2016

Natural Products From Functional Screening of Soil Metagenomic Libraries

Hala Iqbal

NATURAL PRODUCTS FROM FUNCTIONAL SCREENING OF SOIL METAGENOMIC LIBRARIES

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Hala Iqbal

June 2016

NATURAL PRODUCTS FROM FUNCTIONAL SCREENING OF SOIL METAGENOMIC LIBRARIES

Hala Iqbal, Ph.D.

The Rockefeller University 2016

Bacteria are a prolific source of therapeutically and industrially relevant natural products. Traditional routes of natural product discovery focus on cultured bacteria, a strategy that eludes 99% of the microbiome. Metagenomics offers a route to study uncultured bacteria by extracting DNA present in environmental samples and cloning this eDNA into a vector to make metagenomic libraries. Libraries can then be screened in various ways to find natural products with desired characteristics and activities. Functional screening uses phenotypically identifiable traits such as changes in color or antibiotic activity to discover clones of interest in a metagenomic library. This mode of screening is contingent upon the bacterial host being able to heterologously express genes present in library clones, and therefore the choice of host is critical in determining the natural products found. In this thesis, I explored the use of two heterologous hosts for functional screening: *Ralstonia metallidurans* and *Streptomyces albus*, and found that both are powerful hosts for finding a wide range of natural products.

Functional screening of a 700,000 cosmid clone library in *R. metallidurans* for antibiotic activity led the discovery of a diverse set of antibacterially active enzymes. Clones displaying zones of inhibition of growth of an assay strain in top agar bacterial overlays were predicted through bioinformatic analysis to encode peptidases, lipases and cell-wall lytic enzymes conferring antibiotic activity, and confirmed using transposon mutagenesis and subcloning. These antibacterial activities could not be replicated when clones were transformed into *Escherichia coli*, indicating the powerful ability of *R. metallidurans* to find diverse antibacterially active enzymes in metagenomic screens.

Streptomyces is an excellent candidate for finding small molecules from functional metagenomic screens because it natively and heterologously produces a diverse range of small molecules. Functional screening in *Streptomyces* has not been successful to date: small libraries have been built in *Streptomyces lividans* and screening has not resulted in many small molecules being found. A collection of 39 Streptomyces strains was analyzed for transformation efficiency and heterologous expression. *S. albus* was identified as a strain with high heterologous expression capabilities. To transform a large metagenomic library into *Streptomyces albus*, a high quality library was built in *E. coli* using an *E. coli – Streptomyces* shuttle vector and gel-purified to remove any empty vector. Transformation of this library into *S. albus* was optimized by

varying mating conditions. A 1.5 million clone library was built in *S. albus*, making this the first metagenomic library reported in *S. albus*, and is approximately 100x larger than any other library reported in *Streptomyces* to date. Preliminary screens of this library for colonies displaying changes in color were used to identify 12 clones, of which 8 contained carotenoid pathways, 1 contained a type III polyketide pathway and 3 contained pathways that could not be easily identified. Clone specific small molecules from these hits will be further analyzed for novel secondary metabolites.

This thesis represents efforts at solving one of the problems limiting the success of functional screening, namely choice of heterologous host. Screens in two hosts, *R. metallidurans* and *S. albus*, are described that lead to the identification of different types of natural products: antibacterially active enzymes and small molecule biosynthetic pathways. This reiterates the importance of heterologous host choice in functional screening of metagenomic libraries for natural products with desired characteristics and activities.

ACKNOWLEDGEMENTS

To Sean Brady, my graduate advisor, for the mentoring. For the ideas, for the rigor in negative controls, for knowing when to cut losses and move on. For always having an open door. To my thesis committee: Vincent Fischetti, Alexander Tomasz, Minkui Luo for the advice and pointed questions.

To my labmates, for all the support. For answering questions, scientific and silly and philosophical, and various combinations thereof. For sharing reagents and techniques and food. For late nights, early mornings, and generosity in taking things out of the autoclave. Moments of rage and tension-ridden games of Settlers of Catan and letting me play with your babies and enabling my Asia Grill (later, dosa) obsession. For the life advice.

To the entitled med students who studied on our floor. For uniting us in our hate.

To my biological family. My parents, for the constant loving. For putting up with unanswered phone calls and photos of me pulling a face. For making me strong.

To my brother, for being my rock. For knowing how to artfully deal with people, for secret languages and unending gchat conversations and most of all, for Arrested Development and Top Chef marathons.

To my Boston family, for being an escape. For Survivor and swimming and treasure hunts. Elaborate desserts and South Indian food and walnuts. For the company of cute, tiny, easy to manipulate humans, for watching them grow up.

To my chosen family, for having my back. My friends, for teaching me to love without restraint, for understanding my non-hugginess, for holding space for my awkward. For beach trips and spontaneous adventures. For all the conversations.

My platonic life partner, for all the memoir titles and despair.

LK, for keeping me going these past few months.

FC for all the edits and Skype sessions. For understanding.

SC and QD, for dreams of Fiji. For nerdery and the one holiday I actually enjoy because of you two.

KC for listening. For plotting (upcoming) revolutions.

KN, CC, AL for putting up with my Harry Potter obsession. For all the links and book recommendations and Cards Against Humanity. For the urgent cat videos.

ES for the adventures.

To Rugby and my various rugby families, for teaching me discipline and fearlessness. For all the fun.

To my communities, for what we're building. For changing the world.

To Audre, my cat, for changing my life.

To all, my infinite love and gratitude.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

aa	Amino acid
Amp	Ampicillin
ATCC	American Type Culture Collection
BLAST	Basic local alignment search tool
CIP	Calf intestinal alkaline phosphatase
CoA	Coenzyme A
cos	λ phage cohesive end segment
DNA	Deoxyribonucleic acid
eDNA	Environmental DNA
ESI	Electrospray ionization
GB	Gigabase
H_2O	Water
HMW	High Molecular Weight
HPLC	High pressure liquid chromatography
HRMS	High resolution mass spectrometry
kb	Kilobase
LB	Luria Broth medium
MB	Megabase
MIC	Minimum Inhibitory Concentration
MS	Mass spectrometry
NMR	Nuclear magnetic resonance
NRPS	Non-ribosomal peptide synthase

ORF	Open Reading Frame
PCR	Polymerase chain reaction
PKS	Polyketide synthase
rpm	Rotations per minute
rRNA	ribosomal ribonucleic acid
TSB	Tryptic Soy Broth medium
TX	Texas

CHAPTER 1: Introduction and Background

1.1 Chapter Summary

The vast majority of bacteria present in environmental samples have never been cultured and therefore have not been available to explore for the ability to produce useful natural products such as biocatalysts, resistance-mediating enzymes, or biosynthetic clusters that produce therapeutically or industrially relevant small molecules. Metagenomic libraries comprising of DNA extracted directly from natural bacterial communities offer access to the genetic information present in the genomes of as yet uncultured bacteria. This chapter highlights recent efforts to recover natural products from metagenomic libraries.

1.2 The Importance of Natural Products

Natural products are broadly defined as molecules that are produced by living organisms as part of their biological process. More specifically, natural products can be divided into two classes: primary and secondary metabolites. Primary metabolites are categorized as molecules that are involved in processes necessary to the survival of an organism. Secondary metabolites are not directly essential to the continued existence of an organism but may confer an evolutionary advantage and

the absence of these molecules may result in impaired functions such as growth or development (Demain & Fang, 2000).

Natural products from both primary and secondary metabolism have had a long history of therapeutic and industrial use. Primary metabolites such as solvents, organic acids and macromolecules such as amino acids, nucleotides, etc. have been employed in industrial processes and as food supplements (Adrio *et al.*, 2010). Some examples include citric acid, which is obtained from fermentation of the fungus *Aspergillus niger* and is used to add a sour taste to food, and vitamins, which are harvested from plant or animal sources as well as microbial fermentation and used as dietary supplements (Ramachandran *et al.*, 2007; Max *et al.*, 2010).

Secondary metabolite natural products, in particular, display a broad range of uses and targets that have been harnessed in pharmaceutical and industrial settings. Because these molecules are not essential to the continued existence of the organism, they are postulated to result in a competitive edge for survival in an environmental niche. Secondary metabolites are used by organisms in processes such as chemical warfare, inter- and intra-species communication, metal scavenging, and symbiosis, which translate to clinically-relevant activities as diverse as antibiosis, analgesia, anti-tumor and anti-parasitic activity (Williams *et al.*, 1989; Demain & Fang, 2000). Some examples of

secondary metabolite natural products with activity widely utilized by humans include Morphine from the opium poppy *Papaver somniferum*, which is hailed therapeutically for its pain-relief properties, and Trabectedin from the marine tunicate *Ecteinascidia turbinate* that is employed as an anti-tumor drug (Tohme *et al.*, 2011; Laux-Biehlmann *et al.*, 2013).

Despite the importance of natural products, natural product research has been in decline in industries such as drug discovery and development (Newman & Cragg, 2012). In the current climate of emerging diseases and drug-resistance, natural product research has been deemphasized in spite of the fact that molecules that are naturally produced, derived, or inspired constitute 64% of FDA-approved drugs since 1981, comprising 69% of all anti-infectives and 75% of anti-tumor drugs (Cragg & Newman, 2013; Harvey *et al.*, 2015). Since natural products are often honed from billions of years of evolutionary selection, these molecules are considered privileged structures because they display complex chemistries as well as precise targets and functions that are frequently more effective than their rationally designed counterparts (DeSimone *et al.*, 2004; Bongarzone & Bolognesi, 2011). This underscores the critical need for and continued relevance of natural product research.

1.3 Bacteria as a Source of Natural Products

Bacteria are a rich source of natural products, valued for their ability to both endogenously and recombinantly produce a range of molecules displaying diverse chemistries and activities. Endogenously produced bacterial secondary metabolites in particular, have been a prolific source of therapeutically relevant compounds including proteins and various classes of small molecules spanning families such as polyketides, ribosomal peptides, non-ribosomal peptides, and terpenoids. These molecules are made from different combinations of relatively simple repeating building blocks, resulting in complex structures and potent bioactivities. Some examples of endogenously produced natural products in bacteria and their activities are: vancomycin, an antibiotic; acarbose, an anti-diabetic drug; mitomycin, with anti-tumor activity; spinosyn, an insecticide; and rapamycin, which is used clinically as an immunosuppressant (Figure 1) (Hata *et al.*, 1956; Walton *et al.*, 1979; Thomson & Woo, 1989; Levine, 2006; Kirst, 2010).

In addition to endogenously produced natural products, bacteria are also used to recombinantly express molecules. High yields of natural products can be obtained because bacteria are well suited to grow and manipulate in large quantities. Examples of recombinantly produced natural products in bacteria in current therapeutic use include insulin, bovine growth hormone, and interferons (Miller, 1979; Newmark, 1980).

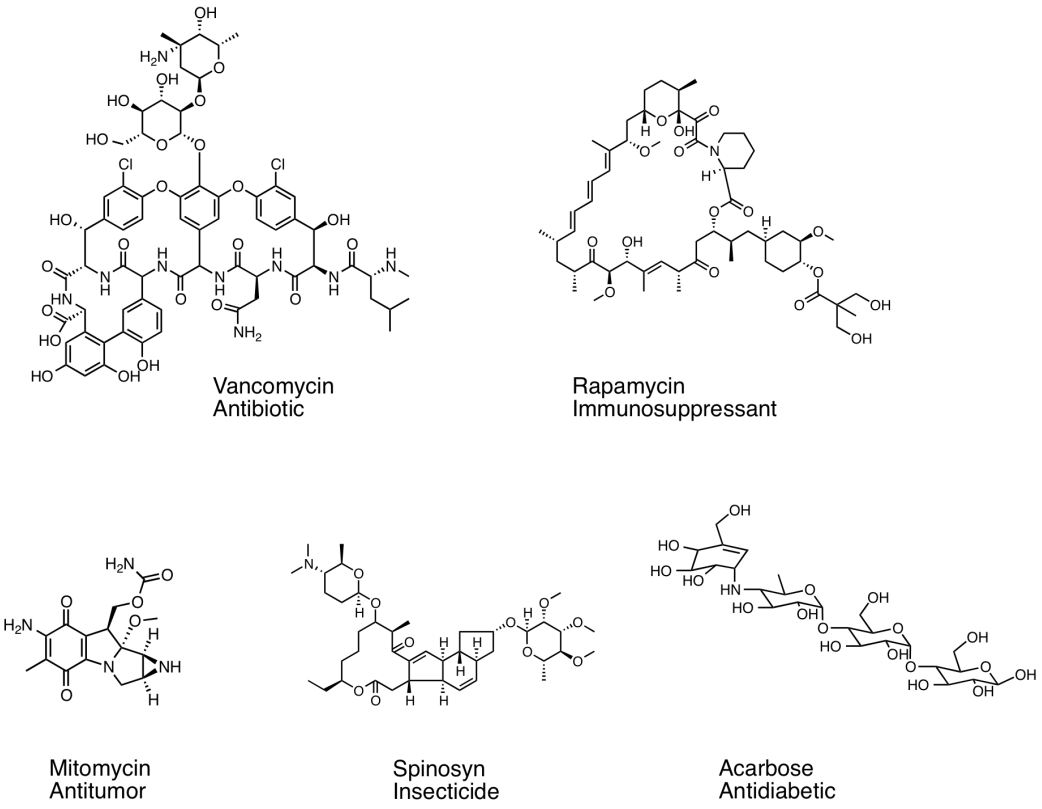


Figure 1. Examples of endogenously produced bacterial natural products with corresponding bioactivities

1.4 Natural Product Discovery in Bacteria

Natural product discovery in bacteria traditionally focuses on finding new molecules and activities from cultured bacteria. Starting with an environmental sample, pure cultures of bacteria are grown in laboratory settings and bioactivity-guided assays are used to identify and extract relevant natural products. The golden age of this method of discovery was in the 1940s – 1960s, when novel natural products were being reported at

unparallel rates (Lewis, 2013). Since then, the discovery of novel natural products has been in decline for many reasons, including high rates of re-discovery of known natural products, and the deprioritizing of natural product discovery programs in the pharmaceutical research industry (Li & Vederas, 2009).

One of the limitations of the traditional route of bacterial natural product discovery that accounts for the high rates of rediscovery of molecules is its heavy reliance on cultured bacteria. Methods as rudimentary as comparing the number of bacteria in an environmental sample grown on plates with the number seen under a microscope reveal vast inconsistencies, with uncultured bacteria greatly outnumbering cultured bacteria grown in laboratory settings. This was known for decades as the “great plate count anomaly” (Lewis *et al.*, 2010). With the advent of genomic sequencing, analysis of 16S genes in environmental samples estimates that the percentage of uncultured bacteria is as high as 99%, meaning that most bacteria in the earth’s biosphere have never been explored for the production of potentially useful natural products (Torsvik *et al.*, 1990; Rappe & Giovannoni, 2003).

1.5 Metagenomics in Natural Product Discovery

Metagenomics offers a strategy to circumvent the limitations of traditional natural product discovery. Whereas it is not possible to characterize the

natural products produced by uncultured bacteria using traditional microbiological methods, it is possible to instead extract microbial DNA directly from an environmental sample (environmental DNA, eDNA) and clone this DNA into easily cultured bacterial host, thereby bypassing the need for culturing bacteria (Figure 2). This general strategy of studying uncultured bacteria genomes has been termed "metagenomics" (Handelsman *et al.*, 1998).

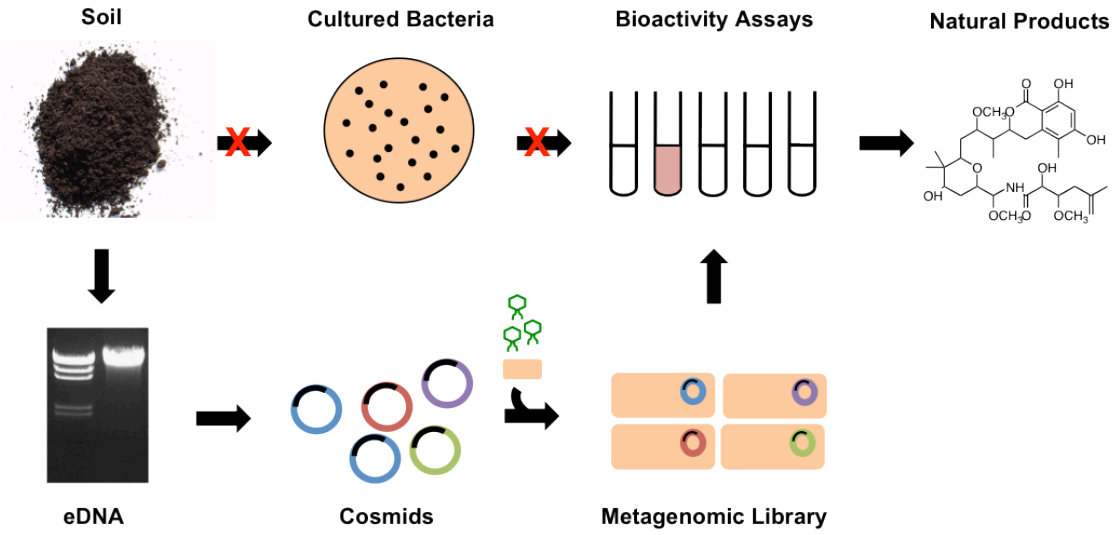


Figure 2. A metagenomic-based approach to natural product discovery. DNA is extracted from environmental samples and used to make metagenomic libraries, which can be screened for natural products. This circumvents the need to culture bacteria in laboratory settings and expands the number of bacteria that can be interrogated for the production of natural products by 100 fold.

Metagenomic libraries are used to capture and store the genetic information in environmental samples. Libraries are made by cloning eDNA from a sample into a suitable cosmid vector, and transforming this cosmid library first into E. coli, and then into a heterologous host. Metagenomic libraries constructed using DNA extracted directly from naturally occurring bacterial populations can then be extensively screened for clones that have the genetic capacity to produce new biocatalysts as well as small molecule natural products.

1.6 Methods for Screening Metagenomic Libraries

Whether looking for novel enzymes or new small molecules, most approaches employed to screen metagenomic libraries can be divided into two general categories: 1. functional screening, which relies on the heterologous expression of eDNA in a model cultured host to yield a phenotype of interest such as color or antibiosis, and 2. homology screening, which relies on DNA sequence similarity to identify clones containing a specific gene of interest (Figure 3). With recent advents in next generation sequencing technology, a third type of sequence-based screening has emerged in which eDNA is sequenced directly from soil samples and used to guide the discovery of biosynthetic clusters.

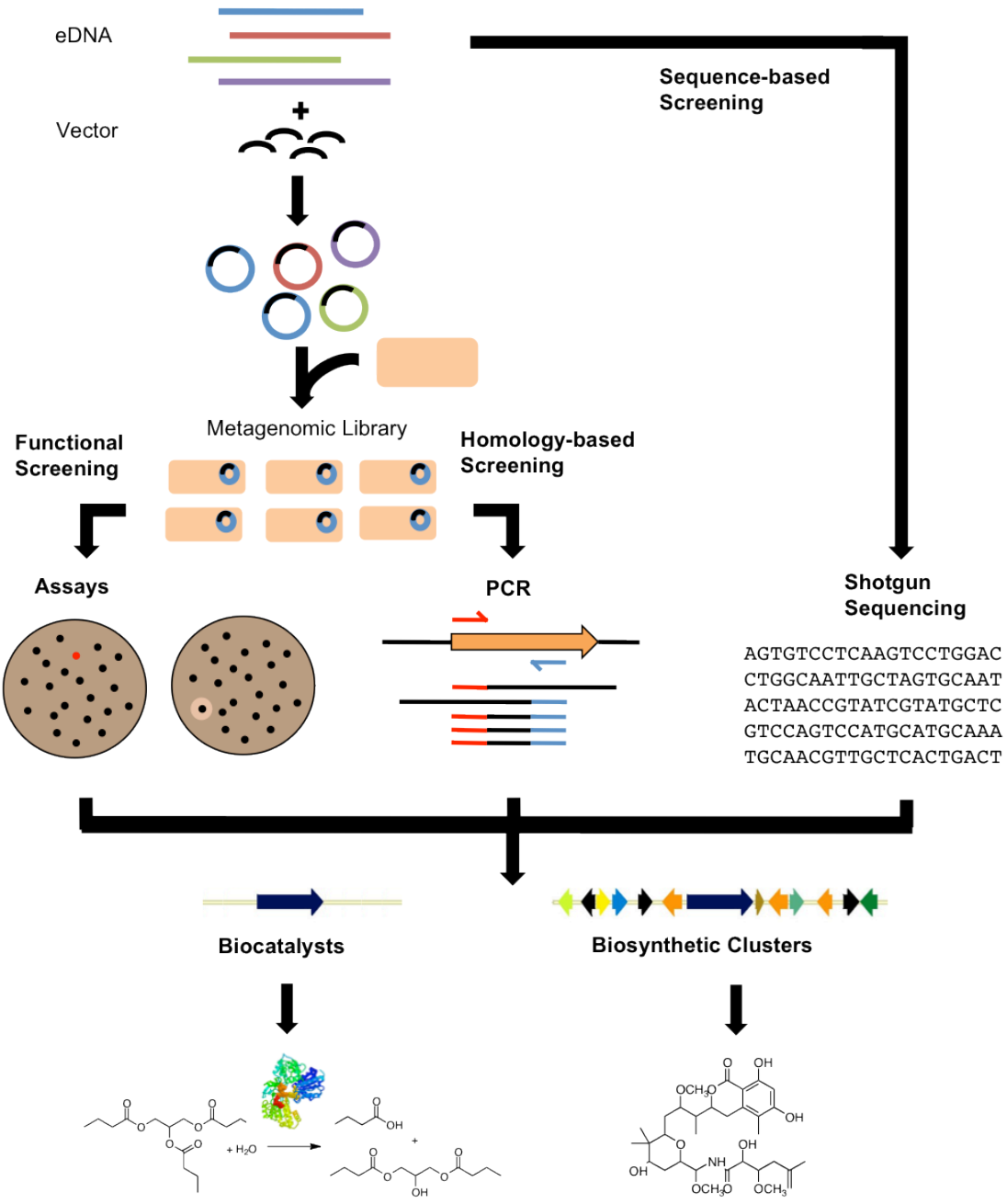


Figure 3. Overview of metagenomic screening methods.

Presented here are recent functional and homology-based metagenomic studies that have identified either novel industrially-relevant bacterial enzymes or collections of enzymes (gene clusters) that encode the biosynthesis of interesting small molecules. As with the search for discrete biocatalysts, the search for collections of biocatalysts that can biosynthesize small molecules of interest has been carried out using both functional and homology-based screening approaches.

1.7 Functional Screening for Discrete Genes

Functional screening of metagenomic libraries for discrete genes such as resistance markers, biocatalysts and other industrially relevant enzymes has often been conducted using simple screens with the desired function, for example enzymatic activity or resistance, as a direct readout to find clones in libraries. In recent years, functional screening efforts have expanded beyond these simple assays to include the use of reporter genes and complementation as tools for identifying metagenomic clones encoding enzymes of interest.

In functional screens for genes conferring resistance to an antibiotic, metagenomic libraries are plated out on media containing concentrations of antibiotic above the MIC to identify clones with genes conferring resistance to the antibiotic. Novel resistance genes that have been found using this methodology include antibiotic-modifying enzymes

and new efflux pumps, many of which show low similarities to known resistance genes and would not have been predicted to confer resistance based on gene annotations (Donato *et al.*, 2010; Parsley *et al.*, 2010; McGarvey *et al.*, 2012).

Similar to screening for resistance, metagenomic libraries can be screened for enzymes of interest in simple direct readout assays for enzymatic activity. Libraries are plated on media containing a substrate for an enzyme of interest, and the appearance of either colored substrate or a halo indicating the processing of a substrate is then used to identify clones encoding the product of the desired enzyme. For example, plates containing 5-bromo-4-chloro-3-indolyl caprylate (X-caprylate) or 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal), both of which yield a blue precipitate upon hydrolysis, have been used to recover novel esterases and β-galactosidases from metagenomic libraries (Wang *et al.*, 2010; Yao *et al.*, 2011). In another recent study, Hu et. al. screened an Antarctic desert soil metagenomic library for esterases using agar plates containing the substrate triglyceride tributyrin (Hu *et al.*, 2012). From a clone surrounded by a clear halo, which indicated tributyrin hydrolysis, they characterized a cold-active esterase only distantly related to reported lipases. Other examples of identifying novel enzymes from metagenomic libraries using this approach include screening for proteases on plates supplemented with skim milk, amylases on starch plates and cellulases on

carboxymethylcellulose plates (Pang *et al.*, 2009; Neveu *et al.*, 2011; Liu *et al.*, 2012).

Unfortunately, most enzymes are not amenable to identification by simply screening for changes in colony appearance. Uchiyama and Miyazaki have developed "gene-expression" reporter assays in an attempt to overcome these limitations. In their product-induced gene-expression (PIGEX) assay, a reporter gene is coupled to a product-sensitive transcription factor such that the presence of the product of a desired enzymatic reaction leads to the transcription of the reporter gene (Uchiyama & Miyazaki, 2010). Their proof of principle study used green fluorescent protein (GFP) placed under the control of a benzoate-sensitive transcription factor to screen for new amidases. In this assay, amidases convert benzamide to benzoate resulting in the expression of GFP by the benzoate-sensitive transcription factor. Using a wastewater sludge metagenomic library, the authors found 11 unique amidases, three of which were distantly related to known amidases. Conceptually similar assays, including a substrate-induced gene-expression (SIGEX) assay designed to identify enzymes involved in the catabolism of a compound of interest and a lacZ-based reporter assay designed to identify attenuators of quorum sensing, have also been used to screen metagenomic libraries (Uchiyama *et al.*, 2005; Schipper *et al.*, 2009; Bijtenhoorn *et al.*, 2011).

In addition to reporter gene assays, complementation has been used as a strategy for isolating novel biocatalysts from metagenomic libraries. In these studies, the ability of metagenomic clones to restore, or complement, a mutation in a reporter strain, is used to detect the expression of an enzyme of interest. This approach was used by Simon et. al. to identify new DNA polymerases from uncultured environmental bacteria. In this study, a glacial ice metagenomic library was transformed into a temperature-sensitive strain of *E. coli* harboring a cold-inactive polymerase mutation that is lethal below 20^{o}C (Simon *et al.*, 2009) [14]. When the library was shifted to 18^{o}C only E. coli containing clones capable of complementing the temperature-sensitive polymerase mutation could grow. In total, nine complementing clones were recovered and they were found to encode either DNA polymerases or domains typical of polymerase enzymes. In another recent example of a complementation screen, a metatranscriptomic library was screened for histidine biosynthesis genes using survival on histidine dropout plates as readout (Kellner *et al.*, 2011).

Table 1: Summary of representative metagenomic screens and corresponding biocatalysts found.

