2015

Metagenomics-Based Tryptophan Dimer Natural Product Discovery and Development Pipeline

Fang-Yuan Chang

METAGENOMICS-BASED TRYPTOPHAN DIMER NATURAL PRODUCT DISCOVERY AND DEVELOPMENT PIPELINE

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Fang-Yuan Chang

June 2015

METAGENOMICS-BASED TRYPTOPHAN DIMER NATURAL PRODUCT DISCOVERY AND DEVELOPMENT PIPELINE

Fang-Yuan Chang, Ph.D.

The Rockefeller University 2015

Most microbial natural product discovery programs rely on the growth of bacteria in the laboratory, yet it is now well established that the vast majority of bacteria in the environment have not been cultured, particularly from the diverse soil microbiota. By extracting DNA directly from soil samples to construct large archived environmental DNA (eDNA) libraries, thousands of genomes from both cultured and as yet uncultured bacteria can be simultaneously screened for gene clusters encoding natural products of interest. Several natural products with pharmaceutically relevant biological activity arise from the dimerization of tryptophans, such as staurosporine, rebeccamycin, and violacein. To discover novel tryptophan dimers (TDs), we have designed a metagenomics-based TD natural product discovery and development pipeline that consists of seven steps: 1) soil eDNA extraction; 2) eDNA library construction; 3) homology-based screening; 4) bioinformatics analysis; 5) heterologous expression; 6) characterization of compounds and their biosynthesis; 7) target identification.

Using a degenerate primer set that targets the CPA synthase gene, one of the conserved genes of tryptophan dimer biosynthesis, we screened the equivalent of ~1 million (over 1 tera base pairs) bacterial genomes from the eDNA libraries, resulting in the discovery of 14 unprecedented TD gene clusters, almost tripling the number of TD gene clusters that have previously been characterized. Using heterologous expression strategies that involve 1) shuttling of pathways into diverse bacterial hosts, 2)

overexpression of positive transcriptional regulator, 3) synthetic refactoring of complete pathways, and 4) co-expression of deficient biosynthetic genes, we successfully expressed nine of the 14 gene clusters. This led to the functional characterization of three novel TD families (*i.e.* indolotryptoline, carboxy-indolocarbazole, and bisindolylmaleimide), consisting of 15 novel natural products (*e.g.* BE-54017s, borregomycins, erdasporines) with therapeutically relevant bioactivities (*e.g.* antitumor, antibacterial). Linking biologically active natural products to their cellular targets remains a challenging and critical process in the development of therapeutic agents and small-molecule probes, especially for cytotoxic agents that might serve as anticancer agents. Using multidrug resistance-suppressed (MDR-sup) fission yeast resistant mutant screening, the molecular target of the indolotryptoline family of TD was identified and validated to be the proteolipid subunits of vacuolar H^+-ATPase (V-ATPase) at a putative binding site that is distinct from the previously described V-ATPase inhibitors. Together, we demonstrate the utility of this pipeline in the isolation, characterization, and development of novel natural products from the soil bacterial metagenome.

Acknowledgments

I would especially like to thank the following people for their support: Jeffrey Craig, Paula Calle, Melinda Ternei, and Shigehiro Kawashima as collaborators for my thesis research; Dr. Sean Brady as my thesis advisor; Dr. Tarun Kapoor and Dr. Howard Hang as my thesis committee members; Dr. Jef Boeke as my thesis external committee member, the rest of the members of the Brady laboratory including Hala Iqbal, Louis Cohen, Alex Milshteyn, Hahk-Soo Kang, Zachary Charlop-Powers, Kipchirchir Bitok, John Biggins, Jeremy Owen, Boojala Reddy, Debjani Chakraborty, Jonette Suiter, Jeffrey Kim, Ryan King, Jacob Banik, Zhiyang Feng, John Bauer, and Michael Clarke-Pearson; my Rockefeller classmates including James Letts, Frej Tulin, Jeff Liesch, Maria Maldonado, Mingzi Zhang, Teresa Davoli, and Jabez Bak; members of the Kapoor laboratory including Sarah Wacker, Lei Tan, Anupam Patgiri, Tommaso Cupido, Corynn Kasap, and Lynn Bidermann; members of the Nurse laboratory, including Atanas Kaykov and Jun Funabiki; and my family, James, Anna, and Sha-mei Chang.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

1-D:	One-dimensional
2-D:	Two-dimensional
AB:	Anza-Borrego (California) desert soil eDNA library
Ad:	Adenylation
A/N:	NCBI accession number
AR:	Sonoran (Arizona) desert soil eDNA library
ATP:	Adenosine triphosphate
BAC:	Bacterial artificial chromosome
Baf:	Bafilomycin A1
BE:	BE-54017
BLAST:	Basic local alignment search tool
bp:	Base-pairs
Bre:	Brefeldin A
$CHCl_3$	Chloroform
CH_3CN:	Acetonitrile
Cla:	Cladoniamide A
Con:	Concanamycin
COSY:	Correlation spectroscopy
CPA:	Chromopyrrolic acid
Da:	Daltons
DMSO:	Dimethylsulfoxide
DNA:	Deoxyribonucleic acid

eDNA:	Environmental DNA
FDA:	The United States Food and Drug Administration
HMBC:	Heteronuclear multiple quantum correlation spectroscopy
HMQC:	Heteronuclear multiple-bond correlation spectroscopy
HMW:	High-molecular weight
H_2O:	Water
HPLC:	High-performance liquid chromatography
hr:	Hour
HRMS:	High resolution mass spectrometry
IC_{50}:	Half maximal inhibitory concentration
IPA:	Indole-3-pyruvic acid
iTOL:	Interactive tree of life
KS:	Ketosynthase
KSβ:	Ketosynthase beta
LB:	Luria broth
LCHR:	Linear plus circular homologous recombination
LC/MS:	Liquid chromatography/mass spectrometry
LLHR:	Linear plus linear homologous recombination
M:	Molar
MeOH:	Methanol
MDR:	Multidrug resistance
MDR-sup:	Multidrug resistance-suppressed
MIC:	Minimum inhibitory concentration

MOPS:	3-(N-morpholino)propanesulfonic acid
NCBI:	National Center for Biotechnology Information
NM:	Chihuahuan (New Mexico) desert soil eDNA library
NMR:	Nuclear magnetic resonance
NTG:	Methylnitronitrosoguanidine
HR-ESI-MS:	High resolution electron spray ionization mass spectrometry
NRPS:	Non-ribosomal peptide synthetase
PCA:	Principal component analysis
PCR:	Polymerase chain reaction
PKS:	Polyketide synthase
PPTase:	Phosphopantetheinyl transferase
rbs:	Ribosomal binding site
RNA:	Ribonucleic acid
rRNA:	Ribosomal RNA
tRNA:	Transfer RNA
SDS:	Sodium dodecyl sulfate
SLIC:	Sequence- and ligation-independent cloning
spp:	Several species
TAR:	Transformation associated recombination
TD:	Tryptophan dimer
TPSA:	Topological polar surface area
V-ATPase:	Vacuolar H^+-ATPase

Chapter 1

1. Introduction and background

1.1 Bacterial natural products and their applications

Natural products are small molecules, typically less than 3000 Da in size, that living organisms produce as part of their biological process. Small molecular weight compounds can broadly be classified into two classes: primary metabolites that are essential for the growth and maintenance of an organism and thus their production are conserved spanning diverse phyla of life (e.g. amino acid, nucleic acid, sugar), verses secondary metabolites that are only made by specific groups (e.g. species, strains) of organisms because they are not directly involved in cellular growth (1). Although the precise role for many of these compounds is yet to be characterized, they are generally presumed to provide the producer with some selective advantage in their living environment (2), such as intercellular communication (*e.g.* quorum sensing) (3) and growth inhibition of competitors (*e.g.* bacteriocin) (Figure 1) (4). Different species of organisms interact with various environments, and thus the chemical structure and biological property of secondary metabolites that they generate are diverse (5). As such, natural product research primarily involves the discovery and characterization of secondary metabolites.

Natural products have long served as a prolific source of novel chemical entities (6). Common source of natural products include plants and fungi, but yet another important source has been bacteria. Bacterial metabolites have themselves been used as therapeutic drugs, such as antibacterial (e.g. vancomycin, daptomycin), anticancer (e.g. doxorubicin, mitomycin C), antifungal (e.g. amphotericin B, nystatin), antiparasitic (e.g.

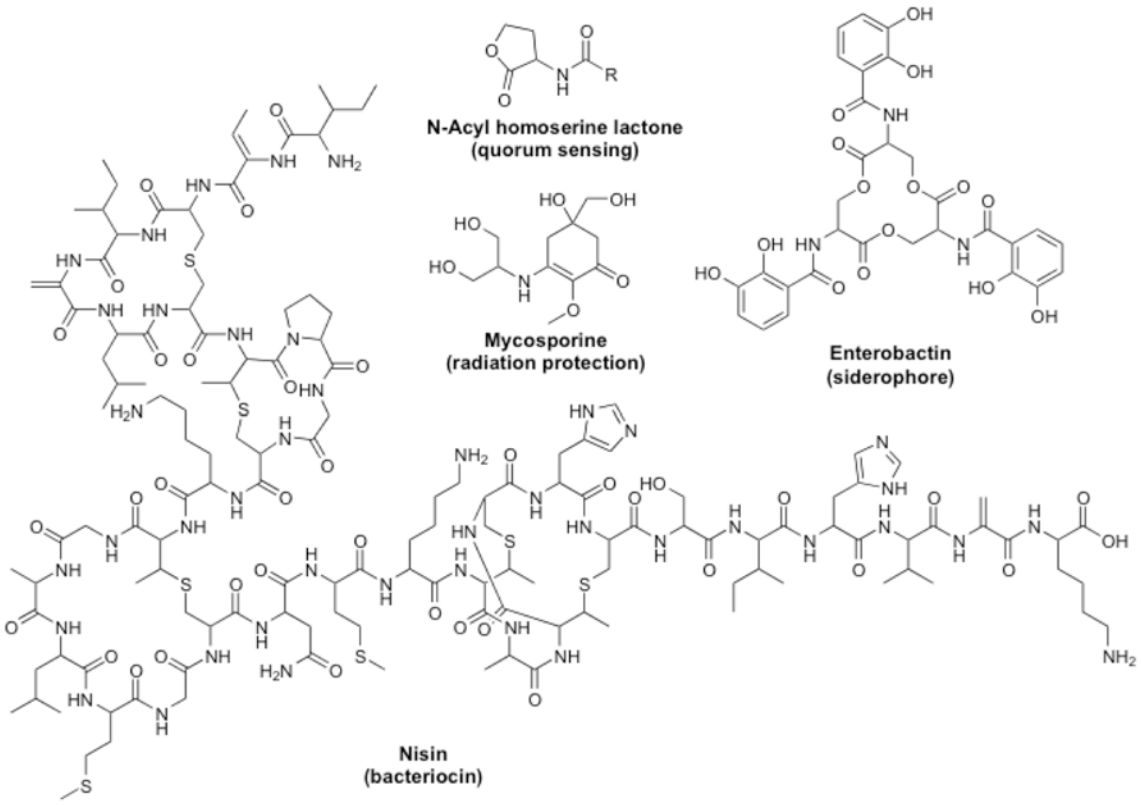


Figure 1. Examples of natural product secondary metabolites with the predicted ecological role in parenthesis.

avermectin), and immunosuppressive (e.g. FK-506, rapamycin) agents, and many more have served as leads or scaffolds to synthetic drugs (Figure 2) (7). As such, approximately 64% of the FDA approved therapeutic agents are derived from or inspired by natural products (6). Because a metabolite must bind to a molecular target to exert its biological activity, natural products are also used in basic biology as small-molecule inhibitors and activators to elucidate biological processes by perturbation, such as FK-506 for the calcineurin-mediated signaling pathway (8) and lactacystin for proteasome activity (9). Furthermore, natural products have found its use as tools for biological

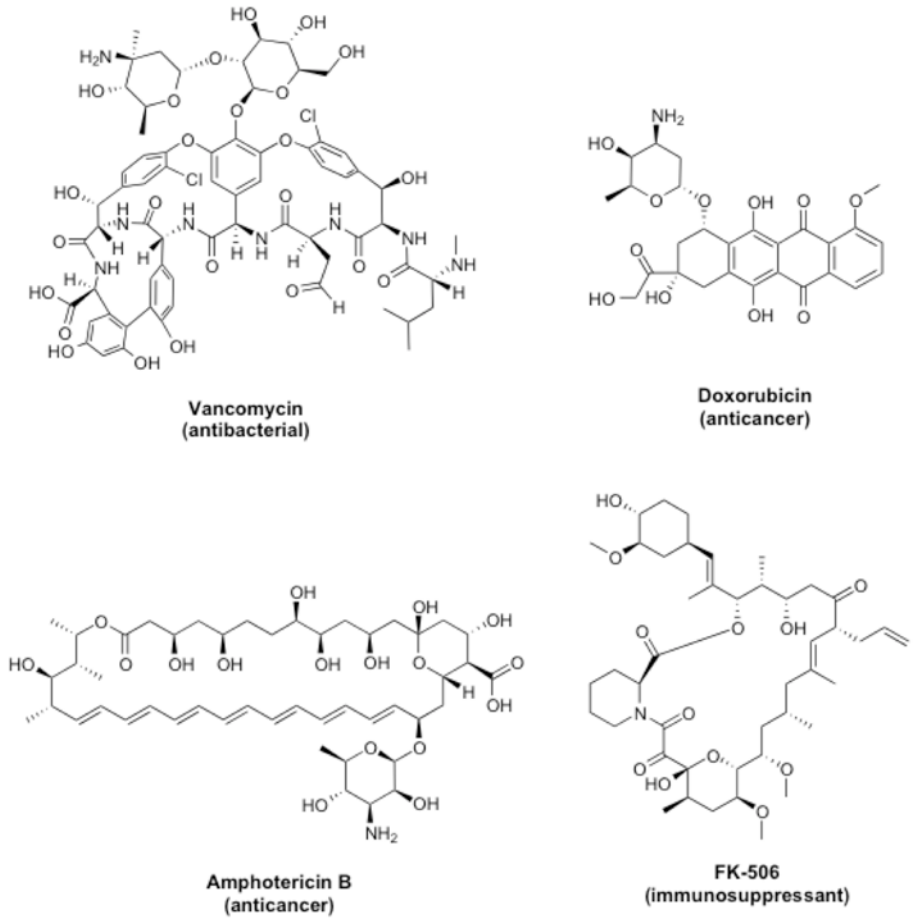


Figure 2. Examples of natural product secondary metabolites that have been FDA approved as therapeutic agents.

engineering, such as resistance markers (e.g. kanamycin, apramycin) and small-molecule mediated promoters (e.g. tetracycline/TetR) (10).

1.2 Limitations of traditional natural product discovery

Considering the significance of natural products in therapeutic, research, and engineering settings, it is disconcerting that the trend of discovering novel natural products is in a steady decline over the past few years (11). A strong repercussion is particularly observed in drug discovery with the decreasing number of novel FDA approved drugs (6). With the emergence of multiple drug resistant pathogens (12) and

various cancer phenotypes increasing the demand for novel therapeutic agents (13), the diminishing returns of natural product isolation calls for a serious reinvestigation into our traditional approaches of natural product discovery.

Soil represents one of the richest environments in bacterial biodiversity, where the number of microbial species in a single gram of soil is roughly equivalent to the number of plant species in the entire Amazon rainforest (14). Since different species produce distinct collections of secondary metabolites, one should theoretically expect the isolation of various natural products from different soil samples. In reality, however, natural product chemists have been repeatedly recovering the same compounds with high duplication rate using conventional methods, which involve the random screening of biologically active compounds from extracts of laboratory-cultured soil bacteria (15). Upon closer inspection of the soil microbiota, it has been found that the fraction of bacterial species that has been cultured in the laboratory is 1%, and thereby traditional natural product discovery techniques do not permit the recovery of metabolites that are produced by the remaining 99% of the as yet uncultured microbial majority (16-19).

1.3 Metagenomics-based natural product discovery

In order to access the metabolites that come from uncultured bacteria, we propose a natural product isolation strategy that avoids microbial culture, and instead harnesses the rapid advancement in DNA sequencing and bioinformatics technology by employing a metagenomic approach. Metagenomics refers to the study of genetic materials that are directly isolated from environmental samples (20, 21). From an environmental sample, DNA can be directly extracted instead of culturing the bacteria (Figure 3). The resulting environmental DNA (eDNA) can then be cloned into *E. coli* to construct eDNA libraries

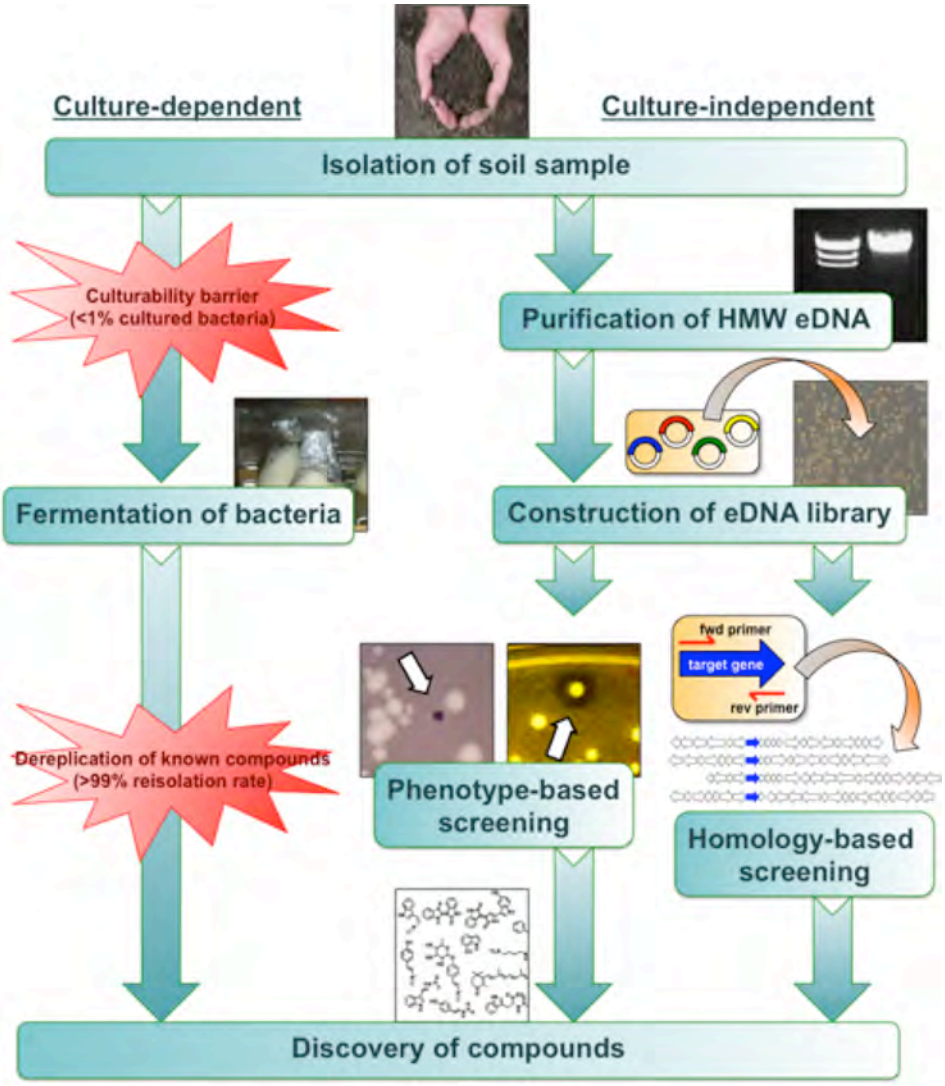


Figure 3. Culture-dependent (Left) versus culture-independent (Right) natural product programs from soil bacteria. Traditional culture-based strategy suffers from low culturability rate of soil microbes and high re-isolation rate of known compounds. Metagenome-based strategy allows access to the biosynthetic potential from the uncultured microbial majority.

(22). For prokaryotes, the genes that are required for natural product biosynthesis, including biosynthetic, regulatory, and resistance genes, are often encoded on one contiguous stretch of DNA in the host bacterial genome (23). Bacterial natural product biosynthetic gene clusters can therefore be captured on a single or multiple overlapping

eDNA clones. High-throughput survey of eDNA libraries then allows for a simultaneous and systematic screening of thousands of genomes from both cultured and uncultured bacteria for biosynthetic gene clusters of interest. The recovered gene clusters can be sequenced for bioinformatics analysis and expressed for small molecule production and isolation.

1.4 Phenotype- versus homology-based screening

Current eDNA library screening approaches can be broadly classified into two types: phenotype-based and homology-based (Figure 3) (24). Phenotype-based approach refers to the screening of eDNA-harboring clones for small molecule-producing phenotypes, which can range from coloration and zone of inhibition of individual colonies grown on solid media, to presence of novel LC/MS peaks in the organic extracts of serially diluted eDNA-harboring bacterial cultures. This screening process is independent of DNA sequence that dictates chemical structure, and therefore permits the discovery of novel compounds that are encoded by hypothetical genes, such as long-chain N-acyl amino acids (25), turbomycins (26), and isocyanide metabolites (27) (Figure 4).

However, this approach is limited to the recovery of natural products encoded by gene clusters that can be captured on a single clone. While natural product gene clusters typically span 10-100 kbp in nucleotide length, most gene clusters that are discovered from this strategy is less than 10 kbp in length (Figure 4) because large, multimillion membered eDNA libraries can currently only be constructed using cosmid-based methods that harbor 30-40 kbp of eDNA insert (28). Also, the gene cluster must be natively expressed in the eDNA-harboring heterologous host for phenotypic detection. Gene

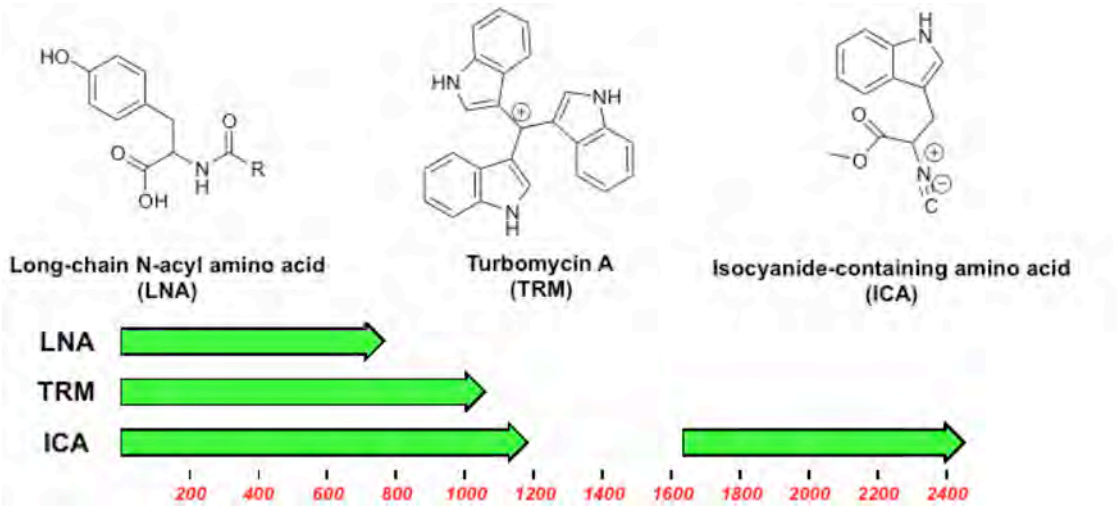


Figure 4. Compounds found from phenotype-based eDNA screening and their corresponding biosynthetic gene clusters.

clusters that satisfy this criterion are limited due to two reasons. First, phylogenetically distant DNA is generally not correctly recognized or realized by the bacterial host, possibly due to the lack of compatible RNA polymerase, transcription factor, codon usage/tRNA, molecular chaperone, post-translational modification enzyme, molecular starting unit, and so forth (29). While smaller eDNA libraries have been constructed in diverse proteobacteria (*Agrobacterium tumefaciens, Burkholderia graminis, Caulobacter vibroides, Pseudomonas putida, Ralstonia metallidurans*) (30) and *Streptomyces lividans* (31), routine large eDNA library construction has thus far been limited to *E. coli*. Second, even phylogenetically related DNA can be tightly repressed because secondary metabolites have the tendency to be produced only in response to certain conditions, such as in low iron environment for siderophores (32). Bacterial whole-genome sequencing demonstrates that the number of putative biosynthetic gene clusters far exceeds the number of characterized secondary metabolites in most cases, suggesting that the majority of gene clusters have remained silent (33).

Homology-based screening serves as an alternative approach. This method requires the identification of a conserved DNA sequence region from previously characterized biosynthetic genes. A set of degenerate primers is designed based on this region and used to survey and recover, by PCR, the eDNA clones that harbor the target conserved region. Although the screening is dependent on known sequence information, it is independent of phenotypic detection, implying that gene clusters that are silent or span multiple overlapping clones can be recovered. Although these gene clusters may not be expressed natively in the eDNA-harboring host, the genetic information can be used for bioinformatics analysis and heterologous expression using alternative bacterial hosts and genetic engineering methods.

In terms of the conventional approaches to genomic natural product discovery (34, 35), homology-based screening is analogous to the data mining of sequenced bacterial genomes for novel biosynthetic gene clusters based on sequence homology. However, even the largest high-throughput eDNA sequencing project (36), generating one billion base pairs of nonredundant sequence, covers the bacterial diversity that is contained in ~0.1 gram of soil (assuming 10,000 species per gram of soil (37) and 1,000 kbp genome size (38)). Thus, homology-based screening serves as an alternative method of surveying the astronomically large sequence space in the environmental genome, without the expenditure of time and labor for high-throughput sequencing.

The choice of target conserved region becomes an important aspect in homology-based screening. For example, nonribosomal peptides and polyketides are two major classes of microbial natural products, and the adenylation and ketosynthase domains have been well established as conserved regions in their respective gene clusters that encode a

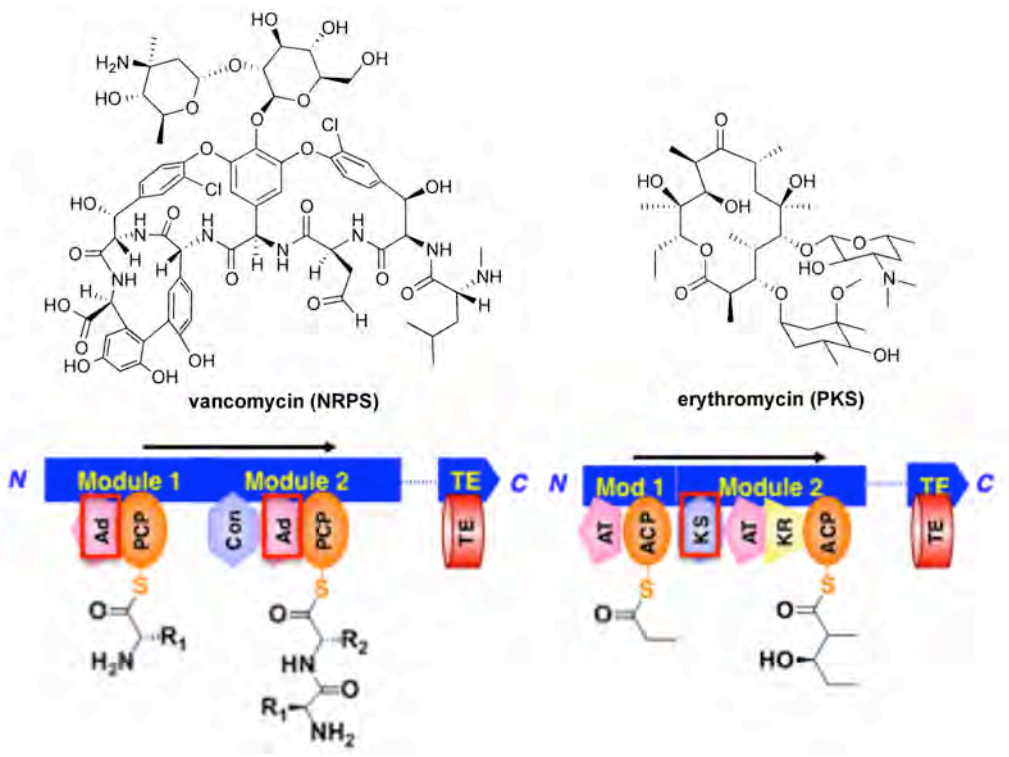


Figure 5. Multimodular nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) that contains a conserved adenylation (Ad) and ketosynthase (KS) domain, respectively (red box).

myriad of structurally diverse compounds (Figure 5) (39, 40). The thorough knowledge regarding the biosynthetic genes of these two natural product classes have greatly facilitated the discovery and generation of additional compounds of these classes (41-43). A similar trend can now be observed for the recently characterized ribosomal peptide class of natural products (44). However, with regards to biosynthetic classification based on sequence homology, the remaining bacterial natural products have been largely unexplored. Finding sequence homology regions for these natural product classes should permit their global screening from the metagenome, and contribute to the expansion of these underexplored natural product classes through the discovery of novel members and the characterization of their biosyntheses.

1.5 Tryptophan dimer natural products

A number of bacterial natural products arise from the coupling of two tryptophans (45). Referred to as tryptophan dimers (TDs), these compounds take on a variety of different chemical structures, depending on how the two tryptophans dimerize and functionalize (Figure 6). TDs are also found to be frequently associated with clinically relevant biological activities, suggesting that the TD motif may be a privileged natural substructure (46). The myriad bioactivities of TD have been believed to come from their ability to target protein kinases or other entities containing ATP-like binding sites, such as DNA topoisomerases, ABC transporters, and intercalative sites of DNA (47). In particular, staurosporine is a potent protein kinase inhibitor (48), and its discovery has spearheaded the research of kinase inhibitors as potential anticancer drugs (47). Meanwhile, rebeccamycin is a TD that is structurally similar to staurosporine, but it inhibits DNA topoisomerase I instead (49). Natural and synthetic derivatives of staurosporine and rebeccamycin have entered clinical trials for cancer, neurodegenerative disorders, and diabetes-associated pathologies (Figure 6) (50-55). From a practical standpoint, TD class is especially amenable to genome-based natural product discovery because the known TD biosynthetic gene clusters are relatively short in length (10-30 kbp), making them genetically tractable for small molecule expression (46). Moreover, the tryptophan moiety renders the compounds UV active and thereby easier to detect and isolate (46). The structural diversity, biological significance, and technical compatibility of TD make it an attractive natural product class for homology-based eDNA screening.

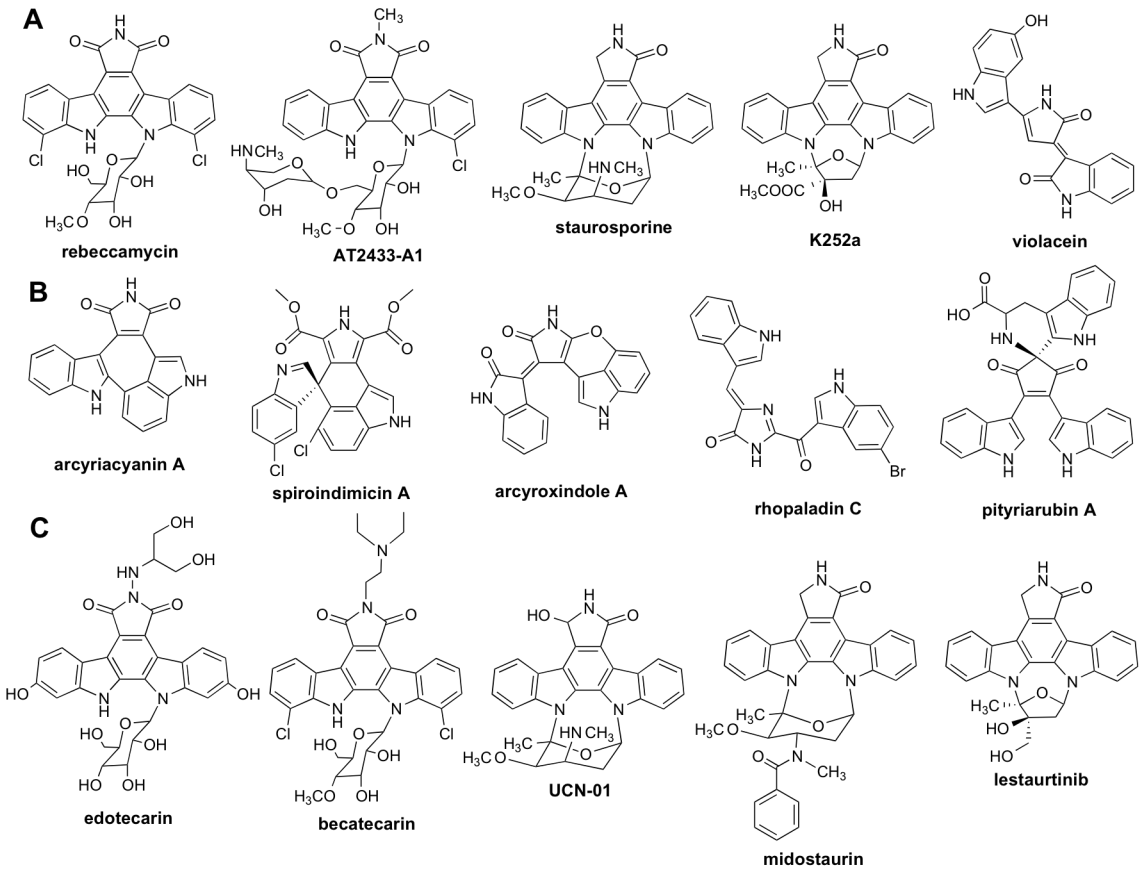


Figure 6. Examples of tryptophan dimers (TDs). **A)** Natural TDs where biosynthesis is known. **B)** Natural TDs where biosynthesis is not known. **C)** Natural and synthetic TDs that have entered clinical trials.

1.6 Tryptophan dimer product discovery and development pipeline

Here we describe a metagenomics-based natural product discovery pipeline that allows for the culture-independent isolation of secondary metabolites, using tryptophan dimer natural product discovery as a model case. The general process involves 1) soil eDNA extraction; 2) eDNA library construction; 3) homology-based screening; 4) bioinformatics analysis; 5) heterologous expression; and 6) characterization of compounds and their biosynthesis (Figure 7). In particular, bioinformatics analysis allows for the rapid dereplication of known compounds and focused recovery of metabolites

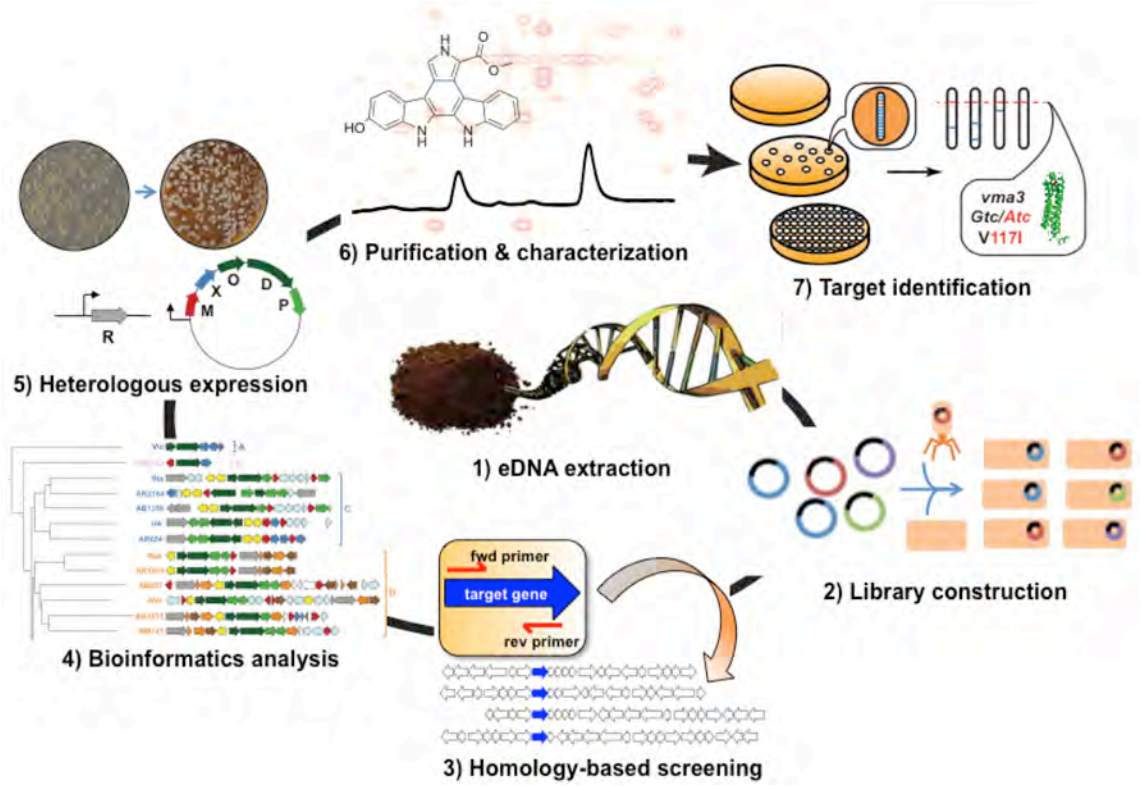


Figure 7. Seven-step overview of the metagenomics-based natural product discovery pipeline, from eDNA library construction to compound characterization.

from particularly interesting natural product families. Considering that biological characterization of natural products is significant for their further development, we also incorporate into the pipeline 7) target identification of compounds (Figure 7). From screening the equivalent of ~1 million (over 1 Tbp) bacterial genomes from the eDNA libraries using this pipeline, we discovered 14 unprecedented TD gene clusters, almost tripling the number of TD gene clusters that have previously been characterized. We functionally characterized nine of the 14 gene clusters by heterologous expression, resulting in the elucidation of three novel TD families that consist of fifteen novel biologically active natural products. The primary resistance mechanism to the

indolotryptoline family of TD was also identified and validated, suggesting the molecular target and binding site. Taken together, we demonstrate how the power of metagenomics can be harnessed, both as a small-molecule discovery tool to isolate novel structural leads, identify their mode of action, and further find derivatives of interesting leads, as well as a chemical ecology investigation tool to survey small molecule biodiversity in the environment and characterize how these compounds are biosynthesized.

Chapter 2

2. Screening for tryptophan dimer biosynthetic gene clusters

2.1 Divergent biosynthesis of TDs

To discover tryptophan dimer (TD) natural products by homology-based screening, a target homology region must first be identified from analyzing the biosynthesis of known TDs. More than 100 bacterial TDs have been isolated from various microbes (45, 46). However, prior to our culture independent discovery efforts, gene clusters have been sequenced and functionally characterized in cultured-based studies for only five bacterial TDs: staurosporine (56), rebeccamycin (57), K252a (58, 59), AT2433-A1 (60), and violacein (61). Despite the significant difference in the five chemical structures, these TDs share the same initial steps in their biosyntheses: 1) the oxidation of tryptophan by an indole-3-pyruvic acid imine (IPA imine) synthase (e.g. StaO, RebO, VioA), and 2) the dimerization of IPA imine by chromopyrrolic acid (CPA) synthase (e.g. StaD, RebD, VioB) to give CPA as a conserved TD intermediate (Figure 8A, 34) (45, 46). Although there is no experimental evidence, it has been proposed based on the requirement of heme that the reaction mechanism of the CPA synthase proceeds via hydrogen abstraction of IPA imine and subsequent radical coupling (Figure 8B). The TD biosynthetic pathways subsequently diverge in their downstream biosynthetic enzymes to generate the different TD structures.

The primary reason that the genes responsible for the production of a particular natural product tends to be clustered together in the prokaryotic genome is presumed to be because bacteria can acquire genetic information by means of horizontal gene transfer (23, 62). The ability of bacteria to obtain foreign gene clusters facilitates the evolution of

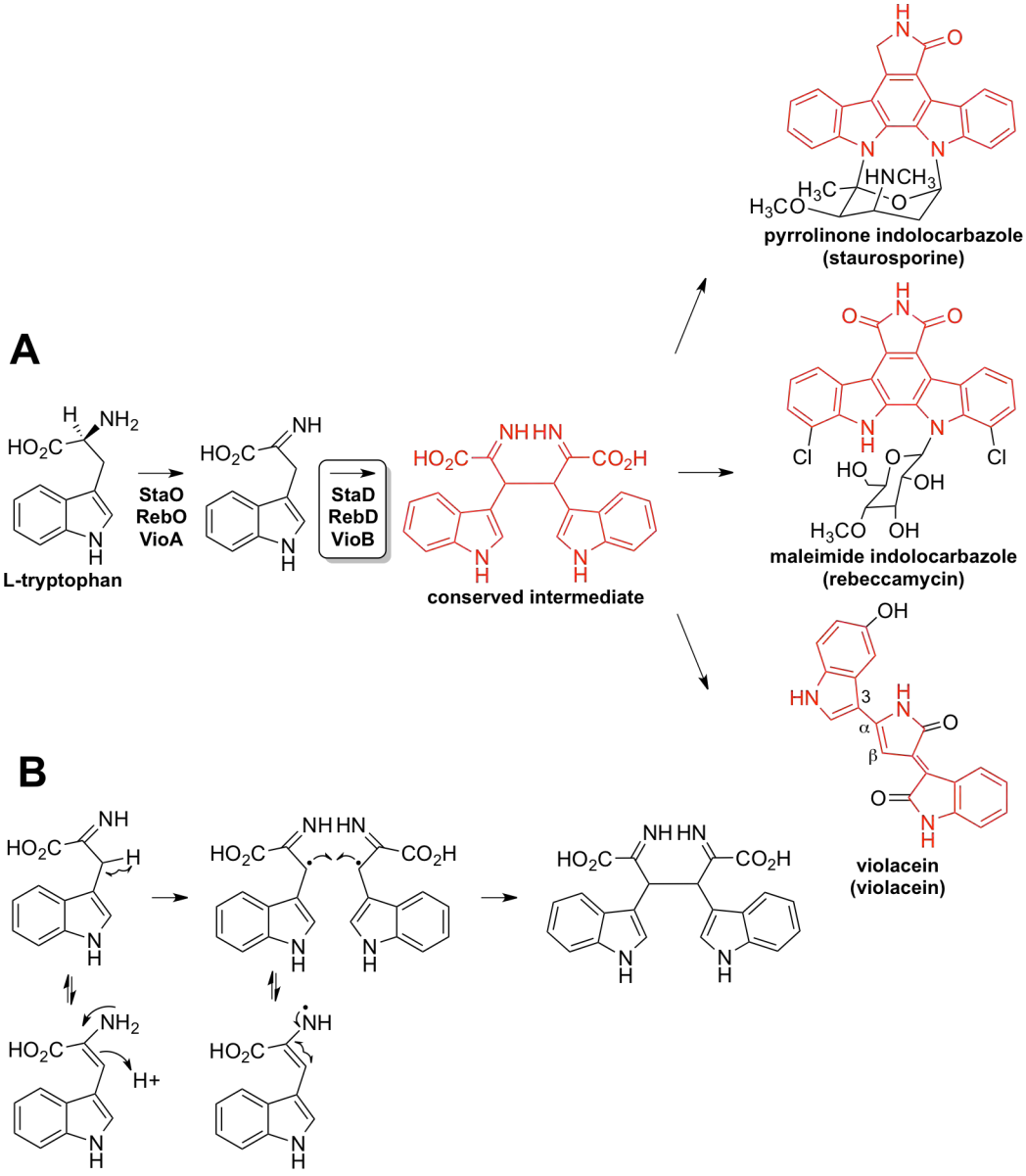


Figure 8. Conserved initial steps by IPA-imine synthase (*e.g.* StaO) and CPA synthase (*e.g.* StaD). **A)** From the two conserved initial steps, the known bacterial TD biosynthetic pathways lead to the production of three TD substructures: pyrrolinone indolocarbazole, maleimide indolocarbazole, and violacein (colored in red). Target region was identified from CPA synthase genes. **B)** Proposed mechanism of CPA synthase-mediated tryptophan dimerization.

secondary metabolites and explains the formation of structural classes that are based on their biosynthesis (63, 64). For nonribosomal peptides and polyketides, their biosynthetic enzymes consist of multiple modules, each contributing to the construction of the end product by loading a particular subunit (Figure 5) (65). Therefore, when these gene clusters are shared amongst various bacteria by horizontal gene cluster, each strain can evolve to produce a different compound tailored to its specific environment simply by adding/deleting, switching around, or modifying the substrate specificity of the individual modules (66, 67). In contrast to the modular logic of NRPS and PKS-based compounds, TDs seem to have evolved by conserving the initial biosynthetic steps that allows for the tryptophans to dimerize, while diversifying the downstream steps to generate various structures (Figure 8).

The biosynthetically characterized bacterial TDs can be classified into three structural families depending on their respective substructures (Figure 8) (68). The first two families are the indolocarbazoles, characterized by a pentacyclic core that is a fusion between an indole and a tricyclic carbazole moiety (46, 69). A pyrrole ring is attached to the center of the pentacyclic core, which is a pyrrolinone ring for one family (pyrrolinone indolocarbazole; *e.g.* staurosporine, K252a) and a maleimide ring for the other family (maleimide indolocarbazole; *e.g.* rebeccamycin, AT2433-A1) (46, 70). The third family is violacein, which is characterized by having an unusual C3-Cβ to C3-Cα carbon connectivity (Figure 8) (71). A number of TD natural products are comprised of one of three substructures (46), suggesting that, although their gene clusters have not yet been characterized, they each contain a homologous set of biosynthetic genes that give rise to its particular TD substructure.

2.2 Target homology region for TD screening

For the identification of an appropriate homology target for genetic screening, a conserved region within one of the genes that are involved in the downstream steps, and therefore the genes that are TD family-specific, should allow a focused screening of compounds containing a particular TD substructure, such as a collection of staurosporine derivatives. However, considering the unexplored biosynthetic potential of the uncultured microbial majority and a multitude of known TDs that do not fall within the characterized TD families (Figure 6B) (34), we wanted to use a homology target that can screen for gene clusters that encode novel TD families as well as derivatives. Since the three characterized TD families all share the same two initial steps in their biosynthesis, we hypothesized that many of the other unprecedented TD families are also biosynthesized by pathway divergence from these two conserved steps. Therefore, in order to conduct a more general screening of TD compounds that include previously uncharacterized TD families, we decided to identify a target region from one of the two conserved, upstream genes from the five known TD gene clusters.

A suitable gene for screening should be well conserved within the known gene clusters of interest (*i.e.* the five known bacterial TD clusters) to increase the probability of recovering positive hits, while also significantly distant from unrelated genes to minimize false positives. The IPA-imine synthase (*i.e. staO, rebO, inkO, atmO, vioA*) and CPA synthase (*i.e. staD, rebD, inkD, atmD, vioB*) genes from the staurosporine (A/N: AB088119.1), rebeccamycin (A/N: AJ414559.1), K252a (A/N: DQ399653.1) AT2433-A1 (A/N: DQ297453.1), and violacein (A/N: AF172851.1) gene clusters were retrieved from the NCBI sequenced genome database and analyzed for sequence homology with

other genes in the sequenced genome database using NCBI BLAST searches. The four IPA-imine synthase genes from the indolocarbazole pathways (*i.e. staO, rebO, inkO, atmO*) were highly conserved (blastx: >90% query cover, ~60% identity), but they showed limited homology to *vioA* (blastx: <20% query cover) and further exhibited background homology (blastx: >90% query cover, ~30% identity) to tryptophan oxidase genes (A/N: WP_020389492.1, *etc*) that do not appear cluster with other TD biosynthetic genes. On the other hand, all five CPA synthase genes were sufficiently conserved (blastx: >95% query cover, ~50% identity) without showing significant background homology to unrelated genes (blastx: <30% query cover). The CPA synthase gene was thereby chosen as the target for homology-based screening.

To identify the conserved regions for PCR primer design, the five characterized CPA synthase genes were aligned by CLUSTALW. Several regions in the DNA sequence were found to be homologous (Figure 9), but ultimately the regions for primer design was chosen based on three criteria. First, the distance between the two regions for forward and reverse primer design is at least 500 bp, such that amplicon sequencing provides enough genetic information to determine positive hits, and no more than 1000 bp to facilitate ease of PCR amplification. Second, to minimize background PCR reactions, the regions should permit the design of primers that are more than 20 bp long. Third, the regions' encoding protein sequences should be as conserved as possible, with DNA sequence degeneracy mostly coming from the third-base wobble in the genetic code. Conservation in protein sequence is prioritized over gene sequence because protein sequence is determined based more on function and less on species-dependent factors, such as codon usage bias (72, 73). Based on these criteria, regions corresponding to

2086-2108 bp and 2635-2655 bp of the *staD* gene were chosen for the design of CPA synthase-specific degenerate PCR primers to amplify CPA gene homologs from the eDNA libraries (Figure 9) (StaDVF: GTS ATG MTS CAG TAC CTS TAC GC; StaDVR: YTC VAG CTG RTA GYC SGG RTG). The fraction of degenerate bases is limited to 17% and 29% for the forward and reverse primers, respectively, yet the primer set still targets all five of the known CPA synthase genes. This primer set should thereby permit the screening of TD biosynthetic gene clusters encoding diverse TD families while minimizing false positives.

