (Wolfgang, Kulasekara et al. 2003; Mathee, Narasimhan et al. 2008) After gel purification, the fragment was ligated into similarly digested pCC1-BAC (Epicentre Biotechnologies). The resulting vector was digested with HpaI and ligated to a DraI fragment from pOJ436 containing an origin of transfer (OriT), integrase, and apramycin resistance gene. (Tobias Kieser 2000) Transformation into EPI300™ *E. coli* (Epicentre Biotechnologies) and selection on chloramphenicol (12.5 μg/mL) and apramycin (50 μg/mL) yielded the capture vector pTARa (TAR-ready BAC with the *Streptomyces* attP integration system, GenBank No.: GQ452294).

3.4.2 TAR Cloning

TAR cloning was initially developed to selectively isolate regions of genomes without the need to construct and screen a genomic library. (Larionov, Kouprina et al. 1994; Larionov, Kouprina et al. 1996; Larionov, Kouprina et al. 1996; Kouprina and Larionov 2006; Kouprina, Noskov et al. 2006; Kouprina and Larionov 2008) The procedures outlined below describe our adaptation of these methods for the isolation of sequenced natural product gene clusters and the assembly of large natural product biosynthetic gene clusters captured on multiple overlapping eDNA clones.

3.4.3 Pathway-Specific Capture Vector Construction

The cycloheximide counter selection cassette (*cyh2*/*bla*) was PCR amplified using pLLX8 as a template following reported protocols. (Mathee,

Narasimhan et al. 2008) The cassette was amplified using primers pLLX8/fw/: 5'-TTTTCTAGAACGCGTTTAATTAAAATCTAAAGTATATATGAGTAAAC-3' and pLLX8/rv/: 5'-CCCTCTAGAGTTAACGTTTAAACAAAAAACGGTGAAAATGGGTGATAG-3'. Each 50 µ L reaction contained 1x FailSafe Buffer B (Epicentre Biotechnologies), 2.5 µM of each primer, 100 ng of pLLX8 template, and 1 U of *Taq* DNA polymerase. Reactions were cycled using the following protocol: initial denaturation (95°C, 2 min), 35 standard cycles of (95°C, 30 sec; 65°C, 30 sec; 72°C, 3 min) and a final extension step (72°C, 7 min). The 2.95 kb PCR product was gel purified prior to capture vector assembly (MinElute™, Qiagen). eDNA clone assembly homology arms were PCR amplified in 25 µL reactions containing 100 ng of template cosmid, 2.5 µM of each primer, 1x FailSafe™ Buffer D (Epicentre Biotechnologies), and 0.5 U *Taq* DNA polymerase. Reactions were cycled using the following protocol: initial denaturation (95°C, 2 min), 35 standard cycles (95°C, 1 min; 60°C, 1 min; 72°C, 1 min) and a final extension step (72°C, 5 min). PCR primers for homology arms were designed to contain 40 bp of homology to the pTARa vector and 40 bp of homology to the counter selection cassette. These homology regions were incorporated to allow pathway-specific capture vector construction using recombination in *S. cerevisiae*. (Mathee, Narasimhan et al. 2008) Upstream homology arm amplification primers contained a sense

primer extension: 5'-ATATTACCCTGTTATCCCTAGCGTAACTATCGATCTCGAG-3', and an antisense primer extension: 5'-CATATATACTTTAGATTTTAATTAAACGCGTTCTAGAAAA-3', which add 40 bp of homology to pTARa and the counter selection cassette, respectively. The downstream targeting sequence sense primer extension is: 5'-CATTTTCACCGTTTTTTGTTTAAACGTTAACTCTAGAGGG-3', which provides homology to the counter selection cassette and the antisense primer extension is: 5'-TAACAGGGTAATATAGAGATCTGGTACCCTGCAGGAGCTC-3', which provides homology to pTARa (Figure 42a). Each primer pair was designed to yield a 600-900 bp amplicon that acts as a homology arm in a pathway-specific capture vector used for a TAR reassembly reaction. (Wolfgang, Kulasekara et al. 2003; Mathee, Narasimhan et al. 2008) Cosmids X16/V48, ZA41/J2 , and 1679/201 were used as templates to generate upstream and downstream homology arms for the PKS, NRPS, and FRI gene clusters respectively. 300 ng of purified *Citrobacter koseri* (GenBank No.: AM229678) genomic DNA (MasterPure™ Complete DNA Purification Kit, Epicentre Biotechnologies) was used as a template to generate upstream and downstream homology arms for the colibactin gene cluster. Each PCR-amplified component was gel purified prior to its use in the assembly of a pathway-specific capture vector.

For the assembly of a pathway-specific capture vector, 200 ng of pTARa was linearized with NheI and added to 200 µg of heat denatured single stranded carrier DNA (heated to 95°C for 10 min then kept on ice), 600 ng of *cyh2*/*bla* counter selection cassette amplicon and 200 ng of an upstream and downstream homology arm amplicon pair prepared as described above. (Mathee, Narasimhan et al. 2008) All components were added to lithium acetate prepared chemically competent CRY1-2 (uracil deficient, *ura*-) yeast, (Gietz and Schiestl 2007) plated on complete synthetic (CSM) uracil dropout agar (Invitrogen) and incubated at 30°C. (Larionov, Kouprina et al. 1994; Larionov, Kouprina et al. 1996; Larionov, Kouprina et al. 1996) Colonies typically began to appear within 24-48 hours. Assembled capture vectors were isolated in bulk by resuspending yeast colonies from a 100 mm CSM-uracil deficient dropout agar plate in 5 mL of 1 x phosphate buffered saline. Plasmid DNA was isolated from 1 mL of this cell suspension (ChargeSwitch™ Yeast Plasmid Isolation Kit, Invitrogen). 100 ng of the purified DNA was transformed into electrocompetent EPI300™ *E. coli* (Epicentre Biotechnologies) and plated on LB agar containing ampicillin (100 µg/mL), chloramphenicol (12.5 µg/mL), and apramycin (50 µg/mL) to yield a pathway-specific capture vector containing a counter-selection cassette.

Table 6: Primers used for pathway-specific capture vector construction

Regions of homology for *in vivo* TAR capture vector assembly in *S. cerevisiae* are shown underlined

TAR primer	Sequence
pLLX8/FW/	5'-TTTTCTAGAACGCGTTTAATTAAAATCTAAAGTATATATGAGTAAAC-3'
pLLX8/RV/	5'-CCCTCTAGAGTTAACGTTTAAACAAAAAACGGTGAAAATGGGTGATAG-3'
PKS(US)/FW/	5'-ATATTACCCTGTTATCCCTAGCGTAACTATCGATCTCGAGTCGACATCCAGCTCAGACAC-3'
PKS(US)/RV/	5'-CATATATACTTTAGATTTTAATTAAACGCGTTCTAGAAAATGCTCGAACTGATCAACGAC-3'
PKS(DS)/FW/	5'-CATTTTCACCGTTTTTTGTTTAAACGTTAACTCTAGAGGGTGCACAATCTGGTGTTCGAG-3'
PKS(DS)/RV/	5'-TAACAGGGTAATATAGAGATCTGGTACCCTGCAGGAGCTCCCGCCTTTTGAACTTCATGT-3'
NRPS(US)/FW/	5'-ATATTACCCTGTTATCCCTAGCGTAACTATCGATCTCGAGGGTGTTCACGTTGTAGCCCT-3'
NRPS(US)/RV/	5'-CATATATACTTTAGATTTTAATTAAACGCGTTCTAGAAAAATGGCCGTAATCACCAGAAG-3'
NRPS(DS)/FW/	5'-CATTTTCACCGTTTTTTGTTTAAACGTTAACTCTAGAGGGGACGTGACGATGGAGATGTG-3'
NRPS(DS)/RV/	5'-TAACAGGGTAATATAGAGATCTGGTACCCTGCAGGAGCTCACTGGTACAGGTTCATGGGC-3'
FRI(US)/FW/	5'-ATATTACCCTGTTATCCCTAGCGTAACTATCGATCTCGAGGTAACGCGACCGGAGCACCATTG-3'
FRI(US)/RV/	5'-CATATATACTTTAGATTTTAATTAAACGCGTTCTAGAAAAGCCGACCACCGCGAAGCACTCT-3'
FRI(DS)/FW/	5'-CATTTTCACCGTTTTTTGTTTAAACGTTAACTCTAGAGGGTGCCGGCCGTTGAGCTTGTTGGTC-3'
FRI(DS)/RV/	5'-TAACAGGGTAATATAGAGATCTGGTACCCTGCAGGAGCTCAGGGGTGGTTCAGCGCCGATGTGG-3'

3.4.4 TAR Cloning and Pathway Assembly

Direct TAR cloning of the colibactin gene cluster from genomic DNA was carried out using reported protocols. (Larionov, Kouprina et al. 1994; Larionov, Kouprina et al. 1996; Larionov, Kouprina et al. 1996; Kouprina, Noskov et al. 2006; Kouprina and Larionov 2008) For eDNA pathway assembly, each cosmid to be used in an assembly reaction was initially linearized by digestion with DraI and the capture vector was linearized by

digestion with PmeI. 200 ng of each linearized cosmid and an equimolar amount (~100 ng) of a linearized pathway-specific capture vector were added to 200 μ L of *S. cerevisiae* spheroplasts prepared as previously reported. (Kouprina and Larionov 2008) The transformed spheroplasts were added to 7 mL of top agar equilibrated to 50°C (1 M sorbitol, 1.92 g/L CSM uracil dropout supplement (Invitrogen), 6.7 g/L yeast nitrogen base (Invitrogen), 2% glucose, 2.5% agar). The top agar containing transformed spheroplasts was overlaid onto CSM dropout agar containing 2.5 μ g/mL cycloheximide. The plates were incubated at 30°C and spheroplast growth was typically seen within 72 hours. The resulting recombinants were patched onto CSM uracil dropout agar with cycloheximide (2.5 μg/mL) for overnight growth at 30°C.

For initial PCR detection of reassembled pathways, a small portion of each yeast patch was resuspended in 10 μ L of 20 mM NaOH and heated at 95°C for 10 minutes. 1.5 μL of the cell lysate was then used as a template in a 50 μ L multiplex PCR reaction following the manufacturer's directions (Multiplex PCR Kit, Q solutionTM, Qiagen). The primer sets used in this analysis were designed to recognize unique regions from each overlapping cosmid clone that was used in an assembly reaction. In the colibactin TAR experiment, PCR primer pairs were designed to detect the previously reported boundaries of the biosynthetic gene cluster. (Nougayrède, Homburg et al. 2006)

3.4.5 Analysis of TAR Recombined Clones

Yeast recombinants that produced PCR amplicons of correct size for all portions of a pathway were grown overnight (30°C, 225 RPM) in 2 mL of SC uracil dropout media or were restruck on CSM uracil dropout agar and grown at 30°C with 2.5 μg/mL cycloheximide and TAR assembled pathways were isolated from these cultures (ChargeSwitch™, Invitrogen). 5 μL of ChargeSwitch™ prepared DNA (1/10 elution volume) was transformed into electrocompetent EPI300™ *E. coli* (Epicentre Biotechnologies) which were outgrown at 30°C for 2 hours (225 RPM) and then plated on LB agar with 12.5 μg/mL chloramphenicol. Whole-cell PCR using the original multiplex PCR pathway detection primers was used to identify *E. coli* colonies containing correctly reassembled gene clusters. DNA was then isolated from 5 mL cultures of PCR positive *E. coli* transformants using alkaline lysis and isopropanol precipitation (CopyControl™ pCC1-BAC Induction Protocol, Epicentre Biotechnologies). *E. coli* transformants containing the colibactin gene cluster were identified using 8 sets of previously reported PCR primers designed to detect different ORFs in the pathway. (Nougayrède, Homburg et al. 2006) Detailed restriction mapping was carried out on each reassembled pathway using an enzyme (PKS, EcoRI; NRPS, EcoRI; FRI, BglII; Colibactin, HindIII) that was predicted to yield restriction fragments that could be easily resolved using agarose gel electrophoresis (1% agarose, 0.5x Tris/Borate/EDTA, 30 V, overnight). In some instances, two different images

of the same digest and electrophoresis experiment were taken at different timepoints to highlight well resolved bands of different sizes (FRI, colibactin). The lambda HindIII and 50 bp molecular weight makers were obtained from New England Biolabs. Full pathway sequencing for each gene cluster was deposited with GenBank under the following accession numbers: NRPS: GQ475282, FRI: GQ475284, PKS: GQ475283, AB649/1850: HM193369.

3.4.6 Fluostatin Characterization

Fluostatin structural elucidation was performed by Dr. Sean Brady. Please reference (Feng, Kim et al. 2010) for details of NMR spectra and HRMS.

3.4.7 Defined Culture Medias

R5 and R2YE were adapted from (Tobias Kieser 2000).

SAM (Streptomyces Antibiotic Medium) and SVM (Streptomyces Vegetative Medium) were adapted from (Lamb, Patel et al. 2006).

CHAPTER 4

4 Characterizing Metagenomic Chemical Diversity

4.1 Introduction

The experiments described in Chapters 2 and 3 demonstrate that it is now possible to systematically recover and functionally reassemble a diverse range of biosynthetic gene clusters from uncultured bacteria using sequence-based screens of large eDNA libraries and TAR. In addition to these technical advances, I became interested in addressing several general questions about natural product biosynthesis in uncultured bacteria: 1) Which uncultured bacteria are associated with natural products found in our environment? 2) Is natural product chemistry conserved between different environments? and 3) How much natural product diversity can be found in nature? These questions are inherently difficult to address due to the complexity of microbiomes, but recent advances in high throughput sequencing and bioinformatics now make it possible. (Margulies, Egholm et al. 2005; Brady and Salzberg 2009) Although 16s rRNA based analyses provide a means of cataloging bacterial diversity, they do not provide insights into the biosynthetic "capacity" of an environmental sample. Early studies have hinted at the biosynthetic diversity present in our environment but were limited to relatively small numbers of sequence variants. (Ginolhac, Jarrin et al. 2004) We set out to

characterize both the phylogeny and biosynthetic capacity of various soil metagenomes using high-throughput sequencing in order to gain broader insight into the chemistry encoded by uncultured bacteria. Three large eDNA libraries were also analyzed using this framework, as they can provide functional access to natural product gene clusters containing sequence variants identified in the screen. Using this approach, we show that several phyla of bacteria not typically associated with secondary metabolism appear to be rich in biosynthetic machinery. We also show that biosynthetic enzymes are highly divergent between different soil microbiomes and approach maximum theoretical evenness, similar to phylogenetic surveys of soil environments. (Dunbar, Ticknor et al. 2000) Together, these results suggest that, in contrast to culture-based approaches, the construction and screening of additional large insert eDNA libraries from soil microbiomes should continue to reveal novel and unique biosynthetic sequence variants that will mostly like not be found in other soil-microbiomes. We anticipate that the experimental framework described here will be a starting point for more sophisticated analyses of natural product chemical diversity in different microbiomes. For example, this approach could be applied to analyze metagenomic chemical diversity under differential physiological states and during pathogenic and symbiotic interactions. (Kroiss, Kaltenpoth et al. 2010) This strategy could also be extended to analyze large numbers of actively expressed biosynthetic enzymes by leveraging recently established

transcriptomics methods. (Wang, Gerstein et al. 2009; Nagalakshmi, Waern et al. 2010) Studies of this nature will help shed further insight into the ecological roles that natural products play in bacterial communities.

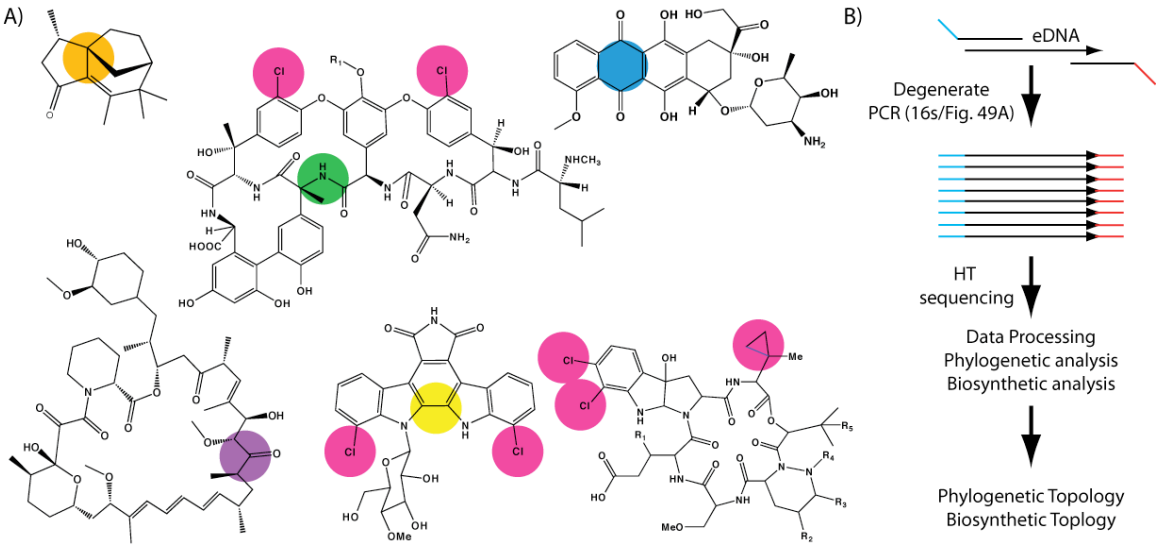


Figure 49: Overview of high-throughput sequencing screens for biosynthetic diversity

eDNA samples and recombinant eDNA libraries were analyzed with degenerate PCR primers designed to target a wide range of canonical biosynthetic motifs (A). These include NRPS (adenylation domains, green), PKS (ketosynthase domains, purple), type II PKS (ketosynthase domains, blue), halogneases (red), oxidative coupling enzymes (yellow), and terpene cyclases (orange). Molecules containing these modifications include albaflavonone (1), glycopeptide-like scaffolds (2), doxyrubicin (3), rapamycin (4), rebeccamycin (5), and kutznerides (6). Adapters and barcodes incorporated into each primer allow for the direct pyrosequencing and simultaneous analysis of multiple samples and biosynthetic enzymes (B).

4.2 Results

4.2.1 Characterizing the Species Diversity of eDNA Libraries.

Small-scale taxonomic surveys of soil-based metagenomic libraries have indicated that a large percentage of cloned eDNA is derived from novel and uncharacterized bacteria. (Rondon, August et al. 2000; Gillespie, Brady et al. 2002; Liles, Manske et al. 2003; Riesenfeld, Goodman et al. 2004) In order to describe the phylogenetic composition of crude eDNA samples and recombinant eDNA libraries in more depth, we analyzed a hypervariable region of 16s ribosomal RNA, a highly conserved component of the 30s prokaryotic ribosomal subunit. (Cole, Chai et al. 2007; Cole, Wang et al. 2009) While other conserved "marker" genes can also be used, 16s rRNA sequencing has become an established metric for comparative metagenomic studies of bacterial populations. (Cole, Chai et al. 2007; Cole, Wang et al. 2009; Wu, Hugenholtz et al. 2009) Also, this ~260bp hypervariable region can be directly mapped to full length 16s rRNA genes with high accuracy allowing robust phylogenetic information to be derived from relatively a small PCR amplicon. The primers we selected for 16s rRNA hypervairable region analysis target more than 72% of all reported prokaryotic 16s rRNA sequences (1,379,424 sequences as of 5/2010) and represent the most comprehensive set of validated degenerate primers available for high throughput sequencing applications. (See Materials and Methods) We elected to analyze each of the recombinant libraries (Anza Borego, Arizona,

Utah) described in earlier chapters, their crude eDNA sources, and six additional samples including eDNA isolated from topsoil collected in New Jersey and Pennsylvania, as well as soybean fields, nalidixic acid treated topsoil, and desert soil from Tanzania.

To analyze the species diversity of recombinant eDNA libraries, they were first processed with ATP-dependent single-stranded nuclease to remove *E.coli* host contaminating genomic DNA. Crude eDNA samples were isolated using reported protocols (Chapters 1 and 2) and further purified with anion exchange resin. (See Materials and Methods) (Brady 2007) We found that this step was necessary to remove co-purified PCR inhibitors (humic acids, polysaccharides, etc.), as many 16s rRNA and biosynthetic amplicons could not be consistently amplified from eDNA purified using standard methods for library construction (Chapters 1 and 2). 454 adapter sequences and unique barcodes were incorporated into each of the 16s rRNA-specific primers. These barcodes allowed us to multiplex the analysis by directly assigning individual pyrosequencing reads to a specific environmental sample. (Hamady, Walker et al. 2008) After determining optimal degenerate PCR amplification conditions, each sample was amplified in three independent reactions. The resulting amplicons were gel purified, fluorometrically analyzed for concentration, and examined using capillary electrophoresis. The 16s rRNA amplicons were then diluted for digital emulsion PCR and each sample was

pooled at equimolar ratios for 454 pyrosequencing according to reported protocols. (Margulies, Egholm et al. 2005) (See Materials and Methods)

After sequencing, the following steps were taken to quality filter the data set prior to generating species assignments: 1) standard flowgram cutoff parameters were used to remove low quality reads from the raw data set (Newbler™, Roche; default parameters) 2) contaminating sequences from the library host (*E. coli*), 3) reads that did not match length thresholds (~260 bp) 4) reads that did not match 454 adapter sequences exactly, indicating sequencing errors or contaminant emulsion PCR products, and 5) reads that contained more than two ambiguous or mismatched bases in primer sequences were all removed. In most phylogenetic analyses of microbiomes, a similarity cutoff of 0.97 is used to define a bacterial species or OTU (operational taxonomic unit). Although this species definition is debated, as bacterial genomes have been shown to be highly dynamic, we elected to use this cutoff to analyze species richness because it is currently the most widely adopted metric for comparative metagenomic analyses of bacterial communities. (Cole, Chai et al. 2007; Cole, Wang et al. 2009) The filtered sequences were split into representative samples using unique barcode identifiers which yielded ~60,000 reads for each recombinant eDNA library and ~20,000 reads for each crude eDNA sample. The number of reads derived from each extract and eDNA library matched our dilution and pooling steps exactly, validating our sequencing preparation procedure.

Two parameters of community structure are generally used to describe microbial populations, α-diversity (the number of species observed within a sample) and β-diversity (the distribution of biological diversity among environments or the number of species shared between two environments). Beta diversity metrics are typically employed to assess the differences between microbial communities in comparative metagenomic studies. (Caporaso, Kuczynski et al. 2010) The output of this analysis is a distance matrix generated by calculating the distance between every pair of samples reflecting the similarity between the samples. Generally, using a basic similarity index, the number of species common to each pair of samples is divided by the total number of recorded species in both samples. (Sorenson 1948) A quantitative refinement of this comparison takes into account the phylogenetic relatedness and relative abundance of each type of sequence/organism in each sample. (Lozupone, Hamady et al. 2007) For this approach, a composite reference OTU\species table (0.97 similarity) is generated for the group of samples being analyzed. This OTU table is taxonomically defined and calibrated by comparing the sequences to curated reference databases. From here, the samples are each analyzed for the numbers and types of reference OTUs identified during species analysis to create an OTU sample description table. A distance matrix can then be calculated from this sample table using standard similarity indices. This distance matrix describing the "relatedness" of the samples can then be

clustered using standard multivariate statistical methods such as principal coordinate analysis (Figure 50a).

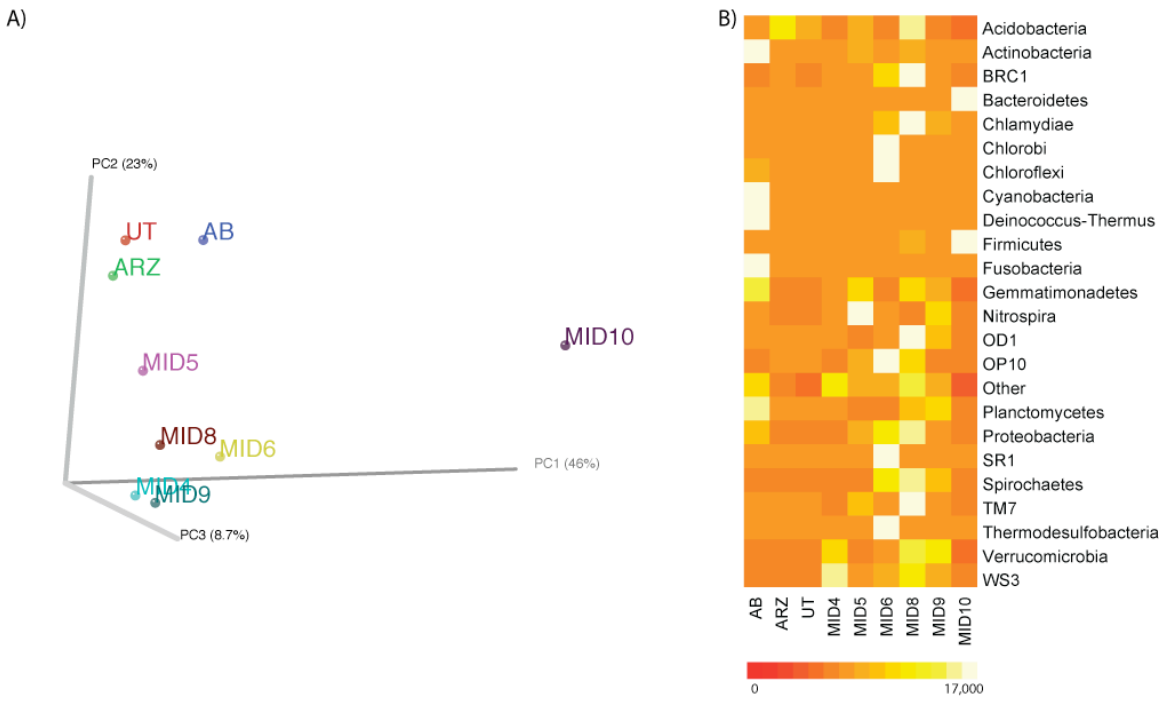


Figure 50: β-diversity analysis of crude eDNA samples

β-diversity analysis of the bacterial species present in 9 crude eDNA samples shows generally similar taxonomic composition with the exception of a nalidixic acid treated sample (MID10) (A) (Please see Materials and Methods for details). A phylum level analysis of these samples (heatmap) shows typical distributions for soil samples (B).