Enzymes Found	Screening Method	eDNA source	Hits	Clones	Insert Size (kb)	Reference
Esterase	Functional	Cotton field soil	1	92,000	3.5	(Yao *et al.*, 2011)
β-galacto-sidase	Functional	Oil field soil	3	12,000	4.8	(Wang *et al.*, 2010)
Esterase	Functional	Antarctic soil	1	10,000	30	(Hu *et al.*, 2012)
Protease	Functional	Desert soil	1 16	30,000 17,000	6 32	(Neveu *et al.*, 2011)
Amylase	Functional	Marine sediment	1	20,000	30 – 40	(Liu *et al.*, 2012)
Cellulase	Functional	Compost soil	4	100,000	33	(Pang *et al.*, 2009)
Amidase	Gene reporter	Wastewater sludge	11	96,000	30 – 40	(Uchiyama & Miyazaki, 2010)
Phenol degradation	Gene reporter	Groundwater	62	152,000	7	(Uchiyama *et al.*, 2005)
Oxido-reductase	Gene reporter	Soil	1	8,000	2.5 – 6	(Bijtenhoorn *et al.*, 2011)
DNA polymerase	Complementation	Glacial ice	9	230,000	4	(Schipper *et al.*, 2009)
Histidine biosynthesis	Complementation	Forest soil	1	13,000		(Simon *et al.*, 2009)
Protease	PCR homology	Grassland soil Grassland soil Wastewater	2 3 5	11,520 30,494 26,800	40 40 30	(Niehaus *et al.*, 2011)
Herbicide degradation	PCR homology	Agricultural and forest soil	437			(Zaprasis *et al.*, 2010)
Copper P-ATPase	PCR homology	Copper waste-sediment	14			(De la Iglesia *et al.*, 2010)
Cellulase	Shotgun sequencing	Cow rumen	27,755			(Hess *et al.*, 2011)
Methyl halide transferase	Synthetic metagenomics	NCBI database	89			(Bayer *et al.*, 2009)
Cellulase	Shotgun sequencing	Compost	800			(Allgaier *et al.*, 2010)

1.8 Functional Screening for Biosynthetic Pathways

Assays for detecting clones of interest harboring biosynthetic pathways for the production of small molecules have focused primarily on simple phenotypically identifiable traits such as antibiosis and color that are often associated with secondary metabolite production. Early functional screens

of metagenomic libraries for were hosted in E. coli (Figure 4). Screening for changes in colony color led to the identification of known blue and pink pigments indigo blue and indirubin, brown pigmented bacterial melanins, tetrapyrrole pigments from heme biosynthesis, the purple-colored violacein, and the novel tri-aryl cation antibiotics turbomycin A and B that displayed both color and antibiotic activity (Brady *et al.*, 2001; MacNeil *et al.*, 2001; Gillespie *et al.*, 2002; Wilkinson *et al.*, 2002; Huang *et al.*, 2009). In top agar overlay assays for antibiotic activity, metagenomic libraries clones were overlaid with a thin layer of top agar containing an assay strain of bacteria. Zones of inhibition of growth around colonies indicating antibiosis of the assay strain were used to identify clones producing N-acyl amino acids and derivatives, as well as isocynanide-functionalized idoles (Brady & Clardy, 2000; Brady *et al.*, 2002; Brady & Clardy, 2005; Brady & Clardy, 2005).

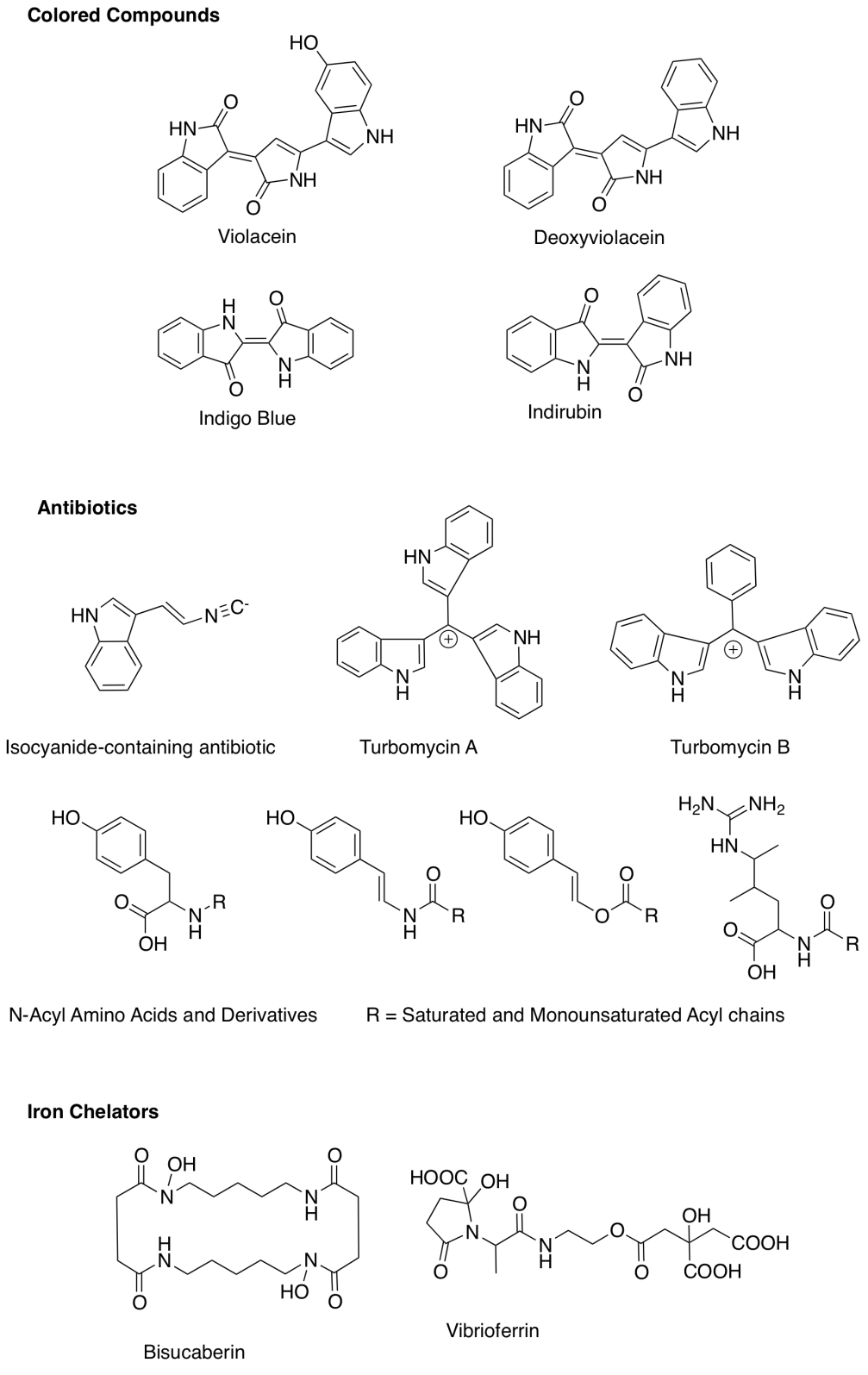
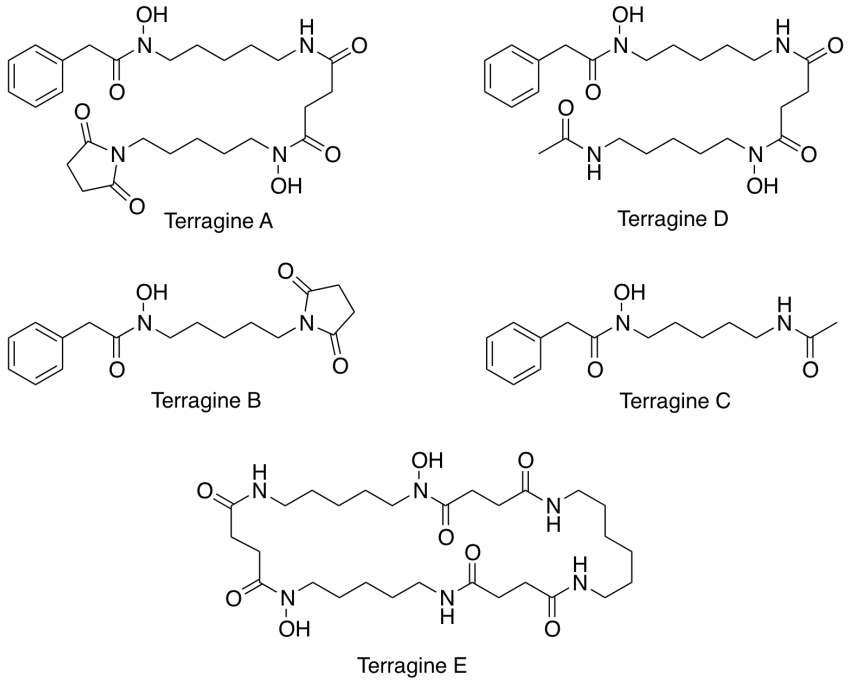


Figure 4. Small molecules found in functional metagenomic screens using *E. coli* as a heterologous host

E. coli is known to display limited heterologous expression capabilities, and because functional screening is contingent upon the metagenomic library host expressing the genes encoded on eDNA inserts to impart a distinct phenotype to the bacterial colony, expanding screening to libraries hosted in other bacterial hosts led to the discovery of novel classes of natural products (Figure 5) (Gabor *et al.*, 2004). The screening of organic extracts of a 1020-clone library hosted in *Streptomyces lividans* led to the identification of the terragine family of molecules (Wang *et al.*, 2000). The soil-dwelling proteobacteria *Ralstonia metallidurans* was identified as a powerful heterologous host, and screening libraries hosted in *R. metallidurans* led to the detection of the colored carotenoid β-carotene, as well as type III polyketide molecules (Craig *et al.*, 2009). Functional screening of libraries hosted in diverse proteobacteria from soil – namely, *Agrobacterium tumefaciens*, *Burkholderia graminis*, *Caulobacter vibrioides* and *Pseudomonas putida* found that the phenotypes and compounds that were detectable depended heavily on the heterologous host used (Craig *et al.*, 2010).

Compounds found in *Streptomyces lividans*



Compounds found in *Ralstonia metallidurans*

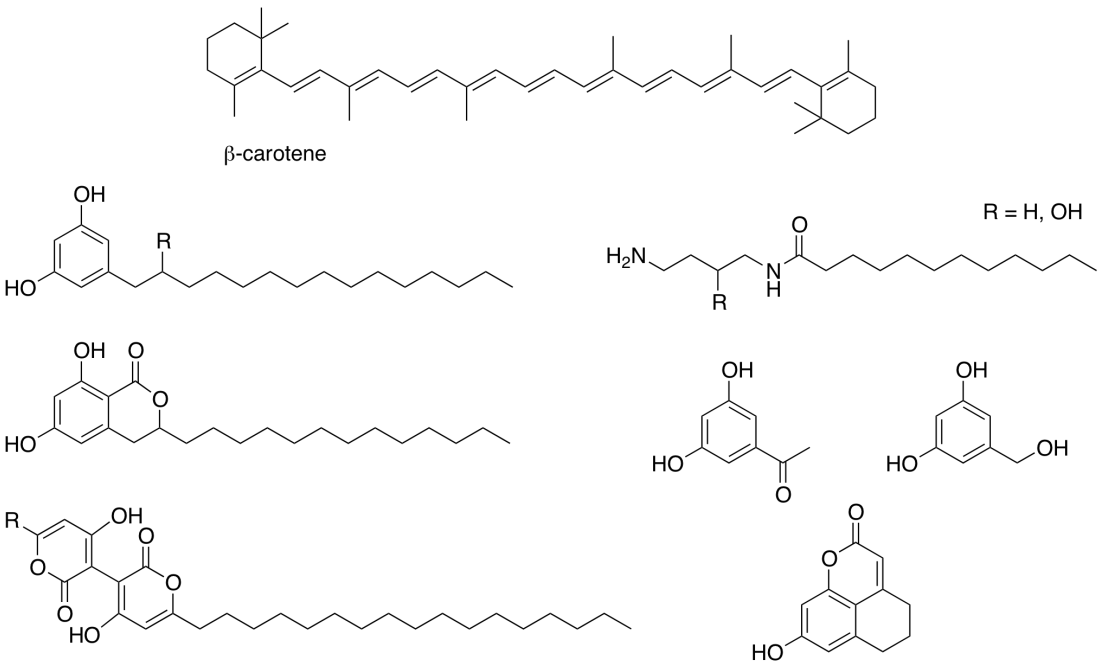


Figure 5. Small molecules found from functional metagenomic screening efforts in alternative heterologous hosts

One of the limits of functional screening has been the small number of high throughput assays available for detecting clones that produce small molecules, and new assays are sorely needed. In an example of a novel screen, Fujita et. al. reported the use of the indicator Chrome Azurol S (CAS), which changes from orange to blue in the presence of iron, to isolate clones encoding siderophores (iron chelators) from marine metagenomic libraries hosted in *E. coli*. In these studies they recovered gene clusters from the known siderophores Bisucaberin and Vibrioferrin (Fujita *et al.*, 2011; Fujita *et al.*, 2012).

1.9 Homology-Based Screening for Discrete Genes

Homology-based screening has typically involved the use of PCR and/or colony hybridization to identify new members of a known family of genes. PCR screening has been used, for example, to identify new proteases, herbicide-degrading genes and copper resistance enzymes (De la Iglesia *et al.*, 2010; Zaprasis *et al.*, 2010; Niehaus *et al.*, 2011). In an extension of this strategy, Wang et. al. generated a library of chimeric lipases with distinct substrate specificities by shuffling pools of eDNA derived PCR amplified lipase gene fragments (Wang *et al.*, 2010).

With advances in next generation sequencing, it has now become possible to bypass experimental hybridization methods and instead use bioinformatics to detect conserved enzymatic sequence motifs in shotgun

sequenced eDNA. This approach is intrinsically higher throughput and allows for more flexible homology searches than PCR. Hess, et. al. shotgun sequenced a 286 Gb cow rumen metagenomic library and searched this dataset for cellulolytic enzymes with potential applications in the biofuel industry (Hess *et al.*, 2011). Their homology search revealed 28,000 potential carbohydrate active genes, of which 90 candidate genes were amplified from cow rumen eDNA, expressed and tested for activity against 10 different carbohydrate substrates. More than half of these enzymes were verified as active on at least one substrate.

The utility of sequence-based screening has also been extended to existing metagenomic datasets. Although the DNA encoding sequences found in public databases are not directly available for functional analysis, enzymes of interest can be characterized using what has been called "synthetic metagenomics" (Bayer *et al.*, 2009). In this approach, genes of interest are codon optimized, chemically synthesized, cloned and then expressed in a heterologous host. Bayer, et. al. bioinformatically mined methyl halide transferase (MHT) enzymes, which can be used to produce agriculturally relevant fumigants, from the NCBI sequence database comprising both cultured and uncultured organisms. Their homology search revealed 89 potential MHT genes, which were chemically synthesized and then expressed in *E. coli.* All but 5 of the enzymes were verified to produce methyl halides in the presence of halide salts. A similar

approach coupling sequence-based screening with gene synthesis, was used by Allgaier et. al. to study glycoside hydrolases from metagenomic sequencing data derived from compost bacteria (Allgaier *et al.*, 2010).

1.10 Homology-based Screening for Small Molecules

In homology-based small molecule discovery efforts, a metagenomic library, or even crude eDNA sequencing data, is probed to identify gene clusters containing conserved sequences that are predicted to be associated with the biosynthesis of a molecule of interest. Complete eDNA derived gene clusters and individual eDNA derived enzymes from metagenomic libraries have been used to generate both new and derivatives of known bioactive secondary metabolites. Enumerated below are some recent examples (Figure 6).

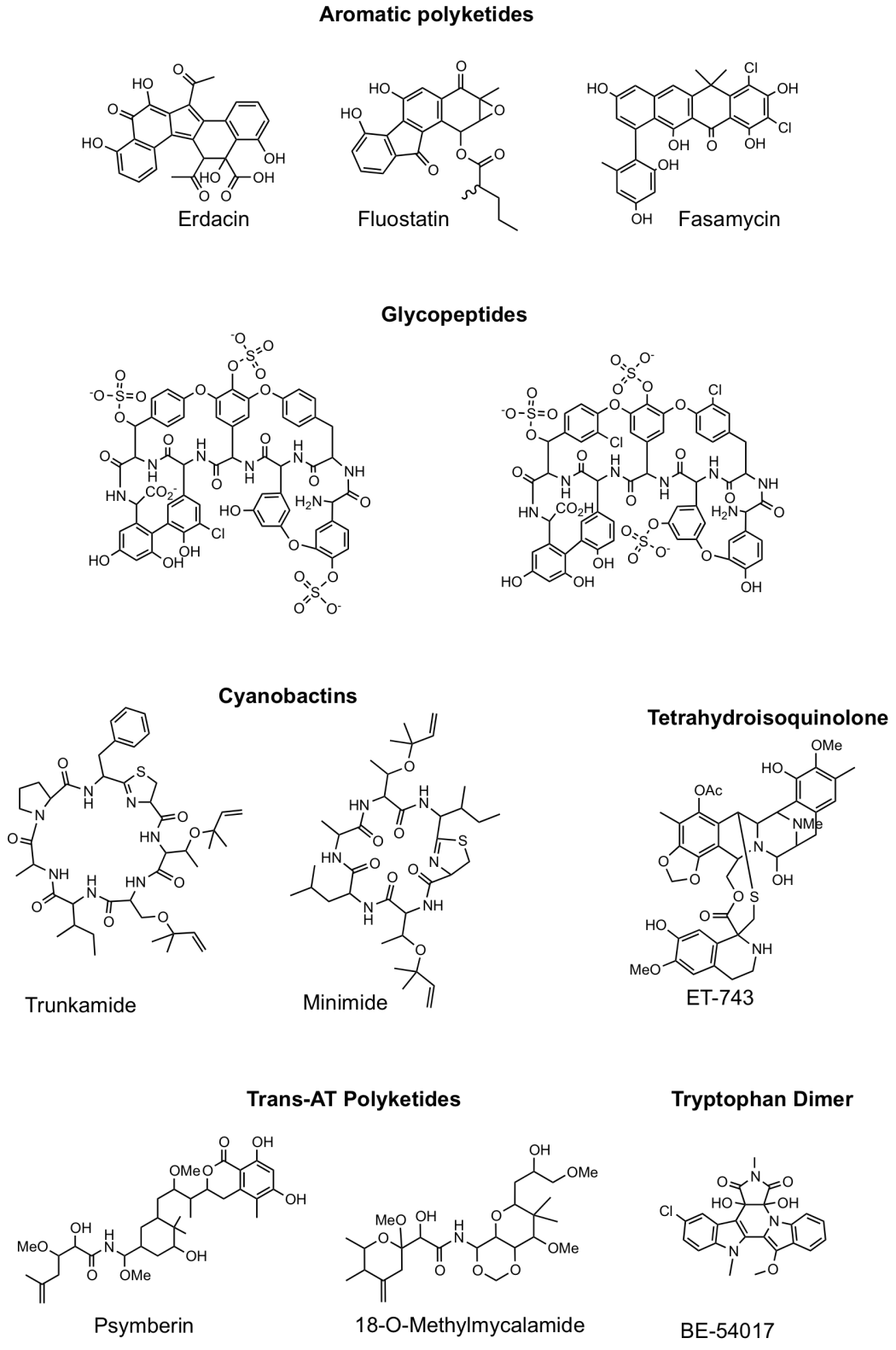


Figure 6. Small molecules found from homology-based screening of metagenomic libraries

1.10.1 Glycopeptides:

Vancomycin and teicoplanin are clinically used glycopeptide antibiotics that exhibit activity against methicillin resistant Gram-positive bacteria. Banik et. al. used degenerate primers based on OxyC, a conserved oxidative coupling enzyme found in vancomycin and teicoplanin-like glycopeptide gene clusters, to identify and recover multiple predicted glycopeptide-encoding gene clusters from soil metagenomic libraries (Banik & Brady, 2008). The recovery of complete biosynthetic gene clusters in these studies required the construction of megalibraries containing in excess of 10,000,000 unique cosmid clones. Recombinant sulfotransferases from one pathway were used *in vitro* to modify the teicoplanin aglycone, producing novel mono, di- and tri-sulfated glycopeptide derivatives. In a subsequent study, tailoring enzymes found in eDNA-derived glycopeptide biosynthetic clusters were expressed in *Streptomyces toyocaensis*, which naturally produces the mono-sulfated glycopeptide A47934 (Banik *et al.*, 2010). This resulted in new glycopeptide derivatives featuring methyl, sulfur and sugar substituents, which were further derivatized in vitro using sulfotransferases. In total 15 new anionic (sulfated) glycopeptide antibiotics were generated in these studies.

1.10.2 Cyanobactins

Cyanobactins are ribosomally-produced cyclic peptides that are prevalent in extracts derived from marine samples, and they frequently display interesting cytotoxic activities (Sivonen *et al.*, 2010). In 2005, two separate groups reported the cloning and heterologous expression of biosynthetic gene clusters for the cyanobactins patellamide from metagenomic libraries of uncultured cyanobacterial symbionts associated with marine *Didemnidae* sponges (Long *et al.*, 2005; Schmidt *et al.*, 2005). Using end sequencing data and PCR primers based on conserved cyanobactin biosynthetic genes, gene clusters that encode both known and novel cyanobactins have subsequently been recovered from other marine symbiont metagenomic libraries. Donia et. al. recovered the complete gene cluster for the known cyanobactin trunkamide on a single fosmid found in an ascidian metagenomic library, and in a similar study, the same group cloned the gene cluster for minimide, which they predicted would be a novel cyanobactin (Donia *et al.*, 2008; Donia *et al.*, 2011). Upon re-engineering and optimization of this gene cluster, minimide was successfully heterologously expressed in *E. coli*. The Schmidt group took advantage of the fact that the structural diversity seen in cyanobactins largely arises from small changes in the gene encoding the ribosomally translated precursor peptide and employed a combination of orthogonal tRNAs loaded with unnatural amino acids, precursor peptide mutagenesis,

and gene shuffling to generate a library of hybrid cyanobactins using the eDNA derived trunkamide biosynthetic machinery (Tianero *et al.*, 2012).

1.10.3 Type II polyketides

A structurally diverse collection of aromatic small molecules, including many antimicrobial and anticancer agents (e.g. tetracycline and doxorubicin), arise from iterative or Type II polyketide synthases (Hertweck *et al.*, 2007). While the gene clusters that code for the biosynthesis of these molecules are very different in their details, they all contain a minimal polyketide synthase that is composed of three highly conserved genes (two ketosynthases and an acyl carrier protein). Both PCR studies and high throughput sequencing efforts have shown that eDNA samples are rich in novel minimal PKS genes (Wawrik *et al.*, 2005; Wawrik *et al.*, 2007). In an effort to identify novel bioactive metabolites, Feng et al. used degenerate primers based on conserved sequences found in minimal polyketide synthase genes to recover polyketide biosynthetic gene clusters captured in soil eDNA libraries. Minimal PKS containing eDNA clones were introduced into model cultured *Streptomyces* hosts for heterologous expression studies. Characterization of the metabolites produced in these studies identified a number of new metabolites with either previously unknown or rare carbon skeletons, one

family of which exhibits activity against antibiotic resistant bacteria (King *et al.*, 2009; Feng *et al.*, 2010; Feng *et al.*, 2011).

1.10.4 Trans-acyltransferse (trans-AT) polyketides:

A number of pharmacologically interesting polyketides isolated from uncultured marine symbionts are predicted to be biosynthesized using freestanding acyltransferases, or trans-ATs (Piel, 2010). A productive strategy for identifying gene clusters encoding these metabolites has been to probe marine metagenomic libraries for trans-AT specific sequences. Trans-AT ketosynthase (KS) domains phylogenetically cluster in accordance with the specific substrate used by the KS domain. Using primers designed to recognize KS domains that utilize acetyl-derived starter units, Fisch et. al. isolated a single amplicon present in a psymberin-producing marine sponge library that was absent in libraries from samples that did not produce the compound (Fisch *et al.*, 2009). They then used this sequence to recover the psymberin gene cluster from a *Psammocinia bulbosa* fosmid metagenomic library. Although no report of the heterologous expression of the complete psymberin gene cluster has yet appeared in the literature, the Piel group has reported the use of eDNA derived tailoring enzymes to modify trans-AT polyketides in vitro. They used an O-methyltransferase from the pederin gene clusters, which they cloned a number of years ago from a beetle symbiont metagenomic

library, to site-specifically methylate the mycalamide A resulting in the production of a hybrid compound 18-O-methylmycalamide with enhanced antitumor activity (Piel, 2002; Zimmermann *et al.*, 2009).

1.10.5 ET-743

Rath et. al recovered the biosynthetic cluster for the anticancer agent ET-743 from a metagenomic library of uncultured tunicate bacterial symbionts (Rath *et al.*, 2011) [46]. The parallels between ET-743 and other tetrahydroisoquinoline structures such as saframycin and safracin led the authors of this study to the hypothesis that ET-743 was of bacterial origin and encoded by a non-ribosomal peptide synthase similar to that seen in other tetrahydroisoquinoline gene clusters. In a cloning-independent strategy, DNA isolated directly from field collected bacterial symbionts found in a tunicate shown to produce ET-743 was 454 pyrosequenced. This data was assembled and candidate ET-743 related nonribosomal peptide synthetase (NRPS) genes were identified by their similarity to saframycin and safracin biosynthetic genes. One NRPS cluster found in these experiments was predicted to contain all of the biosynthetic genes necessary for the assembly of a tetrahydroisoquinoline core. The enzymatic activity of the predicted reductive termination domain seen in this cluster was subsequently confirmed *in vitro* using saframycin

intermediates as substrates, linking the gene cluster to ET-743 biosynthesis in a cloning-independent manner.

1.11 Sequence-based screening – Single Cell Genomics

Single cell genomics has been used to aid sequence-based screening efforts. In this strategy, single bacterial cells are isolated from complex microbial communities and then subjected to multiple displacement amplification (MDA) in order to obtain sufficient genomic DNA for sequencing (Dean *et al.*, 2002). In a study by Grindberg, et al., single cells of *Lyngbya bouillonii* were isolated from cyanobacterial filaments containing a consortium of symbiotic bacteria (Grindberg *et al.*, 2011). DNA from these cells was amplified by MDA, sequenced and confirmed to be *L. bouilonii* by 16S analysis. Biosynthetic genes of the known polyketide/nonribosomal peptide hybrid Apratoxin were identified by screening *in silico* for genes predicted to be involved in the introduction of the β-alkylation seen in this metabolite. Clones containing the apratoxin biosynthetic gene cluster were subsequently recovered from a *Lyngbya bouillonii*-symbiont metagenomic genomic library. In other examples of this approach, fluorescence-activated cell sorting (FACS) has been used to obtain single cells prior to MDA and the sequence data obtained from these cells has shed light on the biosynthetic capacities of individual sponge symbionts (Siegl & Hentschel, 2010; Siegl *et al.*, 2011).

1.12 Expanding the Potential of Functional Screening

In recent years, homology-based screening has emerged at the forefront of metagenomic library screening efforts and has resulted in a higher number of natural products found. However, there are many limitations to homology-based screening that can be addressed by functional screening. For example, one of the shortcomings of homology-based methods lies in its reliance on homology to genes of known compounds. Because of this, homology-based methods are best suited to finding derivatives of known compounds but, unlike functional screening, do not have the capacity to find completely novel natural products. In addition, homology-based methods result in the identification of multitudes of enzymes and biosynthetic pathways, but there is a high rate of attrition in going from genes to molecules. A lot of the pathways may be cryptic clusters and therefore silent or unable to be expressed in the heterologous host of choice, and recent efforts in homology-based screening are attempting to remedy this by turning on pathways using transcription factors and promoters (Bergmann *et al.*, 2007; Kallifidas *et al.*, 2012). This limitation is overcome by functional screening. Because phenotypic differences are used the readout, only clones expressing natural products are identified by this method of screening.

The low rates of success seen in functional screening of metagenomic libraries may be attributed to a number of factors. One critical roadblock has been the choice of heterologous host for the metagenomic libraries. Heterologous hosts must be able to express the genes present on the eDNA insert, and must ideally display high transformation efficiencies, fast growth and absence of pathogenicity. Heterologous expression is limited by intrinsic features of the host bacteria, such as codon usage, lack of substrate building blocks necessary for assembling a natural product, lack of recognition of foreign regulatory elements, promoters and ribosomal binding sites (Liebl *et al.*, 2014). Many studies have confirmed that heterologous expression, and therefore functional screening hits found in a metagenomic library, differ depending on the host used (Craig *et al.*, 2010). Other reasons for the low success rates seen using functional screening include small insert sizes that may not capture an entire gene cluster, small library sizes and low throughput screening methods. The potential of functional screening to find novel natural products has not been fully tapped, and the development of methods to optimize functional screening will lead to higher success rates.