2.3 Construction of eDNA libraries

A major factor to consider for eDNA library construction is the sample type. Many protocols have thus far been reported regarding the extraction of DNA from environmental samples (environmental DNA; eDNA) and its use for the construction of an eDNA library. The eDNA isolation strategies vary depending on the sample type, such as soil (74), marine sponge (75), bromeliad tank water (76), and human feces (gut) (77), with a possible enrichment of a specific cell type (78). An optional enrichment step, such as density gradient centrifugation, is convenient in cases where the cell type of interest is mixed with other cell types in an environmental sample, such as the construction of prokaryotic eDNA libraries from marine sponges. However, although not as drastic as direct cultivation, certain types of bacteria that are more amenable to enrichment will artificially dominate the resulting eDNA library. On the other hand, soil samples mostly consists of prokaryotic cells (79); by employing gentle lysis methods that preferentially disrupts the weaker cell membranes of prokaryotes over eukaryotes (38), eDNA that mostly consists of bacterial genomic DNA can be extracted from the soil without any

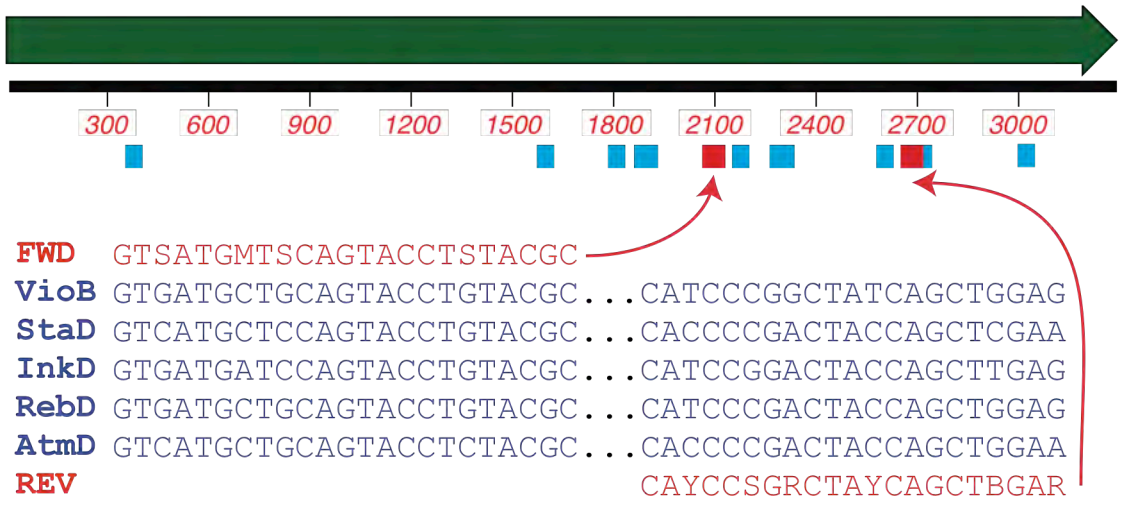


Figure 9. Degenerate PCR primer design for homology-based screening. Conserved regions (blue and red boxes) of *staD*-like CPA synthase genes were identified by CLUSTALW alignment. Regions corresponding to 2086-2108 bp and 2635-2655 bp (red boxes) of the *staD* gene (green arrow) were chosen for the design of a degenerate PCR primer pair (red text) that can target all the five known CPA synthase genes (blue text).

enrichment step. Soil sample is thereby suitable for library construction not only from the standpoint of bacterial biodiversity (refer to Chapter 1.2), but also in that it permits the isolation of eDNA without perturbing the natural bacterial populations.

A soil sample that is suitable for eDNA construction satisfies two criteria: 1) high quality eDNA can be extracted efficiently, and 2) biosynthetic diversity is rich. Empirically, our lab has routinely found arid soils to be particularly capable of yielding a large quantity of eDNA that is both relatively low in short DNA fragment contamination and in inhibitory activity of downstream cloning reactions, and thus has been samples of choice for the construction of large eDNA libraries. Although the precise reason is not clear, we postulate that the relatively low abundance of plant and fungal-based biomass in arid samples minimizes the level of humic acids and other substances that may facilitate

DNA degradation and inhibit eDNA cloning efficiency. In addition, arid soils are relatively rich in a group of bacteria called actinomycetes (80, 81), which is known to be one of the most prolific producers of secondary metabolites (82, 83). As such, a recent report shows that soils from dry area displays the richest observed biosynthetic diversity (84). Although the detailed correlation between soil type and biosynthetic diversity still remains to be investigated, previous empirical observations and experimental data support the construction of eDNA library using desert soil samples.

The eDNA for library construction was isolated from soil samples using total extraction method (Figure 10A) (28). This procedure involves the removal of debris matters from the soil using a sieve, followed by resuspension in a detergent-containing buffer. The sample was shaken with moderate heating to gently lyse the soil bacteria for the liberation of eDNA from the soil. After the removal of soil particulates by centrifugation, eDNA was precipitated from the supernatant using isopropanol. Based on agarose gel electrophoresis analysis, much of this crude eDNA was sized at around 30-100 kbp due to mechanical shearing from the extraction protocol thus far, with detectable levels of smearing that signify the minute presence of short DNA fragments below 30 kbp. Preparatory agarose gel electrophoresis was used to purify high-molecular weight (HMW; >30 kbp) eDNA from crude eDNA.

The construction of a clone library from HMW eDNA has generally been conducted in either a bacterial artificial chromosome (BAC)- (85-87) or cosmid/fosmid-based (88-90) vector system. BAC is a single-copy vector that mimics the nature of chromosome and maintains long insert sizes (91). The clones from BAC eDNA libraries have thereby accommodated up to 200 kbp in insert size, which can capture the majority

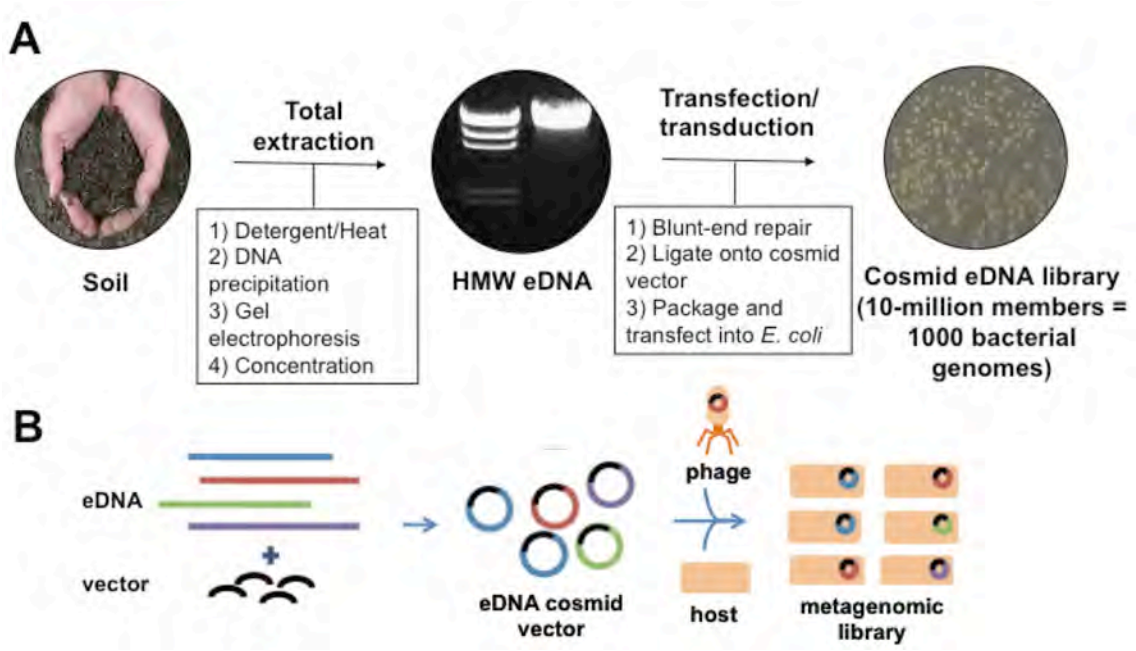


Figure 10. Schematic of eDNA library construction. **A)** High-molecular weight (HMW) eDNA is purified from a soil sample by total extraction and used to construct a library using cosmid (transduction)-based approach. **B)** Cosmid-based approach allows for bacteriophage to select the size of the eDNA cosmid, such that each eDNA clone contains 30-40 kbp of eDNA insert.

of prokaryotic biosynthetic gene clusters. However, due to the mechanical shearing of DNA that arises from standard DNA extraction and purification protocols, a reliable isolation of large DNA suitable for BAC library construction (>100 kbp) proves to be technically challenging (92). Moreover, BAC eDNA libraries are typically contaminated with short insert clones because even a small fraction of short fragment DNA contained in purified HMW eDNA samples will become enriched in the subsequent ligation and transformation steps of library construction (92). As such, much of the previously constructed BAC eDNA libraries range about 10,000-50,000 members, with an average insert size of 100 kbp for smaller and 50 kbp for larger libraries, thereby reaching up to 4 Gbp in total genome size. In contrast, cosmid (multi-copy) or fosmid (single-copy) is a

plasmid that can be packaged into a lambda phage, which can then be transferred into a bacterial host for library construction (transfection/transduction) (93). Because the size of the cosmid is selected by the phage capsid, cosmid libraries have uniform sized inserts of 30-40 kbp (28). In homology-based screening, the total genome size of the library is more important than the maximum insert size because, as long as the coverage of the library is sufficient such that entire biosynthetic gene cluster is captured in multiple overlapping vectors, it does not need to be captured in a single vector. Previous studies show that, in order for the library to capture most of biosynthetic capacity in a soil sample, it should have a total genome size of around 400 Gbp (94). Since a BAC eDNA library of such size remains to be constructed, we have chosen to construct multiple eDNA libraries using a cosmid-based approach. From soil samples collected in the Anza-Borrego desert of California (AB), the Sonoran desert of Arizona (AR), and the Chihuahuan desert of New Mexico (NM), three cosmid-based eDNA libraries were constructed, each consisting of about 15,000,000 unique cosmid clones (Figure 10B). By screening all three of these multimillion-membered eDNA libraries, an equivalent of ~1 million bacterial genomes (over 1 Tbp in total genome size) can be simultaneously surveyed for their biosynthetic potential.

2.4 Screening for eDNA-derived TD gene clusters

To facilitate homology-based screening, each of the 15,000,000 cosmid-based eDNA libraries were archived into unique sublibraries that each consist of 5000 clones (~3000 sublibraries per eDNA library). Matching DNA miniprep and glycerol stock pairs were made for each sublibrary, with the DNA minipreps arrayed such that sets of 8 sublibraries were combined to generate unique "row pools" (~375 row pools per eDNA

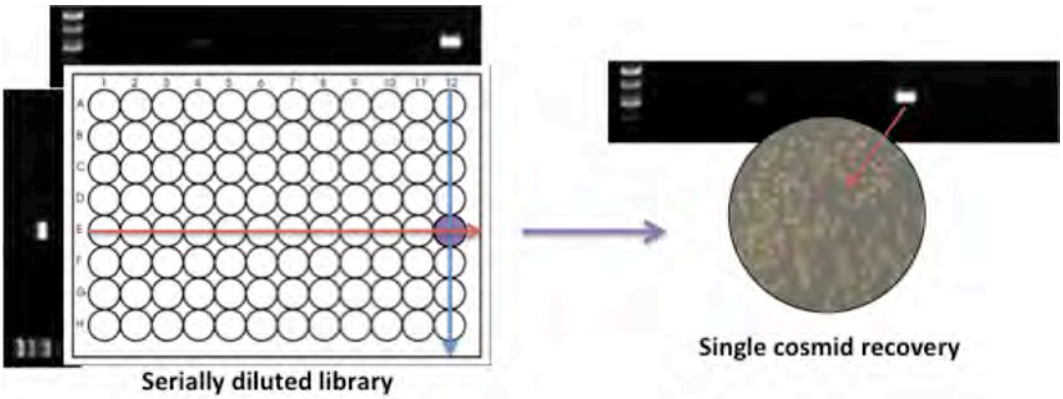


Figure 11. Schematic of eDNA clone recovery by serial dilution. Upon determining the sublibrary containing the clone of interest, the overnight culture of the sublibrary is diluted and arrayed into 96 well plates. The well containing the clone of interest is determined by whole cell PCR of the pooled "rows" and "columns" and subsequently plated onto solid media for single cosmid recovery.

library). To screen for TD gene clusters from the eDNA libraries, the "row pools" were used as PCR templates to detect and amplify ~570bp long CPA synthase PCR products with the previously described degenerate primer set (Figure 9). Approximately 10% of the "row pools" that were screened yielded an amplicon of the correct length. These amplicons were gel purified, re-amplified, and sequenced. Based on the NCBI homology search (blastx), ~80% of the sequenced amplicons were found to be come from CPA synthase genes and were followed up for the recovery of their encoding eDNA clones.

Each of the eDNA clones containing a unique CPA synthase amplicon was isolated using serial dilution strategy (Figure 11). First, a set of specific PCR primers that target each unique CPA synthase amplicon was designed and tested on the DNA minipreps of the 8 sublibraries that constitute the "row pool" hit to determine the sublibrary that contains the target clone. The matching glycerol stock of the eDNA sublibrary was then grown, diluted, and arrayed into sterile 96 well plates, such that each

well contains about 25 cells. The “rows” and “columns” of the wells are pooled and tested by whole cell PCR to determine the well containing the clone of interest. The culture broth from this well was grown on solid media and single colonies were screened by PCR to identify the specific eDNA clone harboring the targeted CPA synthase gene. After the first round of clone recovery, we empirically discovered that CPA synthase amplicon sequences sharing >95% identity came from the same TD gene clusters, and thus the amplicons were clustered at 95% identity in the subsequent rounds of clone recovery to avoid redundancy. In total, 16 unique CPA synthase amplicons and their harboring eDNA cosmid clones were found from the homology-based screening: 9 from the AB, 4 from the AR, and 3 from the NM libraries.

2.5 Bioinformatics analysis of TD gene clusters

The recovered eDNA clones containing the amplicon sequences were *de novo* sequenced using Next-Generation Sequencing technology. The sequences from the cosmid assemblies were annotated by FGENESB for gene (open reading frame) identification and by BLAST (blastp of the gene product) for the bioinformatic prediction of each gene product’s function (Appendix 1). Based on the annotation, all 16 eDNA clones harboring the amplicon sequences were identified to contain putative TD gene clusters, in that each possessed a set of genes, including a putative CPA synthase gene, that is similar to those found in previously annotated TD gene clusters (Figure 12). These putative gene clusters were all less than 27 kbp in length and, with the exception of one gene cluster (AR1973), a cosmid clone harboring the intact, complete gene cluster was found for each TD pathway. Two eDNA clusters, AB1350 and AR1455, were identical in gene content to the characterized staurosporine and rebeccamycin clusters, respectively

(56, 57). The remaining 14 TD gene clusters were unprecedented in gene content, almost tripling the number of bacterial TD gene clusters that have been previously characterized.

For the comparative bioinformatics analysis of the relatively unexplored TD gene clusters, we considered employing phylogenetics used to visualize the evolutionary relationships of bacterial species. In the construction of a phylogenetic tree, a particular gene that is known to undergo a low evolutionary rate is chosen as the phylogenetic marker because its genetic variation will then be strongly linked to ancestral divergence events (80). The gene that is resilient to genetic change tends to be a housekeeping gene that is necessary for cellular viability and conserved across all bacteria. The 16S rRNA gene (83), in particular, has been well-established as a suitable phylogenetic marker, where the comparison of genetic variation in solely the conserved region of the 16S rRNA gene serves as a good proxy for the classification of bacterial species.

The concept of phylogenetics can similarly be translated to natural product biosynthetic gene clusters, where natural product taxonomy can be defined based on the substructures of the compounds encoded by the gene clusters. For example, when a phylogenetic tree was constructed based on the ketosynthase beta (KSβ) domain from a collection of eDNA-derived type II polyketide gene clusters, the clusters were found to group together based on the particular polyketide substructures that they produce (42). Therefore, by using the KSβ domain as a phylogenetic marker or a sequence tag, eDNA-derived gene clusters that encode for chemically novel or clinically relevant polyketide substructures can be rapidly detected on the phylogenetic tree and targeted for heterologous expression (82, 95).

Since the CPA synthase gene is conserved across the 16 eDNA-derived TD gene clusters, we investigated the possibility of the CPA synthase gene as a phylogenetic marker. When the 16 eDNA-derived TD gene clusters and the 5 functionally characterized TD gene clusters were organized according to CPA synthase gene phylogeny, the clusters with similar gene content grouped together. CPA synthase genes, and in turn the TD clusters from which they arise, form 6 distinct clades (Figure 12; Groups A-F). Based on the TD chemical structures encoded by functionally characterized gene clusters, the groupings appear to correlate with the production of distinct TD substructures, suggesting the CPA synthase gene to be a suitable sequence tag for the classification of TD gene clusters (84).

Group A contains violacein-like TD gene clusters, characterized by the presence of *vioE* homologs, which are responsible for introducing the unusual C3-Cβ to C3-Cα carbon connectivity in violacein (89, 94) (Figure 8). The purple pigmentation of violacein enables its use as a reporter (91, 92), and its biological properties, such as antibacterial (93), antitrypanocidal (87), and anticancer (90) activities, are being investigated for medical applications.

Group C contains gene clusters that encode for compounds with a pyrrolinone (monooxygenated) indolocarbazole core, like staurosporine. On the other hand, gene clusters from Group D encode for compounds with a maleimide (dioxygenated) indolocarbazole core, like rebeccamycin. While gene clusters from both groups contain a functionally equivalent cytochrome P450 enzyme (StaP/RebP) that catalyzes the formation of the pentacyclic structure of the indolocarbazole core, the oxidation state of the pyrrole ring of the indolocarbazole core is dictated by the FAD-binding

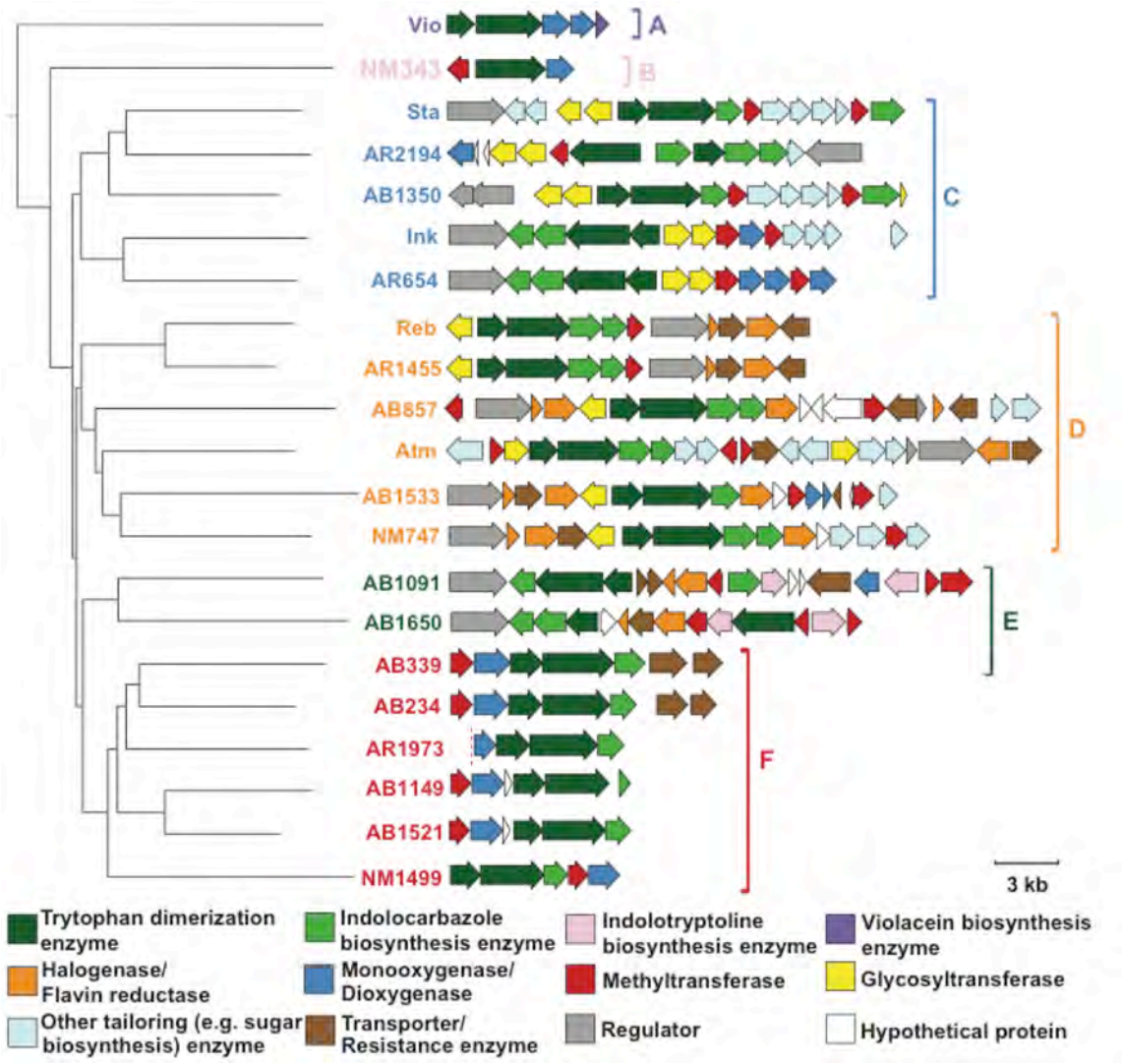


Figure 12. ClustalW-based phylogenetic tree based on culture-derived (Vio, Sta, Ink, Reb, Atm) and eDNA-derived (AB#, AR#, or NM#) CPA synthase genes. TD gene clusters are shown next to each CPA synthase gene. Six functionally distinct groups of TD clusters (A-F) are predicted based on the clustering of the tree. AR1973 is truncated by the vector.

monooxygenase (StaC/RebC) that is functionally and bioinformatically distinct between the two groups (Figure 13) (96, 97). Although the difference in chemical structure is subtle between the Group C and D classes of TD, the biological ramification is significant in that they are known to have different primary molecular targets: Group C

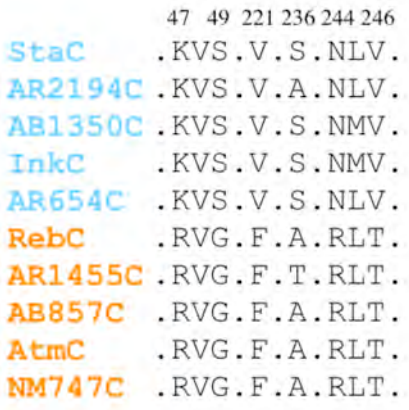


Figure 13. Protein sequence alignment of StaC/RebC-like monooxygenases. The six active site residues that have been identified as being responsible for the functional difference between Group C and Group D are K47/R46, S49/G48, V221/F216, S236/A231, N244/R239, V246/T241.

compounds are protein kinase inhibitors (48), while Group D compounds are DNA topoisomerase I inhibitors (49).

More than 100 TDs have been described from culture-based studies (45, 46). As the majority of these are not associated with a sequenced cluster, predicting whether a newly discovered TD cluster might encode for a novel metabolite is often challenging. Most known TDs are pyrrolinone indolocarbazole-based (Group C) compounds, making it particularly difficult to determine whether eDNA-derived Group C gene clusters (AR2194, AR654) encode for novel metabolites. In contrast, only a handful of maleimide indolocarbazole (Group D) metabolites are known. Group D eDNA-derived gene clusters (AB857, AB1533, TX747) all contain collections of genes that are predicted to allow them to encode for novel maleimide indolocarbazole-based TDs (*e.g.* additional halogenases and sugar tailoring enzymes).

The remaining Group B, E, and F contain no functionally characterized relatives, and their eDNA-derived gene clusters consist of biosynthetic genes that are

unprecedented in known TD biosynthetic pathways, suggesting that gene clusters in these groups could encode for new TD motifs. For this study, we thereby focused on the functional characterization of Group B, E, and F that were predicted based on CPA synthase gene phylogeny to represent novel TD classes.

Chapter 3: Characterization of novel tryptophan dimer classes

3.1 Heterologous expression strategies

Functional characterization of eDNA-derived natural product biosynthetic gene clusters requires heterologous expression, or the production of the pathway-encoding small molecules in a heterologous host. With increasing efficiency in DNA sequence and synthesis, heterologous expression is increasingly becoming more important in natural products research (98, 99). Its application is not just limited to metagenomics-based studies, but has also provided the means for the elucidation of compounds encoded in sequenced genomes (100), the characterization of natural product biosyntheses (101), modification of pathways for the generation of natural product derivatives (102), and optimization of product yield (103). However, despite its significance and the recent advances in its methodology, a reliable approach for heterologous expression that guarantees successful biosynthesis of the natural product does not currently exist. Thus, heterologous expression can be considered to be one of the largest bottlenecks in metagenomics-based natural product discovery efforts. To address the various factors that hinder the expression of gene clusters in a heterologous setting, the heterologous expression platform in our natural product discovery pipeline takes a multipronged approach that incorporates the use of a diverse set of bacterial hosts, as well as methodologies in genetic engineering to manipulate and synthetically refactor cryptic gene clusters (Figure 14).

One of the prominent reasons to the failure of small molecule production in a heterologous host is the host's inability to recognize foreign DNA. All bacterial heterologous hosts are restricted by their intrinsic abilities to recognize and utilize the

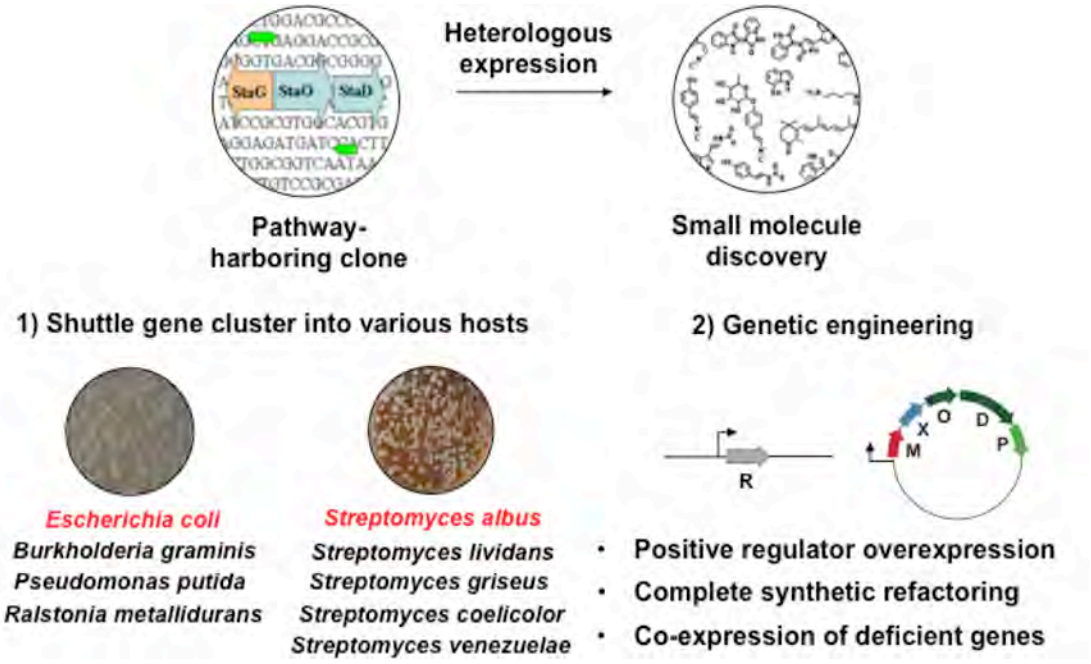


Figure 14. Overview of heterologous expression strategies used in our study: 1) Use of various hosts; 2) Genetic engineering by a) overexpression and co-expression of positive acting transcriptional regulators; b) induced expression of all biosynthetic genes within the cluster (synthetic refactoring); c) synthetic refactoring and co-expression of deficient genes.

foreign DNA captured within eDNA libraries (29). Factors that need to be compatible with the heterologous host include transcriptional elements (104, 105), such as regulatory elements, promoters, and ribosomal binding sites, as well as translational issues (106, 107), such as codon usage and protein folding chaperones. As such, the use of a heterologous producer that is the closely related to the native host should more likely lead to successful expression than a phylogenetically distant counterpart (98). However, because the native hosts are unknown for gene clusters isolated from eDNA, one viable strategy for the heterologous expression of these gene clusters should be to introduce

them into multiple heterologous hosts, thereby increasing the chance of small molecule production in at least one host (Figure 14; #1).

Based on previous metagenomic studies, the majority of soil samples are mostly populated by bacteria from the phylum *Proteobacteria* (108). This has thereby motivated previous studies in our laboratory to design a broad-host range vector pJWC1 (109), which was then used to construct multiple soil eDNA libraries in six different *Proteobacteria* (*Agrobacterium tumefaciens, Burkholderia graminis, Caulobacter vibroides, Escherichia coli, Pseudomonas putida, Ralstonia metallidurans*) for the parallel phenotype-based screening of natural products (30). Various positive hits were recovered from different hosts, but this (30) and various other studies have found the *Ralstonia* (109) and *Burkholderia* (110) spp., in particular, to be prolific *Proteobacteria* hosts for small molecule production. Considering that much of the soil microbiota consists of *Proteobacteria*, we predicted that a large portion of eDNA may be recognized by one of the previously established *Proteobacteria* hosts (30). The eDNA-derived TD gene clusters were therefore introduced into different *Proteobacteria* hosts (*E. coli*, *P. putida*, *R. metallidurans*, *B. graminis*) by means of retrofitting the pathway-harboring cosmids with the previously constructed broad-host range vector pJWC1 (109), followed by introduction of each pathway-harboring clone via electroporation. However, none of the eDNA-derived TD clusters produced any metabolites when natively introduced into the four *Proteobacteria* hosts.

Aside from *Proteobacteria*, actinomycetes, or more specifically bacteria from the genus *Streptomyces*, have been explored as heterologous hosts for biosynthetic gene clusters (111, 112). This is because they are long known to be prolific producers of

natural products (82, 83) and are thereby ample in cellular machineries that facilitate the expression of biosynthetic gene clusters and the production of the encoding small molecules. Moreover, based on the annotation of the biosynthetic genes and the unrelated flanking genes that are captured in the TD pathway-harboring cosmids (Appendix 1; refer to NCBI database with the corresponding A/N for flanking genes), most of the eDNA inserts harboring the TD gene clusters resemble *Streptomyces* spp. and other related actinomycetes bacterial species as their phylogenetic origin. Furthermore, a significant number of known TD compounds (*i.e.* rebeccamycin, staurosporine) have been produced in *Streptomyces* spp. as heterologous hosts (56, 57, 113, 114). *Streptomyces* spp. should thereby be suitable candidates as heterologous hosts for the eDNA-derived TD gene clusters.

To introduce the TD gene clusters into various *Streptomyces* spp. (*S. albus, S. lividans, S. coelicolor, S. venezuelae, S. griseus*), the TD pathway-harboring cosmids were retrofitted with the *E. coli*/*Streptomyces* shuttle vector pOJ436 (115) and subsequently integrated into the host chromosome by bacterial conjugation. The transformation of *S. albus* with the rebeccamycin-like pathway AR1455 resulted in the production of the expected compound, rebeccamycin, suggesting *S. albus* to be an appropriate heterologous host for TD gene clusters. In addition, in *S. albus*, robust production of clone-specific metabolites were observed for pathway AB1650, while limited small molecule production was detected for pathway AB1091, both of which are classified in Group E based on CPA synthase gene phylogeny (Figure 12).

In the case of AB1091, the level of small molecule production was insufficient for structural elucidation. For previous heterologous expression studies, it has been possible

to increase metabolite production from biosynthetic gene clusters by overexpressing positive acting transcriptional regulators found within gene clusters of interest (116, 117). A putative transcription factor from the AB1091 cluster, as predicted based on BLAST annotation, was therefore cloned into a Streptomyces expression vector (pIJ10257) (118) under the control of the strong constitutive ermE* promoter and co-transformed into *S. albus* with the cosmid AB1091 (Figure 14; #2a). The constitutive expression of the putative pathway-specific positive regulator using an artificial promoter resulted in increased production of clone-specific metabolites to a level that permitted their isolation and structural characterization. The functional characterization of the AB1650 and AB1091 gene clusters from Group E is described later in this chapter.

The AB1091 pathway represents one of the many examples of gene clusters that, at their native and unmanipulated state, do not yield their encoding natural products at sufficient levels for functional characterization in laboratory fermentation conditions. The presence of such "cryptic" gene clusters, or putative biosynthetic gene clusters with little or no small molecule production at their native state, are not surprising, since the expression of secondary metabolite gene clusters tends to be only necessary under specific conditions (*e.g.* iron limiting environment for siderophores) and are therefore normally tightly controlled within the native host (33, 119, 120).

The cryptic biosynthetic gene clusters are often repressed at the transcriptional level (121, 122), presumably because small molecule production level amplifies in each step of the process (*i.e.* transcription, translation, enzymatic catalysis) and therefore regulating the very first step provides the simplest and most responsive control. To overcome such transcriptional barrier, the silent gene clusters can be manipulated by

genetic engineering. One method, as described before for pathway AB1091, is to overexpress the putative transcriptional activator that is present in the gene cluster by cloning the activator gene under the control of an artificial promoter that the heterologous host can recognize (Figure 14; #2a).

However, in some cases, the putative regulator may be inactive, irrelevant, or simply not present in the gene cluster. The alternative genetic engineering-based approach is to thereby redesign the entire gene cluster, such that all of the putative biosynthetic genes in the pathway are detached from their native regulatory elements and placed under the control of well-defined genetic parts (Figure 14; #2b). This methodology, originally coined by the synthetic biology community as "synthetic refactoring," is increasingly being used in the natural products community for heterologous expression of biosynthetic pathways (98, 123, 124). We chose *E. coli* as the heterologous host for our synthetic refactoring pursuits because it is the most well-established host for this type of studies (125, 126). Complete synthetic refactoring becomes increasingly more difficult and cumbersome for large gene clusters, so this method was successfully used for the functional characterization of relatively small (<10 kb, ~5 genes) TD gene clusters, namely the AB339, AB234, AB1149, AR1973, AB1521, and NM1499 pathways from Group F and the NM343 pathway from Group B (Figure 12). For the NM343 pathway, an additional process was employed for its heterologous expression, where a biosynthetic gene that was predicted to be deficient was co-expressed with the synthetically refactored NM343 cluster (Figure 14; #2c). Details regarding the characterization of the Group F and B gene clusters are described later in this chapter.

3.2 Group E TD class: indolotryptoline – BE-54017

Group E classification of the CPA synthase gene phylogeny contains gene clusters AB1650 (127) and AB1091 (128). The first gene cluster that we investigated, AB1650 (*abe*), consisted of a complete set of conserved indolocarbazole biosynthetic genes (*abeO, D, C, P*), as well as two monooxygenases (*abeX1, X2*), three methyltransferases (*abeM1, M2, M3*), and a halogenase (*abeH*) (Figure 15). The presence of the two predicted monooxygenases was unprecedented in known TD biosynthetic pathways, suggesting that the AB1650 gene cluster likely encodes the biosynthesis of a novel TD substructure.

For heterologous expression, the native AB1650 pathway was introduced into *S. albus* by bacterial conjugation. Cultures of *S. albus* transformed with AB1650 produced one major clone-specific metabolite **1**, as well as four minor clone-specific metabolites **3-6** (Figure 16). The compounds were extracted from the aqueous bacterial culture broth by partitioning using an organic solvent (ethyl acetate), fractionated using normal phase flash chromatography, and purified using reverse phase HPLC.

The chemical structures of the purified compounds **1** and **3-6** were determined using a combination of HR-ESI-MS and 1-D and 2-D NMR data (Figure 17, Appendix 2, 3). HR-ESI-MS predicts the molecular formula of compound **1** to be $C_{23}H_{18}N_3O_5Cl$. The strong M+2 peak observed in the spectrum confirms the presence of a chlorine atom in the molecule. Based on the integration and chemical shifts of ^{1}H 1-D NMR, compound **1** consists of 7 aromatic protons (δ7.16, 7.26, 7.27, 7.52, 7.73, 8.09, 8.23), 2 hydroxyl protons (6.40, 6.67), and 3 sets of methyl protons (2.97, 4.09, 4.18). ^{13}C and HMQC NMR spectra further indicate the presence of 3 methyl carbons (δ25.3, 34.1, 62.8), 2

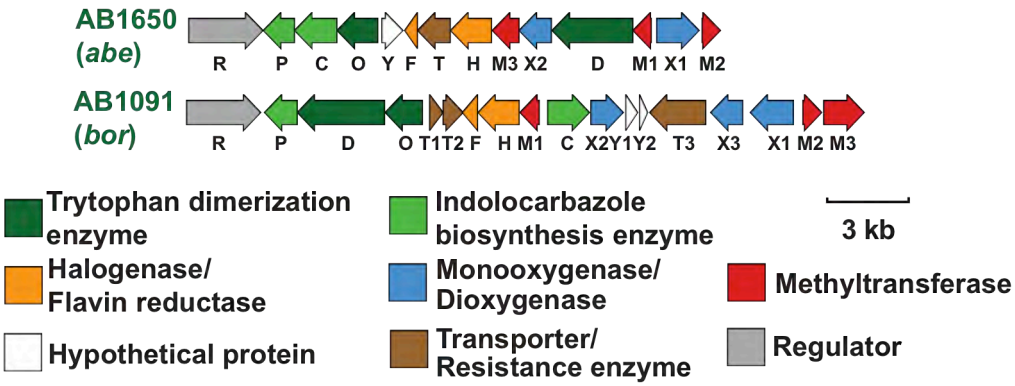


Figure 15. Gene annotation of AB1650 (*abe*) and AB1091 (*bor*) gene clusters from Group E.

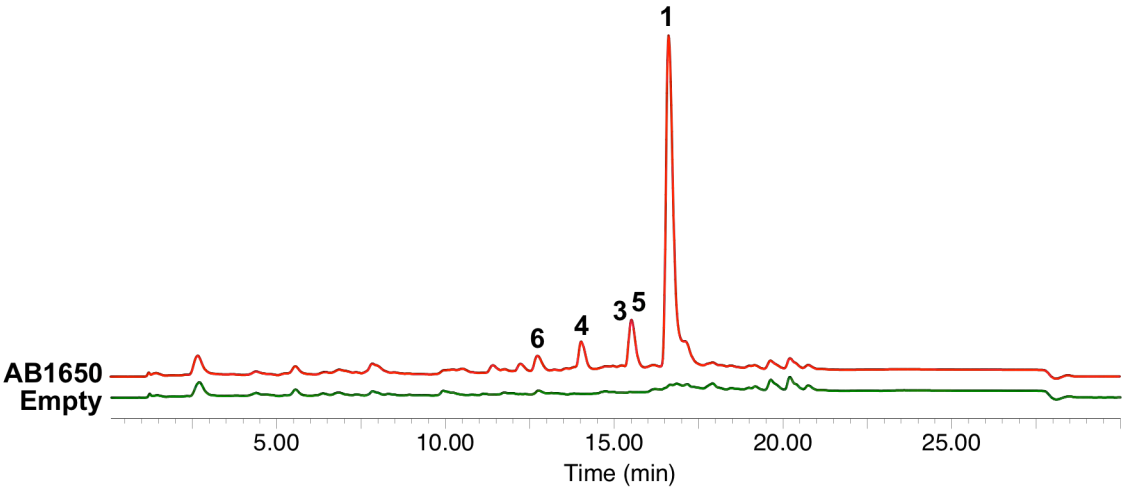


Figure 16. Analytical HPLC-UV chromatograms of culture broth extracts of *S. albus* harboring an empty vector as a negative control (green) and the AB1650 (*abe*) pathway. Compounds **1-6** represent the metabolites that are produced by the *abe* gene cluster.

oxygen-substituted tertiary carbons (δ75.6, 87.7), 7 disubstituted olefinic carbons (δ112.4, 116.7, 118.9, 121.0, 121.3, 123.7, 125.0), 9 trisubstituted olefinic carbons (δ104.6, 127.0, 127.2, 137.8, 123.4, 138.1, 114.7, 132.9, 139.3), and 2 carbonyl carbons (δ171.7, 174.7). Analysis of COSY and HMBC spectral data from **1** establishes three substructures (Figure 18; **A, B, C**).

Substructure A. The methyl protons (H_3-14, δ2.97) show HMBC correlations to the two carbonyl carbons (C-5, δ174.7; C-7, δ171.7). On the basis of empirical chemical

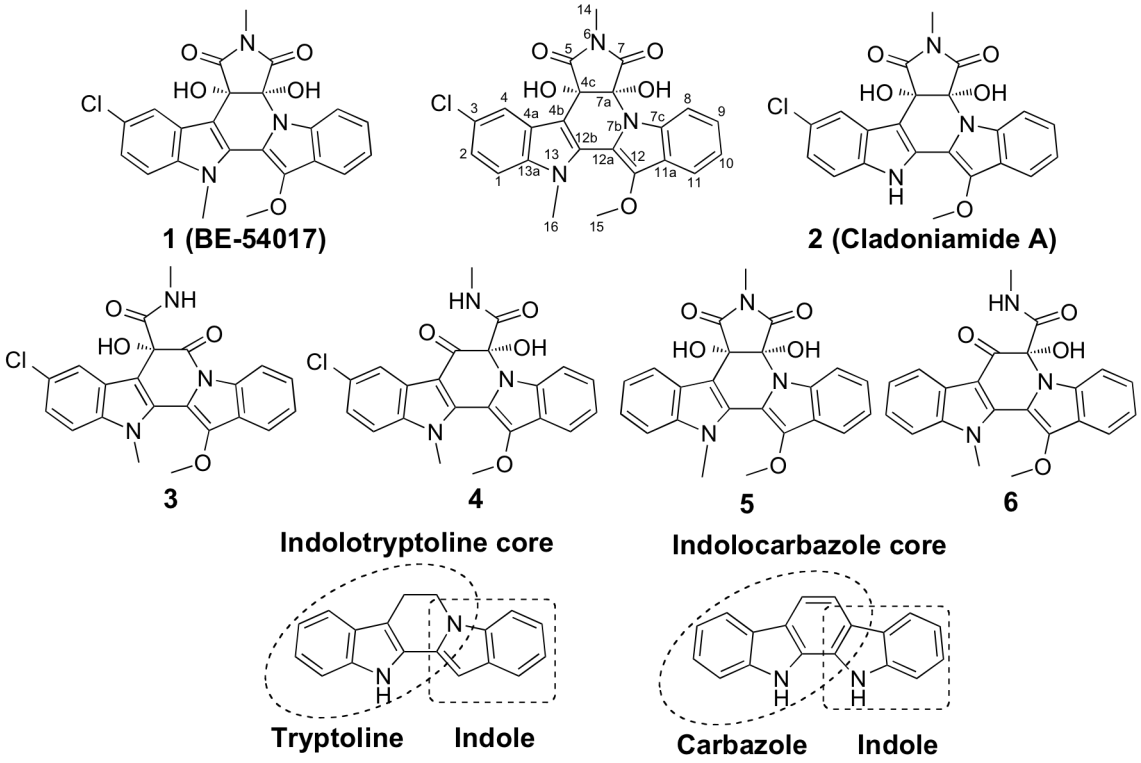


Figure 17. The eDNA-derived *abe* gene cluster encodes the biosynthesis of BE-54017 (**1**), as well as novel derivatives **3-6** as minor metabolites. BE-54017 (**1**) and cladoniamide (**2**) share an indolotryptoline core. Here we show that indolotryptolines arise from indolocarbazole precursors.

shift arguments (129), these two groups are separated by a nitrogen to form the *N*-methyl imido group. These protons also exhibit HMBC correlations to the two oxygen-substituted tertiary carbons (C-4c, δ75.6; C-7a, δ87.7). The two hydroxyl protons (δ6.40, 6.67) similarly show HMBC correlations to the two carbonyl carbons and the two oxygen-substituted tertiary carbons, thereby establishing the dihydroxysuccinimide substructure (substructure **A**; Figure 18).

Substructure B. ^{1}H-^{1}H COSY experiments define a two-carbon aromatic spin system (H-1/H-2) that can be placed adjacent to the trisubstituted olefinic carbons C-3 and C-13a by HMBC correlations from H-1 and H-2 to C-3 and C-13a. The absence of

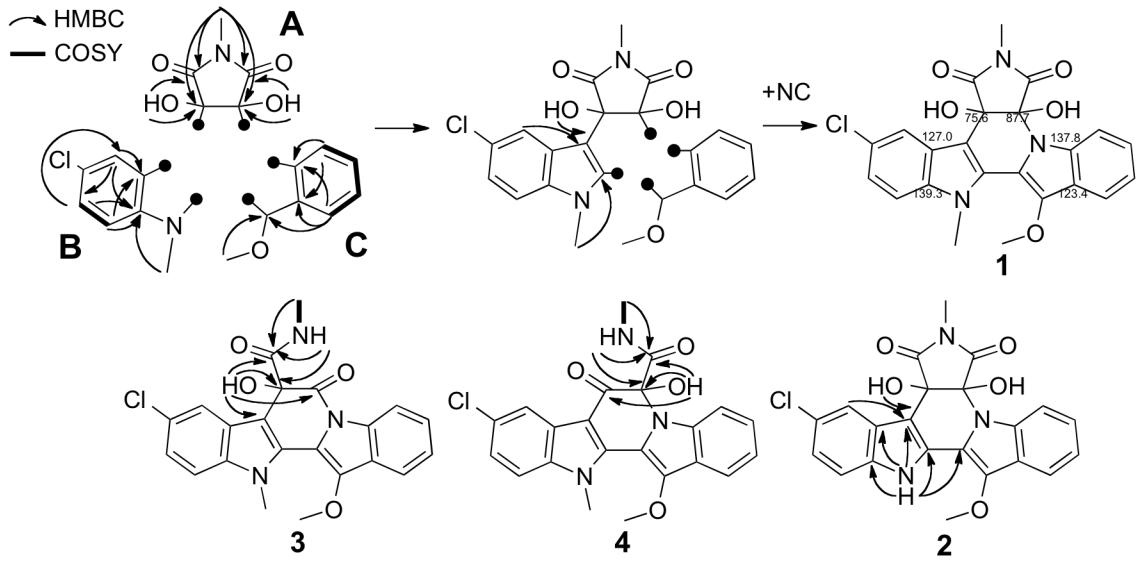


Figure 18. Key 2-D NMR (HMBC and COSY) correlations observed in the structural elucidation of **1-4**.

proton and carbon signals associated with C-3 aside from this four-carbon substructure defines the remaining C-3 substitution to be chloride. An HMBC correlation from H-4 to C-2 and C3 extends it into a five-carbon substructure. An additional HMBC correlations shown in Figure 18, in particular HMBC correlations from H-1 and H-4 to C-4a, establish the six-membered aromatic ring. The substructure is further expanded by an HMBC correlation from the methyl protons (δ4.18) to C-13a, elucidating the methylation of the indolic nitrogen (substructure **B**; Figure 18).

Substructure C. ^{1}H-^{1}H COSY establishes a four-carbon spin system (H-8, H-9, H-10, H-11), which is extended into a six-carbon aromatic ring by HMBC correlations from H-8 and H-11 to C-7c and C-11a. HMBC correlations from the methyl protons (δ4.18) to C-11a and C-12 and from H-11 to C-12 place the methoxylated C-12 adjacent to C-11a (substructure **C**; Figure 18).

HMBC correlations from the methyl H_3-16 to C-12b and from H-4 to C-4b extends substructure **B** to an indole and connects to substructure **A** by an HMBC correlation from OH-4c to C-4b. This leaves one carbon and one nitrogen atom to bridge substructure **A+B** to **C**, based on the molecular formula. The C-7a and C-7c carbons have higher chemical shifts than their symmetric counterparts C-4c and C-4a, indicative that they are surrounded by a nitrogen heteroatom (129). The placement of the nitrogen atom at N-7b adjacent to C-7a and C-7c and the placement of the remaining carbon at C-12a complete the structure determination of **1** to be the antitumor substance BE-54017. Although HMBC correlation signals were not observed for C-12a in compound **1**, we have later isolated compound **2**, a desmethyl derivative of **1**, which show HMBC correlations from NH-13 to C-13a, C-4a, C-4b, C-12b, and C-12a, validating the placement of the C-12a in **1**. Furthermore, the NMR, UV, and HR-ESI-MS data for **1** were identical to previously reported data for BE-54017 (130), confirming the chemical structure.

Minor metabolites **3** and **4** each differ from **1** by the loss of 28 mass units, corresponding to "CO." The NMR spectra for **3** and **4** are identical to **1** except for signals corresponding to substructure **A**. ^{1}H NMR spectra of **3** and **4** each displays a loss of one hydroxyl proton and a gain of one amido proton. A new COSY spin system is present between the amido proton and H_3-14 that now shows HMBC correlation to only one of the two carbonyl carbons (Figure 18). HMBC correlations from the C-4c hydroxyl proton to C-7a, C-5, C-4c, and C-4b in compound **3** and from the C-7a hydroxyl proton to C-7, C-7a, and C-4c in compound **4** allowed us the define the position of the *N*-methylamide in substructure **A** and elucidate their chemical structures (Figure 17). On the basis of

comparisons of HR-ESI-MS and 2-D NMR data, compounds **5** and **6** were determined to be the deschloro analogues of compounds **1** and **4**, respectively. Compounds **3-6** are novel natural products.

BE-54017 (**1**) and a related compound cladoniamide A (**2**) share a unique pentacyclic core structure consisting of an indole fused to a tryptoline moiety (Figure 17). In accordance to the nomenclature of the related indolocarbazole TD substructure, we named this pentacyclic core as the indolotryptoline TD substructure. Previous studies on TD biosynthesis have suggested that the indolotryptoline core might arise either from the oxidation of an indolocarbazole precursor or from an indolocarbazole-independent pathway (45, 131). To elucidate the origin of the indolotryptoline substructure and assign specific functions to the individual genes found in the *abe* gene cluster, we carried out a transposon mutagenesis study on cosmid AB1650. Individual transposon mutants were sequenced to identify clones with insertions in key biosynthetic genes (Figure 19). This collection of transposon mutants was then conjugated back into *S. albus*, and the major clone-specific metabolites found in the culture broth extracts of each mutant were structurally characterized.