As seen in Figure 50a, the soil-derived crude eDNA samples we analyzed showed relatively similar taxonomic profiles with the exception of the broad-spectrum antibiotic (nalidixic acid) treated sample (MID10). Recent studies have shown that different environmental bacterial and viral communities can contain highly even species distributions despite showing conserved functional footprints within similar microbiomes. (Dunbar, Ticknor et al.

2000; Dinsdale, Edwards et al. 2008) For soil microbiomes, the taxonomic evenness has been reported to be as high as 0.9 (Shannon index, Equation 1). (Dunbar, Ticknor et al. 2000) This indicates that there are large numbers of rare and low abundance bacterial species in soil microbiomes that are sampled only once in sequence surveys. Shannon evenness was calculated for each of the samples using a 0.97 similarity cutoff (Table 7, Equation 1). One advantage of this approach over simple numerical diversity based on clustering is that it takes both species diversity and the relative abundance of each species into account when comparing complex samples.

Table 7: Species evenness of eDNA extracts

Shannon indices were calculated at 0.97 similarity (Equation 1).

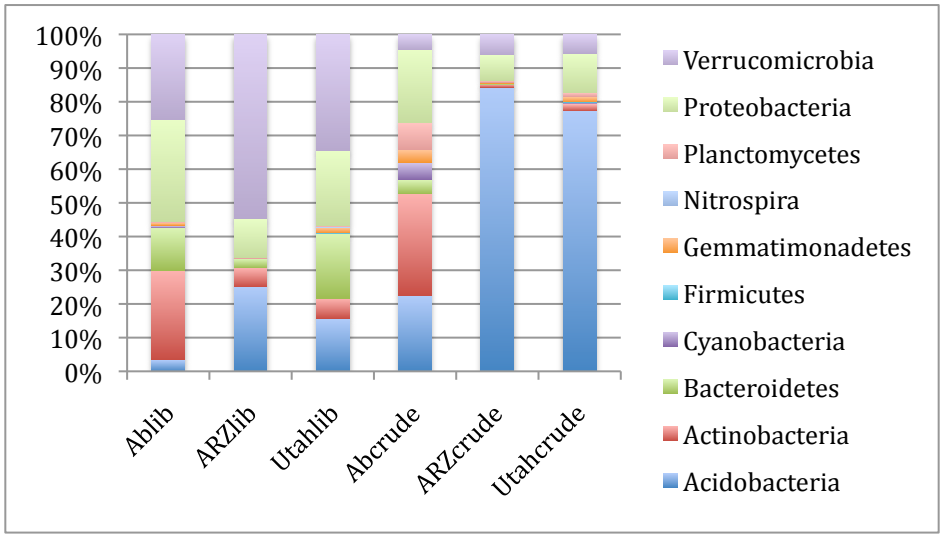
Sample	H'/Hmax (E, evenness)
AB	0.88334
ARZ	0.75975
Utah	0.81083
MID4	0.79283
MID5	0.8489
MID6	0.93501
MID8	0.8157
MID9	0.85599
MID10	0.62757

Similar to other soil microbiome studies, the high level of phylogenetic evenness (theoretical maximum = 1) uncovered in this analysis indicates that despite containing larger proportions of particular phyla, our soil-based eDNA extracts contain high numbers of rarely sampled species within these groups and are generally not dominated by single bacterium. Although this

speaks toward the incredible bacterial diversity present in soil microboimes, it also makes it very challenging to determine statistically significant richness (α-diversity) estimates without substantial sequencing depth. The main purpose of our crude eDNA analysis, however, was to instead gain a taxonomic footprint that would describe the phylum-level distributions of bacteria in each of our eDNA extracts. This taxonomic description can be determined with high accuracy from a much lower number of sequences and was attainable for all of the samples analyzed using a single high-throughput sequencing run (~400,000 reads). (Lozupone, Hamady et al. 2007) To analyze our samples at the phylum level and in more detail, the data was formatted and compared to a curated database of all available annotated 16s rRNA sequences via alignment with PyNAST and INFERNAL. (Cole, Chai et al. 2007; Cole, Wang et al. 2009; Caporaso, Bittinger et al. 2010; Caporaso, Kuczynski et al. 2010) This allowed the rapid taxonomic assignment of each sequencing read as defined by reference sequences in the Ribosomal Database Project (RDP II) (Figures 51 and 52). (Cole, Chai et al. 2007; Cole, Wang et al. 2009) This analysis indicated that our samples contained high percentages of Verrucomicrobial, Proteobacterial and Acidobacterial assignments which are typical of soil microbiomes (Figure 50b). (Torsvik, Øvreås et al. 2002; Daniel 2005)

To determine if our library cloning techniques capture an even representation of bacterial diversity from crude eDNA samples, we examined

the taxonomic assignments for crude eDNA extracts and resulting eDNA libraries for the Anza Borego (California), Arizona, and Utah samples. As seen in the phylum-level assignments, the recombinant libraries derived from these samples contained fewer Acidobacterial assignments and a slightly enriched proportion of Verrucomicrobial assignments compared to crude eDNA extracts (Figure 51).



	Anza Borego	Arizona	Utah
Acidobacteria	0.190043059	0.59168593	0.61633968
Actinobacteria	0.038406881	0.046218037	0.040408389
Bacteroidetes	0.085689392	0.027190508	0.192826705
Cyanobacteria	0.043671881	0	0.005546745
Firmicutes	0.000492186	0.001572052	0.000301751
Gemmatimonadetes	0.029274094	0.002725716	0.000918514
Nitrospira	0	1.69173E-05	0.00110973
Planctomycetes	0.077287281	0.002635211	0.008013166
Proteobacteria	0.085001253	0.039219009	0.105673733
Verrucomicrobia	0.208484738	0.485974438	0.289264272

Figure 51: Phylogenetic analysis of crude eDNA and associated library samples. Phylum-level distributions of eDNA libraries and their crude samples (A). Euclidian distances between the relative abundance of major phyla detected in crude eDNA and library samples show an enrichment of Verrucomicrobial and depletion of Acidobacterial clones (B).

Most model prokaryotic and eukaryotic hosts display an intrinsic bias in the types of DNA (GC/AT content, etc.) that can be stably maintained in clone libraries, so this result is not entirely unexpected. (Comeron and Aguade 1998) Although sequenced Acidobacteria contain relatively similar levels of nucleotide DNA content (average values of 0.578 and 0.518 GC content for *Acidobacterium*/*Escherichia* respectively) it is possible that the heterologous expression of certain genes or gene clusters driven by native promoters may be toxic to *E. coli*. This would prevent the stable maintenance of clones containing these genes and would shift their relative abundance. α-diversity (species richness) estimates of our library samples indicate that between ~5,000-20,000 distinct bacterial species are likely captured in each recombinant eDNA library (Table 8). In this case, rarefaction analysis indicated that it was possible to generate robust α-diversity estimates due to the much lower complexity contained within our eDNA libraries versus crude eDNA extracts.

Table 8: Species richness of eDNA libraries

Library Sample	α-diversity estimator (Chao1, Equation 2)
Anza-Borego	5488.68
Arizona	7775.63
Utah	20675.75

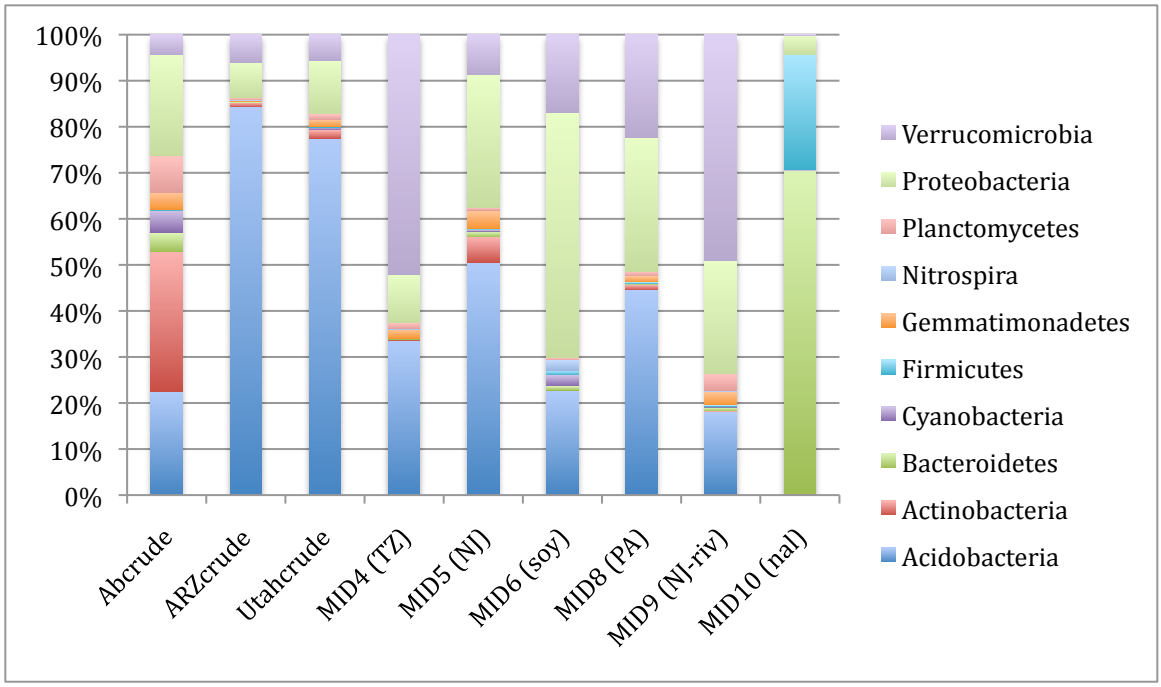


Figure 52: Phylogenetic analysis of crude eDNA samples

16s rRNA analysis of crude eDNA show that the samples contained high proportions of Proteobacteria, Verrucomicrobia, and Acidobacteria.

In general, the phylum-level analyses of crude eDNA samples indicate that eDNA isolation methods capture a representative snapshot of bacterial diversity as compared to reported analyses of soil microbiomes (Figure 52). (Torsvik, Salte et al. 1990; Daniel 2005) Although the resulting recombinant libraries appear to contain fewer Acidobacterial clones and a slight enrichment of Verrucomicrobial assignments compared to crude eDNA extracts, the remaining major bacterial divisions appear to be proportionally represented (Figure 51b). Also, the majority of 16s rRNA reads found in the recombinant eDNA libraries map to uncultured bacteria as defined by reference sequence databases indicating that our libraries should provide access to novel biosynthetic sequence space which has been empirically

demonstrated in Chapters 1-3. (Cole, Chai et al. 2007; Cole, Wang et al. 2009) The Anza Borego (California) eDNA library contains a higher percentage of Actinobacterial assignments compared to other extracts. Actinomycetes have been a productive source of natural product chemical diversity using fermentation-based approaches (Chapter 1: Figure 1). This sample therefore provided a convenient control in the following experiments to test if the analogy of Actinomycete-based biosynthetic richness extends to uncultured bacterial communities.

4.2.2 Characterizing Biosynthetic Diversity in Soil Microbiomes

The complexity of terrestrial microbiomes has made the analysis of total biosynthetic enzyme diversity challenging. Early studies focused on analyzing groups of sequences from single environmental samples. (Metsä-Ketelä, Halo et al. 2002; Ginolhac, Jarrin et al. 2004; Daniel 2005; Gentry, Wickham et al. 2006) This was mainly a limitation imposed by sequencing throughput but these studies showed that a number of novel biosynthetic homologues could be detected in environmental samples using degenerate PCR. (Seow, Meurer et al. 1997; Ginolhac, Jarrin et al. 2004) Due to dramatic advances in high-throughput sequencing technology, the large-scale, parallel analysis of millions of sequences from various environmental samples can now be achieved. (Margulies, Egholm et al. 2005) To characterize the chemical diversity present in our environmental samples, we designed a panel of modified PCR primers targeting different biosynthetic

enzymes for high-throughput pyrosequencing analysis. (Margulies, Egholm et al. 2005) These primers include regions of degenerate homology for the selected biosynthetic target and a 454 adapter oligomer which allows for the direct sequencing of the PCR amplicons. Additionally, unique hamming barcodes were incorporated into each primer to multiplex the analysis and to identify the source of individual sequencing reads in the data set. (See Materials and Methods) (Hamady, Walker et al. 2008)

The domains we chose to analyze include nonribosomal peptide synthetases (AD: adenylation domains), type I polyketide synthases (KS: ketosynthase domains), terpene cyclases, type II polyketide synthases (KSβ), FADH-dependent halogenases, class-specific oxidative tailoring enzymes (indolocarbazole), and a recently described broad-spectrum antibiotic resistance mechanism (nitric oxide synthases) (Figure 49). These biosynthetic enzymes were selected because they are responsible for catalyzing the formation of a large percentage of the natural products that have been characterized from cultured organisms. (Laatsch 2009) While we chose these particular enzymes to span a range of chemical diversity, additional enzymes and degenerate primers could easily be incorporated into this framework for future studies (Figure 26).

To design high-throughput sequencing-compatible primers, a multiple sequence alignment of publicly available homologues for each class of enzyme was used to find regions of nucleotide conservation, ideally less than 500 bp

apart, containing no homopolymer runs, and with similar predicted oligonucleotide annealing temperatures. These parameters were chosen to be compatible with emulsion PCR and pyrosequencing protocols. (Margulies, Egholm et al. 2005) Each primer was tested with a gradient of annealing temperatures (55-75°C) and a panel of buffers (FailSafe™ A-K, Epicentre; Phusion HF/GC™, Finnzymes) in order to find suitable conditions for the consistent amplification of homologues from library and crude eDNA samples. We elected to analyze library samples with additional biosynthetic primer sets because they provide the opportunity to functionally access gene clusters containing any novel sequence variants identified in this screen. Crude eDNA samples could also be analyzed using these additional primer sets, but we examined fewer functional groups (16s, AD, KS) in these cases due to limiting amounts of remaining sample for crude library eDNA preparations. Once an amplicon of expected size was generated for each functional group using optimized conditions, the PCR product was gel purified and cloned, and, on average, 10 recombinants were sequenced to verify their identity. Adenylation domains, ketosynthase domains, KSβ domains, terpene cyclases, nitric oxide synthases, halogenases, and oxidative coupling enzymes were all successfully amplified from library samples using this strategy. Similarly, adenylation and ketosynthase domains were successfully amplified from all crude eDNA samples using this approach. (See Materials and Methods).

Each amplicon generated in the PCR screen was gel purified a second time, fluorometrically quantitated (PicoGreen QuantIT™ Invitrogen), and analyzed via capillary electrophoresis (DNA 7500™ Agilent Technologies). The purpose of these steps is to avoid potential primer contamination in downstream emulsion PCR steps. (Williams, Peisajovich et al. 2006) Some biosynthetic targets yielded amplicons which were >500 bp due to the lack of conserved regions in sequence alignments that were closer in proximity. This led to a clear amplification bias as determined by biotinylated sequence capture and enrichment during pilot experiments (data not shown). (Margulies, Egholm et al. 2005) Therefore, to compensate for the decreased emulsion PCR efficiency, an increased number of relative molecules of these templates were added to the final pools prior to sequencing preparation. This amplicon pooling strategy worked as expected in all cases except for type I ketosynthase amplicons which yielded a slightly reduced number of total reads. (Margulies, Egholm et al. 2005) Library samples were screened with a modified procedure that allowed us to assign 817 unique barcodes to each pooled library subarray (~40,000 clones) from each eDNA library. This framework allowed for the rapid identification and recovery of cosmids containing sequences of interest by revealing where in an arrayed eDNA library a given sequence was found (Figure 53). In this approach, primers were aliquoted into 384 well plates and each individual row of the arrayed libraries were PCR amplified (AD, KS) in separate reactions (1634 total PCR

reactions). From here, samples were pooled into master aliquots which were subsequently gel purified, fluorometrically quantitated and quality checked via capillary electrophoresis as described earlier. The amplicons were then diluted and pooled at proportional ratios for emulsion PCR and pyrosequencing. (Margulies, Egholm et al. 2005)

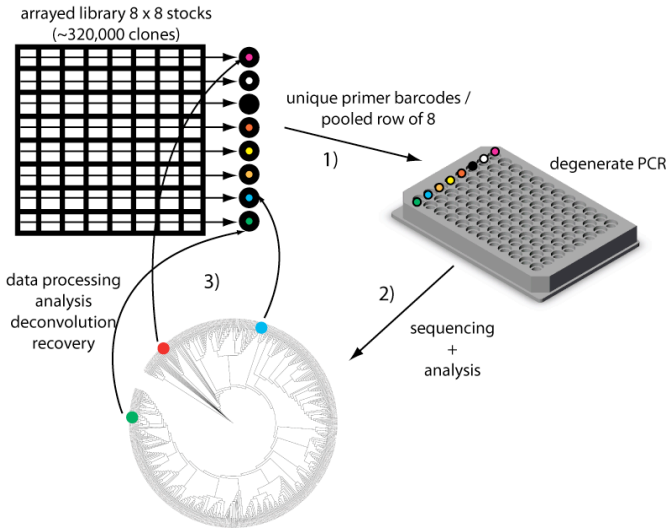


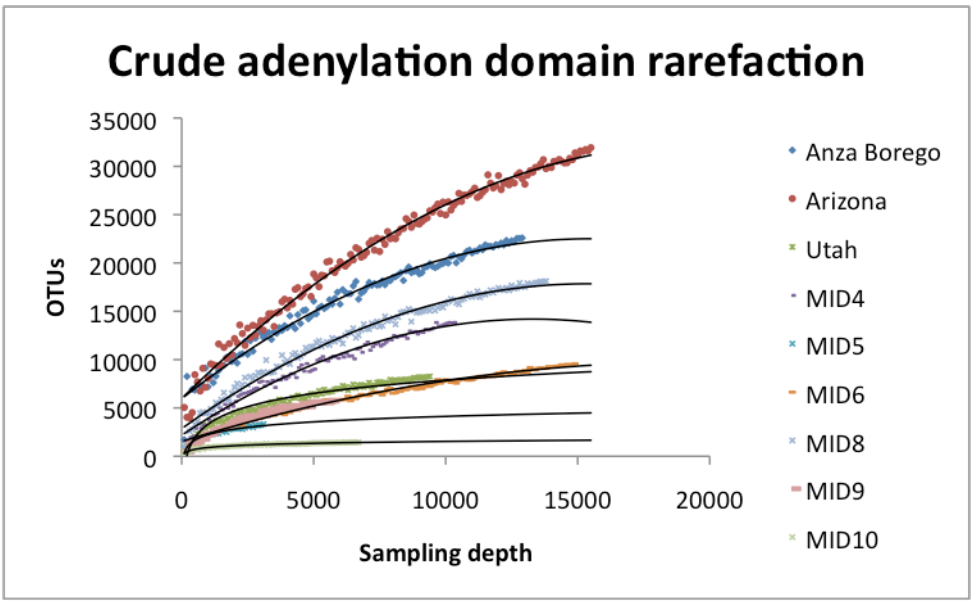
Figure 53: Barcoded clone recovery strategy

Unique hamming barcodes assigned to each pooled row of an arrayed eDNA library are used to individually amplify biosynthetic sequence variants (1). After sequencing and analysis (2), the barcodes associated with a given pyrosequencing read can reveal the specific location within a library where a cosmid containing the sequence is found (3). This screening strategy should provide a rapid method to identify and recover groups of clones for future heterologous expression efforts.

Approximately 400,000 reads from the sequencing experiment could be mapped to biosynthetic amplicons derived from screening library and crude eDNA samples. The remaining 400,000 reads of the pyrosequencing run were allotted to 16s rRNA analysis of the eDNA extracts and libraries. To parse

this mixed dataset, a custom PERL script was used to first remove any reads that did not match quality cutoffs or length requirements as defined by the expected amplicons size ranges for each functional group. Hamming barcode correction was also used to deconvolute and assign each sequencing read to its appropriate biosynthetic functional group, sample, and arrayed library subfraction for downstream analysis. (Hamady, Walker et al. 2008) Rarefaction plots were generated from adenylation domain sequences that were grouped at 0.97 similarity (Figure 54). Robust α-diversity (richness) estimates were difficult to obtain from crude eDNA extracts for ketosynthase domains, similar to species analysis, because the sequence variants approached the theoretical maximum Shannon evenness of 1 in certain cases. Rarefaction analysis indicated that the crude eDNA extracts were undersampled for ketosynthase homolgoues (R^2=1 for a linear curve-fit). A larger proportion of the sequencing experiment could be easily assigned to ketosynthase domain homologues to address this in future experiments. (See Materials and Methods) In a similar strategy to crude eDNA species analysis, however, we were primarily interested in deriving a biosynthetic footprint (β-diversity) from each sample which was easily attainable with a much lower number of reads (Figure 55). (Lozupone, Hamady et al. 2007) Less than 2% of these experimentally derived biosynthetic homologues map to sequenced bacterial genomes (similarity cutoff = 0.97). Sequenced and

cultured bacteria therefore appear to represent only a limited subset of organisms that encode natural products in soil microbiomes.



	Adenylation domains
Anza Borego	22750.77
Arizona	32043.99837
Utah	8248.850227
MID4 (TZ)	13815.65868
MID5 (NJ)	3343.511538
MID6 (soy)	9649.75
MID8 (PA)	18108.31584
MID9 (NJR)	5746.687166
MID10(nal)	1440.489247

Figure 54: Adenylation domain analysis of crude eDNA samples

Rarefaction curves were generated from subsampled OTU tables generated in the primary clustering analysis (0.97 similarity). The same OTU tables were used for richness estimates (Equation 2).

The next analysis measures we undertook were β-diversity comparisons of adenylation and ketosynthase domains between the crude eDNA samples (Figure 54). We hypothesized that biosynthetic enzyme

clustering patterns should match phylogenetic profiles if the genes are evenly distributed throughout all phyla and if they are tightly linked to their specific bacterial sources. Interestingly, the biosynthetic beta indices show little correlation to taxonomically derived data (Figure 54). While it is possible that the primer sets used in the analysis are biased toward specific phyla of bacteria, the independent phylogenetic classification of experimentally derived sequences (Section 4.2.3) indicates that both the adenylation and ketosynthase domain primers are not biased. It has been shown that natural product biosynthetic enzymes are often highly conserved within the same phylum of cultured bacteria. (Ayuso-Sacido and Genilloud 2005) These primer sets should therefore function with relatively even efficiency within a bacterial phylum. Taxonomically similar samples should consequently yield related groups of biosynthetic sequences. As seen in the analysis, this does not appear to be true for the soil microbiomes we analyzed (Figure 55). The large scale sequencing of bacterial genomes has revealed that, on average, smaller genomes contain proportionally fewer canonical natural product gene clusters. This trend could extend to phyla of bacteria which would contain species with lower numbers of secondary metabolic genes. This phenomenon should still, in theory, maintain β-diversity clustering patterns for samples that are taxonomically similar which we did not observe on a global scale (Figure 55). In some instances, samples that are phylogeneticially distinct (eg.MID6/Utah, MID6/MID10), contain more highly related biosynthetic

sequence variants compared to taxonomically similar samples. One possible explanation for this trend is that individual shared bacterial species which do not represent the macro-scale phylogenetic composition of a sample are causing biosynthetic enzyme clustering shifts. This is highly unlikely for several reasons. First, the samples contained a low percentage of shared species although the overall taxonomic composition of the soils was generally conserved (high Shannon evenness). While it is possible that these primer sets exclusively recognize only the small percentage of shared bacterial species present in each sample, which could interfere with correlations, it is improbable. Second, these biosynthetic enzyme profiles represent tens of thousands of sequence variants which are most likely not derived from single, or small numbers of shared bacterial species. For the samples that are taxonomically similar which contain distinct biosynthetic profiles, it is possible that the primer sets recognize only certain subsets of bacterial species within a phylum for particular samples. This is also highly unlikely as described earlier. (Ayuso-Sacido and Genilloud 2005) Overall, it does not appear as though soil microbiomes encode biosynthetic genes that are tightly correlated with specific bacterial phylogenetic compositions (Figure 55). This would suggest that other factors are influencing the distribution and diversity of biosynthetic enzymes found in soil microbiomes.