1.13 Future Prospects

The reservoir of potentially useful products encoded by the earth's microbiome is still largely underexplored, as only a small minority of bacterial species has been cultured in the laboratory. Metagenomic methods have begun to provide access to both biocatalysts and secondary metabolites encoded within the genomes of these previously inaccessible bacteria. In the years to come, advances in sequencing technologies, bioinformatics prediction tools, heterologous expression methods and synthetic biology will undoubtedly increase the efficiency and utility of this general approach.

1.14 References

Allgaier M, Reddy A, Park JI*, et al.* (2010) Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community. *PloS one* **5**: e8812.

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

Banik JJ, Craig JW, Calle PY & Brady SF (2010) Tailoring enzyme-rich environmental DNA clones: a source of enzymes for generating libraries of unnatural natural products. *Journal of the American Chemical Society* **132**: 15661-15670.

Bayer TS, Widmaier DM, Temme K, Mirsky EA, Santi DV & Voigt CA (2009) Synthesis of methyl halides from biomass using engineered microbes. *Journal of the American Chemical Society* **131**: 6508-6515.

Bergmann S, Schumann J, Scherlach K, Lange C, Brakhage AA & Hertweck C (2007) Genomics-driven discovery of PKS-NRPS hybrid metabolites from Aspergillus nidulans. *Nature chemical biology* **3**: 213-217.

Bijtenhoorn P, Mayerhofer H, Muller-Dieckmann J*, et al.* (2011) A novel metagenomic short-chain dehydrogenase/reductase attenuates Pseudomonas aeruginosa biofilm formation and virulence on Caenorhabditis elegans. *PloS one* **6**: e26278.

Bongarzone S & Bolognesi ML (2011) The concept of privileged structures in rational drug design: focus on acridine and quinoline scaffolds in neurodegenerative and protozoan diseases. *Expert opinion on drug discovery* **6**: 251-268.

Brady SF & Clardy J (2000) Long-Chain N-Acyl Amino Acid Antibiotics Isolated from Heterologously Expressed Environmental DNA. *Journal of the American Chemical Society* **122**: 12903-12904.

Brady SF & Clardy J (2005) N-acyl derivatives of arginine and tryptophan isolated from environmental DNA expressed in Escherichia coli. *Organic letters* **7**: 3613-3616.

Brady SF & Clardy J (2005) Cloning and heterologous expression of isocyanide biosynthetic genes from environmental DNA. *Angewandte Chemie (International ed in English)* **44**: 7063-7065.

Brady SF, Chao CJ & Clardy J (2002) New natural product families from an environmental DNA (eDNA) gene cluster. *Journal of the American Chemical Society* **124**: 9968-9969.

Brady SF, Chao CJ, Handelsman J & Clardy J (2001) Cloning and heterologous expression of a natural product biosynthetic gene cluster from eDNA. *Organic letters* **3**: 1981-1984.

Cragg GM & Newman DJ (2013) Natural products: a continuing source of novel drug leads. *Biochimica et biophysica acta* **1830**: 3670-3695.

Craig JW, Chang FY & Brady SF (2009) Natural products from environmental DNA hosted in Ralstonia metallidurans. *ACS Chem Biol* **4**: 23-28.

Craig JW, Chang FY, Kim JH, Obiajulu SC & Brady SF (2010) Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Appl Environ Microbiol* **76**: 1633-1641.

De la Iglesia R, Valenzuela-Heredia D, Pavissich JP, Freyhoffer S, Andrade S, Correa JA & Gonzalez B (2010) Novel polymerase chain reaction primers for the specific detection of bacterial copper P-type ATPases gene sequences in environmental isolates and metagenomic DNA. *Letters in applied microbiology* **50**: 552-562.

Dean FB, Hosono S, Fang L*, et al.* (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* **99**: 5261-5266.

Demain AL & Fang A (2000) The natural functions of secondary metabolites. *Adv Biochem Eng Biotechnol* **69**: 1-39.

DeSimone RW, Currie KS, Mitchell SA, Darrow JW & Pippin DA (2004) Privileged structures: applications in drug discovery. *Combinatorial chemistry & high throughput screening* **7**: 473-494.

Donato JJ, Moe LA, Converse BJ, Smart KD, Berklein FC, McManus PS & Handelsman J (2010) Metagenomic analysis of apple orchard soil reveals antibiotic resistance genes encoding predicted bifunctional proteins. *Appl Environ Microbiol* **76**: 4396-4401.

Donia MS, Ravel J & Schmidt EW (2008) A global assembly line for cyanobactins. *Nature chemical biology* **4**: 341-343.

Donia MS, Ruffner DE, Cao S & Schmidt EW (2011) Accessing the hidden majority of marine natural products through metagenomics. *Chembiochem : a European journal of chemical biology* **12**: 1230-1236.

Feng Z, Kim JH & Brady SF (2010) Fluostatins produced by the heterologous expression of a TAR reassembled environmental DNA derived type II PKS gene cluster. *Journal of the American Chemical Society* **132**: 11902-11903.

Feng Z, Kallifidas D & Brady SF (2011) Functional analysis of environmental DNA-derived type II polyketide synthases reveals structurally diverse secondary metabolites. *Proc Natl Acad Sci U S A* **108**: 12629-12634.

Fisch KM, Gurgui C, Heycke N*, et al.* (2009) Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting. *Nature chemical biology* **5**: 494-501.

Fujita MJ, Kimura N, Yokose H & Otsuka M (2012) Heterologous production of bisucaberin using a biosynthetic gene cluster cloned from a deep sea metagenome. *Molecular bioSystems* **8**: 482-485.

Fujita MJ, Kimura N, Sakai A, Ichikawa Y, Hanyu T & Otsuka M (2011) Cloning and heterologous expression of the vibrioferrin biosynthetic gene cluster from a marine metagenomic library. *Bioscience, biotechnology, and biochemistry* **75**: 2283-2287.

Gabor EM, Alkema WB & Janssen DB (2004) Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol* **6**: 879-886.

Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, Rondon MR, Clardy J, Goodman RM & Handelsman J (2002) Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol* **68**: 4301-4306.

Grindberg RV, Ishoey T, Brinza D*, et al.* (2011) Single cell genome amplification accelerates identification of the apratoxin biosynthetic pathway from a complex microbial assemblage. *PloS one* **6**: e18565.

Handelsman J, Rondon MR, Brady SF, Clardy J & Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5**: R245-249.

Harvey AL, Edrada-Ebel R & Quinn RJ (2015) The re-emergence of natural products for drug discovery in the genomics era. *Nature reviews Drug discovery* **14**: 111-129.

Hata T, Hoshi T, Kanamori K, Matsumae A, Sano Y, Shima T & Sugawara R (1956) Mitomycin, a new antibiotic from Streptomyces. I. *The Journal of antibiotics* **9**: 141-146.

Hertweck C, Luzhetskyy A, Rebets Y & Bechthold A (2007) Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork. *Natural product reports* **24**: 162-190.

Hess M, Sczyrba A, Egan R*, et al.* (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science (New York, NY)* **331**: 463-467.

Hu XP, Heath C, Taylor MP, Tuffin M & Cowan D (2012) A novel, extremely alkaliphilic and cold-active esterase from Antarctic desert soil. *Extremophiles : life under extreme conditions* **16**: 79-86.

Huang Y, Lai X, He X, Cao L, Zeng Z, Zhang J & Zhou S (2009) Characterization of a deep-sea sediment metagenomic clone that produces water-soluble melanin in Escherichia coli. *Marine biotechnology (New York, NY)* **11**: 124-131.

Kallifidas D, Kang HS & Brady SF (2012) Tetarimycin A, an MRSA-active antibiotic identified through induced expression of environmental DNA gene clusters. *Journal of the American Chemical Society* **134**: 19552-19555.

Kellner H, Luis P, Portetelle D & Vandenbol M (2011) Screening of a soil metatranscriptomic library by functional complementation of Saccharomyces cerevisiae mutants. *Microbiological research* **166**: 360-368.

King RW, Bauer JD & Brady SF (2009) An environmental DNA-derived type II polyketide biosynthetic pathway encodes the biosynthesis of the pentacyclic polyketide erdacin. *Angewandte Chemie (International ed in English)* **48**: 6257-6261.

Kirst HA (2010) The spinosyn family of insecticides: realizing the potential of natural products research. *The Journal of antibiotics* **63**: 101-111.

Laux-Biehlmann A, Mouheiche J, Veriepe J & Goumon Y (2013) Endogenous morphine and its metabolites in mammals: history, synthesis, localization and perspectives. *Neuroscience* **233**: 95-117.

Levine DP (2006) Vancomycin: a history. *Clin Infect Dis* **42 Suppl 1**: S5-12.

Lewis K (2013) Platforms for antibiotic discovery. *Nature reviews Drug discovery* **12**: 371-387.

Lewis K, Epstein S, D'Onofrio A & Ling LL (2010) Uncultured microorganisms as a source of secondary metabolites. *The Journal of antibiotics* **63**: 468-476.

Li JW & Vederas JC (2009) Drug discovery and natural products: end of an era or an endless frontier? *Science (New York, NY)* **325**: 161-165.

Liebl W, Angelov A, Juergensen J*, et al.* (2014) Alternative hosts for functional (meta)genome analysis. *Applied microbiology and biotechnology* **98**: 8099-8109.

Liu Y, Lei Y, Zhang X, Gao Y, Xiao Y & Peng H (2012) Identification and phylogenetic characterization of a new subfamily of alpha-amylase enzymes from marine microorganisms. *Marine biotechnology (New York, NY)* **14**: 253-260.

Long PF, Dunlap WC, Battershill CN & Jaspars M (2005) Shotgun cloning and heterologous expression of the patellamide gene cluster as a strategy to achieving sustained metabolite production. *Chembiochem : a European journal of chemical biology* **6**: 1760-1765.

MacNeil IA, Tiong CL, Minor C*, et al.* (2001) Expression and isolation of antimicrobial small molecules from soil DNA libraries. *J Mol Microbiol Biotechnol* **3**: 301-308.

Max B, Salgado JM, Rodriguez N, Cortes S, Converti A & Dominguez JM (2010) Biotechnological production of citric acid. *Brazilian journal of microbiology : [publication of the Brazilian Society for Microbiology]* **41**: 862-875.

McGarvey KM, Queitsch K & Fields S (2012) Wide variation in antibiotic resistance proteins identified by functional metagenomic screening of a soil DNA library. *Appl Environ Microbiol* **78**: 1708-1714.

Miller WL (1979) Use of recombinant DNA technology for the production of polypeptides. *Advances in experimental medicine and biology* **118**: 153-174.

Neveu J, Regeard C & DuBow MS (2011) Isolation and characterization of two serine proteases from metagenomic libraries of the Gobi and Death Valley deserts. *Applied microbiology and biotechnology* **91**: 635-644.

Newman DJ & Cragg GM (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod* **75**: 311-335.

Newmark P (1980) Engineered E. coli produce interferon. *Nature* **283**: 323.

Niehaus F, Gabor E, Wieland S, Siegert P, Maurer KH & Eck J (2011) Enzymes for the laundry industries: tapping the vast metagenomic pool of alkaline proteases. *Microbial biotechnology* **4**: 767-776.

Pang H, Zhang P, Duan CJ, Mo XC, Tang JL & Feng JX (2009) Identification of cellulase genes from the metagenomes of compost soils and functional characterization of one novel endoglucanase. *Current microbiology* **58**: 404-408.

Parsley LC, Consuegra EJ, Kakirde KS, Land AM, Harper WF, Jr. & Liles MR (2010) Identification of diverse antimicrobial resistance determinants carried on bacterial, plasmid, or viral metagenomes from an activated sludge microbial assemblage. *Appl Environ Microbiol* **76**: 3753-3757.

Piel J (2002) A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. *Proc Natl Acad Sci U S A* **99**: 14002-14007.

Piel J (2010) Biosynthesis of polyketides by trans-AT polyketide synthases. *Natural product reports* **27**: 996-1047.

Ramachandran S, Singh SK, Larroche C, Soccol CR & Pandey A (2007) Oil cakes and their biotechnological applications--a review. *Bioresource technology* **98**: 2000-2009.

Rappe MS & Giovannoni SJ (2003) The uncultured microbial majority. *Annual review of microbiology* **57**: 369-394.

Rath CM, Janto B, Earl J*, et al.* (2011) Meta-omic characterization of the marine invertebrate microbial consortium that produces the chemotherapeutic natural product ET-743. *ACS Chem Biol* **6**: 1244-1256.

Schipper C, Hornung C, Bijtenhoorn P, Quitschau M, Grond S & Streit WR (2009) Metagenome-derived clones encoding two novel lactonase family proteins involved in biofilm inhibition in Pseudomonas aeruginosa. *Appl Environ Microbiol* **75**: 224-233.

Schmidt EW, Nelson JT, Rasko DA, Sudek S, Eisen JA, Haygood MG & Ravel J (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**: 7315-7320.

Siegl A & Hentschel U (2010) PKS and NRPS gene clusters from microbial symbiont cells of marine sponges by whole genome amplification. *Environmental microbiology reports* **2**: 507-513.

Siegl A, Kamke J, Hochmuth T, Piel J, Richter M, Liang C, Dandekar T & Hentschel U (2011) Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges. *Isme J* **5**: 61-70.

Simon C, Herath J, Rockstroh S & Daniel R (2009) Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. *Appl Environ Microbiol* **75**: 2964-2968.

Sivonen K, Leikoski N, Fewer DP & Jokela J (2010) Cyanobactins-ribosomal cyclic peptides produced by cyanobacteria. *Applied microbiology and biotechnology* **86**: 1213-1225.

Thomson AW & Woo J (1989) Immunosuppressive properties of FK-506 and rapamycin. *Lancet* **2**: 443-444.

Tianero MD, Donia MS, Young TS, Schultz PG & Schmidt EW (2012) Ribosomal route to small-molecule diversity. *Journal of the American Chemical Society* **134**: 418-425.

Tohme R, Darwiche N & Gali-Muhtasib H (2011) A journey under the sea: the quest for marine anti-cancer alkaloids. *Molecules (Basel, Switzerland)* **16**: 9665-9696.

Torsvik V, Goksoyr J & Daae FL (1990) High diversity in DNA of soil bacteria. *Appl Environ Microbiol* **56**: 782-787.

Uchiyama T & Miyazaki K (2010) Product-induced gene expression, a product-responsive reporter assay used to screen metagenomic libraries for enzyme-encoding genes. *Appl Environ Microbiol* **76**: 7029-7035.

Uchiyama T, Abe T, Ikemura T & Watanabe K (2005) Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat Biotechnol* **23**: 88-93.

Walton RJ, Sherif IT, Noy GA & Alberti KG (1979) Improved metabolic profiles in insulin-treated diabetic patients given an alpha-glucosidehydrolase inhibitor. *British medical journal* **1**: 220-221.

Wang GY, Graziani E, Waters B, Pan W, Li X, McDermott J, Meurer G, Saxena G, Andersen RJ & Davies J (2000) Novel natural products from soil DNA libraries in a streptomycete host. *Organic letters* **2**: 2401-2404.

Wang K, Li G, Yu SQ, Zhang CT & Liu YH (2010) A novel metagenome-derived beta-galactosidase: gene cloning, overexpression, purification and characterization. *Applied microbiology and biotechnology* **88**: 155-165.

Wang Q, Wu H, Wang A, Du P, Pei X, Li H, Yin X, Huang L & Xiong X (2010) Prospecting metagenomic enzyme subfamily genes for DNA family shuffling by a novel PCR-based approach. *J Biol Chem* **285**: 41509-41516.

Wawrik B, Kerkhof L, Zylstra GJ & Kukor JJ (2005) Identification of unique type II polyketide synthase genes in soil. *Appl Environ Microbiol* **71**: 2232-2238.

Wawrik B, Kutliev D, Abdivasievna UA, Kukor JJ, Zylstra GJ & Kerkhof L (2007) Biogeography of actinomycete communities and type II polyketide synthase genes in soils collected in New Jersey and Central Asia. *Appl Environ Microbiol* **73**: 2982-2989.

Wilkinson DE, Jeanicke T & Cowan DA (2002) Efficient molecular cloning of environmental DNA from geothermal

sediments. *Biotechnology Letters* **24**: 155-161.

Williams DH, Stone MJ, Hauck PR & Rahman SK (1989) Why are secondary metabolites (natural products) biosynthesized? *J Nat Prod* **52**: 1189-1208.

Yao J, Fan XJ, Lu Y & Liu YH (2011) Isolation and characterization of a novel tannase from a metagenomic library. *Journal of agricultural and food chemistry* **59**: 3812-3818.

Zaprasis A, Liu YJ, Liu SJ, Drake HL & Horn MA (2010) Abundance of novel and diverse tfdA-like genes, encoding putative phenoxyalkanoic acid herbicide-degrading dioxygenases, in soil. *Appl Environ Microbiol* **76**: 119-128.

Zimmermann K, Engeser M, Blunt JW, Munro MH & Piel J (2009) Pederin-type pathways of uncultivated bacterial symbionts: analysis of o-

methyltransferases and generation of a biosynthetic hybrid. *Journal of the American Chemical Society* **131**: 2780-2781.

CHAPTER 2: Antibacterial enzymes from the functional screening of metagenomic libraries hosted in *Ralstonia metallidurans*

2.1 Chapter Summary

Phenotype-based screening of bacterial metagenomic libraries provides an avenue for the discovery of novel genes, enzymes and metabolites that have a variety of potential clinical and industrial uses. Here we report the identification of a functionally diverse collection of antibacterially active enzymes from the phenotypic screening of 700,000 cosmid clones prepared from Arizona soil DNA and hosted in *Ralstonia metallidurans*. Environmental DNA clones surrounded by zones of growth inhibition in a bacterial overlay assay were found, through bioinformatics and functional analyses, to encode enzymes with predicted peptidase, lipase and glycolytic activities conferring antibiosis. The antibacterial activities observed in our *R. metallidurans*-based assay could not be replicated with the same clones in screens using *Escherichia coli* as a heterologous host, suggesting that the large-scale screening of metagenomic libraries for antibiosis using phylogenetically diverse hosts should be a productive strategy for identifying enzymes with functionally diverse antibacterial activities.

2.2 Introduction

Analyses of bacterial culture broths have traditionally been a route for the discovery of novel small molecules and enzymes (Demain & Sanchez, 2009; Trincone, 2011). While productive, these studies are limited by our inability to culture the vast majority of bacteria from the environment. The metabolomes and proteomes of a more representative sample of environmental bacteria can be accessed using functional metagenomic approaches that involve the extraction of DNA directly from environmental samples (environmental DNA, eDNA), the cloning of this DNA into model cultured bacteria and, finally, the phenotypic screening of these clones in diverse assays (Handelsman *et al.*, 1998; Iqbal *et al.*, 2012).

To date, the majority of metagenomic library screens targeting antibiosis have relied on top agar overlay assays on *Escherichia coli*-based libraries to identify antibacterially active small molecules (Figure 7A). Large-scale functional screens using *E. coli*-based metagenomic libraries have seldom reported the discovery of heterologously expressed antibacterially active enzymes, despite the fact that cultured bacteria have been a prolific source of antimicrobial proteins (Veiga-Crespo *et al.*, 2007; Thallinger *et al.*, 2013). Phage endolysins, in particular, have recently garnered attention for their potential roles in enzyme-based antibiotic therapies (Fischetti, 2010; Thallinger *et al.*, 2013).

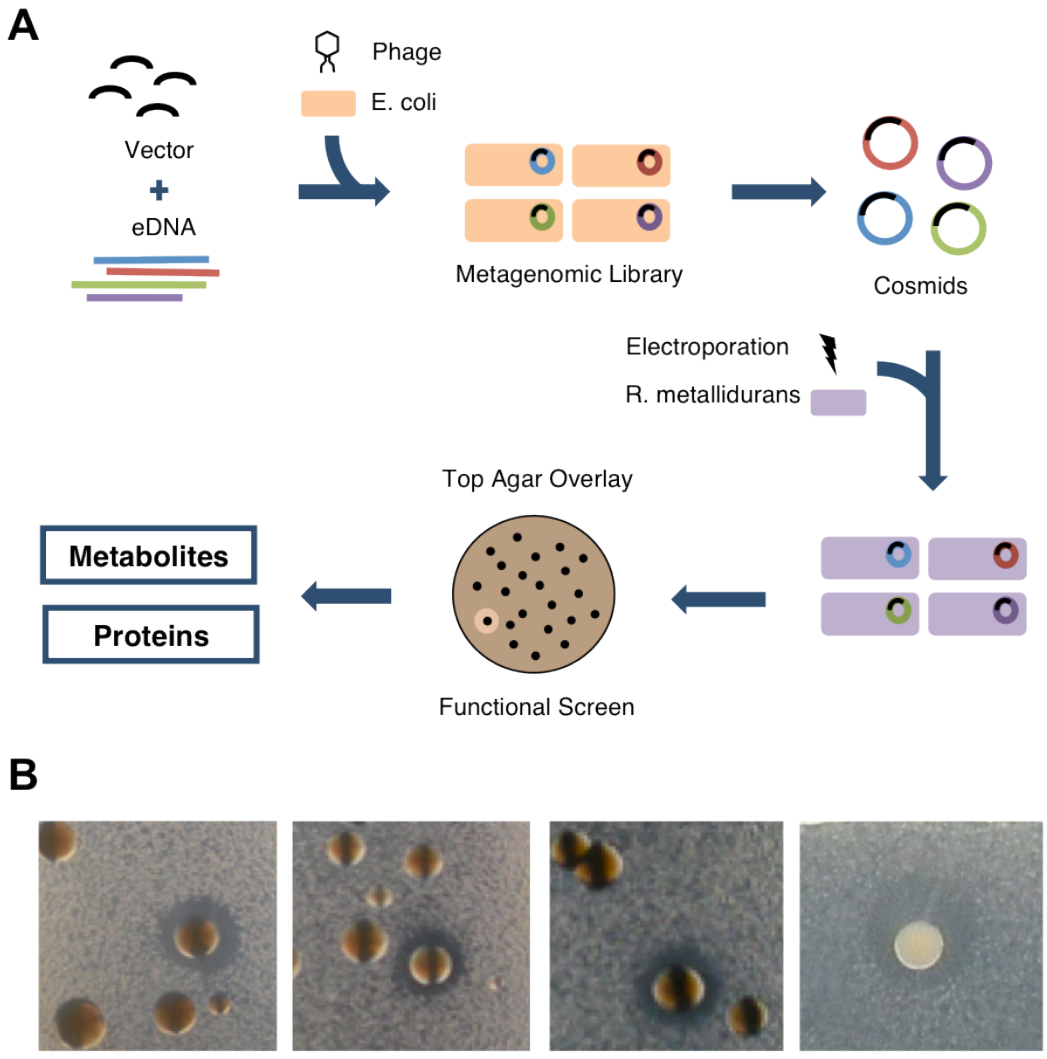


Figure 7. (A) Overview of metagenomic library construction and screening methodology. In this study, eDNA extracted from soil samples was ligated to a shuttle cosmid vector and introduced into *E. coli* by phage transduction. DNA prepared from these libraries was electroporated into R. metallidurans and screened for antibiosis using the top agar overlay method. **(B)** Representative antibacterially active hits obtained by screening soil metagenomic libraries hosted in R. metallidurans using the top agar overlay method. R. metallidurans-based libraries were arrayed onto LB-agar plates and overlaid with a layer of top agar containing a B. subtilis assay strain; zones of growth inhibition around individual colonies indicate heterologous expression of the antibacterial phenotype.

While most metagenomic functional screening has used *E. coli* as a host, the utility of *E. coli* as a heterologous host in functional metagenomic screens is likely to be limited due to its limited heterologous expression capacity (Gabor *et al.*, 2004). Since the success of functional metagenomic screening is contingent upon the ability of the library host to heterologously express genes found on foreign eDNA, the identification of transcriptionally diverse model hosts that can express these foreign genes will likely be critical to the overall success of metagenomic screening strategies. We hypothesized that by changing the host used in phenotypic metagenomic library screens it might be possible to begin to identify the diverse antibacterial enzymes that are undoubtedly encoded within soil microbiomes.

Here we show that enzymes conferring antibiosis can be found in *R. metallidurans* hosted soil DNA libraries. *R. metallidurans* is a gram negative beta proteobacteria that we have explored as an alternative host for small molecule-based functional metagenomic studies because of its previously described heterologous expression capabilities, its genetic tractability and the ease with which it can be grown in the laboratory (Craig *et al.*, 2010). Based on recent interest in lytic enzyme-based antibiotic therapies, the large-scale screening of environmental DNA libraries hosted in *R. metallidurans* may be a productive strategy for identifying enzymes

with diverse activities for potential use as novel therapeutics (Fischetti, 2010; Thallinger *et al.*, 2013).

2.3 Results & Discussion

For this study, DNA extracted directly from soil collected in the Sonoran Desert (Arizona, USA) was used to construct a cosmid library in a broad host cosmid shuttle vector. This metagenomic library comprised of 700,000 unique clones and was predicted to contain ~22.5 Gb of eDNA. The library was originally constructed in *E. coli* using lambda phage packaging and transfection and then cosmid DNA from this *E. coli*-based parent library was electroporated into *Ralstonia metallidurans* [Fig 6A]. We selected *R. metallidurans*, a soil dwelling beta proteobacteria, as the library host for this study because it had previously shown improved heterologous expression capabilities in functional screening studies compared to *E. coli* (Craig *et al.*, 2010).

Previous *R. metallidurans*-based functional metagenomic studies have focused on the identification of antibacterially active clones producing organic extractable small molecules (Craig *et al.*, 2009; Craig *et al.*, 2010). In light of recent reports highlighting the successful use of bacterial and phage enzymes as potential antibacterial therapies, we sought to explore whether antibacterially hits identified in *R. metallidurans*-based eDNA library antibiosis screens that did not produce small

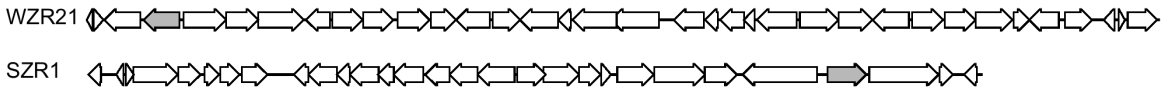
molecules might be a source of eDNA encoded antibacterially active enzymes. Top agar overlays were carried out on the 700,000 membered *R. metallidurans* soil eDNA library to identify clones exhibiting antibiosis activities against a *B. subtilis* assay strain [Fig 7B]. Of the 19 clones we identified as hits in our primary antibiosis assay, cosmid DNA isolated from six clones (SZR1, SZR5, WZR9, WZR11, WZR18 and WZR21) showed the ability to confer antibacterial activity to *R. metallidurans* upon retransformation. In small scale fermentation studies none of these clones showed the presence of clone specific small molecule in culture broth extracts and these clones were therefore examined both bioinformatically and functionally for the ability to encode for antibacterial active enzymes.

Cosmid DNA from each reproducible antibacterial hit was de novo sequenced and annotated using the online SoftBerry software package to identify open reading frames. Putative gene functions were assigned by alignment to the BLAST and PFAM databases. Based on predicted gene functions, only two clones contained genes encoding obvious antibacterial enzymes. Clones SZR1 and WZR21 encoded enzymes with high similarity to previously well-characterized antibacterially active enzymes targeting peptidogylcan bonds [Fig 8A, B]. The cell wall lytic homolog from the first clone, WZR21, is most closely related (54% identity) to *N*-acetylmuramoyl-L-alanine-amidases that cleave the amide bond connecting *N*-acetylmuramic acid to the cross-linked peptides present in

the bacterial cell wall. The cell wall lytic homolog from the second clone, SZR1, is most closely related (31% identity) to lytic transglycosylases that cleave the glycosyl bond between N-acetyl glucosamine and N-acetyl muramic acid sugars in the bacterial cell wall [Fig 8C].