Consistent with what is known about the biosynthesis of indolocarbazoles (Figure 8), transposon insertions in predicted indolocarbazoles biosynthetic genes resulted in either the absence of organic-extractable clone-specific metabolites (*abeO, D*) or the production of the known indolocarbazole intermediate 3-chloro-chromopyrrolic acid (*abeC, P*) (Figure 19). Homologues of the two predicted oxidoreductases, *abeX1* and *abeX2*, do not appear in any known TD biosynthetic gene clusters. Disruption of *abeX2* results in the accumulation of compound **8**, and disruption of *abeX1* leads to the

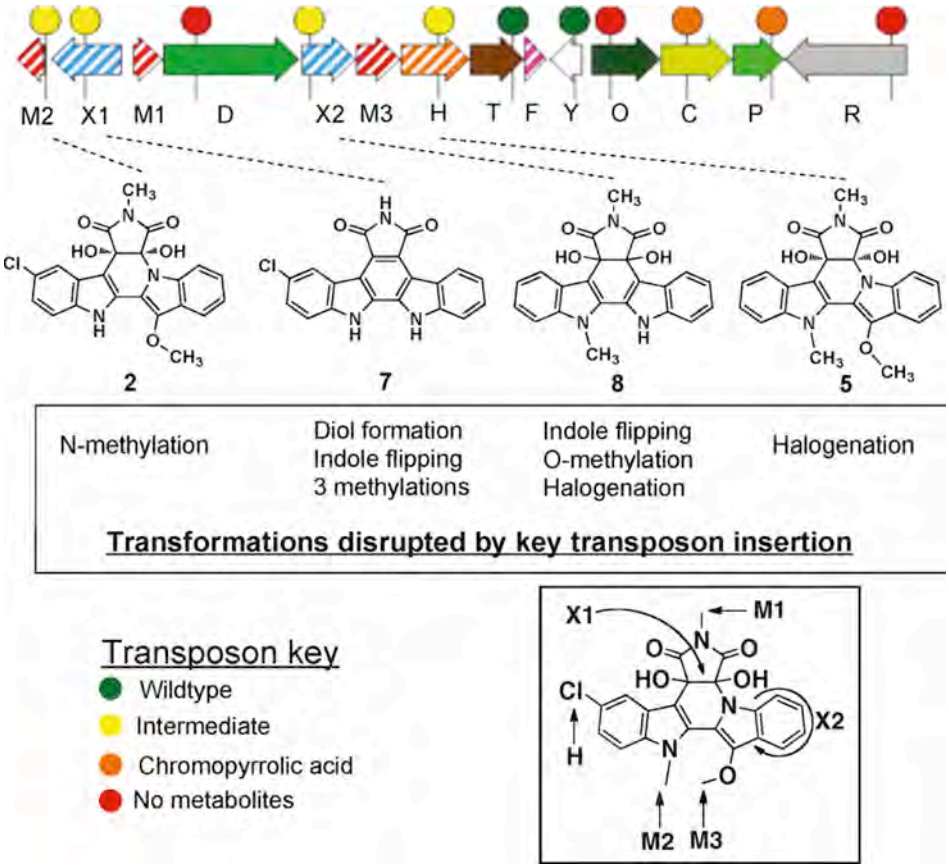


Figure 19. Major metabolites produced by select transposon mutants. Sites on BE-54017 where key biosynthetic enzymes are predicted to act are indicated in the inset.

accumulation of the simple indolocarbazole 3-chloroarcyriaflavin (**7**) as determined based on comparison with previously published spectral data (114). The isolation of compound **7** confirmed the indolocarbazole origin of BE-54017.

Based on HR-ESI-MS data, the molecular formula for compound **8** differs from compound **5** by "CH_2O". A comparison of the ^1H NMR spectra of **5** and **8** showed that **8** does not contain the C-12 methoxy protons and that it has a new NH hydrogen. Most of the HMBC correlations seen in **8** also appeared in **5**, but new HMBC correlations from the C-7a OH to C-7b and from the NH to C-7b, C-12a, and C-11a indicated that the indole in **8** has not been flipped.

Compound **8** contains the C-4c/C-7a diol seen in BE-54017, but has not undergone a rearrangement of the indolocarbazole core, while compound **7** contains neither the diol nor the flipped indole. AbeX1 and AbeX2 show homology to class-A flavoprotein monooxygenases, and therefore the oxidation reactions they carry out are predicted to proceed through epoxides (Figure 20) (132). The mutagenesis results coupled with homology arguments allowed us to construct a biosynthetic scheme for the formation of the indolotryptoline core in BE-54017 from indolocarbazole **7** (Figure 20). Consistent with the biogenesis as previously proposed (131), the diol is introduced by AbeX1 through the epoxidation of the C-4c/C-7a double bond. AbeX2 is then responsible for promoting the rearrangement of the indole via the introduction of a second epoxide at C-7b/C-12a. The opening of this epoxide is accompanied by fragmentation of the C-7a/C-7b bond followed by rotation of the indole around C-12a/C-12b and finally the formation of the C-7a/N-7b bond.

Whether the proposed epoxide hydrolysis and indole rearrangement reactions occur spontaneously, are catalyzed by the monooxygenases themselves, or are catalyzed by other enzymes is not clear. While AbeY shares the same α/β-hydrolase fold superfamily as most epoxide hydrolases (133), our studies suggest that it is dispensable in the biosynthesis of compounds **1** and **3-6**, as all five compounds are produced by the *abeY* knockout mutant. The appearance of small quantities of low-molecular-weight clone-specific metabolites in extracts from cultures of the *abeY* mutant suggests that while AbeY is not required, it may be involved in enhancing the efficiency of one or more biosynthetic transformations, similar to the reported function of the FAD-binding monooxygenase, RebC/StaC, in rebeccamycin/staurosporine biosynthesis (134).

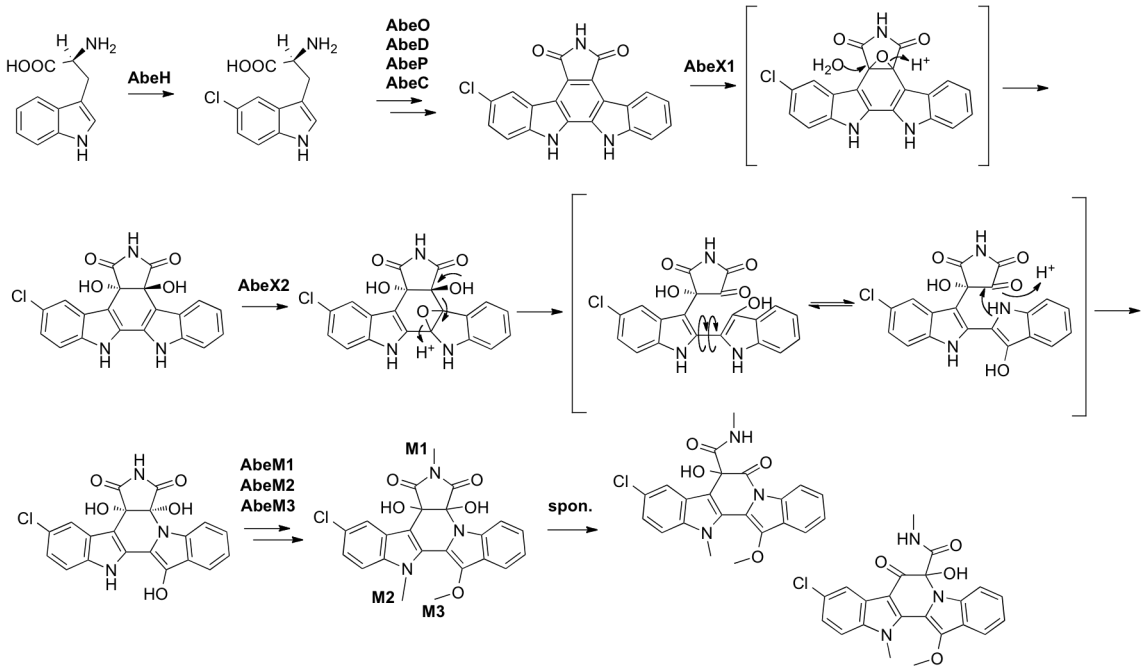


Figure 20. Two monooxygenases, AbeX1 and AbeX2, are responsible for the conversion of an indolocarbazole precursor into the indolotryptoline core of BE-54017. BE-54017 requires three methylations; the exact timing of these transformations is not known.

Alternatively, the role of AbeY may be complemented by the host's endogenous biosynthetic machinery.

A detailed accounting of the functionality that appears on the compounds produced in our transposon mutagenesis experiments allowed us to assign functions to the three methyltransferases and the predicted halogenase found in the *abe* gene cluster (Figure 19). Disruption of *abeH*, a homologue of the tryptophan halogenase found in the rebeccamycin pathway, led to the accumulation of deschloro derivative **5**, thereby confirming its role as a halogenase. The absence of the chloride substituent on compound **8** suggested that the transposon insertion in *abeX2* disrupts the expression of downstream genes in the same operon. The gene *abeM3* is positioned between *abeX2* and *abeH* and is

therefore predicted to be transcriptionally silenced in this transposon mutant. Since both the N-6 and N-13 methylations appear on **8**, *abeM3* is predicted to be responsible for the methylation of the C-12 hydroxyl in BE-54017. The regiospecificity of the two remaining *N*-methyltransferases, *abeM1* and *abeM2*, was inferred from the accumulation of the O-12,N-6 dimethylated derivative **2** (cladoniamide A) in the *abeM2* transposon mutant. AbeM1, AbeM2, and AbeM3 are therefore N-6, N-13, and O-12 specific methyltransferases, respectively (Figure 19 inset). No gene in the *abe* cluster could be linked to the hydrolysis of the *N*-methylsuccinimide, suggesting that this reaction is either carried out by the host or occurs spontaneously during the fermentation process.

Compounds **1-8** were assayed for antiproliferative activity against human colon cancer HCT116 cells. While **3-8** were not active below 8 μg/mL, **1** and **2** exhibited potent antiproliferative activities [IC_{50} (μg/mL): 0.079 for **1**, 0.0088 for **2**] (Table 1). Crystallographic studies have shown that the planar structure of indolocarbazoles is important for topoisomerase/DNA (rebeccamycin) and kinase (staurosporine) binding (48, 49). The C-4c/C-7a diol seen in indolotryptolines causes the *N*-methylsuccinimide to bend out of the bisindole plane, which may afford these compounds the ability to bind different cellular targets for their antitumor activity. Considerable progress in generating novel indolocarbazole analogues by combinatorial synthesis has been made, and the identification of AbeX1 and AbeX2 provides new tools for producing structurally and functionally diverse bisindole metabolites (114, 135).

In summary, our metagenomic natural product discovery pipeline has allowed us to clone and characterize the first indolotryptoline biosynthetic gene cluster and establishes indolotryptoline as an interesting lead structure for its antiproliferative

Table 1. IC_{50} data summary of BE-54017 compounds from the *abe* gene cluster (against human colon cancer HCT116 cells)

Compound	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
IC_{50} (μg/mL)	0.079	0.0088	20	21	8.7	29	13	11

activity. After the publication of this finding (127), the biosynthetic gene cluster of cladoniamides was discovered and reported by a different laboratory (136).

3.3 Group E TD class: indolotryptoline – borregomycins

The second gene cluster in Group E from cosmid AB1091 (*bor*) (128) resembles, but is noticeably distinct from, the BE-54017 (*abe*) gene cluster (Figure 15). As seen in the *abe* cluster, the *bor* cluster contains a complete set of conserved indolocarbazole biosynthetic genes (*borO, D, C,* and *P*), a single halogenase (*borH*), and homologs of the two oxidoreductase genes (*borX1, borX2*) that are known to encode the transformation of indolocarbazole precursors into indolotryptolines. Additionally, the *bor* cluster uniquely contains a third oxidoreductase gene (*borX3*) that is not seen in the *abe* cluster. Similar to the *abe* cluster, three methyltransferase genes (*borM1, M2,* and *M3*) are present in the *bor* cluster; however, *borM2* and *borM3* do not show significant sequence identity to any of the *abe* methyltransferases. The *bor* cluster was thus likely to encode for a new indolotryptoline-based metabolite.

In an effort to access the metabolite(s) encoded by the *bor* gene cluster, the eDNA clone AB1091 was conjugated into *S. albus*. However, most clone-specific metabolites were produced by *S. albus* at such low levels that it was not possible to isolate sufficient quantities for structural elucidation. Therefore, as previously mentioned, a putative positive acting transcription factor from the *bor* cluster, *borR*, was cloned into a

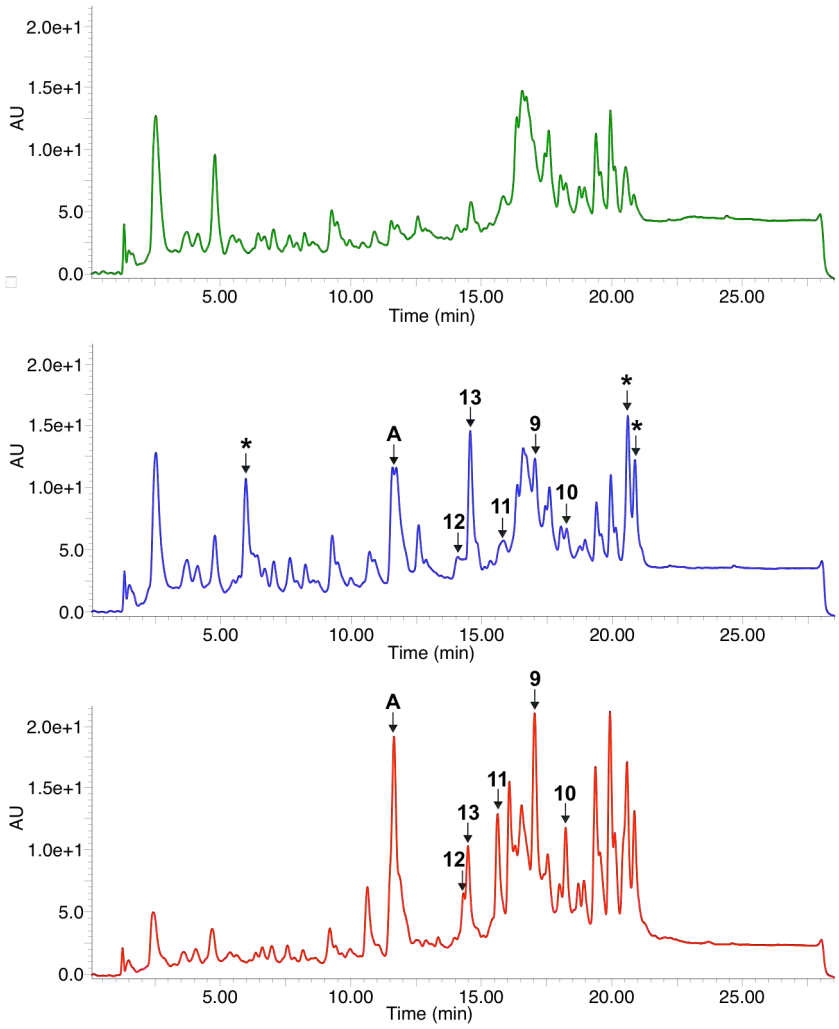


Figure 21. Analytical HPLC-UV chromatograms of culture broth extracts of *S. albus* harboring an empty vector as negative control (green), the *bor* pathway alone (blue), or *S. albus* harboring the *bor* pathway as well as the *borR* overexpression construct (red). Compounds **9-13** represent all of the detected metabolites that are specific to the *bor* cluster. The shunt product **13** accumulates with the *bor* cluster alone, while the final products **1** and **2**, accumulate at higher levels with *borR* overexpression. Peaks marked with an asterisk (*) are present in negative control and are thus not clone-specific. Peaks marked "**A**" are not present in negative control, but these low molecular weight compounds (179 and 195 mass units) are also seen in extracts from *S. albus* cultures harboring other indolotryptoline gene clusters and are therefore likely common shut products of indolotryptoline biosynthesis.

Streptomyces expression vector (pIJ10257) under the control of the strong constitutive ermE* promoter and co-transformed into *S. albus* along with the cosmid AB1091. Constitutive expression of *borR* in the *S. albus* background containing AB1091 resulted in increased production of clone-specific metabolites to a level that permitted their isolation and structural characterization (Figure 21).

In the presence of the *borR* overexpression construct, the *bor* gene cluster encodes for the production of five detectable clone-specific metabolites in *S. albus* (**9-13**) (Figure 22). The methodology for the structural elucidation of **9-13** was similar to that for compounds **1-8** with the combination of HR-ESI-MS and 1-D and 2-D NMR data (Figure 23, Appendix 2, 3). HR-ESI-MS predicts molecular formulas of $C_{23}H_{17}N_3O_4Cl_2$, $C_{22}H_{15}N_3O_4Cl_2$, $C_{21}H_{13}N_3O_4Cl_2$ for **10, 11,** and **12,** respectively. The strong M+2 and M+4 peaks observed in these MS spectra confirm the presence of two chlorines in each molecule. The molecular formula differences of "CH_2" and the similarity in their UV spectra suggest that the structures of **10-12** differ by their methylation patterns. This is confirmed by the differences in the number of methyl protons and methyl carbons in the 1-D ^1H and ^{13}C NMR spectra.

The numbers of unique chemical shifts seen in the ^1H and ^{13}C NMR spectra and the ^1H signal integration ratios indicate that **12** is symmetric with one set of methyl protons along the plane of symmetry (3H, δ2.98, s). Each half of **12** contains one hydroxyl proton (2H, δ5.58, brs), three aromatic protons (2H, δ7.14, dd; 2H, δ7.49, d; 2H, δ8.14, d) and one indolic amino proton (2H, δ10.76, brs). ^{13}C and HMQC NMR spectra indicate that **12** has one methyl carbon along the plane of symmetry (δ25.4). Each half of **12** contains one carbonyl carbon (δ175.9), one oxygen substituted tertiary carbon

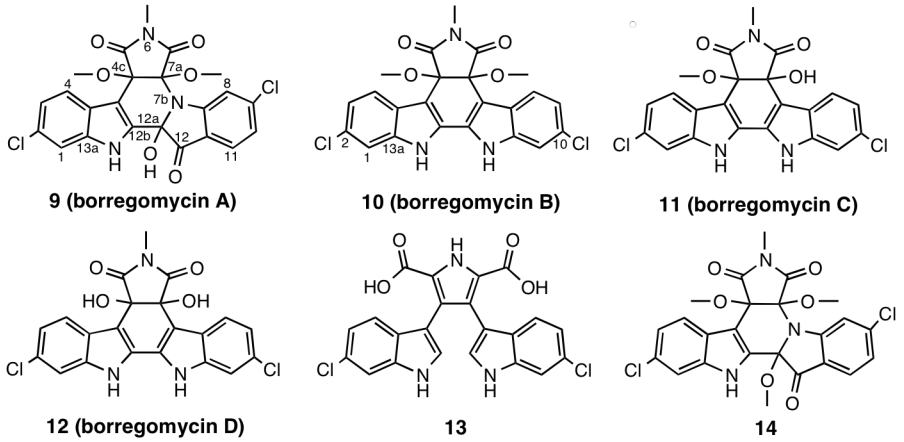


Figure 22. Borregomycins produced by the *bor* gene cluster. Indolotryptoline-based borregomycin A (**9**), dihydroxyindolocarbazole-based borregomycins B-D (**10-12**), dichlorochromopyrrolic acid (**13**), and *O*-methyl-borregomycin A (**14**) were isolated from *S. albus* harboring the eDNA-derived *bor* gene cluster.

(δ77.2), five tri-substituted olefinic carbons (δ110.1, 126.9, 128.4, 128.6, 138.7) and three di-substituted olefinic carbons (δ112.3, 121.8, 123.7). Analysis of COSY and HMBC spectral data from **12** establishes two substructures (**A** and **B**; Figure 23).

Substructure A. The methyl protons (H-14, δ2.98) on the plane of symmetry show an HMBC correlation to the carbonyl carbons (C-5/C-7, δ175.9). Based on symmetry and empirical chemical shift arguments, these two groups are separated by a nitrogen atom to form the *N*-methyl imido group. An HMBC correlation from the hydroxyl protons (δ5.58) to the oxygen substituted tertiary carbon (C-4c/C-7a, δ77.2) and the carbonyl (C-5/C-7, δ175.9) establishes the dihydroxysuccinimide substructure (**Substructure A**; Figure 23).

Substructure B. ^{1}H-^{1}H COSY experiments define a two-carbon spin system (H-8/H-9) that can be placed adjacent to C-7c and C-10 by HMBC correlations from H-8 and H-9 to C-7c and C-10. An HMBC correlation from H-11 to C-9 and C-10 extends this

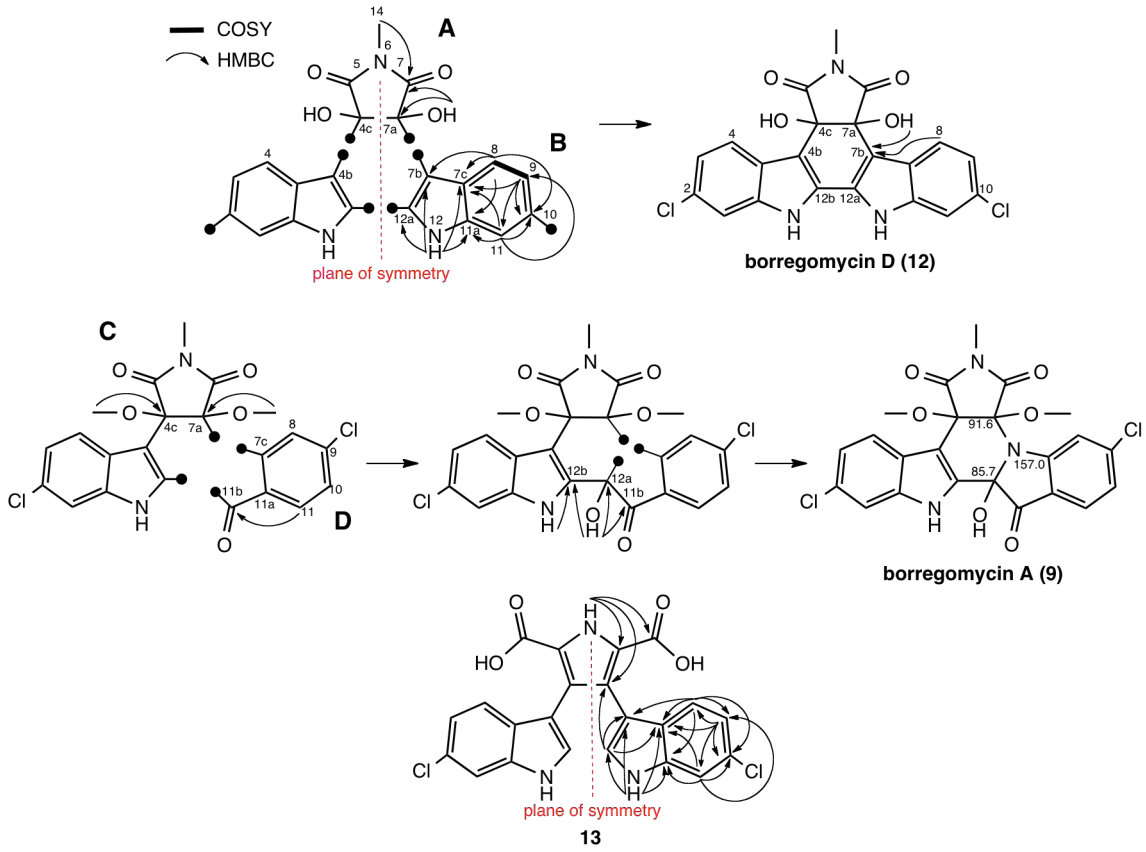


Figure 23. Key HMBC correlations observed in the structural elucidation of borregomycins.

four-carbon substructure into a five-carbon substructure. Additional HMBC correlations, in particular HMBC correlations from H-11 and H-8 to C-11a, establish the six membered aromatic ring. HMBC correlations from the indolic amino proton to the C-7c and C-11a in the aromatic ring along with the remaining two tri-substituted carbons, C-7b and C-12a, form the indole substructure observed on each side of the plane of symmetry (**Substructure B**; Figure 23). Substructures **A** and **B** are connected across the plane of symmetry by HMBC correlations from OH-4c/OH-7a and H-4/H-8 to C-4b/C-7b. The placement of the chlorides at C-2/C-10 completes the structure determination of **12**.

Similar NMR arguments to those used to establish the structure of **12** were used to define the di- and tri- methylated dihydroxyindolocarbazole structure **10** and **11**, which are additionally methylated on either the C-4c or the C-4c and C-7a hydroxyl groups, respectively. In each case, the position of the new methyl group was elucidated by the replacement of the hydroxyl proton signal with a new methyl proton signal that shows an HMBC correlation to either the C-4c or C-7a oxygen substituted tertiary carbon.

Based on HR-ESI-MS, the molecular formulas of compound **9** is $C_{23}H_{17}N_3O_6Cl_2$. The spectral data for **9** closely resembles that of the trimethylated **10**, but differs by the appearance of a carbonyl carbon (δ194.4) and an oxygen substituted tertiary carbon (δ85.7) in place of two olefins, as well as the appearance of a hydroxyl proton (1H, δ7.72, brs) in place of one of the indolic amino proton. Similar arguments to those used in the structure determination of **12** allow for the construction of the trimethylated dihydroxysuccinimide substructure (**Substructure C**; Figure 23). However, unlike the trimethylated **10**, one of the H-11 aromatic protons in **9** show an HMBC correlation to the new carbonyl carbon (C-11b) instead of a tri-substituted olefin, indicating that C-11a is substituted with C-11b to form **Substructure D** (Figure 23). HMBC correlations from the new hydroxyl proton (C12a-OH) to the new oxygen substituted tertiary carbon (C-12a), the carbonyl carbon (C-11b) from **Substructure D** and C-12b from **Substructure C** connect these two fragments as shown in Figure 23. The final structure of **9** is completed by placing the remaining nitrogen (N-7b) required by the molecular formula into the structure to form an indolotryptoline. The placement of the nitrogen satisfies the remaining unsaturations required by the unsaturation index, and it is supported by the downfield shift of the three carbon atoms neighboring the nitrogen in the final structure

(Figure 23). The disappearance of one indolic amino proton is consistent with the flipping of the indole moiety seen in the final structure.

Compounds **9-12** have been named borregomycin A-D after the geographic origin of the soil from which the *bor* cluster was cloned (Anza Borrego desert; AB library). Two unnamed compounds, **13** and **14**, were also isolated in this study (Figure 22). The structure of compound **13** ($C_{22}H_{13}N_3O_4Cl_2$) was solved by comparison of its spectral data with the previously published spectral data (137) for chlorinated chromopyrrolic acid and confirmed by inspection of the HMBC and COSY NMR data (Figure 23). As for compound **14** ($C_{24}H_{19}N_3O_6Cl_2$), the NMR spectral data for **14** only differ from that for **9** by the appearance of an additional methoxy group (δ3.01). Based on an HMBC correlation from these additional methyl protons to C-12a, **14** was determined to be the C-12a methoxy analog of **9** (Figure 22). This compound is not seen in the crude extract (Figure 21) and is therefore assumed to be an artifact that arises from methanolysis of borregomycin A during the isolation protocol.

On the basis of their final structures, sequence homology arguments, and precedents from previous studies of TD biosynthesis, the biosynthesis of borregomycins A-D can be rationalized as outlined in Figure 24. Similar to the biosynthesis of rebeccamycin and BE-54017, the biosynthesis of the borregomycins is likely initiated by the chlorination of L-tryptophan by the halogenase BorH, followed by the construction of an indolocarbazole using the set of conserved indolocarbazole biosynthetic enzymes BorO, D, P and C. The resulting 2,10-halogenation pattern is not seen in any previously reported indolotryptoline or indolocarbazole based metabolite. This is consistent with the fact that BorH is more closely related to known tryptophan 6-halogenases than halogenases found

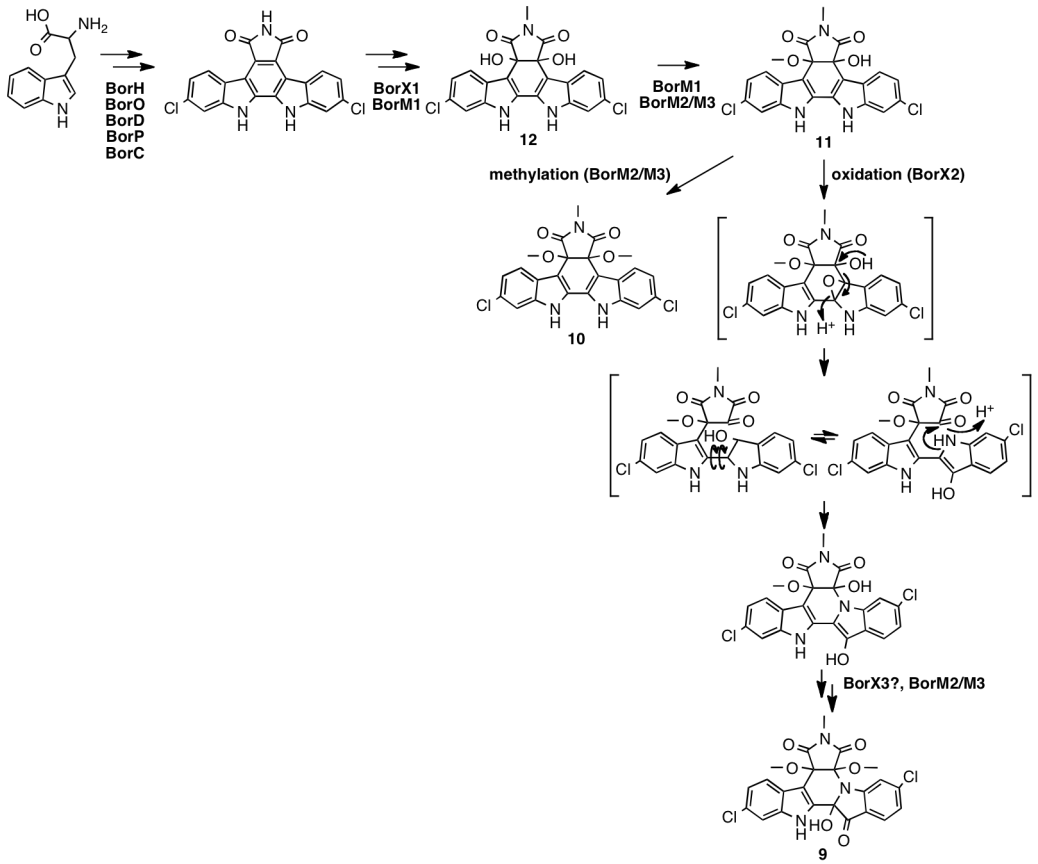


Figure 24. Proposed scheme for the biosynthesis of the borregomycins. The borregomycins are predicted to arise form a branched biosynthetic pathway with one branch of the pathway giving rise to the indolotryptoline-based metabolite (**9**) and a second branch giving rise to dihydroxyindolocarbazole-based metabolites (**10-12**). Compound **13** is a known shunt product in tryptophan dimer biosynthesis.

in other TD gene clusters. On the basis of sequence homology to enzymes from BE-54017 biosynthesis, oxidoreductase BorX1 is predicted to install the C-4c/C-7a hydroxyl groups and putative *N*-methyltransferase BorM1 is predicted to methylate N-6 to generate borregomycin C (**12**). One of the two remaining methyltransferases unique to the *bor* cluster, collectively referred to as BorM2/BorM3, is then predicted to *O*-methylate **12** on either the C-4c or C-7a hydroxyl to yield dimethylated **11**. The regiospecificity of this *O*-methylation reaction is not known.

The concurrent production of both borregomycin A (**9**) and borregomycin B (**10**) from **11** can be rationalized by the existence of a branch in the biosynthetic scheme where **11** either initially undergoes oxidation by BorX2 or further methylation by BorM2/M3. Should BorM2/M3 act directly on **11** to generate the trimethylated borregomycin B (**10**), transformation of the indolocarbazole into an indolotryptoline by BorX2 is inhibited as a result of both the C-4c and C-7a hydroxyls being blocked with methyls, rendering **10** as the terminal product of one branch of the *bor* pathway. If however the C-7a hydroxyl of **11** is available for deprotonation, BorX2 can promote the rearrangement of the indole via epoxide-driven fragmentation of the C-7a/C-7b bond followed by rotation of the indole around the C-12a/C-12b axis and formation of the C-7a/N-7b bond. The resulting indolotryptoline intermediate is then predicted to undergo methylation on the C-7a hydroxyl (BorM2/M3), followed by oxidation across the C-12/C-12a bond to yield borregomycin A (**9**).

Branched biosynthetic pathways are quite common in natural product biosynthesis. In most cases, however, they either generate collections of compounds from the same structural family or relatives of a single major product that are labeled as intermediates, degradation products or shunt products (138, 139). In rarer cases, for example in methymycin/pikromycin biosynthesis in *Streptomyces venezuelae* (140) and saliniketal/rifamycin biosynthesis in *Salinispora arenicola* (141), pathways have been found to produce terminal biologically active metabolites that are representative of different structural classes.

Borregomycins were assayed for cytotoxicity against a panel of model organisms (Table 2). Both borregomycin A (**9**) and borregomycin B (**10**) exhibit low micromolar

Table 2. Cytotoxicity data summary of the borregomycins.

Organism	9 (A)	10 (B)	11 (C)	12 (D)	13	14
Human HCT116 IC_{50}, μM	1.2	1.4	1.9	3.9	48	1.1
S. aureus USA300 MIC, μg/mL	>25	0.20	0.39	3.1	>25	>25
B. subtillis Sr168 MIC, μg/mL	>25	0.20	1.6	3.1	>25	>25
E. coli EC100 MIC, μg/mL	>25	>25	>25	>25	>25	>25
S. cerevisae W303 MIC, μg/mL	>25	>25	>25	>25	>25	>25

antiproliferative activities against human colon cancer HCT116 cells, comparable to the activity reported for the extensively studied indolocarbazole rebeccamycin (IC_{50}: 0.72 μM) (142). Borregomycin B (**10**) is also a Gram-positive antibiotic, however, at the highest concentration tested, borregomycin A (**9**) did not inhibit the growth of any of the bacteria we tested (Table 2). The *bor* pathway therefore not only encodes for metabolites from distinct structural families, but also metabolites with distinct bioactivities.

Close structural relatives of indolotryptolines, including tryptoline-based (143), indolo-β-carboline (fascaplysin)-based (144) and indolocarbazole-based (47) compounds have all been explored as protein kinase inhibitors. The final products of the two *bor* biosynthetic pathway branches, borregomycin A (**9**) and B (**10**), were therefore tested for inhibitory activity against a panel of 59 diverse disease-relevant kinases (KinaseProfiler, Millipore). Borregomycin B did not exhibit an IC_{50} of less than 10 μM against any kinases in the panel. At 10 μM, the most inhibited kinases PRAK, IGF-1R, and PI3KCδ showed residual activities of 66%, 67%, and 69%, respectively (Appendix 4). At the same concentration, borregomycin A exhibited a sub-10 μM IC_{50} against a single kinase, CaMKI (Figure 25). When borregomycin A was assayed against a more extensive panel of CaMKI-related kinases, it most strongly inhibited the CaMKIIδ kinase (IC_{50}: 3.4 μM;

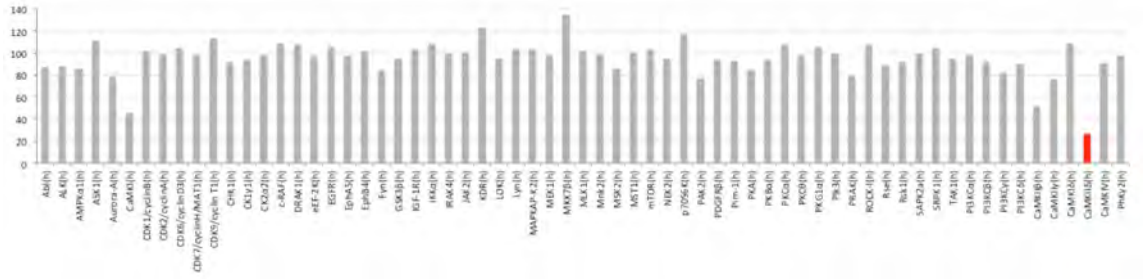


Figure 25. KinaseProfiler results. Residual protein kinase activity in the presence of 10 μM borregomycin A is shown for a diverse set of disease-relevant kinases (n = 2).

Appendix 4). Elevated levels of CaMKII, which is a Ca^{2+}/calmodulin-dependent serine/threonine kinase that is involved in various signal transduction pathways, have been implicated in diseases such as heart failure (145) and cancer (146). As borregomycin A is active at comparably low micromolar levels against both HCT116 cells and CaMKIIδ *in vitro*, the inhibition of CaMKIIδ may be a key contributing factor to its *in vitro* cytotoxicity.

Fascaplysin, a simple indolo-β-carboline, inhibits Cdk4-D1 kinase, however its planar structure also renders it a DNA intercalator (147). This has inspired the synthesis of nonplanar fascaplysin analogs in an attempt to eliminate the DNA intercalating property, thereby reducing its toxicity to normal cells (148). Similarly, the planar borregomycin B appears to be a general cytotoxin, while borregomycin A, which is rendered nonplanar by BorX2 catalyzed indole rearrangement followed by oxidation, appears to show more specificity in both the cell-based cytotoxicity assay and the kinase activity assay. The observed kinase inhibitory activity of borregomycin A suggests that oxidized indolotryptolines may warrant further investigation as potential lead structures for the development of novel protein kinase inhibitors.

Based on the presence of a novel oxidoreductase BorX3 in the *bor* gene cluster, we initially predicted the C-12/C-12a oxidation in borregomycin A to be catalyzed by BorX3 (Figure 24). However, recent mutational studies with the cladoniamide gene cluster (*cla*) suggests that the C-12/C-12a oxidation arises spontaneously in the absence of a methyltransferase that caps the C-12 hydroxyl (OH-12) moiety (149). Since the *bor* pathway lacks the OH-12 *O*-methyltransferase, this suggests that the C-12/C-12a oxidation occurs spontaneously and is not dependent on the BorX3 oxidoreductase.

While the *abe* and *cla* pathways that lack the C-4c and C-7a hydroxyl (OH-4c and OH-7a) *O*-methyltransferases yield indolotryptoline-based compounds with the hydrolyzed *N*-methylsuccinimide ring (**3, 4, 6**), compounds detected from the *bor* pathway lacks these structures. It has been found that these hydrolyzed indolotryptoline structure are unstable and undergo further spontaneous hydrolytic/oxidative degradation (150). Therefore, while the *bor* pathways permits oxidation across C-12/C-12a in the absence of OH-12 *O*-methyltransferase, the methyl capping of the C-4c and C-7a hydroxyls by the OH-4c and OH-7a *O*-methyltransferases in the *bor* pathway not only introduces a branch point in the pathway to produce compounds **9** and **10**, but also protects the structural integrity of the borregomycins from the hydrolysis of the *N*-methylsuccinimide ring.

The characterization of the *abe* gene cluster from the previous study establishes indolotryptoline as an interesting lead structure for its antiproliferative activity. However, most bacterial TD natural products that have been characterized to date are members of the indolocarbazole family, and indolotryptoline-based TD compounds have seldom been seen (46). Here we show that homology-based screening, coupled with the phylogenetic

profiling of eDNA-derived gene clusters, allows for the directed discovery of borregomycins, novel members within previously rare families (indolotryptoline) of bioactive natural products.

3.4 Group F TD class: carboxy-indolocarbazole

Group F gene clusters (84) are comprised of single operons containing three conserved indolocarbazole biosynthesis genes (*espO, D, P*), a predicted methyltransferase (*espM*) and a FAD-binding monooxygenase (*espX*) (Figure 12, 26). Although Group F gene clusters were recovered from all three metagenomic libraries (AB234, AB339, AB1149, AB1521, AR1973, TX1499) (Figure 12), no clusters with the same gene content were found in the NCBI database of sequenced bacterial genomes. The TD cluster from clone AB339, which we have called the *esp* cluster, was selected as a representative member of Group F for heterologous expression studies.

Initial heterologous expression efforts, including the introduction of the *esp* cluster into model bacterial hosts and induced expression of the *esp* operon under a T7 promoter in *E. coli* did not yield any detectable clone specific small molecules. Therefore, we chose to synthetically refactor the *esp* cluster by individually cloning each *esp* gene in front of a T7 promoter and inducing the expression of the *esp* genes in *E. coli* using the Novagen Duet vector system.

As homologs of *espO, D* and *P* are seen in a number of well-studied TD indolocarbazole biosynthetic gene clusters, their functions are predictable (134, 151). Co-expression of these three genes in *E. coli* resulted in the low level production of the expected indolocarbazole-based intermediates (**18-20**), indicating that *esp* is, in fact, a TD gene cluster (Figure 26A, *espODP* 24 hr). Addition of *espM*, a predicted

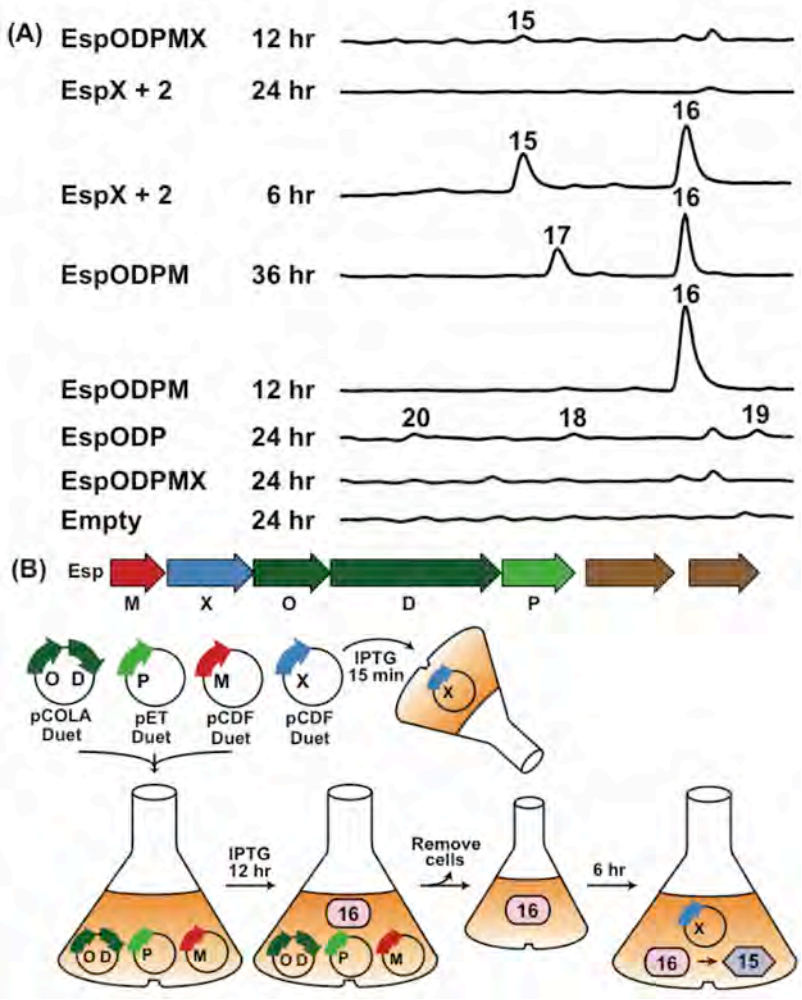


Figure 26. A) HPLC-UV traces of organic extracts from *E. coli* cultures expressing various combinations of *esp* genes. "+ **16**" refers to the addition of compound **16** in the form of spent medium from EspODPM cultures. **B**) Schematic of the method used for the complete refactoring of the *esp* cluster.

methyltransferase gene, to *espODP* led to the appearance of compound **2** (*espODPM* 12 hr). After extended incubation periods, either in culture broth or as a purified compound in DMSO, **16** spontaneously oxidized to **17** (*espODPM* 36 hr). Interestingly, when *espX*, a predicted FAD-binding monooxygenase gene, was co-expressed with *espODPM,* no clone specific metabolites were detected in the culture broth extract (*espODPMX* 24 hr), mimicking the result from the expression of the native *esp* operon in *E. coli*.

Suspecting that the product of the entire *esp* operon might be rapidly degraded, we investigated the EspX-catalyzed transformation reaction by feeding compound **16** in the form of spent culture broth from cultures of *E. coli* expressing *espODPM* to EspX-expressing *E. coli* cultures (Figure 26B). Within a narrow time window (<6-8 hr) we observed the accumulation of a new compound (**15**) in concert with the disappearance of **16** (Figure 26A, EspX + 2, 6 hr). After longer incubations, neither **15** nor **16** could be detected in the culture broth. Retrospectively, we re-examined the *espODPMX* monoculture at shorter time points and were able to detect very small qualities of **15** in these *E. coli* cultures (Figure 26A, EspODPMX 12hr). Ultimately, the temporal control over individual *esp* gene expression that was possible in our refactoring study permitted the isolation of a natural product that would otherwise have been too transiently present to identify. Compound **15** was purified from the culture broth of *espX* expressing cultures fed with spent "*espODPM*" culture broth, while compounds **16** and **17** were isolated from *E. coli* cultures expressing *espODPM*.

The structures of compounds **15-17** were elucidated using a combination of HR-ESI-MS and 1-D and 2-D NMR data. The NMR spectra of compound **17** were similar to those of the known indolocarbazole, K252c (staurosporine aglycone) (134) with two major differences. First, the proton signal from H-7 of compound **17** integrates to 1H instead of 2H and is shifted farther downfield than the corresponding proton signal from K252c. Second, **17** has an additional carbonyl carbon and a set of proton and carbon signals corresponding to a methyl group in its NMR spectra that are not seen in the spectra of K252c. Based on HR-ESI-MS, **17** has the molecular formula $C_{22}H_{15}N_3O_3$, which differs from the molecular formula of K252c by the loss of a proton and the gain

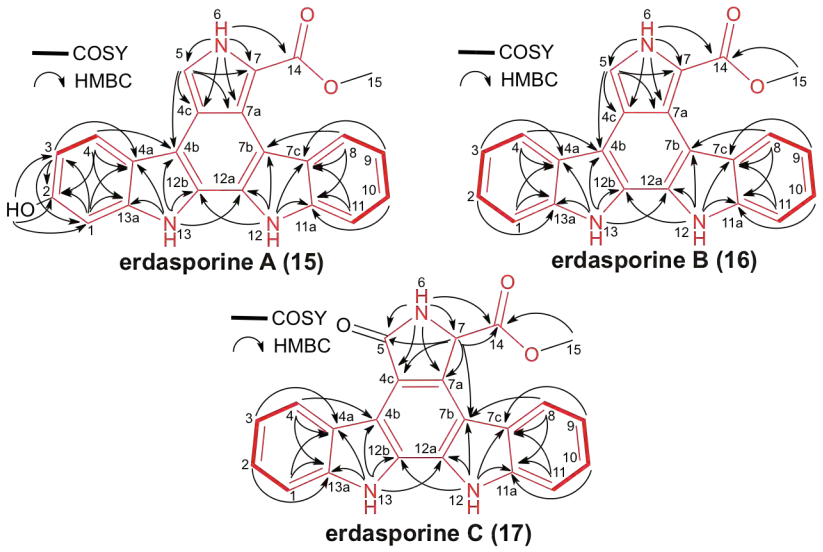


Figure 27. 2-D NMR correlations observed for the structural determination of erdasporine A-C (**15-17**).

of a methylcarboxy (CH_3CO_2) unit, thereby suggesting that compound **17** is a modified version of K252c with a methylcarboxy unit substituted at C-7. Extensive 2-D NMR correlations as shown, especially HMBC correlations from H_3-15 to C-14 and from H-7 to C-14, confirm the chemical structure of **17** (Figure 27).

Based on HR-ESI-MS, compound **16** has the molecular formula $C_{22}H_{15}N_3O_2$, which differs from the molecular formula of **17** by the loss of an oxygen atom. The NMR spectra of **16** closely resemble **17**, but differ by (i) the replacement of the C-7 tertiary carbon (δ60.6) seen in **17** with an olefinic carbon (δ113.0) and (ii) the replacement of the C-5 carbonyl carbon (δ173.4) seen in **17** with an olefinic carbon (δ115.4). These differences are accompanied by the loss of the proton signal (δ5.94) seen at C-7 in **17** and the gain of a proton signal (δ8.27) at C-5. These changes indicate that the oxygen at C-5 of **17** is replaced by a proton such that compound **16** is the deoxygenated form of **17**.

This final structure is supported by the extensive 2-D NMR correlations as shown (Figure 27).

Compound **15** has a UV spectrum that closely resembles that of **16**, but **15** has the molecular formula $C_{22}H_{15}N_3O_3$, which differs from that of **16** by the addition of an oxygen atom. The NMR spectra of **15** are similar to that of **16**, but differ by the replacement of the H-2 aromatic proton seen in **16** (δ7.31) with a hydroxyl proton (δ9.22). The downfield shift of C-2 from δ123.7 in **16** to 154.2 and the HMBC correlations from OH-2 to C-1, C-2, and C-3 suggest that compound **15** is the C-2 hydroxylated form of **16.** This chemical structure is confirmed by the extensive 2-D NMR correlations as shown (Figure 27).

Compounds **15-17** have novel methylcarboxylated indolocarbazole structures and were named erdasporine A-C, respectively (Figure 28). The erdasporines were found to be potent cytotoxins with low μM activity against bacteria and human cell lines (Figure 28).

Based on our heterologous expression studies and previous studies on indolocarbazole biosynthesis with well-characterized transient indolocarbazole intermediates (134, 151, 152), erdasporine biosynthesis is proposed to initially proceed like many other indolocarbazole pathways, with EspO, D, and P giving rise to dicarboxy-indolocarbazole (**21**) from two tryptophans (Figure 29). In the absence of additional enzymes, this intermediate is known to spontaneously decarboxylate and oxidize to form three indolocarbazole products (**18-20**). In many previously characterized indolocarbazole pathways, an FAD-binding monooxygenase (*e.g.* StaC/RebC) directs the formation of a single indolocarbazole product (monooxygenated **18** for Group C clusters

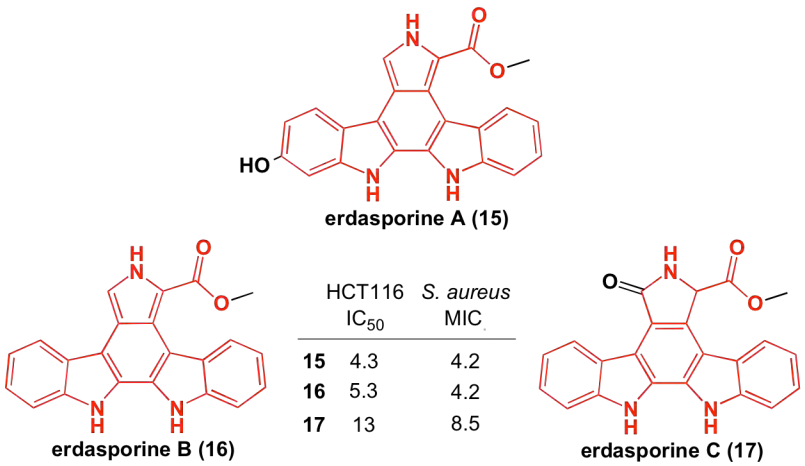


Figure 28. Chemical structure and cytotoxicity data of erdasporine A-C (**15-17**) encoded by the *esp* gene cluster. The carboxy-indolocarbazole core that is representative of Group F is colored in red. Cytotoxicity (μM) against human HCT116 cells and *Staphylococcus aureus* is shown.