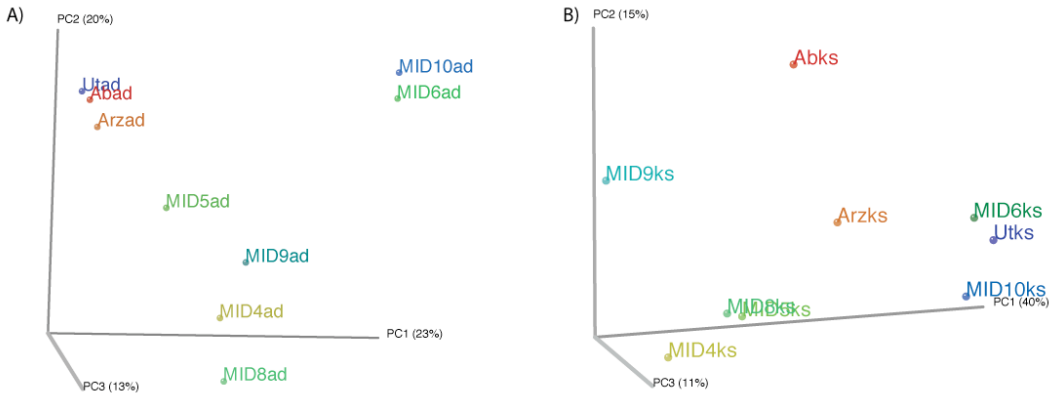


Figure 55: Adenylation and ketosynthase domain β-diversity in crude eDNA

β-diversity analysis of adenylation and ketosynthase domains present in crude samples shows no correlation to species characterizations of the same samples. This indicates that there is little biosynthetic functional conservation between the samples.

Pairwise Sorenson similarity comparisons of biosynthetic enzyme classes that were analyzed in our eDNA libraries approached theoretical minima and revealed that each library contained almost entirely distinct groups of biosynthetic sequence variants (Table 9, Equation 3). (Sorenson 1948) (See Materials and Methods)

Table 9: Sorenson similarity indices

* The Utah library did not yield any true biosynthetic oxidase (StaO-like) sequence variants (Equation 3).

Sample	AD	KS	Hal	Type II PKS	Terp cyc	bNOS	StaO
AB-Arz	0.0037	0.0096	0.0042	0.0709	0.0241	0.0042	0.2555
AB-Utah	0.0043	0.0000	0.0221	0.0472	0.0066	0.0117	*
Arz-Utah	0.0123	0.0173	0.0154	0.0161	0.0570	0.0036	*

Recent experiments have shown that related environmental bacterial and viral microbiomes display low functional evenness (similar metabolic profiles) despite containing variable taxonomic distributions. (Dinsdale, Edwards et al. 2008) Specifically, these studies have shown that a small number of dominant primary metabolic genes are found in related environmental microbiomes regardless of taxonomic composition and diversity. As seen here, it does not appear as though soil microbiomes are dominated by a small set of common secondary metabolic sequences in contrast to environmental surveys of primary metabolism. (Dinsdale, Edwards et al. 2008) We hypothesize that the divergence in secondary metabolic clustering is influenced by independent factors such as specific subgroups of sympatric bacteria or, potentially, viral populations and horizontal gene transfer events. Many natural product gene clusters contain transposon-like genetic elements which indicate that horizontal gene transfer between sympatric microbes may be responsible for natural product gene diversity patterns. (Frigaard, Martinez et al. 2006) In some dramatic cases such as the colibactin gene cluster described in Chapter 3, the nucleotide sequence of the entire 56 kb gene cluster is 100% conserved between different species of bacteria (*E. coli* and *C. koseri*), indicating a relatively recent horizontal acquisition. The continued screening of cultured enteric bacterial isolates have shown that this gene cluster is present in a large number of diverse bacterial hosts. (Putze, Hennequin et al. 2009) The unique clustering

of biosynthetic sequence variants could be explained by selective pressures which favor the frequency of certain natural product biosynthetic enzymes over others regardless of taxonomic background. These populations of sequence variants could, in theory, be mobilized by horizontal gene transfer events rather than the presence of specific bacterial species. (Frigaard, Martinez et al. 2006) It has recently been shown that conserved microbial consortia can contain entirely distinct viral populations. (Reyes, Haynes et al. 2010) Studies also suggest that virus populations paradoxically contain a large and unstudied reservoir of secondary metabolic genes. (Dinsdale, Edwards et al. 2008) There are an estimated 10^{31} phage on the planet, so the potential diversity of genes that can be transferred by this population is enormous. (Hendrix 2002; Breitbart and Rohwer 2005) An analysis of the viromes present in soil samples could reveal insights into the specific types of natural product genes being mobilized by viral populations which could, in turn, influence microbial secondary metabolism. The more detailed analyses of different soil viromes and transposable elements associated with natural product gene clusters will be required to examine these potential factors in more detail.

We next estimated the number of unique biosynthetic homologues (0.97 similarity) found in each of our eDNA libraries (Table 9). The eDNA library samples contained a lower overall number of sequence variants compared to crude samples making this analysis possible. (Please see

Appendix) It is important to note that the similarity cutoff used for this approach (0.97) does not necessarily indicate a functionally and biosynthetically distinct enzyme. These richness estimates should, however, provide a baseline value of biosynthetic enzyme diversity for future comparative analyses. The idolocarbazole oxidase homologues yielded many sequence variants which showed little homology to known biosynthetic genes upon more detailed BLAST analysis. We hypothesize that these sequences originate from contaminating primary metabolic enzymes, which contain the same nucleotide conservations found in degenerate PCR primers targeting biosynthetic oxidative coupling enzymes. For the purposes of biosynthetic enzyme richness estimates, this enzyme class was not included due to these ambiguous reads. As noted earlier, the Anza Borego eDNA sample contains a high percentage of actinobacterial assignments relative to other extracts as determined by 16s rRNA analysis. Fermentation-based studies would suggest that this library sample should contain a large number of natural product biosynthetic enzymes as the culture-dependent analysis of this bacterial phylum has been very productive. (Chapter 1: Figure 1) Many of the primers designed for this experiment were also derived from sequenced Actinomycetes in public databases. We therefore hypothesized that the analysis of the Anza Borego extract could yield a proportionally higher number of biosynthetic sequence variants compared to other eDNA samples. Interestingly, the analysis revealed that, despite having a higher proportion

of Actinomycetes compared to other samples, the Anza Borego library does not contain a higher number of biosynthetic enzyme variants. This result suggests that additional bacterial phyla that are not normally associated with the production of secondary metabolites are contributing to the biosynthetic enzyme diversity found in our other eDNA libraries. Although there is a surprisingly similar number of biosynthetic sequence variants for each of the libraries (Table 10), these groups of sequences are almost entirely distinct as seen in Sorenson similarity comparisons (Table 9). Together, these results indicate that, in contrast to culture-based discovery efforts, the construction and screening of additional soil-based eDNA libraries will continue to yield large numbers of novel biosynthetic sequences and gene clusters that will most likely not be found in other samples.

Table 10: Biosynthetic enzyme diversity estimates

α-diversity (richness) estimates were calculated with 0.97 similarity cutoffs for each biosynthetic read assigned to library samples (Chao1, Equation 2).

	Anza Borego	**Arizona**	**Utah**
Adenylation Domains	4819.1	15426.2	8303.7
Ketosynthase Domains	765.4	1035.7	530.4
Halogenases	2286.2	3470.7	3920.4
Type II PKS	2347.0	2102.4	2779.2
Terpene cyclases	583.9	413.0	274.0
Nitric oxide synthases	7207.4	5225.6	8165.5

We next analyzed the biosynthetic sequences by comparing them to known homologues using BLAST. (Altschul, Gish et al. 1990) BLAST

analysis can serve as a general indicator of whether a sequence variant, and the biosynthetic gene cluster which contains it, are related to characterized metabolites. The results from this analysis were stored in a database and will provide a source of novel screening targets that can be computationally searched and subsequently isolated from eDNA libraries for future studies. (See Materials and Methods) One benefit of using high-throughput sequencing as a screening tool is the ability to uncover rare variants in a mixture of similar sequences. Namely, for some biosynthetic enzymes an individual pool of 4,000-5,000 cosmids (the most basic screening unit in our libraries) likely contains more than one homologue. Sequencing a small number of cloned PCR products in these cases will likely only reveal the most highly represented member of a mixed pool due to differences in dynamic range after PCR amplification. Using high-depth sequencing, rare homologues in complex mixtures of sequences can be detected. Furthermore, the barcodes corresponding to pooled rows/columns of library subarrays (Figure 53) allow for the more efficient identification and recovery of biosynthetic sequences compared to the manual screening and fractionation methods utilized in traditional clone recovery efforts (Chapter 1-3).

To produce a visually intuitive molecular phylogenetic comparison of sequence variants identified in the screen, we generated sequence alignments for each biosynthetic enzyme using CLUSTALW and MUSCLE. (Thompson, Gibson et al. 2002; Edgar 2004) Contaminating reads were first removed

from the data set by clustering all sequences with reference biosynthetic sequences using BLAST and uclust. (Edgar 2004) From here, strict gap opening and extension penalties were used for both pairwise and multiple sequence alignment steps during CLUSTALW analysis (50:50) to generate a multiple sequence alignment. A manually curated database of known biosynthetic sequences which were trimmed based on predicted primer binding sites was also added to these alignments to serve as reference points. For tree curation, an outlying group of random 454 pyrosequencing reads and related biosynthetic enzyme regions that were not predicted to be amplified were added to each data set prior to alignment. These control reads allowed for the easy identification of outlying branches and were used to quickly inspect the quality of the multiple sequence alignments and resulting phylogenetic trees. The visualization of large numbers of aligned sequences becomes challenging when the dataset contains more than a few thousand nodes. (O'Donoghue, Gavin et al. 2010) We therefore reduced the complexity of the dataset by grouping sequences with greater than 0.75 similarity into individual branches prior to phylogenetic tree construction. Some biosynthetic functional groups contained a small percentage of sequences that could not be directly assigned to biosynthetic genes. These reads generally clustered on outlying branches that could be easily identified and removed during tree curation steps. For ease of visualization, after clustering, a representative sequence was chosen for each OTU based on the

frequency of occurrence. Each branch and node of an output alignment was then colored according to the sample barcode which identified where a read was derived from. As seen by the output of this analysis, the eDNA libraries contain sequence variants that cluster with known biosynthetic enzymes corresponding to characterized natural products and they also contain large numbers of "rare" and unexplored biosynthetic sequence variants. We hypothesize that the further analysis of these rare sequence variants should yield biosynthetic pathways from uncultured bacteria that encode structurally novel secondary metabolites.

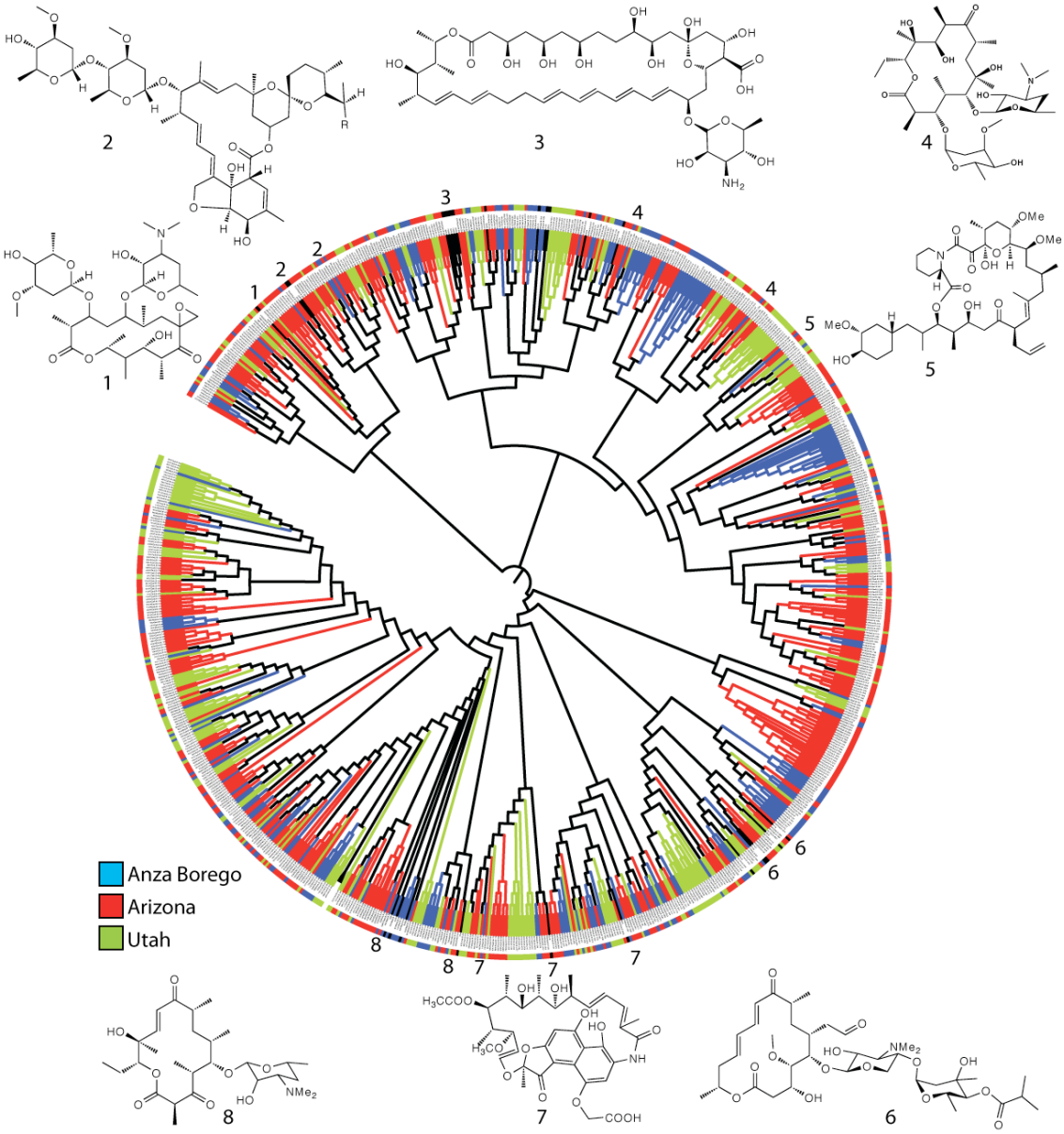


Figure 56: Type I PKS relative sequence similarity

Relative sequence similarity of ketosynthase domains amplified from eDNA libraries. Sequence origins are colored according to the legend. For reference sequences, the following compounds are included: 1) oleandomycin, 2) avermectin, 3) nystatin, 4) erythromycin, 5) FK506b, 6) niddamycin, 7) rifamycin, and 8) pikromycin. In some cases, multiple KS domains from the same reference gene cluster were predicted to be amplified by primers used in this experiment and are also shown.

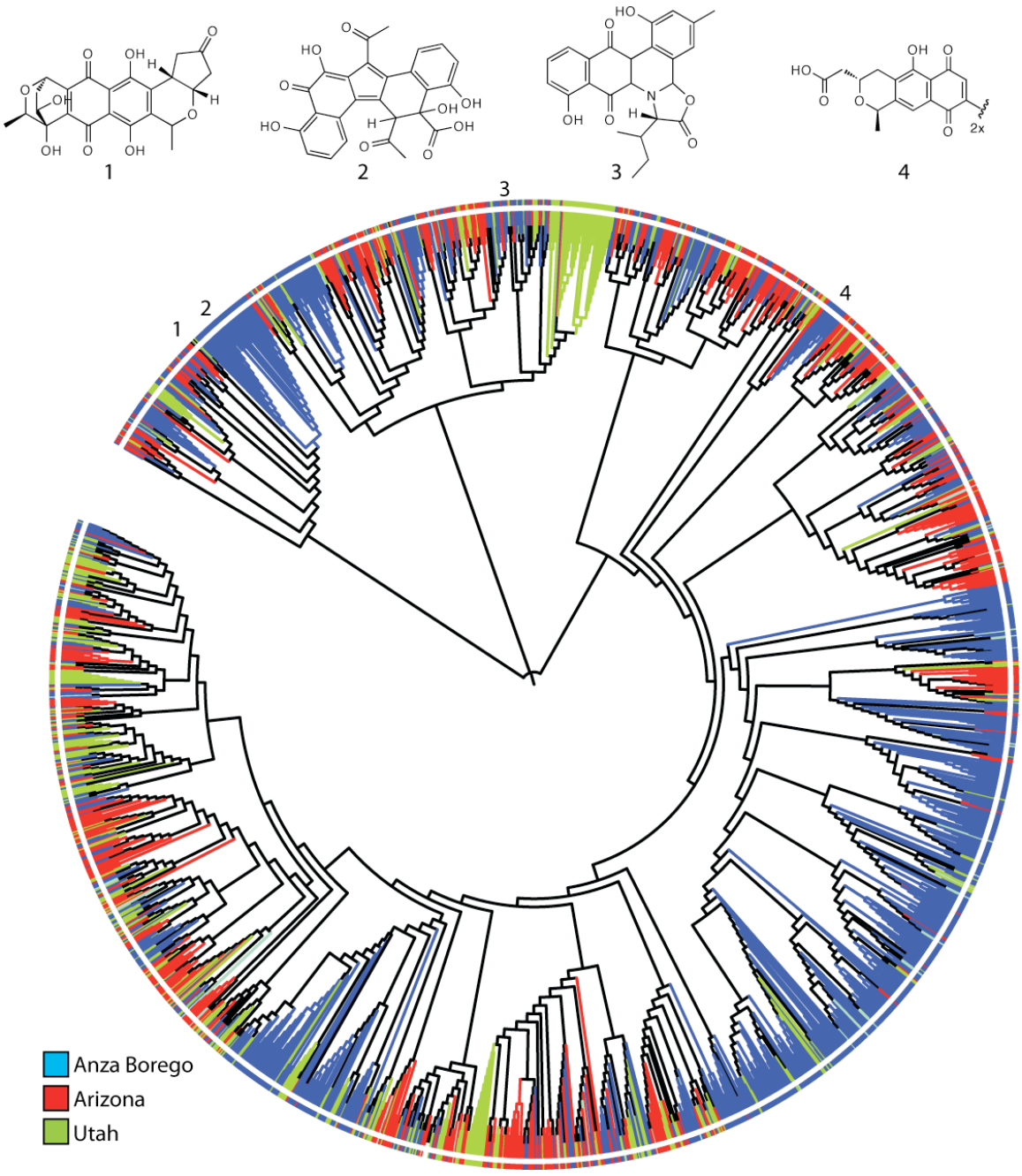


Figure 57: Type II PKS relative sequence similarity

Relative sequence similarity of type II PKS (KSb) domains amplified from eDNA libraries. Sequence origins are colored according to the legend. The following reference sequences and compounds are shown: 1) griseorhodin, 2) erdacin, 3) jadomycin, and 4) actinorhodin

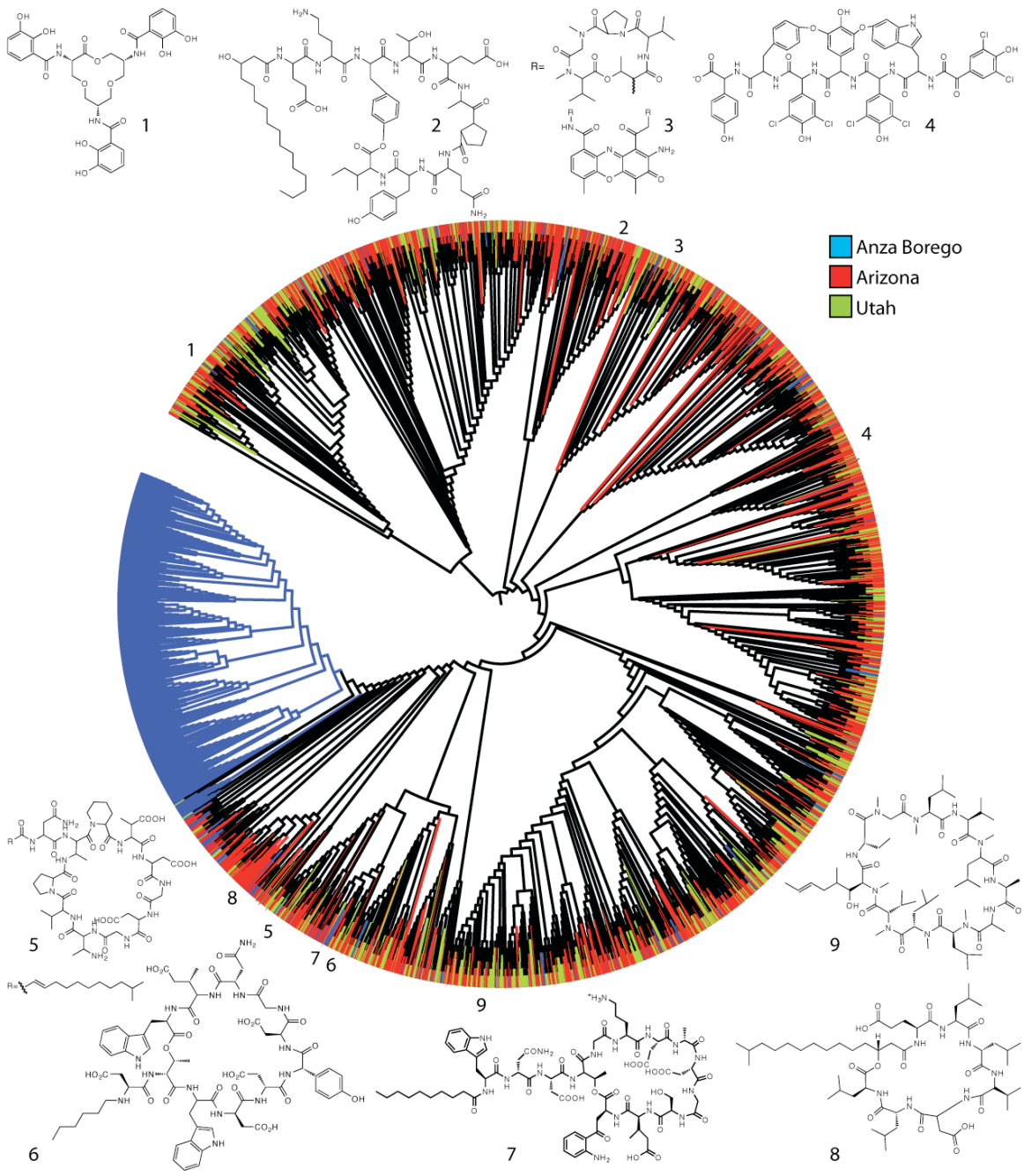


Figure 58: Adenylation domain (NRPS) relative sequence similarity

Relative sequence similarity of adenylation domains amplified from eDNA libraries. Sequence origins are colored according to the legend. For reference sequences, the following compounds are included: 1) enterobactin, 2) fengycin, 3) actinomycin, 4) complestatin, 5) friulimicin, 6) calcium dependent antibiotic, 7) daptomycin, 8) surfactin, and 9) ciclosporin. Adenylation domains from the same pathway generally clustered tightly and are listed as single representative numbers at this scale.