PFAM-trimmed functional domains present in the predicted cell wall lytic enzymes from SZR1 and WZR21 were phylogenetically compared to soluble lytic transglycosylase and amidase domains from functionally characterized antibacterial phage endolysins [Fig 8D]. The predicted cell wall lytic domain from clone SZR1 displayed similarity to the lytic transglycosylase domain of the *Pseudomonas aeruginosa* Bacteriophage B3 endolysin predicted to participate in host cell lysis (Braid *et al.*, 2004). This same analysis indicated high similarity between the predicted N-acetyl-muramic-acid-L-alanine amidase domain from clone WZR21 and the antibacterially active phage amidase domain plyM13 found in one of the only other reported metagenomic screens to identify enzymes with cell wall lytic activity. In this study, animal fecal metagenomes were screened using a two-step functional assay in which phage genes were first identified by their proximity to hemolytic phage holins using a blood agar-based assay and then recombinantly expressed and tested for lytic activity against a heat-killed bacterial assay strain (Schmitz *et al.*, 2010).

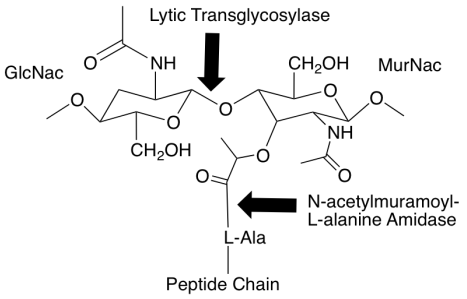
A



B

Clone	Annotation	Species	% Identity	Accession No.
WZR21	N-acetylmuramoyl-L-alanine amidase	*Desulfovibrio sp. U5L*	54	ZP_10078644.1
SZR1	Lytic Transglycosylase	*Clostridium sp. CAG:1013*	31	WP_016405493.1

C



D

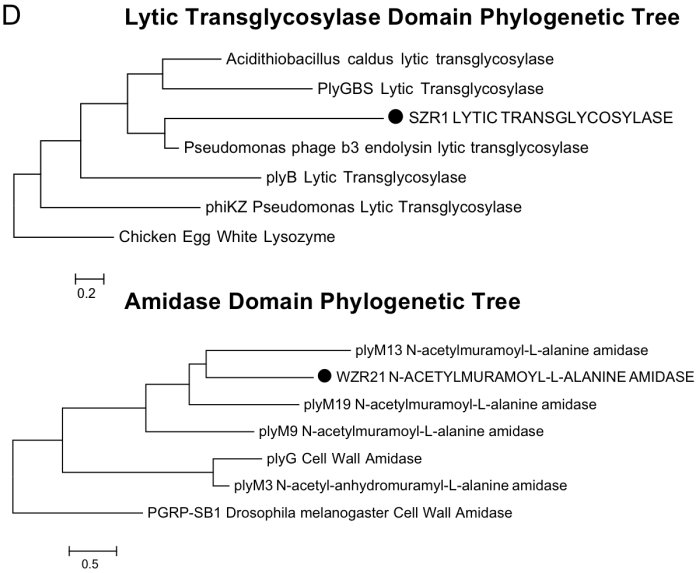
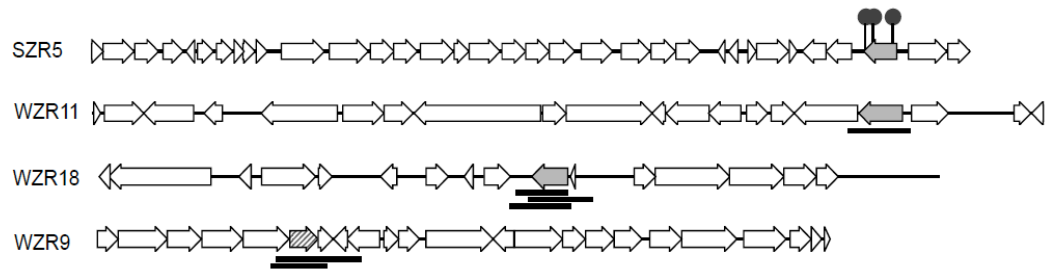


Figure 8. (A) Sequence maps from antibacterially active clones containing putative cell wall degrading enzymes (grey). (B) Annotation of putative cell wall degrading enzymes associated with antibacterial activity. Species and % identity columns describe the top BLAST hit to the predicted cell wall lytic enzyme encoding gene. (C) Repeat structure of the peptidoglycan layer of the cell wall of gram-positive bacteria. Arrows indicate the points of action of the cell wall degrading enzymes found in this study. (D) ClustalW-derived phylogenetic trees comparing functional domains of a representative set of cell wall lytic transglycosylases and amidases with the predicted enzymatic domains found in our phenotypic screen (highlighted in capital letters with black circles). Protein sequences were trimmed to PFAM families PF01464 (SLT Transglycosylase) or PF01520 (Amidase) domains.

For the four antibacterially active clones (WZR9, WZR11, WZR18 and WZR21) where no open reading frames were found to bear strong relationship to unequivocal antibacterially active enzymes, the genetic elements required for antibacterial activity were identified through either subcloning (WZR9, WZR11 and WZR18) or transposon mutagenesis (WZR21) experiments, followed by top agar overlay screening to identify either antibacterial active subclones or inactive transposon mutants.

Antibacterial activity assays run on subclone libraries made using sheared cosmid DNA from clones WZR9, WZR11 and WZR18 led to the identification of a single antibacterially active enzyme on each clone. In these studies antibacterially active subclones were recovered and sequenced until the overlapping region on the recovered clones was reduced to the point of a single open reading frame (ORF) [Fig 9A]. The antibacterial associated ORF found on clones WZR11 and WZR18 are predicted to encode for proteases, while the antibacterial ORF from WZR9 was predicted to encode for a lipase [Figure 9A and B]. Antibacterial assays run on transposon mutants of SZR5 yielded three unique mutants that lacked the antibacterial phenotype. The transposon insertion in all three mutants was found in an ORF that encodes for a predicted protease.

A



B

Clone	Enzyme	MEROPS Family	Species	% Identity	Accession No.
SZR5	Metallopeptidase	M20F	*Myxococcus xanthus*	54	YP_634146.1
WZR18	Metallopeptidase	M14B	*Leptospira santarosai*	40	WP_004466397.1
WZR11	Serine Protease	S8A	*Ferroglobus placidus*	38	YP_003434691.1
WZR9	Lipolytic	-	Uncultured bacterium	50	AAX37296.1

Figure 8. (A) Sequence maps from clones containing antibacterially active proteases (solid grey ORFs) and lipolytic enzymes (hashed grey ORF). Antibiosis was confirmed either by the insertion of transposons (red circles) leading to loss of activity, or by the expression of the antibiosis phenotype by random shotgun subclones (black lines). (B) Annotation of putative proteases and lipolytic enzymes associated with antibacterial activity. Species and % identity columns describe the closest homolog from the MEROPS database for proteases and from the BLAST database for the lipolytic enzyme.

While proteases and lipases have been found in previous metagenomic screens using *E. coli* as a heterologous host, they have rarely been identified using direct screens for antibiosis. For example, previous attempts to clone proteases from environmental samples have focused on identifying industrially relevant enzymes for applications such as laundry detergent using milk agar plate assays (Kennedy *et al.*, 2011; Niehaus *et al.*, 2011). Similarly, industrially relevant lipases and esterases

have been isolated in previous metagenomic studies using lipase-directed screening methods, including degradation of tributyrin or related techniques (Kennedy *et al.*, 2011; Reyes-Duarte *et al.*, 2012). There is only one case in which lipolytic enzymes conferring antibiosis were identified in an *E. coli*-based metagenomic study (Yung *et al.*, 2011).

The predicted proteases from clones SZR5, WZR11 and WZR18 were aligned to the MEROPS database – a curated collection of proteases organized into families of related enzymes by sequence similarities (Rawlings *et al.*, 2012). This analysis indicated that the SZR5 and WZR18 proteases belong to the M20F and M14B subfamilies of metallopeptidases, respectively and the WZR11 protease belongs to the S8A subfamily of serine proteases [Fig 9B]. In a BLAST search against GenBank, the predicted lipase from clone WZR9 shows highest similarity (50% identity) to an esterase found in a soil functional metagenomic screen for ester-hydrolysis activity (Kim *et al.*, 2006). The three enzymes found in the only *E. coli*-based metagenomic screen reporting antibacterial lipases show only low sequence identity (10% identity to CcAb1, 9% to CcAb2, 23 % to UaAb1), to the lipolytic enzyme in clone WZR9 found in our *R. metallidurans*-based screen.

It is somewhat surprising that enzymes with antibacterial activity have not been reported more frequently in metagenomic screens given that a number of bacteria (e.g *Achromobacter lyticus* and *Myxococcus*

xanthus) are known to produce bacteriolytic proteases, and genomes of some sequenced bacteria for example predatory bacteria such as *Bdellovibrio bacteriovirus* are rich in proteolytic enzymes (Sudo & Dworkin, 1972; Li *et al.*, 1998; Rendulic *et al.*, 2004). We hypothesized that this disparity in the finding of antibacterial enzymes in metagenomic screens may be a function of choice of heterologous host. To assess whether or not the putative enzymes identified in our *R. metallidurans*-based screens could have been discovered using the same techniques but with *E. coli* as the heterologous host, we performed the same top agar overlay assay on *E. coli* transformed with each of the six antibacterially active eDNA clones we identified. In these assays, none of the *E. coli*-based clones displayed the antibacterial phenotype observed in our *R. metallidurans*-based screens. This experiment provides additional support to the idea that bacterial hosts differ in their abilities to heterologously express foreign genetic material, and underscores the importance of diverse hosts in functional screenings of metagenomic libraries.

The six clones found in our *R. metallidurans*-based metagenomic screen for antibacterial activity contain a lipase, proteases and cell wall lytic enzymes that are predicted to hydrolyze the three key linking chemistries (e.g. ester, amide and glycosidic bonds) present in the bacterial cell (Figure 10). In light of the attention that lytic enzymes have recently garnered as potential therapeutics, obtaining a diverse

assortment of novel antibacterial enzymes for development as lead agents has become increasingly relevant. The diversity of antibacterial enzymes found in this study suggests that the phenotypic screening of soil metagenomes using various heterologous hosts, including *R. metallidurans*, may prove useful for identifying therapeutically relevant antibacterial enzymes in future large-scale screens of environmental samples.

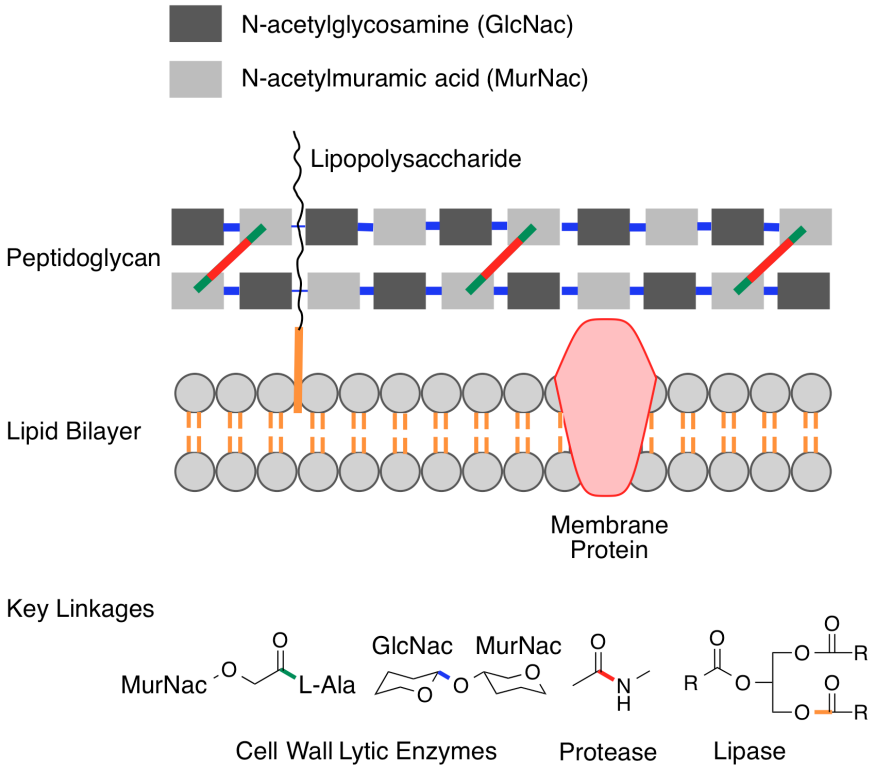


Figure 10. The enzymes found in our metagenomic screen target the key "linking chemistries" (amide, ester and glycolytic bonds) found in bacterial cell walls.

2.4 Materials and Methods

2.4.1 eDNA Library Construction

DNA extracted from soil collected in the Sonoran Desert of Arizona (USA) was used to create a 700,000-membered cosmid-based metagenomic library in *E. coli* (Brady, 2007). To obtain crude eDNA, 250 g of soil was passed through a 1/8 inch sieve to remove rocks and large debris, and then incubated in lysis buffer (100 mM Tris-HCl, 100 mM NaEDTA, 1.5 M NaCl, 1% (w/v) cetyl trimethyl ammonium bromide, 2% (w/v) SDS, pH 8.0) (1:1 wt:vol) at 70^{o}C for 2 hours (Zhou *et al.*, 1996). Heat lysed samples were centrifuged (4,000 g, 10 min) to remove soil particulates. Crude eDNA was precipitated from the supernatant by the addition of 0.7 volume isopropanol and collected by centrifugation (4,000 g, 10 min, 4^{o}C). The pellet was washed with 70% ethanol and the eDNA resuspended in minimum volume of TE buffer. eDNA was separated from the remaining soil material by ethidium bromide-free agarose (1%) gel electrophoresis (1 hour at 100V, 5 hours at 20V). High molecular weight eDNA (>25 kb) was extracted from the gel by electroelution, concentrated by isopropanol precipitation and blunt-end repaired (End-It, Epicentre Biotechnologies). Blunt-ended eDNA was ligated with either the previously reported broad-host-range cosmid vector pJWC1 or pJSS, a pJWC1 derivative with a DNA linker containing a ScaI cloning site [TGGCCTGTCATGAGCAGGATC] replacing the *sacB* gene (Craig *et al.*,

2009). Cosmids vectors were prepared for ligation by digestion with ScaI and dephosphorylation with calf-intestinal alkaline phosphatase. Ligation reactions were packaged with lambda phage packaging extracts (MaxPlax - Lambda Packaging Extracts, Epicentre) and transfected into *E. coli* EPI 300 (TransforMax, Epicentre Biotechnologies) grown to OD_{600} 1.0. Transfection reactions were selected overnight in LB broth containing 20 µg/mL tetracycline (37^{o}C with shaking). In total, 5 x 10^5 and 2 x 10^5 unique cosmid clones were constructed using vectors pJWC1 and pJSS respectively, constituting 22.5 GB eDNA. The library was stored as 15% glycerol stocks.

To transform libraries into *Ralstonia metallidurans*, cosmid DNA was miniprepped (Qiagen) from overnight cultures inoculated from library glycerol stocks and pooled together the next day in equivalent volumes. 1 – 2 µg of DNA was transformed by electroporation (0.8 kV/1.0 mm cuvette) into 80 µl aliquots of electrocompetent *R. metallidurans* CH34 cells prepared as previously described (Taghavi *et al.*, 1994). Following electroporation, cells were diluted in 1 mL SOC medium and incubated for 3 hours (30^{o}C with shaking) before plating onto LB-tetracycline plates (20 mg/mL). The resulting *R. metallidurans*-based library comprising two times coverage of the original eDNA cosmid library was scraped from the selection plates after two days at 30°C and archived as glycerol stocks.

2.4.2 Functional Screening of Libraries in *R. metallidurans*

The *R. metallidurans*-based library was screened for clones with antibacterial activity in a top agar overlay assay against *Bacillus subtilis*. For this assay, the library was diluted directly from glycerol stocks and plated onto 150 mm LB-tetracycline (20 mg/mL) plates at a density of 1,000 – 1,500 clones per plate. Colonies were allowed to mature at 30°C overnight and then incubated at room temperature for 4 – 5 days to allow for heterologous expression. A thin layer of 0.7% top agar (10 – 12 mL) was overlaid onto the plates. The top agar was inoculated with a 1:200 dilution of tetracycline-resistant *B. subtilis* 1E9 (Bacillus Genome Stock Center, Ohio) grown to OD_{600} 0.5. Plates were incubated at 30°C overnight. To cleanly obtain the naturally kanamycin-resistant *R. metallidurans* clones without residual *B. subtilis* assay strain contamination, colonies forming zones of growth inhibition in the *B. subtilis* lawn were picked from the assay plates and struck for single colonies on LB plates containing tetracycline 20 µg/mL and kanamycin 30 µg/mL. DNA was miniprepped from overnight cultures of single colonies, retransformed into *R. metallidurans* and patched onto LB-tetracycline (20 µg/mL) plates. These plates were re-assayed using top agar overlays; clones that showed the antibiosis phenotype were considered true hits and archived as 15% glycerol stocks in *R. metallidurans* and *E. coli*.

2.4.3 Sequencing and Bioinformatics

High quality cosmid DNA was obtained from antibacterially active hits by miniprepping overnight *E. coli* EPI 300 cultures induced with CopyControl Induction Solution (Epicentre Biotechnologies). DNA was sequenced using 454 GS-FLX Titanium pyrosequencing technologies (MSKCC Genomics Core Laboratory) and assembled on GS De Novo Assembler software (Roche) (GenBank Accession numbers KF835381-KF835386 for SZR1, SZR5, WZR9, WZR11, WZR18 and WZR21, respectively). Sequences were annotated using the online tool SoftBerry to predict open reading frames, and alignments to BLAST and PFAM databases were used to predict gene function (Altschul *et al.*, 1990; Solovyev & Salamov, 2011; Punta *et al.*, 2012). For phylogenetic analysis, predicted antibacterial protein found in clones SZR1, WZR21 and WZR9 were trimmed based on their alignment to PFAM families PF01464 (SLT Transglycosylase), PF01520 (Amidase) or PF07859 (Alpha/beta hydrolase), respectively. Phylogenetic trees were constructed from ClustalW alignments using the MEGA5 program with the Neighbor-Joining method and 1000 bootstrap replications (Tamura *et al.*, 2011).

2.4.4 Subclone Library Construction

Cosmid DNA (2 µg) was sheared to 3 kb using Blue miniTubes in the Covaris S220 Focused-ultrasonicator. Sheared DNA was blunt-end

repaired, ligated into ScaI-digested dephosphorylated pJWC1, and transformed into electrocompetent *E. coli* EPI 300 cells using the same methods described in the eDNA library construction section. Subclone libraries were transformed into electrocompetent *R. metallidurans* and assayed in a top agar overlay assay for antibacterial activity. Clones displaying zones of growth inhibition were struck onto LB-tetracycline-kanamycin plates. Single colony cultures were miniprepped and transformed into electrocompetent *E. coli* EPI 300 to obtain sufficient DNA for sequencing with vector specific primers.

2.4.5 Transposon Mutagenesis

E. coli-based random transposon mutagenesis libraries were created using the HyperMu transposon system (HyperMu <Kan-1>, Epicentre). DNA prepared from overnight cultures of the transposon mutagenesis library was transformed into electrocompetent *R. metallidurans* cells and selected on LB-tetracycline plates. The resulting *R. metallidurans-based* transposon mutant libraries were assayed using top agar overlays to identify clones that failed to display a zone of growth inhibition. DNA isolated from LB-tetracycline overnight cultures of these clones was transformed into *E. coli* EPI 300 cells, and DNA prepped from overnight cultures of transformants was Sanger sequenced using HyperMu specific primers.

2.4.6 Top Agar Overlays in *E. coli*

Cosmid DNA from antibacterial clones identified in our original functional screen in *R. metallidurans* were transformed into electrocompetent *E. coli* EPI300 and assayed for the ability to confer the same antibiosis phenotype in *E. coli*. Each antibacterially active clone transformed into *E. coli*, and negative control (empty pJWC1 vector), were struck onto LB-tetracycline (20 µg/mL) plates to obtain single colonies. Plates were incubated at 30^{o}C overnight, followed by 3 days at room temperature and then assayed using top agar overlays as described above.

2.5 References

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

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

Braid MD, Silhavy JL, Kitts CL, Cano RJ & Howe MM (2004) Complete genomic sequence of bacteriophage B3, a Mu-like phage of Pseudomonas aeruginosa. *J Bacteriol* **186**: 6560-6574.

Craig JW, Chang FY & Brady SF (2009) Natural products from environmental DNA hosted in Ralstonia metallidurans. *ACS Chem Biol* **4**: 23-28.

Craig JW, Chang FY, Kim JH, Obiajulu SC & Brady SF (2010) Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Appl Environ Microbiol* **76**: 1633-1641.

Demain AL & Sanchez S (2009) Microbial drug discovery: 80 years of progress. *The Journal of antibiotics* **62**: 5-16.

Fischetti VA (2010) Bacteriophage endolysins: a novel anti-infective to control Gram-positive pathogens. *International journal of medical microbiology : IJMM* **300**: 357-362.

Gabor EM, Alkema WB & Janssen DB (2004) Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol* **6**: 879-886.

Handelsman J, Rondon MR, Brady SF, Clardy J & Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5**: R245-249.

Iqbal HA, Feng Z & Brady SF (2012) Biocatalysts and small molecule products from metagenomic studies. *Current opinion in chemical biology* **16**: 109-116.

Kennedy J, O'Leary ND, Kiran GS, Morrissey JP, O'Gara F, Selvin J & Dobson AD (2011) Functional metagenomic strategies for the discovery of novel enzymes and biosurfactants with biotechnological applications from marine ecosystems. *Journal of applied microbiology* **111**: 787-799.

Kim YJ, Choi GS, Kim SB, Yoon GS, Kim YS & Ryu YW (2006) Screening and characterization of a novel esterase from a metagenomic library. *Protein expression and purification* **45**: 315-323.

Li S, Norioka S & Sakiyama F (1998) Bacteriolytic activity and specificity of Achromobacter beta-lytic protease. *J Biochem* **124**: 332-339.

Niehaus F, Gabor E, Wieland S, Siegert P, Maurer KH & Eck J (2011) Enzymes for the laundry industries: tapping the vast metagenomic pool of alkaline proteases. *Microbial biotechnology* **4**: 767-776.

Punta M, Coggill PC, Eberhardt RY*, et al.* (2012) The Pfam protein families database. *Nucleic Acids Res* **40**: D290-301.

Rawlings ND, Barrett AJ & Bateman A (2012) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* **40**: D343-350.

Rendulic S, Jagtap P, Rosinus A*, et al.* (2004) A predator unmasked: life cycle of Bdellovibrio bacteriovorus from a genomic perspective. *Science (New York, NY)* **303**: 689-692.

Reyes-Duarte D, Ferrer M & Garcia-Arellano H (2012) Functional-based screening methods for lipases, esterases, and phospholipases in metagenomic libraries. *Methods Mol Biol* **861**: 101-113.

Schmitz JE, Schuch R & Fischetti VA (2010) Identifying active phage lysins through functional viral metagenomics. *Appl Environ Microbiol* **76**: 7181-7187.

Solovyev V & Salamov A (2011) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. *Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies,*(Li RW, ed.) 61-78. Nova Science Publishers.

Sudo S & Dworkin M (1972) Bacteriolytic enzymes produced by Myxococcus xanthus. *J Bacteriol* **110**: 236-245.

Taghavi S, van der Lelie D & Mergeay M (1994) Electroporation of Alcaligenes eutrophus with (mega) plasmids and genomic DNA fragments. *Appl Environ Microbiol* **60**: 3585-3591.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M & Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum

likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution* **28**: 2731-2739.

Thallinger B, Prasetyo EN, Nyanhongo GS & Guebitz GM (2013) Antimicrobial enzymes: an emerging strategy to fight microbes and microbial biofilms. *Biotechnology journal* **8**: 97-109.

Trincone A (2011) Marine biocatalysts: enzymatic features and applications. *Marine drugs* **9**: 478-499.

Veiga-Crespo P, Ageitos JM, Poza M & Villa TG (2007) Enzybiotics: a look to the future, recalling the past. *Journal of pharmaceutical sciences* **96**: 1917-1924.

Yung PY, Burke C, Lewis M, Kjelleberg S & Thomas T (2011) Novel antibacterial proteins from the microbial communities associated with the sponge Cymbastela concentrica and the green alga Ulva australis. *Appl Environ Microbiol* **77**: 1512-1515.

Zhou J, Bruns MA & Tiedje JM (1996) DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62**: 316-322.

CHAPTER 3: Functional Screening of Soil Metagenomic Libraries using *Streptomyces* as a Heterologous Host

3.1 Chapter Summary

Functional screening of metagenomic libraries offers a route to study natural products made by uncultured bacteria. *Streptomyces* is postulated as an excellent heterologous host for screening for small molecules, but past attempts at building and screening libraries in *Streptomyces* have been unfruitful. For functional screening to be successful, the library host must be able to heterologously express biosynthetic genes present in library clones to produce natural products imparting phenotypic differences. *S. albus* was identified from a diverse collection of *Streptomyces* as a strain that is good at heterologous expression. To build a metagenomic library in *S. albus*, a high quality library was first built in *E. coli*. An *E. coli* – *Streptomyces* shuttle vector stable for large DNA inserts was constructed for this library. Soil environmental DNA (eDNA) was cloned into the vector and library DNA was gel purified to remove empty vector contamination, resulting in a high quality library. Mating conditions were varied to optimize transformation efficiency. These methods were used to build a 1.5 million clone library in *S. albus*. Preliminary screens of this library for changes in colony color led to the identification of 12 reconfirmed hits with 8 carotenoid pathways, 1 type III polyketide pathway

and 3 clones with unknown biosynthetic pathways. Hits will be further analyzed to identify the structure of clone-specific molecules being produced.

3.2 Introduction

Bacterial are a rich source of natural products with diverse chemistries and activities (Clardy *et al.*, 2006, Iqbal *et al.*, 2012). It is estimated that 99% of bacteria in environmental samples are uncultured, and therefore cannot be surveyed by traditional culture-dependent methods for natural product discovery (Schloss & Handelsman, 2006). Metagenomics offers a route to study uncultured bacteria by extracting the total DNA present in an environmental sample (environmental DNA, eDNA), cloning eDNA into a vector to make a metagenomic library, and expressing this library in a model cultured bacterial host (Rondon *et al.*, 2000). Metagenomic libraries can be screened in various ways for desired biocatalysts or biosynthetic clusters, allowing access to the reservoir of natural products produced by uncultured bacteria (Leis *et al.*, 2013).

Functional or activity-based screening represents one approach for identifying clones of interest in metagenomic libraries. Clones with desired activities are identified by phenotypic changes imparted by the molecules encoded on and produced by the metagenomic library clone. Common phenotypic characteristics used to identify clones of interest include

change in color or morphology of the host bacteria. Other traits that have been used as output in functional screens include antibiotic activity against an assay strain of bacteria, catalytic activity on enzymatic substrate, resistance to antibiotics or toxic compounds, and auxotrophic complementation (Chistoserdovai, 2010).

Whereas functional screening was initially the most commonly used screening method for finding biosynthetic clusters from metagenomic libraries, homology and sequence-based screening methods have overtaken functional screening in terms of popularity and success rates (Iqbal *et al.*, 2012). Because homology and sequence-based screening methods rely on similarity to known biosynthetic pathways, functional screening remains the only method of screening that has the potential to uncover truly novel compounds. However, the success of the method has been limited by three major bottlenecks: 1. Library hosts that are able to heterologously express biosynthetic genes, 2. Small DNA insert libraries that do not capture entire biosynthetic clusters for making secondary metabolites on single clones, 3. Intrinsically low hit rates because of the small percentage (approximately 5%) of bacterial genomes devoted to secondary metabolism (Ruiz *et al.*, 2010, Calteau *et al.*, 2014, Ziemert *et al.*, 2014). These limitations can be overcome by strategic choice of library heterologous host based on characteristics and activities being screened

for, developing methods for large DNA insert cloning, and selection-based enrichment of libraries before screening.