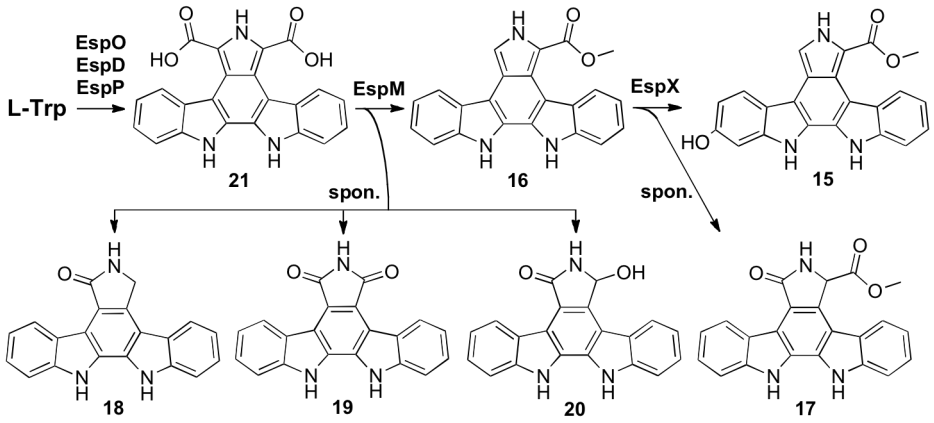


Figure 29. Proposed biosynthetic scheme for the erdasporines. Key steps include methylation of the dicarboxy-indolocarbazole (**21**) by EspM to generate **16** and the hydroxylation of **16** by EspX to give **15.**

and dioxygenated **19** for Group D clusters). In the *esp* cluster, methylation by EspM appears to preclude these spontaneous oxidative decarboxylation events, resulting in the formation of **16** and, with time, **17**. In the presence of EspX, compound **16** undergoes oxidation to form **15** as the product of the *esp* cluster.

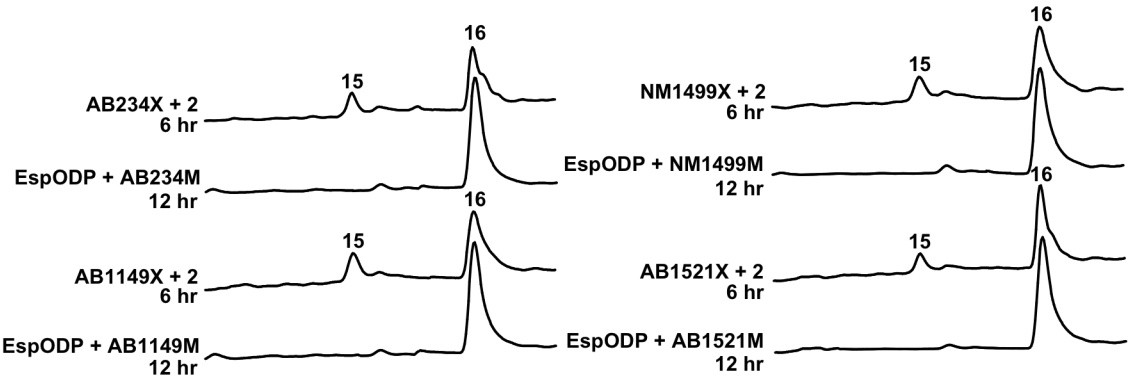


Figure 30. HPLC-UV traces of organic extracts from *E. coli* cultures expressing the indicated EspM-like methyltransferase and EspX-like monooxygenase in the EspODP background. "+ **16**" refers to the addition of compound **16** in the form of spent medium from EspODP culture co-expressing each pathway-specific methyltransferase.

The remaining eDNA-derived Group F clusters are predicted to contain the same five genes as the *esp* cluster. While EspODP homologs are functionally equivalent across all characterized TD clusters, the exact role of the EspM-like methyltransferases and the EspX-like monooxygenases in each Group F gene cluster was not certain. The remaining four complete Group F clusters were therefore characterized by expressing each unique EspM-like methyltransferase (AB234M, AB1149M, NM1499M, AB1521M) and unique EspX-like monooxygenase (AB234M, AB1149M, NM1499M, AB1521M) gene in the EspODP expression system that was used to characterize the original AB339 *esp* gene cluster. In each case, the expression of the pathway-specific methyltransferase resulted in the accumulation of **16** and the subsequent expression of the monooxygenase led to the production of **15** (Figure 30). Based on these studies, all Group F clusters were determined to be functionally equivalent to the *esp* cluster.

The functional characterization of the *esp* cluster defines a new TD group (Figure 12, Group F) that is based on a carboxy-indolocarbazole core substructure (Figure 28).

Bioinformatics analysis of the genes surrounding eDNA-derived Group F clusters suggests that they likely originate from diverse species within the phylum actinobacteria (Appendix 1; refer to NCBI database with the corresponding A/N for flanking genes). Although *esp*-like clusters were found in all three soil eDNA libraries examined, neither sequence nor functional analyses of cultured bacteria have identified this family of TDs, suggesting that these secondary metabolites are likely largely produced by as yet uncultured microbes.

3.5 Group B TD class: bisindolylmaleimide

Group B gene cluster from cosmid NM343 (*mar*) (153) is predicted to contain three or four biosynthetic genes: a CPA synthase (*marB*), a dioxygenase (*marC*), and a methyltransferase (*marM*) (Figure 12, 31). In addition, a *vioE* homolog gene (*marE*) is present immediately downstream of *marC*, which may also be involved in small molecule production. The dioxygenase MarC is unprecedented in known TD biosynthetic pathway and might direct the formation of a novel TD core structure.

For initial heterologous expression efforts, the native *mar* cluster was introduced into a variety of model hosts for expression studies, but no clone-specific metabolites were detected in the culture broths. In an effort to address potential transcriptional inefficiencies of *mar* gene cluster promoters in these hosts, we synthetically refactored the gene cluster by individually cloned the *mar* biosynthetic genes in front of T7 promoters and introduced these constructs into *E. coli*. Unfortunately, this also failed to result in the production of any detectable clone-specific small molecules by *E. coli* cultures (Figure 32, MarBCEM).

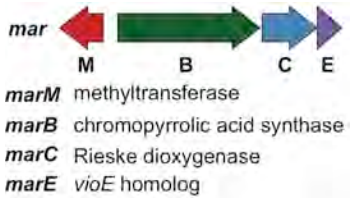


Figure 31. eDNA-derived *mar* biosynthetic gene cluster that encodes for methylarcyriarubin.

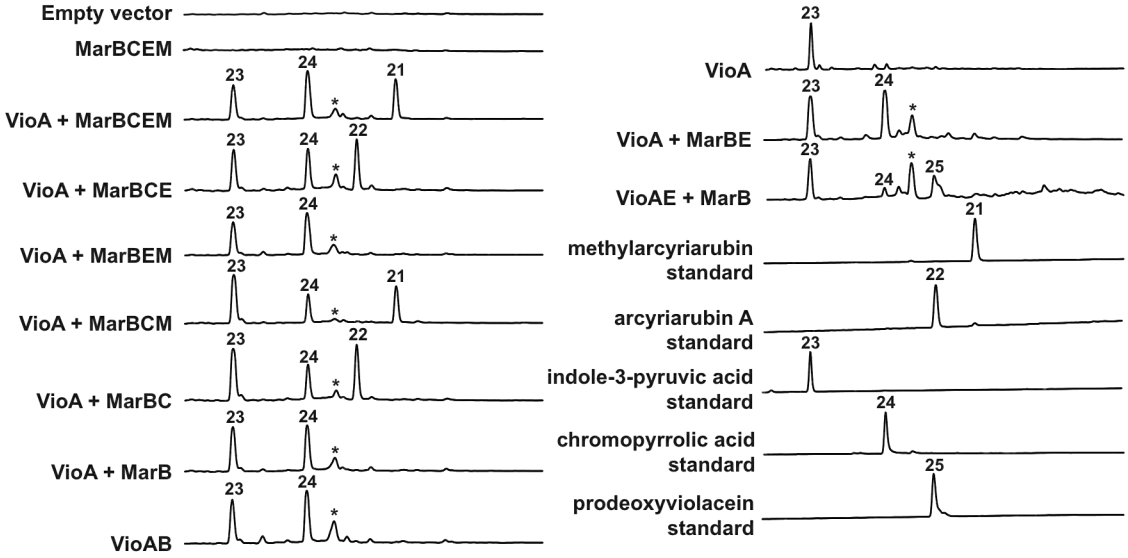


Figure 32. HPLC-UV traces of culture broth extracts from *mar* gene cluster expression studies in *E. coli*. IPA imine is reported to undergo spontaneous deamination to form IPA (**23**). The peak marked with an asterisk (*) is an uncharacterized by-product of the co-expression of an IPA imine synthase with a CPA synthase in *E. coli*.

Functionally characterized CPA synthases from other TD biosynthetic clusters have been found to accept oxidized tryptophan (IPA imine), but not tryptophan itself, as a substrate (154). Neither the *mar* cluster nor the *E. coli* genome contains an IPA imine synthase homolog. Therefore, if MarB functions as a CPA synthase as predicted by its high sequence identity to known CPA synthases, an IPA imine synthase would have to be supplied *in trans* for *mar* biosynthesis to proceed in a heterologous expression setting. A number of sequenced bacterial genomes contain isolated predicted IPA imine synthase

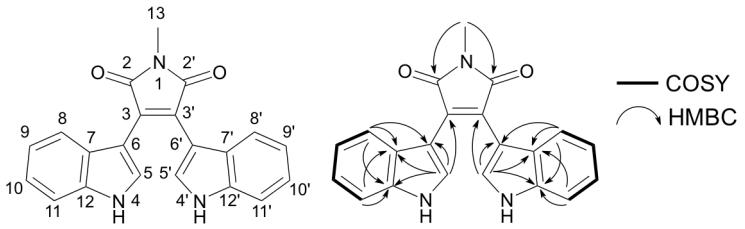


Figure 33. Numbering scheme and correlations observed in the HMBC and ^{1}H-^{1}H COSY NMR spectra of **21.**

genes, suggesting that IPA imine production may be encoded outside secondary metabolism in a variety of bacteria. Therefore, we co- expressed the IPA imine synthase *vioA* from the violacein cluster (155) with the rest of the *mar* biosynthetic genes in *E. coli*. This resulted in the production of a clone-specific metabolite (**21**) (Figure 32, VioA + MarBCEM), which we had not observed in any previous TD studies, along with the expected TD intermediates, IPA (**23**) and CPA (**24**) (152, 154).

Compound **21** was purified from large-scale cultures of *E. coli* transformed with the VioA + MarBCEM expression constructs. The UV, HR-ESI-MS, and NMR (Figure 33) data of **21** was consistent (156, 157) with the structure of the methylated bisindolylmaleimide methylarcyriarubin (**21**). Although methylarcyriarubin has been made synthetically, to the best of our knowledge, this is the first reported case of methylarcyriarubin being isolated as a natural product.

To elucidate the role of the individual *mar* biosynthetic genes in the biogenesis of methylarcyriarubin (**21**), culture broth extracts from *E. coli* strains expressing different combinations of the *mar* biosynthetic genes were characterized in detail. In *E. coli* lacking *marM* (Figure 32, VioA+MarBCE), arcyriarubin A (**22**), the desmethyl form of methylarcyriarubin (**21**), was produced along with **23** and **24**. Without the dioxygenase *marC* (Figure 32, VioA+MarBEM; VioA+MarB), accumulation of **23** and **24** was

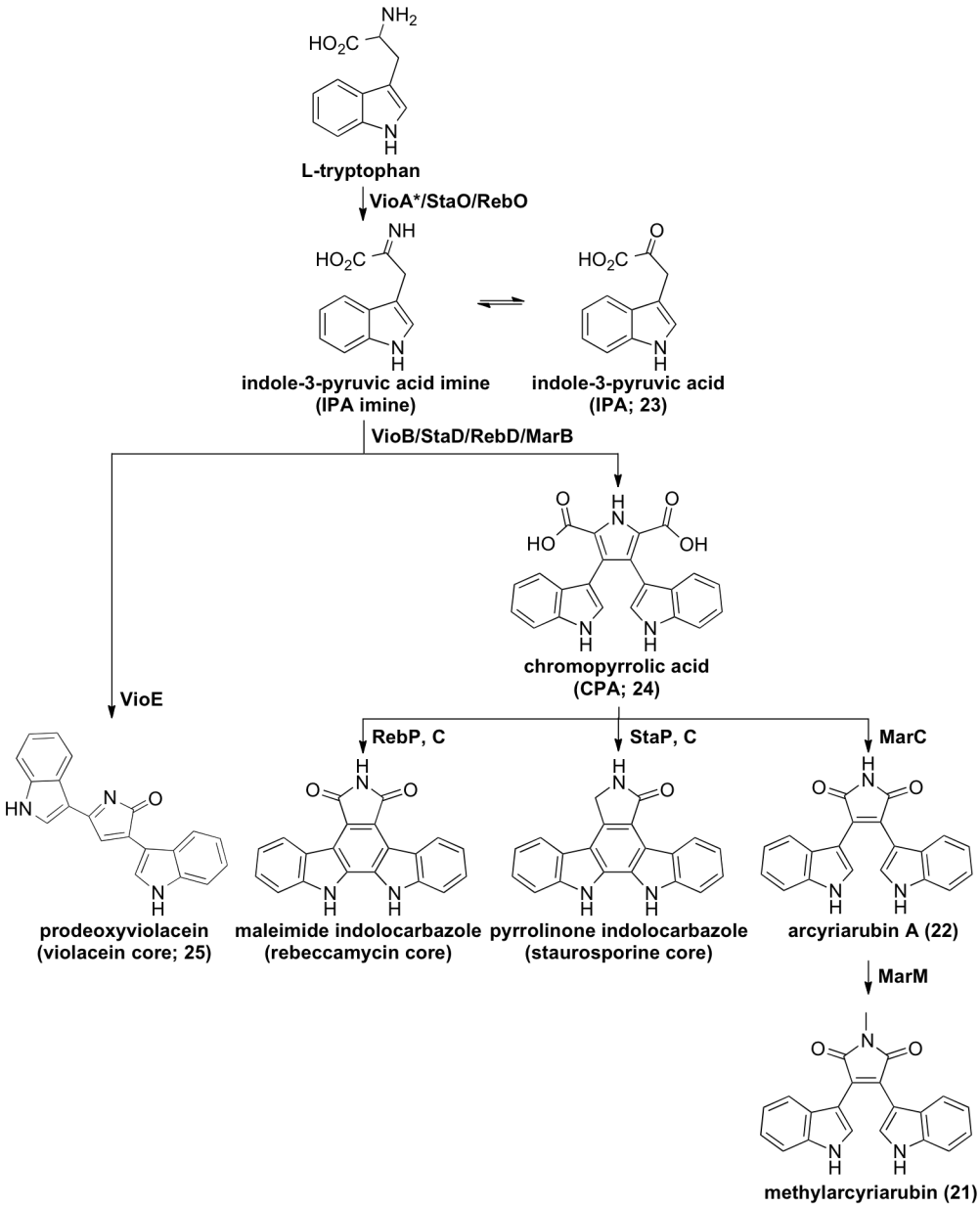


Figure 34. Biosynthesis of methylarcyriarubin alongside with that of other known bacterial tryptophan dimers, namely violacein and indolocarbazoles. VioA-like IPA imine synthase (*) is predicted to be encoded in the endogenous host's genome.

observed. Replacement of MarB with the characterized CPA synthase from the violacein pathway (VioB) resulted in an identical metabolic profile (Figure 32, VioA+MarB; VioAB). No difference in metabolic profile was observed between *E. coli* expressing all

the *mar* biosynthetic genes (Figure 32, VioA+MarBCEM) and a strain lacking the predicted *vioE* homolog, *marE* (Figure 32, VioA+MarBCM). Based on these expression studies, the biosynthesis of methylarcyriarubin is predicted to share the same two initial biosynthetic transformations as all other biosynthetically characterized bacterial tryptophan dimers (Figure 34). Specifically, tryptophan is first oxidized to IPA imine by an IPA imine synthase that is found outside the *mar* cluster in the endogenous host's genome. MarB then dimerizes IPA imine to give CPA. The *mar* biosynthetic pathway is then predicted to diverge from known tryptophan dimer pathway in that MarC appears to function as a bisindolylmaleimide synthase by converting CPA into arcyriarubin A. MarM is predicted to methylate the bisindolylmaleimide to yield methylarcyriarubin (**21**).

MarC is functionally similar to RebC from rebeccamycin biosynthesis in that they both produce a maleimide moiety from CPA. RebC is an FAD-binding monooxygenase that reacts in tandem with RebP to produce maleimide indolocarbazole (96, 151, 158). Based on co-crystallization studies (97, 159-161), the likely substrate of RebC was found to be 7-carboxy-K252c, which is produced by RebP via the C-5 and C-5' aryl-aryl coupling of CPA. This led to a proposed mechanism for RebC involving hydroxylation of 7-carboxy-K252c at the α-carbon of the carboxyl group to facilitate decarboxylation and yield a pyrrole-diol moiety, followed by an oxidative rearrangement to generate the maleimide (Figure 35). While MarC is also responsible for generating a maleimide moiety, it is predicted, based on sequence homology, to be a Rieske type dioxygenase. Rieske type dioxygenase contains a [2Fe-2S] iron-sulfur cluster, instead of FAD, as the cofactor (162). In MarC dependent bisindolylmaleimide biosynthesis, we propose that MarC hydroxylates the C2-C3 and C2'-C3' olefins of CPA in two successive oxidations

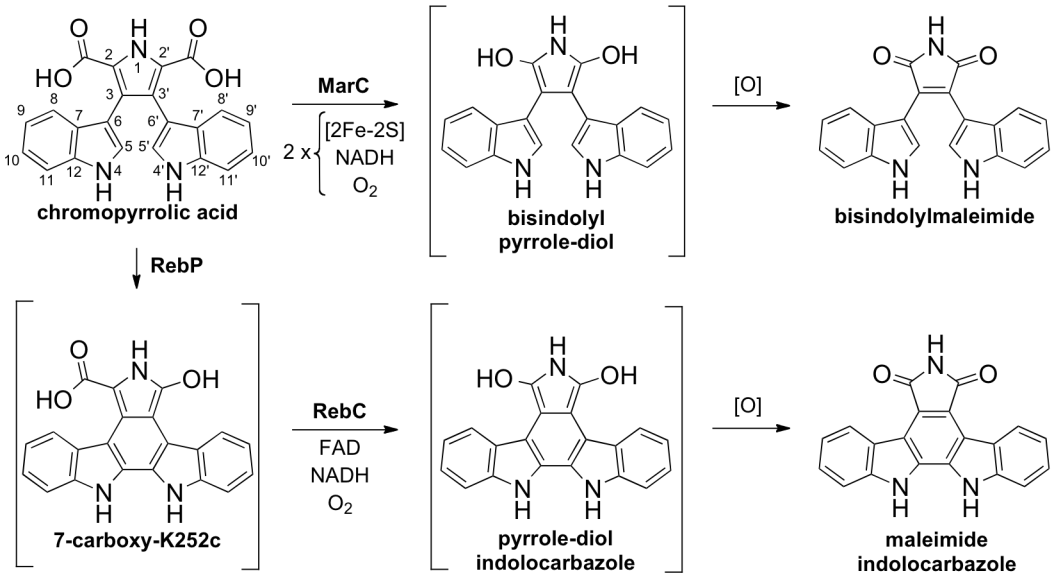


Figure 35. Comparison of the proposed enzymatic oxidative mechanism between bisindolylmaleimide and indolocarbazole in the biosynthesis of a maleimide moiety.

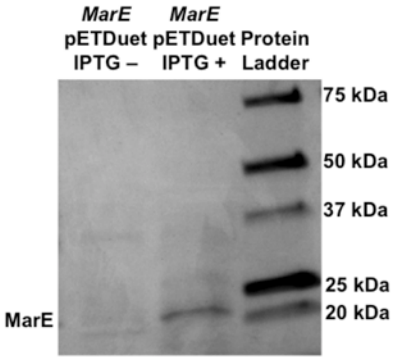


Figure 36. Soluble protein extract of *marE*/pETDuet harboring *E. coli* with or without IPTG induction (200 rpm, 37 °C, 2 hours after OD_{600} ~0.5). IPTG-dependent expression of a protein with the size corresponding to MarE (~22 kDa) is observed.

to facilitate two decarboxylations and generate, without relying on a second enzyme, the pyrrole-diol that can similarly undergo oxidative rearrangement to yield the maleimide (Figure 35). While our data supports this simple model, the involvement of unknown host factors cannot be ruled out.

Based on our *in vivo* analyses, no function can yet be assigned to MarE, a predicted VioE homolog. Gel analysis of the soluble protein extract of *marE* harboring *E. coli* confirms that soluble MarE is produced in this system (Figure 36). In violacein biosynthesis, VioE is predicted to produce prodeoxyviolacein (**25**) by "interacting" with a transient intermediate of unknown structure that is produced by the violacein CPA synthase (Figure 34) (94, 163, 164). VioE is a small protein (191 aa) that lacks any functionally characterized homologs, any known catalytic residues or any recognized cofactor-binding motifs. Accordingly, the mechanistic details of its role in violacein biosynthesis remain unclear. MarE shares high sequence identity to VioE, however in vivo MarE cannot complement the function of VioE in the production of prodeoxyviolacein (**25**) (Figure 32, VioA+MarBE versus VioAE+MarB). Whether MarE is inactive in the *mar* cluster, plays a role that is redundant in *E. coli*, or functions in the biosynthesis of an as yet unidentified metabolite is subject for further investigation.

Bisindolylmaleimide natural products share a common 3,4-di-1H-indol-3-yl-1H-pyrrole-2,5-dione core structure (165) (Figure 37, red). The discovery of the simplest bisindolylmaleimide natural product, arcyriarubin A (166), from a slime mold (*Arcyria denudate*) spearheaded the extensive study of this class of compounds, with more than 2,400 and 4,000 bisindolylmaleimide-related references in the PubMed and SciFinder databases, respectively. Bisindolylmaleimide analogs, with activities in cancer (167-169), diabetes (170), cardiovascular (171) and neurodegenerative (172) disease models have now been synthesized, some of which have advanced into clinical trials (173-175) (Figure 37A). By the heterologous expression of the *mar* cluster, we hereby functionally characterized the first bisindolylmaleimide biosynthetic pathway.

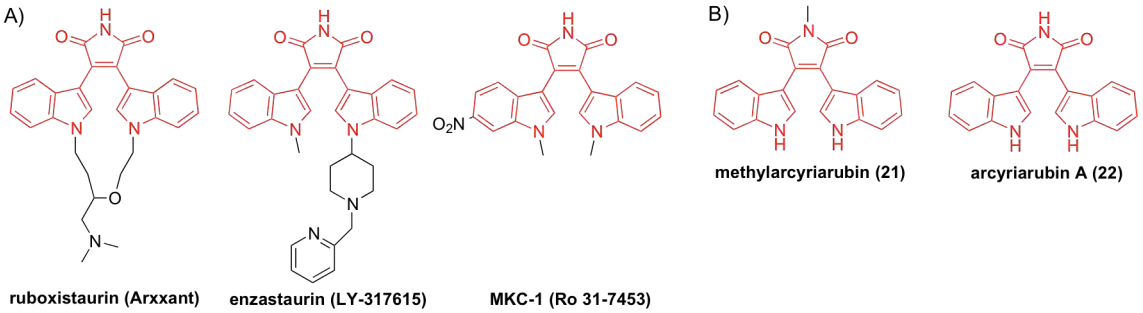


Figure 37. Bisindolylmaleimide compounds, with the 3,4-di-1H-indol-3-yl-1H-pyrrole-2,5-dione core structure colored in red. **A)** Synthetic derivatives that have entered into clinical trials for drug therapy. **B)** Compounds that are encoded by the eDNA-derived *mar* biosynthetic gene cluster.

A diverse collection of bisindolylmaleimides have been chemically synthesized and tested for bioactivity, with a particular focus on kinase inhibitory activity (176, 177). Arcyriarubin A (**22**) (Figure 37B) is a submicromolar inhibitor of protein kinase C (IC_{50}: 0.55 μM) (178). Interestingly, the addition of the *N*-methyl to give methylarcyriarubin (**21**) (Figure 37B) abolishes protein kinase C inhibitory activity (IC_{50}: > 100 μM) (178), but leads to activity against mitogen-stimulated protein kinase $p70^{s6k}/p85^{s6k}$ (IC_{50}: 8 μM) (179). A number of synthetic bisindolylmaleimide derivatives, including ruboxistaurin, enzastaurin, and MKC-1, have undergone or are currently in clinical trials as potent and specific kinase inhibitors for use as cancer and diabetes therapies (Figure 37A) (173-175).

Indolocarbazole tryptophan dimers, which differ from bisindolylmaleimides by the coupling of the C-5 and C-5' indole carbons, have also been extensively explored as kinase inhibitors (47). The additional C-C coupling forces indolocarbazoles to bind in a planar conformation, while the more flexible bisindolylmaleimides have been observed to bind in a nonplanar conformation (180, 181). This conformational flexibility is believed

to be responsible for trends observed in tryptophan bioactivities. Bisindolylmaleimides tend to be less potent, but more specific kinase inhibitors compared to their indolocarbazole counterparts (178, 182). Known indolocarbazole biosynthetic enzymes have been used, both *in vitro* in chemoenzymatic synthesis (183) and *in vivo* in combinatorial biosynthesis (113, 114), to generate many un-natural indolocarbazole analogs. Considering the significant clinical relevance of the bisindolylmaleimides, the identification of the *mar* biosynthetic gene cluster should facilitate the generation of additional collections of potentially pharmaceutically relevant tryptophan dimers using biosynthetic approaches.

3.6 Expansion of bacterial tryptophan dimer biosynthetic scheme

Using various heterologous expression strategies illustrated here, including the shuttling of pathways in multiple hosts, overexpression of predicted pathway-specific transcriptional regulators, complete synthetic refactoring of gene clusters, and the co-expression of predicted deficient genes, we functionally characterized nine of the 14 unprecedented gene clusters (AB1650, AB1091, AB339, AB234, AR1973, AB1149, AB1521, NM1499, NM343), resulting in the elucidation of three novel TD families that consist of biologically active natural products including BE-54017, cladoniamide A, borregomycins, erdasporines, and methylarcyriarubin. The dimerization of tryptophan by the tandem action of an IPA imine synthase and a CPA synthase generates a tremendously versatile intermediate (CPA) that appears to serve as a substrate for the biosynthesis of a diverse array of dimer substructures. The discovery of three novel TD families, namely indolotryptoline, carboxy-indolocarbazole, and bisindolylmaleimide, adds new branches to the meta-biosynthetic scheme of bacterial TDs (Figure 38). As

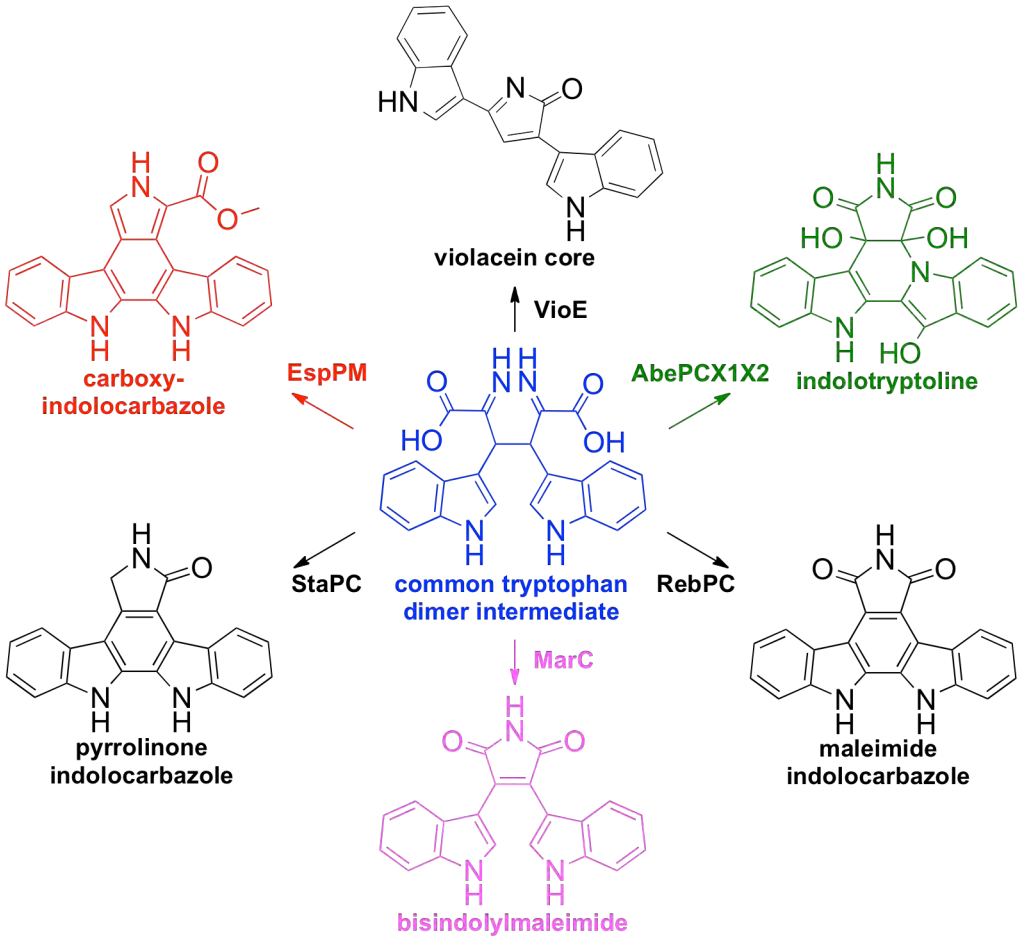


Figure 38. Bacterial tryptophan dimer biosynthetic pathways that diverge from a common tryptophan dimer intermediate (highlighted in blue). Indolotryptolines branch out via formation of an indolocarbazole precursor, followed by two successive oxidative steps (green; Group E), carboxy-indolocarbazole branch out via indolocarbazole formation, followed by methylation of a carboxyl unit (red; Group F), and bisindolylmaleimide pathway branch out via oxidative decarboxylation (pink; Group B).

only a small subset of known TD natural products have had their biosynthetic gene clusters characterized thus far, it is likely that the current global biosynthetic scheme is still incomplete and that there are more divergent tryptophan dimer pathways awaiting for discovery and characterization.

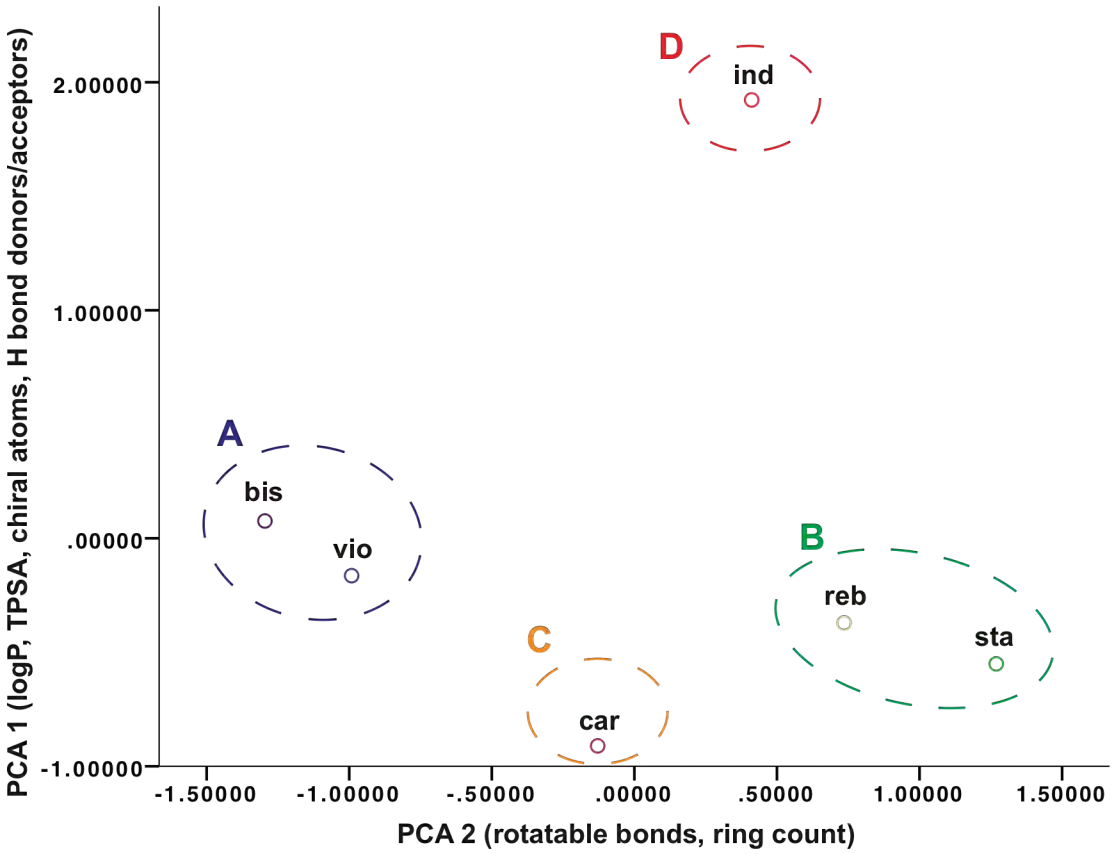


Figure 39. Chemical space of tryptophan dimer natural product family as represented by PCA plot of basic physiochemical properties. reb = rebeccamycin-like maleimide indolocarbazole; sta = staurosporine-like pyrrolinone indolocarbazole; vio = violacein; bis = bisindolylmaleimide; car = carboxy-indolocarbazole; ind = indolotryptoline.

To qualitatively evaluate the exploration of the tryptophan dimer chemical diversity, the basic physiochemical properties of the three known and three novel TD families were calculated using JChem (ChemAxon) and used as variables for principle component analysis (PCA) to map out their structural variability using SPSS Statistics software (IBM). Maximal variation was observed when PCA axis 1 consisted mainly of variables: log, TPSA, chiral atoms, and H bond donors/acceptors, while PCA axis 2 consisted mainly of variables: rotatable bonds and ring count. The resulting PCA plot

representing the chemical space of the TD core structures gives rise to four general groups (A, B, C, and D) that cluster based on structural relatedness (Figure 39). Group A consists of the bisindolylmaleimide and violacein families with relatively flexible structures, group B consists of the pyrrolinone and maleimide indolocarbazole families with relatively rigid structure, and group C and D each consists of carboxy-indolocarbazole and indolotryptoline cores as singletons, respectively. The PCA-based chemical space illustrates that the discovery of the bisindolylmaleimide family has allowed for the deep mining of the chemical space related to the violacein family, while the discovery of the carboxy-indolocarbazole and indolotryptoline families has permitted expansion into the previously unexplored tryptophan dimer natural product chemical space.

Chapter 4: Investigation of tryptophan dimer's mode of action

4.1 Resistant mutant screening

Upon the discovery of novel bioactive natural products, identifying their molecular targets become a critical step for their development as therapeutic agents and small-molecule probes that allow temporal and spatial modulation of target function (184, 185). However, linking bioactive small molecules to their cellular targets remains challenging and a general methodology for the rapid and systematic investigation of the compound's mode of action has not yet been established. This is especially true when studying cytotoxic natural products that might serve as anticancer agents.

One approach for investigating the mode of action of a small molecule involves the selection and full genome sequencing of mutants that acquire compound resistance (186). Upon the identification of resistance conferring mutations, a compound's effect on the activity of both the mutant and wildtype gene products can be used to suggest the mutated protein as the putative molecular target. In addition, since the resistance conferring mutations perturb the interaction of the compound to the putative target, the mutations can be mapped onto the target's structure to elucidate the putative binding site of the compound, information that would not be available in gene deletion- or overexpression-based genome-wide screening studies.

Because bacteria are fast-growing and genetically tractable, resistant mutant screening approach is commonly employed for mode of action of antimicrobial natural products (187, 188). However, its application to antitumor natural product mode of action studies has been limited (189) due to the time consuming, costly, and cumbersome nature of carrying out these experiments using human cells (184, 185). Antitumor agent target

identification studies have turned to *in vitro* affinity-based methods (190, 191), but since the screening is conducted outside of the cellular environment, the positive hits may not be physiologically relevant. A rapid and systematic resistant mutant screening approach that is compatible with antitumor agents should thereby be preferable.

4.2 MDR-sup fission yeast as model organism

Yeasts are an attractive eukaryotic model for antineoplastic mode of action studies because of their small genomes, fast growth rates, and genetic tractability (192). Most previous mode of action studies using yeast have turned to the budding yeast *Saccharomyces cerevisiae* as a model organism (184, 185). In contrast to budding yeast, fission yeast (*Schizosaccharomyces pombe*) maintains many more of the basic cancer-relevant cellular processes present in human cells (*e.g.* cell division, DNA replication, heterochromatin assembly), making it a potentially more general model for mode of action studies (193). Unfortunately, *S. pombe*'s lack of sensitivity to many cytotoxins due to its robust multidrug resistance (MDR) response has limited its use for genome-wide screening (194). A recent study identified five major contributors to fission yeast's MDR phenotype (four drug-efflux transporters and a transcription factor) and showed that their deletion results in the increased sensitivity of *S. pombe* to a wide range of chemical toxins (195). This MDR-suppressed (MDR-sup) strain of *S. pombe* should be particularly well suited for antiproliferative natural product mode of action studies because of its broad sensitivity to cytotoxins and its reduced ability to acquire drug resistance through uninformative, non-specific MDR mechanisms. We thereby sought to establish a system for the genome-wide resistant mutant screening to antiproliferative natural products using MDR-sup *S. pombe* as a model organism.

4.3 Indolotryptoline as subject for mode of action study

Natural products have traditionally served as the most prolific source of new antineoplastic agents. Natural products or natural product inspired structures comprise approximately 80% of FDA approved anticancer drugs (6). The natural products screening model that led to the discovery of most anticancer agents is based on the premise that there is a positive correlation between small molecule structural novelty and the ability of compounds to interact with new molecular targets (196). Ultimately, novel compounds that have successfully transitioned into clinically useful cancer chemotherapy drugs (*e.g.* doxorubicin, mitomycin, paclitaxel) tend to exhibit very potent (nM - pM) antiproliferative activity against diverse cancer cell lines *in vitro* (196, 197). Indolotryptoline-based natural product cytotoxins satisfy both of these criteria (*i.e.* structural uniqueness and high potency), making them appealing targets for mode of action studies (127).

The most extensively studied tryptophan dimers are the indolocarbazoles, which differ from indolotryptolines by the presence of a tricyclic carbazole in place of the tryptoline moiety (Figure 17). A number of synthetic derivatives of the indolocarbazoles staurosporine and rebeccamycin, which are protein kinase (48) and DNA topoisomerase I (49) inhibitors, respectively, have been introduced into clinical trials as cancer therapeutic agents (46, 47). Indolotryptolines show nanomolar *in vitro* cytotoxicity which is similar to the activity seen for the most potent natural indolocarbazoles; however, structural differences between these two compound families suggest they likely have distinct modes of action (127). There have been two previous reports on the potential molecular target of indolotryptoline-based metabolites, but in both instances, assays were

carried out against only a single molecular target of interest. The first study, as described in Chapter 3.3, tested the indolotryptoline borregomycin A against a panel of protein kinases *in vitro* and found it to be most active against the pCaMKII∂ kinase (128). The second study identified the indolotryptoline BE-54017 as a hit in a high-throughput screen for compounds that inhibit vacuolar H^+-ATPase (V-ATPase) activity in a human cell line (198). We reasoned that a resistant mutant selection approach would provide a more systematic *in vivo* analysis of the mode of action of indolotryptolines and potentially even provide residue level details about the binding site of this family of natural products. We have thereby chosen BE-54017 (**1**) and cladoniamide A (**2**), two most potent members of the indolotryptoline family of antiproliferative natural products, as subjects for the resistant mutant screening-based mode of action study.

4.4 Mutations conferring indolotryptoline resistance

BE-54017 and cladoniamide A show similar *in vitro* cytotoxicity against human cell and MDR-sup *S. pombe* (IC_{50} BE-54017: 15 nM; cladoniamide A: 32 nM), suggesting the potential for a common mode of action in both organisms. In contrast, the *S. pombe* strain from which MDR-sup *S. pombe* was derived (MDR-active *S. pombe*) is approximately 50 times less sensitive to these natural products (IC_{50} BE-54017: 780 nM; cladoniamide A: 1600 nM) and is likely prone to off target effects not seen in human cells. As an initial step to resistant mutant screening study, methylnitronitrosoguanidine (NTG) mutagenized MDR-sup *S. pombe* was plated on media containing either BE-54017 or cladoniamide A (Figure 40). For both metabolites, concentrations ranging from 2 to 6 fold above the IC_{50} (BE-54017: 35-70 nM; cladoniamide A: 89-178 nM) were used for resistant mutant selections, which yielded between one and ten resistant colonies per

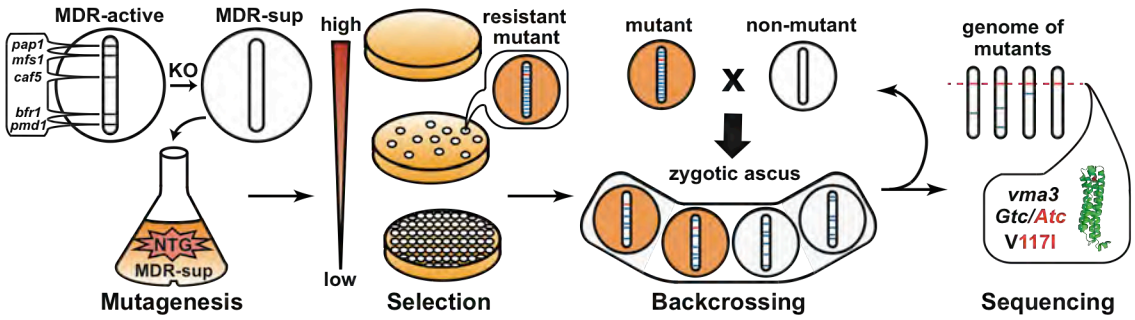


Figure 40. Four-step schematic of molecular target identification using *S. pombe*. **1) Mutagenesis:** The *S. pombe* strain with multidrug resistance (MDR) response suppressed (MDR-sup) through five-gene knockouts (KO) is randomly mutagenized with methylnitronitrosoguanidine (NTG). **2) Selection:** NTG mutagenized cells are plated on drug containing solid media to select for resistant mutants. **3) Backcrossing:** Drug-resistant mutants (orange) are crossed with un-mutagenized MDR-sup (white), resulting in the formation of ascus containing four spore progeny with a reduction in the number of mutations. Progeny retaining drug resistance (orange) are maintained. **4) Sequencing:** The genome of each backcrossed mutant is sequenced and compared to the un-mutagenized parent genome to identify the specific mutation that confers drug resistance.

100,000 mutagenized cells. For each target compound, we picked 12 resistant colonies for further analysis (BE-54017 = *S. pombe* strains B1-B12, cladoniamide A = *S. pombe* strains C1-12). These 24 mutants were re-examined for indolotryptoline resistance and tested for cross-resistance to the unrelated cytotoxin, brefeldin A. Strains that failed to show resistance when re-screened against indolotryptolines (B2, B9, C1, C11) and those showing cross-resistance to brefeldin A (C7) were abandoned. All of the remaining strains were found to be resistant to both BE-54017 and cladoniamide A, suggesting a

A

Mutant ID	V-ATPase complex *vma3*	*vma11*	Zinc transporter *zhf1*	Off genes	Fold difference (relative to non-mutant) BE	Cla	Baf	Con	Bre
B12	V59I	–	–	SPAC25B8.11	x8.0	x8.0	x0.48	x1.2	x1.1
B7	V117I	–	–	–	x4.2	x3.9	x1.0	x1.0	x0.94
C6	–	G44D	–	–	x2.0	x2.1	x0.92	x0.89	x0.97
B1	–	–	W324L	qcr6	x6.1	x6.2	x3.4	x1.9	x1.0
B5	–	–	W324L	apl5 SPCC162.02c	x6.3	x6.2	N/D	N/D	N/D
C3	–	–	W324L	SPBC725.04 SPBC18H10.20c SPCC61.04c	x6.1	x6.0	N/D	N/D	N/D
B C8	M130I	–	–	N/D	x3.8	x3.6	x1.9	x0.88	x0.96
B10, C12	–	A36V	–	N/D	x1.8	x1.8	x0.93	x1.0	x1.1
B11	–	G44D	–	N/D	x1.9	x2.0	N/D	N/D	N/D
B3	–	–	G28D	N/D	x5.8	x5.6	N/D	N/D	N/D
8 remaining clones	–	–	W324L	N/D	x6.2*	x6.1*	N/D	N/D	N/D

C

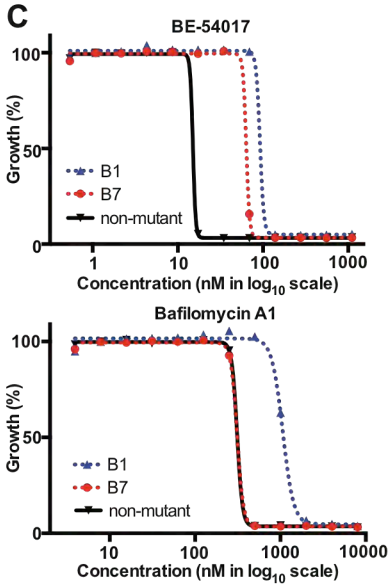


Figure 41. Data from BE-54017 (B#) and cladoniamide A (C#) resistant mutant strains. Compilation of data from either **A)** fully sequenced or **B)** *vma3*, *vma11*, or *zhf1* PCR sequenced resistant mutants. Genes containing mutations, the amino acid change encoded by each mutations and the fold-difference in whole-cell cytotoxicity (IC_{50}) relative to un-mutagenized strain are shown. BE = BE-54017; Cla = Cladoniamide A; Baf = Bafilomycin A1; Con = Concanamycin A; Bre = Brefeldin A; N/D = not determined; number in asterisk (*) represents the average value. **C)** Examples of dose response curves used to determine whole-cell cytotoxicity and fold-differences in IC_{50}. Curves for BE-54017 and bafilomycin A1 against mutant B1, mutant B7 or un-mutagenized MDR-sup *S. pombe* are shown.

common mode of action and drug binding site. The IC_{50} for these resistant mutants ranged from 2 to 8 fold above the IC_{50} determined for un-mutagenized MDR-sup *S. pombe* (Figure 41, Appendix 5).

Random NTG mutagenesis results in the introduction of multiple mutations into the genome of each strain, complicating the identification of the specific mutations that are relevant to the mode of action. A key benefit of using fission yeast as a model for the resistance selection approach is that the complexity of the mutagenized genetic

background can be significantly simplified through backcrossing with un-mutagenized yeast (Figure 40). Backcrossing serves to replace non-drug resistance associated mutations with wildtype allelles from the un-mutagenized strain, thereby preventing irrelevant mutations from complicating downstream genome-wide bioinformatics analyses. Six representative mutant strains showing varying levels of resistance (B1, B5, B7, B12, C3, C6) were backcrossed four to six times with un-mutagenized MDR-sup *S. pombe* (Figure 41). Consistently, two of the four progeny from each backcross retained drug resistance, suggesting that a single mutation was responsible for the observed resistant phenotype in each strain. The final backcrossed clones were subjected to Illumina whole-genome sequencing and the resulting reads were mapped onto the un-mutagenized MDR-sup genome to identify differences in protein coding sequences. Upon comparison to the un-mutagenized MDR-sup genome, the six backcrossed strains were found to contain between one and four point mutations (Figure 41A). While various one off mutations were observed in this strain collection, all resistant strains had mutations in either the zinc transporter gene, *zhf1* (B1, B5, C3) or in genes encoding the c (*vma3*; B7, B12) or c' (*vma11*; C6) subunits of the vacuolar H^+-ATPase complex (V-ATPase), suggesting that zinc transporter and V-ATPase activity were likely linked to the molecular mechanism of indolotryptoline cytotoxicity. In no cases did we detect mutations that might traditionally be associated with a generic MDR-like phenotype, highlighting one of the key advantages of using the MDR-sup strain.

The *zhf1*, *vma3* and *vma11* genes from the thirteen resistant strains that were not analyzed by Illumina whole-genome sequencing were PCR amplified and individually sequenced (Figure 41B). All thirteen strains were found to contain a mutation in one of

these three genes. Eight strains contain the same *zhf1* mutation (W324L) observed previously, making it the most common mutation we detected. Two strains were found to contain previously identified *vma3/11* mutations and the remaining three strains contained new variants of *vma3* (C8), *vma11* (B10) and *zhf1* (B3).

The genetic tractability of yeast allows for a specific mutation of interest to be easily introduced into a clean background to provide a genetic validation of its role in conferring resistance. To confirm the relevance of the *vma* and *zhf* mutations to indolotryptoline resistance, each unique *vma3/11* point mutations (B7, B10, B12, C6, C8) and the common W324L *zhf1* mutation (B1) were introduced into the un-mutagenized MDR-sup *S. pombe* strain by homologous recombination. The resulting strains, which were shown by PCR amplicon sequencing to harbor the desired mutation, were all found to be resistant to indolotryptolines at the same level as the randomly mutagenized clones containing the same mutation. The individual *vma3, vma11* and *zhf* mutations we detected are therefore necessary and sufficient for conferring resistance to indolotryptoline-based compounds.