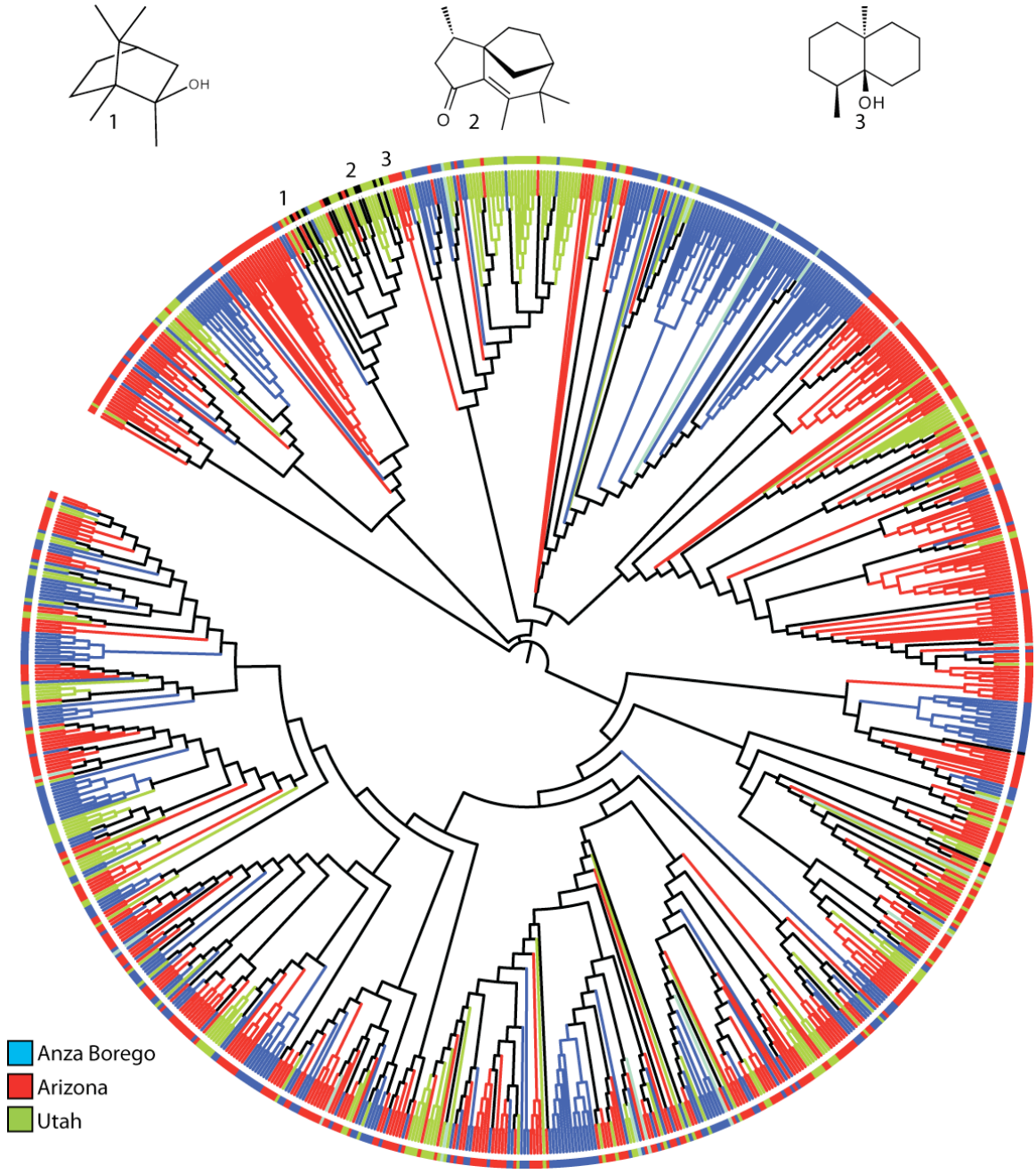


Figure 59: Terpene cyclase relative sequence similarity

Relative sequence similarity of terpene cyclases amplified from eDNA library samples. Reference bacterial terpenes are shown: 1) 2-methylisoborneol, 2) albaflavenone, and 3) geosmin. Sequence origins are colored according to the legend.

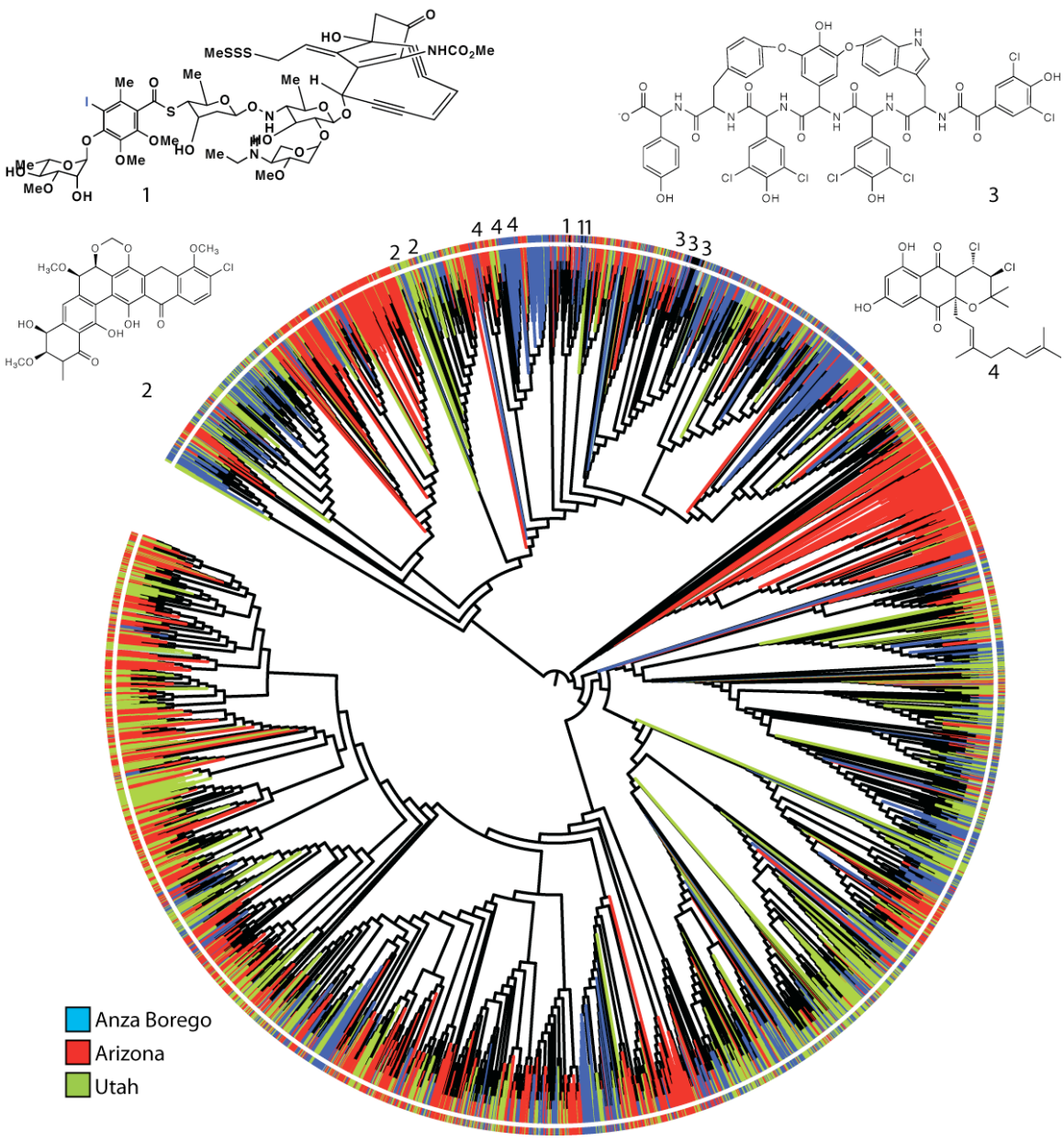


Figure 60: Halogenase relative sequence similarity

Relative sequence similarity of FAD-dependent halogenase homologues amplified from eDNA library samples. Due to the large number of halometabolites, 4 representative members were selected and structurally related halometabolites were listed with the same corresponding compound number: 1) calicheamycin, 2) lysolipin, 3) complestatin, and 4) napyradiomycin. Sequence origins are colored according to the legend.

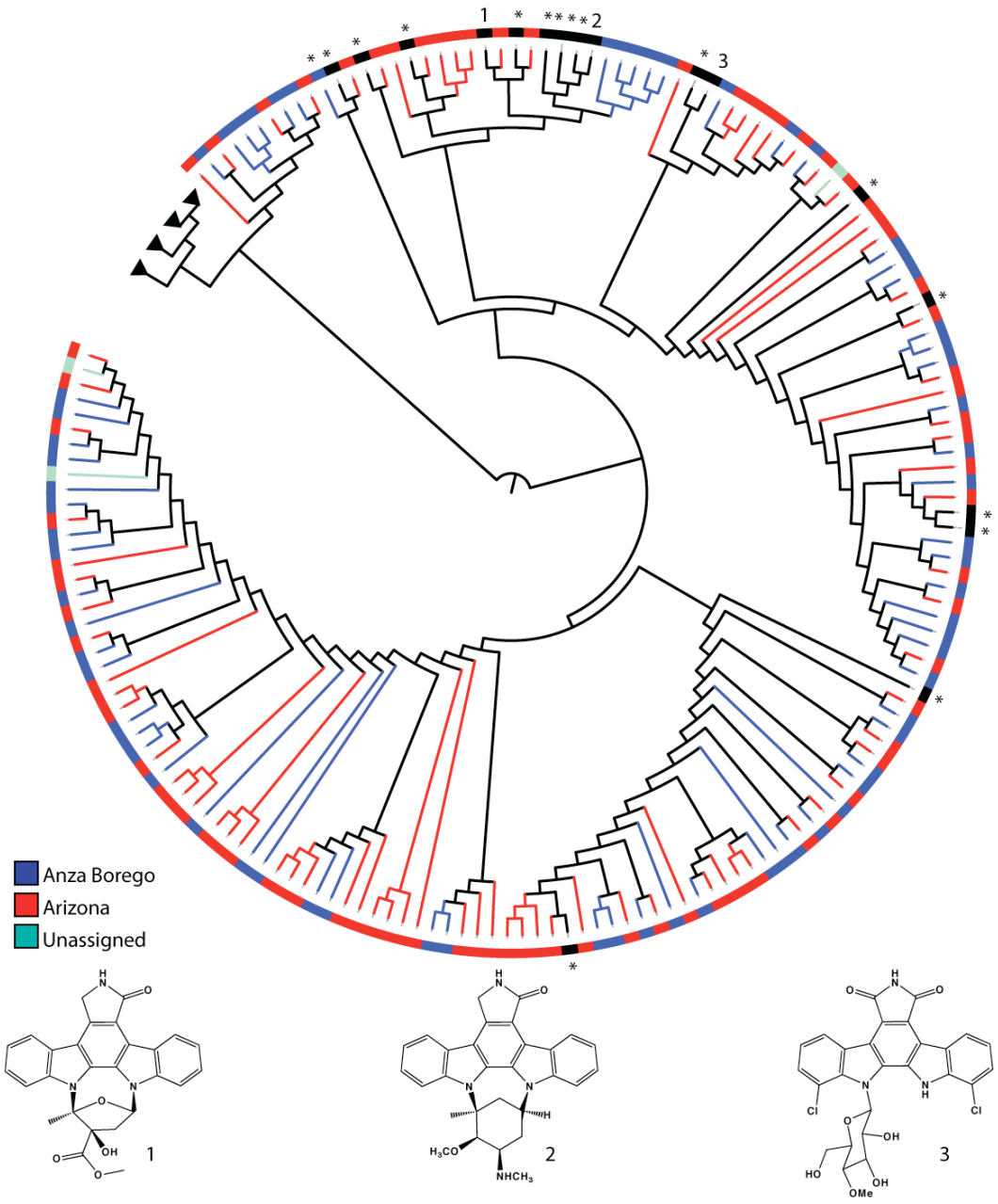


Figure 61: Indolocarbazole oxidase relative sequence similarity

Relative sequence similarity between biosynthetic oxidases involved in indolocarbazole biosynthesis. Black arrows (collapsed branches) highlight a clade which contained outlying sequences that could not be directly associated with biosynthetic genes. Reference compounds are shown: 1) K252a, 2) staurosporine, and 3) rebeccamycin. * indicate sequences variants that were recovered from eDNA libraries by Fang Chang and Paula Calle, which have been confirmed to contain indolocarbazole-like biosynthetic gene clusters. eDNA library samples are colored according to the legend.

The oxidative enzyme variants (indolocarbazole oxidase) contained large numbers of what appear to be primary metabolic sequences which were amplified during the screen. The analysis also did not yield any oxidase sequences from the Utah eDNA library which was empirically confirmed by Fang Chang using manual screening efforts. Using the analysis pipeline described here, it was easy to identify contaminating reads as a single major outlying branch after sequence alignment. These branches were subsequently collapsed prior to tree visualization (Figure 61). A set of control sequences from confirmed biosynthetic gene clusters which were recovered by Fang Chang and Paula Calle, were used during the curation of this particular data set to validate the clustering and analysis steps. As seen in the output tree, these cosmids were all identified in the screen along with novel variants that were detected in our eDNA libraries (Figure 61). It is important to stress that each of the sequence similarity analyses was performed on recombinant eDNA libraries. This general strategy therefore provides an opportunity to rapidly discover and isolate large numbers of biosynthetic gene clusters from eDNA libraries in contrast to manual screening approaches (Chapters 1-3).

The adenylation domain (NRPS) analysis yielded a major outlying branch that is unique for the Anza Borego sample (Figure 58, blue). These sequences were analyzed in more detail and all are predicted to be biosynthetic in origin based on nearest neighbor homology comparisons using BLAST. (Altschul, Gish et al. 1990) We originally hypothesized that these

adenylation domain sequences were preferentially amplified due to the relatively large percentage of Actinomycetes present in the Anza Borego (California) sample. As noted in the following section, however, an analysis of these reads instead suggests that they, along with many other outlying biosynthetic reads, are derived from two relatively unstudied bacterial phyla, Planctomycetes and Gemmatomonadetes.

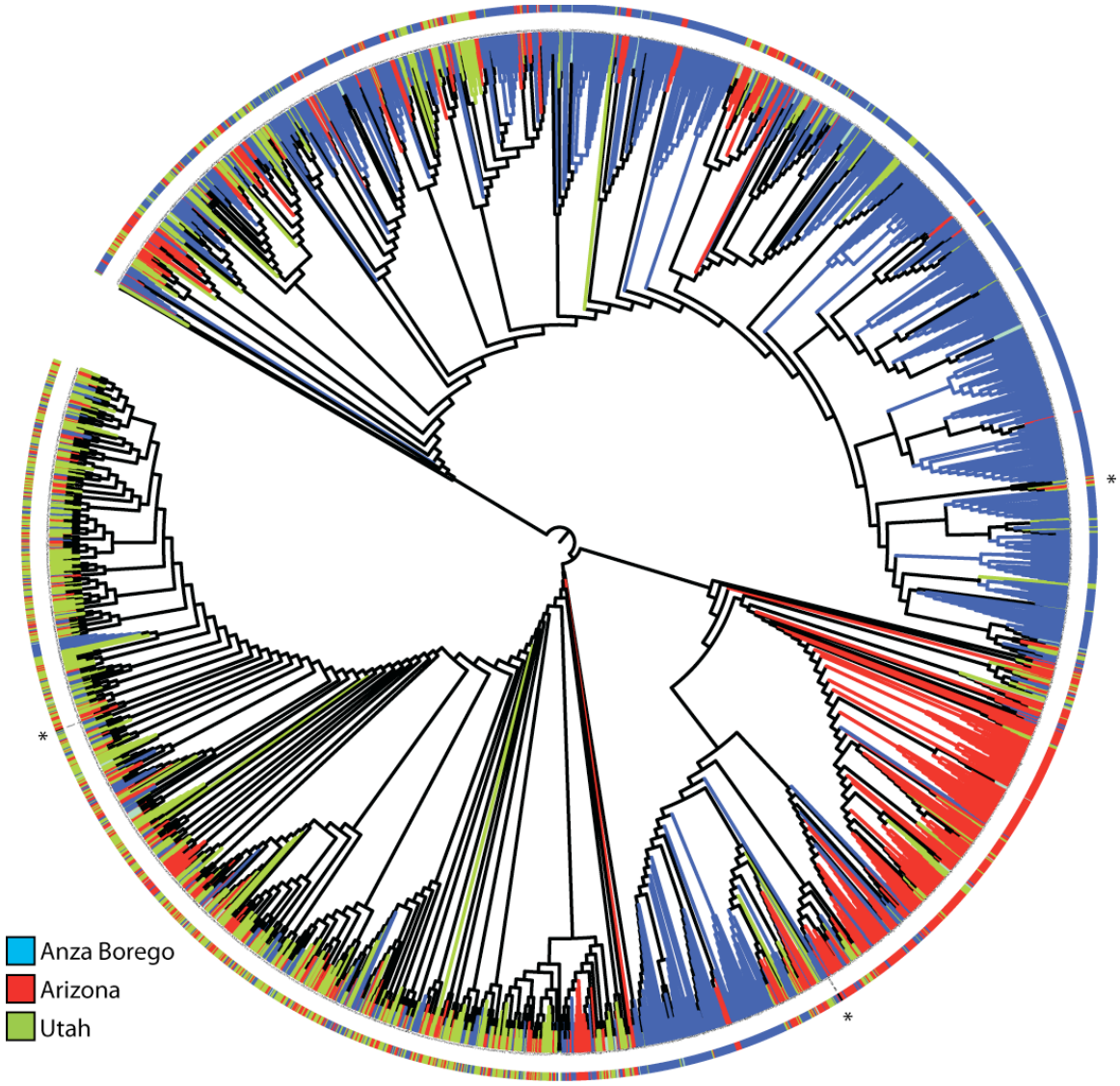


Figure 62: Bacterial nitric oxide synthase relative sequence similarity

Relative sequence similarity for putative nitric oxide synthases are shown. Samples are labeled according to the legend and reference sequences marked (*).

As an interesting complement to the search for novel classes of natural products, many of which are potent antibiotics, it is now clear that the environment contains large and diverse reservoirs of antibiotic resistance genes. (Riesenfeld, Goodman et al. 2004; Dantas, Sommer et al. 2008; Sommer, Dantas et al. 2009; Donato, Moe et al. 2010) It has been recently

demonstrated that the environment contains naturally occurring bacteria that can subsist on high levels of antibiotics as their sole carbon source. (Dantas, Sommer et al. 2008) To probe this phenomenon in more detail, several groups have begun to explore the antibiotic resistance genes present in our environment. These screens have been generally executed using functional assays, however, and were therefore limited by the inherent heterologous expression capabilities of library hosts, typically *E. coli*. (Sommer, Dantas et al. 2009) We elected to therefore use our high-throughput sequencing analysis pipeline to gain deeper insight into a relatively unstudied form of bacterial broad-spectrum antibiotic defense which was recently elucidated. Bacterial nitric oxide synthases (NOS) are heme-based monooxygenases that oxidize L-arginine to nitric oxide (NO) and citrulline. Although mammalian NOSs have been well characterized, the roles that NOSs play in bacteria have only recently begun to be analyzed. (Kers, Wach et al. 2004; Crane 2008; Sudhamsu and Crane 2009; Crane, Sudhamsu et al. 2010) Bacterial antibiotics including lactams, aminoglycosides, and quinolines exert their toxicity, in large part, by promoting the formation of reactive oxygen species (ROS). (Kohanski, Dwyer et al. 2007) Recent studies have specifically shown that NO produced by bacterial NOSs can provide potent broad-spectrum antibiotic resistance for bacterial hosts by both modifying toxins and protecting against antibiotic induced formation of ROS. (Gusarov, Shatalin et al. 2009) As seen in the

relative sequence similarity analysis, each of our eDNA libraries contains thousands of putative bacterial NOSs as defined by curated reference sequences and BLAST homology (Figure 62). The nitric oxide synthase sequence variants also appear to cluster distinctly based on the source eDNA library. There have been no large scale studies of the diversity of bacterial NOSs in uncultured environmental bacteria as their role in antibiotic resistance has only recently been uncovered. This screen indicates that soil environments contain large numbers of distinct bacterial NOSs that could provide broad-spectrum antibiotic resistance to a diverse range of naturally occurring bacteria. The high-throughput screening and functional analysis of additional pools of antibiotic resistance genes found in our eDNA libraries could reveal insights into the mechanisms uncultured bacteria have developed to protect against antibiotic challenge and oxidative stress. Studies of these resistance reservoirs and the mechanisms by which they are transferred to different bacterial hosts will be equally important in determining how to best cope with the increased prevalence of antibiotic-resistant pathogens. (Dantas, Sommer et al. 2008; Sommer, Dantas et al. 2009)

Together, the α-, β-diversity, and relative sequence similarity analyses suggest that the additional construction and screening of soil-based eDNA libraries should continue to yield large numbers of unique biosynthetic sequence variants and, by extension, gene clusters for future natural product

discovery efforts. The large-scale recovery of cosmids containing these unexplored biosynthetic sequence variants should reveal insight into the sources of and types of metabolites encoded in the "rare" biosynthetic biosphere.

4.2.3 Linking Phylogeny and Biosynthetic Function in Uncultured Bacteria

Linking phylogeny and function in metagenomically derived DNA sequences is challenging. (Ginolhac, Jarrin et al. 2004; Riesenfeld, Goodman et al. 2004; Choi and Kim 2007; Stepanauskas and Sieracki 2007) This is primarily because DNA isolated directly from environmental samples is comprised of sheared mixtures of genomic DNA from a diverse range of bacterial species. Sequenced bacterial genomes have revealed that canonical natural product gene clusters are typically found throughout a bacterial chromosome and are not tightly linked with maker genes such as 16s rRNA. (Omura, Ikeda et al. 2001; Bentley, Chater et al. 2002; Ikeda, Ishikawa et al. 2003; Jiang, Tetzlaff et al. 2009; Soror, Rao et al. 2009; Chou, Fanizza et al. 2010) The phylogenetic origin of most biosynthetic gene clusters cannot, therefore, be derived from simply sequencing cis-linked maker genes unless they are randomly captured within a clone of interest. (Brady and Clardy 2005) Furthermore, many bacterial divisions contain only partially sequenced representatives as opposed to complete bacterial genomes. In these cases, and in the absence of a direct homologue, sequence comparisons cannot

be used to determine the phylogenetic origins of biosynthetic gene variants found in metagenomic DNA. (Wu, Hugenholtz et al. 2009) In order to overcome this, we applied a recently developed algorithm that utilizes an interpolated markov model (IMM) of oligonucleotide usage frequencies found in reference sequences to assign taxonomies. (Brady and Salzberg 2009) Several additional indirect phylogenetic classification methods have been described but the majority of these approaches show dramatic reductions in accuracy for sequences under 1,000 bp in length or in cases where highly similar homologues do not exist in sequenced organisms. (Brady and Salzberg 2009) This method, PHYMM, uses IMMs to characterize oligonucleotides of variable length which are typical of a phylogenetic grouping. The analysis is first calibrated against a reference sequence database containing annotated bacterial sequences which have well defined phylogenetic origins. Footprints generated in this primary step can then be used to assign short metagenomic reads to their closest phylogenetic match. While this approach still does not function for candidate environmental phyla that lack any sequence information beyond 16s rRNA, it can accurately assign phylogenies in the absence of a direct homologue, thereby providing a preferred method compared to homology-based strategies. In instances where there are close homologues present in reference genomes, PHYMM combines IMM similarity scores with BLAST (PHYMM-BLAST) increasing the accuracy of taxonomic assignments to ~99.9% at the phylum level.

In order to process the data from the screen, we grouped the sequences into biosynthetic core (adenylation, ketosynthase, terpene cyclase), tailoring (halogenase, oxidase), and resistance enzymes to determine if there were any differences in taxonomic assignments between these types of genes. We also separated each read by its source in order to see if there were different taxonomic assignments associated with each sample. As a general validation of the lack of bias induced by our biosynthetic primer sets, PHYMM analysis yielded classifications from the majority of major bacterial divisions observed in 16s species analyses (Figure 63).

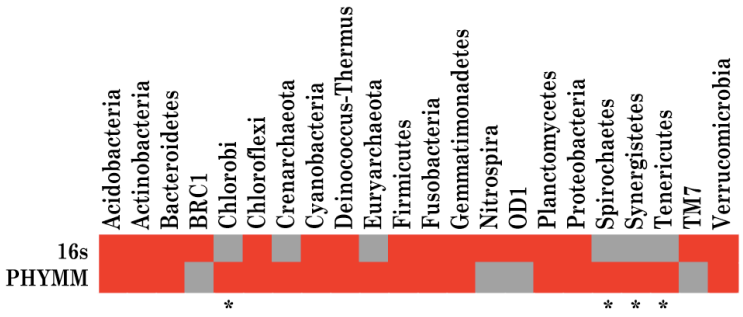


Figure 63: Major bacterial divisions detected with screening primers

Phylogenetic assignments generated using 16s rRNA sequencing and PHYMM-based classifications of biosynthetic sequences indicate that the primers used for the analysis were not biased toward particular phyla. (*) indicate phyla which contained very few assignments but were still within 90% confidence approximations for PHYMM analysis.