In this work, we describe methods for overcoming problems associated with heterologous host choice for finding small molecule biosynthetic clusters in soil metagenomic libraries. Since functional screening is contingent upon the identification of phenotypic differences between library clones, the heterologous host must be able to generate the molecules imparting phenotype that are encoded by metagenomic eDNA. The success of functional screening therefore is dependent upon the heterologous expression capabilities of the library host. The heterologous expression of foreign DNA varies widely in bacteria, depending on a myriad of factors such as the absence of intrinsic biosynthetic machinery and substrates in the host, differences in codon usage and GC content, lack of recognition of regulatory elements such as promoters and ribosome binding sites, improper folding of proteins, etc (Terpe, 2006).

E. coli is the most commonly used heterologous host for metagenomics libraries due to ease of handling and genetic manipulation, as well as high transformation efficiencies for building and storing metagenomic libraries. However, Gabor et. al. used bioinformatic analysis and *in silico* modeling of the factors involved in heterologous expression to estimate that *E. coli* is able to express only 40% of genes from foreign

DNA (Gabor *et al.*, 2004). Of critical importance is the fact that this number was found to vary widely, with *E. coli* only being able to express 7% of genes from Actinobacteria, a phylum that has historically been a prolific source of therapeutically relevant small molecules natural products and antibiotics (Watve *et al.*, 2001, Manivasagan *et al.*, 2014).

One solution to the limitation posed by heterologous expression in functional screening has been to expand the set of bacteria used as metagenomic library hosts. Craig et. al. used *Ralstonia metallidurans* – a soil dwelling bacteria with the ability to thrive in harsh, toxic environments– as a host and were able to find novel small molecules using functional screening methods for color and antibiotic activity (Craig *et al.*, 2009). In a follow up study using diverse soil-dwelling proteobacteria, namely *Pseudomonas putida*, *Burkholderia graminis*, *Agrobacterium tumefaciens* and *Caulobacter vibrioides*, as heterologous hosts, Craig et. al. showed that eDNA clones impart varying phenotypes to different bacterial hosts including absence of phenotype, and therefore choice of host was a factor in finding library hits (Craig *et al.*, 2010). In a similar study, Martinez et. al. showed that heterologous expression capabilities differed between *E. coli*, *Streptomyces lividans* and *P. putida* (Martinez *et al.*, 2004). These results underscore the importance of the choice of metagenomic library host to the success of finding hits in functional screening.

Streptomyces are gram positive, aerobic, soil-dwelling bacteria of the phylum *Actinobacteria* that have long been considered a promising candidate as a heterologous host for finding small molecule biosynthetic clusters from metagenomic libraries. *Streptomyces* have a long history of yielding therapeutically relevant secondary metabolites, and as many as 80% of antibiotics in clinical use are sourced from the genus (Procopio *et al.*, 2012, Gomez-Escribano & Bibb, 2014). In addition to antibiotics, *Streptomyces* also natively produce small molecules with antifungal, anticancer and immunosuppressant activities (Figure 11). These natural products span a large number of structurally diverse families of compounds including polyketides, nonribosomal peptides, ribosomal peptides and terpenoids (Lucas *et al.*, 2013).

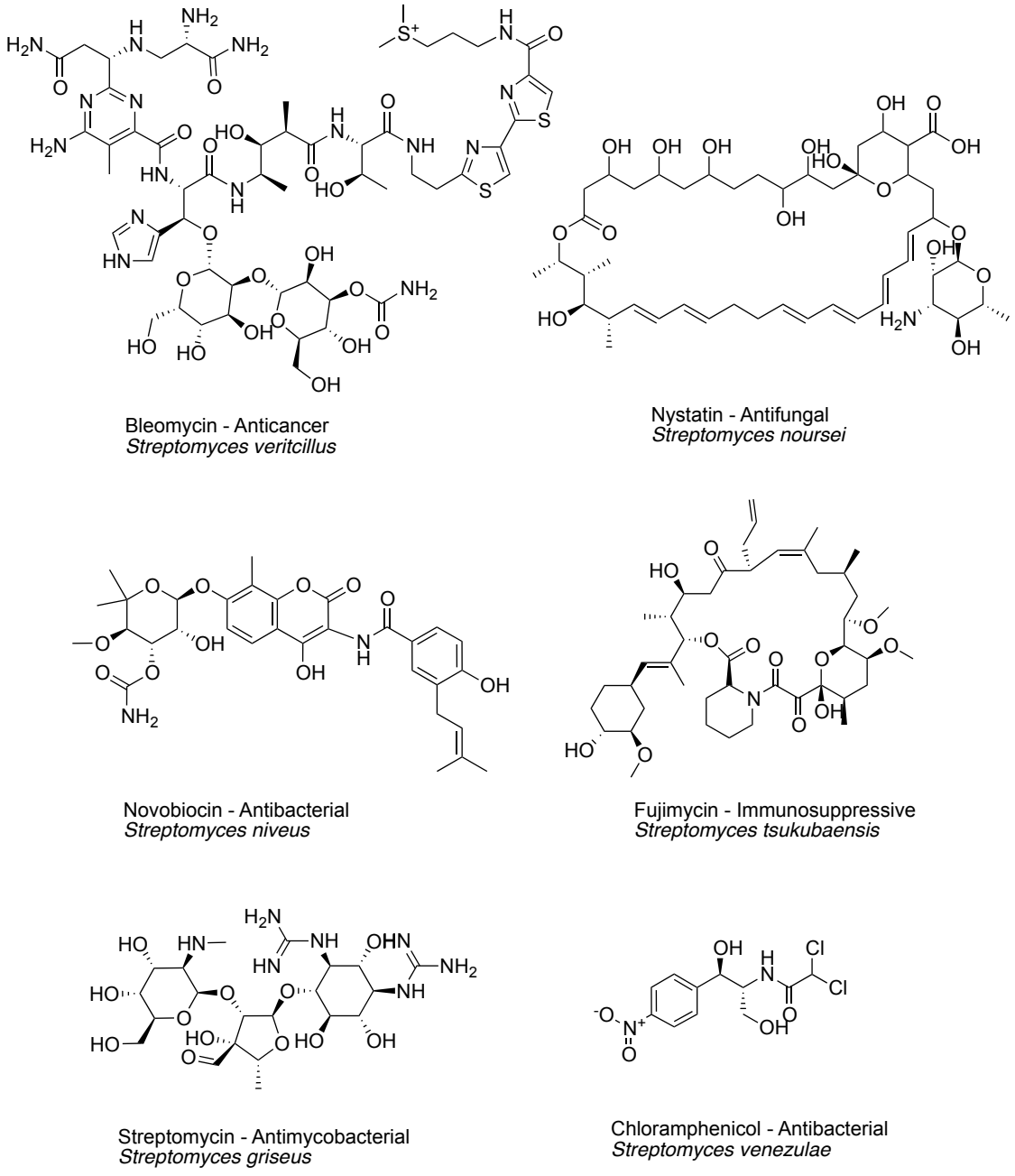


Figure 10. Natural Products from *Streptomyces* in Clinical Use.

There are numerous reasons why *Streptomyces* are postulated to be excellent heterologous hosts for metagenomic libraries. The fact that so many strains natively produce diverse small molecules indicates that *Streptomyces* have the biosynthetic machinery to assemble the core of secondary metabolites from various families of small molecules, as well as the tailoring enzymes for adding auxiliary features to compounds. In addition, various strains of *Streptomyces* have been used to heterologously express a large number of compounds from individual genomic and metagenomic library clones, for example the peptidyl nucleoside Gougerotin, the polyketide antibiotic Tetracycline, and the indolotryptoline Borregomycin (Binnie *et al.*, 1989, Gomez-Escribano & Bibb, 2012, Chang & Brady, 2013, Niu *et al.*, 2013, Gomez-Escribano & Bibb, 2014). *Streptomyces* are non-pathogenic, relatively easy to grow, and various tools have been developed to make the genus more genetically tractable (Wexler & Johnston, 2010). Finally, *Streptomyces* are especially promising as a heterologous host for the functional screening of soil metagenomic libraries, as Actinobacteria comprise a large percentage of the bacteria present in soil environmental samples (Fierer *et al.*, 2007, Reddy *et al.*, 2012).

Conversely, there are many challenges associated with the use of *Streptomyces* as a metagenomic library heterologous host. Compared to other bacteria such as *E. coli* and *R. metallidurans*, *Streptomyces* grow

slowly and require rich media that is prone to contamination (Kieser *et al.*, 2000). *Streptomyces* exhibit complex growth profiles and display a lot of variation in phenotype and production of secondary metabolites depending on factors such as media, growth conditions and aeration. In addition, transformation of DNA into *Streptomyces* is cumbersome and inefficient, which is especially a problem for building large metagenomic libraries. Plasmid stability in *Streptomyces* has also been a problem: large DNA inserts (>30 kb) are unstable in most high copy number plasmids, while low copy number plasmids display poor yields of secondary metabolites as well as library DNA, and are therefore ineffective as metagenomic library hosts (Fong *et al.*, 2007).

There have been a small number of attempts at functional screening of metagenomic libraries using *Streptomyces* as a heterologous host (Table 2) (Wang *et al.*, 2000, Courtois *et al.*, 2003, Martinez *et al.*, 2004, Martinez *et al.*, 2005, Lussier *et al.*, 2011, McMahon *et al.*, 2012). *Streptomyces lividans* has been used as the strain of choice in all reported libraries to date for its ease of handling. Compared to other strains of *Streptomyces*, *S. lividans* grows relatively fast and shows high conjugation efficiency, which is conducive to the transformation of metagenomic libraries (Mazodier *et al.*, 1989). *S. lividans* is frequently chosen over its close relative *Streptomyces coelicolor* because it displays lower levels of endogenous proteolytic and methylation-dependant restriction activity,

allowing it to accept foreign methlyated DNA and heterologously express proteins more easily than *S. coelicolor* (Lewis *et al.*, 2010, Niu *et al.*, 2013). In addition, *S. lividans* is postulated to have the metabolic potential for small molecule production, as it natively produces the small molecules actinorhodin, undecylprodigiosin and calcium-dependent antibiotic (CDA) and has 25 biosynthetic clusters in its genome (Ruckert *et al.*, 2015). Because of these reasons, *S. lividans* is commonly chosen for the heterologous expression of individual biosynthetic clusters, although anecdotally as well as in our experience, it has not been able to heterologously express biosynthetic pathways that have been introduced into the bacteria. This is also reflected in the low number of natural products that have been found by screening metagenomic libraries hosted in *S. lividans*.

Table 2. Summary of Metagenomic Library Screening Efforts in *Streptomyces*

Library Host	Clones	Insert Size (kb)	Screening Method	NP Found	Reference
S. lividans	1020		HPLC	Terragine A – E	(Wang *et al.*, 2000)
S. lividans	800	50	HPLC	2 Fatty dienic alcohols	(Courtois *et al.*, 2003)
S. lividans	23600	35	Hemolytic Assay Color screen	Phospholipase, peptidase Melanin pigments, Transcription factor	(McMahon *et al.*, 2012)
S. lividans	2000		Lipolytic Activity Proteolytic Activity		(Lussier *et al.*, 2011)
S. lividans	800	40	Color Screen	Transcription factor	(Martinez *et al.*, 2005)
S. lividans	13000	11 – 85			(Martinez *et al.*, 2004)

Despite the potential ascribed to functional screening of metagenomic libraries hosted in *Streptomyces*, previous attempts have been largely unsuccessful both in terms of building large metagenomic libraries as well as finding novel small molecules. The libraries screened to date have been small, with the largest reported library comprising 826 MB of DNA in 23,600 clones, compared to the 300 GB libraries routinely made using *E. coli* as a heterologous host (Kang & Brady, 2013). The size of libraries has been limited by poor transformation efficiencies in

Streptomyces. The process of protoplast preparation for electroporation is cumbersome and protoplast yields are low (Kieser *et al.*, 2000). The most commonly used method is conjugal mating from *E. coli*, but this method displays lower transformation efficiencies and is frequently plagued by contamination. Attempts to counter low transformation efficiencies have included the individual transformation of full-length library clones into *Streptomyces*, but even using highthroughput methods involving robots, resulting libraries have been small, on the order of 10,000 clones (Martinez *et al.*, 2004). Another problem with many of the libraries reported has been empty vector contamination, wherein self-ligated or uncut vector present in library ligation reactions transforms easily because of its small size and therefore propagates faster, and can end up constituting the bulk of a library.

In addition to library size, screening assays have also proved a limitation to the success of metagenomic screening in *Streptomyces*. Many libraries have been screened for the production of small molecules by growing individual clones and comparing HPLC traces of organic extracts, a labor and time-intensive lowthroughput process that is not conducive to screening large libraries. In spite of the inefficiencies, this screening method has led to the discovery of novel secondary metabolites: two fatty dienic alcohols, and the closely related compounds terragine A – E (Wang *et al.*, 2000, Courtois *et al.*, 2003). Other screening

assays for color or enzymatic activities have largely been unsuccessful in terms of finding novel small molecules and enzymes.

In the present work, I describe the construction and preliminary screening of a 1.5 million member cosmid clone library in *Streptomyces albus*, comprising 60 GB of genetic information. This is the first library reported in *S. albus* almost 100 X larger than any other library in *Streptomyces* to date. I solved the problems plaguing *Streptomyces* library construction and screening by: 1. Identifying a strain of *Streptomyces* that displayed high transformation frequency conducive to library transformation and high heterologous expression capabilities to allow for efficient functional screening, 2. Building a high quality *E. coli* library for transformation into the chosen strain of *Streptomyces*. An *E. coli* - *Streptomyces* shuttle vector displaying DNA insert stability was constructed for this purpose and the resulting library was gel-purified to remove empty vector contamination 3. Optimizing the *E. coli* – *Streptomyces* conjugation protocol to allow the transformation of the largest reported library into *Streptomyces* (Figure 12). Preliminary screens of this library for colored clones resulted in 12 reconfirmed hits, comprising 8 carotenoid pathways, 1 Type III polyketide pathway and 3 clones with unknown biosynthetic pathways. This library construction and screening method represents an avenue of overcoming the bottleneck of

heterologous expression in the functional screening of metagenomic libraries for small molecules.

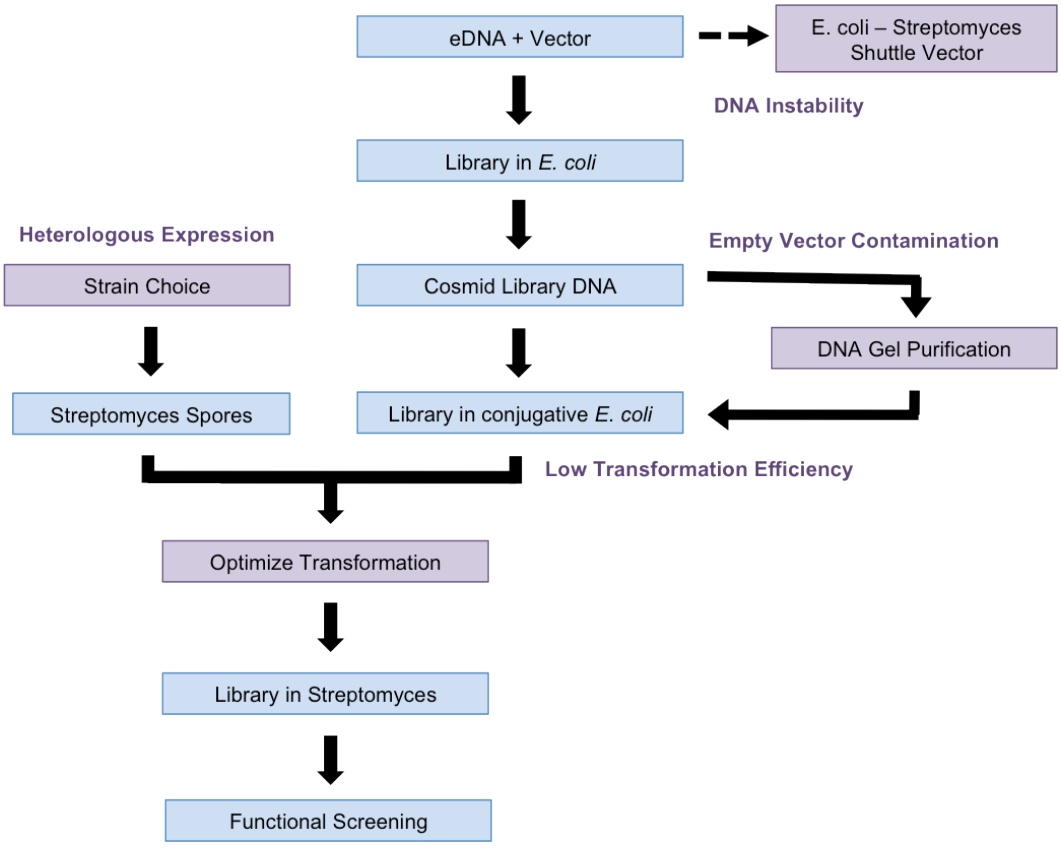


Figure 12. Flowchart depicting the problems and solutions involved in building metagenomic libraries in *Streptomyces*. The prevailing problems with hosting libraries in *Streptomyces* (purple text) include heterologous expression, DNA instability, empty vector contamination of libraries, and low transformation efficiencies. These were solved (purple boxes) by choosing a highly transforming strain of *Streptomyces* displaying high heterologous expression capabilities, building a high quality library in *E. coli* using a stable shuttle vector and gel-purifying the DNA to remove empty vector, and optimizing the transformation efficiency of library mating reactions.

3.3 Results and Discussion

3.3.1. *Streptomyces* Strain Choice

In choosing which strain of *Streptomyces* to focus efforts towards development as a metagenomic library host, the following two characteristics are critical: the transformation efficiency displayed by the strain, and its intrinsic heterologous expression capabilities. High rates of transformation of eDNA clones into the strain are needed for the effective transfer of large metagenomic libraries, and the ability to heterologously expression genes encoded on the eDNA clone is necessary for finding natural products by functional screening.

Different strains of *Streptomyces* have been known to display varying transformation efficiencies and have been postulated to differ in their heterologous expression capabilities (Mazodier *et al.*, 1989). Whereas the transformation efficiency of strains can easily measured by counting the number of exconjugants obtained from a mating experiment with a known number of *Streptomyces* spores, heterologous expression capabilities have only been anecdotally accounted and, to the best of our knowledge, have never been quantified or compared between different strains of *Streptomyces*.

To identify which strain of *Streptomyces* to use as a metagenomic library host for finding small molecule biosynthetic clusters, we first measured the transformation frequencies of a collection of *Streptomyces* strains, and then tested strains displaying high transformation frequencies for their ability to express genes from biosynthetic pathways and produce small molecules. To assess strains for both these qualities, we compiled a test library of 97 full-length metagenomic library cosmid clones harboring biosynthetic clusters, hosted in the *E. coli* strain S17.1 capable of conjugal mating into *Streptomyces*. This test library contained both uncharacterized and known type II polyketides pathways, and most importantly, contained full-length cosmids with no empty vector contamination to accurately reflect the transformation efficiency expected of large eDNA clones. Type II polyketides were chosen as the small molecules for this test set because a large number of the compounds in this family, as well as intermediates and shunt products, have conjugated ring systems and are therefore colored (Staunton & Weissman, 2001). The ability to visually identify colonies producing colored compounds provided an easy first-pass readout of heterologous expression.

3.3.2 Testing *Streptomyces* strains for Transformation Frequency

Phylogenetic analysis of 16S rRNA sequences from the *Streptomyces* genus was used to guide the selection of a diverse collection of 39 strains

of *Streptomyces*, chosen from different branches of the resulting phylogenetic tree (Figure 13). Strains were chosen to include commonly used laboratory strains, strains anecdotally known to be good at heterologous expression, strains producing small molecule natural products, and sequenced strains identified as harboring a large number of cryptic biosynthetic clusters.

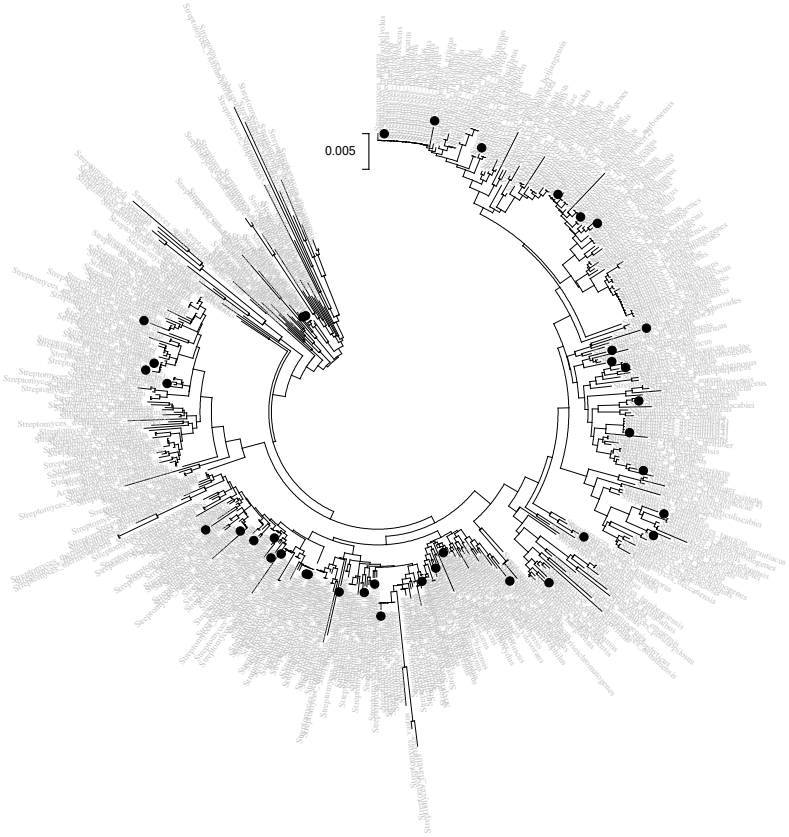


Figure 13. Phylogenetic tree of 16S rRNA sequences of the genus *Streptomyces*. Sequences were downloaded from the Ribosome Database Project and aligned using MUSCLE. MEGA was used to construct the tree. Black dots mark the 39 strains of *Streptomyces* tested for transformation efficiency and heterologous expression capabilities.

To test the *Streptomyces* strain collection for mating efficiency, standard transformation reactions for the conjugation of 10^8 spores of all strains with the *E. coli*-based type II polyketide test library were set up. Transformation frequencies were calculated based on the number of exconjugants seen (Table 3). Plates from 14 strains showing transformation frequency greater than or equal to 10^{-6} (> 150 exconjugants, 1.5 x fold coverage of the test library) were scraped and used to make individual spore stocks of the type II polyketide library in highly transforming strains of *Streptomyces*.

Table 3. Strains of *Streptomyces* and corresponding transformation frequencies

Streptomyces Strain	Transformation Frequency	Streptomyces Strain	Transformation Frequency
S. rimosus	10^{-4}	S. capoamus	10^{-7}
S. coelicolor	10^{-4}	S. nodosus	10^{-7}
S. bikiniensis	10^{-4}	S. bottropenis	10^{-7}
S. vinaceusdrappus	10^{-4}	S. roseosporus	10^{-7}
S. griseoviridus	10^{-4}	S. albidoflavus	10^{-7}
S. lividans	10^{-4}	S. ghanaensis	10^{-7}
S. tendae	10^{-4}	S. avermitilis	10^{-7}
S. malachiticus	10^{-5}	S. alboniger	10^{-7}
S. coerulescens	10^{-5}	S. achromogenes	10^{-7}
S. coeruleorubidus	10^{-5}	S. globisporous	10^{-7}
S. albus	10^{-6}	S. melanosporalaciens	10^{-7}
S. violatus	10^{-6}	S. almquisti	10^{-7}
S. antibioticus	10^{-6}	S. fumanus	10^{-7}
S. viridochromogenes	10^{-6}	S. spheroides	10^{-8}
S. peucetius	10^{-7}	S. longispuroflavus	10^{-8}
S. pseudovenezulae	10^{-7}	S. fradiaei	10^{-8}
S. venezulae	10^{-7}	S. hygroscopicus	0
S. tauricus	10^{-7}	S. platensis	0
S. spectabilis	10^{-7}	S. lincolnensis	0
S. griseus	10^{-7}		

3.3.3 Testing *Streptomyces* strains for Heterologous Expression

Library spore stocks of the type II polyketide test library in highly transforming strains of *Streptomyces* individually plated out onto R5A rich media to test for heterologous expression. As negative control, spore stock from vector pWEB436 mated into in the highly transforming strains of *Streptomyces* was simultaneously plated on R5A media. Colonies were allowed to mature for 14 days, after which plates were visually inspected

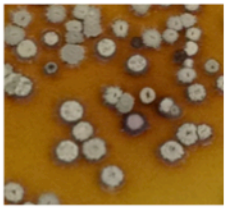
for changes in color and compared to negative control plates to account for natural variation in colony color. Four strains of *Streptomyces* (*S. albus*, *S. lividans*, *S. coerulescens* and *Streptomyces viridochromogenes*) showed colored hits (Figure 14A), indicating that genes from type II polyketide pathways were being heterologously expressed. *S. lividans*, *S. coerulescens* and *S. viridochromogenes* showed colonies with the same phenotype – a bright pink-violet color – with the hit rates of colored colonies corresponding to 1 clone in the library being expressed. In the case of *S. albus*, 3 phenotypes were seen: pink-violet and blue, the hit rates of which corresponded to 1 clone each, and a dark brown phenotype which showed a high hit rate indicating multiple clones (Table 4). To de-replicate the brown phenotypes seen, 30 brown colonies were grown in R5A liquid media for 14 days and extracted using neutral and acidified Ethyl Acetate. HPLC-MS analysis of these extracts revealed the presence of 4 different chemotypes with peaks corresponding to clone-specific small molecules (Figure 14 B).

Table 4. Heterologous expression of Type II Polyketide clones by highly transforming strains of *Streptomyces*.

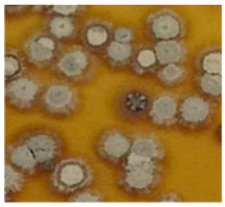
***Streptomyces* Strain**	**Hit Rate**	**Clones**	**Phenotype**
S. albus	0.67% 1% 12%	1 1 4	Pink-Violet Blue Brown
S. antibioticus			
S. bikiniensis			
S. coelicolor			
S. coeruleorubidus			
S. coerulescens	2%	1	Pink-Violet
S. griseoviridus			
S. lividans	2%	1	Pink-Violet
S. malachiticus			
S. rimosus			
S. tendae			
S. vinaceusdrappus			
S. violatus			
S. viridochromogenes	1.33%	1	Pink-Violet

With a total of 6 different clones being expressed, *S. albus* displayed the best heterologous expression capability out of all the highly transforming strains tested. It was a clear choice for *Streptomyces* strain host for functional screening of metagenomic libraries, and for further optimization of mating protocols.