4.5 V-ATPase proteolipid subunit as putative target of indolotryptoline

Previous studies have observed a tight functional connection between zinc toxicity and V-ATPase activity. The inactivation of the V-ATPase in diverse organisms from yeasts to plants, by either a gene knockout or a small-molecule inhibitor, is known to result in increased sensitivity to zinc, indicating that the V-ATPase plays a critical role in zinc homeostasis (199, 200). As would be expected for a V-ATPase inhibitor, we observed that MDR-sup *S. pombe* becomes more sensitive to BE-54017 and cladoniamide A with increasing concentrations of zinc in the growth medium (Figure 41).

A

$[ZnCl_2]$ (mM)	Fold difference (relative to no $ZnCl_2$)				
	BE	Cla	Baf	Con	Bre
0.01	x0.88	x0.95	x0.96	x1.1	x1.0
0.05	x0.50	x0.58	x0.64	x0.78	x0.97
0.1	x0.30	x0.35	x0.48	x0.63	x1.0
0.2	x0.22	x0.24	x0.24	x0.45	x0.93

B

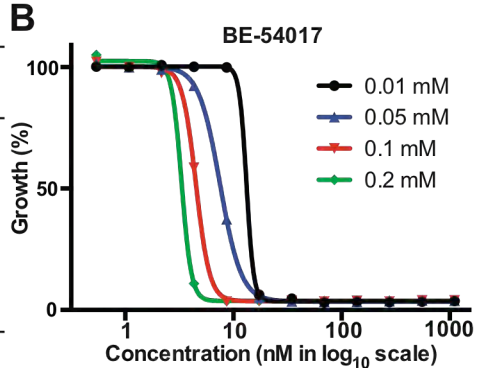
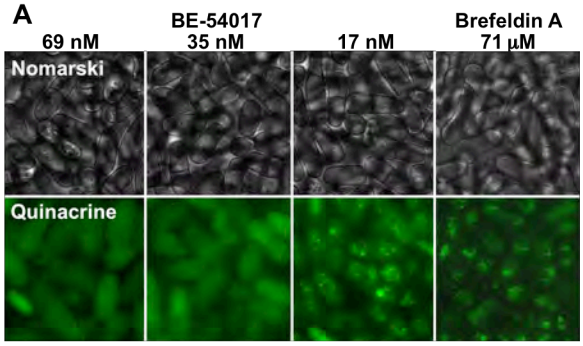


Figure 42. Sensitivity of the un-mutagenized MDR-sup *S. pombe* to cytotoxins in the presence of increasing concentrations of zinc. **A)** Summary of fold-differences in whole-cell cytotoxicity (IC_{50}) relative to media without the addition of $ZnCl_2$. **B)** Dose response curve for un-mutagenized MDR-sup when exposed to BE-54017 in the presence of different concentrations of $ZnCl_2$.

Although V-ATPase activity in budding yeast has been investigated *in vitro* using isolated vacuolar membranes (201), the purification of similar membrane fractions from *S. pombe* has proved challenging, thereby precluding comparable *in vitro* biochemical analyses (202). However, because V-ATPase is responsible for maintaining the acidification of cellular organelles, *in vivo* V-ATPase activity can be monitored using acidic staining dyes (*e.g.* quinacrine) that form fluorescent puncta in vacuoles at reduced pH (203). As previously reported for known V-ATPase inhibitors (204, 205), the addition of indolotryptolines to culture media at sub-minimum inhibitory concentrations resulted in the loss of fluorescent puncta formation in a dose dependent manner in un-mutagenized MDR-sup *S. pombe* (Figure 43A). *In vivo* V-ATPase inhibition by indolotryptolines (MIC BE-54017: 69 nM; cladoniamide A: 143 nM) occurs at concentrations similar to those needed for whole-cell cytotoxicity (MIC BE-54017: 35



B

Mutant ID		B1	B7	B10	B12	C6	C8
Fold difference (relative to non-mutant)	BE	x1	x4	x2	x8	x2	x4
	Cla	x1	x4	x2	x8	x2	x4

Figure 43. Visualization of *in vivo* V-ATPase activity by acidic organelle staining. **A)** Nomarski and fluorescent images of quinacrine-stained un-mutagenized MDR-sup *S. pombe* upon incubation with BE-54017 (BE) or brefeldin A (Bre). Fluorescent puncta formation is inhibited at higher concentrations of BE, with no puncta observed in >95% of the cells at 69 nM (MIC). Bre has no effect on puncta formation. **B)** Summary of fold-difference in *in vivo* V-ATPase MIC relative to un-mutagenized MDR-sup. Cla = Cladoniamide A.

nM; cladoniamide A: 72 nM). Acidic staining experiments were carried out using strains containing each unique *vma3* (B7, B12, C8) and *vma11* (B10, C6) mutation, as well as a representative strain with the common W324L *zhf1* (B1) mutation (Figure 43B, Appendix 6). All *vma3* and *vma11* mutants showed increased tolerance to puncta disruption by indolotryptolines, while the *zhf1* mutant responded like non-mutant MDR-sup in these experiments. Taken together, these experiments indicate that indolotryptolines inhibit V-ATPase activity and that the primary deleterious effect of V-ATPase inhibition is likely the downstream disruption of zinc homeostasis.

Indolotryptoline resistant strains were tested for cross-resistance to the well-characterized plecomacrolide-type V-ATPase inhibitors, bafilomycin and concanamycin (Figure 41, Appendix 5). The W324L Zhf1 mutant confers resistance to both known V-ATPase inhibitors. With the exception of the Vma3 M130I mutant, which confers a two-

fold increase in resistance to bafilomycin, V-ATPase mutants did not show cross-resistance to either bafilomycin or concanamycin. The fact that V-ATPase mutants are largely compound class specific, while Zhf1 mutants confer resistance irrespective of the compound class, further supports a model in which the V-ATPase, and not zinc transporter, is the likely molecular target of indolotryptolines; zinc toxicity is likely the downstream deleterious consequence of V-ATPase inhibition.

4.6 Putative indolotryptoline binding site in V-ATPase

The V-ATPase is a multi-protein complex that consists of two domains: a peripheral ATP binding domain V_1 and a membrane associated proton translocating pore domain V_0 (Figure 44A) (206, 207). In yeast, the V_0 domain contains a hexameric cylinder that is known as the proteolipid ring. The catalytic V_1 subunit hydrolyzes ATP, which drives the rotation of the proteolipid ring that in turn allows for protons to translocate across the membrane. The proteolipid ring is thought to be composed of c (*vma3*), c' (*vma11*) and c'' (*vma16*) subunits that assemble in a 4:1:1 (c:c':c'') stoichiometry (206-208). As is the case with plecomacrolide resistance conferring mutations, indolotryptoline conferring mutations are found in the proteolipid subunits. Based on the fact that four of the five *vma3/11* mutations that confer resistance to indolotryptolines do not show cross-resistance to plecomacrolide-type metabolites and that the fifth mutation only confers low-level resistance to bafilomycin (Figure 41), the indolotryptoline-binding site is likely distinct from, but proximal to, the plecomacrolide-binding pocket in the proteolipid ring.

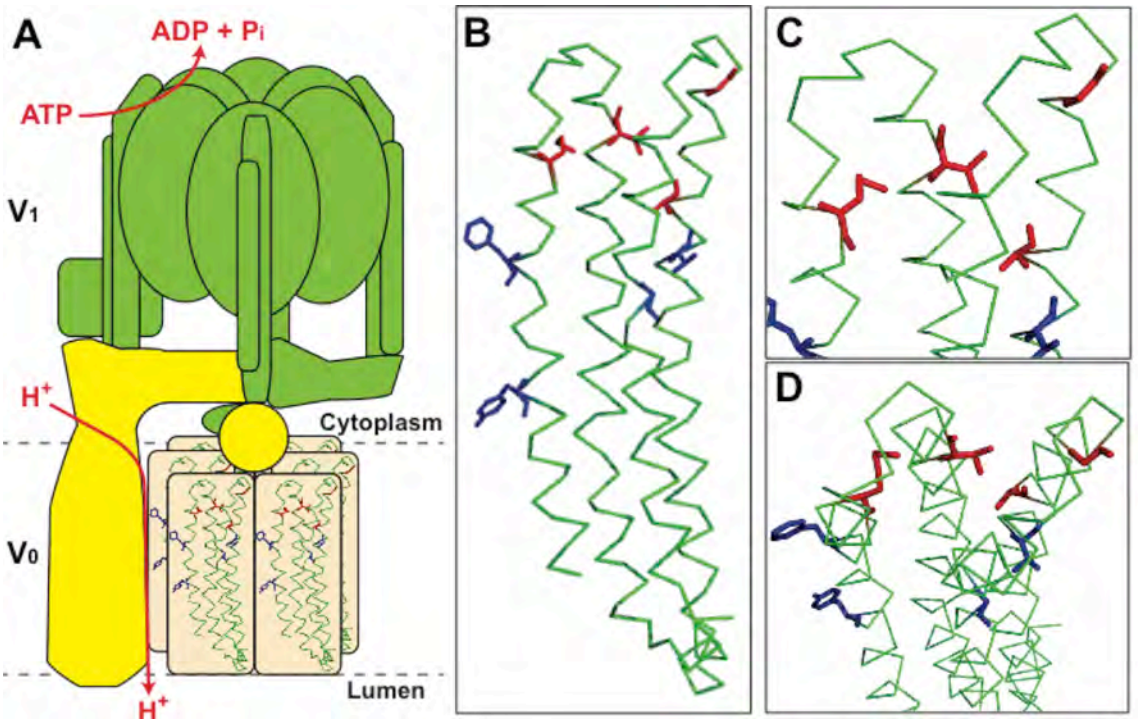


Figure 44. Structure of V-ATPase. **A)** Schematic of yeast V-ATPase architecture, which consists of the catalytic V_1 domain (green) and the membrane translocating V_0 domain (yellow and beige). Proteolipid subunits c/c'/c'' (beige) are part of the V_0 domain and form a hexameric ring structure. **B, C, D)** Side (**B**), close up (**C**) and top (**D**) view of the crystal structure of a proteolipid subunit from the *E. hirae* V-ATPase (PDB ID: 2BL2). The peptide backbone is represented as a green stick model. Residues found to be involved in conferring resistance to indolotryptolines are shown in red. Residues reported to be involved in plecomacrolide resistance in previous studies are shown in blue.

The detailed molecular structure of the yeast V-ATPase is not known; however, the proteolipid substructure from the vacuolar Na^+-ATPase from the bacterium, *Enterococcus hirae*, has been determined by X-ray crystallography (209). Despite phylogenetic and functional differences, the high sequence homology seen between the *E. hirae* proteolipid subunits and those found in eukaryotic V-ATPase complexes has led to the use of the *E. hirae* V-ATPase structure as a model for eukaryotic V-ATPase studies

(208, 210). In particular, the residues that confer resistance to V-ATPase inhibitors have been mapped onto *E. hirae* V-ATPase structure to predict putative inhibitor binding sites. Residues reported to confer resistance to plecomacrolide-like inhibitors generally cluster midway between the cytoplasmic and luminal face of proteolipid ring, and point out from the four helix bundle that makes up each of the proteolipid subunit (208). In contrast, indolotryptoline resistance conferring mutations map near the cytoplasmic face and mostly point into the center of the bundle (Figure 44), suggesting indolotryptolines bind at a unique site within the proteolipid four-helix bundle near its cytoplasmic face. Based on the significant structural differences between indolotryptolines and known V-ATPase inhibitors (Figure 44) (211), it is not surprising that they would bind at a distinct site.

Tryptophan dimers, including the closely related indolocarbazole family of natural products, have been observed to interact with proteins containing nucleotide-binding sites (*i.e.* ATP or DNA) through mimicry of a nucleotide base (46-49). While the V-ATPase uses ATP, the proteolipid subunits of V-ATPase are not known to contain a nucleotide-binding site, suggesting that tryptophan dimer binding motifs can extend beyond simple nucleotide mimicry. This should encourage the continued discovery and mode of action analysis of new natural tryptophan dimers. Disruption of the planar indolocarbazole ring system through the sp^3 hybridization of the two carbons at the base of the pyrrole ring in the indolotryptoline structure is likely a key factor in the altered binding mode of this class of tryptophan dimers. Ultimately, an indolotryptoline-proteolipid co-crystal structure will likely be required to confirm these binding motif hypotheses.

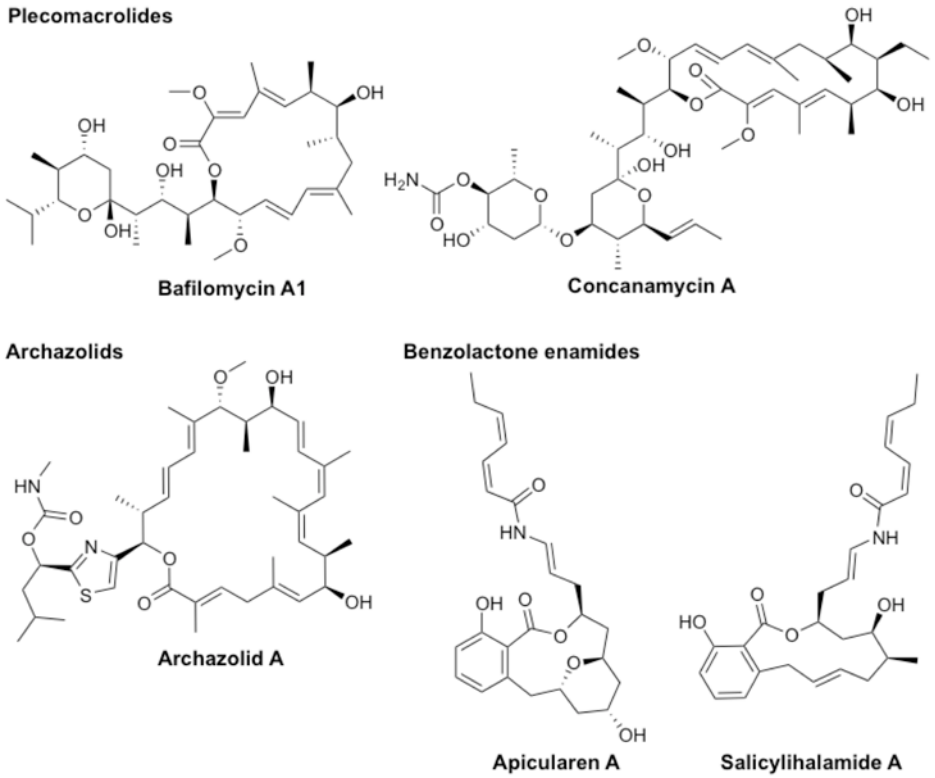


Figure 45. Chemical structure of previously characterized V-ATPase inhibitors.

4.7 Summary of MDR-sup *S. pombe* resistant mutant screening

Through whole-genome sequencing of the indolotryptoline resistant mutants, we identified point mutations in the proteolipid (c/c') subunits of V-ATPase and the zinc transporter Zhf1 that conferred resistance to this family of natural products. Acid vacuole staining, cross-resistance studies, and direct c/c' subunit mutagenesis indicate that the primary resistance mechanism to indolotryptolines is the mutation in V-ATPase and that the V-ATPase, as opposed to the zinc transporter, is the likely molecular target of indolotryptolines. The mapping of the resistance conferring mutations onto a model V-ATPase proteolipid structure predicts that indolotryptolines bind at the interface of the four helix bundle of a single proteolipid subunit near its cytoplasmic face, in a site that is distinct from previously described V-ATPase inhibitors.

V-ATPase is a highly conserved protein complex in eukaryotes that plays a role in acidifying a variety of organelles (206, 207). V-ATPase inhibitors have been explored as cancer therapeutic agents due to their cytotoxicity towards diverse cancer cell lines (212, 213). Cell lines from cancers that are especially malignant, aggressive and unresponsive to current therapies are known to be sensitive to V-ATPase inhibitors, possibility due to the involvement of an acid microenvironment in tumor progression and multidrug resistance (212, 213). The unique indolotryptoline V-ATPase binding site proposed in this study should help guide the development of a new class of V-ATPase inhibitors that can be explored for anticancer activity.

Resistant mutant analysis provides insight to both the molecular target and binding site of the small molecule, while not requiring any chemical modification of the target compound, preliminary prediction of the target, or construction of a custom-made genetic library. As such, the resistant mutant selection of MDR-sup *S. pombe* serves as a convenient alternative approach for mode of action studies of cytotoxic natural products. The characterization of a cytotoxic natural product's molecular target using human cells has traditionally been costly, time consuming, and technically cumbersome. Resistant mutant screening using MDR-sup *S. pombe* should serve as a powerful and generally applicable alternative target identification technique that fits nicely into diverse drug discovery pipelines, including this metagenomics-based TD discovery platform, for gene-level target identification of cytotoxins.

Chapter 5: Discussions

5.1 Conclusions and future directions

We have hereby illustrated the metagenomics-based natural product discovery and development pipeline that provides a means of exploring the biosynthetic capacity of thousands of bacterial genomes simultaneously regardless of their culturability. Its utility was demonstrated by recovering 16 unprecedented gene clusters by homology-based screening that targets the CPA synthase gene contained within TD pathways. Heterologous expression of eDNA-derived gene clusters permitted the functional characterization of three novel TD families with members that exhibit clinically relevant bioactivities. One of the TD family, indolotryptoline, was further developed by elucidating its molecular mode of action.

This study focuses on the discovery of TD natural products, and while iterations of this approach should lead to the isolation of additional compounds from the TD class, this pipeline can be generalized to other natural product classes, depending on the design of the degenerate primer set for PCR screening. Homology-based screening strategies have been employed to isolate novel natural products from the isocyanide natural product (214), type II polyketide (42, 82, 95, 215, 216), and glycopeptide-type non-ribosomal peptide (41, 43, 88) class in our laboratory, and from the cyanobactin-type ribosomal peptide (217) and the trans-acyltransferase (trans-AT) polyketide (218) class by other groups (Figure 46A) . As for non-ribosomal peptides and polyketides, studies are currently underway in our laboratory to sequence and catalog the conserved adenylation and ketosynthase domains for the recovery of these types of gene clusters on an on-

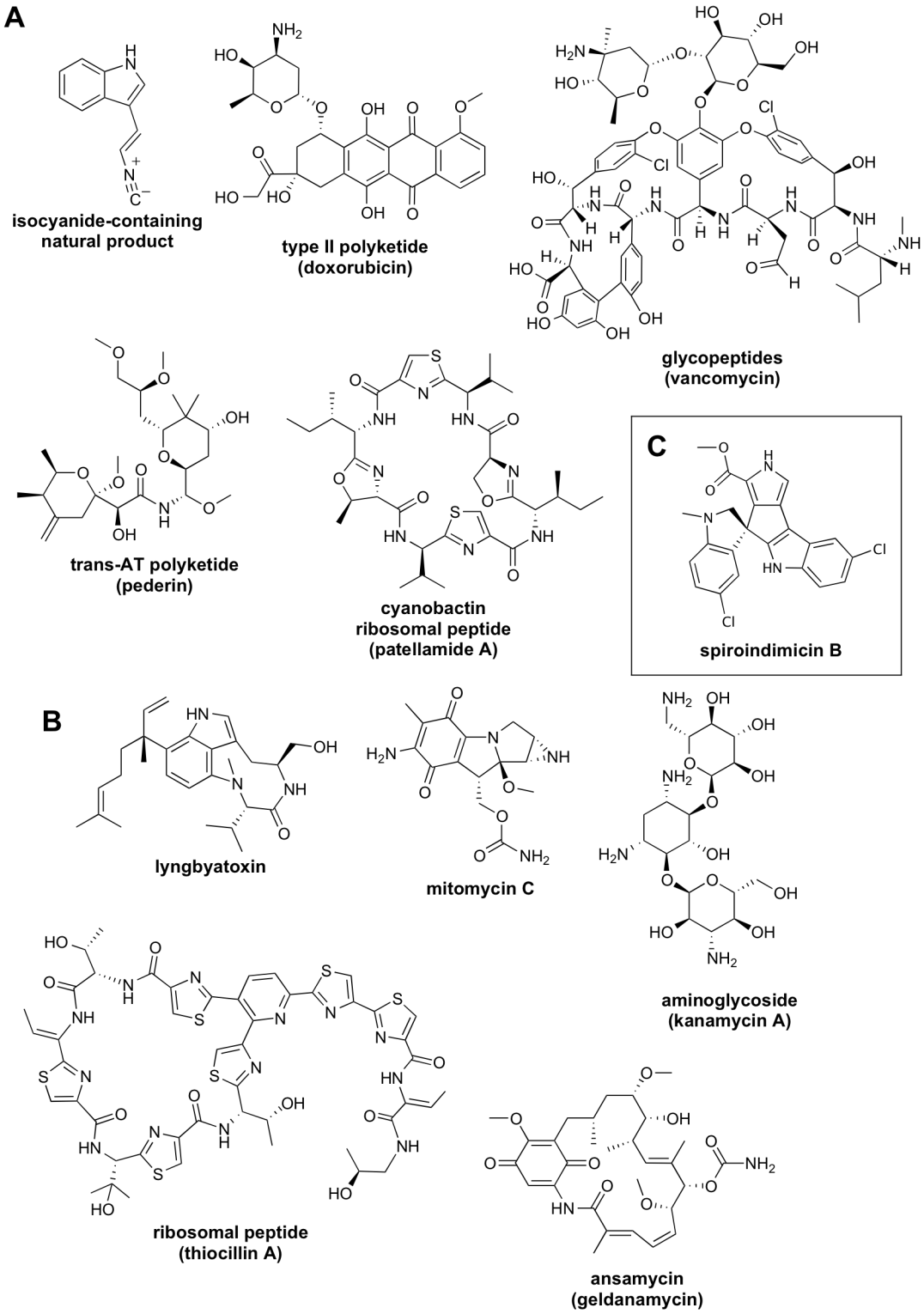


Figure 46. Chemical structure of natural product classes that have been (**A**) or could be (**B**) surveyed by homology-based screening. **C** is a marine-derived TD.

demand basis (43, 219). Of other natural product classes that could be of interest for homology-based screening include lyngbyatoxin-type non-ribosomal peptides (since isoprene-containing compounds are generally underexplored in bacteria and the gene cluster is relatively small, thereby amenable for genetic manipulation) (220), mitomycin-type natural products (clinical relevance and relatively unique biosynthetic pathway) (221), aminoglycosides (clinical relevance) (222), ansamycins (clinical relevance, common 3-amino-5-hydroxybenzoic acid (AHBA) biosynthetic motif) (223), and the various types of ribosomal peptide natural products (RiPPs; growing class of bacterial metabolites) (224) (Figure 46B). Moreover, since natural products discovered from marine bacteria are generally found to have drastically different chemical structures compared to soil bacterial metabolites (225), including the recently discovered marine actinomycetes-derived TD compound spiroindimicins (Figure 46C) (226), screening TD gene clusters in eDNA libraries constructed from marine samples should also be fruitful.

The seven-step overview (Figure 7) of our metagenomic natural product discovery and development pipeline can be further generalized into four phases: 1) metagenomic library construction, 2) gene cluster screening, 3) natural product heterologous expression and production, and 4) natural product characterization and development. Improvements and alternative strategies to these four stages will be discussed as future directions. As described previously, soil eDNA library construction can routinely be conducted in our laboratory (28) and requires no major advancement for homology-based screening, aside from the possibility of DNA extraction from other environmental samples (*e.g.* marine samples). BAC (85-87) and P1 phage vector (227)-based library construction approaches may be explored for the accommodation of larger

insert sizes (>40 kb), but it is not as important, considering that the complete gene cluster does not need to be contained in a single cosmid for homology-based screening. Therefore, the rest will be devoted to the discussion of expanding on the remaining three phases. In particular, we examine the utility of CPA synthase gene phylogeny and violacein reporter construction in gene cluster screening, as well explore the roles of synthetic biology in the heterologous expression of cryptic gene clusters and pathway engineering for the generation of natural product analogs as a compound development strategy.

5.2 CPA-guided analysis of metagenomic TD biodiversity

The characterization of unique TD substructures from distinct clades in the CPA synthase gene phylogenetic tree suggests that CPA synthase gene sequences diverge in concert with the functional outputs of their respective TD gene clusters, making CPA gene sequences alone a good marker to guide the discovery of novel TDs. Since the publication of our first CPA-guided TD discovery study (127), several reports have been made regarding the use of CPA targeting primer set for the detection and discovery of TD gene clusters from various genetic sources, establishing the CPA synthase gene as a suitable sequence tag for TD gene clusters (226, 228, 229). To further explore TD diversity in the environment, DNA extracted directly from 20 soil samples from geographically distinct sites in New Mexico was screened by PCR using our CPA synthase degenerate primers, and the resulting PCR amplicons were sequenced (Appendix 7). A phylogenetic analysis of these sequences shows that they form new clades both within (*e.g.* NMCC27) and outside (*e.g.* NMCC11) the characterized TD

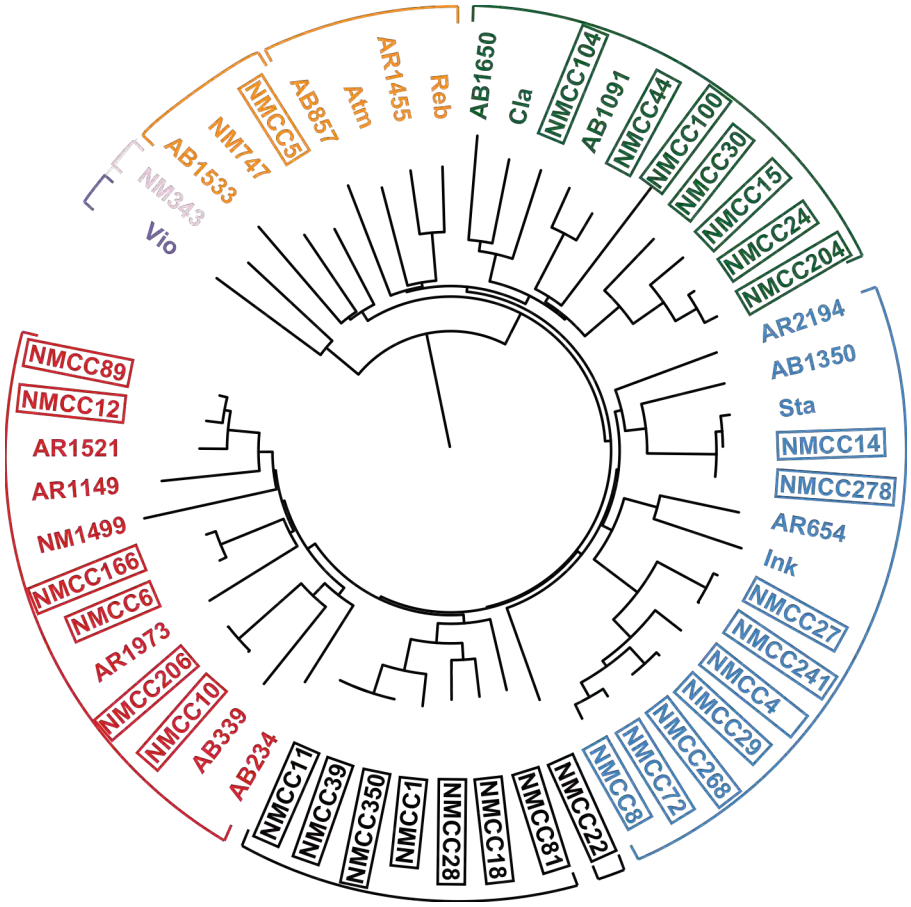


Figure 47. ClustalW-based phylogenetic tree of trimmed CPA synthase gene amplicons. The eDNA-derived sequences (boxed) fall into known TD groups A-E (colored), and form new clades (black) that encode potentially novel TD core substructures.

groups, suggesting these CPA synthase homologs may appear in gene clusters that encode novel TD with known substructures as well as novel TD substructures (Figure 47).

Our study has thereby only uncovered the tip of the iceberg of bacterial TD biodiversity and within the massive pool of the bacterial metagenome lies even greater number of unprecedented TDs and their corresponding biosynthetic pathways that remain to be discovered. For subsequent TD discovery efforts, the CPA synthase gene amplicons can be sequenced in parallel in multiple environmental samples to identify the

appropriate soil sample for eDNA library construction, either based on the diversity of novel clades for the general discovery of new TD substructures, or the plethora of novel members of a specific clade for the focused discovery of new derivatives from a biologically significant TD class.

Natural product discovery programs have long relied on the random screening of culture broth extracts to identify metabolites with bioactive properties. Unfortunately, it has become increasingly difficult to isolate new members of biomedically interesting classes of compounds using random screening strategies due to the phenomenon of redundant isolation (230). Homology-based screening of metagenomes, guided by phylogenetic profiling, offers a solution to this dilemma by allowing the targeted recovery of novel biosynthetic pathways, from the level of the selection of soil samples for eDNA library construction to the choice of eDNA-derived gene clusters for heterologous expression efforts. In particular, CPA synthase gene-guided analysis not only allows for the exclusion of sequences that are likely associated with gene clusters encoding for previously encountered metabolites, but also predict the type of the TD substructure that they encode. Thus, CPA synthase gene phylogenetics should greatly facilitate future metagenomic TD natural product discovery programs.

5.3 Violacein reporter-based screening of TD gene clusters

Homology-based screening of gene clusters from eDNA libraries suffers from several limitations. First, the collection of eDNA-derived TD gene clusters that are uncovered from the library will necessarily be biased by primer design because the primer sequence cannot possibly be degenerate enough to amplify all possible CPA synthase genes, while specific enough to not be overwhelmed by false positives. Future

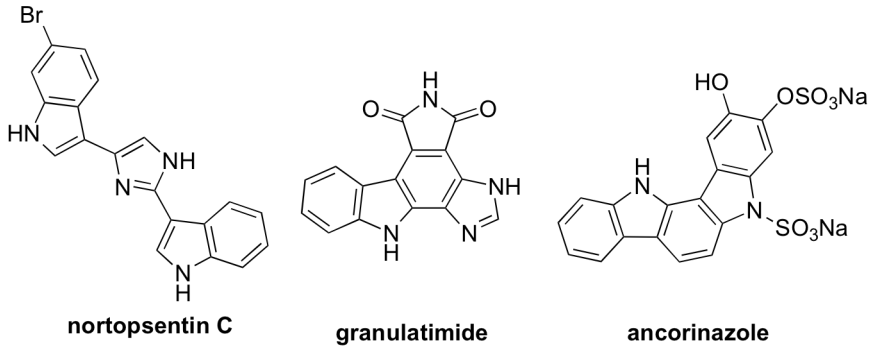


Figure 48. TD- and TD-like natural products that are predicted to be biosynthesized via IPA-imine synthase, but not CPA synthase, mediated reactions.

directions to TD screening efforts should therefore include the design of other degenerate primers to broaden the recovery of TD gene clusters with diverse CPA synthase gene sequences, but on the other hand, a screening methodology that is independent of sequence bias is also preferable.

Second, while certain biosynthetic genes may be functionally conserved within a particular natural product class, the genes with no conserved regions in their DNA sequences for primer design cannot be targeted for homology-based screening. As described in Chapter 2.2, this was observed in the case of IPA-imine synthase genes (Figure 8), where the *staO*/*rebO* genes from indolocarbazole clusters do not share significant sequence homology to the *vioA* gene from the violacein cluster, although these genes are known to be functionally equivalent (94, 163, 164). Despite that the known bacterial TD gene clusters contain both the IPA-imine synthase and the CPA synthase genes, some known TD and TD-like natural products can be predicted to be biosynthesized via IPA-imine synthase-, but not CPA synthase-mediated catalysis, such as nortopsentin A (similar to violacein, but the internal ring consist of two, instead of one,

nitrogen atoms) (231), granulatimide (coupling of tryptophan and histidine) (232), and ancorinazole (*anti*, instead of *syn*, dimerization of indoles) (233) (Figure 48). Therefore, a screening approach that can target IPA-imine synthase and other genes should expand the capability of recovering diverse TD and TD-like natural product gene clusters.

Third, while homology-based screening allows for biosynthetic gene clusters that do not necessarily have to be expressed in the library host, this implies that a significant number of gene clusters that are recovered in this manner are cryptic, as demonstrated in our study, thereby requiring considerable time and cost expenditures for heterologous expression efforts. An approach that can selectively screen for functional gene clusters should be useful.

Taking into account these limitations, we envision that a violacein reporter-based complementation screening should serve as an appropriate alternative strategy to metagenomic TD discovery. As described previously, violacein is a purple TD compound that also consists of IPA-imine synthase and CPA synthase in its pathway (Figure 8) (94, 163, 164). As such, a bacterial colony expressing violacein can easily be detected on an agar plate by purple pigmentation. A reporter construct can thus be made that contains the violacein gene cluster (155), but with one of the genes knocked out, such as the IPA-imine synthase gene. The co-expression of the cosmid eDNA clone along with the reporter construct should allow for a visual detection of a bacterial colony expressing a functional IPA-imine synthase gene, which will complement the reporter construct and yield a purple-pigmented phenotype (Figure 48). The three previously mentioned limitations that restrict homology-based screening are nonexistent in this approach, since the screening criterion is based on phenotype and not on sequence. Moreover, as opposed

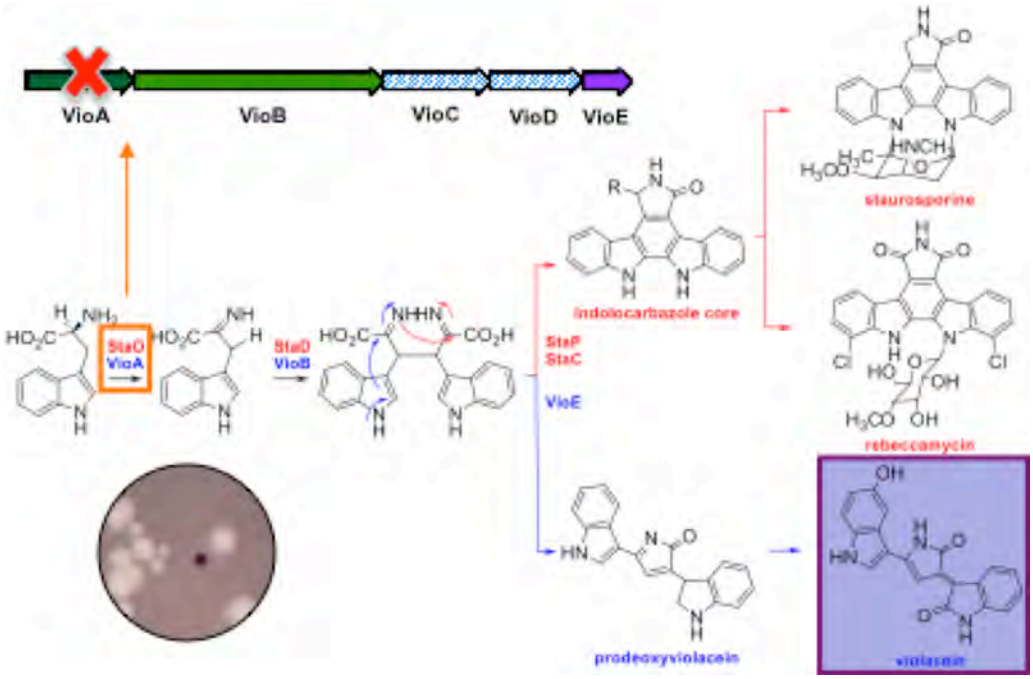


Figure 49. Reporter-based complementation screening of TD gene clusters. The *vioA*-deficient violacein gene cluster is used as a reporter to detect eDNA clones that express a functional IPA-imine synthase gene.

to the traditional phenotype-based screening, the encoding small molecule does not have to provide any phenotype, such as coloration or zone of inhibition, to the bacterial colony for detection, and the entire gene cluster does not have to be contained in a single cosmid, since only the IPA-imine synthase gene has to be expressed for pigmentation.

Preliminary study of reported-based screening using IPA-imine synthase gene-deficient construct in a small *E. coli*-based eDNA library resulted in the recovery of several putative biosynthetic gene clusters, one of which is even predicted to encode for small molecules of a different natural product class, namely isocyanide-containing natural products (27, 214, 234). By cloning the violacein pathway operon under the control of a host-specific artificial promoter, we also found that violacein can be

produced in different heterologous hosts, including *S. albus*. By conducting this reporter-based screening in various hosts, we anticipate the discovery of novel natural product gene clusters that we could otherwise not find using homology-based screening. Considering the isolation of an isocyanide natural product gene cluster, we could also engineer different reporter constructs with each biosynthetic gene of the violacein pathway knocked out, and recover a variety of gene clusters in that matter, with the intent of combining eDNA-derived genes from non-TD related pathways to reconstruct and build a synthetic violacein pathway.

5.4 Synthetic biology in heterologous natural product expression

Due the limited capability for any particular bacterial host to recognize and utilize foreign DNA, heterologous expression becomes a major bottleneck in metagenomic and other genome-driven natural product discovery programs. This is well portrayed in our study, where we have yet to express five of the 14 novel gene clusters. Here we discuss some of the strategies overlapping with the field of synthetic biology for the heterologous expression of gene clusters, particularly the cryptic clusters that remain recalcitrant to expression using traditional methods. Approaches can broadly be divided into the engineering of biosynthetic pathways or the host strain.

Pathway manipulation is not only critical to the refactoring of cryptic gene cluster, but also in the general assembly of overlapping fragments. As mentioned previously, the complete biosynthetic gene cluster does not have to be contained in a single eDNA cosmid clone for its detection by homology-based screening. However, if the entire gene cluster spans multiple overlapping cosmids, genetic engineering tools must then be used to combine and assemble these separate fragments into one seamless entity for

heterologous expression. Such technique was not used in our study because all of the TD gene clusters were found to be contained on a single cosmid clone. However, if required, our laboratory routinely uses a method called transformation associated recombination (TAR) for the assembly of multiple overlapping clones (215). The robust homologous recombination machineries found in budding yeast (*Saccharomyces cerevisiae*) are exploited in this approach (235). By the transformation of linearized overlapping cosmid clones along with a capture vector carrying homology arms corresponding to sequences flanking the gene cluster in *S. cerevisiae*, these genetic elements undergo *in vivo* homologous recombination to yield a stable plasmid containing the complete intact gene cluster (236). Another method that utilizes *S. cerevisiae* is "DNA assembler," which also allows for the stitching of an entire gene cluster using *in vivo* homologous recombination (237). Bacterial *in vivo* homologous recombination methods, such as linear plus circular homologous recombination (LCHR) (238, 239), and linear plus linear homologous recombination (LLHR) (100), have also been developed for the assembly or direct cloning of biosynthetic gene clusters.

However, if the assembled gene cluster remains cryptic in a heterologous host, the pathway can be engineered to modify their native genetic elements, a process known as synthetic refactoring (125, 126). Much of the successful refactoring efforts for natural product biosynthetic gene clusters have been the replacement of native promoters with strong artificial promoters to drive transcription at high levels (123, 124, 240). Techniques not only include the cloning of individual genes or operons in expression vectors, as demonstrated in our study, but also using homologous recombination to exchange the promoter regions of the gene cluster. Synthetic refactoring becomes

increasingly time consuming and cumbersome with increasing size of gene cluster and number of genes/operons, and therefore the refactoring of natural product gene clusters has long been a challenging task. The *in vivo* homologous recombination approaches that allows for the rapid assembly of genetic elements, as previously described, should help facilitate this process. Moreover, since synthetic refactoring involve several artificial genetic elements, *in vitro* genetic engineering tools, such as Gibson assembly (241), sequence- and ligation-independent cloning (SLIC) (242), and Golden Gate assembly (243), should also prove to be useful for the construction of "synthetic" gene clusters.

Nevertheless, even with the use of strong artificial promoters, small molecule may still not be produced, or produced at very low levels. Issues at the transcriptional level include the fact that maximal transcription of all biosynthetic genes often do not lead to high-level small molecule production because different genes tend to have varying optimal transcriptional levels to minimize negative effects of high transcription, including metabolic flux and toxicity of the encoding protein or small molecules (244, 245). In addition, premature transcription termination and other negative transcriptional effects encoded outside of promoter regions, can still occur (246). The artificial gene orientation in the "synthetic" gene clusters may also cause problems (126), such as polymerase collision. Addressing all of these problems is challenging, especially for eDNA-derived gene clusters that lack detailed information regarding expression in their native settings, but future advances in synthetic biology should help overcome these issues.

Other factors to consider beyond transcription include ribosomal binding sites and codon usage for the proper translation (105), molecular chaperones for the proper protein folding (107), auxiliary proteins for the proper functioning of the enzyme (*e.g.*

phosphopantetheinyl transferase (PPTase) for NRPS and PKS) (247), stability of biosynthetic enzymes that may be susceptible to proteolytic activity (98), availability of necessary precursors (247, 248), and self-resistance. Ribosomal binding sites (rbs) and codon usage can be addressed using pathway manipulation by encoding rbs alongside with the promoters (105) and codon optimizing or randomizing the coding sequences (126), respectively. The remaining issues may be more appropriately addressed as part of the strain optimization process, a second synthetic biology-based approach to heterologous expression.

As in the case of the *mar* (NM343) cluster with the IPA-imine synthase, some gene clusters may not have all the biosynthetic genes necessary for small molecule production encoded within the gene cluster and rely on elements encoded within the endogenous host. Other prominent examples include deoxysugar biosynthetic genes, which are not present in certain actinomycetes-derived gene clusters that encode for compounds involving the attachment of that sugar for their biosynthesis (249). Similarly, other factors can also be considered as a deficiency in certain biochemical elements in a particular heterologous host that prevents small molecule production, including the lack of molecular chaperones, auxiliary proteins, stability factors, resistance elements. Even transcriptional and translational deficiencies can be attributed to the lack of the proper polymerases, transcriptional regulators, and transfer RNAs (tRNAs) in the heterologous host (98). Specific elements that are predicted to be necessary can be co-expressed with the gene cluster on a case-by-case basis, like the IPA-imine synthase gene for the *mar* cluster. Alternatively, considering that many of these elements are necessary for the expression of various biosynthetic gene clusters, they can be stably maintained in the

bacteria for the design of a universal heterologous host. Commercially available *Rosetta* (Novagen) *E. coli* strain that expresses rare tRNAs (250) and BAP1 *E. coli* strain with a chromosomally integrated Sfp PPTase for non-ribosomal peptide and polyketide biosynthesis (247), are some of the examples of bacterial hosts that have been optimized for heterologous expression.

The consideration of necessary biosynthetic precursors not just involves the addition of the appropriate biosynthetic genes, but also the engineering of the host strain to tune their metabolic flux. For example, several *Streptomyces* heterologous hosts have their genome minimized, or their endogenous secondary metabolite pathways knocked out, such that common biosynthetic precursors are not channeled into endogenous pathways (112, 251, 252). Another prominent example is the tuning of *E. coli* and *S. cerevisiae* strains for high-level production of farnesyl pyrophosphate (FPP) in their native ergosterol pathway, such that the anti-malarial drug artemisinin can be produced in these hosts at increased levels by the introduction of the appropriate biosynthetic genes (253, 254).

The necessity in our current natural product discovery pipeline to introduce the eDNA-derived pathways into multiple heterologous hosts implies that all of the possible factors that contribute to the successful heterologous expression of foreign biosynthetic gene clusters have not yet been characterized and cannot yet be implemented in a single bacterial strain. However, advances in synthetic biology should allow us to increase both the success rate and speed in the heterologous expression of eDNA-derived natural product pathway by permitting further improvements in both pathway manipulation and host optimization.

5.5 Metagenomic toolbox for natural product analog biosynthesis

The eDNA-derived TD gene clusters that have functionally been characterized to produce natural products can now be further leveraged to yield small molecules that will ordinarily not be made by the endogenous host. Such metabolites produced through the use of biosynthetic enzymes and thereby have structural resemblance to natural products, but are not exactly the compound encoded by the naturally-occurring biosynthetic gene clusters, are referred to as natural product analogs (255) or unnatural products (256). Here we do not consider total synthesis, biomimetic synthesis, and semisynthesis that utilize organic synthesis as a means to generate compounds, but discuss on the use of biosynthetic enzymes contained in the eDNA-derived gene clusters for small molecule production.

One of the key features of TD biosynthesis is the high reactivity of TD intermediates (*e.g.* activated tryptophan*)*. Once the biosynthetic pathway is allowed to proceed by the oxidation of tryptophan by IPA-imine synthase (*e.g.* StaO) or the dimerization of oxo-tryptophan by CPA synthase (*e.g.* StaD), many chemical reactions have been known to occur spontaneously, such as the cyclization of the pyrrole ring to form CPA from the conserved tryptophan dimer intermediate (Figure 34, 38) (45). As such, a number of biosynthetic enzymes, like VioE in violacein (94, 163, 164) and RebC/StaC in rebeccamycin/staurosporine biosynthesis (134), have been suggested to exhibit no catalytic activity, but nonetheless serve as a structural scaffold that directs the formation of a single desired product from a collection of shunt products that would

otherwise form spontaneously. Likely owing to the reactivity of these TD intermediates, a simple random mutagenesis of a violacein-producing bacterial strain in a previous study resulted in the production of various unnatural TDs (257). Using this strategy, referred to as mutasynthesis (256), a library of randomly mutagenized eDNA-derived gene clusters can similarly be introduced into the heterologous host for the production and isolation of novel unnatural products. The mutagenesis of biosynthetic genes can also be directed. For example, a targeted mutagenesis of a set of residues converted the FAD-binding monooxygenase RebC (from rebeccamycin pathway) to have StaC (from staurosporine pathway)-like functionality, thereby allowing for RebC to produce a staurosporine-like pyrrolinone indolocarbazole core instead of the usual maleimide indolocarbazole core, and vice versa (96, 97). As we gain further knowledge regarding the eDNA-derived TD biosynthetic enzymes, we should also be able to engineer them to have particular functions of interest, ultimately leading to the production of natural product analogs.

Instead of engineering the biosynthetic enzymes, the substrates that are to be fed into the enzymes can be modified as well, taking into advantage of substrate promiscuity in certain biosynthetic enzymes. This is referred to as chemoenzymatic synthesis, which is actually a more general term used to describe a process that utilizes a biosynthetic enzyme to conduct a particular chemical transformation on a compound (258). However, chemoenzymatic synthesis is the appropriate terminology to describe this unnatural product generation approach because a defined set of compounds (substrates) is generally fed into the biosynthetic enzyme for small molecule production. For example, a past study has grown a rebeccamycin-producing strain in a cultural broth containing high concentrations of bromide for the isolation of brominated rebeccamycin analogs (259).

Another study has fed different natural and synthetic indolocarbazole and indolocarbazole-like aglycons into the culture broth of bacterial host expressing the glycosyltransferase RebG, for the enzymatic generation of glycosylated indolocarbazole and indolocarbazole-like analogs (260). A similar strategy can also be done with eDNA-derived biosynthetic enzymes and gene clusters by feeding various substrates.

Lastly, TD analogs have been produced in the past by combinatorial biosynthesis. In this approach, the biosynthetic genes from staurosporine and rebeccamycin pathways were mixed and match in a heterologous host (*S. albus*) for the production of natural product analogs, including compounds with the sugar moiety from staurosporine, but the aglycon moiety from rebeccamycin, and vice versa (114). Taking into advantage the substrate promiscuity of the glyosyltransferase StaG, sugar biosynthetic genes from unrelated gene clusters have also been mixed and matched to generate analogs with various sugars attached (113). The potential for the generation of novel natural product analogs increases exponentially with the recovery of more TD biosynthetic genes from the eDNA library. The discovery and characterization of eDNA-derived TD gene clusters thereby not only allow us to isolate novel natural products, but also expand the metagenomic toolbox for the generation of even more natural product analogs.

Chapter 6: Materials and methods

Soil environmental DNA (eDNA) library construction. The three eDNA megalibraries, each consisting of over 10,000,000 unique cosmid clones, were constructed from the soil samples collected from the Anza-Borrego desert of California (AB), the Sonoran desert of Arizona (AR) and the Chihuahuan desert of New Mexico (NM), using published methods (28). Briefly, soil was resuspended in lysis buffer (100 mM Tris-HCl, 100 mM EDTA, 1.5 M NaCl, 1% (w/v) CTAB, 2% (w/v) SDS, pH 8.0) and heated for 2 hr at 70 °C. Soil particulates were removed by centrifugation (30 min, 4000 X g, 4 °C). Crude eDNA was precipitated from the resulting supernatant through the addition of 0.7 vol of isopropanol, pelleted (30 min, 4000 X g, 4 °C), washed with 70% ethanol, and pelleted once more (10 min, 4000 X g, 4 °C) to yield crude eDNA.

High molecular weight (HMW; ≥25 kb) eDNA was purified from crude eDNA by agarose gel electrophoresis (1% agarose gel, 16 hr, 20 V). The electroeluted (2 hr, 100V) HMW eDNA was concentrated (100 KDa molecular weight cut off), blunt-ended (End-It), ligated into cosmid vector, packed into λ phage (MaxPlax), and transfected into *E. coli* (EC100, Epicentre). The eDNA libraries were archived as unique sublibraries, each containing 4000-5000 clones. Matching DNA miniprep and glycerol stock pairs were generated for each sublibrary. DNA minipreps were arrayed such that sets of 8 sublibraries were combined to generate unique "row pools."

eDNA library homology guided screening of chromopyrrolic acid (CPA) synthase gene. A degenerate primer set was designed based on conserved regions of known CPA synthase genes from culture-based studies (accession no.: *vioB* AF172851.1, *staD* AB088119.1, *rebD* AJ414559.1, *inkD* DQ399653.1, *atmD* DQ297453.1.) Primers:

StaDV-F: GTS ATG MTS CAG TAC CTS TAC GC, StaDV-R: YTC VAG CTG RTA GYC SGG RTG (Table 3). The eDNA libraries were screened by performing PCR on miniprep DNA from each of the unique "row pools". Each 20 μL reaction consisted of 8.3 μL of water, 10 μL of FailSafe PCR Buffer G (Epicentre), 0.5 μL each of StaDVF and StaDVR primers (final concentration of 2.5 μM each), 0.5 μL of template "row pool" eDNA (100 ng), and 0.2 μL *Taq* DNA polymerase (New England Biolabs). PCR cycling conditions were as follows: 1 cycle of 95°C for 5 min; 7 cycles of 95°C for 30 sec, 65°C for 30 sec with 1°C decrement per cycle to 59°C, 72°C for 40 sec; 30 cycles of 95°C for 30 sec, 58°C for 30 sec, 72°C for 40 sec; 1 cycle of 72°C for 7 min; 4°C hold. Amplicons of the correct size (~561 base pairs) were gel purified, re-amplified and sequenced using the same degenerate primers. Amplicons that were confirmed to be CPA synthase gene sequences based on BLASTX homology searches (NCBI) were used to guide the recovery of the corresponding cosmid clones from within our eDNA megalibraries.