It is important to note that PHYMM will not assign phylogenetic classifications from bacterial phyla that do not contain sequenced representative genomes which are used in the original training set. In order to accommodate a wider range of assignments, we therefore also included in the training set draft genomes of phylogenetically defined bacteria which are

actively being sequenced. In all cases and for all classes of functional groups, the analysis assigned a high proportion of biosynthetic sequences to Actinomycetes, Planctomycetes, and Gemmatomonadetes (Figure 64). Actinomycetes have been a cornerstone of natural product discovery and the original degenerate primer designs were based on several Actinomycete sequences so this result was expected. The latter two phyla, Planctomycetes and Gemmatomonadetes, were not predicted based on their known properties or their relative abundance in each of the samples according to species analysis. These two phyla consistently appeared in proportionately high assignment percentages in blind PHYMM analyses of biosynthetic reads, regardless of which biosynthetic functional group was analyzed (Figure 64). A significant percentage of reads were also assigned to Verrucomicrobiae but their relatively high taxonomic abundance in each sample made it difficult to assess whether these bacteria are enriched in biosynthetic enzymes or whether they simply contained low numbers of biosynthetic sequences across multiple species within this phylum. Taken together, we hypothesize that Planctomycetes and Gemmatomonadetes represent promising candidates for future natural product discovery efforts due to the enriched number of biosynthetic sequences that appear to be associated with their uncultured and unsequenced counterparts. When analyzing groups of sequences that do not cluster with known homologues based on sequence similarity (i.e. outlying branches on the adenylation domain sequence similarity analysis

(Figure 58)), these reads map almost exclusively to Verrucomicrobiae, Planctomycetes, and Gemmatomonadetes. While it is possible that the primers used in this screen preferentially amplify sequences from these phyla of bacteria, a close analysis of predicted binding sites in sequenced representatives do not suggest that this should occur (*Gemmatimonas aurantiaca* and *Rhodopirellula baltica SH 1* (GenBank Nos. AP009153, BX119912)). Efforts will be focused on the large-scale recovery of groups of these clones in order to sequence additional cis-linked genes within cosmids which may be useful as phylogenetic markers when processed through PHYMM. We could also, in theory, continue to "walk" overlapping cosmids until marker genes are located within an insert to quantitatively determine the origins of these novel biosynthetic sequences. Additional analyses of different classes of biosynthetic enzymes, a more diverse range of microbiome samples, and continued sequencing could help support these findings as well.

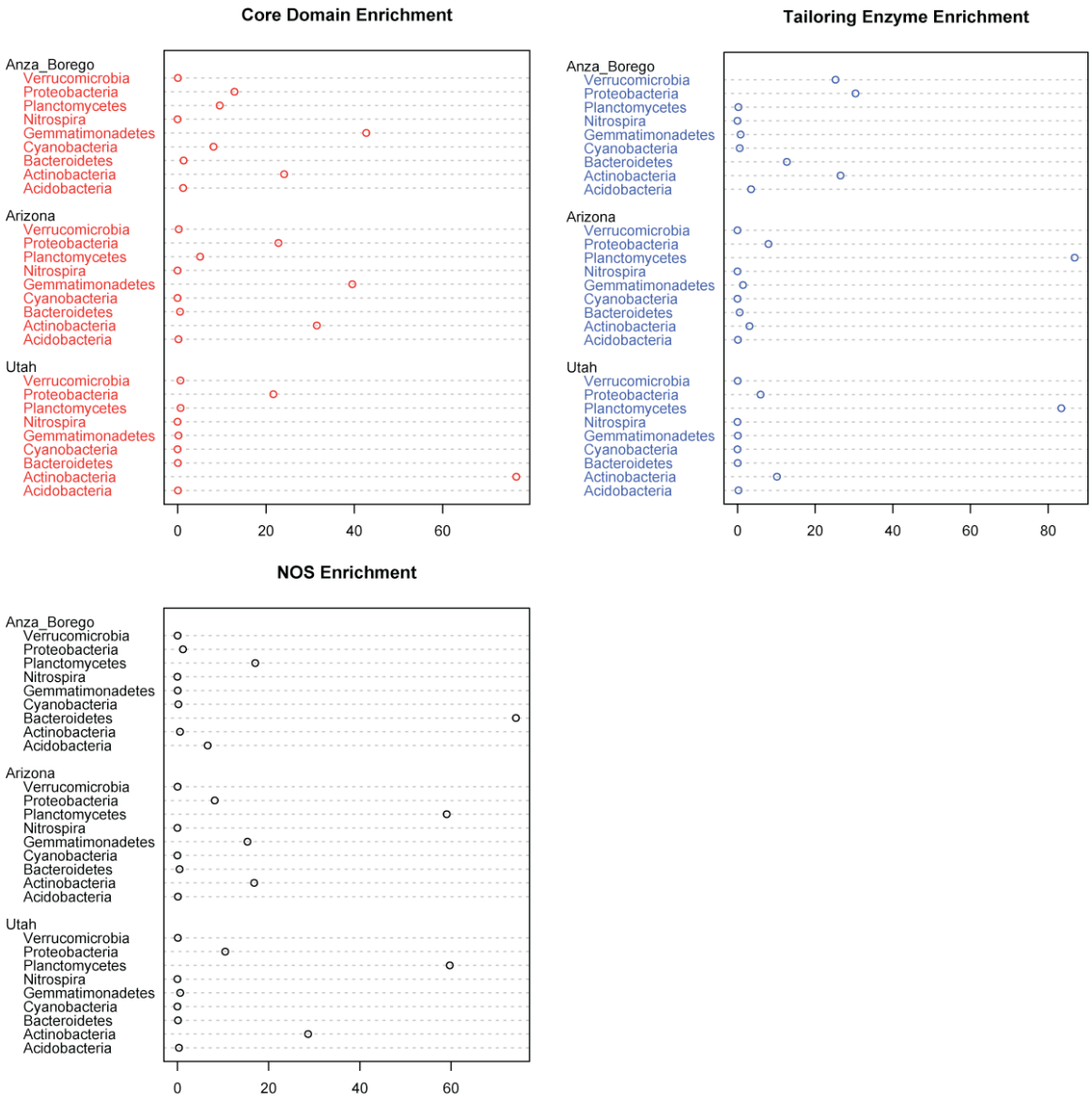


Figure 64: Biosynthetic richness estimates

Richness calculations were normalized for each biosynthetic enzyme class across all samples to remove heterogeneity in the dataset. x-axis values display relative enrichment of biosynthetic sequence variants assigned to particular phyla using PHYMM compared to their abundance based on 16s rRNA species analysis. Sequences were grouped according to core (ketosynthase, adenylation, terpene cyclase) and tailoring (halogenase, oxidase) classifications. In each case, Gemmatomonadetes and Planctomycetes appear to be enriched in these biosynthetic sequences.

4.3 Discussion and Future Directions

Previous studies have shown that soil microbiomes contain a diverse array of novel bacterial species. (Gans, Wolinsky et al. 2005) The large-scale sequencing of 16s rRNA genes has been used for comparative phylogenetics, but this approach does not describe the biosynthetic "capacity" of an environmental sample. By systematically analyzing our crude eDNA samples for both taxonomic and biosynthetic topology using a high-throughput sequencing strategy, we show that there is high biosynthetic functional evenness and low similarity between different soil microbiomes. These results also suggest that additional factors, apart from bacterial species composition, are most likely influencing the biosynthetic diversity present in soil microbiomes. Interestingly, the overall richness of biosynthetic enzyme variants appears to be conserved between our recombinant eDNA libraries despite containing non-redundant sequences and different phylogenetic compositions. These data indicate that additional biosynthetic enzyme screens of different soil microbiomes will continue to yield large numbers of distinct biosynthetic sequences that will most likely not be found in other samples. The results presented here also suggest that bacterial phyla which are not normally associated with the production of secondary metabolites are most likely contributing to the natural product enzyme diversity found in many of our eDNA samples. By utilizing a short-read phylogenetic classification strategy, we show that two recently discovered phyla of bacteria

that are not typically examined for the production of natural products (Planctomycetes and Gemmatomonadetes) may be enriched in biosynthetic gene systems. Model bacterial hosts from these phyla should prove useful for future natural product discovery efforts by providing culture-based discovery platforms and, potentially, heterologous expression hosts for metagenomically derived natural product gene clusters. These putative phylogenetic assignments can be quantitatively confirmed by recovering groups of library clones containing additional genes that can be used for direct phylogenetic classification.

The framework presented here provides a way to gain more comprehensive insight into the chemical diversity encoded by natural bacterial communities and provides direct access to large numbers of natural product gene clusters in eDNA libraries containing biosynthetic sequence variants. It also begins to describe, on a larger scale, the types of naturally occurring bacteria that are associated with biosynthetic enzyme diversity for natural products research. We envision that this approach could be used for more sophisticated analyses of microbiomes in future experiments. For example, the roles that natural products play in specific pathogenic and symbiotic interactions could be examined on a large scale using the system described here. This strategy could also be easily extended to study the types of biosynthetic genes that are actively transcribed under different settings by analyzing the transcriptome of complex bacterial communities or eDNA

libraries under varying cultivation conditions. (Wang, Gerstein et al. 2009; Nagalakshmi, Waern et al. 2010) In general, the large-scale recovery and analysis of groups of biosynthetic sequence variants identified in screens of this nature should prove very useful for studying natural product biosynthesis in uncultured bacterial communities.

4.4 Materials and Methods

4.4.1 eDNA Sample Preparation

Crude eDNA extracts that were not used for library construction were prepared from 50 g (dry weight) of soil prepared according to manufacturers instructions (PowerMax™ Soil DNA Isolation Kit, MoBio). For crude eDNA extracts prepared using standard eDNA library purification methods, one additional round of purification was performed with ~100μg of crude eDNA sample using anion exchange resin (QiaexII™, Qiagen). We found that this step was necessary to remove co-purified PCR inhibitors (humic acids, polysaccharides, etc.) in order to consistently amplify biosynthetic sequences from crude eDNA extracts. Final eluates were kept frozen at -20°C in 10 mM Tris – 1mM EDTA. 0.5μL (~100ng) of the purified samples were used for downstream PCR screening applications. Library samples were combined from pooled master aliquots corresponding to 320,000 cosmid clones at equimolar ratios for bulk screening purposes. 2.5 μL of pooled library subfractions corresponding to ~40,000 cosmid clones were diluted into 25 μL total volume of 10 mM Tris – 1mM EDTA in 384 well plates for high resolution barcoded screening purposes (Adenylation and type I ketosynthase

domains). Library samples in heat sealed (Velocity 11™) 384 well plates were treated with ATP-dependent single stranded nuclease for 2 x 14 hours in 40 μL reactions following manufacturers instructions (PlasmidSafe™, Epicentre) to remove contaminating host (*E. coli*) genomic DNA prior to screening efforts. Supplemental nuclease was added after the first 14 hour incubation. Single stranded nuclease reactions were stopped by heating at 70°C for 1 hour. 1 μL of each nuclease reaction was run in parallel to untreated samples on a 1% agarose gel to qualitatively analyze the *E. coli* genomic DNA removal process. Plates containing treated library samples were then sealed and kept at 4°C prior to screening and archived at -20°C for long-term storage.

4.4.2 Screening Procedure

Bacterial 16s rRNA hypervariable region PCR primers (16sv4/fw/: 5'-AYTGGGYDTAAAGNG -3', 16sv4/rv/: 5'- TACNVGGGTATCTAATCC -3') were used for phylogenetic analysis (Cole, Wang et al. 2009). These 16s rRNA primers target 1,042,316 and 901,964 sequences, respectively, out of 1,379,424 annotated prokaryotic 16s rRNA sequences, and represent the most comprehensive set of validated degenerate primers reported to date. (Cole, Wang et al. 2009) These primers also allowed the rapid taxonomic classification of each sample using alignments to the curated Ribosomal Database. (Cole, Wang et al. 2009) In total, 9 soil samples were multiplexed for crude sample analysis including desert soil from Anza-Borego California, Arizona, Utah, and Tanzania; soil from New Jersey and Pennsylvania; a

soybean field, and a nalidixic acid treated sample. (Polianskaia, Ivanov et al. 2008) For adenylation and ketosynthase domain analysis of library samples, a series of unique error-correcting Hamming barcodes (Hamady, Walker et al. 2008) were incorporated into each primer to identify the pooled row from each library sub-array (~40,000 clones) that a given sequence was identified in. Biosynthetic primers targeted conserved motifs including adenylation domains (AD), ketosynthase domains (KS), type II polyketide synthases (tII), FAD dependent halogenases (hal), terpene synthases (ts), oxidative coupling biosynthetic enzymes (staO), and nitric oxide synthases (bNOS).

Table 11: High-throughput sequencing primers

Enzyme target	Primer sequence
Adenylation/fw/	5'-GCSTACSYSATSTACACSTCSGG-3'
Adenylation/rv/	5'-SASGTCVCCSGTSCGGTA-3'
Ketosynthase/fw/	5'-TSAAGTCSAACATCGGBCA-3'
Ketosynthase/rv/	5'-CGCAGGTTSCSGTACCAGTA-3'
TIIpks/fw/	5'-TSGCSTGCTTCGAYGCSATC-3'
TIIpks/rv/	5'-TGGAANCCGCCGAABCCGCT-3'
Hal/fw/	5'-TTCCCSCGSTACCASATCGGSGAG-3'
Hal/rv/	5'-GSGGGATSWMCCAGWACCASCC-3'
TerpCyc/fw/	5'-ACTGGTAYGTBTGGGTBTTCT-3'
TerpCyc/rv/	5'-SRCVGTGTKCTCGAACTGSTG-3'
StaO/fw/	5'-ACSMTSYTSTTYGGBGCSTGG-3'
StaO/rv/	5'-SGSBGGRTAGTAGGTCTGSC-3'
bNOS/fw/	5'-CCATRTACCAKCCGTTRAASGGVGC-3'
bNOS/rv/	5'-CARYTSRTYGGTYCGSTAYGCBGG-3'

Each PCR primer incorporated a 454 FLX Titanium adapter A (5'-CGTATCGCCTCCCTCGCGCCATCAG-3') and adapter B (5'-CTATGCGCCTTGCCAGCCCGCTCAG-3') to allow direct pyrosequencing of

the PCR products. (Margulies, Egholm et al. 2005) We designed the PCR analysis so the resulting amplicons would be approximtately 500bp in length, would contain few homonucleotide stretches, and so that novel sequences could be amplified easily using standard PCR conditions. These considerations were designed mainly with 454 pyrosequencing and emulsion PCR limitations in mind. (Williams, Peisajovich et al. 2006) To determine optimal amplification conditions, a temperature gradient was used in the following cycling program: initial denaturation (98°C, 2 min), 30 cycles of (98°C, 10 sec; (55-75°C gradient), 30 sec; 72°C, 30 sec) and a final extension step (72°C, 5 min). In cases where gradient cycling conditions did not consistently yield an amplicon, a touchdown PCR program was used with the following parameters: initial denaturation (98°C, 2 min), 7 touchdown cycles (98°C, 10 sec; (72°C dt -2°C/cycle), 30 sec; 72°C, 30 sec), followed by 30 standard cycles of (98°C, 10 sec; 58°C, 30 sec; 72°C, 30 sec), and a final extension step (72°C, 5 min). A panel of buffers (FailSafe™ A-K, Epicentre Biotechnologies; PhusionGC/HF™, Finnzymes) and DNA polymerases were tested under these conditions. Phusion DNA polymerase™ in combination with GC buffer (Finnzymes) was found to be optimal for all screens. Halogenase, type II pks, terpene cyclase, and adenylation domain primers amplified optimally with an annealing temperature of 70°C while the remaining biosynthetic primers yielded amplicons consistently using the

touchdown PCR protocol. 16s rRNA amplification reactions were amplified with a 55°C annealing temperature.

Optimized conditions were used to screen crude eDNA samples and pooled library samples in triplicate with the following reaction conditions. Each 20 µl PCR reaction contained 1µl of purified eDNA template, 0.5 µM each primer, 200 µM of each dNTP, 1X Phusion™ GC Buffer (Invitrogen), 0.2 U Phusion™ DNA polymerase (Invitrogen) and 3% DMSO. For adenylation and ketosynthase domain analysis of arrayed eDNA libraries, each unique barcoded primer was aliquoted into 384 well plates (Genetix™) at 50µM concentration. Each pooled row of a library subarray was aliquoted into 384 well plates and treated to remove *E. coli* genomic DNA as described earlier. A mastermix was aliquoted into 384 well PCR plates (1634 reactions in total), and 0.25 µL of primer and 0.5 µL of library template were transferred into PCR plates using an automated liquid handling robot (Perkin-Elmer Janus™) to a total volume of 10µL. These samples were then PCR cycled according to the optimized protocols.

For experiment validation, 10 randomly selected pools from each barcoded library PCR reaction were checked via gel electrophoresis (1uL sample) to confirm that the amplification reactions yielded expected amplicon sizes. Amplified samples were then individually spectroscopically analyzed (Nanodrop™), pooled, and purified according to manufacturer's recommendations. (Qiagen MinElute™ PCR purification columns). As a

further purification step, the pooled and column purified samples were gel purified in bulk (Qiagen MinElute™ kit) using agarose gel electrophoresis (1% agarose, 0.5x Tris/Borate/EDTA, 150 V, 45 min) with gentian violet staining to avoid ethidium bromide mutagenesis. Each gel purified amplicon was checked fluorometrically (PicoGreen QuantIT™ Invitrogen) and via capillary gel electrophoresis (DNA 7500™ Agilent Technologies) prior to pooling at proportional molar ratios for 454 Titanium pyrosequencing according to manufacturers protocols (Roche, library sequencing protocol). All sequencing was performed at MSKCC (Memorial Sloan Kettering Cancer Center) in the Core Genomics Facility. Library samples were sequenced unidirectionally from primer A while crude samples were sequenced unidirectionally from primer B.

4.4.3 Phylogeny Reconstruction and Biosynthetic Profiling

4.4.3.1 Phylogenetic profiling

The deconvolution and hamming barcode correction scripts were adapted from publicly available computational workflows.(Caporaso, Kuczynski et al. 2010) Quality filtered (Newbler™, default parameters, Roche) and sample deconvoluted (custom Python program) sequencing reads were clustered at an OTU cutoff of 0.97 for species analysis using uclust. (Edgar, unpublished; (Edgar 2004)) All matches that exceeded 0.99 similarity for the library host, *E. coli,* were removed prior to phylogenetic analysis. The OTU clustering step was used to generate a composite

representative phylogenetic profile for all samples. PyNAST (Caporaso, Bittinger et al. 2010) alignment was utilized to compare each representative OTU to the greenegenes and RDP database of curated 16s rRNA sequences(DeSantis, Hugenholtz et al. 2006). From here, the taxonomic profile of the OTU reference set was used as a baseline to determine the relatedness of each sample. Briefly, an OTU table outlining which samples contained reads that mapped to the composite reference OTU set, including the number of times a sequence mapped to these OTUs, was used for all β and α-diversity (Chao1) calculations. Specifically, this OTU table was used to generate a distance matrix representing the similarity of sequences\species found in each sample set. The distance matrix was then analyzed using principal coordinate analysis to visualize the relatedness of each sample. Rarefaction analysis and α-diversity Chao1 estimates were also generated from these OTU tables (Equation 2). Many related phylogenetic analysis steps have been incorporated into the latest distribution of QIIME which was used to compare the output of our experiments. (Caporaso, Kuczynski et al. 2010) Phylum level heatmaps were generated from the OTU table output using R. (The Comprehensive R Archive Network CRAN)

4.4.3.2 Biosynthetic profiling

Each biosynthetic enzyme was processed to remove short and low quality reads and to deconvolve the samples based on their barcodes prior to analysis. These scripts were similar to those used for 16s rRNA analysis. This

data was then used to generate a composite 0.97 similarity reference sequence set to compare the relatedness of biosynthetic enzymes found in each of the samples. In a similar strategy to species analysis, a table representing the number of times a reference OTU was found in a given sample was used to create a distance matrix from which α and β diversity metrics were calculated. As noted in the text, α-diversity metrics were difficult to obtain for crude eDNA samples for ketosynthase homologues, as the biosynthetic enzymes approached theoretical maximum Shannon evenness (~1) and rarefaction indicated that the sequences were significantly undersampled. This was partially due to the diversity of the homologues detected in crude samples and also due to the slightly decreased emulsion PCR efficiency and resulting number of sequencing reads for this functional group. Future experiments could adjust for these factors easily by dedicating a larger proportion of a sequencing run to analyze type I ketosynthase domains. Beta indices could be calculated using this distance matrix, however, and principal coordinate analysis was used to generate the clustering plots. All principal coordinate plots were visualized using KiNG (Kinemage, Next Generation).

Some biosynthetic functional groups required several additional processing steps as many of the reads could not be directly mapped to biosynthetic homologues upon close analysis. This was most likely caused by contaminant sequences which were coamplified with biosynthetic

homologues. In these cases, experimental sequencing reads were first clustered with BLAST using a 0.5 similarity and 0.01 e-value cutoff against a manually curated reference sequence database of known biosynthetic homologues. This metric was empirically determined to yield a majority proportion of biosynthetic reads in addition to a small percentage of ambiguous reads. In some cases (Oxidase, terpene cyclase), uclust was used with a 0.5 similarity cutoff because very few characterized biosynthetic homologues existed in public databases preventing large-scale homology based clustering using BLAST. (Edgar, unpublished, (Edgar 2004)) Each of the curated reference sequence sets was trimmed according to predicted primer binding sites in order to increase the accuracy of the alignments as the pyrosequencing reads were approximately 350bp in average length. We also added random 454 pyrosequencing reads and biosynthetic reads which did not map to regions predicted to be amplified using the primer sets in this experiment. These reads were added as negative controls to see how well the alignment, clustering, and tree construction steps functioned. From here, a stringent multiple sequence alignment (CLUSTALW: gap/ext penalty 50, pwgap/ext penalty 50) was generated to produce a phylogenetic tree representing the relatedness of experimental reads to the reference sequences. The phylogenetic tree was initially curated by removing clearly divergent branches that did not contain biosynthetic sequences as determined by manual BLAST analysis against the full non redundant

protein database. (Altschul, Gish et al. 1990) Using these strict alignment metrics, it was easy to visualize outlying negative control sequences during tree curation. All major sequence groups that did not contain reference sequences were analyzed in detail using BLAST. For the majority of these "rare" branches, the closest homologues were biosynthetic in origin and they could be easily classified. In some cases, the reads did not directly map to characterized biosynthetic genes and were instead most closely related to hypothetical proteins. Each of these hypothetical branches was analyzed by viewing the genomic context that closest matches were found in. In many cases, the closest homologues of these "rare" and unidentified branches were clearly biosynthetic based on their cis-linked genomic sequences although they were mis-annotated as hypothetical proteins in public databases. These branches were therefore included in output molecular phylogenetic analyses. The sensitivity of the cutoff parameters can be easily adjusted if we wish to analyze more highly divergent gene sequences but we were primarily interested in canonically biosynthetic genes for this analysis. All phylogenetic trees were visualized using iTOL (Letunic and Bork 2007).

Shannon evenness calculations were performed using Equation 1. α-diversity (chao1) calculations were performed using Equation 2 where S_{obs} is the number of observed species, n_1 is the number of species observed once, and n_2 is the number of species observed multiple times. This index is particularly useful for data sets that contain low abundance classes, as is

likely the case in soil microbiomes. (Chao 1984) Pairwise Sorenson similarity calculations were generated using Equation 3 with OTU tables (0.97 similarity) from initial clustering steps. These OTU tables were rareified to remove sample heterogeneity in order to normalize the calculation to similar numbers of representative sequences. This procedure avoids distribution shifts that are caused by differences in sampling depth versus true sample characteristics.

Equation 1: Shannon evenness (E)

H' = Shannon diversity index: S = number of species, p_i = proportion of a given species to the total number of individuals in a community

$$E = \frac{H'}{lnS} : H' = \sum_{i=1}^{S} (p_i ln p_i)$$

Equation 2: The general form of the Chao1 estimator

S_{obs} = the number of observed species, n_1 = the number of species observed once, and n_2 = the number of species observed multiple times

$$S_{Chao} = S_{obs} + \frac{n_1^2}{2n_2}$$

Equation 3: Sorenson's similarity index

C = the number of shared units, A + B = the total number of observed units

$$QS = \frac{2C}{A+B}$$

4.4.3.3 Phylogenetic classification of short reads

For phylogentic analysis of raw reads, we utilized a recently developed method which relies on interpolated markov models of oligonucleotide usage

patterns as compared to a reference sequence database. (Brady and Salzberg 2009) Raw reads were processed according to published protocols and the output was filtered with to a 90% confidence threshold. (Brady and Salzberg 2009) We elected to choose this method for taxonomic assignment as many bacterial divisions contain only partially sequenced representatives precluding simple homology comparisons. Biosynthetic richness was calculated by simply taking the proportion of PHYMM-assigned taxonomies from the raw biosynthetic reads and normalizing this value against the phylogenetic profile generated from 16s rRNA analyses. Despite potential bias toward Actinomycetes (because the majority of sequences used in the primer design were derived from sequenced Actinomycetes) the results indicate that Planctomycetes and Gemmatomonadetes are potentially enriched with biosynthetic sequences. Future clone recovery efforts will be required to support these predictions by sequencing groups of cosmids associated with these sequence variants.

4.4.3.4 Homology based screening database

Each of the reads was initially renamed according to the sample, and library subfraction it was amplified from based on barcode assignments detected in the sequences. The renaming of the reads was performed using sample demultiplexing scripts in QIIME. (Caporaso, Kuczynski et al. 2010) For detailed biosynthetic profiling of each sequencing read, we analyzed the data using BLASTx with a custom database comprised of non-redundant

protein(nr), environmental (env_nr) and additional global ocean sampling (gos-nr) sequences. The composite sequence data set was split using custom PERL programs into framesizes that could be loaded into memory for the BLAST calculations. This process was outsourced (Amazon EC2™) and the results were stored in XML format for direct entry into our database. This database also provides direct links to any related molecular structures that are associated with close sequence homologues as determined by BLAST. (Altschul, Gish et al. 1990) Sebastian Jayaraj developed and implemented this framework using Web2Py.