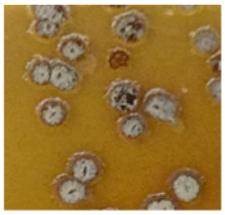
A



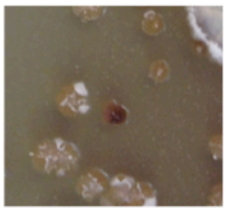
S. albus
Pink-Violet



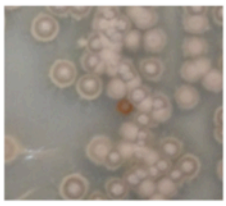
S. albus
Blue



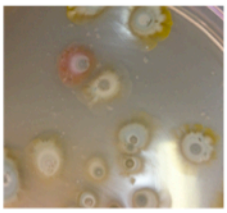
S. albus
Brown



S. coerulescens
Pink-Violet



S. lividans
Pink-Violet



S. viridochromogenes
Pink-Violet

B

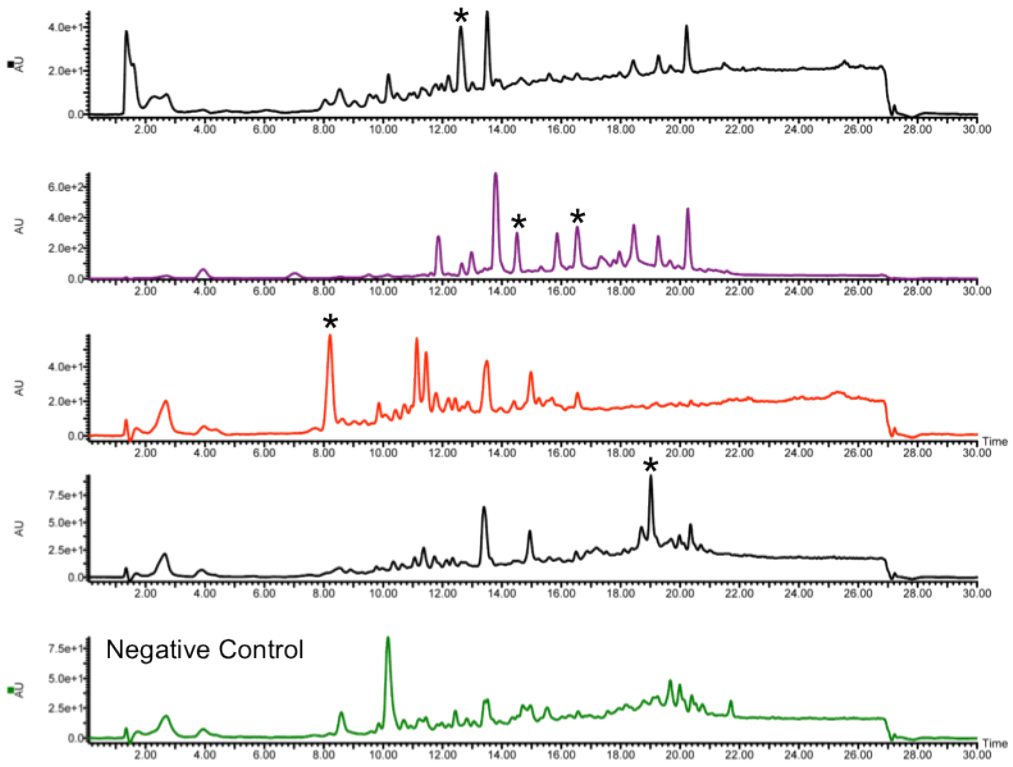


Figure 14. Heterologous expression capabilities of highly transforming strains of *Streptomyces.* (A) Colored phenotypes found in screen for heterologous expression capabilities of strains of *Streptomyces*. A collection of clones containing Type II polyketide pathways was mated into highly transforming strains of *Streptomyces*. 3 phenotypes were seen: pink-violet colored clones in *S. albus*, *S. coerulescens*, *S. lividans* and *S. viridochromogenes*, and blue and brown colored clones in S. albus. (B) Dereplication of brown colored clones seen in *S. albus*. Neutral ethyl acetate extractions of brown colored clones revealed the presence of 4 different chemotypes corresponding to different clones. Clone-specific speaks are marked with an asterisk.

3.3.4 Construction of a High Quality Metagenomic Library in *E. coli*

To successfully transform a large metagenomic library into *Streptomyces*, it was imperative to begin with a high quality library in *E. coli*. Features of this library included the use of a new *E. coli – Streptomyces* shuttle vector displaying large insert stability, and the absence of empty vector contamination.

Vector choice is critical to metagenomic library DNA stability in *Streptomyces*. Replicative vectors with high copy numbers in *Streptomyces* display instability for large DNA inserts (>30 kb) (Fong *et al.*, 2007). *Streptomyces* integrative vectors on the other hand, integrate into the genome and are low copy number and therefore stable for large DNA inserts, but are hard to build libraries with in *E. coli* because of low plasmid DNA yields. To solve the problem of DNA instability, the *E. coli* -

Streptomyces shuttle vector pWEB436 was constructed by ligating a fragment containing the *E. coli* origin of replication, Ampicillin selective marker, and cos site from the replicative cosmid cloning vector pWEB to the segment of the integrative *Streptomyces* vector pOJ436 containing the *Streptomyces* origin of replication, ΦC31 attP integration site, integrase and Apramycin selective marker. The resulting vector pWEB436 was replicative in *E. coli* and integrative in *Streptomyces*, allowing for the efficient building of a library in *E. coli* and stable propagation in *Streptomyces* (Figure 15).

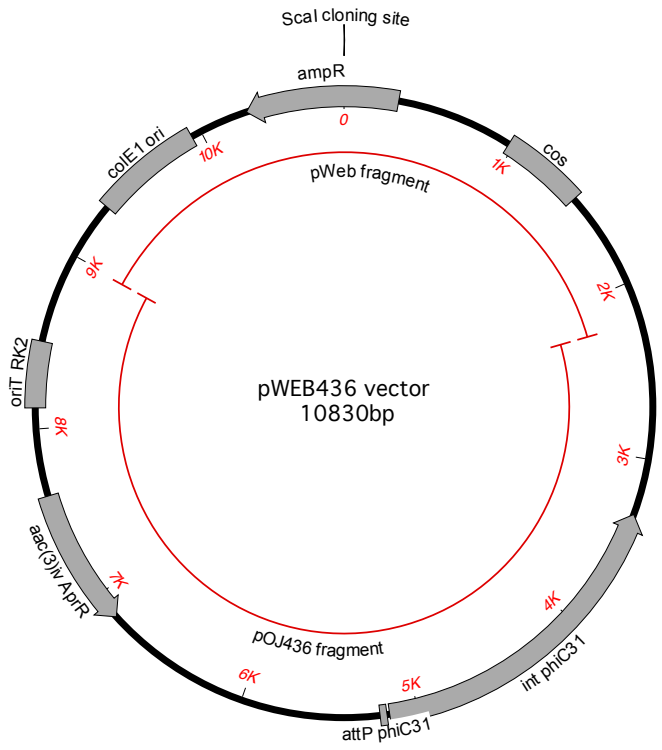


Figure 15. pWEB436 *E. coli* – *Streptomyces* shuttle vector. Vector was constructed from pWEB and pOJ436 to be replicative in *E. coli* and integrative in *Streptomyces*.

The most challenging part of creating a high quality library in *E. coli* is avoiding empty vector. Empty vector is inescapable in cloning; it can enter the library through uncut DNA in vector preparation, or through self-ligation of un-dephosphorylated vector in ligation reactions. Empty vector displays a higher transformation rate in libraries because it is significantly smaller than cosmid clones (10 kb vs 40 kb) and so can propagate easily throughout and take over a library (Hanahan, 1983). To transform a library without empty vector into *Streptomyces*, the library was first built in *E. coli* using standard protocols: eDNA was extracted from the soil sample, cloned into pWEB436 and resulting cosmids were transfected into *E. coli* EC100. Empty vector was removed from the library at this point by prepping library DNA and running this DNA and on an agarose gel to separate out higher molecule weight cosmid DNA from vector DNA. The band corresponding to cosmid DNA gel was cut and gel purified by electroelution; this DNA was then transformed into the conjugative strain of *E. coli* S17. In this way, empty vector was removed from the library before the library was transformed into *Streptomyces* to solve the problem of empty vector contamination.

Texas desert soil eDNA was cloned into pWEB436 to build and transform a 1.5 million clone library into *E. coli* S17.1 as described. Transformants in S17.1 were randomly picked, miniprepped and found to contain full-length cosmids only and no empty vector. Exconjugants from

the library mating reaction into *Streptomyces* were similarly checked for full-length cosmids to confirm lack of appreciable vector contamination in the high quality library in *E. coli*.

3.3.5 Optimizing Transformation Efficiency in *S. albus*

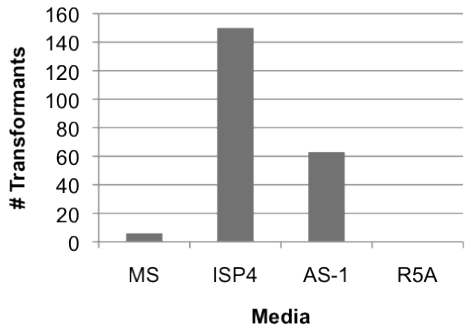
The transformation frequency of *S. albus* using standard mating conditions (10^{-6}) was too low to transform a 1.5 million clone library using a manageable number of transformation reactions and plates. The mating conditions varied to increase transformation frequency were: composition of media used for mating plates, concentration of $MgCl_2$ in the media, ratio of *Streptomyces* : *E. coli* used in the mating reaction, and the conjugative strain of *E. coli* used for mating.

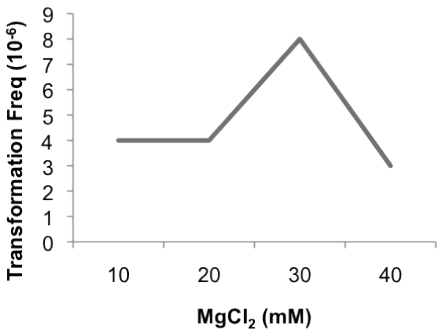
Media choice and $MgCl_2$ concentration were found to result in varying transformation efficiencies. The media that resulted in the highest number of exconjugants was ISP4, higher than the number seen on other commonly used, rich *Streptomyces* media such as R5A, MS and AS-1 (Figure 16A). In addition, increasing the concentration of $MgCl_2$ in the media from 10mM used in standard mating reactions to 30 mM resulted in higher mating efficiencies (Figure 16B).

The ratio and amount of *Streptomyces* and *E. coli* cells used in mating reactions were also varied to optimize the number of exconjugant colonies per mating plate. Plates were able to support up to 25x more cells

than used in the standard mating reaction without contamination and loss of antibiotic selection. 25x and 8x respective increases of *Streptomyces* and *E. coli* compared to the standard mating reaction, corresponding to 5 x 10^9 *S. albus* spores and 1.6 x 10^8 *E. coli* library cells, were found to result in a 67x fold increase in the number of exconjugants per plate (Figure 16C).

Finally, the conjugative strain of *E. coli* used for mating the metagenomic library into Streptomyces was also varied. The methylating strain S17.1 resulted in the highest transformation frequency into *S. albus* compared to mating using the non-methylating strain ET 12567 carrying helper plasmid pUZ8002, and the tri-parental helper mating strain DHα carrying the helper plasmid pRK2013 (Figure 16D) (Figurski & Helinski, 1979, Flett *et al.*, 1997).





Strep / E. coli	1	2	8	16
1	10	25		
2.5	20	100	100	200
10	50	100	800	1200
25	50	75	2000	2000

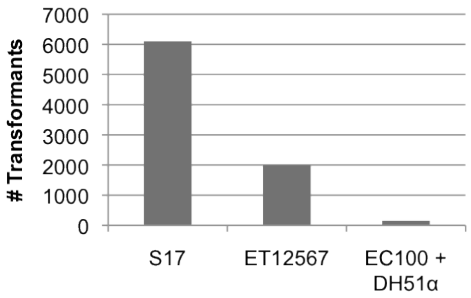


Figure 16. Optimizing the mating efficiency of *S. albus* with *E. coli* for library construction. (A) Different media were used for transformation plates. ISP4 displayed the highest number of exconjugants. (B) $MgCl_2$ concentration was varied. Highest transformation efficiencies were seen at 30 mM $MgCl_2$. (C) Different combinations of the ratio and amount of *S. albus* spores and E. coli cells were found to result in vastly improved number of exconjugants per plate. (D) *E. coli* conjugative strains S17.1, ET 12567 and DH5α strain with tri-parental mating plasmid were tested for transformation efficiency. *E. coli* S17.1 displayed the highest number of excojugants per mating plate.

The cumulative effect of changing mating conditions was a 300 x fold increase in transformation efficiency, with almost 7000 exconjugants obtained per plate. This allowed the 1.5 million member metagenomic

library to be transformed from *E. coli* into *S. albus* using 250 transformation reactions and plates.

3.3.6 Preliminary Metagenomic Library Screen in *S. albus*

Transformation plates from mating reactions of the TX desert eDNA library into *S. albus* were visually examined after 7 – 14 days for colonies displaying changes in phenotype, before making library spore stocks. Plates had approximately 10,000 exconjugants each, and the high density of colonies meant that plates could only be screened for colored clones off the mating plates. Three different colored phenotypes were consistently seen: clones displaying a yellow color, clones producing orange/red diffusible compounds, and clones displaying a brown phenotype (Figure 17).

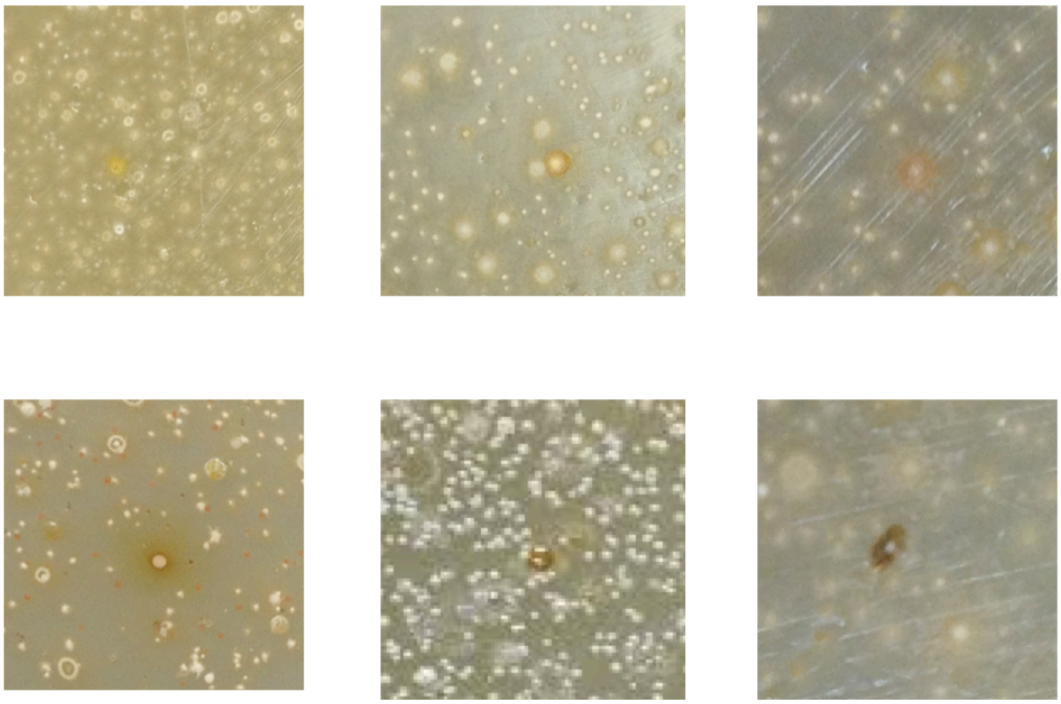


Figure 17. Examples of colored phenotype hits from screening TX soil metagenomic library in *S. albus*. Colonies range from yellow/orange to red to brown.

A total of 48 colored clones were picked from the library transformation plates and struck out for singles on ISP4 plates. Of these clones, 16 retained phenotype both when re-struck and also when the cosmid conferring phenotype was miniprepped, retransformed into *E. coli* S17.1 cells and re-mated into *S. albus*. After dereplicating, a total of 12 clones comprising 6 yellow/orange, 2 red and 4 brown clones, were categorized as unique, reconfirmed hits (Table 5).

Table 5. Unique, reconfirmed hits from phenotypic screening of TX soil metagenomic library in *S. albus*.

Clone	Color	Size (kb)	Biosynthetic Pathway
J1	Brown	34.1	
J10	Orange	31.8	Carotenoid
L10	Yellow	28.4	Carotenoid
M4	Brown	27.1	
M13	Brown	30.9	Type III Polyketide
N1	Red	24.4	Carotenoid
P1	Orange	30.3	Carotenoid
P8	Orange	36.1	Carotenoid
P11	Orange	32.2	Carotenoid
R15	Orange	25.9	Carotenoid
R16	Red	31.9	Carotenoid
R27	Brown	35.7	

3.3.7 Sequencing and Bioinformatic Analysis of *S. albus* library hits

High quality cosmid DNA was recovered from unique reconfirmed hits found in the *S. albus* metagenomic library screen and sequenced using the IonTorrent Personal Genome Machine (PGM). The sequencing obtained was assembled, and annotated using Meta GeneMark to first identify open reading frames (ORFs). BLAST and PFAM homology searches were used to assign function to ORFs. Sequences were analyzed using antiSMASH to identify biosynthetic clusters and align for homology to known pathways producing small molecules (Medema *et al.*, 2011). Clusters were marked as complete if the clone contained primary

metabolic genes flanking the secondary biosynthetic genes comprising the cluster.

Sequence analysis of hits revealed 8 clones with carotenoid pathways, 6 complete and 2 incomplete clusters, comprising core tetraterpenoid synthetic genes along with various combinations of tailoring enzymes (Figure 18). In addition, one clone was found to contain a complete type III polyketide biosynthetic cluster (Yu *et al.*, 2012). The 3 other clones did appear to harbor any clearly recognizable biosynthetic pathways for producing colored compounds in their cosmid sequences (Tables 7 – 9).

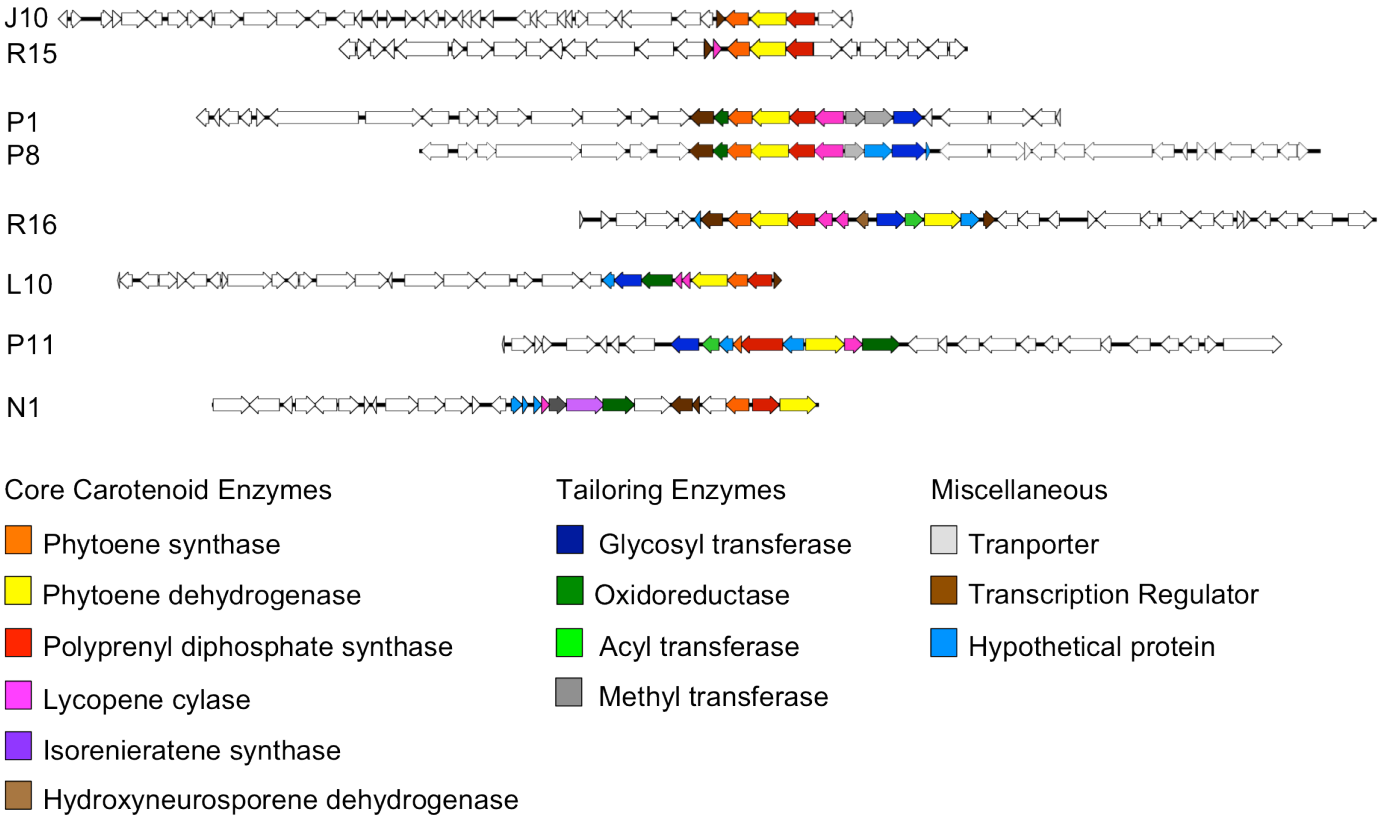


Figure 18. Sequence maps of carotenoid biosynthetic clusters found in phenotypic screen of TX soil metagenomic library.

Table 6. Open reading frames and annotations in clone M13, a type III polyketide pathway clone found in phenotypic screening of soil metagenomic library.

ORF	Size (aa)	Annotation	Organism	% Identity
1	402	Hypothetical protein	*Streptomyces viridochromogenes*	37
2	210	Membrane protein	*Catelliglobosispora koreensis*	68
3	319	Membrane protein	*Catelliglobosispora koreensis*	48
4	175	Membrane protein	*Hamadaea tsunoensis*	44
5	183	PadR transcriptional regulator	*Actinoplanes friuliensis*	69
6	130	Hypothetical protein	*Uncultured bacterium*	30
7	353	Hypothetical protein	*Microbispora rosea*	48
8	78	Hypothetical protein	*Streptomyces bicolor*	58
9	363	Aminotransferase	*Kibdelosporangium sp.*	54
10	402	Cytochrome P450	*Streptomyces actuosus*	47
11	252	Hypothetical protein	*Kibdelosporangium sp.*	44
12	223	Acetyltranferase	*Kibdelosporangium sp.*	64
13	469	Monooxygenase	*Saccharopolyspora erythraea*	57
14	73	Hypothetical protein	*Kibdelosporangium sp.*	48
15	370	Hypothetical protein	*Kibdelosporangium sp.*	56
16	157	Asparagine synthase	*Kibdelosporangium sp.*	73
17	470	Asparagine synthase	*Kibdelosporangium sp.*	67
18	127	FeS assembly SUF protein	*Sphingomonas sp.*	40
19	356	Amidohydrolase	*Mesorhizobium sp*	48
20	247	Short chain dehydrogenase	*Streptomyces sp.*	60

21	136	Phosphopantetheinyl transferase	*Acidobacteriaceae bacterium*	35
22	328	3-oxo-acyl ACP synthase	*Kibdelosporangium sp.*	70
23	495	Beta-ketoacyl synthase	*Kibdelosporangium sp.*	73
24	311	Glycine cleavage system protein	*Thermomonospora curvata*	50
25	169	Hydroxymyristol ACP-dehydratase	*Streptomyces sp.*	77
26	95	Acyl carrier protein	*Streptomyces sp.*	75
27	251	3-ketoacyl-ACP reductase	*Kibdelosporangium sp.*	70
28	309	ACP S-malonyltransferase	*Streptomyces sp.*	61
29	142	Phosphopantetheinyl transferase	*Thermomonospora curvata*	41
30	170	4-hydroxybenzoyl-CoA thioesterase	*Thermomonospora curvata*	137
31	457	Tryptophan synthase	*Kibdelosporangium sp.*	76
32	461	FAD-linked oxidase	*Kribbella catacumbae*	66
33	448	Cytochrome P450	*Kibdelosporangium sp.*	70
34	299	Thiamine biosynthesis protein	*Kibdelosporangium sp.*	59
35	400	Trehalose-phosphate synthase	*Dactylosporangium aurantiacum*	80

Table 7. Open reading frames and annotations of clone J1, found in phenotypic screening of soil metagenomic library with an unknown biosynthetic pathway

ORF	Size (aa)	Annotation	Organism	% Identity
1	275	Sugar transporter	*Gemmata obscuriglobus*	81
2	1347	Hypothetical protein	*Gemmata obscuriglobus*	68
3	346	Ketoacyl reductase	*Gemmata obscuriglobus*	81
4	73	Hypothetical protein	*Gemmata obscuriglobus*	57
5	195	Hypothetical protein	*Gemmata obscuriglobus*	66
6	389	Alcohol dehydrogenase	*Gemmata obscuriglobus*	86
7	271	Cyclase	*Gemmata obscuriglobus*	75
8	398	L-sorbosone dehydrogenase	*Gemmata obscuriglobus*	72
9	294	Manganese catalase	*Hymenobacter norwichensis*	69
10	203	Isochorismatase hydrolase	*Gemmata obscuriglobus*	71
11	120	PAS/PAC sensor histidine kinase	*Gemmata obscuriglobus*	59
12	232	Serine-glyoxylate aminotransferase	*Gemmata obscuriglobus*	70
13	139	Hypothetical protein	*Gemmata obscuriglobus*	53
14	146	Thioesterase	*Uncultured bacteria*	47
15	369	tRNA adenylyltransferase	*Gemmata obscuriglobus*	89
16	317	NAD-dependent epimerase	*Gemmata obscuriglobus*	78
17	331	GDP-mannose dehydratase	*Gemmata obscuriglobus*	87
18	370	Hypothetical protein	*Gemmata obscuriglobus*	58
19	332	Quinone oxidoreductase	*Gemmata obscuriglobus*	77
20	298	Hypothetical protein	*Gemmata obscuriglobus*	82
21	252	Hydroxyphenylpyruvate dioxygenase	*Gemmata obscuriglobus*	83
22	159	Hypothetical protein	*Thioploca ingrica*	48
23	276	Hypothetical protein	*Scytonema hofmanni*	47

24	314	Hypothetical protein	*Blastopirellula marina*	26
25	1238	Hypothetical protein	*Blastopirellula marina*	35
26	511	Cryotchrome C-554	*Blastopirellula marina*	50
27	254	Hypothetical protein	*Blastopirellula marina*	60

Table 8. Open reading frames and annotations of clone M4, found in phenotypic screening of TX soil metagenomic library, containing an unknown biosynthetic pathway

ORF	Size (aa)	Annotation	Organism	% Identity
1	76	S1/P1 Nuclease	*Methylosinus trichosporium*	44
2	150	Thioredoxin	*Gemmatimonadetes bacterium*	56
3	451	Hypothetical protein	*Gemmatimonadetes bacterium*	33
4	284	AraC domain-containing protein	*Gemmatimonadetes bacterium*	42
6	214	Hypothetical protein	*Actinomycetospora chiangmaiensis*	72
7	246	Aminoglycoside 3'-phosphotransferase	*Pseudomona aeruginosa*	56
8	56	DNA polymerase II	*Gemmatimonadetes bacterium*	91
9	213	Chromosomal replication initiator protein	*Gemmatimonadetes bacterium*	69
10	287	Chromosomal replication initiator protein	Gemmatimonadetes bacterium	73
11	105	50S ribosomal protein L34	*Gemmatimonas sp.*	80
12	131	Ribonuclease P protein	*Gemmatimonadetes bacterium*	58
13	69	Hypothetical protein	*Gemmatimonadetes bacterium*	87
14	359	Insertase	*Gemmatimonadetes bacterium*	52
15	231	Insertase	*Gemmatimonadetes bacterium*	74
16	121	Hypothetical protein	*Gemmatimonas aurantiaca*	40
17	149	Hypothetical protein	*Gemmatimonas aurantica*	36
18	277	Beta-lactamase domain protein	*Gemmatimonadetes bacterium*	61
19	409	tRNA modification GTPase	*Gemmatimonadetes bacterium*	70

Table 9. Open reading frames and annotations of clone R27, found in phenotypic screening of soil metagenomic library with an unknown biosynthetic pathway

ORF	Size (aa)	Annotation	Organism	% Identity
1	460	Hypothetical protein	*Paenibacillus polymyxa*	27
2	261	Formamidopyrimidine-DNA glycosylase	*Solirubrobacterales bacterium*	46
3	296	Hypothetical protein	*Patulibacter medicamentivorans*	38
4	376	Oxidoreductase	*Microthrix parvicella*	70
5	179	Protein-disulfide isomerase	*Frankia sp.*	50
6	140	Nucleoside disphosphate kinase	*Marine actinobacterium*	70
7	345	Rod shape-determining protein MreB	*Acidithrix ferrooxidans*	80
8	278	Cell shape-determining protein MreC	*Ferrimicrobium acidiphilum*	39
9	158	Rod shape-determining protein MreD	*Selenomonas*	33
10	642	Penicillin-bining protein PbpB	*Ferrimicrobium acidiphilum*	42
11	374	Rod shape-determining protein RodA	*Acidithrix ferrooxidans*	47
12	547	Ribonuclease	*Actinobacterium acidi*	71
13	107	50S Ribosomal protein L21	*Marine actinobacterium*	65
14	81	50S Ribosomal protein L21	*Actinobacterium acidi*	76
15	422	GTPase CgtA	*Marine actinobacterium*	66
16	364	Glutamate 5-kinase	*Marine actinobacterium*	62
17	265	Hypothetical protein	*Microthrix parvicella*	61
18	285	Carboxylate-amine ligase	*Jiangella alkaliphilia*	66
19	95	Carboxylate-amine ligase	*Nocardiopsis baichengensis*	54

20	563	ABC transporter	*Actinopolymorpha alba*	57
21	149	ABC transporter	*Actinopolymorpha alba*	57
22	463	ABS transporter	*Actinopolymorpha alba*	63
23	80	Hypothetical protein	*Frankia sp.*	51

3.3.8 Carotenoid Hits

Carotenoids are a large group of compounds, with over 600 different molecules produced by a range of microorganisms from plants to fungi to bacteria (Stahl & Sies, 2005). Carotenoids are tetraterpenoids made of 8 isoprenoid units attached together in a head to tail fashion, with various combinations of modifications such as oxidation, glycosylation and acetylation, conferred by tailoring enzymes (Paniagua-Michel *et al.*, 2012). Most carotenoids absorb visible light in the 400 – 550 nm range, imparting a red, orange or yellow color to the compounds (Stahl & Sies, 2005). This property is harnessed by photosynthetic organisms, including bacteria such as cyanobacteria, in the harvesting of light and in other photosynthetic processes such as photooxidative protection (Cogdell *et al.*, 2000, Takaichi & Mochimaru, 2007). In non-photosynthetic bacteria, carotenoids are involved in processes such as membrane stabilization, protection against oxidative damage and virulence (Gruszecki & Strzalka, 2005, Liu *et al.*, 2005). Recent therapeutic interest in carotenoids has focused on their antioxidative protective properties towards human cells. Carotenoids have especially garnered attention as dietary supplements

that are postulated to reduce the risk of developing diseases mediated by oxidative stress (Fiedor & Burda, 2014).