Recovery of cosmid clones harboring tryptophan dimer (TD) gene clusters. Cosmid clones containing CPA synthase genes were recovered from the archived eDNA libraries using a serial dilution approach. For each amplicon of interest, a specific PCR primer set was designed to recognize the sequence of that particular amplicon. These primers were used to identify, from a given "row pool," the corresponding sublibrary that contains the clone of interest. The sublibrary glycerol stock was resuspended into LB to an OD_{600} of 0.5, diluted 2 x 10^5 fold and arrayed as 60 μL aliquots (about 25 cells) into 4 sterile 96 well plates. Upon overnight growth, the well containing the clone of interest was identified by whole cell PCR. The culture broth from this well was then spread onto LB plates and single colonies were screened by colony PCR to identify the specific clone

harboring the targeted CPA synthase gene. Cosmid clones were *de novo* sequenced at the Sloan Kettering Institute DNA Sequencing Core Facility using 454 pyrosequencing technology (Roche). Clone assemblies were annotated using FGENESB (Softberry) or CloVR (261) for gene prediction and BLASTP (NCBI) for protein homology relationships.

Phylogenetic tree construction of CPA synthase genes. The ClustalW alignment was performed on the sequences of culture-derived and eDNA-derived CPA synthase genes using MacVector version 12.0.3 (Open Gap Penalty: 10.0; Extend Gap Penalty: 5.0; Pairwise Alignment Mode: Slow). The corresponding phylogenetic tree was constructed from the alignment using the non-TD pathway related CPA synthase-like hypothetical gene Riv7116_4841 from *Rivularia sp. PCC 7116* (accession no.: CP003549.1) as an outgroup for rooting (Best Tree Mode; Tree Building Method: Neighbor Joining; Distance: Tajima-Nei).

Retrofitting and conjugation of cosmid clones into *S. albus*. Each cosmid clone containing the TD pathway (*e.g.* AB1650, AB1091) was digested with *AanI* and ligated with the 6.8 kb *DraI* fragment from the *E. coli*/Streptomyces shuttle vector pOJ436. This fragment contains the origin of transfer (*oriT*), apramycin resistance marker, and the Streptomyces φC31 integration system (115). The retrofitted pathways were each transformed into *S. albus* by conjugation with *E. coli* S17.1 using published methods (115). Exconjugants were selected on mannitol soy flour medium (MS) using an apramycin (25 μg/mL) and naladixic acid (25 μg/mL) overlay. Successful exconjugants were re-struck on MS plates and grown for an additional 5 days before harvesting spores.

Isolation and purification of AB1650 specific metabolites. Tryptone soya broth

(Oxoid) seed cultures grown for 18 h at 30 °C (200 rpm) were used to inoculate 8 liters of R5A media (3). After 12 days (30 °C, 200 rpm) the entire culture was extracted with 2 volumes of ethyl acetate. The resulting crude extract was partitioned by silica gel flash chromatography using a $CHCl_3$:MeOH step gradient. Compounds **1** and **3 - 6** were purified from the 99:1 ($CHCl_3$:MeOH) fraction using reversed phase HPLC (linear gradient from 55:45 H_2O:MeOH to 30:70 H_2O:MeOH over 30 min, XBridge C18, 150 X 10 mm, 5 μm). Compound **1** (21.5 mg) eluted with 36:64 H_2O:MeOH. Compound **4** (12.7 mg) eluted with 45:55 H2O:MeOH. Compound **6** (3.9 mg) eluted with 48:52 H_2O:MeOH. Compounds **3** (1.8 mg) and **5** (5.3 mg) co-eluted with 42:58 H_2O:MeOH and were separated from each other using an isocratic (48:52 H_2O:MeOH) HPLC step. NMR data was recorded on a 600-MHz spectrometer (Bruker). HRMS data was obtained using Thermo LTQ-Orbitrap mass spectrometer (Rockefeller University). Analytical LC/MS data was recorded on a Waters Micromass ZQ instrument.

Transposon mutagenesis. A library of transposon mutants was generated using the HyperMu <KAN-1> insertion kit from Epicentre following the manufacturer's protocol. Transposon insertions were located by single pass sequencing (MCLAB) using a transposon-specific primer (Epicentre). *S. albus* recombinants harboring AB1650 cosmids containing transposon insertions in individual genes of interest were grown in R5A media (2 liters) as described above. Ethyl acetate extracts from these cultures were initially fractioned by silica gel flash chromatography using the methods outlined above. Compound **2** (2.9 mg), compound **7** (1.6 mg) and compound **8** (1.9 mg) eluted with 99:1 $CHCl_3$:MeOH fractions and were purified by reversed phase HPLC using 65:35, 60:40, and 70:30 H_2O:CH_3CN isocratic conditions, respectively. HRMS data was obtained using

a Waters LCT Premier XE mass spectrometer (Sloan-Kettering Institute).

Bioactivity assays. For cytotoxicity assays against bacteria and yeast, overnight cultures (37°C, 200 rpm) were diluted 10^{-6} fold and arrayed as 100 μl aliquots into sterile 96 well microtiter plates. An ampicillin control and compounds resuspended in dimethylsulfoxide (DMSO) were added at an initial concentration of 25 μg/mL (or 50 μg/mL in some cases) and serially diluted 2-fold across the plate such that the final concentrations were 25, 13, 6.3, 3.1, 1.6, 0.78, 0.39, 0.20, 0.098, 0.049, 0.024, 0.012 μg/mL. A DMSO control was similarly diluted across one row of the plate. Plates were incubated at 30 °C for 24 hr. The lowest concentration with no observable growth (OD < 0.05) is reported as the minimum inhibitory concentration (MIC) for each metabolite.

For cytotoxicity assay against human colon tumor cell line HCT116 (ATCC: CCL-247), the cells were grown in McCoy's 5A Media (modified, Invitrogen) supplemented with 10% fetal bovine serum and 1% (w/v) Penicillin/Streptomycin (37 °C with 5% CO_2). Cells were seeded as 100 μl aliquots into sterile 96 well plates at approximately 1,000 cells per plate and incubated for 24 hr before adding compounds. A DMSO control and compounds resuspended in DMSO were dissolved in fresh media and added to the cells at final concentrations of 50, 25, 13, 6.3, 3.1, 1.6, 0.78, 0.39, 0.20, 0.098, 0.049, 0.024 μg/mL and grown for additional 72 hr. The cell density in each well was then determined with a crystal violet assay (262). Briefly, the cells were washed with phosphate buffered saline and fixed with 4% formaldehyde in phosphate buffered saline (10 min, 24 °C). After an additional wash with phosphate buffered saline, the cells were stained with 0.1% (w/v) filtered crystal violet solution (30 min, 24 °C), washed with water and air-dried. 10% acetic acid was added to the stained cells to extract the dye and

the absorbance at 590 nm was measured using a microplate spectrophotometer (Epoch, BioTek). The normalized absorbance values were plotted and curve fitted using Graphpad Prism to determine the half maximal inhibitory concentrations (IC_{50}).

***BorR* expression.** *BorR* was PCR amplified from AB1091 using Phusion polymerase (New England Biolabs) and the following primer pair: BorR-F: GAGA*CATATG*AAGACTCTGCCGGGTCG, BorR-R: GAGA*TTAATTAA*CTACCGCGCTTCTCGGAG (NdeI and PacI sites added for cloning are shown in italics; Table 3). PCR cycling conditions: 1 cycle of 98°C for 1 min; 40 cycles of 98°C for 10 sec, 63°C for 25 sec, 72°C for 1 min 30 sec; 1 cycle of 72°C for 7 min; 4°C hold. NdeI/PacI digested PCR product was cloned into NdeI/PacI digested pIJ10257. (263). This construct was then moved by conjugation into *S. albus* harboring the *bor* pathway and resulting exconjugants were selected using hygromycin (100 μg/mL). Successful exconjugants were re-struck on MS plates and grown for an additional 5 days before harvesting spores.

Borregomycin production and isolation. *S. albus* AB1091 spore stocks (both with and without the pIJ10257 *borR* expression construct) were used to inoculate 50 mL aliquots of R5A media (57) in 200 mL baffled flasks (10 liters total). After 12 days (30°C, 200 rpm) the cultures were pooled and extracted with 2 volumes of ethyl acetate. The resulting crude extract was dried *in vacuo*, dissolved in 90:10 methanol:water and then partitioned with hexane, methylene chloride, and ethyl acetate using modified Kupchan scheme (264). The methylene chloride fraction was separated by silica gel RediSep flash chromatography (RediSepRf 12 gram silica flash column: 5 min 100% hexane, 35 min linear gradient from 100% hexanes to 100% ethyl acetate, 5 min of 100% ethyl acetate).

Compounds **9** and **14** co-eluted with the 75:25 hexanes:ethyl acetate fraction, **10** eluted with the 70:30 hexanes:ethyl acetate fraction, **11** eluted with the 65:35 hexanes:ethyl acetate fraction, **12** eluted with the 60:40 hexanes:ethyl acetate fraction and **13** eluted with the 10:90 hexanes:ethyl acetate fraction. Compounds were purified from these fractions using isocratic reversed phase HPLC (XBridge C18, 150 X 10 mm, 5 μm). **9** (0.7 mg) and **14** (0.7 mg) were separated from each other using 62:38 methanol:water. **10** (1.4 mg) was purified using 65:35 methanol:water. **11** (1.8 mg) was purified using 57:43 methanol:water. **12** (0.6 mg) was purified using 50:50 methanol:water. **13** (0.5 mg) was purified using 52:48 methanol:water. Analytical LC/MS data was recorded on a Micromass ZQ instrument (Waters). NMR data was obtained using a 600-MHz spectrometer (Bruker). HRMS data was acquired using a LCT Premier XE mass spectrometer (Waters) at the Sloan Kettering Institute Analytical Core Facility and LTQ-Orbitrap mass spectrometer (Thermo Scientific) at the Rockefeller University Proteomics Resource Center. Specific rotation was measured using P-1020 Polarimeter (Jasco).

KinaseProfiler assay. Compounds were submitted to EMD Millipore to determine their inhibitory activity against a panel of kinases (KinaseProfiler) and for IC_{50} (IC_{50}Profiler) measurements. Radiometric and homogenous time-resolved fluorescence (HTRF) based methods were used to measure the incorporation of phosphate into the substrate in the presence of the compound for protein and lipid kinase assays, respectively. The activity values were normalized against readings taken in the absence of any added compound. Please refer to the EMD Millipore website (www.millipore.com) for additional protocol details.

Synthetic refactoring of the *esp* gene cluster. The biosynthetic genes from the *esp* gene cluster in cosmid clone AB339 were amplified using the manufacturer's recommended *Phusion Hot Start Flex* DNA polymerase reaction conditions (New England Biolabs). PCR primers are listed in Table 3. PCR cycling conditions were as follows: 1 cycle of 95°C for 5 min; 30 cycles of 95°C for 10 sec, 62°C for 30 sec, 72°C for 30 sec/kb; 1 cycle of 72°C for 7 min; 4°C hold. The resulting amplicons were digested and cloned into the following Duet vectors: EspM NdeI/MfeI site of pCDFDuet-1; EspX NcoI/HindIII site of pCDFDuet-1; EspO NcoI/HindIII site of pCOLADuet-1; EspD NdeI/MfeI site of pCOLADuet-1; EspP NcoI/HindIII site of pETDuet-1.

Induced expression analysis of refactored *esp* gene cluster. Electrocompetent *E. coli* BL21 cells were transformed with constructs containing various combinations of *esp* genes, grown in LB medium in the presence of the appropriate antibiotics (spectinomycin 100 μg/mL; kanamycin 30 μg/mL; ampicillin 100 μg/mL) and induced at OD_{600} of 0.5 by the addition of IPTG to a final concentration of 0.1 mM. After growth for between 6 and 36 h (200 rpm, 25 °C), the culture was extracted with ethyl acetate and dried *in vacuo*. Upon resuspension in methanol, the samples were subjected to reversed phase LC/MS analysis (150 X 4.6 mm, 5 μm XBridge C18: linear gradient of 80:20 water:methanol to 0:100 water:methanol). Analytical LC/MS was obtained using Micromass ZQ mass spectrometer (Waters).

Isolation and purification of the erdasporines. For compound **15**, 2 liters of EspODPM expressing *E. coli* BL21 culture was grown for 12 hr (200 rpm, 25 °C) after IPTG induction. The culture was pelleted by centrifugation (10 min, 4000 X g, 25 °C) and the resulting supernatant was added to 2 liters of EspX expressing *E. coli* BL21 culture that

had been grown for 15 minutes after IPTG induction. The combined culture was then grown for 6 hr (200 rpm, 25 °C) before extraction with 2 volumes of ethyl acetate.

For compounds **16** and **17**, 2 liters of EspODPM expressing *E. coli* BL21 culture grown for 12 hr (compound **16**) or 36 hr (compound **17**) after IPTG induction was extracted with 2 volumes of ethyl acetate.

Organic extracts were separated by silica gel RediSep flash chromatography (RediSepRf 12 gram silica flash column: 3 min 100% chloroform, 27 min linear gradient from 100% chloroform to 90:10 chlorofom:methanol). Compound **16** eluted with 99:1, **17** eluted with 96:4 and **15** eluted with 95:5. Compounds **15-17** were purified from these fractions using isocratic reserved phase HPLC (150 X 10 mm, 5 μm XBridge C18). **16** (2.3 mg) was purified using 64:38 water:acetonitrile. **17** (1.1 mg) was purified using 68:32 water:acetonitrile. Compound **15** was first fractionated using 70:30 water:acetonitrile and subsequently purified (0.4 mg) using silica gel flash chromatography (0.5 gram of Silica Gel 60 packed in a glass pipette) with an isocratic 70:30 hexane:ethyl acetate mobile phase. HRMS data was obtained using a LTQ-Orbitrap mass spectrometer (Thermo Scientific). NMR data was acquired using a 600-MHz spectrometer (Bruker). Specific rotation was measured using a P-1020 Polarimeter (Jasco).

Induced expression analysis of eDNA-derived Group F clusters. The pathway-specific methyltransferase and monooxygenase genes from the Group F clusters found in cosmid clones AB234, AB1149, NM1499 and AB1521 were amplified using the same PCR reaction and cycling conditions that was done for clone AB339 amplification reactions. PCR primers are listed in Table 3. The methyltransferase and the monooxygenase

amplicons were cloned into the NdeI/MfeI or NcoI/HindIII sites of pCDFDuet-1, respectively.

The heterologous expression was conducted in a similar manner to that used to study the *esp* gene cluster. Briefly, 25 mL of each methyltransferase/EspODP co-expressing *E. coli* BL21 culture was grown for 12 hr after IPTG induction. This culture was pelleted by centrifugation and the resulting supernatant was added to 25 mL of a monooxygenase expressing *E. coli* BL21 culture that had been grown for 15 minutes after IPTG induction. The cultures were then grown for an additional 6 hr and then extracted with ethyl acetate. Extracts were subjected to reversed phase LC/MS analysis as described above.

Cloning of genes from the *mar* and violacein (*vio*) gene clusters. Individual genes were amplified from the *mar* and *vio* clusters using clones NM343 and CSL51 (155) as templates, respectively, using *Phusion Hot Start Flex* DNA polymerase kit (New England Biolabs). Primers are shown in Table 3. PCR cycling conditions: 1 cycle of 95°C for 5 min; 30 cycles of 95°C for 10 sec, 62°C for 30 sec, 72°C for 30 sec/kb; 1 cycle of 72°C for 7 min; 4°C hold. Gel purified amplicons were restriction digested and cloned into the following Duet (Novagen) vectors: MarB NcoI/SalI sites of pCOLADuet-1; MarC NcoI/HindIII sites of pETDuet-1; MarE NdeI/KpnI sites of pETDuet-1; MarM NcoI/SalI sites of pCDFDuet-1; VioA NdeI/MfeI sites of pCOLADuet-1; VioB NcoI/HindIII sites of pCOLADuet-1; VioE NdeI/KpnI sites of pETDuet-1.

Heterologous expression of *mar* and *vio* biosynthetic genes. For expression studies, electrocompetent *E. coli* BL21 cells were transformed with Duet vectors harboring various combinations of *mar* and *vio* biosynthetic genes and grown in LB medium with

required antibiotic combination for selection (30 μg/mL kanamycin, 100 μg/mL ampicillin, 100 μg/mL spectinomycin). Gene expression was induced in cultures grown to an OD_{600} of 0.5 with the addition of IPTG (final concentration of 0.1 mM). Thirty-six hours (200 rpm, 25 °C) post induction, the cultures were extracted with ethyl acetate that was acidified to pH ~3-4 with the addition of hydrochloric acid. Extracts dried *in vacuo* were dissolved in methanol and subjected to reversed phase LC/MS analysis (150 X 4.6 mm, 5 μm XBridge C18: linear gradient of 80:20 water:methanol to 0:100 water:methanol with 0.1% formic acid). Commercially available methylarcyriarubin (*i.e.* Bisindolylmaleimide V, Santa Cruz Biotechnology), arcyriarubin A (*i.e.* Bisindolylmaleimide IV, Santa Cruz Biotechnology), and indole-3-pyruvic acid (Sigma-Aldrich) were run on the LC/MS, under the same conditions. Because chromopyrrolic acid and prodeoxyviolacein were not commercially available, these standards were prepared from the heterologous expression of well-defined biosynthetic genes from the violacein pathway. Chromopyrrolic acid and prodexoyviolacein were produced from VioAB and VioABE expressing *E. coli* cultures, respectively. They were each purified from culture broth extracts by HPLC using conditions based on previous violacein pathway studies (94, 163, 164). Analytical LC/MS data was acquired using Micromass ZQ mass spectrometer (Waters).

Large-scale production and isolation of methylarcyriarubin (21) from *E. coli* cultures expressing *mar* genes. Cultures of VioA + MarBCEM expressing *E. coli* BL21 cells (2 liters) grown for 36 hr (200 rpm, 25 °C) after IPTG induction was extracted with ethyl acetate (4 liters). This extract was initially fractionated by silica gel RediSep flash chromatography (RediSepRf 12 gram silica flash column: 3 min 100% chloroform, 27

min linear gradient from 100% chloroform to 85:15 chlorofom:methanol). Compound **21** eluted with 99:1 chloroform:methanol. Compound **21** was then purified (1.6 mg/L) from the 99:1 fraction by 65:35 water:acetonitrile isocratic reserved phase HPLC (150 X 10 mm, 5 μm XBridge C18). LTQ-Orbitrap mass spectrometer (Thermo Scientific) and 600-MHz spectrometer (Bruker) were used to acquire HRMS and NMR data, respectively, for structure elucidation studies.

MarE protein expression analysis. A liquid culture (100 mL) of *E. coli* harboring *vioE*/pETDuet-1 was grown to an OD_{600} of 0.5. The culture was subsequently split into two, with one uninduced and the other induced with the addition of IPTG (final concentration of 0.1 mM). Two hours (200 rpm, 37 °C) post induction, an aliquot (1 mL) was removed from each culture and the cells were pelleted by centrifugation (1 min, 13,000 X g). The cells were resuspended in native purification buffer (200 µL; 0.5 M NaCl, 50 mM NaH_2PO_4, pH 8.0) and lysed by sonication (30 cycles of 1 sec pulse on and 2 sec pulse off, 45% amplitude; Sonic Dismembrator, Fisher Scientific). The cell debris and the insoluble proteins were collected by centrifugation (15 min, 13,000 X g, 4 °C). Aliquots (20 µL) of the supernatant were mixed with SDS loading buffer (50 mM Tris-HCl pH 6.8, 2% SDS, 10% glycerol, 1% β-mercaptoethanol, 12.5 mM EDTA, 0.02% bromophenol blue), heated (10 min, 95 °C) and run on a polyacrylamide gel (4-20% Mini-PROTEAN TGX Gel with Precision Plus Protein Dual Color Standards, Bio-Rad). Gels were stained with coomassie (Coomassie Brilliant Blue R-250 Staining Solution, Bio-Rad) and imaged using Gel Doc XR+ System (Bio-Rad).

Reagents and strains for resistant mutant screening. The multidrug resistance-suppressed (MDR-sup) *S. pombe*, SAK84 and SAK690, and the MDR-active *S. pombe*

SAK1, from which MDR-sup *S. pombe* was derived, were generously provided by Dr. Tarun M. Kapoor (Laboratory of Chemistry and Cell Biology, The Rockefeller University). The genotypes (265) of these strains are listed in Table 4. The indolotryptoline-based compounds, BE-54017 and cladoniamide A, were isolated from *Streptomyces albus* harboring the *abe* gene cluster, as described previously. Bafiolomycin A1, concanamycin A, and brefeldin A were purchased from a commercial supplier (Santa Cruz Biotechnology).

***Schizosaccharomyces pombe* whole-cell cytotoxicity assay.** Freshly struck multidrug resistance-suppressed (MDR-sup) *S. pombe* SAK84 was inoculated into liquid YE4S media and grown (30˚C, 300 rpm) to log phase (OD_{595} = 0.5). The culture was diluted 50-fold and distributed as 100 μl aliquots into a sterile 96 well microtiter plate. BE-54017 or cladoniamide A resuspended in DMSO were added to the first well at the initial concentration of 0.50 μg/ml and were serially diluted 2-fold across the plate (final concentrations 0.50, 0.25, 0.13, 0.063, 0.031, 0.016, 0.0078, 0.0039, 0.0020, 0.0010, 0.00050, 0.00025 μg/ml). A compound-free DMSO control was similarly diluted across the plate. After outgrowth (36 hr, 30˚C, 300 rpm), the absorbance (OD_{595}) of each well was measured using a microplate reader (Epoch Microplate Spectrophotometer; BioTek). Using Graphpad Prism, the normalized absorbance values were plotted and curve-fitted to determine the half maximal inhibitory concentration (IC_{50}) for each indolotryptoline. The same method (using different initial drug concentrations) was used to determine IC_{50}s for bafilomycin A1, concanamycin A and brefeldin A against resistant and non-mutant strains.

Selection of indolotryptoline resistant *S. pombe* mutants. 20 ml of log phase *S. pombe* SAK84 was pelleted by centrifugation (3000 x g, 3 min) and resuspended in TM buffer (50 mM Tris, 50 mM maleic acid, 7.5 mM $(NH_4)_2SO_4$, 0.4 mM $MgSO_4$, pH6.0) containing 50 μg/ml methylnitronitrosoguanidine (NTG) to randomly mutagenize the genome. After 30 min at 32˚C (250 rpm), the mutagenized cells was pelleted, washed twice with 10 ml sterile water, resuspended in 20 ml fresh YE4S, and allowed to recover for 3 hr (32˚C, 250 rpm). The culture was adjusted to OD_{595} = 0.5 and 150 μl aliquots were spread onto YE4S plates containing different concentrations of the indolotryptolines. After 72 hr at 32˚C, resistant clones were picked from plates containing either ~10 (BE-54017: 0.031 μg/ml; cladoniamide A: 0.078 μg/ml) or ~50 (BE-54017: 0.016 μg/ml; cladoniamide A: 0.039 μg/ml) colonies. Each strain was then re-assessed for indolotryptoline resistance and cross-resistance to brefeldin A using the whole-cell cytotoxicity assay described above.

Backcrossing of indolotryptoline resistant mutants. Resistant strains were crossed with non-mutant MDR-sup strain SAK690, which only differs in genetic background from SAK84 by having a different mating type (h-). Both resistant and non-mutant *S. pombe* strains grown on YE4S plates were resuspended in water to produce suspensions with an OD_{595} = ~1. These were mixed in equal volumes and 10 μl aliquots were spotted onto an SPA plate. After 40 hr at 25˚C, the mixture was struck onto a YE4S plate. Using a dissecting microscope/micromanipulator (Axioskop 40, Zeiss), zygotic asci (mating products), were isolated from the YE4S plate and incubated (37˚C, 6 hr) to permit native digestion of the ascus wall. The four spores from each zygotic ascus were then separated and individually grown on YE4S plate (30˚C, 5 days). The resulting colonies were tested

for resistance to indolotryptolines and resistant colonies were used in subsequent rounds of backcrossing. For additional rounds of backcrossing, the progeny that retained resistance were crossed, depending on its mating type, with either SAK84 or SAK690.

Whole-genome sequencing and bioinformatics. Backcrossed mutants were grown in YE4S and genomic DNA was isolated from these cultures using zymolyase treatment followed by phenol/chloroform extraction (266). Genomic DNA from six resistant mutants and two non-mutant strains (SAK84 and SAK690) was sequenced at the Rockefeller University Genomics Resource Center using Illumina HiSeq 2000 technology (50 bp single-end, ~150 million reads in total). Reads from resistant mutants were compared to non-mutant samples to identify resistant mutant specific somatic mutations that altered the wildtype amino acid sequence with $\geq$ 4x coverage and >50% mutation allele frequency. In brief, the variant detection pipeline consisted of the mapping of Illumina reads to the *S. pombe* genome (193) using BWA, removal of duplicates, indel-based realignment using GATK, base quality score recalibration, mutation calling for single nucleotide variant using GATK Unified Genotyper, and annotation using SnpEff (Sloan Kettering Bioinformatics Core).

PCR sequencing of V-ATPase proteolipid subunit genes. A fresh colony of each resistant strain that was not sequenced by HiSeq was resuspended in 0.2% SDS and heated at 95°C for 10 min. One μl of this crude cell lysate was used as template in PCR reactions designed to amplify the *vma3, vma11* and *zhf1* genes (*Phusion Hot Start Flex* DNA polymerase kit, New England Biolabs). Primers are shown in Table 3. PCR cycling conditions: 1 cycle of 95°C for 2 min; 30 cycles of 95°C for 10 sec, 58°C for 30 sec,

72°C for 30 sec/kb; 1 cycle of 72°C for 7 min; 4°C hold. The resulting amplicons were sequenced from both ends using the same set of primers that were used for PCR.

Targeted mutagenesis of *S. pombe* genome. *Vma3* or *vma11* specific recombination cassettes containing point mutations of interest were amplified from the appropriate resistant mutant using the same *Phusion Hot Start Flex* PCR conditions described previously. Primers were designed (Table 3) to generate amplicons with ~500 bp homology arms flanking each side of the point mutation of interest. The PCR cassette was introduced into SAK84 by lithium acetate-assisted transformation (266). The transformation reaction was spread onto YE4S plates containing defined concentrations of indolotryptoline to select for strains conferring drug resistance. The acquisition of drug resistance and point mutation were confirmed by whole-cell cytotoxicity assay and PCR sequencing, respectively.

V-ATPase activity assay by acidic organelle staining. Log phase *S. pombe* was added to fresh indolotryptoline containing media (final concentrations = 0.50, 0.25, 0.13, 0.063, 0.031, 0.016, 0.0078, 0.0039 μg/ml) to give 5 ml cultures with OD_{595} = 0.15. These cultures were grown at 30°C for 1 hr, pelleted (2 min, 3000 x g) and washed with YE4S buffered with 50 mM MOPS (pH 7.6). The cell pellet was then resuspended in buffered YE4S containing 200 μM quinacrine and staining was allowed to proceed at room temperature for 10 min. After washing with buffered YE4S, the samples were resuspended in the same media and transferred to an 8-chambered coverglass. Cells were imaged at the Rockefeller University Bio-Imagining Resource Center under a fluorescent microscope using 100x objective lens with DIC optics for Nomarski imaging or with FITC filter set for quinacrine visualization (DeltaVision Image Restoration Microscope

System with Olympus IX-70 base microscope, Applied Precision). A minimal inhibitory concentration (MIC) was defined as the dose at which the formation of fluorescent puncta was inhibited in >95% of the cells.

Mapping of resistance-conferring residues onto a V-ATPase structure. The *Enterococcus hirae* Na^+-ATPase proteolipid subunit, NtpK (PDB ID: 2bl2), was imaged using PyMOL. A CLUSTALW alignment of the *S. pombe* V-ATPase proteolipid subunits Vma3 and Vma11 and NtpK was created to map residues between proteins from the two organisms. Residues that confer resistance to indolotryptoline and plecomacrolide compounds were identified in NtpK based on this alignment. Side chains for resistance conferring residues were converted to those seen in wildtype Vma3 or Vma11 and then represented as colored sticks on the NtpK structure.

TD diversity analysis from crude eDNA. Topsoil was collected from 20 distinct sites in New Mexico. Crude eDNA was extracted from each sample using the same initial protocol as described for eDNA library construction. However, instead of purifying HMW eDNA by gel electrophoresis, crude eDNA was cleaned with two rounds of column based purification (PowerClean DNA Clean-Up Kit, MO-BIO).

For crude eDNA screening, the forward primer of the degenerate primer (StaDV-F) was modified at the 5' end to permit the direct 454 sequencing (Roche) of amplicons from all 20 samples simultaneously. Forward primers each contained a 454 sequencing adapter tag (CGTATCGCCTCCCTCGCGCCATCAG), followed by a unique 8 base pair barcode for each soil sample and then the StaDV-F degenerate sequence. A unique StaDV-F/StaDV-R pair was then used to amplify CPA synthase gene fragments from each soil sample. Each 20 μL PCR reaction consisted of 8.3 μL of water, 10 μL of FailSafe PCR

Buffer D (Epicentre), 0.5 μL each of modified StaDV-F and StaDV-R primers (final concentration of 2.5 μM each), 0.5 μL of template crude eDNA (100 ng) and 0.2 μL *Taq* DNA polymerase (New England Biolabs). PCR cycling conditions were as follows: 1 cycle of 95°C for 5 min; 30 cycles of 95°C for 30 sec, 59°C for 30 sec, 72°C for 40 sec; 1 cycle of 72°C for 7 min; 4°C hold. Amplicons of the correct size (561 base pairs) were gel purified and processed for single-end read sequencing using the 454 GS-GLX Titanium platform at the Sloan Kettering Institute DNA Sequencing Core Facility.

The raw reads were initially processed using the Qiime software suite (version 1.6) which utilizes size cutoff, quality cutoff, insertion/deletion removal and chimera removal filters to retain only high quality reads. Only reads with lengths >500 base pairs were retained and they were subsequently trimmed to 450 base pairs (from the 3' end where sequencing errors occur most frequently). Reads were then clustered at 95% identity and only the amplicon sequences that were populated with more than 30 reads from a single soil sample were retained for phylogenetic analysis. Consensus amplicons that were either unrelated to CPA synthase genes or were potentially chimeric based on BLASTX homology searches (NCBI) were also removed, leaving 31 sequences for analysis (Appendix 7).

A ClustalW alignment was performed on these 31 amplicon sequences plus all known full-length CPA synthase genes using MacVector version 12.0.3 (Open Gap Penalty: 10.0; Extend Gap Penalty: 5.0; Pairwise Alignment Mode: Slow). Using this alignment as a guide, full-length CPA synthase genes were trimmed to match the 450 bp amplicon sequences and a final phylogenetic tree was constructed using the hypothetical gene Riv7116_4841 as an outgroup for rooting (Best Tree Mode; Tree Building Method:

Neighbor Joining; Distance: Tajima-Nei). This tree was reformatted to a circular display using iTOL.

Table 3. PCR primer list. Underlined sequences indicate the restriction sites added for ligation.

Gene	Sequence
StaDV-F	GTSATGMTSCAGTACCTSTACGC
StaDV-R	YTCVAGCTGRTAGYCSGGRTG
BorR-F	GAGAcatATGAAGACTCTGCCGGGTCG
BorR-R	GAGAttaattaaCTACCGCGCTTCTCGGAG
EspM-F	GAGAcatATGACTTGGAGCCCAGGAATG
EspM-R	GAGAcaattgGCCATTGTGGTCATCGCG
EspX-F	GAGAacatgtctATGGTGCATGACGTTGACGTG
EspX-R	GAGAaagcttCTCATCCCACCTCACTGAAG
EspO-F	GAGAccatggcaATGAGGTGGGATGAGGCG
EspO-R	GAGAgcggccgcCACTCATCACGACACCTCC
EspD-F	GAGAcatATGAGTGTTTTTGATCTGCCCC
EspD-R	GAGAcaattgCGTCATTGTTCGTCCTCGG
EspP-F	GAGAtcatgatcATGACGCAGCGCGGTACA
EspP-R	GAGAaagcttCGTTCACAGCGGATCGAC
234M-F	GAGAcatATGACAGAACAGCGGCTGAC
234M-R	GAGAcaattgTCTCAAGCCGTCTTGACCTC
234X-F	GAGAacatgtctATGAAGTTCGACGTTGACGTGC
234X-R	GAGAaagcttTCACGTCCACCACTCTTTTTCG
1149M-F	GAGAcatATGGATGACACCAACCAGCAAAC
1149M-R	GAGAcaattgATTGCTTCGCGGGTGATGC
1149X-F	GAGAacatgtctATGAAACACGACGTCGATGTCC
1149X-R	GAGAaagcttCGATGGGTATTCGAGCCG
1499M-F	GAGAcatATGACATCCGGACCCGATC
1499M-R	GAGAcaattgTTGGATGGTCGGCGGTCA
1499X-F	GAGAacatgtctATGCAAGACGACGTGGAAGTG
1499X-R	GAGAaagcttGGTGGAAGATCAGCAAGCC
1521M-F	GAGAcatATGGACGACACCAACCAGC

1521M-R	GAGAcaattgTGGGTGACACGGGTACGA
1521X-F	GAGAacatgtctATGGAAACCCCAGACGTTGATG
1521X-R	GAGAaagcttCTTCATGGCAGTCTCCTTGC
MarB-F	GAGAccatggcaATGAGCATCCTGGAATTTCCGC
MarB-R	GAGAgtcgacCCTCACAAGAGTGGAACGG
MarC-F	GAGAtcatgatcATGCTGAGCGCCGAAGACA
MarC-R	GAGAaagcttCTCATGCGGTCTCCTTGC
MarE-F	GAGAcatATGAGCGCCGCCCGC
MarE-R	GAGAggtaccGAGGATTGTTGGTCTGCTGAC
MarM-F	GAGAacatgtctATGACAACTCAGGGAACGCC
MarM-R	GAGAgtcgacGCTCAGCGTTCTTTCGTGC
VioA-F	GAGAcatATGACAAACTATTCCGACATTTGC
VioA-R	GAGAcaattgGGAAATCCAGAATGCTCATGC
VioB-F	GAGAccatggcaATGAGCATTCTGGATTTCCCC
VioB-R	GAGAaagcttTGCATATCAAGCCTCTCTAGAC
VioE-F	GCGCcatATGCCGATGCCTGTCCAC
VioE-R	GCGCggtaccCACAAACGGAACAGGACTCAGT
seq-vma3-F	CGACATTGTAAAAGCCAGCT
seq-vma3-R	TCCCACCATAGAGATTCTC
seq-vma11-F	CAACGAAATACTACATCGACA
seq-vma11-R	TGATTAGCCTTAGAGAAAGTC
seq-zhf1-F	ATATAGCAAGTTTGCGCCTC
seq-zhf1-R	GTGACACAATAGATTAACCACG
mut-vma3-F	CGATACGACATTGTAAAAGCC
mut-vma3-R	CGTGAAGTACATGCTTATACG
mut-vma11-F	AGAACTTGTGCCAAAAGTCC
mut-vma11-R	GCCTTAGAGAAAGTCAACAAG
mut-zhf1-F	TTGTGGTAAACGCGATTAGTG
mut-zhf1-R	CTAACGAGAAGAATCAAACC

Table 4. *Schizosaccharomyces pombe* strain list.

Name	Genotype
SAK1	h+; *ade6-M210 leu1 ura4-D18*
	h+; *ade6 leu1 pap1::kanr bfr1::hygr pmd1::natr caf5::kanr mfs1::natr erg5**
SAK84	*dnf2**
	h-; *ade6 leu1 pap1::kanr bfr1::hygr pmd1::natr caf5::kanr mfs1::natr erg5**
SAK690	*dnf2**

* indicates frameshift mutation

Appendix

Appendix 1: Tryptophan dimer gene cluster annotation

AB1650 gene cluster (A/N: JF439215)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	M2	230	methyltransferase type 11	49/60	Saccharopolyspora erythraea NRRL 2338
2	X1	533	PheA/TfdB family FAD-binding monooxygenase	36/50	Pseudomonas fluorescens Pf-5
3	M1	231	methyltransferase type 12	35/49	Actinosynnema mirum DSM 43827
4	D	1015	chromopyrrolic acid synthase RebD	52/66	Lechevalieria aerocolonigenes
5	X2	407	monooxygenase FAD-binding protein	53/66	Stackebrandtia nassauensis DSM 44728
6	M3	336	O-methyltransferase, family protein 2	49/65	Mycobacterium smegmatis str. MC2 155
7	H	515	tryptophan 6-halogenase KtzR	64/77	Kutzneria sp. 744
8	T	412	cation/H+ antiporter	45/58	Streptomyces bingchenggensis BCW-1
9	F	160	flavin reductase KtzS	57/66	Kutzneria sp. 744
10	Y	264	alpha/beta hydrolase	47/58	Amycolatopsis mediterranei U32
11	O	513	L-amino acid oxidase StaO	53/70	Streptomyces sp. TP-A0274
12	C	533	putative monooxygenase RebC	57/70	Lechevalieria aerocolonigenes
13	P	392	cytochrome P450	56/69	Salinispora arenicola CNS-205
14	R	932	ATP-dependent transcription regulator LuxR	36/51	Salinispora arenicola CNS-205

AB1091 gene cluster (A/N: JX827455)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	R	958	putative regulatory protein	42/56	Actinomadura melliaura
2	P	421	putative cytochrome P450 enzyme	45/59	Actinomadura melliaura
3	D	1096	chromopyrrolic acid synthase	60/71	Streptomyces uncialis
4	O	480	tryptophan 2-monooxygenase	55/66	Streptomyces venezuelae ATCC 10712
5	T1	174	ABC transporter, ATP-binding protein	32/48	Streptomyces griseoflavus Tu4000
6	T2	258	ABC transporter ATP-binding protein	36/51	Streptomyces scabiei 87.22
7	F	196	flavin reductase domain-containing protein	59/71	Salinispora tropica CNB-440
8	H	529	putative tryptophan halogenase	71/82	Streptomyces sp. Tu6071
9	M1	248	methyltransferase	60/72	Streptomyces uncialis
10	C	538	monooxygenase	60/72	Streptomyces sp. TP-A0274
11	X2	417	flavin monooxygenase	61/72	Streptomyces uncialis
12	Y1	153	hypothetical protein	33/50	Microlunatus phosphovorus NM-1
13	Y2	104	monooxygenase (low homology)	35/52	Streptomyces hygroscopicus subsp. jinggangensis 5008
14	T3	721	drug exporter of the RND superfamily-like protein	58/72	Modestobacter marinus
15	X3	411	cytochrome P450	46/63	Frankia sp. EUN1f
16	X1	555	flavin monooxygenase	58/67	Streptomyces uncialis
17	M2	242	methyltransferase	45/59	Streptomyces uncialis
18	M3	520	O-methyltransferase	50/63	Saccharomonospora azurea NA-128

AB339 gene cluster (A/N: KF551865)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	M	352	O-methyltransferase	53/70	Streptomyces tubercidicus
2	X	550	FAD-binding monooxygenase	54/66	Streptomyces venezuelae ATCC 10712
3	O	499	tryptophan 2-monooxygenase	57/72	Streptomyces venezuelae ATCC 10712
4	D	1104	chromopyrrolic acid synthase StaD	61/70	Streptomyces sp. TP-A0274
5	P	468	cytochrome P450	54/66	Salinispora arenicola CNS-205

AB234 gene cluster (A/N: KF551864)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	M	347	O-methyltransferase	64/75	Kutzneria sp. 744
2	X	553	FAD-binding monooxygenase	64/74	Kutzneria sp. 744
3	O	515	tryptophan 2-monooxygenase	57/71	Streptomyces venezuelae ATCC 10712
4	D	1105	chromopyrrolic acid synthase StaD	62/71	Streptomyces sp. TP-A0274
5	P	421	cytochrome P450 StaP	56/69	Streptomyces longisporoflavus

AB1149 gene cluster (A/N: KF551867)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	M	339	O-methyltransferase	63/76	Kutzneria sp. 744
2	X	542	FAD-binding monooxygenase	64/74	Kutzneria sp. 744
3	Y	120	hypothetical protein	35/50	Actinoplanes friuliensis DSM 7358
4	O	531	tryptophan 2-monooxygenase	56/71	Streptomyces venezuelae ATCC 10712
5	D	1100	chromopyrrolic acid synthase StaD	60/70	Streptomyces sp. TP-A0274
6	P	423	cytochrome P450 StaP	58/70	Streptomyces sp. TP-A0274

AB1521 gene cluster (A/N: KF551869)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	M	340	O-methyltransferase	60/70	Kutzneria sp. 744
2	X	548	FAD-binding monooxygenase	61/72	Kutzneria sp. 744
3	Y	122	hypothetical protein	33/50	Actinoplanes friuliensis DSM 7358
4	O	535	tryptophan 2-monooxygenase	57/70	Streptomyces venezuelae ATCC 10712
5	D	1205	chromopyrrolic acid synthase StaD	61/69	Streptomyces sp. TP-A0274
6	P	427	cytochrome P450 StaP	55/68	Salinispora arenicola CNS-205

NM1499 gene cluster (A/N: KF551861)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	O	515	tryptophan 2-monooxygenase	59/71	Streptomyces venezuelae ATCC 10712
2	D	1096	chromopyrrolic acid synthase StaD	56/67	Streptomyces sp. TP-A0274
3	P	407	cytochrome P450 StaP	55/65	Streptomyces sp. TP-A0274
4	M	337	O-methyltransferase	55/69	Kutzneria sp. 744
5	X	525	FAD-binding monooxygenase	61/73	Kutzneria sp. 744

AR1973 gene cluster (A/N: KF551873)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	X	346	FAD-binding monooxygenase (truncated)	51/60	Kutzneria sp. 744
2	O	536	tryptophan 2-monooxygenase	56/69	Streptomyces venezuelae ATCC 10712
3	D	1091	chromopyrrolic acid synthase StaD	59/72	Kutzneria sp. 744

NM343 gene cluster (A/N: KF551863)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	M	351	hydroxyneurosporene-O-methyltransferase	52/67	Synechococcus sp. PCC 6312
2	B	1142	chromopyrrolic acid synthase StaD	44/57	Streptomyces sp. TP-A0274
3	C	445	Rieske (2Fe-2S) domain-containing protein	49/64	Pusillimonas noertemannii
4	E	198	prodeoxyviolacein synthase VioE	46/60	Pseudoalteromonas sp. 520P1

AR1455 gene cluster (A/N: KF551872)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	G	410	N-glycosyltransferase	79/83	Lechevalieria aerocolonigenes
2	O	477	L-tryptophan oxidase	86/91	Lechevalieria aerocolonigenes
3	D	1096	chromopyrrolic acid synthase	80/87	Lechevalieria aerocolonigenes
4	C	529	monooxygenase	80/85	Lechevalieria aerocolonigenes
5	P	397	P450 protein	80/85	Lechevalieria aerocolonigenes
6	M	273	methyltransferase	82/91	Lechevalieria aerocolonigenes
7	R	924	regulatory protein	67/76	Lechevalieria aerocolonigenes
8	F	170	flavin reductase	81/88	Lechevalieria aerocolonigenes
9	U	420	integral membrane transporter	74/84	Lechevalieria aerocolonigenes
10	H	530	tryptophan halogenase	85/91	Lechevalieria aerocolonigenes
11	T	468	integral membrane transporter	82/86	Lechevalieria aerocolonigenes

AB1350 gene cluster (A/N: KF551868)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	N2	85	cytochrome P450	61/74	Nonomuraea longicatena
2	C	545	monooxygenase	76/85	Streptomyces purpureus
3	MB	277	methyltransferase	75/86	Streptomyces purpureus
4	E	207	epimerase	74/85	Streptomyces sp. TP-A0274
5	I	369	aminotransferase	82/89	Amycolatopsis orientalis
6	K	73	ketoreductase	61/78	Streptomyces sp. TP-A0274
7	K	273	ketoreductase	70/78	Streptomyces sp. TP-A0274
8	J	473	dehydratase	75/85	Amycolatopsis decaplanina
9	MA	272	methyltransferase	64/74	Streptomyces sp. TP-A0274
10	P	410	cytochrome P450	62/72	Streptomyces sanyensis
11	D	1025	chromopyrrolic acid synthase	65/73	Streptomyces sp. TP-A0274
12	MA	500	tryptophan oxidase	71/84	Streptomyces sp. TP-A0274
13	G	433	glycosyltransferase	75/85	Salinispora arenicola
14	N	407	cytochrome P450	83/89	Streptomyces purpureus
15	R1	677	transcriptional regulator	41/55	Streptomyces uncialis
16	R2	357	transcriptional regulator	48/58	Streptomyces uncialis

AB857 gene cluster (A/N: KF551866)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	M1	281	methyltransferase RebM	50/63	Lechevalieria aerocolonigenes
2	R	908	transcriptional regulator	38/52	Actinomadura melliaura
3	F	178	flavin reductase	54/70	Lechevalieria aerocolonigenes
4	H1	530	tryptophan halogenase RebH	75/86	Lechevalieria aerocolonigenes
5	G	434	glycosyltransferase	60/72	Lechevalieria aerocolonigenes
6	O	489	tryptophan oxidase	63/76	Lechevalieria aerocolonigenes
7	D	1093	chromopyrrolic acid synthase	55/68	Kutzneria albida DSM 43870
8	C	538	FAD-monooxygenase	57/67	Actinomadura melliaura
9	P	425	cytochrome P450	49/63	Salinispora arenicola
10	H2	515	tryptophan halogenase	68/79	Streptomyces sp. FxanaC1
11	Z1	191	hypothetical protein	36/55	Calothrix sp. PCC 7103
12	Z2	195	putative SWIM Zn-finger	39/50	Rhodococcus qingshengii
13	E	637	SNF2 related domain helicase	56/67	Salinispora arenicola
14	M2	355	methyltransferase	44/57	Streptomyces lavendulae
15	T1	508	MFS transporter	38/58	Kutzneria sp. 744
16	R2	146	MarR trancriptional regulator	59/72	Actinoalloteichus spitiensis
17	T2	456	ion antiporter	57/71	Amycolatopsis alba
18	I	275	indole-3-glycerol-phosphate synthase	55/63	Streptomyces albulus
19	L	447	phospho-2-dehydro-3-deoxyheptonate aldolase	72/79	Saccharopolyspora spinosa

AB1533 gene cluster (A/N: KF551870)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	R	875	transcriptional regulator	43/58	Actinomadura melliaura
2	F	190	flavin reductase	63/72	Lechevalieria aerocolonigenes
3	U	426	PyrJ1 membrane transporter	52/67	Streptomyces rugosporus
4	H1	530	tryptophan halogenase RebH	75/85	Lechevalieria aerocolonigenes
5	G	412	glycosyltransferase	63/73	Lechevalieria aerocolonigenes
6	O	490	tryptophan oxidase	69/82	Lechevalieria aerocolonigenes
7	D	1093	chromopyrrolic acid synthase	61/73	Kutzneria albida DSM 43870
8	C	40	FAD-monooxygenase	70/85	Kutzneria albida DSM 43870
9	P	410	cytochrome P450	58/70	Kutzneria albida DSM 43870
10	H2	512	tryptophan halogenase	74/83	Streptomyces sp. FxanaC1
11	Z	214	hypothetical protein	37/54	Cyanobacterium PCC 7702
12	M1	273	methyltransferase RebM	62/76	Lechevalieria aerocolonigenes
13	X1	253	cytochrome P450 hydroxylase	53/63	Saccharopolyspora spinosa
14	X2	130	cytochrome P450 hydroxylase	52/66	Arthrobacter globiformis
15	M2	123	glyoxalase resistance	76/84	Arthrobacter crystallopoietes
16	M2	343	methyltransferase	45/59	Streptomyces davawensis JCM 4913
17	Y	285	aldo/keto reducatse	72/82	Mycobacterium parascrofulaceum

NM747 gene cluster (A/N: KF551862)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	R	903	transcriptional regulator	45/56	Lechevalieria aerocolonigenes
2	F	198	flavin reductase	63/71	Lechevalieria aerocolonigenes
3	H1	530	tryptophan halogenase RebH	78/87	Lechevalieria aerocolonigenes
4	T	466	transporter	55/72	Saccharothrix espanaensis DSM 44229
5	G	440	glycosyltransferase	66/76	Lechevalieria aerocolonigenes
6	O	485	tryptophan oxidase	67/79	Lechevalieria aerocolonigenes
7	D	1109	chromopyrrolic acid synthase	63/74	Kutzneria albida DSM 43870
8	C	533	FAD-monooxygenase	66/76	Lechevalieria aerocolonigenes
9	P	416	cytochrome P450	59/72	Actinomadura melliaura
10	H2	512	tryptophan halogenase	75/84	Streptomyces sp. FxanaC1
11	Z	218	hypothetical protein	32/48	Calothrix sp. PCC 7103
12	I	380	aminotransferase	66/77	Amycolatopsis rifamycinica
13	J	439	CalS14 dehydratase	56/66	Micromonospora echinospora
14	M	329	methyltransferase	53/64	Streptomyces davawensis JCM 4913
15	Y	356	glucose-1-phosphate thymidylyltransferase	67/79	Streptomyces coelicoflavus

AR2194 gene cluster (A/N: KF551874)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	X	415	cytochrome P450	59/71	Streptomyces sulphureus
2	N	407	cytochrome P450 StaN	51/70	Salinispora pacifica
3	G	381	glycosyltransferase	64/75	Salinispora pacifica
4	M1	129	methyltransferase	63/71	Salinispora pacifica
5	M2	127	methyltransferase	77/86	Streptomyces purpureus
6	D	1092	chromopyrrolic acid synthase	63/74	Kutzneria albida DSM 43870
7	C1	546	flavin monooxygenase	57/66	Streptomyces uncialis
8	O	502	tryptophan oxidase	59/72	Streptomyces sp. TP-A0274
9	C2	538	FAD-binding monooxygenase	62/72	Streptomyces purpureus
10	P	438	cytochrome P450	54/67	Kutzneria albida DSM 43870
11	I	158	indole-3-glycerol-phosphate synthase	69/81	Promicromonospora sukumoe
12	R	805	transcriptional regulator	38/52	Nonomuraea longicatena

AR654 gene cluster (A/N: KF551871)

N	Gene	Size (aa)	Homolog	ID/SM (%)	Origin
1	R	905	transcriptional regulator	42/57	Nonomuraea longicatena
2	P	401	cytochrome P450	68/77	Nonomuraea longicatena
3	C	529	FAD-monooxygenase InkE	72/79	Nonomuraea longicatena
4	D	995	chromopyrrolic acid synthase	64/72	Nonomuraea longicatena
5	O	496	tryptophan oxidase	69/78	Nonomuraea longicatena
6	G	438	glycosyltransferase	67/75	Nonomuraea longicatena
7	N	404	cytochrome P450	79/84	Nonomuraea longicatena
8	M1	378	methyltransferase	69/81	Nonomuraea longicatena
9	X1	394	cytochrome P450 hydroxylase	81/88	Nonomuraea longicatena
10	X2	399	cytochrome P450 hydroxylase	49/61	Nonomuraea longicatena
11	M2	297	methyltransferase	65/73	Nonomuraea longicatena
12	X3	413	cytochrome P450	40/55	Bradyrhizobium japonicum

Appendix 2: Compound spectral data summary

BE-54017 A (1): $[\alpha]^{18}_{D}$ -364 (*c* 0.1, methanol); UV (methanol) λ_{max} 209, 243, 320 (sh), 350, 373 nm; ^{1}H NMR (600 MHz, acetone-d_6) δ2.97 (3H, s, H_3-14), 4.09 (3H, s, H_3-15), 4.18 (3H, s, H_3-16), 6.40 (1H, bs, OH-4c), 6.67 (1H, bs, OH-7a), 7.16 (1H, t, *J* = 7.2 Hz, H-10), 7.26 (1H, m, H-2), 7.27 (1H, m, H-9), 7.52 (1H, d, *J* = 8.7 Hz, H-1), 7.73 (1H, d, *J* = 8.0 Hz, H-11), 8.09 (1H, d, *J* = 2.0 Hz, H-4), 8.23 (1H, d, *J* = 8.5 Hz, H-8); ^{13}C NMR (150 MHz, acetone-d_6) δ25.3 (C-14), 34.1 (C-16), 62.8 (C-15), 75.6 (C-4c), 87.7 (C-7a), 104.6 (C-4b), 112.4 (C-1), 114.7 (C-12a), 116.7 (C-8), 118.9 (C-11), 121.0 (C-4), 121.3 (C-10), 123.4 (C-11a), 123.7 (C-2), 125.0 (C-9), 127.0 (C-4a), 127.2 (C-3), 132.9 (C-12b), 137.8 (C-7c), 138.1 (C-12), 139.3 (C-13a), 171.7 (C-7), 174.7 (C-5); HR-ESI-MS *m/z* 452.1019 $[M+H]^+$ (calcd for $C_{23}H_{19}N_3O_5Cl$, 452.1013).