CHAPTER 5

5. Concluding Remarks

It is now apparent that the uncultivated bacterial majority harbors a large reservoir of biosynthetic genes that cannot be examined using traditional fermentation-based strategies. As shown here, the heterologous expression of natural product gene clusters isolated directly from the environment can provide access to the chemistry encoded by uncultured bacteria. These results specifically demonstrate that sequenced-based screens are a powerful complementary discovery strategy when paired with eDNA libraries of sufficient size and a robust gene cluster reassembly method, TAR. This framework overcomes a major barrier which heretofore prevented the functional characterization of a wide range of natural product gene clusters isolated from metagenomic sources. The general utility of TAR was demonstrated by functionally reassembling a large eDNA-derived natural product gene cluster to uncover Fluostatins F-H. This platform should form the basis for future natural product discovery efforts from uncultured bacteria. Using a high-throughput sequencing screen, we have also shown that uncultured bacteria from soil microbiomes encode an incredibly diverse range of biosynthetic genes that are largely unexplored. The biosynthetic

enzymes found in different environments are distinct and do not appear to be correlated with classically defined secondary metabolite encoding phyla (i.e. Actinomycetes). This suggests that the continued screening of different microbiomes using this approach should reveal novel phylogenetic sources of natural product chemical diversity.

Heterologous expression barriers represent the most significant remaining challenge preventing the large-scale discovery of novel natural products from uncultured bacteria. The heterologous production of a secondary metabolite requires the functional and coordinated transcription and translation of multiple foreign genes in a selected host using a particular culturing condition. Promoter activation, ribosome binding site recognition, product toxicity, and differences in primary metabolite diversity are just some of the factors that can impede heterologous expression. Also, the metabolite must be stable using a selected isolation strategy and generated in sufficient quantities to be detected using modern analytic approaches. Natural product isolation efforts are further confronted by the fact that neither the structure of the molecule nor the required precursors are known for many cryptic gene clusters. Even fully expressed gene clusters, as determined by transcriptional, proteomic, and functional readouts, can been recalcitrant to natural product characterization efforts due to these challenges. (Nougayrède, Homburg et al. 2006; Chung, Lim et al. 2008) A large number of cryptic biosynthetic gene clusters in cultured model

organisms have not been associated with the production of a small molecule despite extensive fermentation, expression, and engineering efforts. Similarly, only a fraction of the biosynthetic pathways isolated from eDNA libraries are predicted to be functionally expressed in their native form in a heterologous host as seen in earlier chapters. The use of alternative promoters for transcriptional activation efforts has not provided a universal solution to access metabolites encoded by cryptic gene clusters. Also, modifying large natural product gene clusters containing multiple ORFs presents a significant technical challenge.

At a fundamental level, we understand very little about how natural product gene clusters are regulated in their native context, making rational efforts to heterologously activate biosynthetic pathways difficult. The continued examination of natural product diversity in uncultured bacteria will therefore rely on additional developments which build upon the platform presented in this thesis. These include 1) additional phylogenetically diverse heterologous expression hosts, 2) high-throughput assays to enrich biosynthetic sequences from complex metagenomic libraries, 3) methods to detect small quantities of heterologously expressed molecules, 4) multiplex methods to modify large biosynthetic gene clusters, 5) scalable methods to test heterologous transcription and translation conditions, and 6) a more fundamental understanding of how biosynthetic gene clusters are regulated. I

have included, in the following section, a brief discussion of some promising approaches that can be applied to address many of these challenges.

5.1 Future directions

5.1.1 Quantitatively Linking Phylogeny and Function

The bacterial diversity present in nature makes the selection of a heterologous host challenging because they are inherently limited in their ability to functionally process foreign DNA. The selection of a phylogenetically related host will most likely increase the chances of functionally expressing an eDNA-derived natural product gene cluster. (Craig, Chang et al. 2010) The species origins of metagenomically-derived biosynthetic gene clusters cannot be easily determined, however, due to the lack of cis-linked marker genes and the fact that many bacterial divisions have few sequenced representatives for comparison. (Wu, Hugenholtz et al. 2009) As demonstrated in Chapter 4, high-throughput sequencing can be applied to draw broad correlations between uncultured bacterial phyla and biosynthetic enzyme diversity. More quantitative determinations of bacterial phylogeny and biosynthetic function will guide the development of specific heterologous hosts for future natural product discovery efforts. While we outlined a potential method of quantitatively linking phylogeny and function in eDNA libraries by recovering overlapping cosmids until canonical marker genes are found, this approach is not easily scalable. The physical compartmentalization of single environmental bacteria allows functional and

phylogenetic genes to be simultaneously assayed in their natural bacterial context. Recent studies have used FACS, microdroplet encapsulation, magnetic enrichment (BEAMing), and microfluidics to examine phylogeny and function in uncultured environmental bacteria. (Diehl, Li et al. 2006; Li, Diehl et al. 2006; Ottesen, Hong et al. 2006; Stepanauskas and Sieracki 2007) A recent study of the termite midgut microbiome using this single-cell approach demonstrated the power of linking genes of interest to a single bacterium within a complex ecosystem. (Ottesen, Hong et al. 2006) In this body of work, the authors revealed the quantitative phylogenetic origins of genes found in the termite midgut microbiome by digitally analyzing single environmental bacteria in a microfluidic device (Figure 65). By applying these analysis techniques to biosynthetic genes, it should be possible to assay biosynthetic function and phylogeny in a cis-linked and quantitative manner for large numbers of environmental bacteria. Large-scale studies along these lines could quantitatively describe the phylogenetic origins of many biosynthetic systems which will help drive the rational development of future heterologous expression and culture-based natural product discovery hosts.

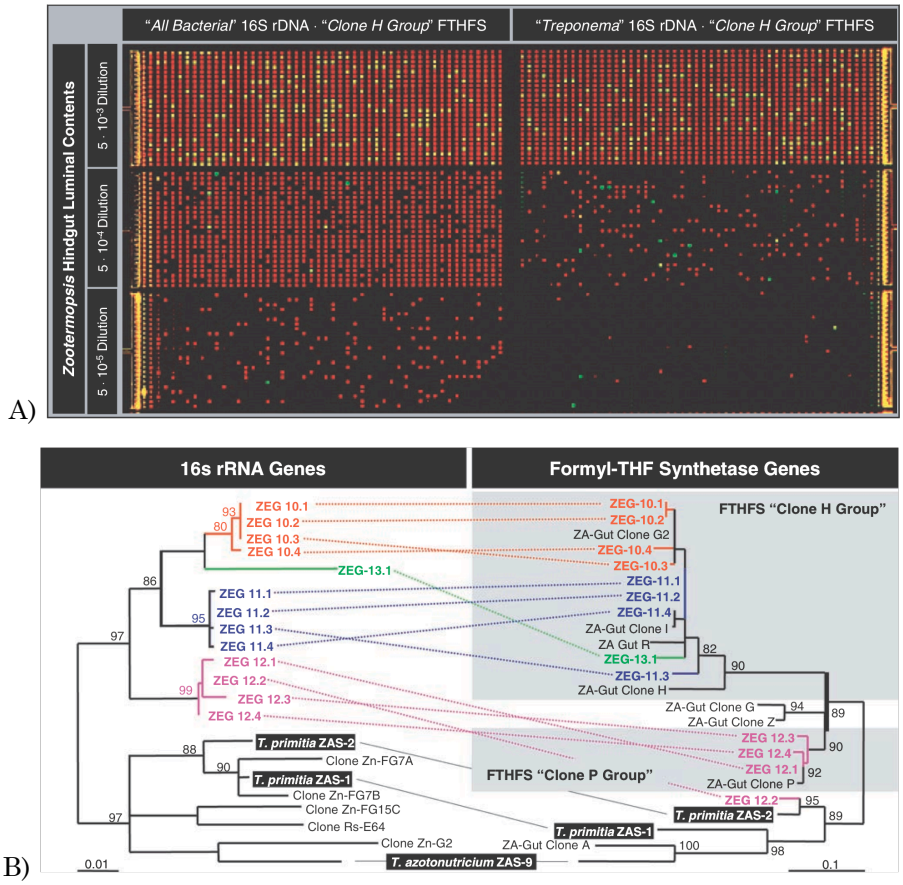


Figure 65: Linking phylogeny and function using microfluidic digital PCR

The authors of this study simultaneously assayed phylogeny (red, 16s rRNA) and function (green, formyl-THF synthetase) in single uncultured bacteria from the termite midgut using a microfluidic device (A). As seen by indirect phylogeny inferences, broad associations remained generally accurate but quantitative analysis showed that many of these genes are associated with other taxonomic goups and appear to have been horizontally transferred (B). *Figure from* (Ottesen, Hong et al. 2006)

5.1.2 High-Throughput Sequencing and Gene Synthesis

Reductions in sequencing cost and increases in throughput have already made a significant impact on our ability to characterize biosynthetic gene clusters isolated from eDNA libraries (Chapter 2-4). In general, the sequence-based screening method described in this thesis provides a way to enrich and isolate natural product encoding gene clusters from mixtures of

eDNA-derived cosmids. The cost and throughput of sequencing has already developed to a point where it is feasible to randomly sequence large collections of eDNA-derived cosmids to obtain complete biosynthetic gene cluster information without the need for enrichment strategies. Our largest eDNA libraries contain in excess of 300,000 genome equivalents of total DNA. 16s rRNA analyses indicate that between 5,000-20,000 unique bacterial species were likely captured in these libraries. 20,000 bacterial genome equivalents of DNA would require 80 GB of sequencing at 1x coverage for an average bacterial genome 4 MB in size. Modern sequencing platforms can now produce more than 200 GB of data in a single experimental run. (Shendure, Porreca et al. 2005) Although the *de novo* assembly of large contigs from short-read data generated by these technologies was previously challenging, novel algorithms combined with tag-directed assembly methods now make it possible. (Zerbino and Birney 2008; Ng, Turner et al. 2009; Turner, Lee et al. 2009; Hiatt, Patwardhan et al. 2010) 5,000,000 cosmid clones, between ½ to 1/3 of our largest eDNA libraries, can theoretically be fully sequenced in a single experiment using current high-throughput sequencing platforms (200 GB/run). The most recent sequencing methods are predicted to further increase throughput by up to two orders of magnitude (20 TB/run). (Polonsky, Stolovitzky et al. 2007) During a proof of principle experiment, we generated 20,000 single read sequences (~15 kb of data) from a randomly selected group of nonenriched transposon mutagenized cosmids

and this relatively small data set revealed numerous biosynthetic clones (data not shown). A comprehensive screen using modern high throughput sequencing technology will clearly reveal many more biosynthetic gene clusters by extension.

The dramatic advances in next-generation sequencing technology have been accompanied by drops in the cost of DNA synthesis. As mentioned in Chapter 3, gene cluster total synthesis is a powerful way to overcome many of the heterologous expression and cloning limitations imposed by working with natural DNA sequences. In addition, this approach allows convenient cloning and expression features to be incorporated into gene clusters to make downstream combinatorial biosynthetic strategies more easily accessible. (Reisinger, Patel et al. 2006) This approach is still not feasible, however, for engineering large collections of biosynthetic gene clusters due to cost considerations ($0.39/bp as of 5/2010). The primary expense during gene synthesis, however, is actually due to the need to sequence multiple synthons for quality control measures. This is mainly due to error rates (errors/bp) caused by oligonucleotide coupling chemistries. (Czar, Anderson et al. 2009) The price of gene synthesis will therefore continue to drop as sequencing technology advances. In addition, recently developed microchip-based DNA synthesis approaches can obtain >20 kb of sequence for as little as $1 USD. (Tian, Gong et al. 2004) The continued refinement of these techniques will undoubtedly lead to the more widespread adoption of synthetically

engineered biosynthetic gene clusters as they can be codon optimized for heterologous expression and are free of many of the cloning, screening, recovery, engineering, and assembly challenges that face natural biosynthetic pathways. By coupling high-throughput sequencing efforts, described earlier, with gene synthesis, it is possible to imagine directly sequencing and constructing natural product gene clusters without the need to isolate biosynthetic clones, assemble large natural DNA constructs, or engineer large multi-ORF gene clusters for heterologous expression efforts (Figure 66).

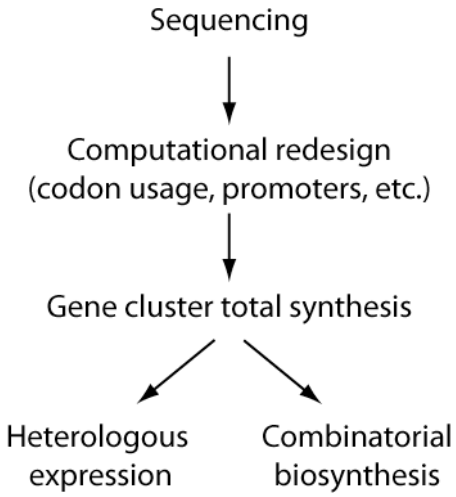


Figure 66: Theoretical sequencing-based workflow for natural product discovery

5.1.3 Synthetic Biology and Natural Product Discovery

The heterologous production of secondary metabolites in tractable fermentation hosts provides an opportunity to generate sustainable and scalable sources of these small molecules which are often difficult to obtain. Many natural products are either synthetically intractable or are not cost effective to synthesize or isolate in large quantities, making their functional

characterization difficult. Even natural products that have well defined functions often suffer from limited availability. (Kirkpatrick, Raja et al. 2003; Trost and Dong 2008) Recently, groups have begun to explore synthetic biology concepts with the goal of engineering fermentation strains that can efficiently produce large quantities of secondary metabolites. (Martin, Pitera et al. 2003; Ro, Paradise et al. 2006; Chang, Eachus et al. 2007; Tsuruta, Paddon et al. 2009) In contrast to pathway-specific strategies, synthetically engineered base strains can provide a general platform for the high throughput discovery of natural products from both cultured and uncultured sources. Because these engineering efforts typically involve extensive genetic manipulations, the majority of these studies have been executed in model laboratory hosts such as *E. coli*, *S. cerevisiae*, and *S. avermitilis*. (Martin, Pitera et al. 2003; Ro, Paradise et al. 2006; Chang, Eachus et al. 2007; Tsuruta, Paddon et al. 2009; Komatsu, Uchiyama et al. 2010) The most notable example of this approach focused on generating a sustainable source of the anti-malarial sesquiterpene artemisinin. (Martin, Pitera et al. 2003; Ro, Paradise et al. 2006; Chang, Eachus et al. 2007; Tsuruta, Paddon et al. 2009) In this study, Keasling and coworkers systematically tuned the expression of genes in the native *S. cerevisiae* and *E. coli* ergosterol pathway to produce high levels of FPP (farnesyl pyrophosphate), a universal precursor for sesquiterpene biosynthesis. By downregulating *erg9* expression, the enzyme that normally converts FPP to squalene in the ergosterol pathway,

the authors were able to generate high levels of cytosolic FPP. From here, amorphadiene synthase (ADS) was engineered into the pathway to convert FPP into amorphadiene. With three additional enzymatic steps, they were able to generate large quantities of artemisinic acid, a direct precursor to artemisinin. The continued refinement of a synthetically altered *S. cerevisiae* strain from this study has led to the production of unprecendented levels of artemisinic acid. Through extensive genetic engineering, the authors were able to achieve a direct fermentation-based strategy to convert simple sugars into a complex natural product (Figure 67). This strategy not only overcame a critical supply need, as plant-sourced artemisinin suffers from high price variability, but also generated a base strain that should be capable of producing large quantities of a diverse range of isoprenoids.

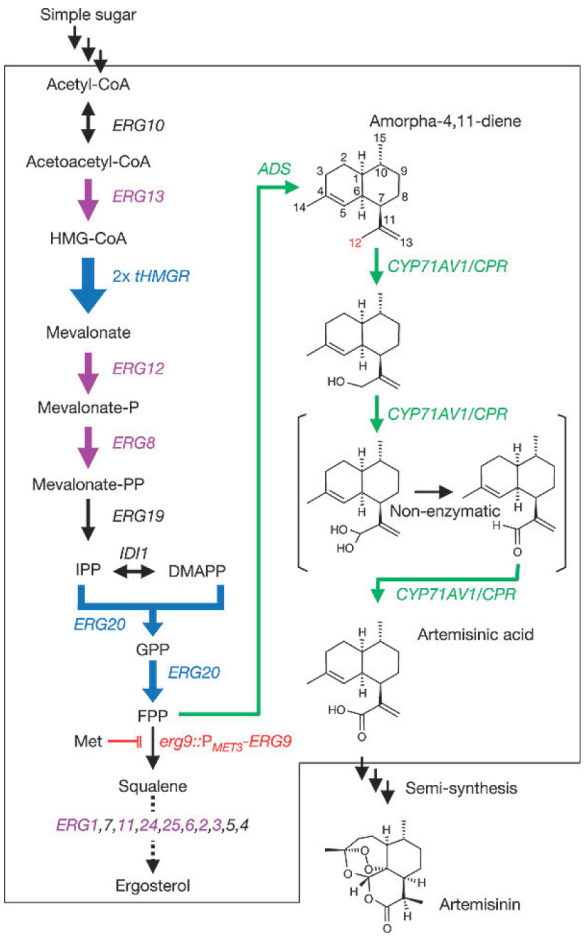


Figure 67: Heterologous production of artemisinin

This schematic shows an example of synthetically tuning central metabolism in *S. cerevisiae* (the ergosterol pathway) toward the increased production of precursors for the antimalarial terpene artemisinin. *Adapted from* (Ro, Paradise et al. 2006; Tsuruta, Paddon et al. 2009)

A similar engineering strategy has recently been applied to a genetically tractable Streptomycete, *S. avermitilis*, in order to produce a broad-range heterologous expression strain capable of generating high quantities of precursor molecules for NRPS/PKS and isoprenoid-like biosynthetic systems. (Komatsu, Uchiyama et al. 2010) In this study, the authors engineered 1.4 Mb of non-essential deletions in *S. avermitilis* to increase precursor flux toward the heterologous production of secondary

metabolites. The authors specifically targeted biosynthetic gene clusters found within the sequenced *S. avermitilis* genome in order to remove competing sources of precursor utilization for heterologously introduced biosynthetic gene clusters. After these deletions were made, this base strain was capable of heterologously producing aminoglycoside, polyketide, and beta-lactam natural products at higher titers than their natural host strains. The rational tuning of central metabolism in additional genetically tractable heterologous hosts could provide a powerful broad-range platform for the heterologous expression of a diverse range of biosynthetic gene clusters from eDNA. This approach could also be coupled with codon optimization and synthetic strategies, outlined previously, to provide a scalable heterologous expression platform for both culture-based and metagenomic natural product discovery efforts. As described previously, the continued development of novel heterologous expression hosts, either synthetically engineered or naturally-derived, will clearly benefit future natural product discovery efforts.

APPENDIX

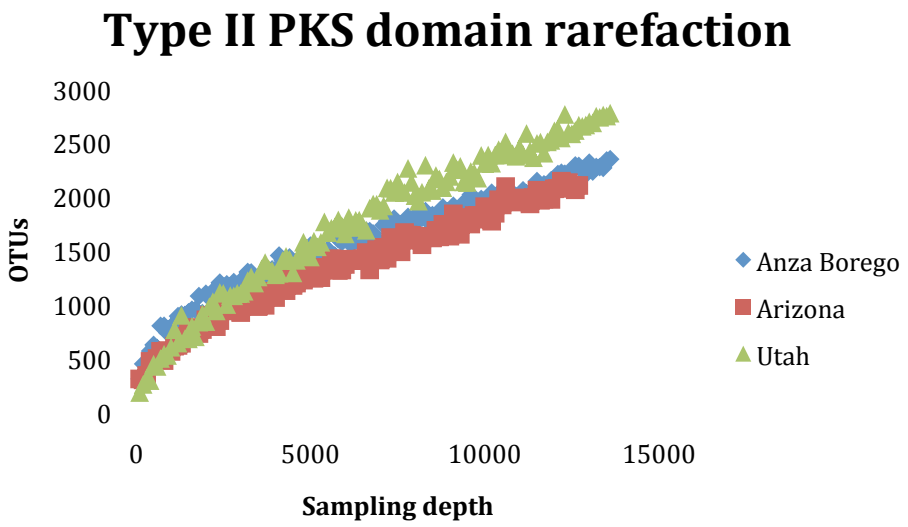


Figure 68: Type II PKS domain rarefaction analysis

Rarefaction analysis of type II polyketide synthase (KSβ) domain amplicons derived from library samples. The number of unique sequences (defined as OTUs in this case) are shown at 97% similarity along the y-axis. Each sample is colored according to the source of the eDNA library.

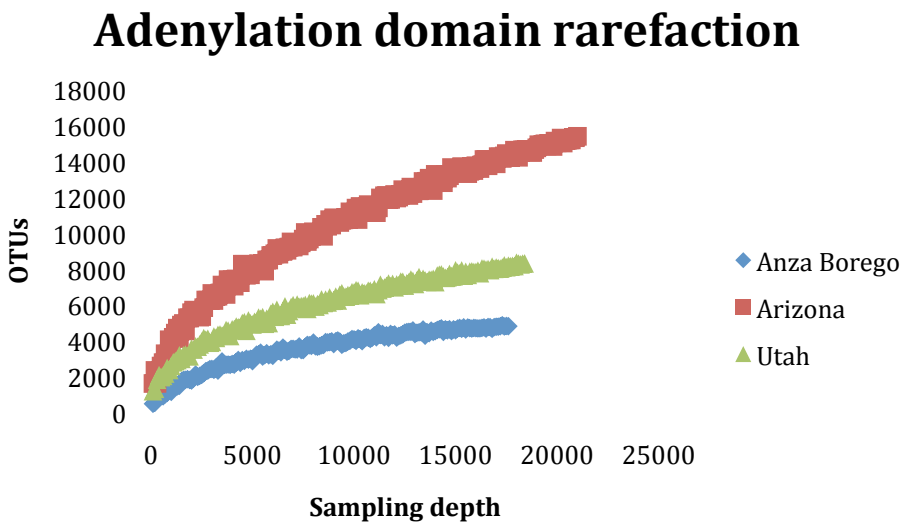


Figure 69: Adenylation domain (NRPS) rarefaction analysis

Rarefaction analysis of adenylation domain sequences derived from library samples. The number of unique sequences (defined as OTUs in this case) are shown at 97% similarity on the y-axis. Each sample is colored according to the source of the eDNA library template used in amplification reactions.

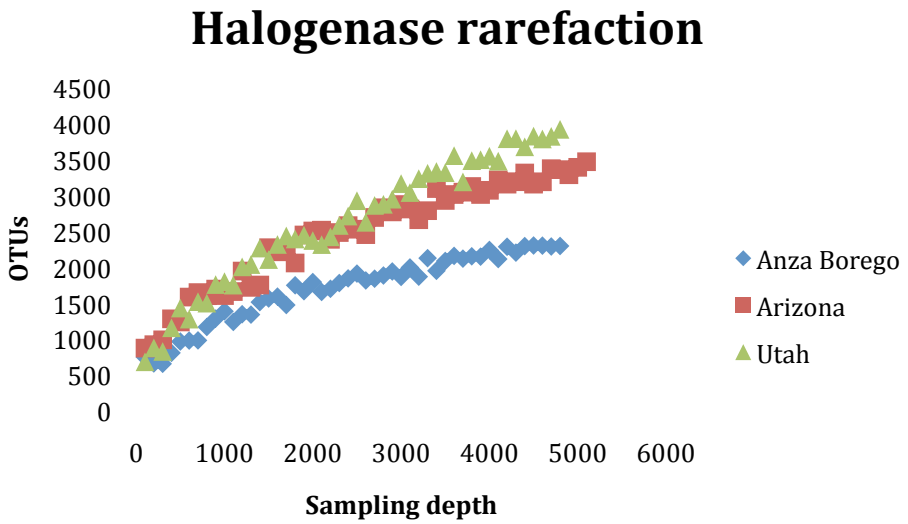


Figure 70: Flavin-depedent halogenase rarefaction analysis

Rarefaction analysis of flavin-dependent halogenase domain amplicons derived from eDNA library samples. The number of unique sequences (defined as OTUs in this case) are shown at 97% similarity along the y-axis. Each sample is colored according to the source of the eDNA library used in amplification reactions.