To identify the carotenoids being produced by metagenomic library hits, clones were grown in R5A liquid media for 10 days, and pellet-extracted using methanol and acetone to obtain carotenoid compounds. Extracts were resuspended in acetone and analyzed by HPLC-MS (Figure 19). A combination of low resolution and high resolution mass spectrometry, UV spectra, comparison to known standards, and consideration of the tailoring enzymes present in the respective biosynthetic clusters were used to identify known compounds produced by hit library clones. 3 clones J10, P1 and P8 produced the known carotenoid β-carotene, 2 clones J10 and R15 produced isorenieratene, and 1 clone R16 produced rhodopin and its hydroxylated derivative dihydroxylycopene (Table 10, Figure 20).

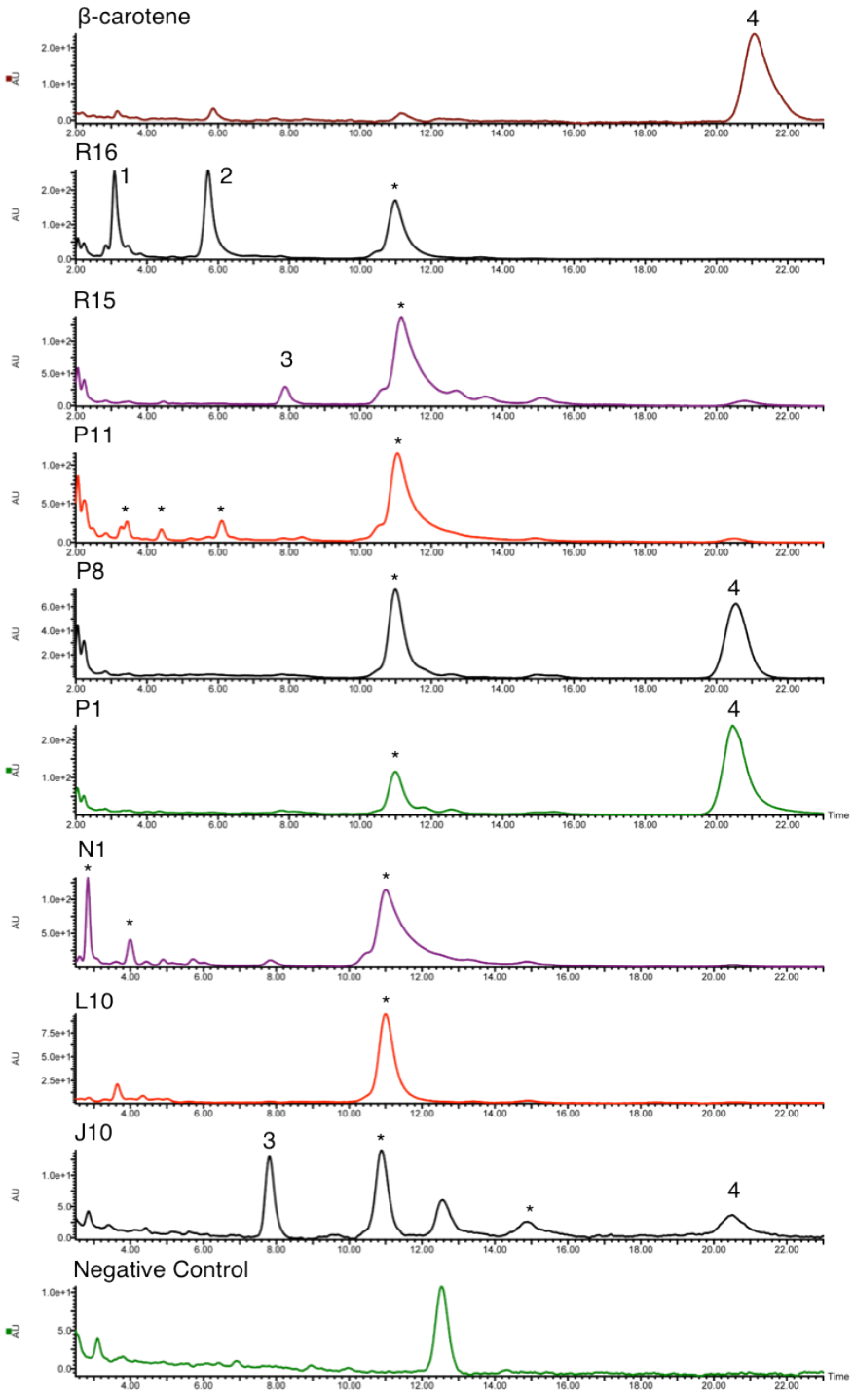


Figure 19. HPLC traces of carotenoid hits. Hits were grown in R5A media and pellet extracted using acetone. Peaks corresponding to known compounds are numbered. Peaks corresponding to unidentified compounds are starred.

Table 10. Known carotenoid compounds produced by reconfirmed hits found by phenotypic screening of TX soil metagenomic libraries.

Retention Time	Clone	LR/HR Mass	Formula	Compound	Structure
3.1	R16	572.4599	$C_{41}H_{60}O_2$	Dihydroxylycopene	1
5.5	R16	554.4481	$C_{40}H_{58}O$	Rhodopin	2
7.8	J10	528.3747	$C_{40}H_{48}$	Isorenieratene	3
	R15	528.3766			3
20.5	J10	536.84	$C_{40}H_{56}$	β-carotene	4
	P1	536.87			4
	P8	536.89			4

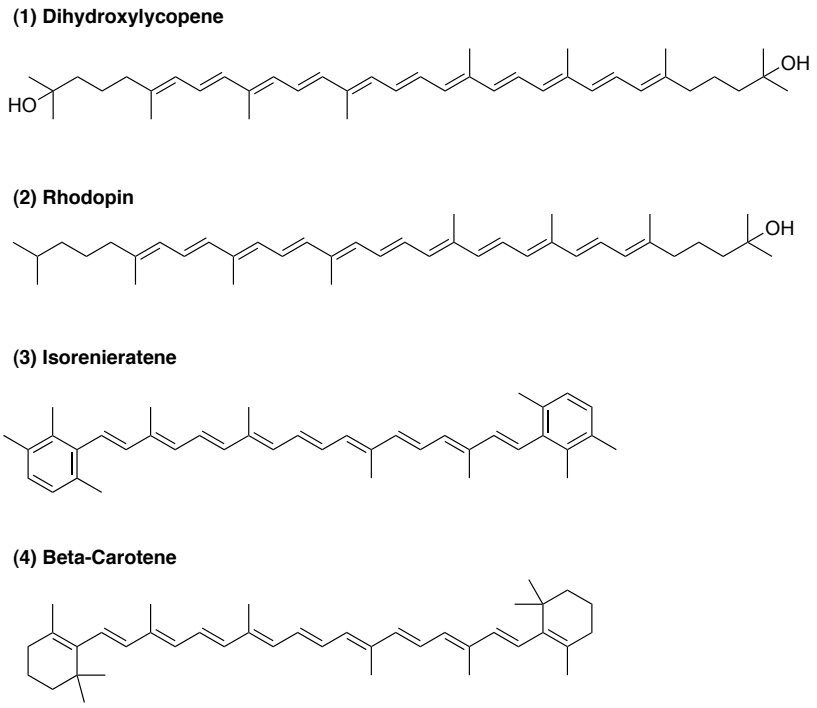


Figure 20. Structures of known carotenoids found in phenotypic screening of TX soil metagenomic libraries

A number of HPLC peaks in carotenoid extracts could not be assigned compounds using the criteria listed above (Table 11). These peaks will be HPLC-purified and analyzed using high-resolution mass spectrometry to assign exact masses. Sequencing data indicates that clone P11 is both a complete pathway and therefore most likely to produce an interesting carotenoid. This clone will be grown and extracted on a large scale (15 L of culture) and peaks with retention times 3.4, 4.4 and 6.0 mins (Figure 19, Table 11) will be analyzed using high-resolution MS and NMR to elucidate the structure of the molecules corresponding to clone-specific peaks.

Table 11. HPLC peaks corresponding to unknown compounds made by carotenoid hits

Retention Time	Clone	LR Mass
2.8	N1	
3.4	P11	
4.0	N1	
4.4	P11	
6.0	P11	
11.0	J10	641.43
	L10	641.45
	N1	641.46
	P8	641.45
	P11	641.46
	R15	641.46
	R16	641.43
15.0	J10	

3.3.9 Type III Polyketide and Unknown Pathway Hits

Metagenomic library hit M13 harboring a type III polyketide pathway was tested for the production of clone specific small molecules. Clone M13 was grown in liquid R5A media for 7 days and confluent culture was extracted using neutral and acidified ethyl acetate. HPLC analysis revealed the presence of 6 clone-specific peaks in neutral extracts (Figure 21). Large-

scale cultures of this clone will be grown, extracted, and analyzed using HRMS and NMR to identify the molecules being produced.

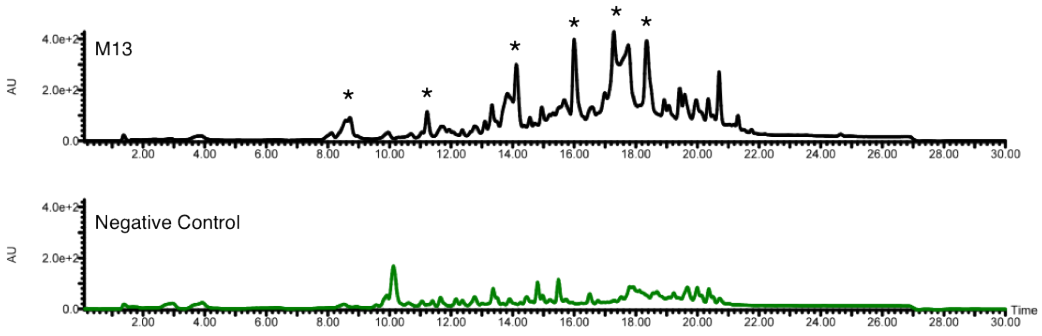


Figure 21. HPLC trace of clone M13. Clone specific peaks are starred.

To test metagenomic library hits that did not harbor recognizable biosynthetic pathways for the production of clone specific small molecules, clones J1, M4 and R27 were grown in liquid ISP4 media for 14 days, following which confluent cultures were extracted using neutral and acidified ethyl acetate. Two out of the 3 clones, J1 and R27, displayed clone specific peaks in neutral extracts (Figure 22). These clones will be further grown, extracted and analyzed using HRMS and NMR to elucidate the structure of the compounds being produced.

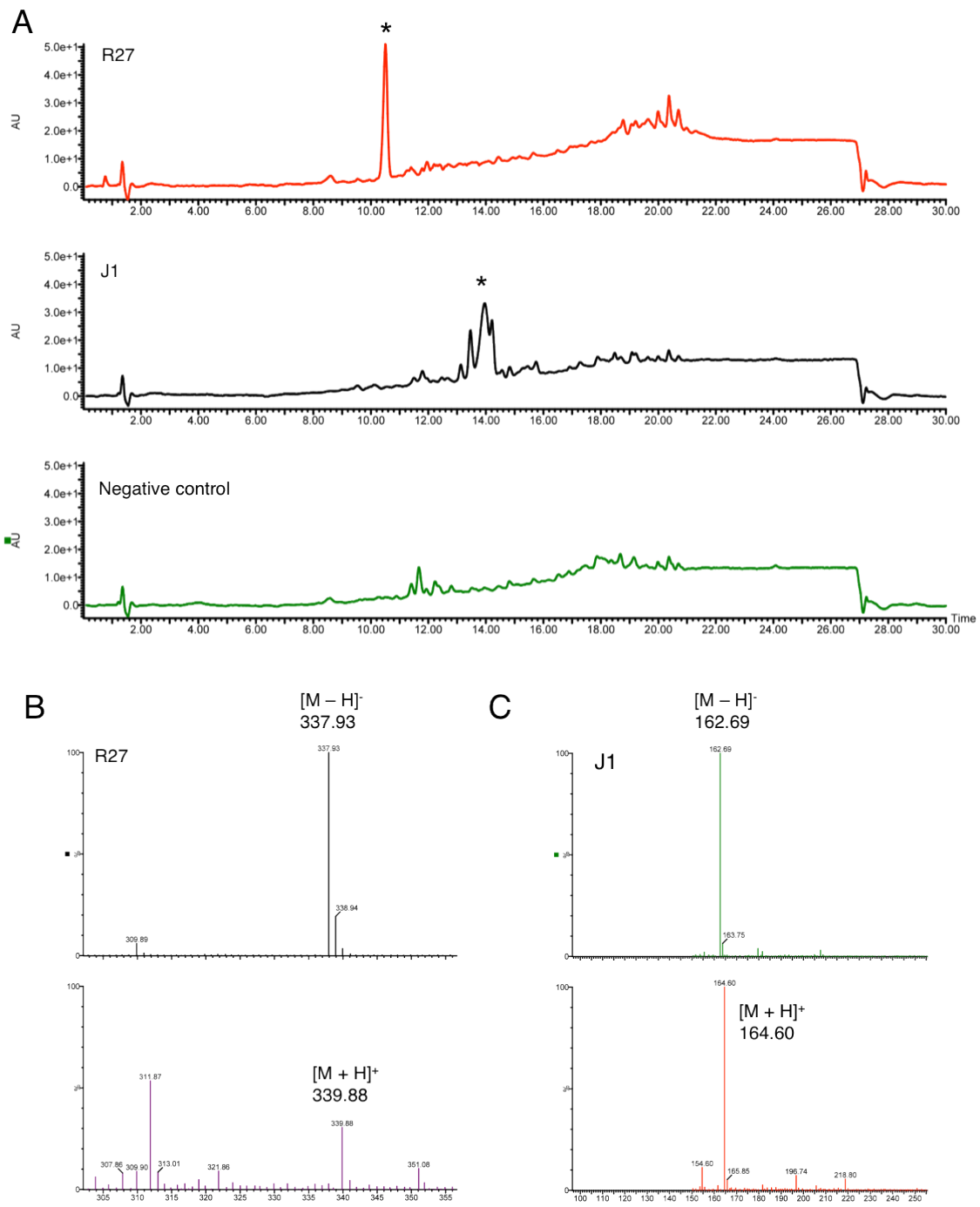


Figure 22. HPLC-MS analysis of hits containing unknown biosynthetic pathways (A) HPLC traces of neutral ethyl-acetate extracts of clones J10 and R27. Clone specific peaks are starred. (B) MS trace of clone R27 in ES + and – ion modes. (C) MS trace of clone J1 in ES + and – ion modes

3.4 Conclusion

This study represents efforts towards solving one of the bottlenecks limiting the success of functional screening of metagenomic libraries, namely: heterologous expression. Functional screening is contingent upon biosynthetic pathways encoded on metagenomic clones being expressed in the heterologous host. Streptomyces, which is a prolific source of natively and heterologously small molecules, is a promising candidate for functional screening metagenomic libraries for secondary metabolites. However, previous attempts at building and screening libraries in Streptomyces have been largely unsuccessful. Small libraries have been built in *S. lividans*, which displays poor heterologous expression capabilities, and have led to the identification of only the small molecules Terragines A-E and two fatty dienic alcohols.

In this study, I describe the construction and functional screening of a 1.5 million clone metagenomic library hosted in *S. albus*. *S. albus* was identified as having the best heterologous expression capabilities from a collection of 39 strains of *Streptomyces* including *S. lividans*. This is the only metagenomic library reported in *S. albus* to date, and is approximately 100 x larger than any other library using *Streptomyces* as a heterologous host. Preliminary functional screening of this library for colored phenotypes led to the identification of 12 reconfirmed hits, of

which 8 contained carotenoid pathways, 1 contained a type III polyketide pathway, and 3 contained genes that were not immediately recognizable as biosynthetic pathways. Small molecules produced by these clones will be extracted and investigated using HRMS and NMR for the production of novel secondary metabolites.

3.5 Future Developments

Functional metagenomics represents a promising method to access the natural products made by the uncultured majority of environmental bacteria by using a DNA-based, culture independent approach. This thesis represents an attempt at solving one of the bottlenecks of functional screening of metagenomic libraries, namely, heterologous expression of the metabolites encoded by environmental DNA, by using different bacterial library hosts. In chapter 2 of this thesis, I describe the construction and screening of a 700,000 clone library hosted in the β-proteobacteria *Ralstonia metallidurans* and find a diverse set of enzymes conferring antibacterial activity that do not display activity when the corresponding cosmids are hosted in *E. coli*. In chapter 3 of this thesis, I describe the construction of a 1.5 million clone library in *Streptomyces albus*, which is the only library built in *S. albus* to date and is almost 100 x larger than any other library reported to be hosted in *Streptomyces*.

Preliminary screening of this library for colored compounds led to the identification of diverse small molecules, some of which have never been found in functional metagenomic screens of libraries.

Further directions for this work include the use of various other assays to functionally screen this metagenomic library hosted in *Streptomyces*. Preliminary screens of this library for colored colonies were conducted on media optimized for mating instead of secondary metabolite expression. Screening this library on rich media such as R5A that is optimized for expression could lead to the identification of other colored phenotypes corresponding to novel small molecules. In addition, this library can be screened by top agar overlays for clones conferring antibiosis, indicating the presence of antibacterially active secondary metabolites. Another interesting screen would be to interrogate this library for clones resistant to antibiotics by plating the library spore stock on plates containing above-MIC levels of antibiotics. This screen could lead to the identification of novel resistance elements or resistance elements associated with biosynthetic clusters, and could also elucidate what kinds of genes are found near resistance elements.

The use of different heterologous hosts for metagenomic libraries is one way to overcome the bottleneck of heterologous expression in functional screening, and the work presented in this thesis indicates that heterologous host choice based on desired characteristics of natural

products being screened for is a productive strategy for finding small molecules and biocatalysts. In addition to heterologous host choice, other ways to overcome the bottleneck of heterologous expression in functional screening include the tailoring of bacterial hosts to make them more amenable to the expression of heterologous genes. In a recent study in *E. coli* for example, the expression of alternative sigma factors was used to aid in the recognition of heterologous promoters, and this approach was used to screen for genes conferring ethanol tolerance (Gaida *et al.*, 2015). A similar strategy involving the expression of diverse transcription factors and microbial hormones, for example A-factor in *Streptomyces griseus* and other γ-butyrlactones that control the switch from primary to secondary metabolism in *Streptomyces*, could be developed to expand the range of secondary metabolites expressed that can be expressed in different bacterial hosts (Ohnishi *et al.*, 2005, Takano, 2006). Another example of a strategy to adapt bacterial hosts for the purpose of functional screening involves genome minimization, in which inessential genes are deleted. This strategy was used in *Streptomyces avermitilis* to increase the concentration of secondary metabolites produced (Komatsu *et al.*, 2010), and could be used to improve the detection of phenotypes in functional screening.

In addition to heterologous expression, there are many other bottlenecks that limit the success of functional metagenomic screening.

One limitation is the fact that most biosynthetic clusters encoding small molecules are > 30 kb, whereas most metagenomic libraries are made using cosmid cloning techniques and therefore contain eDNA inserts of size < 30 kb. The major impediment to large insert cloning is the extraction of high molecular weight DNA from environmental samples. Most protocols for the extraction of eDNA result in DNA being sheared in the process, leading to clonable DNA of size 30 kb. Gentler extraction protocols will need to be developed that can overcome this limitation without compromising the quality of DNA obtained. In addition, vectors and cloning strategies for the stable propagation of large DNA inserts need to be developed. Large insert cloning would vastly improve the capacity to capture an entire biosynthetic gene cluster in a metagenomic library clone, which would increase the probability of small molecules with activity or color being produced, therefore leading to increased hit rates in functional screening.

Another limitation of functional screening is the fact that most assays used to find natural products from metagenomic libraries are cumbersome and lowthroughput, and so large libraries cannot be efficiently screened. One method to overcome this limitation is to develop of pre-screening selection strategies to enrich libraries for clones encoding biosynthetic genes. An example is the use of complementation strategies to select for metagenomic clones harboring PPT-ase genes that are

commonly found in biosynthetic clusters (Charlop-Powers *et al.*, 2013). This and similar selection strategies can be used to overcome the inefficiency of screening large libraries and allow faster screening of very large libraries. This method has the potential to improve functional screening.

In addition to improving the process of functional screening as described above, the usage of functional screening can also be expanded in terms of the source of environmental DNA used to make libraries. For example, functional screening has mostly concentrated on using desert soil for building and screening libraries and functional screening of environmental DNA from extreme environments has mostly been limited to small screens for biocatalysts (Iqbal, *et al.*, 2012). Expanding screening to antibacterial assays or assays for colored compounds could potentially lead to the discovery of small molecules with interesting activities. Similarly, pre-extraction strategies for the enrichment of microbiota itself, for example, for bacteria that are able to grow using certain carbon sources, or bacteria with high G-C content, or metabolically active microbiota, could also be viable methods by which to study an interesting subset of environmental bacteria.

Functional screening is a powerful method for studying uncultured bacteria in environmental samples using a DNA-based metagenomic approach. Unlike other screening methods, functional screening has the

potential to uncover completely novel classes of natural products. However, the rates of success of functional screening in finding natural products have been low. The methods described in this thesis and in this section have the potential to improve upon functional screening and bolster its use in accessing therapeutically and clinically relevant natural products produced by the uncultured majority of bacteria that comprise the environmental microbiota.

3.6 Materials and Methods

3.6.1 *Streptomyces* Strain Collection

For phylogenetic and molecular evolutionary analysis of *Streptomyces* to chose strains for investigation, 16S rRNA sequences from the genus *Streptomyces* were obtained from the Ribosomal Database Project (Release 11, version 3, https://rdp.cme.msu.edu/index.jsp). Type strain sequences from individual isolates (size > 1200 bp, good quality) were downloaded (518 sequences) and aligned using the Maximum Composite Likelihood algorithm (MUSCLE) in MEGA (version 5.2.2) (Tamura *et al.*, 2007). The phylogenetic tree was drawn using the Neighbor-Joining method. Strains of *Streptomyces* were chosen from different branches of the tree to maximize phylogenetic diversity.

39 strains of *Streptomyces* were obtained from ATCC and the Agricultural Research Service Culture (ARCS/NRRL) Collection (http://nrrl.ncaur.usda.gov/). Strains were grown in liquid YEME media (Yeast Extract 3 g/L, Bacto-peptone 5 g/L, Malt Extract 3 g/L, Glucose 10 g/L, Sucrose 170 g/L) and struck for singles on ISP4 media (Soluble Starch 10 g/L, K_2HPO_4 1 g/L, $MgSO_4.7H_2O$ 1 g/L, NaCl 1 g/L, $(NH_4)_2SO_4$ 1 g/L, $CaCO_3$ 2 g/L, Agar 20 g/L, Trace salt solution 1 ml/L added after autoclaving: $FeSO_4.7H_2O$ 0.1 g/L, $MnCl_2.4H_2O$ 0.1 g/L, $ZnSO_4.7H_2O$ 0.1 g/L). Strains were scratched out on ISP4 or, if no sporulation was seen, MS (20 g/L Mannitol, 20 g/L Soya flour, Tap water) media.

3.6.2. *Streptomyces* Spore Stocks

Lawns of *Streptomyces* on plates scratched lightly with wooden sticks were allowed to sporulate for 10 – 14 days. To make spore stocks, each plate was flooded with 4 ml dH_2O and scraped using a flame-sterilized spreader. The water and bacteria mixture was filtered through an autoclaved 10 ml syringe plugged with cotton wool to remove *Streptomyces* mycelia. Filtrate was centrifuged at 4000 rpm for 5 minutes to remove agar and residual mycelia, and the spore pellet was resuspended in minimum amount of 10% glycerol. OD_{450} spectroscopic measurements of a known volume of spore stock diluted in dH_2O were

used to estimate the number of spores/ml (OD_{450} 0.3 = 10^8 spores) (Kieser *et al.*, 2000).

3.6.3 Standard Mating into *Streptomyces*

An overnight culture of clones or library pools in *E. coli* S17.1 (ATCC 47055) was diluted 1:50 and grown to OD_{600} 0.5. *E. coli* cells were centrifuged (4000 g, 20 minutes), washed with cold LB twice and resuspended in 1/100 volume cold LB. 10^8 *Streptomyces* spores were calculated by measuring OD_{450} (10^8 spores = OD_{450} 0.3) and added to 0.5 ml Tryptic Soy Broth (TSB) (Oxoid). Spores were heat shocked for 10 mins at 50°C and were spun down. The supernatant was removed, and 0.5 ml of the washed and concentrated library pool in *E. coli* was added to the spores. The transformation mixture was spun down, resuspended in 100 ul of the supernatant and plated on 150 mm plates with 30 ml ISP4 solid media made with 10 mM $MgCl_2$. Plates were incubated at 30°C for 12 hours, then each plate was overlaid with 3 mg Apramycin (or antibiotic appropriate for *Streptomyces* exconjugant selection) and 1.5 mg Nalidixic Acid (to eliminate *E. coli*) in 2 ml dH_2O. Plates were allowed to dry and then incubated at 30°C for 10 days to allow exconjugants to sporulate. Spore stocks of the plates were made, as needed, in 10% glycerol stocks and stored at -20°C.