Cladoniamide A (2): $[\alpha]^{18}_{D}$ -331 (*c* 0.1, methanol); UV (methanol) λ_{max} 212, 238, 328 (sh), 349, 370 nm; ^{1}H NMR (600 MHz, acetone-d_6) δ2.94 (3H, s, H_3-14), 4.11 (3H, s, H_3-15), 6.40 (1H, bs, OH-4c), 6.73 (1H, bs, OH-7a), 7.10 (1H, t, *J* = 7.6 Hz, H-10), 7.13 (1H, dd, *J* = 8.6, 2.0 Hz, H-2), 7.21 (1H, t, *J* = 7.5 Hz, H-9), 7.45 (1H, d, *J* = 8.6 Hz, H-1), 7.69 (1H, d, *J* = 8.0 Hz, H-11), 8.02 (1H, d, *J* = 1.6 Hz, H-4), 8.19 (1H, d, *J* = 8.5 Hz, H-8), 10.93 (1H, bs, NH-13); ^{13}C NMR (150 MHz, acetone-d_6) δ24.4 (C-14), 62.0 (C-15), 75.8 (C-4c), 87.9 (C-7a), 103.7 (C-4b), 113.9 (C-1), 114.8 (C-12a), 116.5 (C-8), 118.9 (C-11), 120.8 (C-4), 121.3 (C-10), 122.8 (C-11a), 123.5 (C-2), 124.5 (C-9), 126.7 (C-3), 127.5 (C-4a), 130.2 (C-12b), 137.6 (C-7c), 138.1 (C-12), 138.4 (C-13a), 171.7 (C-7), 174.5 (C-5); HR-ESI-MS *m/z* 438.0844 $[M+H]^+$ (calcd for $C_{22}H_{17}N_3O_5Cl$, 438.0778).

BE-54017 B (3): UV (methanol) λ_{max} 224, 278, 297, 360 nm; ^{1}H NMR (600 MHz, acetone-d_6) δ2.76 (3H, d, *J* = 4.5 Hz, H_3-14), 4.21 (3H, s, H_3-16), 4.22 (3H, s, H_3-15),

6.00 (1H, bs, OH-4c), 7.24 (1H, dd, *J* = 8.8, 2.0 Hz, H-2), 7.38 (1H, t, *J* = 7.3 Hz, H-10), 7.45 (1H, t, *J* = 7.3 Hz, H-9), 7.54 (1H, d, *J* = 8.8 Hz, H-1), 7.77 (1H, d, *J* = 1.8 Hz, H-4), 7.81 (1H, d, *J* = 7.8 Hz, H-11), 8.04 (1H, bs, NH-6), 8.50 (1H, d, *J* = 8.2 Hz, H-8); ^{13}C NMR (150 MHz, acetone-d_6) δ27.2 (C-14), 34.6 (C-16), 63.2 (C-15), 78.0 (C-4c), 103.7 (C-4b), 113.4 (C-1), 117.8 (C-8), 118.2 (C-12a), 120.6 (C-11), 120.8 (C-4), 124.8 (C-2), 126.2 (C-10), 126.8 (C-11a), 127.3 (C-4a), 127.4 (C-3), 128.7 (C-9), 130.7 (C-12b), 136.3 (C-7c), 140.1 (C-13a), 141.4 (C-12), 170.2 (C-7a), 171.4 (C-5); HR-ESI-MS *m/z* 424.1048 $[M+H]^+$ (calcd for $C_{22}H_{19}N_3O_4Cl$, 424.1064).

BE-54017 C (4): UV (methanol) λ_{max} 224, 259, 286, 309, nm; ^{1}H NMR (600 MHz, acetone-d_6) δ2.76 (3H, d, *J* = 4.5 Hz, H_3-14), 4.22 (3H, s, H_3-15), 4.36 (3H, s, H_3-16), 6.93 (1H, bs, OH-7a), 7.15 (1H, t, *J* = 7.5 Hz, H-10), 7.28 (1H, t, *J* = 7.7 Hz, H-9), 7.37 (1H, dd, *J* = 8.7, 2.1 Hz, H-2), 7.66 (1H, d, *J* = 8.7 Hz, H-1), 7.70 (1H, d, *J* = 8.4 Hz, H-8), 7.82 (1H, d, *J* = 8.0 Hz, H-11), 7.96 (1H, bs, NH-6), 8.11 (1H, d, *J* = 2.0 Hz, H-4); ^{13}C NMR (150 MHz, acetone-d_6) δ26.5 (C-14), 34.6 (C-16), 63.1 (C-15), 87.1 (C-7a), 107.2 (C-4b), 113.2 (C-1), 114.3 (C-8), 115.2 (C-12a), 120.5 (C-11), 120.7 (C-4), 121.6 (C-10), 122.8 (C-11a), 125.1 (C-2), 126.3 (C-9), 126.8 (C-4a), 129.6 (C-3), 138.9 (C-7c), 139.5 (C-13a), 140.6 (C-12), 143.3 (C-12b), 168.5 (C-7) 184.7 (C-4c); HR-ESI-MS *m/z* 424.1068 $[M+H]^+$ (calcd for $C_{22}H_{19}N_3O_4Cl$, 424.1064).

BE-54017 D (5): UV (methanol) λ_{max} 210, 245, 323 (sh), 352, 376 nm; ^{1}H NMR (600 MHz, acetone-d_6) δ2.95 (3H, s, H_3-14), 4.07 (3H, s, H_3-15), 4.18 (3H, s, H_3-16), 6.30 (1H, bs, OH-4c), 6.60 (1H, bs, OH-7a), 7.15 (1H, t, *J* = 7.1 Hz, H-10), 7.16 (1H, t, *J* = 7.1 Hz, H-3), 7.25 (1H, t, *J* = 7.7 Hz, H-9), 7.28 (1H, t, *J* = 7.7 Hz, H-2), 7.50 (1H, d, *J* = 8.3 Hz, H-1), 7.71 (1H, d, *J* = 7.9 Hz, H-11), 8.07 (1H, d, *J* = 8.0 Hz, H-4), 8.22 (1H, d, *J* =

8.5 Hz, H-8); ^{13}C NMR (150 MHz, acetone-d_6) δ25.2 (C-14), 33.8 (C-16), 62.7 (C-15), 75.9 (C-4c), 87.8 (C-7a), 105.1 (C-4b), 110.9 (C-1), 115.2 (C-12a), 116.6 (C-8), 118.8 (C-11), 121.7 (C-3), 121.2 (C-10), 121.8 (C-4), 123.6 (C-11a), 123.9 (C-2), 124.7 (C-9), 126.2 (C-4a), 131.5 (C-12b), 137.5 (C-12), 137.7 (C-7c), 140.8 (C-13a), 171.8 (C-7), 174.6 (C-5); HR-ESI-MS *m/z* 418.1383 $[M+H]^+$ (calcd for $C_{23}H_{20}N_3O_5$, 418.1403).

BE-54017 E (6): UV (methanol) λ_{max} 226, 261, 288, 312 nm; ^{1}H NMR (600 MHz, acetone-d_6) δ2.75 (3H, d, J = 4.5 Hz, H_3-14), 4.21 (3H, s, H_3-15), 4.35 (3H, s, H_3-16), 7.14 (1H, t, J = 7.4 Hz, H-10), 7.26 (1H, t, J = 7.6 Hz, H-9), 7.31 (1H, t, J = 7.4 Hz, H-3), 7.38 (1H, t, J = 7.5 Hz, H-2), 7.62 (1H, d, J = 8.2 Hz, H-1), 7.68 (1H, d, J = 8.4 Hz, H-8), 7.80 (1H, d, J = 8.0 Hz, H-11), 7.91 (1H, bs, NH-6), 8.15 (1H, d, J = 7.7 Hz, H-4); ^{13}C NMR (150 MHz, acetone-d_6) δ26.5 (C-14), 34.3 (C-16), 63.1 (C-15), 87.1 (C-7a), 107.9 (C-4b), 111.6 (C-1), 114.3 (C-8), 115.7 (C-12a), 120.3 (C-11), 121.6 (C-10), 121.7 (C-4), 123.0 (C-11a), 124.2 (C-3), 125.2 (C-2), 125.8 (C-4a), 126.0 (C-9), 138.8 (C-7c), 140.1 (C-12), 141.0 (C-13a), 142.3 (C-12b), 168.8 (C-7) 184.7 (C-4c); HR-ESI-MS *m/z* 390.1457 $[M+H]^+$ (calcd for $C_{22}H_{20}N_3O_4$, 390.1454).

BE-54017 F (7): ^{1}H NMR (600 MHz, DMSO-d_6) δ7.37 (1H, t, J = 7.6 Hz, H-9), 7.57 (1H, t, J = 7.6 Hz, H-10), 7.58 (1H, d, J = 8.2 Hz, H-2), 7.83 (1H, d, J = 8.2 Hz, H-11), 7.87 (1H, d, J = 8.6 Hz, H-1), 8.98 (1H, bs, NH-6), 8.99 (1H, m, H-8), 9.00 (1H, m, H-4) 11.07 (1H, bs, NH-12) 11.08 (1H, bs, NH-13); ^{13}C NMR (150 MHz, DMSO-d_6) δ112.1 (C-11), 113.8 (C-1), 114.4 (C-7b), 115.8 (C-4b), 119.8 (C-4c), 120.3 (C-7a), 120.3 (C-9), 121.5 (C-7c), 122.7 (C-4a), 123.2 (C-4), 124.2 (C-3), 124.3 (C-8), 126.4 (C-2), 126.9 (C-10), 129.2 (C-12a), 129.9 (C-12b), 138.9 (C-13a), 140.4 (C-11a), 171.3 (C-7), 171.4 (C-

5); HR-ESI-MS *m/z* 390.1457 $[M+H]^+$ (calcd for $C_{22}H_{20}N_3O_4$, 390.1454). HR-ESI-MS *m/z* 358.0385 [M-H]-(calcd for $C_{20}H_9N_3O_2Cl$, 358.0462).

BE-54017 G (8): ^{1}H NMR (600 MHz, acetone-d_6) δ2.96 (3H, s, H_3-14), 4.19 (3H, s, H_3-16), 5.44 (1H, bs, OH-4c), 5.46 (1H, bs, OH-7a), 7.13 (1H, t, *J* = 7.5 Hz, H-3), 7.13 (1H, t, *J* = 7.5 Hz, H-9), 7.17 (1H, t, *J* = 7.5 Hz, H-10), 7.22 (1H, t, *J* = 7.6 Hz, H-2), 7.47 (1H, d, *J* = 8.0 Hz, H-1), 7.47 (1H, d, *J* = 8.0 Hz, H-11), 8.19 (1H, d, *J* = 8.0 Hz, H-4), 8.19 (1H, d, *J* = 8.0 Hz, H-8), 10.77 (1H, bs, NH-12); ^{13}C NMR (150 MHz, acetone-d_6) δ25.3 (C-14), 31.9 (C-16), 77.6 (C-4c), 77.6 (C-7a), 109.3 (C-4b), 110.2 (C-1), 110.8 (C-7b), 112.5 (C-11), 121.2 (C-3), 121.3 (C-9), 122.5 (C-8), 122.6 (C-4), 123.1 (C-2), 123.4 (C-10), 127.4 (C-12a), 127.6 (C-4a), 127.8 (C-7c), 129.4 (C-12b), 138.9 (C-11a), 139.5 (C-13a), 176.0 (C-7), 176.2 (C-5); HR-ESI-MS *m/z* 410.1119 $[M+H]^+$ (calcd for $C_{22}H_{17}N_3O_4Na$, 410.1117).

Borregomycin A (9): $[\alpha]^{18}_D$ +443 (*c* 0.097, DMSO); UV (MeOH) λ_{max} 275, 302, 365; ^{1}H NMR (600 MHz, DMSO-d_6) δ2.87 (3H, s, H_3-14), 3.62 (3H, s, H_3-15), 3.64 (3H, s, H_3-16), 7.09 (1H, dd, *J* = 8.3, 1.6 Hz, H-10), 7.12 (1H, dd, *J* = 8.4, 1.8 Hz, H-3), 7.44 (1H, d, *J* = 1.6 Hz, H-1), 7.45 (1H, d, *J* = 1.6 Hz, H-8), 7.65 (1H, d, *J* = 8.2 Hz, H-11), 7.71 (1H, d, J = 8.4 Hz, H-4), 7.72 (1H, brs, OH-12a), 11.84 (1H, brs, H-13); ^{13}C NMR (150 MHz, DMSO-d_6) δ24.7 (C-14), 54.0 (C-15), 54.4 (C-16), 79.7 (C-4c), 85.7 (C-12a), 91.6 (C-7a), 105.7 (C-4b), 112.1 (C-1), 115.4 (C-8), 118.5 (C-11a), 120.9 (C-3), 121.6 (C-10), 122.0 (C-4), 122.7 (C-4a), 126.4 (C-11), 128.1 (C-2), 133.3 (C-12b), 137.7 (C-13a), 143.4 (C-9), 157.0 (C-7c), 167.0 (C-5), 171.3 (C-7), 194.4 (C-12); HR-ESI-MS *m/z*500.0403 $[M-H]^-$ (calcd for $C_{23}H_{16}N_3O_6Cl_2$, 500.0416).

Borregomycin B (10): UV (MeOH) λ_{max} 230, 256, 272 (sh), 347 (sh), 364, 386; ^{1}H NMR (600 MHz, acetone-d_6) δ3.02 (3H, s, H_3-14), 3.41 (6H, s, H_3-15/H_3-16), 7.16 (2H, dd, J = 8.6, 1.7 Hz, H-3/H-9), 7.52 (2H, d, J = 1.6 Hz, H-1/H-11), 8.09 (2H, d, J = 8.6 Hz, H-4/H-8), 10.97 (2H, brs, H-12/H-13); ^{13}C NMR (150 MHz, acetone-d_6) δ25.4 (C-14), 54.7 (C-15/C-16), 82.0 (C-4c/C-7a), 108.4 (C-4b/C-7b), 112.4 (C-1/C-11), 122.1 (C-3/C-9), 124.0 (C-4/C-8), 126.4 (C-4a/C-7c), 128.9 (C-2/C-10), 129.3 (C-12a/C-12b), 138.8 (C-11a/C-13a), 174.1 (C-5/C-7); HR-ESI-MS m/z 468.0518[M-H]$^-$ (calcd for $C_{23}H_{16}N_3O_4Cl_2$, 468.0518).

Borregomycin C (11): $[\alpha]^{18}_{D}$ +149 (c 0.13, DMSO); UV (MeOH) λ_{max} 230, 256, 272 (sh), 347 (sh), 363, 384; ^{1}H NMR (600 MHz, acetone-d_6) δ2.95 (3H, s, H_3-14), 3.13 (3H, s, H_3-15), 5.75 (1H, brs, OH-7a), 7.15 (1H, dd, J = 8.6, 1.7 Hz, H-9), 7.18 (1H, dd, J = 8.6, 1.7 Hz, H-3), 7.51 (1H, d, J = 1.5 Hz, H-11), 7.54 (1H, d, J = 1.5 Hz, H-1), 8.13 (1H, d, J = 8.6 Hz, H-3), 8.17 (1H, d, J = 8.5 Hz, H-8), 10.73 (1H, brs, H-12), 10.91 (1H, brs, H-13); ^{13}C NMR (150 MHz, acetone-d_6) δ25.5 (C-14), 53.5 (C-15), 77.8 (C-7a), 81.5 (C-4c), 106.0 (C-4b), 111.8 (C-7b), 112.3 (C-11), 112.5 (C-1), 121.8 (C-9), 122.3 (C-3), 123.5 (C-4), 124.2 (C-8), 126.5 (C-4a), 126.8 (C-7c), 127.3 (C-12a), 128.8 (C-10), 128.9 (C-2), 131.3 (C-12b), 138.7 (C-13a), 138.9 (C-11a), 174.9 (C-5), 175.3 (C-7).HR-ESI-MS m/z456.0494 [M+H]$^+$ (calcd for $C_{22}H_{16}N_3O_4Cl_2$, 456.0518).

Borregomycin D (12): UV (MeOH) λ_{max} 230, 256, 272 (sh), 347 (sh), 363, 384;^{1}H NMR (600 MHz, acetone-d_6) δ2.98 (3H, s, H_3-14), 5.58 (2H, brs, OH-4c/OH-7a), 7.14 (2H, dd, J = 8.6, 1.7 Hz, H-3/H-9), 7.49 (2H, d, J = 1.6 Hz, H-1/H-11), 8.14 (2H, d, J = 8.6 Hz, H-4/H-8), 10.76 (2H, brs, H-12/H-13); ^{13}C NMR (150 MHz, acetone-d_6) δ25.4 (C-14), 77.2 (C-4c/C-7a), 110.1 (C-4b/C-7b), 112.3 (C-1/C-11), 121.8 (C-3/C-9), 123.7 (C-4/C-8),

126.9 (C-4a/C-7c), 128.6 (C-2/C-10), 128.4 (C-12a/C-12b), 138.7 (C-11a/C-13a), 175.9 (C-5/C-7); HR-ESI-MS *m/z*442.0338 $[M+H]^+$ (calcd for $C_{21}H_{14}N_3O_4Cl_2$, 442.0361).

Dichlorochromopyrrolic acid (13): UV (MeOH) λ_{max} 227, 265, 330 (sh); ^{1}H NMR (600 MHz, acetone-d_6) δ6.79 (2H, dd, *J* = 8.5, 1.8 Hz, H-3/H-9), 7.14 (2H, s, H-12a/H-12b), 7.16 (2H, d, *J* = 8.5 Hz, H-4/H-8), 7.31 (2H, d, *J* = 1.6 Hz, H-1/H-11), 10.17 (2H, brs, H-12/H-13), 10.78 (1H, brs, H-6); ^{13}C NMR (150 MHz, acetone-d_6) δ109.7 (C-4b/C-7b), 111.7 (C-1/C-11), 119.8 (C-3/C-9), 121.9 (C-4/C-8), 124.0 (C-4c/C-7a), 125.3 (C-5/C-7), 127.0 (C-2/C-10), 127.1 (C-12a/C-12b), 127.7 (C-4a/C-7c), 137.3 (C-11a/C-13a), 161.6 (HOOC-5/HOOC-7); HR-ESI-MS *m/z*452.0193 $[M-H]^-$ (calcd for $C_{22}H_{12}N_3O_4Cl_2$, 452.0205).

***O*-methyl borregomycin A (14):** $[\alpha]^{18}_D$ +90 (*c* 0.13, DMSO); UV (MeOH) λ_{max} 275, 302, 365;^{1}H NMR (600 MHz, DMSO-d_6) δ2.85 (3H, s, H_3-14), 3.01 (3H, s, OCH_3-12a), 3.34 (3H, s, H_3-15), 3.72 (3H, s, H_3-16), 7.14 (1H, dd, *J* = 8.5, 1.7 Hz, H-3),7.17 (1H, dd, *J* = 8.3, 1.7 Hz, H-10), 7.43 (1H, d, *J* = 1.7 Hz, H-1), 7.54 (1H, d, *J* = 1.8 Hz, H-8), 7.69 (1H, d, J = 8.4 Hz, H-4), 7.71 (1H, d, *J* = 8.2 Hz, H-11), 12.01 (1H, brs, H-13); ^{13}C NMR (150 MHz, DMSO-d_6) δ24.7 (C-14), 51.7 (C-15), 54.3 (OH_3C-12a), 54.7 (C-16), 79.7 (C-12a), 89.7 (C-4c), 92.0 (C-7a), 106.3 (C-4b), 112.2 (C-1), 116.2 (C-8), 119.3 (C-11a), 121.2 (C-3), 122.1 (C-4), 122.5 (C-10), 122.5 (C-4a), 126.4 (C-11), 128.6 (C-2), 131.3 (C-12b), 137.9 (C-13a), 144.2 (C-9), 158.1 (C-7c), 166.5 (C-5), 171.2 (C-7), 194.4 (C-12); HR-ESI-MS *m/z* 514.0574 $[M-H]^-$ (calcd for $C_{24}H_{18}N_3O_6Cl_2$, 514.0573).

Erdasporine A (15): UV (methanol) λ_{max} 231 (sh), 312, 325, 370, 388; ^{1}H NMR (600 MHz, DMSO-d_6) δ3.94 (3H, s, H_3-15), 6.70 (1H, dd, *J* = 8.5, 1.6 Hz, H-3), 6.98 (1H, d, *J* = 1.2 Hz, H-1), 7.23 (1H, t, *J* = 7.6 Hz, H-9), 7.33 (1H, t, *J* = 7.6 Hz, H-10), 7.69 (1H, d,

J = 8.2 Hz, H-11), 8.17 (1H, d, *J* = 3.1 Hz, H-5), 8.28 (1H, d, *J* = 7.8 Hz, H-8), 8.59 (1H, d, *J* = 8.6 Hz, H-4), 9.22 (1H, brs, OH-2), 11.07 (1H, brs, NH-12), 11.17 (1H, brs, NH-13), 12.64 (1H, brs, NH-6); ^{13}C NMR (150 MHz, DMSO-d_6) δ50.9 (C-15), 96.1 (C-1), 108.7 (C-7b), 108.8 (C-3), 109.7 (C-4b), 111.0 (C-7), 111.6 (C-11), 115.2 (C-5), 117.1 (C-4c), 117.9 (C-7a), 118.4 (C-4a), 119.2 (C-9), 120.3 (C-8), 122.8 (C-10), 123.4 (C-7c), 125.0 (C-12a), 125.5 (C-4), 126.9 (C-12b), 137.8 (C-11a), 139.5 (C-13a), 154.2 (C-2), 161.2 (C-14); HR-ESI-MS *m/z* 368.1024 $[M-H]^-$ (calcd for $C_{22}H_{14}N_3O_3$, 368.1035).

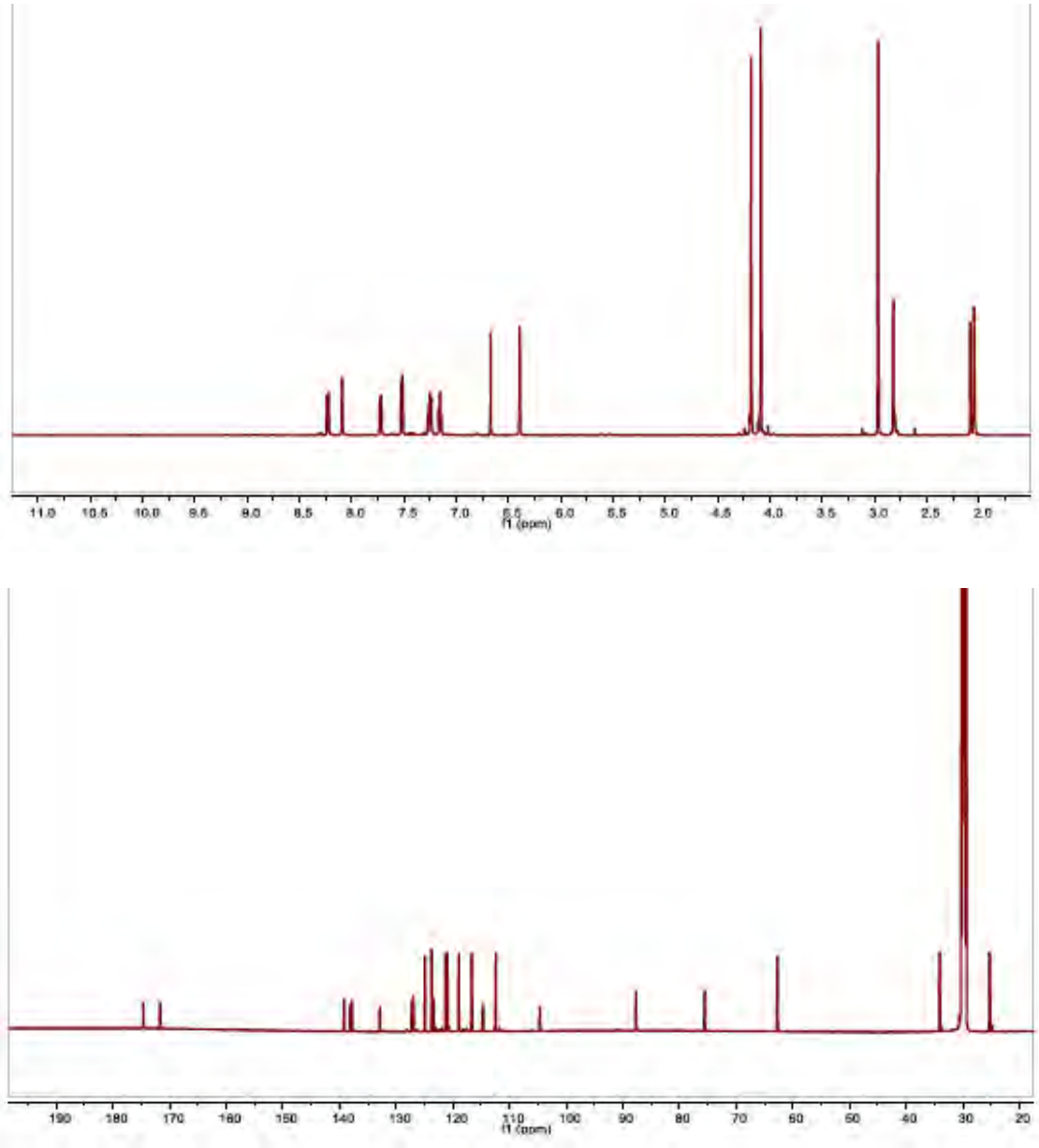
Erdasporine B (16): UV (methanol) λ_{max} 238 (sh), 311, 323, 364, 384; ^{1}H NMR (600 MHz, acetone-d_6) δ3.99 (3H, s, H_3-15), 7.23 (1H, t, *J* = 7.5 Hz, H-3), 7.28 (1H, t, *J* = 7.4 Hz, H-9), 7.31 (1H, t, *J* = 7.2 Hz, H-2), 7.36 (1H, t, *J* = 7.5 Hz, H-10), 7.60 (1H, d, *J* = 8.0 Hz, H-1), 7.66 (1H, d, *J* = 8.0 Hz, H-11), 8.27 (1H, s, H-5), 8.28 (1H, d, *J* = 8.2 Hz, H-8), 9.07 (1H, d, *J* = 8.1 Hz, H-4), 10.63 (1H, bs, NH-12), 10.86 (1H, brs, NH-13), 11.86 (1H, brs, NH-6); ^{13}C NMR (150 MHz, acetone-d_6) δ51.3 (C-15), 111.3 (C-4b), 111.4 (C-1), 111.7 (C-7b), 112.4 (C-11), 113.0 (C-7), 115.4 (C-5), 118.9 (C-4c), 119.8 (C-7a), 120.0 (C-3), 120.5 (C-9), 121.3 (C-8), 123.7 (C-2), 124.3 (C-7c), 124.7 (C-12a), 124.8 (C-10), 126.2 (C-4a), 126.4 (C-4), 130.9 (C-12b), 139.4 (C-11a), 139.5 (C-13a), 162.1 (C-14); HR-ESI-MS *m/z* 354.1237 $[M+H]^+$ (calcd for $C_{22}H_{16}N_3O_2$, 354.1243).

Erdasporine C (17): $[\alpha]^{18}_D$ ±0 (*c* 0.1, methanol); UV (methanol) λ_{max} 220 (sh), 293, 331, 343, 359; ^{1}H NMR (600 MHz, acetone-d_6) δ3.65 (3H, s, H_3-15), 5.94 (1H, s, H-7), 7.28 (1H, t, *J* = 7.3 Hz, H-3), 7.32 (1H, t, *J* = 7.5 Hz, H-9), 7.46 (1H, m, H-2), 7.48 (1H, m, H-10), 7.65 (1H, d, *J* = 8.1 Hz, H-1), 7.70 (1H, d, J = 8.1 Hz, H-11), 7.85 (1H, brs, NH-6), 8.38 (1H, d, *J* = 7.9 Hz, H-8), 9.41 (1H, d, J = 7.9 Hz, H-4), 10.76 (1H, brs, NH-12), 10.96 (1H, brs, NH-13); ^{13}C NMR (150 MHz, acetone-d_6) δ52.8 (C-15), 60.6 (C-7), 111.8

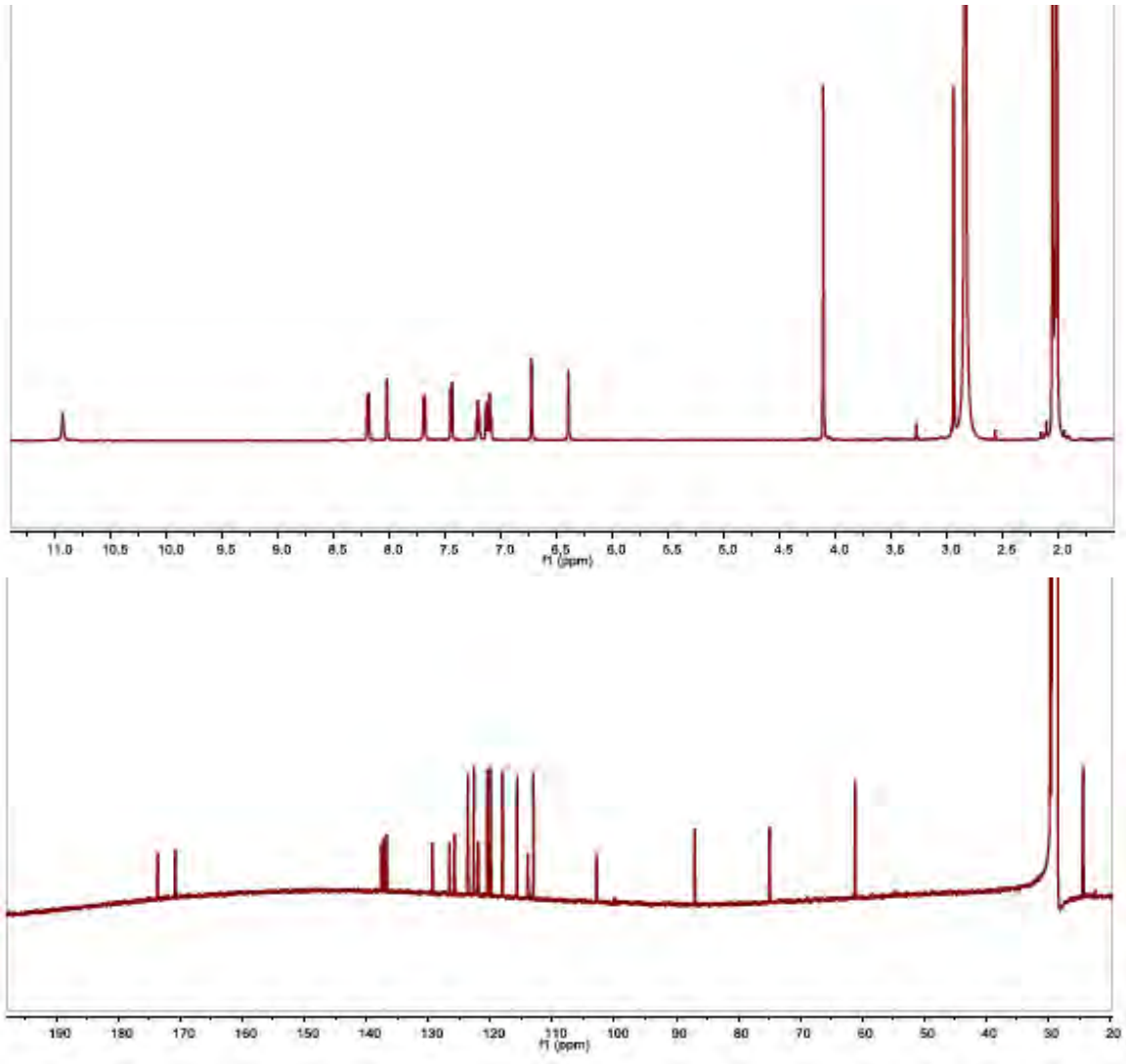
(C-1), 112.4 (C-11), 115.9 (C-7b), 117.5 (C-4b), 120.3 (C-4c), 120.4 (C-3), 120.9 (C-9), 123.3 (C-8), 123.4 (C-7c), 124.3 (C-4a), 126.3 (C-10), 126.5 (C-2), 126.8 (C-4), 127.3 (C-12b), 129.5 (C-12a), 131.8 (C-7a), 140.7 (C-11a), 140.9 (C-13a), 171.4 (C-14), 173.4 (C-5); HR-ESI-MS *m/z* 370.1186 $[M+H]^+$ (calcd for $C_{22}H_{16}N_3O_3$, 370.1192).

Methylarcyriarubin (21): UV (methanol) λ_{max} 227, 276, 377, 475; ^{1}H NMR (600 MHz, DMSO-d_6) δ3.05 (3H, s, H_3-13), 6.63 (2H, t, J = 7.5 Hz, H-10), 6.80 (2H, d, J = 8.0 Hz, H-11), 6.97 (2H, t, J = 7.5 Hz, H-9), 7.37 (2H, d, J = 8.1 Hz, H-8), 7.75 (2H, s, H-5), 11.66 (2H, bs, N-4);^{13}C NMR (150 MHz, DMSO-d_6) δ24.0 (C-13), 105.6 (C-6), 111.7 (C-8), 119.3 (C-10), 120.9 (C-11), 121.6 (C-9), 125.3 (C-12), 127.1 (C-3), 129.1 (C-5), 136.0 (C-7), 171.8 (C-2); HR-ESI-MS *m/z* 342.1237 $[M-H]^+$ (calcd for $C_{21}H_{16}N_3O_2$, 342.1243).

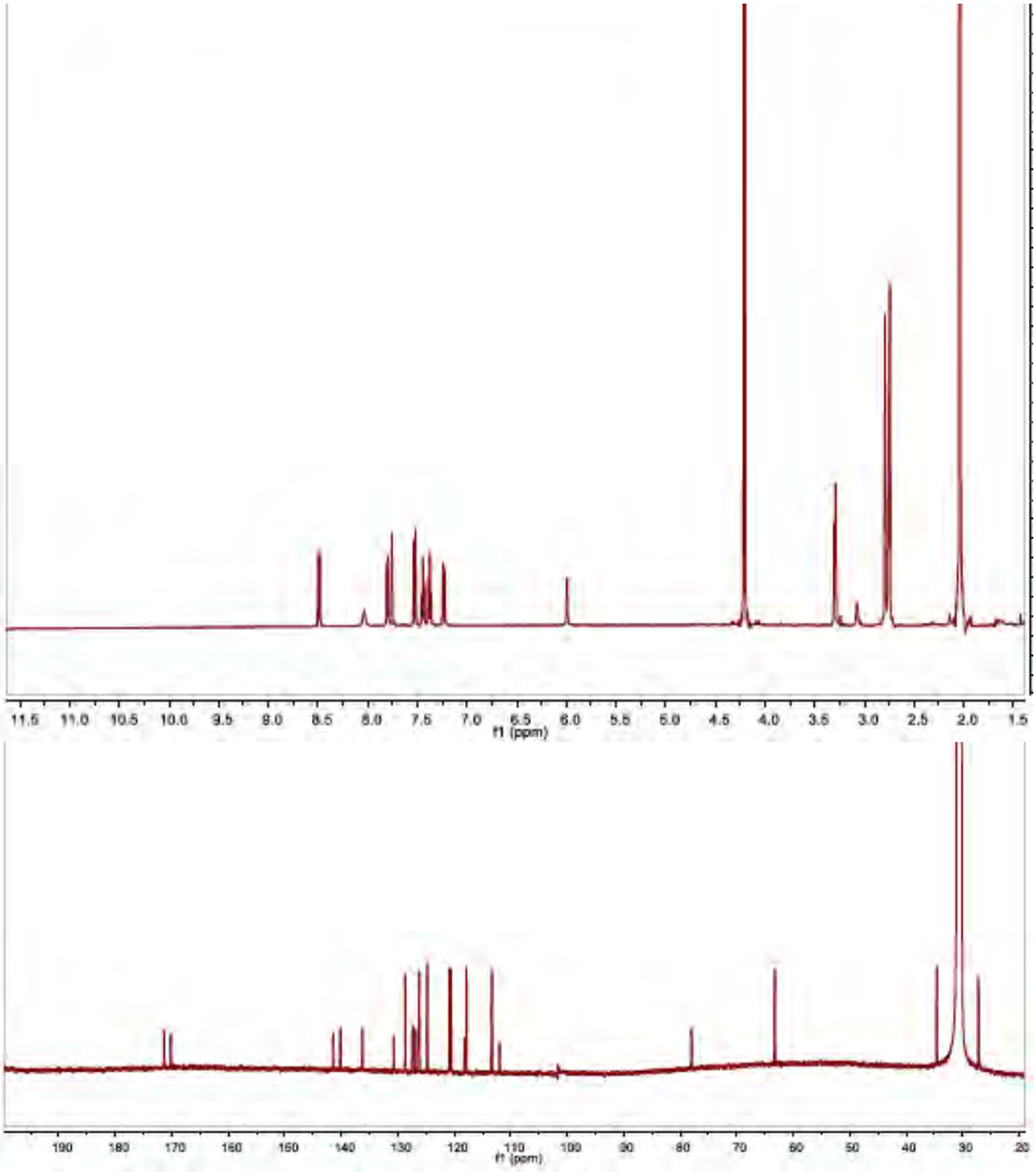
Appendix 3: 1-D NMR spectra



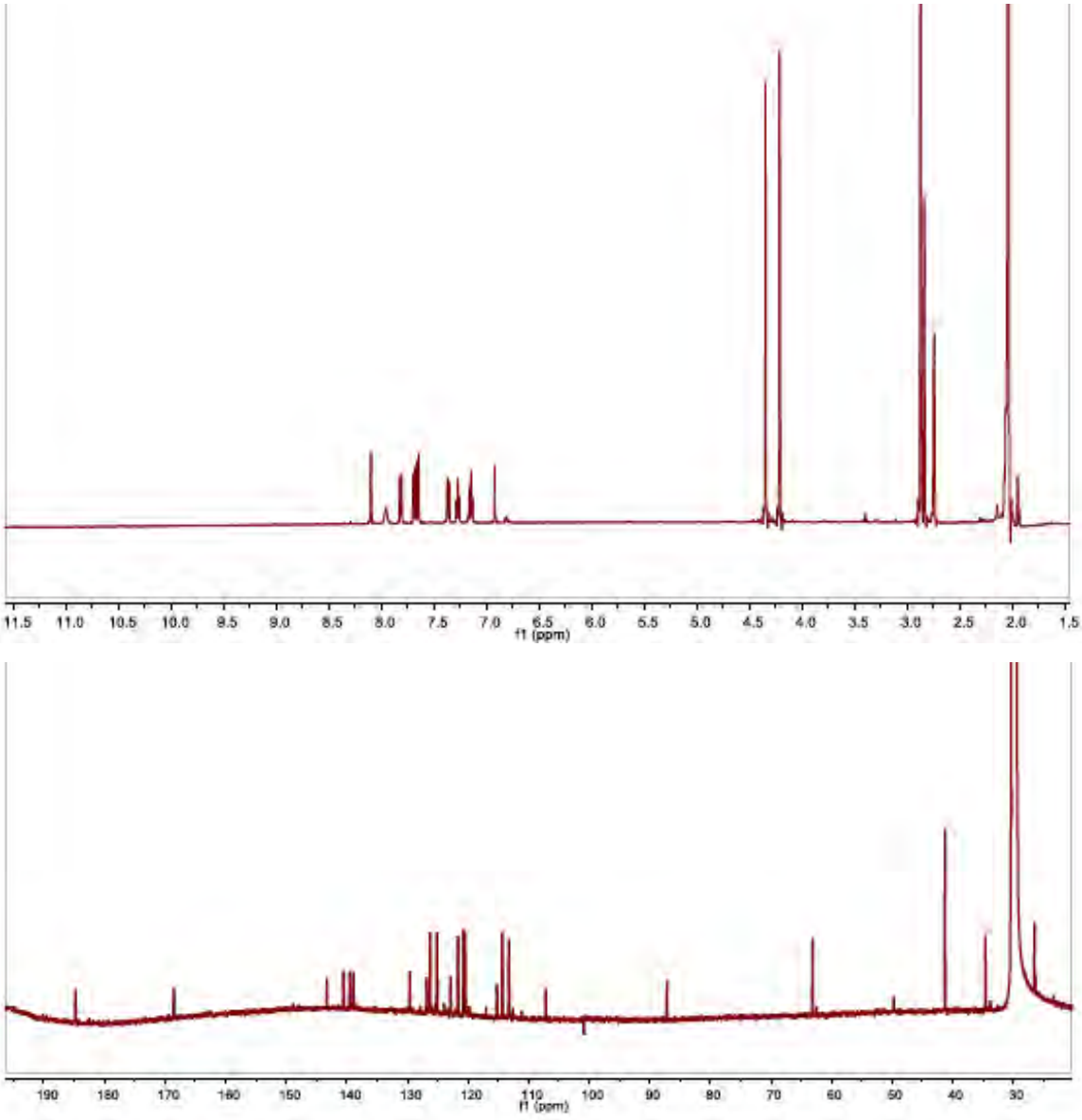
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **1** in acetone-d_6.



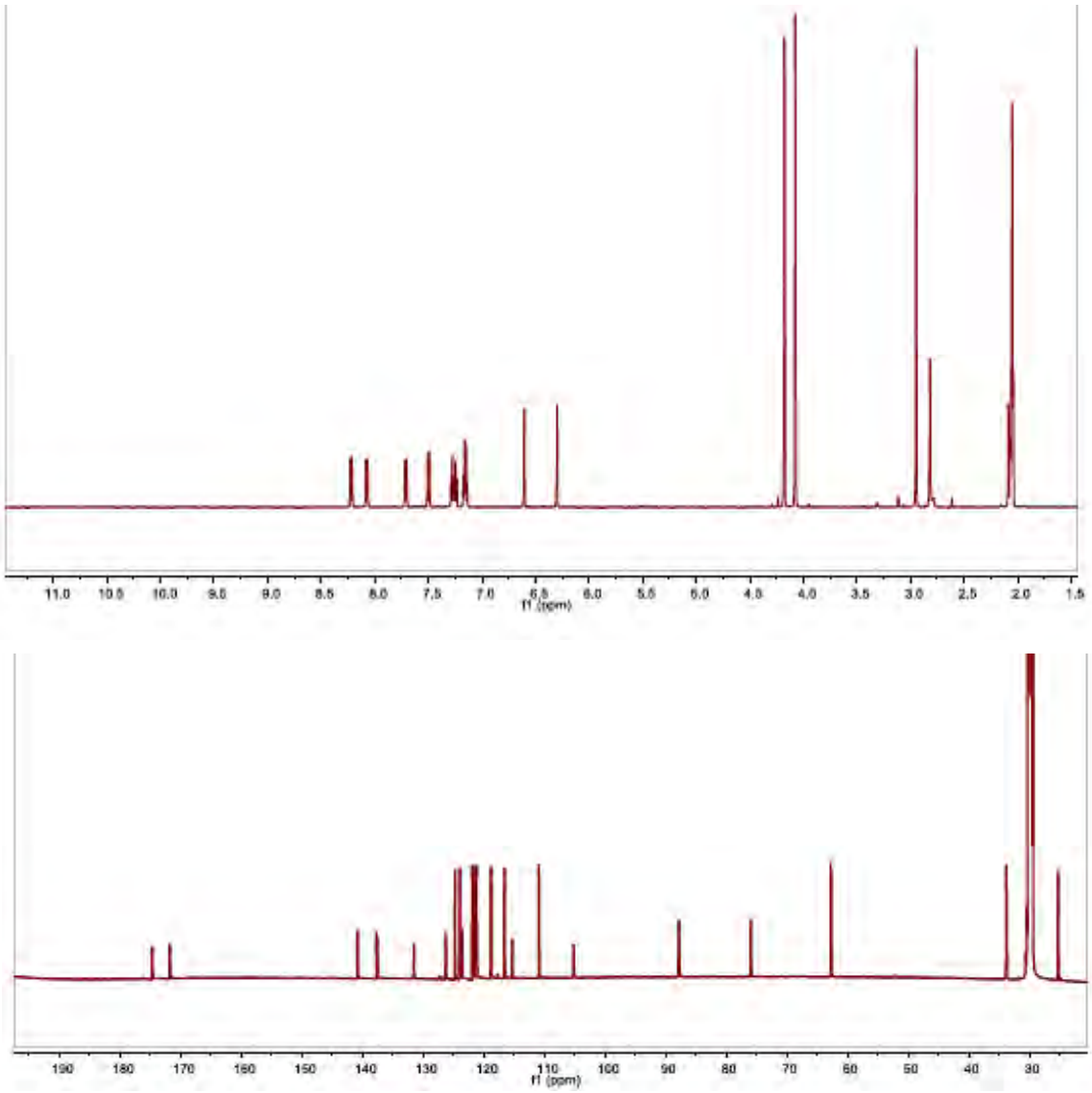
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **2** in acetone-d_6.