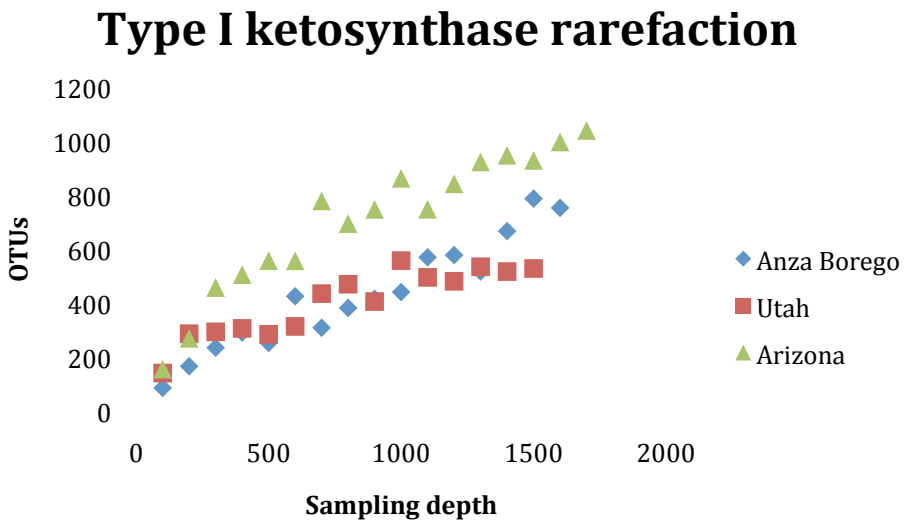


Figure 71: Type I ketosynthase domain rarefaction analysis

Rarefaction analysis of type I ketosynthase domain sequences amplified from eDNA library samples. The number of unique sequences (defined as OTUs in this case) are shown at 97% similarity on the y-axis. Each sample is colored according to the source of the eDNA library template.

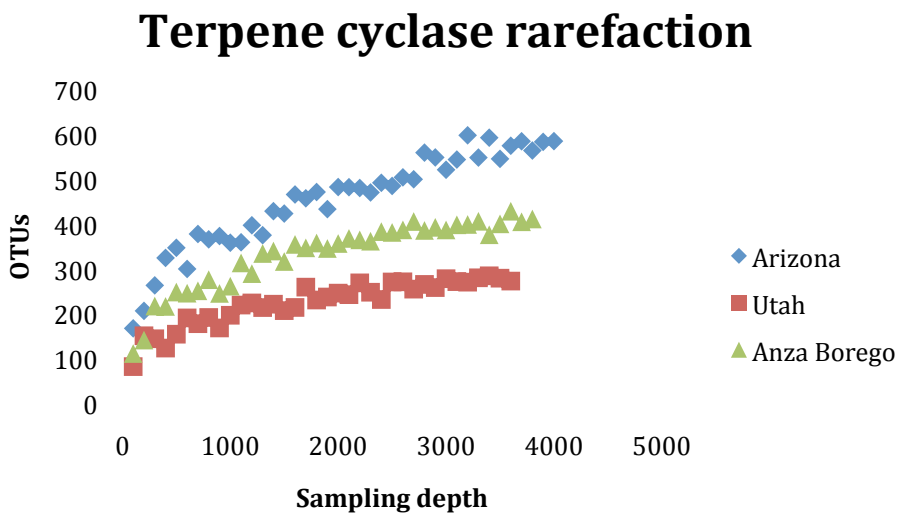


Figure 72: Terpene cyclase rarefaction analysis

Rarefaction analysis of terpene cyclase sequences derived from eDNA library samples. The number of unique sequences (defined as OTUs in this case) are shown at 97% similarity on the y-axis. Each sample is colored according to the source of the eDNA library template used in amplification reactions.

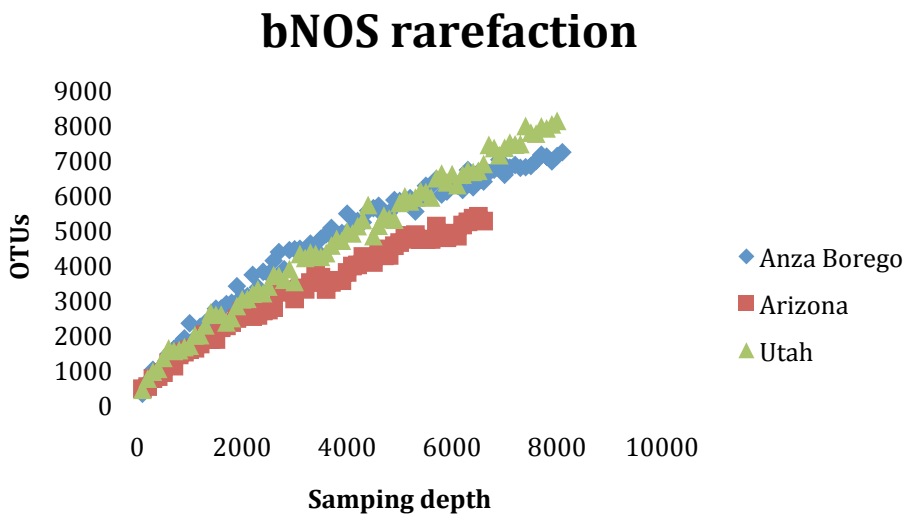


Figure 73: Bacterial nitric oxide synthase rarefaction analysis

Rarefaction analysis of bacterial nitric oxide synthase domain sequences amplified from eDNA library samples. The number of unique sequences (defined as OTUs in this case) are shown at 97% similarity on the y-axis. Each sample is colored according to the source of the eDNA library template used in amplification reactions.

GLOSSARY

BLAST (Basic Local Alignment Search Tool): An algorithm that finds regions of local similarity between sequences. This program, along with its immediate derivatives, compares nucleotide or protein sequences to databases and calculates the statistical significance of matches. BLAST can be used to infer evolutionary and functional relationships between sequences as well as help identify members of gene families.

CLUSTALW: A general purpose multiple sequence alignment program for DNA or proteins. This algorithm calculates a best match for a set of sequences and lines them up so that the identities, similarities, and differences can be seen.

PHYMM: A computational analysis tool developed by Brady, A. et al. which allows for the phylogenetic classification of short sequencing reads (~100bp) derived from high-throughput sequencing experiments. This technique provides a robust method of assigning phylum-level classifications by utilizing well-defined reference genomes as a training set for interpolated markov analysis. I adapted this tool by creating a training set of sequences including all partially and fully sequenced bacterial genomes and sub-sequences with known phylogenetic origins.

QIIME (Quantitiative Insights Into Microbial Ecology): A suite of computational tools that allows for the filtering, classification, and analysis of phylogenetic sequencing data (primarily 16s rRNA-based data) from metagenomic samples. Many of the analytical tools created for this thesis were recently incorporated into this suite. It is recommended to refer to this series of tools as a comprehensive reference when analyzing novel sets of data, especially for phylogenetic classification efforts.

MUSCLE: A rapid multiple sequence alignment technique developed by Robert Edgar that is of comparable accuracy to alternative multiple sequence alignment algorithms such as CLUSTALW. In some cases, CLUSTALW produced higher quality alignments but was much more computationally intensive. Both of these tools were used for multiple sequence alignments and the results were compared in each case.

iTOL (Interactive Tree of Life): A java-based advanced programming interface (API) developed by the European Molecular Biology Laboratory for analyzing phylogenetic information derived from multiple sequence alignments and analyses. The API is easily accessible from EMBL and was used for the visualization of large data sets derived from the high throughput sequencing experiment and multiple sequence alignments generated with CLUSTALW using this data.

PERL (Practical Extraction and Report Language): A well-established programming language that was used for the parsing and analysis of sequence data derived from the high-throughput sequencing experiment.

PYTHON: A high-level programming language used for many scripting applications during analysis.

RDB (Ribosomal Database): This initiative was created by the Center for Microbial Ecology at Michigan State University and serves as a comprehensive collection of 16s ribosomal RNA sequences. This online resource also now contains many of the analytical tools (Chao1, rarefaction, clustering, taxonomic assignments) used in the phylogenetic analysis of our eDNA samples. This suite should also be referenced as a comprehensive resource of phylogenetic classification tools for future analysis.

NCBI (National Center for Biotechnology Information): A comprehensive resource of nucleotide and protein sequences used for homology comparisons. BLAST analysis was performed using this data set and an API provided by NCBI (netblast).

REFERENCES

Abulencia, C. B., D. L. Wyborski, et al. (2006). "Environmental whole-genome amplification to access microbial populations in contaminated sediments." Appl Environ Microbiol **72**(5): 3291-3301.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

August, P. R., T. H. Grossman, et al. (2000). "Sequence analysis and functional characterization of the violacein biosynthetic pathway from Chromobacterium violaceum." J Mol Microbiol Biotechnol **2**(4): 513-519.

Ayuso-Sacido, A. and O. Genilloud (2005). "New PCR primers for the screening of NRPS and PKS-I systems in actinomycetes: detection and distribution of these biosynthetic gene sequences in major taxonomic groups." Microb Ecol **49**(1): 10-24.

Baker, B. J., G. W. Tyson, et al. (2006). "Lineages of acidophilic archaea revealed by community genomic analysis." Science **314**(5807): 1933-1935.

Banik, J. J. and S. F. Brady (2008). "Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary." Proc Natl Acad Sci U S A **105**(45): 17273-17277.

Bauer, J. D., R. W. King, et al. (2010). "Utahmycins a and B, azaquinones produced by an environmental DNA clone." J Nat Prod **73**(5): 976-979.

Bentley, S. D., K. F. Chater, et al. (2002). "Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2)." Nature **417**(6885): 141-147.

Bertrand, H., F. Poly, et al. (2005). "High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction." J Microbiol Methods **62**(1): 1-11.

Bertrand, K. P., K. Postle, et al. (1983). "Overlapping divergent promoters control expression of Tn10 tetracycline resistance." Gene **23**(2): 149-156.

Bierman, M., R. Logan, et al. (1992). "Plasmid cloning vectors for the conjugal transfer of DNA from Escherichia coli to Streptomyces spp." Gene **116**(1): 43-49.

Blasiak, L. C., F. H. Vaillancourt, et al. (2006). "Crystal structure of the non-haem iron halogenase SyrB2 in syringomycin biosynthesis." Nature **440**(7082): 368-371.

Blodgett, J. A. V., P. M. Thomas, et al. (2007). "Unusual transformations in the biosynthesis of the antibiotic phosphinothricin tripeptide." Nat Chem Biol **3**(8): 480-485.

Borodovsky, M., R. Mills, et al. (2003). "Prokaryotic gene prediction using GeneMark and GeneMark.hmm." Curr Protoc Bioinformatics **Chapter 4**: Unit4 5.

Brady, A. and S. L. Salzberg (2009). "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models." Nat Methods **6**(9): 673-676.

Brady, S. F. (2007). "Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules." Nat Protoc **2**(5): 1297-1305.

Brady, S. F., J. D. Bauer, et al. (2007). "Natural products from isnA-containing biosynthetic gene clusters recovered from the genomes of cultured and uncultured bacteria." J Am Chem Soc **129**(40): 12102-12103.

Brady, S. F., C. J. Chao, et al. (2002). "New natural product families from an environmental DNA (eDNA) gene cluster." J Am Chem Soc **124**(34): 9968-9969.

Brady, S. F., C. J. Chao, et al. (2004). "Long-chain N-acyltyrosine synthases from environmental DNA." Appl Environ Microbiol **70**(11): 6865-6870.

Brady, S. F., C. J. Chao, et al. (2001). "Cloning and heterologous expression of a natural product biosynthetic gene cluster from eDNA." Org Lett **3**(13): 1981-1984.

Brady, S. F. and J. Clardy (2000). "Long-chain N-acyl amino acid antibiotics isolated from heterologously expressed environmental DNA." J Am Chem Soc **122**(51): 12903–12904.

Brady, S. F. and J. Clardy (2004). "Palmitoylputrescine, an antibiotic isolated from the heterologous expression of DNA extracted from bromeliad tank water." J Nat Prod **67**(8): 1283-1286.

Brady, S. F. and J. Clardy (2005). "Cloning and heterologous expression of isocyanide biosynthetic genes from environmental DNA." Angew Chem Int Ed Engl **44**(43): 7063-7065.

Brady, S. F. and J. Clardy (2005). "N-acyl derivatives of arginine and tryptophan isolated from environmental DNA expressed in Escherichia coli." Org Lett **7**(17): 3613-3616.

Brady, S. F. and J. Clardy (2005). "Systematic investigation of the Escherichia coli metabolome for the biosynthetic origin of an isocyanide carbon atom." Angew Chem Int Ed Engl **44**(43): 7045-7048.

Brady, S. F., L. Simmons, et al. (2009). "Metagenomic approaches to natural products from free-living and symbiotic organisms." Nat Prod Rep **26**(11): 1488-1503.

Breitbart, M. and F. Rohwer (2005). "Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing." Biotechniques **39**(5): 729-736.

Cane, D. E., C. T. Walsh, et al. (1998). "Harnessing the biosynthetic code: combinations, permutations, and mutations." Science **282**(5386): 63-68.

Caporaso, J. G., K. Bittinger, et al. (2010). "PyNAST: a flexible tool for aligning sequences to a template alignment." Bioinformatics **26**(2): 266-267.

Caporaso, J. G., J. Kuczynski, et al. (2010). "QIIME allows analysis of high-throughput community sequencing data." Nat Methods **7**(5): 335-336.

Chang, M. C. Y., R. A. Eachus, et al. (2007). "Engineering Escherichia coli for production of functionalized terpenoids using plant P450s." Nat Chem Biol **3**(5): 274-277.

Chao, A. (1984). "Nonparametric Estimation of the Number of Classes in a Population." Scand J Statist **11**: 265-270.

Choi, I.-G. and S.-H. Kim (2007). "Global extent of horizontal gene transfer." Proc Natl Acad Sci U S A **104**(11): 4489-4494.

Chou, W. K., I. Fanizza, et al. (2010). "Genome mining in Streptomyces avermitilis: cloning and characterization of SAV_76, the synthase for a new sesquiterpene, avermitilol." J Am Chem Soc **132**(26): 8850-8851.

Chung, E. J., H. K. Lim, et al. (2008). "Forest soil metagenome gene cluster involved in antifungal activity expression in Escherichia coli." Appl Environ Microbiol **74**(3): 723-730.

Clardy, J. and S. F. Brady (2007). "Cyclic AMP directly activates NasP, an N-acyl amino acid antibiotic biosynthetic enzyme cloned from an uncultured beta-proteobacterium." J Bacteriol **189**(17): 6487-6489.

Cole, J. R., B. Chai, et al. (2007). "The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data." Nucleic Acids Res **35**(Database issue): D169-172.

Cole, J. R., Q. Wang, et al. (2009). "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis." Nucleic Acids Res **37**(Database issue): D141-145.

Comeron, J. M. and M. Aguade (1998). "An evaluation of measures of synonymous codon usage bias." J Mol Evol **47**(3): 268-274.

Corey, E. J. and W. D. Li (1999). "Total synthesis and biological activity of lactacystin, omuralide and analogs." Chem Pharm Bull (Tokyo) **47**(1): 1-10.

Court, D. L., J. A. Sawitzke, et al. (2002). "Genetic engineering using homologous recombination." Annu Rev Genet **36**: 361-388.

Courtois, S., C. M. Cappellano, et al. (2003). "Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products." Appl Environ Microbiol **69**(1): 49-55.

Courtois, S., A. Frostegård, et al. (2001). "Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation." Environ Microbiol **3**(7): 431-439.

Craig, J. W., F.-Y. Chang, et al. (2009). "Natural products from environmental DNA hosted in Ralstonia metallidurans." ACS Chem Biol **4**(1): 23-28.

Craig, J. W., F.-Y. Chang, et al. (2010). "Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria." Appl Environ Microbiol **76**(5): 1633-1641.

Crane, B. R. (2008). "The enzymology of nitric oxide in bacterial pathogenesis and resistance." Biochem Soc Trans **36**(Pt 6): 1149-1154.

Crane, B. R., J. Sudhamsu, et al. (2010). "Bacterial nitric oxide synthases." Annu Rev Biochem **79**: 445-470.

Curtis, T. P., W. T. Sloan, et al. (2002). "Estimating prokaryotic diversity and its limits." Proc Natl Acad Sci U S A **99**(16): 10494-10499.

Czar, M. J., J. C. Anderson, et al. (2009). "Gene synthesis demystified." Trends Biotechnol **27**(2): 63-72.

Daniel, R. (2005). "The metagenomics of soil." Nat Rev Microbiol **3**(6): 470-478.

Dantas, G., M. O. Sommer, et al. (2008). "Bacteria subsisting on antibiotics." Science **320**(5872): 100-103.

Davies, A. M., R. Tata, et al. (2005). "Crystal structure of a putative phosphinothricin acetyltransferase (PA4866) from Pseudomonas aeruginosa PAC1." Proteins **61**(3): 677-679.

Davies, A. M., R. Tata, et al. (2007). "l-Methionine sulfoximine, but not phosphinothricin, is a substrate for an acetyltransferase (gene PA4866) from Pseudomonas aeruginosa: structural and functional studies." Biochemistry **46**(7): 1829-1839.

Demidov, V. V., N. O. Bukanov, et al. (2000). "Duplex DNA capture." Curr Issues Mol Biol **2**(1): 31-35.

DeSantis, J., T Z, P. Hugenholtz, et al. (2006). "NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes." Nucleic Acids Res **34**(Web Server issue): W394-399.

DeSantis, T. Z., P. Hugenholtz, et al. (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." Appl Environ Microbiol **72**(7): 5069-5072.

Diehl, F., M. Li, et al. (2006). "BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions." Nat Methods **3**(7): 551-559.

Dinsdale, E. A., R. A. Edwards, et al. (2008). "Functional metagenomic profiling of nine biomes." Nature **452**(7187): 629-632.

Donato, J. J., L. A. Moe, et al. (2010). "Metagenomic analysis of apple orchard soil reveals antibiotic resistance genes encoding predicted bifunctional proteins." Appl Environ Microbiol **76**(13): 4396-4401.

Dunbar, J., L. O. Ticknor, et al. (2000). "Assessment of microbial diversity in four southwestern United States soils by 16S rRNA gene terminal restriction fragment analysis." Appl Environ Microbiol **66**(7): 2943-2950.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Eid, J., A. Fehr, et al. (2009). "Real-time DNA sequencing from single polymerase molecules." Science **323**(5910): 133-138.

Feng, Z., J. H. Kim, et al. (2010). "Fluostatins produced by the heterologous expression of a TAR reassembled environmental DNA derived type II PKS gene cluster." J Am Chem Soc **132**(34): 11902-11903.

Fischbach, M. A., J. R. Lai, et al. (2007). "Directed evolution can rapidly improve the activity of chimeric assembly-line enzymes." Proc Natl Acad Sci U S A **104**(29): 11951-11956.

Fischbach, M. A. and C. T. Walsh (2009). "Antibiotics for emerging pathogens." Science **325**(5944): 1089-1093.

Ford, T. C. and D. Rickwood (1982). "Formation of isotonic Nycodenz gradients for cell separations." Anal Biochem **124**(2): 293-298.

Frigaard, N. U., A. Martinez, et al. (2006). "Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea." Nature **439**(7078): 847-850.

Fu, H., D. A. Hopwood, et al. (1994). "Engineered biosynthesis of novel polyketides: evidence for temporal, but not regiospecific, control of cyclization of an aromatic polyketide precursor." Chem Biol **1**(4): 205-210.

Fukuchi, N., K. Furihata, et al. (1992). "Rotihibin A, a novel plant growth regulator, from Streptomyces ap." Biosci Biotechnol Biochem **56**(5): 840-841.

Galbraith, E. A., D. A. Antonopoulos, et al. (2004). "Suppressive subtractive hybridization as a tool for identifying genetic diversity in an environmental metagenome: the rumen as a model." Environ Microbiol **6**(9): 928-937.

Gans, J., M. Wolinsky, et al. (2005). "Computational improvements reveal great bacterial diversity and high metal toxicity in soil." Science **309**(5739): 1387-1390.

Gellert-Mortimer, S. T., G. N. Clarke, et al. (1988). "Evaluation of Nycodenz and Percoll density gradients for the selection of motile human spermatozoa." Fertil Steril **49**(2): 335-341.

Gentry, T. J., G. S. Wickham, et al. (2006). "Microarray applications in microbial ecology research." Microb Ecol **52**(2): 159-175.

Gibson, D. G. (2009). "Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides." Nucleic Acids Res **37**(20): 6984-6990.

Gibson, D. G., G. A. Benders, et al. (2008). "Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome." Science **319**(5867): 1215-1220.

Gibson, D. G., G. A. Benders, et al. (2008). "One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome." Proc Natl Acad Sci U S A **105**(51): 20404-20409.

Gibson, D. G., J. I. Glass, et al. (2010). "Creation of a bacterial cell controlled by a chemically synthesized genome." Science **329**(5987): 52-56.

Gietz, R. D. and R. H. Schiestl (2007). "High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method." Nat Protoc **2**(1): 31-34.

Gietz, R. D. and R. H. Schiestl (2007). "Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method." Nat Protoc **2**(1): 38-41.

Gillespie, D. E., S. F. Brady, et al. (2002). "Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA." Appl Environ Microbiol **68**(9): 4301-4306.

Ginolhac, A., C. Jarrin, et al. (2004). "Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones." Appl Environ Microbiol **70**(9): 5522-5527.

Gokhale, R. S. and C. Khosla (2000). "Role of linkers in communication between protein modules." Curr Opin Chem Biol **4**(1): 22-27.

Green, B. D. and M. Keller (2006). "Capturing the uncultivated majority." Curr Opin Biotechnol **17**(3): 236-240.

Guan, C., J. Ju, et al. (2007). "Signal mimics derived from a metagenomic analysis of the gypsy moth gut microbiota." Appl Environ Microbiol **73**(11): 3669-3676.

Gusarov, I., K. Shatalin, et al. (2009). "Endogenous nitric oxide protects bacteria against a wide spectrum of antibiotics." Science **325**(5946): 1380-1384.

Hamady, M., J. J. Walker, et al. (2008). "Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex." Nat Methods **5**(3): 235-237.

Handelsman, J., M. R. Rondon, et al. (1998). "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products." Chem Biol **5**(10): R245-249.

Harrison, D. E., R. Strong, et al. (2009). "Rapamycin fed late in life extends lifespan in genetically heterogeneous mice." Nature **460**(7253): 392-395.

Hendrix, R. W. (2002). "Bacteriophages: evolution of the majority." Theor Popul Biol **61**(4): 471-480.

Herouet, C., D. J. Esdaile, et al. (2005). "Safety evaluation of the phosphinothricin acetyltransferase proteins encoded by the pat and bar sequences that confer tolerance to glufosinate-ammonium herbicide in transgenic plants." Regul Toxicol Pharmacol **41**(2): 134-149.

Hiatt, J. B., R. P. Patwardhan, et al. (2010). "Parallel, tag-directed assembly of locally derived short sequence reads." Nat Methods **7**(2): 119-122.

Homburg, S., E. Oswald, et al. (2007). "Expression analysis of the colibactin gene cluster coding for a novel polyketide in Escherichia coli." FEMS Microbiol Lett **275**(2): 255-262.

Hornung, A., M. Bertazzo, et al. (2007). "A genomic screening approach to the structure-guided identification of drug candidates from natural sources." Chembiochem **8**(7): 757-766.

Howitz, K. T., K. J. Bitterman, et al. (2003). "Small molecule activators of sirtuins extend Saccharomyces cerevisiae lifespan." Nature **425**(6954): 191-196.

Huang, Y., X. Lai, et al. (2009). "Characterization of a deep-sea sediment metagenomic clone that produces water-soluble melanin in Escherichia coli." Mar Biotechnol (NY) **11**(1): 124-131.

Ikeda, H., J. Ishikawa, et al. (2003). "Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis." Nat Biotechnol **21**(5): 526-531.

Imamura, N., K. Kakinuma, et al. (1982). "Biosynthesis of vineomycins A1 and B2." J Antibiot (Tokyo) **35**(5): 602-608.

Ingham, C. J., A. Sprenkels, et al. (2007). "The micro-Petri dish, a million-well growth chip for the culture and high-throughput screening of microorganisms." Proc Natl Acad Sci U S A **104**(46): 18217-18222.

Jannasch, H. W. and G. E. Jones (1959). "Bacterial Populations in sea water determined by different methods of enumeration." Limnol Oceanography **4**(4): 128-139.

Jayaraj, S., R. Reid, et al. (2005). "GeMS: an advanced software package for designing synthetic genes." Nucleic Acids Res **33**(9): 3011-3016.

Jiang, J., C. N. Tetzlaff, et al. (2009). "Genome mining in Streptomyces avermitilis: A biochemical Baeyer-Villiger reaction and discovery of a new branch of the pentalenolactone family tree." Biochemistry **48**(27): 6431-6440.

Joseph, S. J., P. Hugenholtz, et al. (2003). "Laboratory cultivation of widespread and previously uncultured soil bacteria." Appl Environ Microbiol **69**(12): 7210-7215.

Kaeberlein, T., K. Lewis, et al. (2002). "Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment." Science **296**(5570): 1127-1129.

Kalyuzhnaya, M. G., A. Lapidus, et al. (2008). "High-resolution metagenomics targets specific functional types in complex microbial communities." Nat Biotechnol **26**(9): 1029-1034.

Kalyuzhnaya, M. G., M. E. Lidstrom, et al. (2008). "Real-time detection of actively metabolizing microbes by redox sensing as applied to methylotroph populations in Lake Washington." ISME J **2**(7): 696-706.