3.6.4 *Streptomyces* Strain Choice

A type II polyketide test library was made from 97 type II polyketide clones. Overnight cultures of individual clones in *E. coli* S17.1 were grown, pooled together and used to make a 15% glycerol stock of the type II polyketide library. Overnight cultures of the test library pool and vector control pWEB436 in *E. coli* S17.1 were mated into the collection of 39 *Streptomyces* strains and plated on 150 mm ISP4 + 10 mM $MgCl_2$ plates. Exconjugant colonies were allowed to sporulate for 14 days, following which colonies were counted and used to calculate transformation frequency in reference to the 10^8 *Streptomyces* spores used in the mating reaction. Plates from strains displaying transformation frequencies > 10^{-6} were scraped and used to make spore stocks. OD_{450} values were used to estimate spores/ml, and spore stocks were appropriately diluted to obtain 1000 colonies on R5A plates (Sucrose 100 g/L, K_2SO_4 0.25 g/L, $MgCl_2.6H_2O$ 10.12 g/L, Glucose 10 g/L, Casamino acids 0.1 g/L, Yeast Extract 5 g/L, MOPS 21 g/L, NaOH 2 g/L, Agar 20 g/L, R2YE Trace elements solution 2ml/L – $ZnCl_2$ 40 mg/L, $FeCl_3.6H_2O$ 200 mg/L, $CuCl_2.2H_2O$ 10 mg/L, $MnCl_2.4H_2O$ 10 mg/L, $Na_2B_4O_7.10H_2O$ 10 mg/L, $(NH_4)_6Mo_7O_{24}.4H_2O$ 10 mg/L).

After 14 days of growth on R5A plates, colonies were visually inspected for changes in color by comparing to the natural variation seen on vector control plates. To de-replicate *S. albus* phenotypes, 30 brown

colonies and a negative control colony from the pWEB436 vector mating plate were picked into 50 ml R5A liquid media and allowed to grow for 14 days. 20 ml of confluent cultures were extracted with 10 ml neutral Ethyl Acetate, followed by 10 ml acidified Ethyl Acetate + 0.5 ml 88% Formic Acid. Extracts were dried down in a SpeedVac, resuspended in 50 ul Methanol and run on a Waters LCMS with a C_{18} column (Xbridge 10 x 150 mm) as follows: 1.5 ml/min flow rate, solvents dH2O + 0.1% formic acid and Methanol + 0.1% formic acid used. 5 minutes at 90:10 dH_2O:Methanol, 15 minutes linear gradient from 90:10 dH_2O:Methanol to 100% Methanol, 5 minutes at 100% Methanol, 1 minute from 100% Methanol to 90:10 dH_2O:Methanol, 4 minutes at 90:10 dH_2O:Methanol. Mass Spectrometry data was collected on a Waters Micromass ZQ instrument. Chemotypes were de-replicated by comparing HPLC traces and masses of clone-specific peaks.

3.6.5 *Streptomyces* Cosmid DNA Miniprep

Streptomyces clones were inoculated into 50 ml Tryptic Soy Broth (TSB) liquid media and allowed to grow 1 – 2 days till confluence. 4 ml of cultures were centrifuged, resuspended in 1.2 ml of P1 buffer (Qiagen) + 1 mg/ml Lysozyme and incubated at 37^{o}C for 2 hours or until solution cleared. DNA was miniprepped using the Qiagen Spin Miniprep kit (Qiagen) according to standard protocol with minor modifications: 1.2 ml

and 1.68 ml of buffers P2 and N3 were used, and P2 reaction was allowed to proceed for 5 minutes. One miniprep column was used for each cosmid. DNA immobilized on the miniprep column was washed 2 x with 0.5 ml PB buffer and 6 x with 0.75 ml PE buffer. DNA was eluted out in 40 ul EB warmed to 50^{o}C.

To obtain high quality DNA for further analysis, *Streptomyces* minipreps were transformed into *E. coli* EC100 cells. Electrocompetent EC100 cells were prepared as previously described with minor modifications (Sharma & Schimke, 1996). Briefly, an overnight culture of cells was inoculated in salt-free YENB media (7.5% Yeast, 8% Nutrient Broth) and allowed to grow to OD_{600} 0.7. Cells were spun down for 20 minutes at 4000 g and rinsed twice with 50 ml cold water. The pellet was washed with cold 20 ml 10% glycerol and resuspended in 1/1000 volume 10% glycerol. Aliquots of electrocompetent cells were frozen using liquid Nitrogen, and thawed out individually for use. 20 ul of DNA from *Streptomyces* minipreps was added to each aliquot of electrocompetent EC100 cells, electroporated (0.8 kV in 1.0 mm cuvettes), bathed in warm SOC and incubated at 37^{o}C for 1 hour with shaking. The entire transformation reaction was spun down, resuspended in 100 ul of the supernatant, and plated on LB + Apramycin (75 ug/ml) plates. Colonies were picked and overnight cultures were miniprepped to obtain high yield, high quality DNA. Cosmid DNA was run on 0.7% Agarose gel with

Ethidium Bromide to ensure that it was the correct size and that no truncations were seen.

3.6.6 eDNA Extraction from Soil

The eDNA library was made using Texas desert soil (TX, USA) as previously described (Brady, 2007). Briefly, 250 g of soil was sieved through a 1/8 inch mesh screen to remove debris, into two Nalgene bottle. 15 ml lysis buffer (100 mM Tris-HCl, 100 mM Na EDTA, 1.5 M NaCL, 1% w/v cetyl trimethyl ammonium bromide, 2% w/v SDS, pH 8.0) was added and inverted well to mix. Bottles were incubated at 70^{o}C for 2 hours, and soil lystae was centrifuged at 4000 g for 10 minutes at 4^{o}C. The supernatant was centrifuged again for 20 minutes to remove any residual debris. eDNA was isolated by isopropanol precipitation: 0.7 volumes of isopropanol was added to the crude lysate, bottles were gently inverted to mix and incubated at room temperature for 30 mins. DNA was pelleted by centrifuging at 4000 g for 30 minutes at 4^{o}C. The pellet was rinsed with 100 ml of 70% ethanol, centrifuged at 4000 g for 30 minutes at 4^{o}C, and allowed to air dry for 2 – 4 hours at room temperature. DNA was resuspended in 5 – 15 ml TE (10 mM Tris, 1 mM EDTA, pH 8) warmed to 50^{o}C.

Gel purification was used to obtain High Molecular Weight DNA (> 25 kb) from the crude soil eDNA prep. eDNA was run on a 0.7 % Ethidium

Bromide-free agarose gel at 100 V for 1 hour, then 20 V for 5 hours. The band corresponding to High Molecular Weight DNA was cut from the gel and transferred to dialysis tubing (MWCO 10,000) for electroelution at 100 V for 3 hours into 0.5x TBE (45 mM Tris, 45 mM boric acid, 1 mM EDTA, pH 8). HMW DNA was concentrated in an Amicon centrifugal concentrator (MWCO 30,000), followed by rinsing with and resuspending of the eDNA in TE. eDNA was blunt-end repaired for cloning using the End-It DNA End Repair kit (Epicentre Biotechnologies).

3.6.7 Construction of vector pWEB436

The pWEB436 *E. coli-Streptomyces* shuttle vector was made from *E. coli* replicative vector pWEB (Epicentre Biotechnologies) and *Streptomyces* integrative vector pOJ436. Vectors pOJ436 and pWEB were cut using DraI and HpaI restriction enzymes respectively (New England Biolabs) and gel purified (Qiagen) on a 0.7% agarose gel to remove unrequired DNA fragments. Cut vector fragments were ligated (FastLink-DNA ligation kit, Epicentre Biotechnologies), transformed into EC100 cells and selected on LB plates containing Ampicillin (100 ug/ml) and Apramycin (75 ug/ml). The resulting vector was checked for correct size and orientation using restriction digest reactions.

3.4.8 Construction of eDNA library in *E. coli*

Vector pWEB436 was prepped for cloning by digestion with ScaI restriction enzyme (New England Biolabs) and overnight dephosphorylation with Calf Intestinal Phosphatase (New England Biolabs). Blunt end repaired eDNA was ligated to pWEB436 using the FastLink DNA ligation kit (Epicentre Biotechnologies). The ligation reaction was titered by transfection into *E. coli* EC100 cells grown with 10 mM $MgSO_4$ to OD_{600} 1.0 and plating dilutions on Apramycin plates (75 ug/ml). The number of colonies obtained on titer plates was used to calculate library clones per ligation reaction. Transfection reactions were scaled up and stored as 15% glycerol stocks of library pools with 5000 colonies each and library megapools with 500,000 clones each.

3.6.9 Transformation of the eDNA library into *E. coli* S17.1

To obtain DNA for transforming the library into *E. coli* S17.1 cells for further mating into *Streptomyces*, overnight cultures of library megapools were midiprepped (Machery-Nagel). 50 ug of DNA from each megapool was run separately on a 0.7% Ethidium Bromide-free agarose gel for 2 hours at 80V, followed by 4-6 hours at 40 V. The sides of the gel were cut and stained with Ethidium Bromide and the gel was reassembled to guide excision of the band corresponding to cosmid DNA. The cosmid DNA-containing gel slice was transferred into Dialysis Membrane tubing (12-14 kD, 25 mm width, 16 mm diameter, Spectrumlabs). 8 ml of 0.5 X TBE

Buffer was added to the tubing. Electroelution was allowed to proceed for 3 hours at 100V followed by 1 minute at 30V with the electrodes switched to dislodge any DNA stuck to the membrane. DNA was concentrated in a 30,000 molecular weight cut off concentrator (Pall Corporation) by centrifuging at 4000 rpm, followed by two rinses with 10 ml TE Buffer (10mM Tris, 1mM EDTA, pH 8.0) until a final volume of 0.5 ml was reached.

Library Megapool DNA was transformed into *E. coli* S17.1 cells. Competent cells were made from cultures grown in YENB (7.5 g/L Yeast Extract, 8 g/L Nutrient Broth) as previously described. Briefly, an overnight culture of cells was inoculated in YENB allowed to grown to OD_{600} 0.7. Cells were centrifuged for 20 minutes at 4000 g and rinsed twice with 50 ml cold water taking care not to dislodge the pellet. The pellet was rinsed with cold 20 ml 10% glycerol, resuspended in 1/1000 volume 10% glycerol and frozen in liquid nitrogen as 100 ul aliquots of electrocompetent cells. For each transformation reaction, 2 ug of DNA was electroporated into an 75 ul aliquot of S17.1 competent cells (0.8 kV in 1.0 mm cuvettes) and bathed in 1 ml warm SOC (Hi Media) followed by 1 hour incubation at 37^{o}C with shaking. Dilution titers of the transformation reaction were plated on LB with Apramycin (75 ug/ml) and used to calculate the number of reactions needed to obtain 1.5 x fold coverage of the library. Plates from

the transformation reactions were scraped into 1 ml LB and pooled to make 15% glycerol stocks of library megapools in S17.

3.6.10 Optimizing Mating into *Streptomyces albus*

To optimize mating the library into *S. albus*, modifications of the standard Streptomyces were made as follows. The standard protocol described above was plated on different solid media: MS, ISP4, AS-1 (Yeast Extract 1 g/L, L-Alanine 0.2 g/L, L-Arginine 0.2 g/L, L-Asparagine 0.5 g/L, Soluble Starch 5 g/L, NaCl 2.5 g/L, Na_2SO_4 10 g/L, Agar 20 g/L), and R5A. The mating reaction was also plated on ISP4 media with different concentrations of $MgCl_2$: 10 mM, 20 mM, 30 mM and 40 mM. The ratio of *Streptomyces* and *E. coli* used in the mating reaction was varied. *Streptomyces* spores were calculated by measuring OD450 values (OD_{450} 0.3 = 10^8 spores) with a ratio of 1 being 10^8 spores. The number of *E. coli* in a given culture was calculated by measuring OD_{600} values, and a ratio of 1 corresponded to 5 ml of culture at OD_{600} 0.5 before being washed and concentrated 10 x fold.

To test different mating strains of *E. coli*, gel-purified TX library DNA was electroporated into *E. coli* ET 12567 cells as described and selected on LB plates with Kanamycin (30 ug/ml) and Apramycin (75 ug/ml). Plates were scraped and used to make 15% glycerol stocks. An overnight cultures of this library was grown in LB + Kanamycin (30 ug/ml) +

Chloramphenicol (12.5 ug/ml) + Apramycin (75 ug/ml) and the standard *Streptomyces* mating protocol was followed. For triparental matings using the *E. coli* helper strain DH5α (pRK2013), overnight cultures of TX library in EC100 and DH5α cells were mixed in equivalent amounts and used in the standard Streptomyces mating protocol.

3.6.11 Screening the TX *S. albus* library and Reconfirming Hits

S. albus – *E. coli* mating plates of the TX library were screened visually for phenotypic variation. Plates were incubated at 30^{o}C and, starting 7 days after the mating protocol, colonies were checked every 2 days by holding plates against light and dark backgrounds. Any hits found were struck out for singles on ISP4 plates. Colonies retaining hit phenotype after 10 days were inoculated into 50 ml TSB, grown till confluence, and DNA was extracted using the *Streptomyces* cosmid DNA miniprep protocol described above. DNA obtained from minipreps was transformed into electrocompetent *E. coli* EC100 cell, and overnight cultures of transformants were miniprepped to obtain high yield DNA for further analysis.

To confirm that hits were conferred by cosmid DNA, DNA from all hits was retransformed into *E. coli* S17.1 cells by electroporating 1 ug of DNA mixed with 75 ul aliquot of *E. coli* S17.1 cells (0.8 kV in 1.0 mm cuvettes). The transformation reaction was allowed to recover at 37^{o}C in

SOC media (Hi Media) for 1 hour with shaking. Overnight cultures of hits retransformed into *E. coli* S17.1 were re-mated into *S. albus* using the standard *Streptomyces* - *E. coli* mating protocol and checked for preservation of colored phenotype before being classified as reconfirmed hits.

3.6.10 Sequencing and Bioinformatics

High quality DNA was miniprepped from reconfirmed hits and sequenced on the IonTorrent Personal Genome Machine (PGM) platform (Life Technologies) using an Ion 318 chip. Sequencing was assembled on Newbler GS *De Novo* Assembly Software (version 2.6, Roche). Individual cosmids from reconfirmed hits were sequenced using Sanger sequencing with primers designed to amplify off the vector sequences flanking the eDNA insert site (pWEB436-Fwd primer sequence: CGTAAGATGCTTTTCTGTGACTGG; pWEB436-Rev primer sequence: CGGTATTATCCCGTGTTGACGC). End sequencing data was aligned to sequences obtained via PGM sequencing to identify contigs corresponding to reconfirmed hits and to validate that contigs covered the entire eDNA insert. For clones with gaps in sequencing, primers were designed from the ends of contigs to amplify missing sequence by primer walking. Complete cosmid sequences were annotated as follows: open reading frames were identified using MetaGeneMark

(http://exon.gatech.edu/GeneMark/meta_gmhmmp.cgi) and genes were predicted by homology searches to the BLAST (blast.ncbi.nlm.nih.gov) and pfam (pfam.xfam.org) databases.

3.6.13 Carotenoid Isolation and Identification

Spore stocks of carotenoid hits and negative control pWEB436 vector mated into *S. albus* were used to start liquid cultures in R5A media. Cultures were grown at 30^{o}C with 200 rpm shaking for 10 days. To extracts carotenoids, cultures were spun down for 20 minutes at 4000 g and the pellet was rinsed with water. The pellet was washed with 0.5 volume methanol to remove superfluous metabolites. 0.5 volume acetone was added to the pellet, and cells were vortexed well and incubated at room temperature with shaking for 2-3 hours. Acetone extracts were filtered and dried down in a SpeecVac. Extracts were resuspended in 0.002 volume acetone, and stored at -20^{o}C and covered in foil to protect from light degradation as much as possible.

To identify carotenoids, extracts were run on a Waters LCMS with a C_{18} column (Xbridge 10 x 150 mm): 1.5 ml/min flow rate for 30 minutes using a 92% Acetonitrile, 6% Methanol, 2% Propanol isocratic solvent system. β-carotene (Sigma) was run alongside hits as a known standard. Low-resolution mass spectrometry data was obtained on Waters Micromass ZQ instrument. For high-resolution mass spectrometry data,

clone-specific peaks were purified on an Agilent Technologies 1200 Series HPLC using the solvent conditions described above. HRMS data was collected using Waters LCT Premier XE mass spectrometer (Sloan-Kettering Institute).

3.6.12 Characterization of Type III Polyketide and Unknown Pathway Hits

To analyze small molecules produced by type III polyketide and unknown pathway clones, 50 ml cultures of hits and pWEB436 negative control in *S. albus* were grown (30^{o}C, 200 rpm shaking). The type III polyketide clone was grown in R5A for 7 days, and unknown clones were grown in liquid ISP4 media for 21 days until confluent. Cultures were extracted with 15 ml neutral Ethyl Acetate, followed by 15 ml acidified Ethyl Acetate + 1 ml 88% Formic Acid. Extracts were dried down and resuspended in 0.002 volume methanol and run on a Waters LCMS with a C_{18} column (Xbridge 10 x 150 mm) as follows: 1.5 ml/min flow rate, solvents dH2O + 0.1% formic acid and Methanol + 0.1% formic acid used. 5 minutes at 90:10 dH_2O:Methanol, 15 minutes linear gradient from 90:10 dH_2O:Methanol to 100% Methanol, 5 minutes at 100% Methanol, 1 minute from 100% Methanol to 90:10 dH_2O:Methanol, 4 minutes at 90:10 dH_2O:Methanol. Mass Spectrometry data was collected on a Waters Micromass ZQ instrument.

3.7 References

Binnie C, Warren M & Butler MJ (1989) Cloning and heterologous expression in Streptomyces lividans of Streptomyces rimosus genes involved in oxytetracycline biosynthesis. *J Bacteriol* **171**: 887-895.

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

Calteau A, Fewer DP, Latifi A, Coursin T, Laurent T, Jokela J, Kerfeld CA, Sivonen K, Piel J & Gugger M (2014) Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *BMC genomics* **15**: 977.

Chang FY & Brady SF (2013) Discovery of indolotryptoline antiproliferative agents by homology-guided metagenomic screening. *Proc Natl Acad Sci U S A* **110**: 2478-2483.

Charlop-Powers Z, Banik JJ, Owen JG, Craig JW & Brady SF (2013) Selective enrichment of environmental DNA libraries for genes encoding nonribosomal peptides and polyketides by phosphopantetheine transferase-dependent complementation of siderophore biosynthesis. *ACS Chem Biol* **8**: 138-143.

Chistoserdovai L (2010) Functional metagenomics: recent advances and future challenges. *Biotechnology & genetic engineering reviews* **26**: 335-352.

Clardy J, Fischbach MA & Walsh CT (2006) New antibiotics from bacterial natural products. *Nat Biotechnol* **24**: 1541-1550.

Cogdell RJ, Howard TD, Bittl R, Schlodder E, Geisenheimer I & Lubitz W (2000) How carotenoids protect bacterial photosynthesis. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* **355**: 1345-1349.

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

Craig JW, Chang FY & Brady SF (2009) Natural products from environmental DNA hosted in Ralstonia metallidurans. *ACS Chem Biol* **4**: 23-28.

Craig JW, Chang FY, Kim JH, Obiajulu SC & Brady SF (2010) Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Appl Environ Microbiol* **76**: 1633-1641.

Fiedor J & Burda K (2014) Potential role of carotenoids as antioxidants in human health and disease. *Nutrients* **6**: 466-488.

Fierer N, Bradford MA & Jackson RB (2007) Toward an ecological classification of soil bacteria. *Ecology* **88**: 1354-1364.

Figurski DH & Helinski DR (1979) Replication of an origin-containing derivative of plasmid RK2 dependent on a plasmid function provided in trans. *Proc Natl Acad Sci U S A* **76**: 1648-1652.

Flett F, Mersinias V & Smith CP (1997) High efficiency intergeneric conjugal transfer of plasmid DNA from Escherichia coli to methyl DNA-restricting streptomycetes. *FEMS Microbiol Lett* **155**: 223-229.

Fong R, Vroom JA, Hu Z, Hutchinson CR, Huang J, Cohen SN & Kao CM (2007) Characterization of a large, stable, high-copy-number Streptomyces plasmid that requires stability and transfer functions for heterologous polyketide overproduction. *Appl Environ Microbiol* **73**: 1296-1307.

Gabor EM, Alkema WB & Janssen DB (2004) Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol* **6**: 879-886.

Gaida SM, Sandoval NR, Nicolaou SA, Chen Y, Venkataramanan KP & Papoutsakis ET (2015) Expression of heterologous sigma factors enables functional screening of metagenomic and heterologous genomic libraries. *Nature communications* **6**: 7045.

Gomez-Escribano JP & Bibb MJ (2012) Streptomyces coelicolor as an expression host for heterologous gene clusters. *Methods Enzymol* **517**: 279-300.

Gomez-Escribano JP & Bibb MJ (2014) Heterologous expression of natural product biosynthetic gene clusters in Streptomyces coelicolor: from genome mining to manipulation of biosynthetic pathways. *J Ind Microbiol Biotechnol* **41**: 425-431.

Gruszecki WI & Strzalka K (2005) Carotenoids as modulators of lipid membrane physical properties. *Biochimica et biophysica acta* **1740**: 108-115.

Hanahan D (1983) Studies on transformation of Escherichia coli with plasmids. *J Mol Biol* **166**: 557-580.

Iqbal HA, Feng Z & Brady SF (2012) Biocatalysts and small molecule products from metagenomic studies. *Current opinion in chemical biology* **16**: 109-116.

Kang HS & Brady SF (2013) Arimetamycin A: improving clinically relevant families of natural products through sequence-guided screening of soil metagenomes. *Angewandte Chemie (International ed in English)* **52**: 11063-11067.

Kieser T, Bibb MJ, Buttner MJ, Chater KF & Hopwood DA (2000) *Practical Streptomyces Genetics*. John Innes Foundation.

Komatsu M, Uchiyama T, Omura S, Cane DE & Ikeda H (2010) Genome-minimized Streptomyces host for the heterologous expression of secondary metabolism. *Proc Natl Acad Sci U S A* **107**: 2646-2651.

Leis B, Angelov A & Liebl W (2013) Screening and expression of genes from metagenomes. *Adv Appl Microbiol* **83**: 1-68.

Lewis RA, Laing E, Allenby N, Bucca G, Brenner V, Harrison M, Kierzek AM & Smith CP (2010) Metabolic and evolutionary insights into the closely-related species Streptomyces coelicolor and Streptomyces lividans deduced from high-resolution comparative genomic hybridization. *BMC genomics* **11**: 682.

Liu GY, Essex A, Buchanan JT, Datta V, Hoffman HM, Bastian JF, Fierer J & Nizet V (2005) Staphylococcus aureus golden pigment impairs neutrophil killing and promotes virulence through its antioxidant activity. *The Journal of experimental medicine* **202**: 209-215.

Lucas X, Senger C, Erxleben A, Gruning BA, Doring K, Mosch J, Flemming S & Gunther S (2013) StreptomeDB: a resource for natural compounds isolated from Streptomyces species. *Nucleic Acids Res* **41**: D1130-1136.

Lussier FX, Chambenoit O, Cote A, Hupe JF, Denis F, Juteau P, Beaudet R & Shareck F (2011) Construction and functional screening of a metagenomic library using a T7 RNA polymerase-based expression cosmid vector. *J Ind Microbiol Biotechnol* **38**: 1321-1328.

Manivasagan P, Kang KH, Sivakumar K, Li-Chan EC, Oh HM & Kim SK (2014) Marine actinobacteria: an important source of bioactive natural products. *Environmental toxicology and pharmacology* **38**: 172-188.

Martinez A, Kolvek SJ, Hopke J, Yip CL & Osburne MS (2005) Environmental DNA fragment conferring early and increased sporulation and antibiotic production in Streptomyces species. *Appl Environ Microbiol* **71**: 1638-1641.

Martinez A, Kolvek SJ, Yip CL, Hopke J, Brown KA, MacNeil IA & Osburne MS (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**: 2452-2463.

Mazodier P, Petter R & Thompson C (1989) Intergeneric conjugation between Escherichia coli and Streptomyces species. *J Bacteriol* **171**: 3583-3585.

McMahon MD, Guan C, Handelsman J & Thomas MG (2012) Metagenomic analysis of Streptomyces lividans reveals host-dependent functional expression. *Appl Environ Microbiol* **78**: 3622-3629.

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E & Breitling R (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* **39**: W339-346.

Niu G, Li L, Wei J & Tan H (2013) Cloning, heterologous expression, and characterization of the gene cluster required for gougerotin biosynthesis. *Chem Biol* **20**: 34-44.

Ohnishi Y, Yamazaki H, Kato JY, Tomono A & Horinouchi S (2005) AdpA, a central transcriptional regulator in the A-factor regulatory cascade that leads to morphological development and secondary metabolism in Streptomyces griseus. *Bioscience, biotechnology, and biochemistry* **69**: 431-439.

Paniagua-Michel J, Olmos-Soto J & Ruiz MA (2012) Pathways of carotenoid biosynthesis in bacteria and microalgae. *Methods Mol Biol* **892**: 1-12.

Procopio RE, Silva IR, Martins MK, Azevedo JL & Araujo JM (2012) Antibiotics produced by Streptomyces. *The Brazilian journal of infectious diseases : an official publication of the Brazilian Society of Infectious Diseases* **16**: 466-471.

Reddy BV, Kallifidas D, Kim JH, Charlop-Powers Z, Feng Z & Brady SF (2012) Natural product biosynthetic gene diversity in geographically distinct soil microbiomes. *Appl Environ Microbiol* **78**: 3744-3752.

Rondon MR, August PR, Bettermann AD*, et al.* (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* **66**: 2541-2547.

Ruckert C, Albersmeier A, Busche T*, et al.* (2015) Complete genome sequence of Streptomyces lividans TK24. *J Biotechnol* **199**: 21-22.

Ruiz B, Chavez A, Forero A*, et al.* (2010) Production of microbial secondary metabolites: regulation by the carbon source. *Critical reviews in microbiology* **36**: 146-167.

Schloss PD & Handelsman J (2006) Toward a census of bacteria in soil. *PLoS Comput Biol* **2**: e92.

Sharma RC & Schimke RT (1996) Preparation of electrocompetent E. coli using salt-free growth medium. *BioTechniques* **20**: 42-44.

Stahl W & Sies H (2005) Bioactivity and protective effects of natural carotenoids. *Biochimica et biophysica acta* **1740**: 101-107.

Staunton J & Weissman KJ (2001) Polyketide biosynthesis: a millennium review. *Natural product reports* **18**: 380-416.

Takaichi S & Mochimaru M (2007) Carotenoids and carotenogenesis in cyanobacteria: unique ketocarotenoids and carotenoid glycosides. *Cellular and molecular life sciences : CMLS* **64**: 2607-2619.

Takano E (2006) Gamma-butyrolactones: Streptomyces signalling molecules regulating antibiotic production and differentiation. *Curr Opin Microbiol* **9**: 287-294.

Tamura K, Dudley J, Nei M & Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular biology and evolution* **24**: 1596-1599.

Terpe K (2006) Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to

commercial systems. *Applied microbiology and biotechnology* **72**: 211-222.

Wang GY, Graziani E, Waters B, Pan W, Li X, McDermott J, Meurer G, Saxena G, Andersen RJ & Davies J (2000) Novel natural products from soil DNA libraries in a streptomycete host. *Organic letters* **2**: 2401-2404.

Watve MG, Tickoo R, Jog MM & Bhole BD (2001) How many antibiotics are produced by the genus Streptomyces? *Arch Microbiol* **176**: 386-390.

Wexler M & Johnston AW (2010) Wide host-range cloning for functional metagenomics. *Methods Mol Biol* **668**: 77-96.

Yu D, Xu F, Zeng J & Zhan J (2012) Type III polyketide synthases in natural product biosynthesis. *IUBMB life* **64**: 285-295.

Ziemert N, Lechner A, Wietz M, Millan-Aguinaga N, Chavarria KL & Jensen PR (2014) Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proc Natl Acad Sci U S A* **111**: E1130-1139.