Kalyuzhnaya, M. G., O. Nercessian, et al. (2005). "Fishing for biodiversity: novel methanopterin-linked C transfer genes deduced from the Sargasso Sea metagenome." Environ Microbiol **7**(12): 1909-1916.

Kalyuzhnaya, M. G., R. Zabinsky, et al. (2006). "Fluorescence in situ hybridization-flow cytometry-cell sorting-based method for separation and enrichment of type I and type II methanotroph populations." Appl Environ Microbiol **72**(6): 4293-4301.

Keller, M. and K. Zengler (2004). "Tapping into microbial diversity." Nat Rev Microbiol **2**(2): 141-150.

Kers, J. A., M. J. Wach, et al. (2004). "Nitration of a peptide phytotoxin by bacterial nitric oxide synthase." Nature **429**(6987): 79-82.

Kim, J., Z. Feng, et al. (2010). "Cloning large natural product gene clusters from the environment: piecing environmental DNA gene clusters back together with TAR." Biopolymers **93**.

Kim, J., L. Simmons, et al. (2009). Metagenomic Approaches to Natural Product Discovery. Comprehensive Natural Products II: Chemistry and Biology. B. S. Moore, Elsevier. **2**.

Kim, U. J., H. Shizuya, et al. (1992). "Stable propagation of cosmid sized human DNA inserts in an F factor based vector." Nucleic Acids Res **20**(5): 1083-1085.

King, R. W., J. D. Bauer, et al. (2009). "An environmental DNA-derived type II polyketide biosynthetic pathway encodes the biosynthesis of the pentacyclic polyketide erdacin." Angew Chem Int Ed Engl **48**(34): 6257-6261.

Kirkpatrick, P., A. Raja, et al. (2003). "Daptomycin." Nat Rev Drug Discov **2**(12): 943-944.

Kodumal, S. J., K. G. Patel, et al. (2004). "Total synthesis of long DNA sequences: synthesis of a contiguous 32-kb polyketide synthase gene cluster." Proc Natl Acad Sci U S A **101**(44): 15573-15578.

Kohanski, M. A., D. J. Dwyer, et al. (2007). "A common mechanism of cellular death induced by bactericidal antibiotics." Cell **130**(5): 797-810.

Komatsu, M., T. Uchiyama, et al. (2010). "Genome-minimized Streptomyces host for the heterologous expression of secondary metabolism." Proc Natl Acad Sci U S A **107**(6): 2646-2651.

Kouprina, N., J. Graves, et al. (1997). "Specific isolation of human rDNA genes by TAR cloning." Gene **197**(1-2): 269-276.

Kouprina, N. and V. Larionov (2006). "Selective isolation of mammalian genes by TAR cloning." Curr Protoc Hum Genet **Chapter 5**: Unit 5.17.

Kouprina, N. and V. Larionov (2008). "Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast Saccharomyces cerevisiae." Nat Protoc **3**(3): 371-377.

Kouprina, N., V. N. Noskov, et al. (2006). "Selective isolation of large chromosomal regions by transformation-associated recombination cloning for structural and functional analysis of mammalian genomes." Methods Mol Biol **349**: 85-101.

Kroiss, J., M. Kaltenpoth, et al. (2010). "Symbiotic Streptomycetes provide antibiotic combination prophylaxis for wasp offspring." Nat Chem Biol **6**(4): 261-263.

Laatsch, H. (2009). AntiBase 2009, Wiley-VCH.

Lamb, S. S., T. Patel, et al. (2006). "Biosynthesis of sulfated glycopeptide antibiotics by using the sulfotransferase StaL." Chem Biol **13**(2): 171-181.

Lange, B. M., T. Rujan, et al. (2000). "Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes." Proc Natl Acad Sci U S A **97**(24): 13172-13177.

Larionov, V., N. Kouprina, et al. (1994). "Transformation-associated recombination between diverged and homologous DNA repeats is induced by strand breaks." Yeast **10**(1): 93-104.

Larionov, V., N. Kouprina, et al. (1996). "Specific cloning of human DNA as yeast artificial chromosomes by transformation-associated recombination." Proc Natl Acad Sci U S A **93**(1): 491-496.

Larionov, V., N. Kouprina, et al. (1996). "Highly selective isolation of human DNAs from rodent-human hybrid cells as circular yeast artificial chromosomes by transformation-associated recombination cloning." Proc Natl Acad Sci U S A **93**(24): 13925-13930.

Letunic, I. and P. Bork (2007). "Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation." Bioinformatics **23**(1): 127-128.

Li, M., F. Diehl, et al. (2006). "BEAMing up for detection and quantification of rare sequence variants." Nat Methods **3**(2): 95-97.

Li, Y., M. Wexler, et al. (2005). "Screening a wide host-range, waste-water metagenomic library in tryptophan auxotrophs of Rhizobium leguminosarum and of Escherichia coli reveals different classes of cloned trp genes." Environ Microbiol **7**(12): 1927-1936.

Liles, M. R., B. F. Manske, et al. (2003). "A census of rRNA genes and linked genomic sequences within a soil metagenomic library." Appl Environ Microbiol **69**(5): 2684-2691.

Liles, M. R., L. L. Williamson, et al. (2008). "Recovery, purification, and cloning of high-molecular-weight DNA from soil microorganisms." Appl Environ Microbiol **74**(10): 3302-3305.

Lim, H. K., E. J. Chung, et al. (2005). "Characterization of a forest soil metagenome clone that confers indirubin and indigo production on Escherichia coli." Appl Environ Microbiol **71**(12): 7768-7777.

Liu, W. C., L. Parker, et al. (1970). "Isolation, characterization, and structure of rabelomycin, a new antibiotic." J Antibiot (Tokyo) **23**(9): 437-441.

Lozupone, C. A., M. Hamady, et al. (2007). "Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities." Appl Environ Microbiol **73**(5): 1576-1585.

Lukashin, A. V. and M. Borodovsky (1998). "GeneMark.hmm: new solutions for gene finding." Nucleic Acids Res **26**(4): 1107-1115.

Ma, H., S. Kunes, et al. (1987). "Plasmid construction by homologous recombination in yeast." Gene **58**(2-3): 201-216.

MacNeil, I. A., C. L. Tiong, et al. (2001). "Expression and isolation of antimicrobial small molecules from soil DNA libraries." J Mol Microbiol Biotechnol **3**(2): 301-308.

Marcy, Y., T. Ishoey, et al. (2007). "Nanoliter reactors improve multiple displacement amplification of genomes from single cells." PLoS Genet **3**(9): 1702-1708.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Martin, V. J. J., D. J. Pitera, et al. (2003). "Engineering a mevalonate pathway in Escherichia coli for production of terpenoids." Nat Biotechnol **21**(7): 796-802.

Martinez, A., S. J. Kolvek, et al. (2004). "Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts." Appl Environ Microbiol **70**(4): 2452-2463.

Marziali, A., J. Pel, et al. (2005). "Novel electrophoresis mechanism based on synchronous alternating drag perturbation." Electrophoresis **26**(1): 82-90.

Mathee, K., G. Narasimhan, et al. (2008). "Dynamics of Pseudomonas aeruginosa genome evolution." Proc Natl Acad Sci U S A **105**(8): 3100-3105.

Mchenney, M. A., T. J. Hosted, et al. (1998). "Molecular cloning and physical mapping of the daptomycin gene cluster from Streptomyces roseosporus." J Bacteriol **180**(1): 143-151.

Menzella, H. G., R. Reid, et al. (2005). "Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes." Nat Biotechnol **23**(9): 1171-1176.

Menzella, H. G., S. J. Reisinger, et al. (2006). "Redesign, synthesis and functional expression of the 6-deoxyerythronolide B polyketide synthase gene cluster." J Ind Microbiol Biotechnol **33**(1): 22-28.

Metsä-Ketelä, M., L. Halo, et al. (2002). "Molecular evolution of aromatic polyketides and comparative sequence analysis of polyketide ketosynthase and 16S ribosomal DNA genes from various streptomyces species." Appl Environ Microbiol **68**(9): 4472-4479.

Miao, V., M.-F. Coëffet-Legal, et al. (2005). "Daptomycin biosynthesis in Streptomyces roseosporus: cloning and analysis of the gene cluster and revision of peptide stereochemistry." Microbiology **151**(Pt 5): 1507-1523.

Miller, D. N., J. E. Bryant, et al. (1999). "Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples." Appl Environ Microbiol **65**(11): 4715-4724.

Moore, B. S. and C. Hertweck (2002). "Biosynthesis and attachment of novel bacterial polyketide synthase starter units." Nat Prod Rep **19**(1): 70-99.

Mootz, H. D. and T. W. Muir (2002). "Protein splicing triggered by a small molecule." J Am Chem Soc **124**(31): 9044-9045.

Müller, C., S. Nolden, et al. (2007). "Sequencing and analysis of the biosynthetic gene cluster of the lipopeptide antibiotic Friulimicin in Actinoplanes friuliensis." Antimicrob Agents Chemother **51**(3): 1028-1037.

Murray, A. W. and J. W. Szostak (1983). "Construction of artificial chromosomes in yeast." Nature **305**(5931): 189-193.

Nagalakshmi, U., K. Waern, et al. (2010). "RNA-Seq: a method for comprehensive transcriptome analysis." Curr Protoc Mol Biol **Chapter 4**: Unit 4 11 11-13.

Neumann, C. S., D. G. Fujimori, et al. (2008). "Halogenation strategies in natural product biosynthesis." Chem Biol **15**(2): 99-109.

Newman, D. J. and G. M. Cragg (2004). "Marine natural products and related compounds in clinical and advanced preclinical trials." J Nat Prod **67**(8): 1216-1238.

Newman, D. J. and G. M. Cragg (2007). "Natural products as sources of new drugs over the last 25 years." J Nat Prod **70**(3): 461-477.

Newman, D. J., G. M. Cragg, et al. (2003). "Natural products as sources of new drugs over the period 1981-2002." J Nat Prod **66**(7): 1022-1037.

Ng, S. B., E. H. Turner, et al. (2009). "Targeted capture and massively parallel sequencing of 12 human exomes." Nature **461**(7261): 272-276.

Nguyen, K. T., D. Ritz, et al. (2006). "Combinatorial biosynthesis of novel antibiotics related to daptomycin." Proc Natl Acad Sci U S A **103**(46): 17462-17467.

Nicolaou, K. C., S. M. Dalby, et al. (2009). "Total synthesis of (+)-haplophytine." Angew Chem Int Ed Engl **48**(41): 7616-7620.

Nougayrède, J.-P., S. Homburg, et al. (2006). "Escherichia coli induces DNA double-strand breaks in eukaryotic cells." Science **313**(5788): 848-851.

O'Donoghue, S. I., A. C. Gavin, et al. (2010). "Visualizing biological data-now and in the future." Nat Methods **7**(3 Suppl): S2-4.

Oldenburg, K. R., K. T. Vo, et al. (1997). "Recombination-mediated PCR-directed plasmid construction in vivo in yeast." Nucleic Acids Res **25**(2): 451-452.

Omura, S., H. Ikeda, et al. (2001). "Genome sequence of an industrial microorganism Streptomyces avermitilis: deducing the ability of producing secondary metabolites." Proc Natl Acad Sci U S A **98**(21): 12215-12220.

Ottesen, E. A., J. W. Hong, et al. (2006). "Microfluidic digital PCR enables multigene analysis of individual environmental bacteria." Science **314**(5804): 1464-1467.

Palaniappan, N., B. S. Kim, et al. (2003). "Enhancement and selective production of phoslactomycin B, a protein phosphatase IIa inhibitor, through identification and engineering of the corresponding biosynthetic gene cluster." J Biol Chem **278**(37): 35552-35557.

Podar, M., C. B. Abulencia, et al. (2007). "Targeted access to the genomes of low-abundance organisms in complex microbial communities." Appl Environ Microbiol **73**(10): 3205-3214.

Polianskaia, L. M., K. E. Ivanov, et al. (2008). "[Estimation of abundance dynamics of gram-negative bacteria in soil]." Mikrobiologiia **77**(6): 848-853.

Polonsky, S., G. Stolovitzky, et al. (2007). "DNA Transistor." IBM Research Report **RC24242**(W0704-094): 23.

Pratt, M. R., E. C. Schwartz, et al. (2007). "Small-molecule-mediated rescue of protein function by an inducible proteolytic shunt." Proc Natl Acad Sci U S A **104**(27): 11209-11214.

Putze, J., C. Hennequin, et al. (2009). "Genetic structure and distribution of the colibactin genomic island among members of the family Enterobacteriaceae." Infect Immun **77**(11): 4696-4703.

Qin, J., R. Li, et al. (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature **464**(7285): 59-65.

Rappé, M. S. and S. J. Giovannoni (2003). "The uncultured microbial majority." Annu Rev Microbiol **57**: 369-394.

Rausch, C., T. Weber, et al. (2005). "Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs)." Nucleic Acids Res **33**(18): 5799-5808.

Reisinger, S. J., K. G. Patel, et al. (2006). "Total synthesis of multi-kilobase DNA sequences from oligonucleotides." Nat Protoc **1**(6): 2596-2603.

Reyes, A., M. Haynes, et al. (2010). "Viruses in the faecal microbiota of monozygotic twins and their mothers." Nature **466**(7304): 334-338.

Rickwood, D., T. Ford, et al. (1982). "Nycodenz: a new nonionic iodinated gradient medium." Anal Biochem **123**(1): 23-31.

Riesenfeld, C. S., R. M. Goodman, et al. (2004). "Uncultured soil bacteria are a reservoir of new antibiotic resistance genes." Environ Microbiol **6**(9): 981-989.

Ro, D. K., E. M. Paradise, et al. (2006). "Production of the antimalarial drug precursor artemisinic acid in engineered yeast." Nature **440**(7086): 940-943.

Rondon, M. R., P. R. August, et al. (2000). "Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms." Appl Environ Microbiol **66**(6): 2541-2547.

Rusch, D. B., A. L. Halpern, et al. (2007). "The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific." PLoS Biol **5**(3): e77.

Sagova-Mareckova, M., L. Cermak, et al. (2008). "Innovative methods for soil DNA purification tested in soils with widely differing characteristics." Appl Environ Microbiol **74**(9): 2902-2907.

Savile, C. K., J. M. Janey, et al. (2010). "Biocatalytic Asymmetric Synthesis of Chiral Amines from Ketones Applied to Sitagliptin Manufacture." Science.

Sawitzke, J. A., L. C. Thomason, et al. (2007). "Recombineering: in vivo genetic engineering in E. coli, S. enterica, and beyond." Methods Enzymol **421**: 171-199.

Schloss, P. D. and J. Handelsman (2004). "Status of the microbial census." Microbiol Mol Biol Rev **68**(4): 686-691.

Schmidt, E. W., J. T. Nelson, et al. (2005). "Patellamide A and C biosynthesis by a microcin-like pathway in Prochloron didemni, the cyanobacterial symbiont of Lissoclinum patella." Proc Natl Acad Sci U S A **102**(20): 7315-7320.

Schneekloth, J., John S, F. N. Fonseca, et al. (2004). "Chemical genetic control of protein levels: selective in vivo targeted degradation." J Am Chem Soc **126**(12): 3748-3754.

Schwartz, E. C., L. Saez, et al. (2007). "Post-translational enzyme activation in an animal via optimized conditional protein splicing." Nat Chem Biol **3**(1): 50-54.

Schwarzer, D., R. Finking, et al. (2003). "Nonribosomal peptides: from genes to products." Nat Prod Rep **20**(3): 275-287.

Seow, K. T., G. Meurer, et al. (1997). "A study of iterative type II polyketide synthases, using bacterial genes cloned from soil DNA: a means to access and use genes from uncultured microorganisms." J Bacteriol **179**(23): 7360-7368.

Shao, Z., H. Zhao, et al. (2009). "DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways." Nucleic Acids Res **37**(2): e16.

Sharan, S. K., L. C. Thomason, et al. (2009). "Recombineering: a homologous recombination-based method of genetic engineering." Nat Protoc **4**(2): 206-223.

Sharma, P. K., N. Capalash, et al. (2007). "An improved method for single step purification of metagenomic DNA." Mol Biotechnol **36**(1): 61-63.

Sharma, S., V. Radl, et al. (2007). "Quantification of functional genes from procaryotes in soil by PCR." J Microbiol Methods **68**(3): 445-452.

Shendure, J., G. J. Porreca, et al. (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome." Science **309**(5741): 1728-1732.

Smolina, I., C. Lee, et al. (2007). "Detection of low-copy-number genomic DNA sequences in individual bacterial cells by using peptide nucleic acid-assisted rolling-circle amplification and fluorescence in situ hybridization." Appl Environ Microbiol **73**(7): 2324-2328.

Sogin, M. L., H. G. Morrison, et al. (2006). "Microbial diversity in the deep sea and the underexplored "rare biosphere"." Proc Natl Acad Sci U S A **103**(32): 12115-12120.

Sommer, M. O., G. Dantas, et al. (2009). "Functional characterization of the antibiotic resistance reservoir in the human microflora." Science **325**(5944): 1128-1131.

Sorenson, T. (1948). "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons." Biologiske Skrifter **5**(4): 1-34.

Soror, S. H., R. Rao, et al. (2009). "Mining the genome sequence for novel enzyme activity: characterisation of an unusual member of the hormone-sensitive lipase family of esterases from the genome of Streptomyces coelicolor A3 (2)." Protein Eng Des Sel **22**(6): 333-339.

Spencer, D. M., P. J. Belshaw, et al. (1996). "Functional analysis of Fas signaling in vivo using synthetic inducers of dimerization." Curr Biol **6**(7): 839-847.

Stepanauskas, R. and M. E. Sieracki (2007). "Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time." Proc Natl Acad Sci U S A **104**(21): 9052-9057.

Stockwell, B. R. and S. L. Schreiber (1998). "Probing the role of homomeric and heteromeric receptor interactions in TGF-beta signaling using small molecule dimerizers." Curr Biol **8**(13): 761-770.

Sudhamsu, J. and B. R. Crane (2009). "Bacterial nitric oxide synthases: what are they good for?" Trends Microbiol **17**(5): 212-218.

Tankéré, S. P. C., D. G. Bourne, et al. (2002). "Microenvironments and microbial community structure in sediments." Environ Microbiol **4**(2): 97-105.

Thompson, J. D., T. J. Gibson, et al. (2002). "Multiple sequence alignment using ClustalW and ClustalX." Curr Protoc Bioinformatics **Chapter 2**: Unit 2.3.

Tian, J., H. Gong, et al. (2004). "Accurate multiplex gene synthesis from programmable DNA microchips." Nature **432**(7020): 1050-1054.

Tobias Kieser, M. J. B., Mark J. Buttner Keith F. Chater David A. Hopwood (2000). Practical Streptomyces Genetics. Colney, Norwich NR4 7UH, England, John Innes Centre.

Torsvik, V., J. Goksøyr, et al. (1990). "High diversity in DNA of soil bacteria." Appl Environ Microbiol **56**(3): 782-787.

Torsvik, V., L. Øvreås, et al. (2002). "Prokaryotic diversity--magnitude, dynamics, and controlling factors." Science **296**(5570): 1064-1066.

Torsvik, V., K. Salte, et al. (1990). "Comparison of phenotypic diversity and DNA heterogeneity in a population of soil bacteria." Appl Environ Microbiol **56**(3): 776-781.

Tringe, S. G., C. von Mering, et al. (2005). "Comparative metagenomics of microbial communities." Science **308**(5721): 554-557.

Trost, B. M. and G. Dong (2008). "Total synthesis of bryostatin 16 using atom-economical and chemoselective approaches." Nature **456**(7221): 485-488.

Tsuruta, H., C. J. Paddon, et al. (2009). "High-level production of amorpha-4,11-diene, a precursor of the antimalarial agent artemisinin, in Escherichia coli." PLoS One **4**(2): e4489.

Turner, E. H., C. Lee, et al. (2009). "Massively parallel exon capture and library-free resequencing across 16 genomes." Nat Methods **6**(5): 315-316.

Tyson, G. W., J. Chapman, et al. (2004). "Community structure and metabolism through reconstruction of microbial genomes from the environment." Nature **428**(6978): 37-43.

Uchiyama, T., T. Abe, et al. (2005). "Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes." Nat Biotechnol **23**(1): 88-93.

Uchiyama, T. and K. Watanabe (2007). "The SIGEX scheme: high throughput screening of environmental metagenomes for the isolation of novel catabolic genes." Biotechnol Genet Eng Rev **24**: 107-116.

Uchiyama, T. and K. Watanabe (2008). "Substrate-induced gene expression (SIGEX) screening of metagenome libraries." Nat Protoc **3**(7): 1202-1212.

Van Wagoner, R. M. and J. Clardy (2006). "FeeM, an N-acyl amino acid synthase from an uncultured soil microbe: structure, mechanism, and acyl carrier protein binding." Structure **14**(9): 1425-1435.

Vetcher, L., Z.-Q. Tian, et al. (2005). "Rapid engineering of the geldanamycin biosynthesis pathway by Red/ET recombination and gene complementation." Appl Environ Microbiol **71**(4): 1829-1835.

Wang, G. Y., E. Graziani, et al. (2000). "Novel natural products from soil DNA libraries in a streptomycete host." Org Lett **2**(16): 2401-2404.

Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.

Webster, G., L. Yarram, et al. (2007). "Distribution of candidate division JS1 and other Bacteria in tidal sediments of the German Wadden Sea using targeted 16S rRNA gene PCR-DGGE." FEMS Microbiol Ecol **62**(1): 78-89.

Weissman, K. J. and P. F. Leadlay (2005). "Combinatorial biosynthesis of reduced polyketides." Nat Rev Microbiol **3**(12): 925-936.

Wenzel, S. C., F. Gross, et al. (2005). "Heterologous expression of a myxobacterial natural products assembly line in pseudomonads via red/ET recombineering." Chem Biol **12**(3): 349-356.

Whitman, W. B., D. C. Coleman, et al. (1998). "Prokaryotes: the unseen majority." Proc Natl Acad Sci U S A **95**(12): 6578-6583.

Wilkinson, D., T. Jeanicke, et al. (2002). "Efficient molecular cloning of environmental DNA from geothermal sediments." Biotechnology Letters **24**(2): 155-161.

Williams, R., S. G. Peisajovich, et al. (2006). "Amplification of complex gene libraries by emulsion PCR." Nat Methods **3**(7): 545-550.

Williamson, L. L., B. R. Borlee, et al. (2005). "Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor." Appl Environ Microbiol **71**(10): 6335-6344.

Wolfgang, M. C., B. R. Kulasekara, et al. (2003). "Conservation of genome content and virulence determinants among clinical and environmental isolates of Pseudomonas aeruginosa." Proc Natl Acad Sci U S A **100**(14): 8484-8489.

Woods, R. A. and R. D. Gietz (2001). "High-efficiency transformation of plasmid DNA into yeast." Methods Mol Biol **177**: 85-97.

Wu, D., P. Hugenholtz, et al. (2009). "A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea." Nature **462**(7276): 1056-1060.

Wyatt, M. A., W. Wang, et al. (2010). "Staphylococcus aureus nonribosomal peptide secondary metabolites regulate virulence." Science **329**(5989): 294-296.

Yap, W., X. Li, et al. (1996). "Genetic diversity of soil microorganisms assessed by analysis of hsp70 (dnaK) sequences." Journal of Industrial Microbiology and Biotechnology **17**(3): 179-184.

Yeh, E., S. Garneau, et al. (2005). "Robust in vitro activity of RebF and RebH, a two-component reductase/halogenase, generating 7-chlorotryptophan during rebeccamycin biosynthesis." Proc Natl Acad Sci U S A **102**(11): 3960-3965.

Yin, J., P. D. Straight, et al. (2007). "Genome-wide high-throughput mining of natural-product biosynthetic gene clusters by phage display." Chem Biol **14**(3): 303-312.

Yin, J., P. D. Straight, et al. (2005). "Genetically encoded short peptide tag for versatile protein labeling by Sfp phosphopantetheinyl transferase." Proc Natl Acad Sci U S A **102**(44): 15815-15820.

Yokouchi, H., Y. Fukuoka, et al. (2006). "Whole-metagenome amplification of a microbial community associated with scleractinian coral by multiple displacement amplification using phi29 polymerase." Environ Microbiol **8**(7): 1155-1163.

Zaehner, D. and F. P. Fiedler (1999). "Fifty years of antimicrobials: past perspectives and future trends in The Need for New Antibiotics: Possibly Ways Forward " 53rd Symposium of the Society for General Microbiology **98**: 67-84.

Zengler, K., G. Toledo, et al. (2002). "Cultivating the uncultured." Proc Natl Acad Sci U S A **99**(24): 15681-15686.

Zengler, K., M. Walcher, et al. (2005). "High-throughput cultivation of microorganisms using microcapsules." Methods Enzymol **397**: 124-130.

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

Zhang, K., A. C. Martiny, et al. (2006). "Sequencing genomes from single cells by polymerase cloning." Nat Biotechnol **24**(6): 680-686.

Zhou, J., M. A. Bruns, et al. (1996). "DNA recovery from soils of diverse composition." Appl Environ Microbiol **62**(2): 316-322.