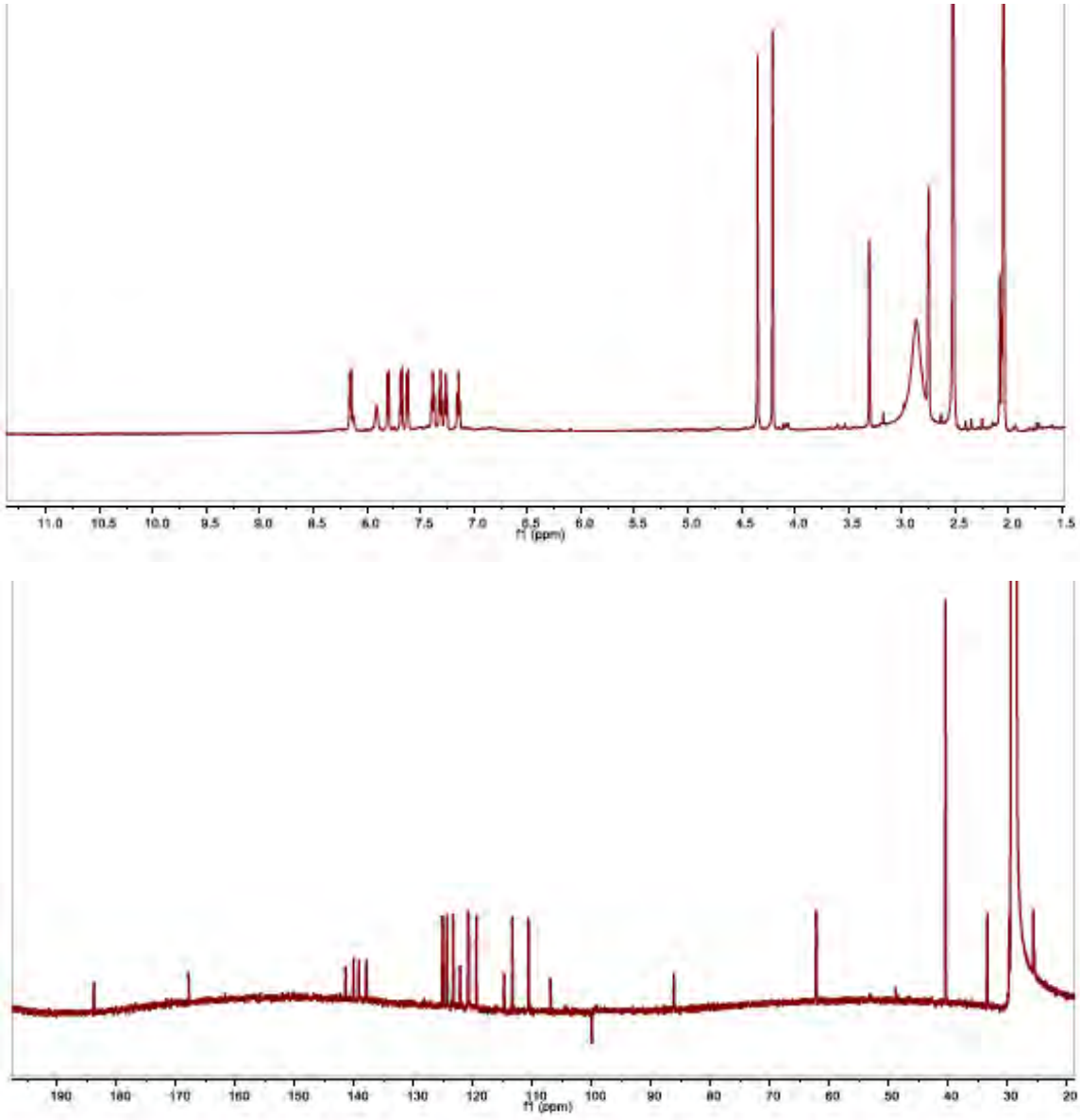
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **3** in acetone-d_6.



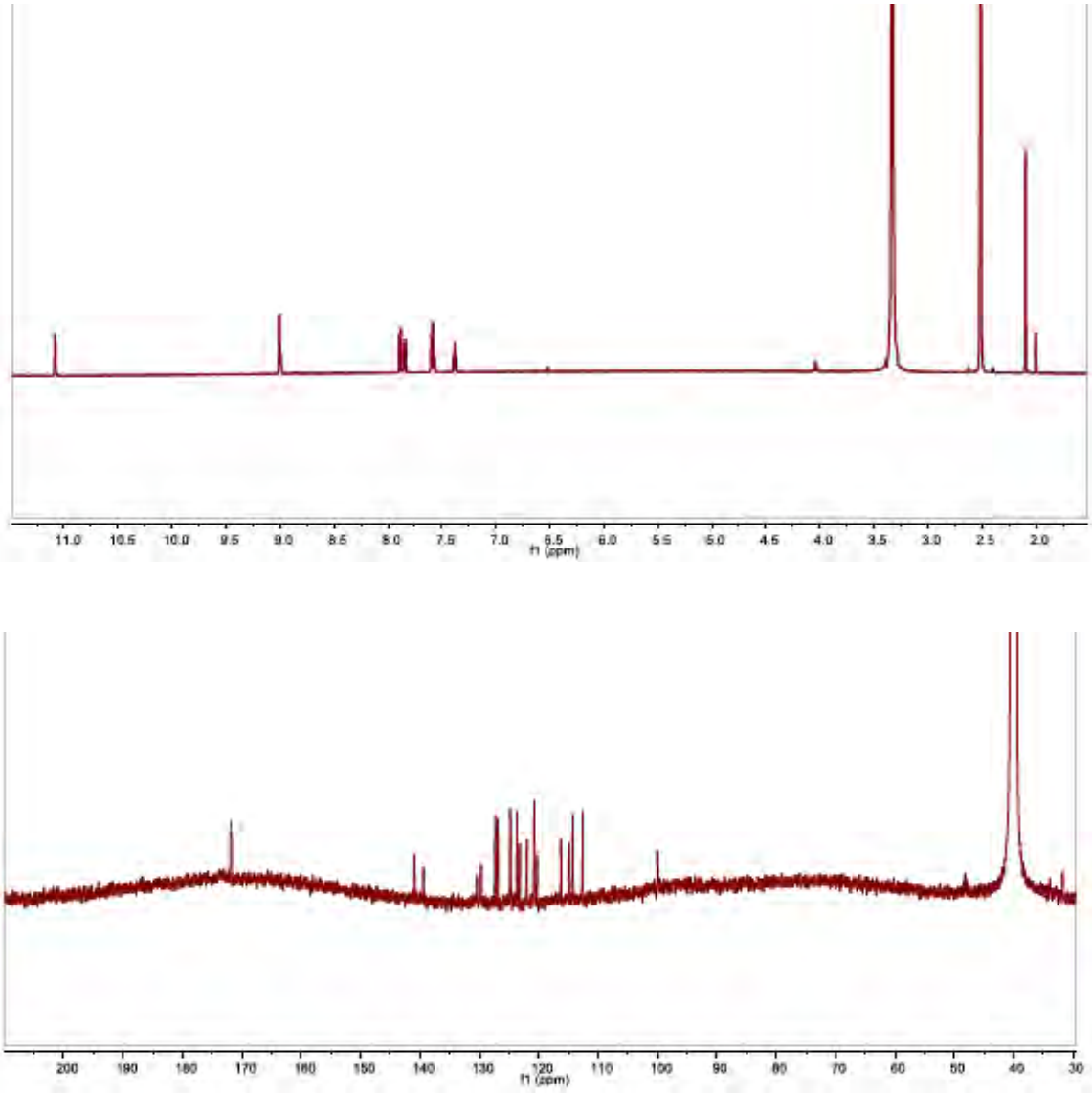
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **4** in acetone-d_6.



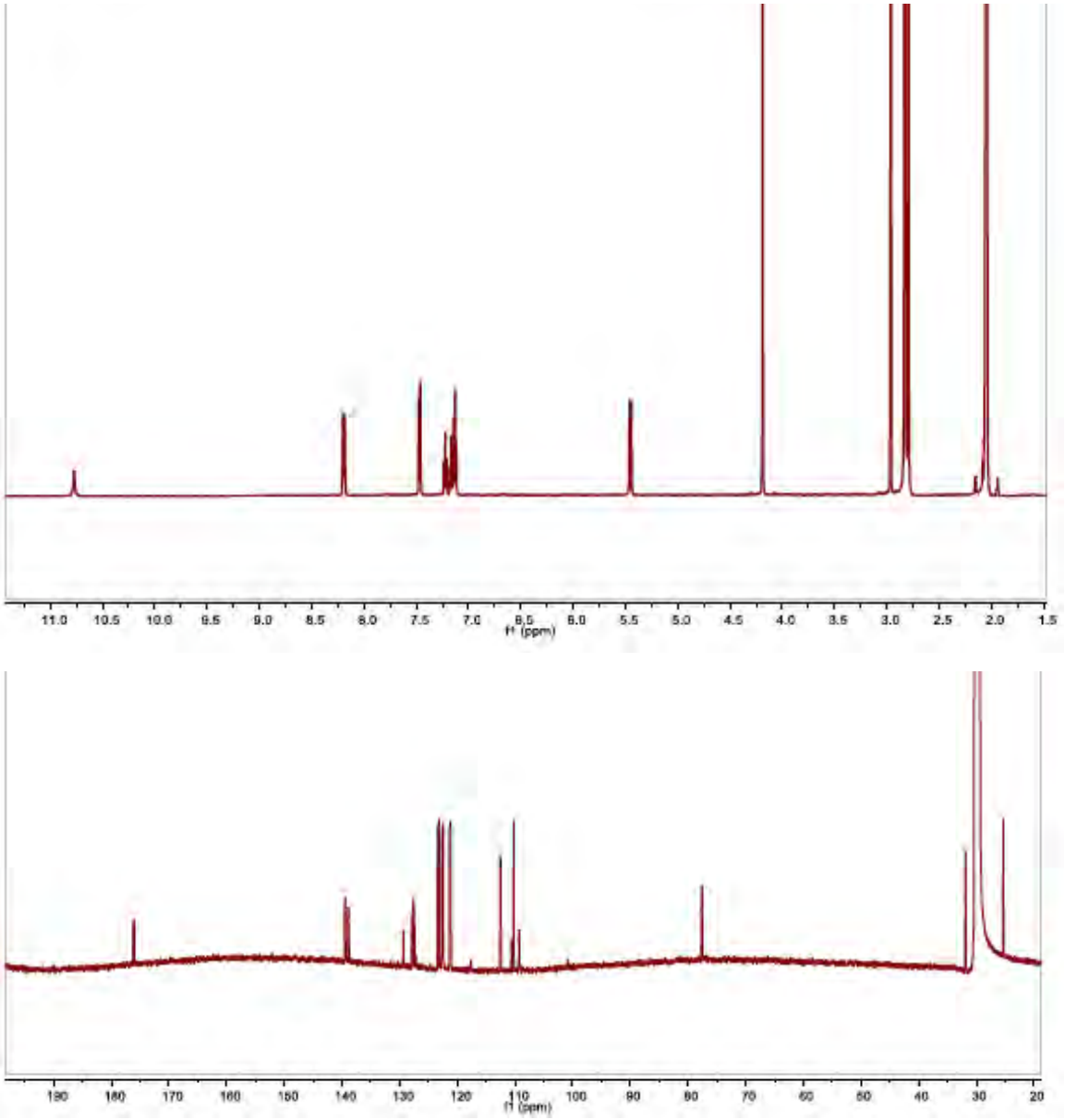
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **5** in acetone-d_6.



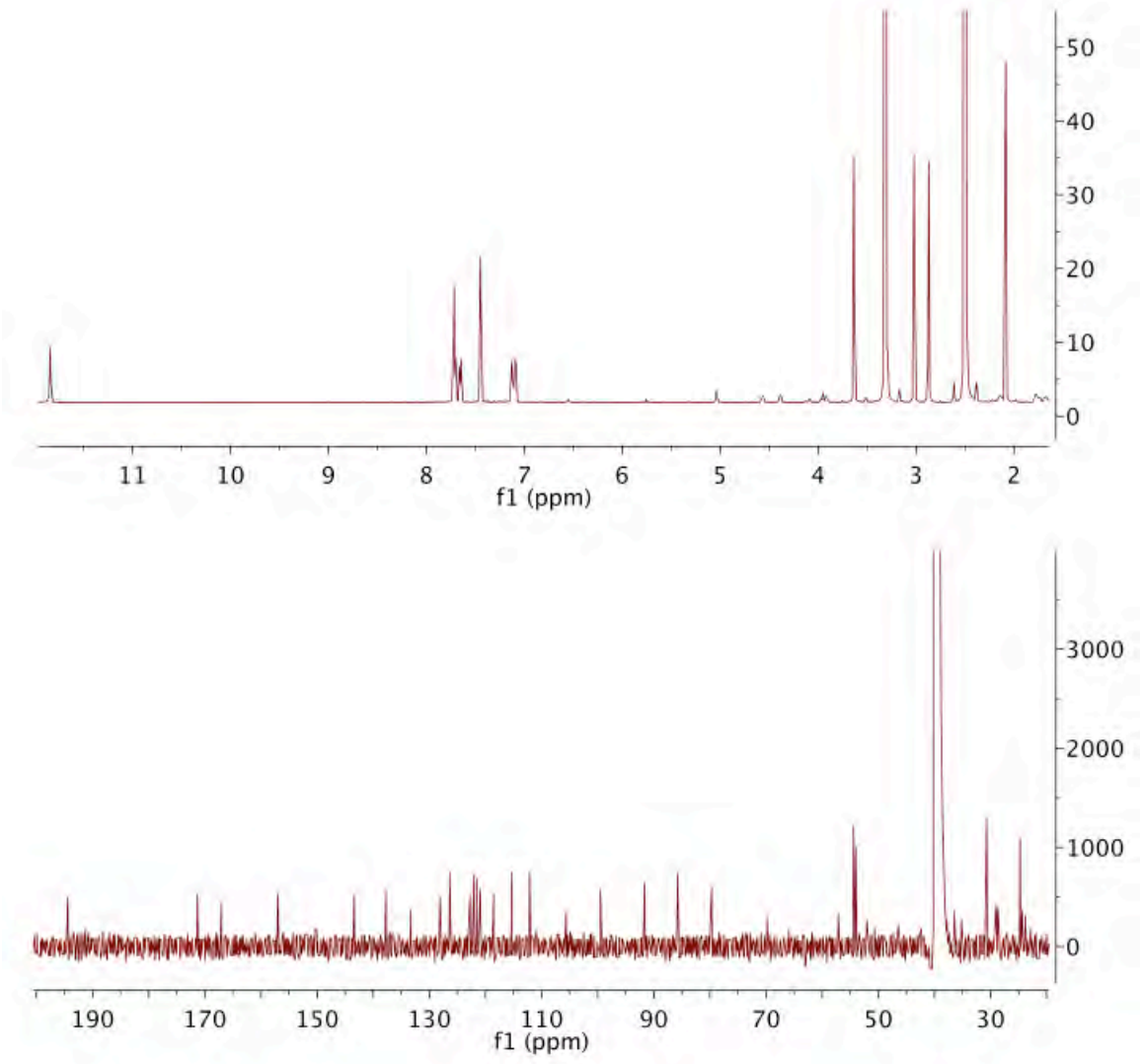
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **6** in acetone-d_6.



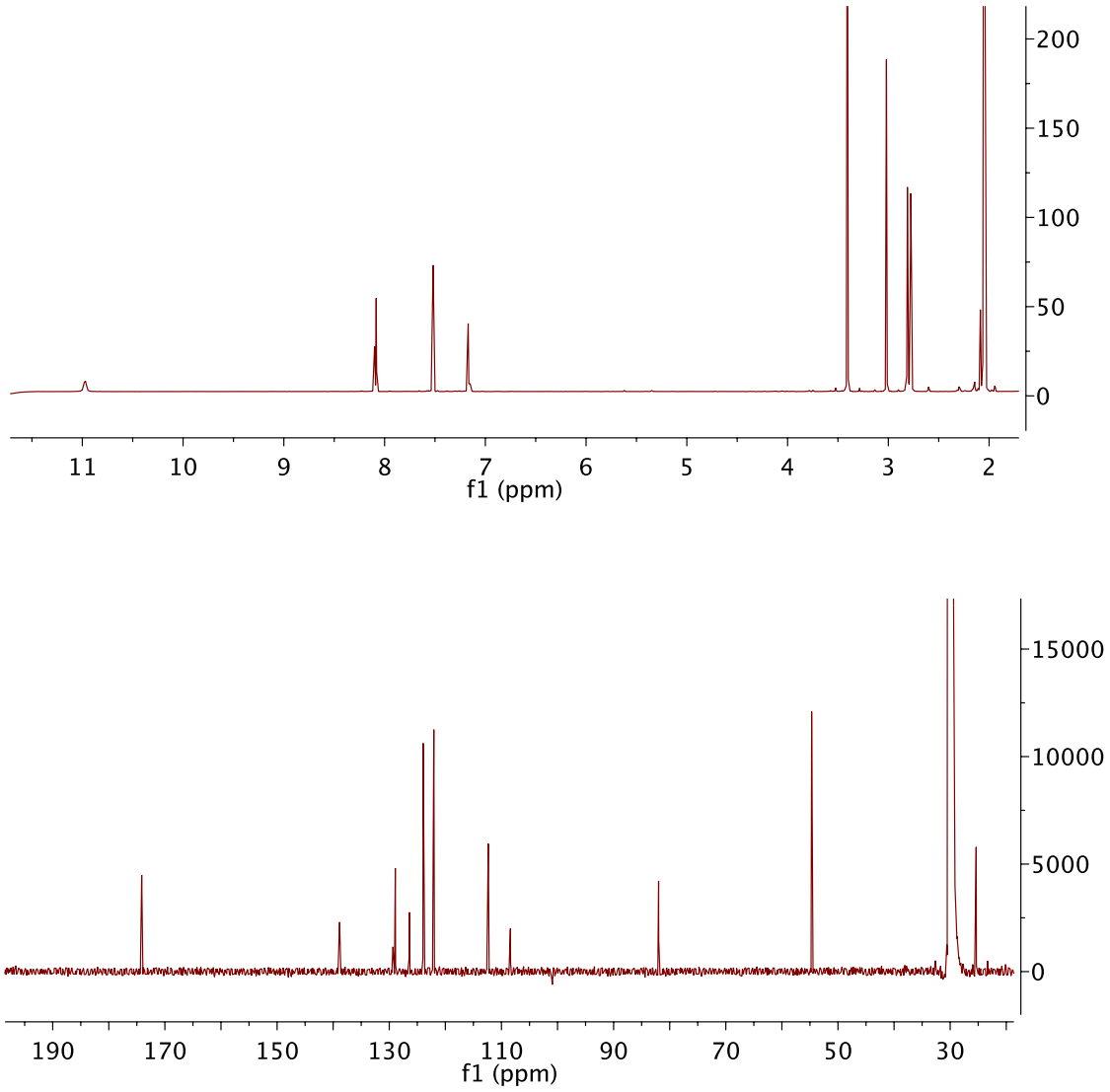
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **7** in DMSO-d_6.



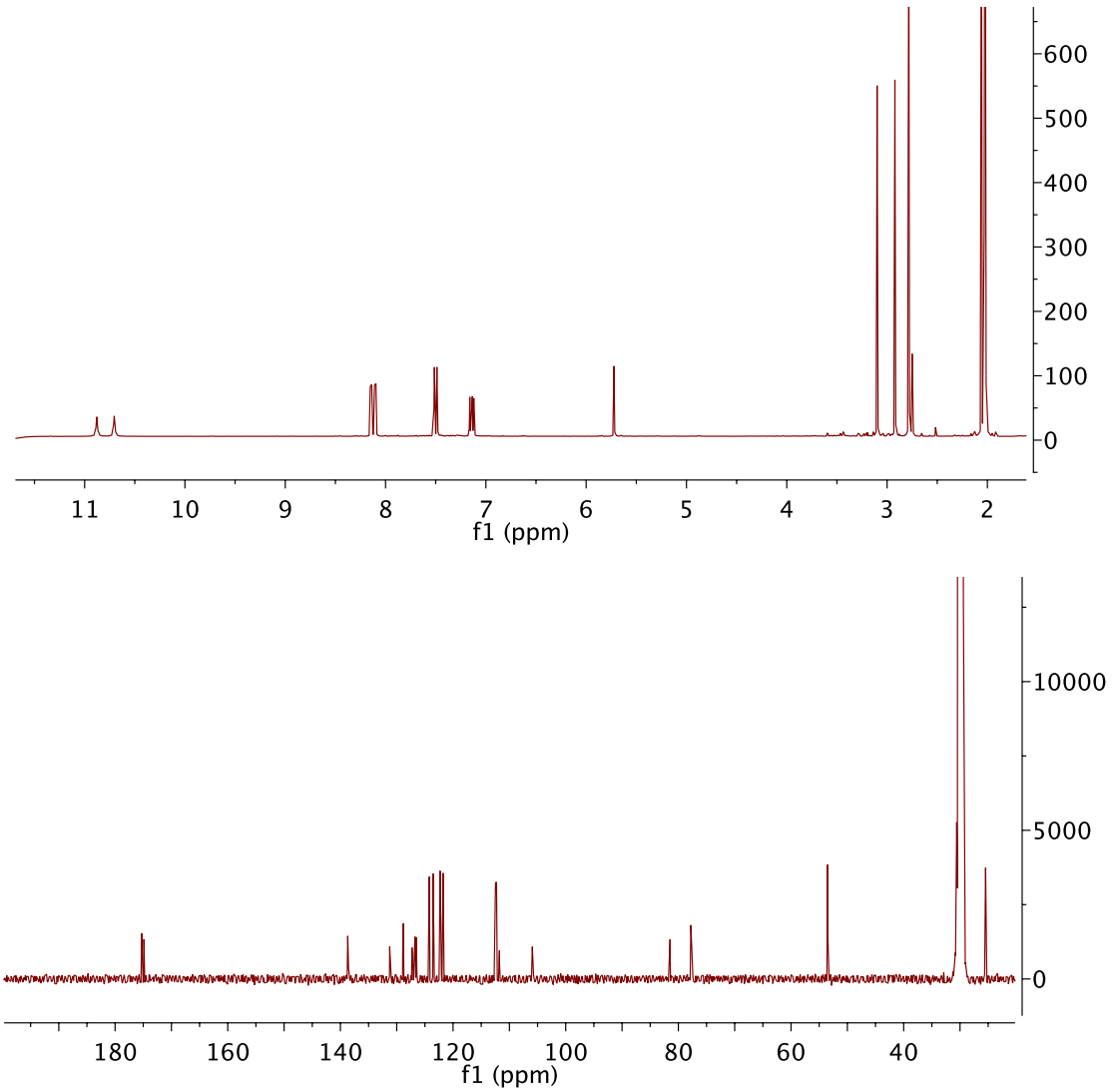
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **8** in acetone-d_6.



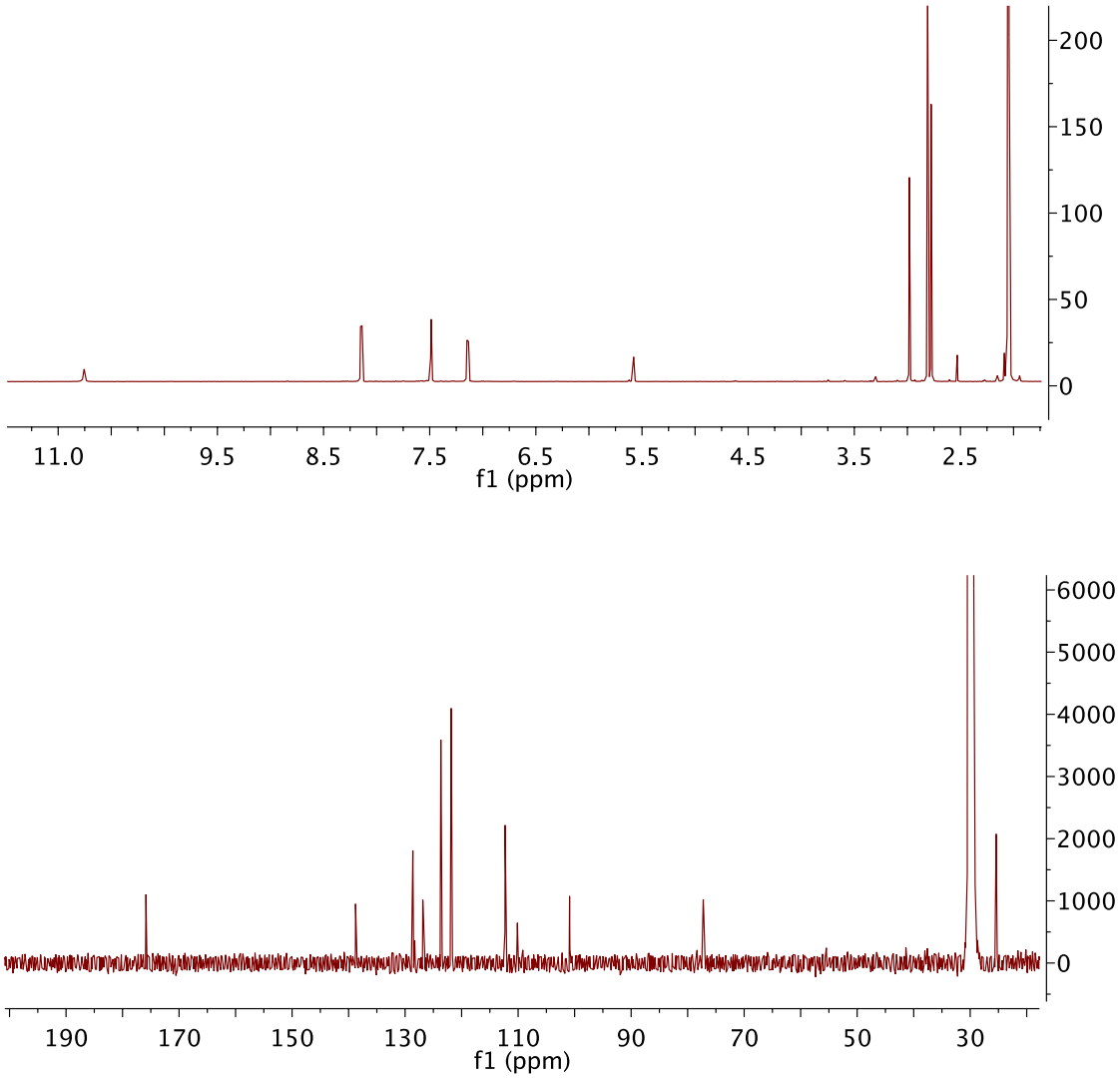
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **9** in DMSO-d_6.



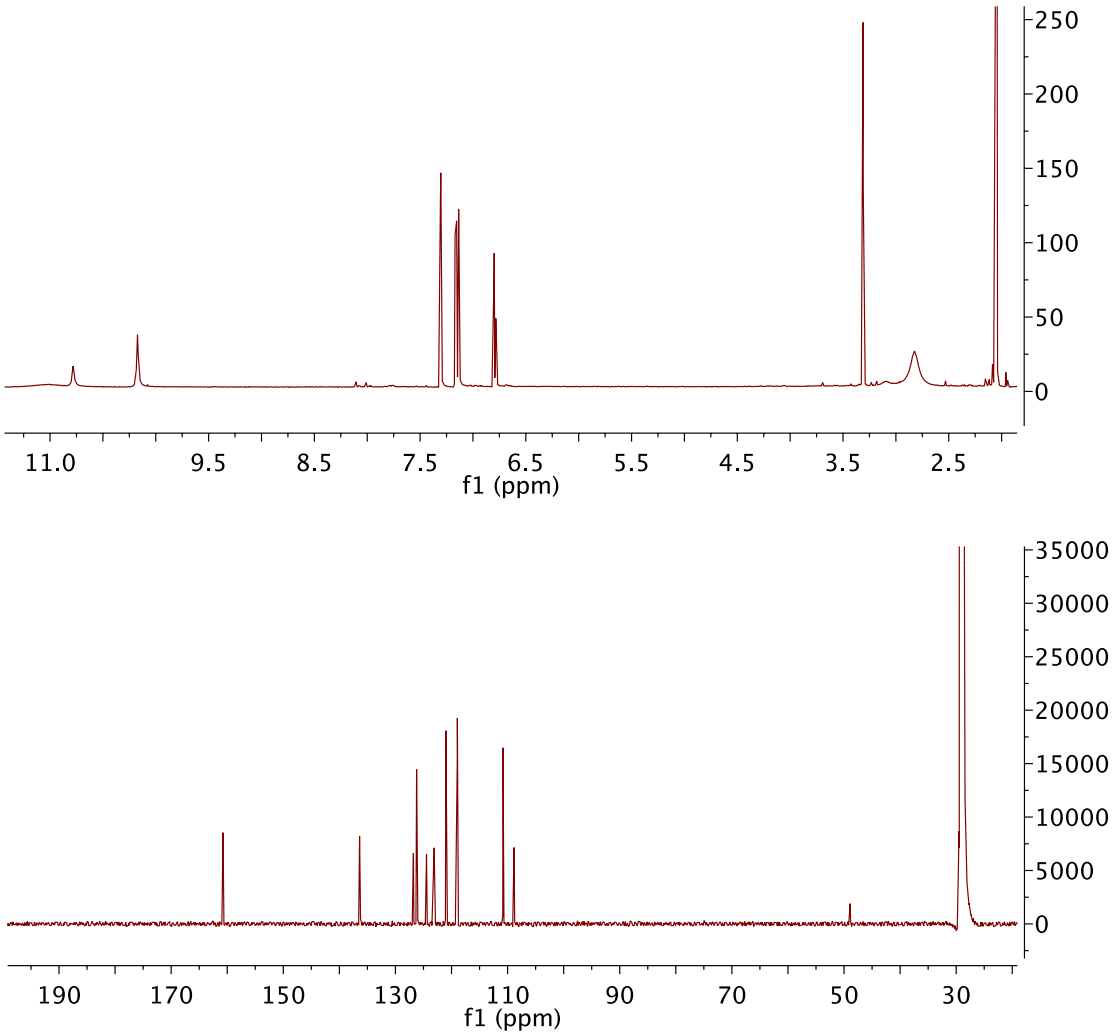
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **10** in acetone-d_6.



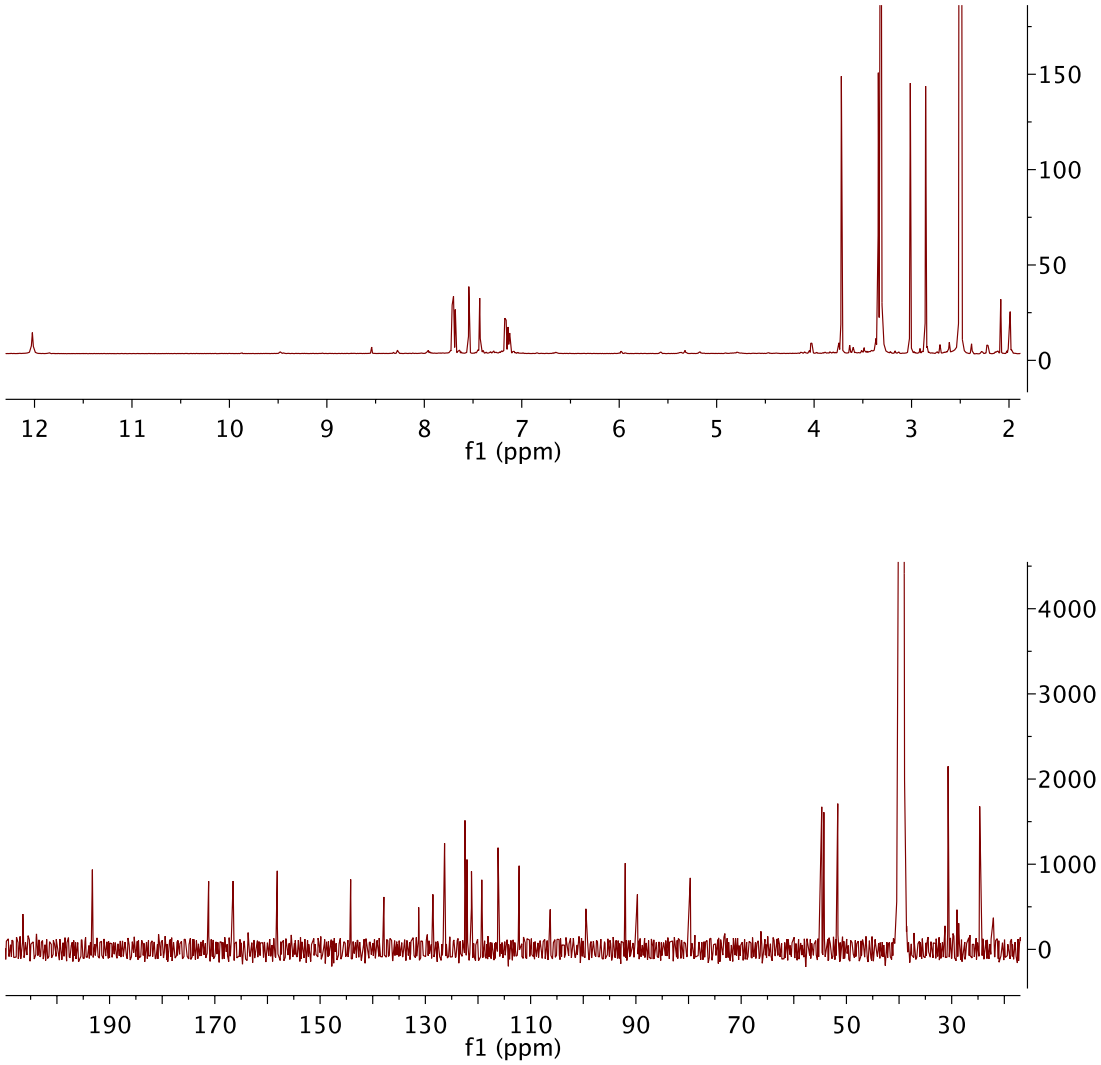
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **11** in acetone-d_6.



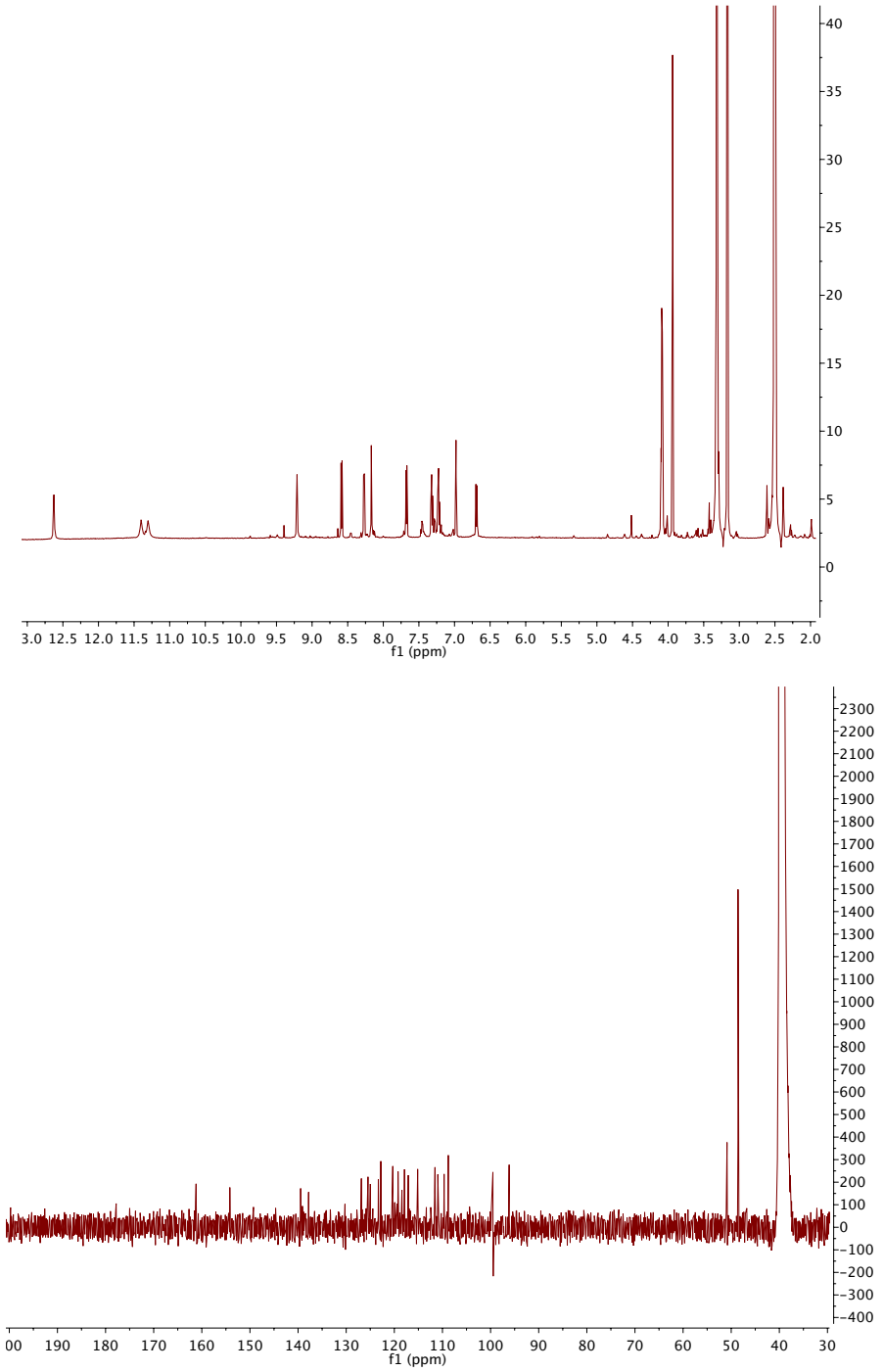
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **12** in acetone-d_6.



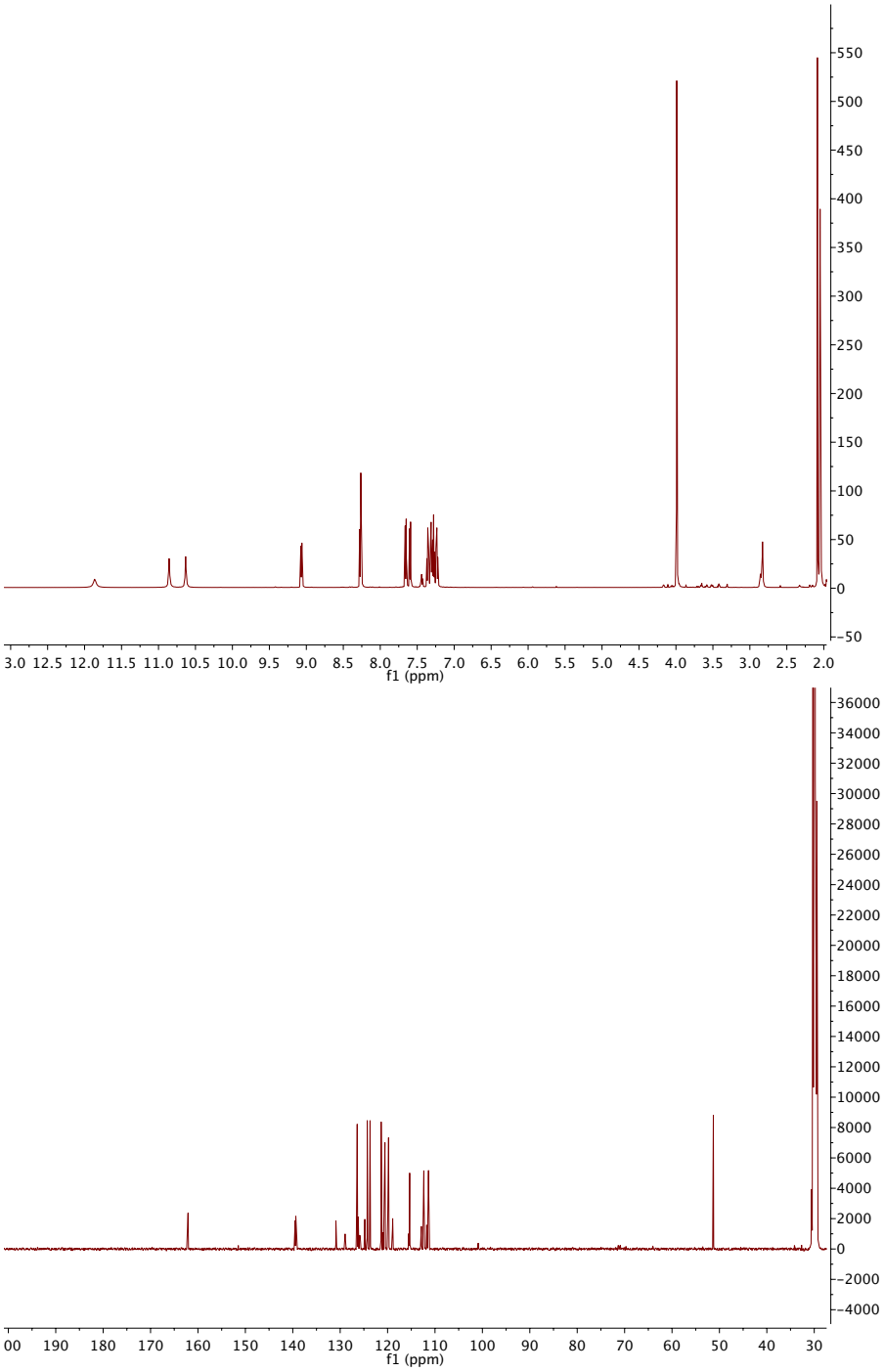
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **13** in acetone-d_6.



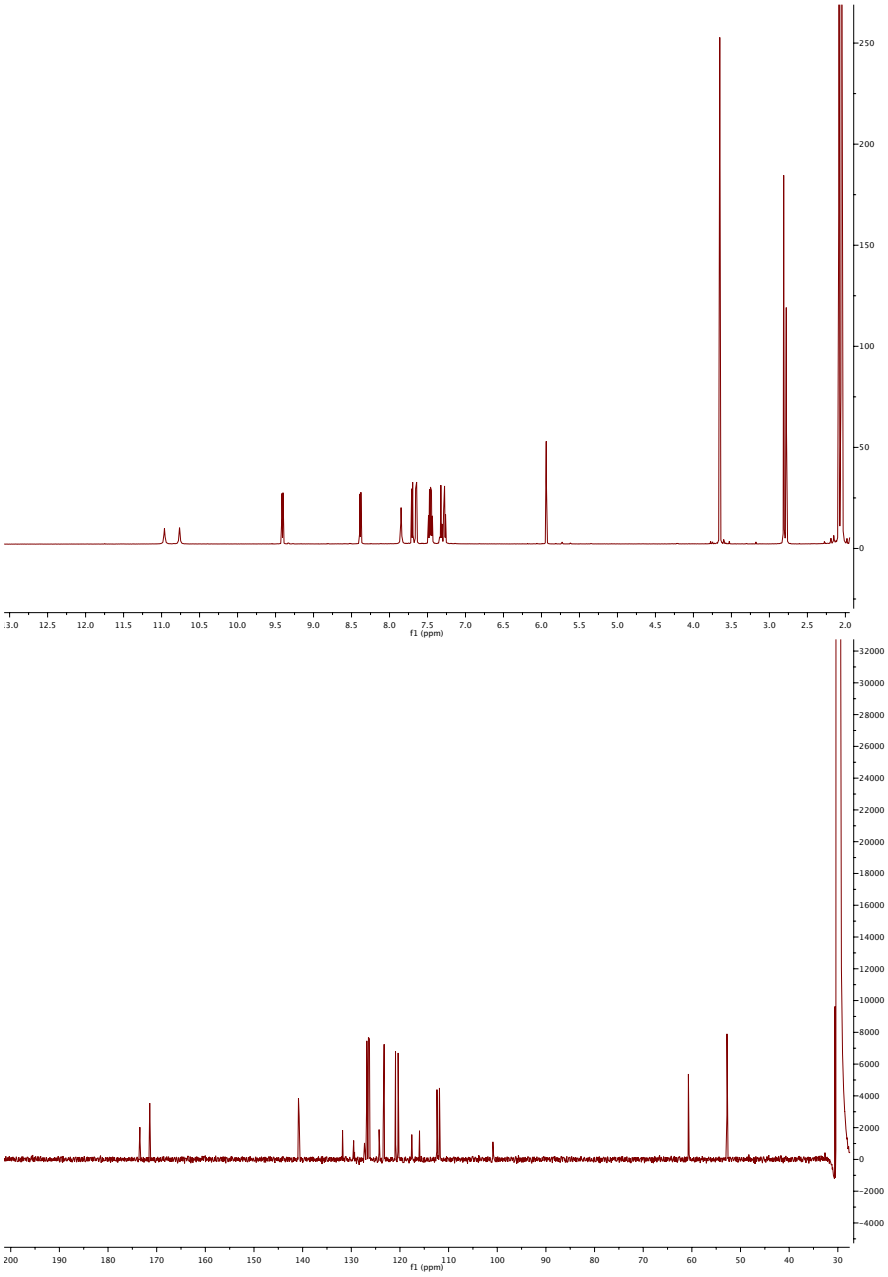
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **14** in DMSO-d_6.



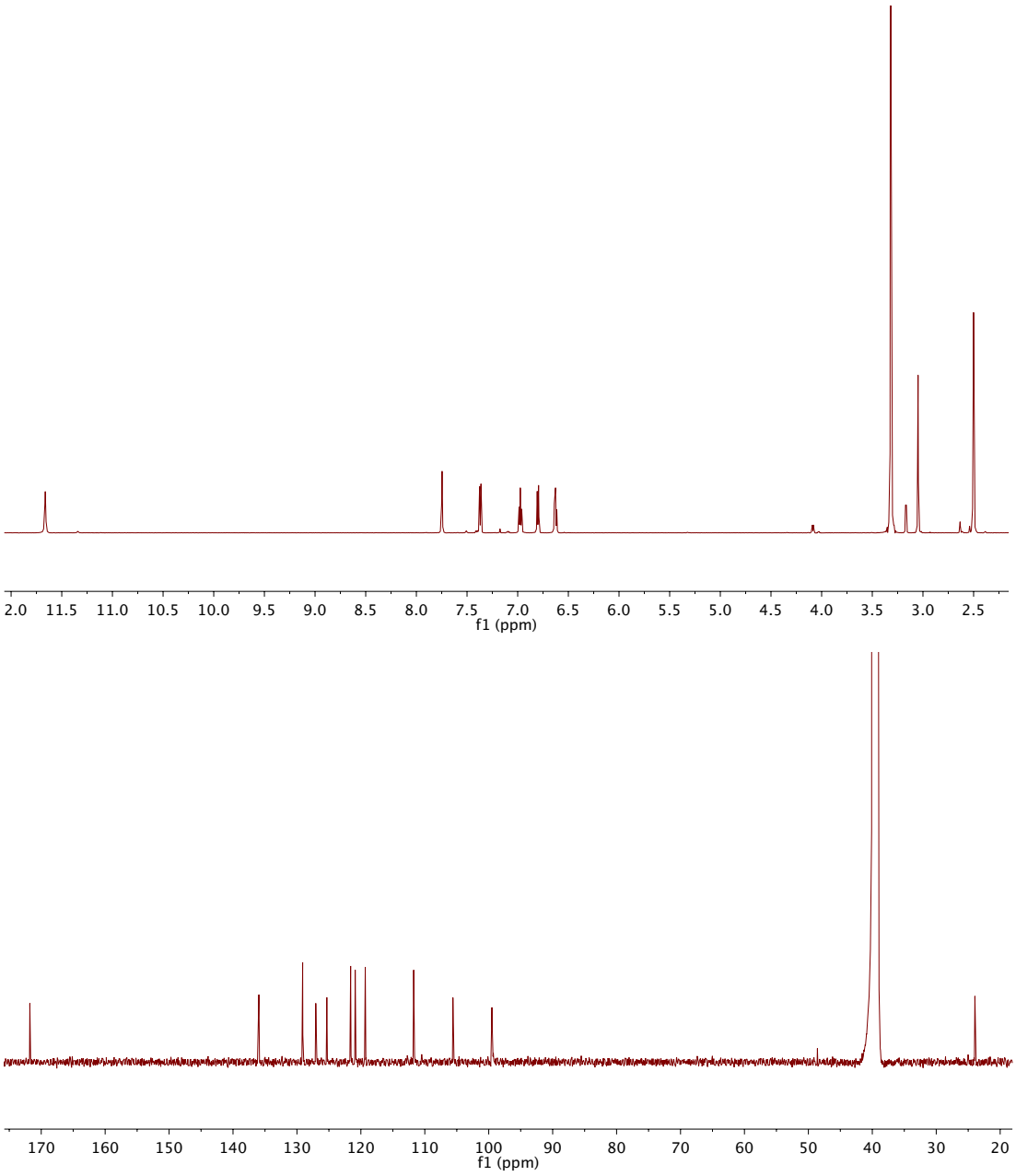
^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **15** in DMSO-d_6.



^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **16** in acetone-d_6.



^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **17** in acetone-d_6.



^{1}H NMR (**top,** 600 MHz) and ^{13}C NMR (**bottom,** 150 HMz) of compound **21** in DMSO-d_6.

Appendix 4: KinaseProfiler results

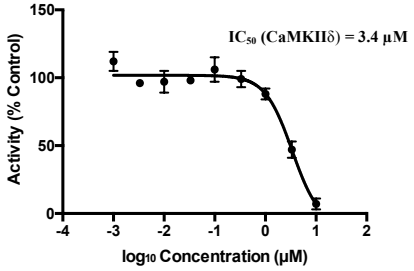
A

Kinase Target	Borrgomycin A @ 10 μM	Borregomycin B @ 10 μM
Abl(h)	87	101
ALK(h)	88	98
AMPKα1(h)	86	102
ASK1(h)	111	109
Aurora-A(h)	78	79
c-RAF(h)	108	97
CaMKI(h)	45	100
CDK1/cyclinB(h)	102	109
CDK2/cyclinA(h)	99	104
CDK6/cyclinD3(h)	105	114
CDK7/cyclinH/MAT1(h)	98	105
CDK9/cyclin T1(h)	113	108
CHK1(h)	92	107
CK1γ1(h)	94	111
CK2α2(h)	98	86
DRAK1(h)	107	113
eEF-2K(h)	97	146
EGFR(h)	106	133
EphA5(h)	97	108
EphB4(h)	102	92
Fyn(h)	84	104
GSK3β(h)	95	72
IGF-1R(h)	104	67
IKKα(h)	108	109
IRAK4(h)	100	114
JAK2(h)	101	125
KDR(h)	123	104
LOK(h)	95	116
Lyn(h)	104	81
MAPKAP-K2(h)	104	118
MEK1(h)	98	108
MKK7β(h)	135	106
MLK1(h)	102	123
Mnk2(h)	99	96
MSK2(h)	86	104
MST1(h)	101	117
mTOR(h)	104	107
NEK2(h)	95	94
p70S6K(h)	117	116
PAK2(h)	77	89
PDGFRβ(h)	94	133
Pim-1(h)	93	76
PKA(h)	85	79
PKBα(h)	94	122
PKCα(h)	107	110
PKCθ(h)	98	109
PKG1α(h)	106	98
Plk3(h)	100	104
PRAK(h)	79	66
ROCK-I(h)	107	111
Rse(h)	89	90
Rsk1(h)	92	106
SAPK2a(h)	100	102
SRPK1(h)	105	103
TAK1(h)	95	91
PI3KCα(h)	98	77
PI3KCβ(h)	92	83
PI3KCγ(h)	82	81
PI3KCδ(h)	90	69

B

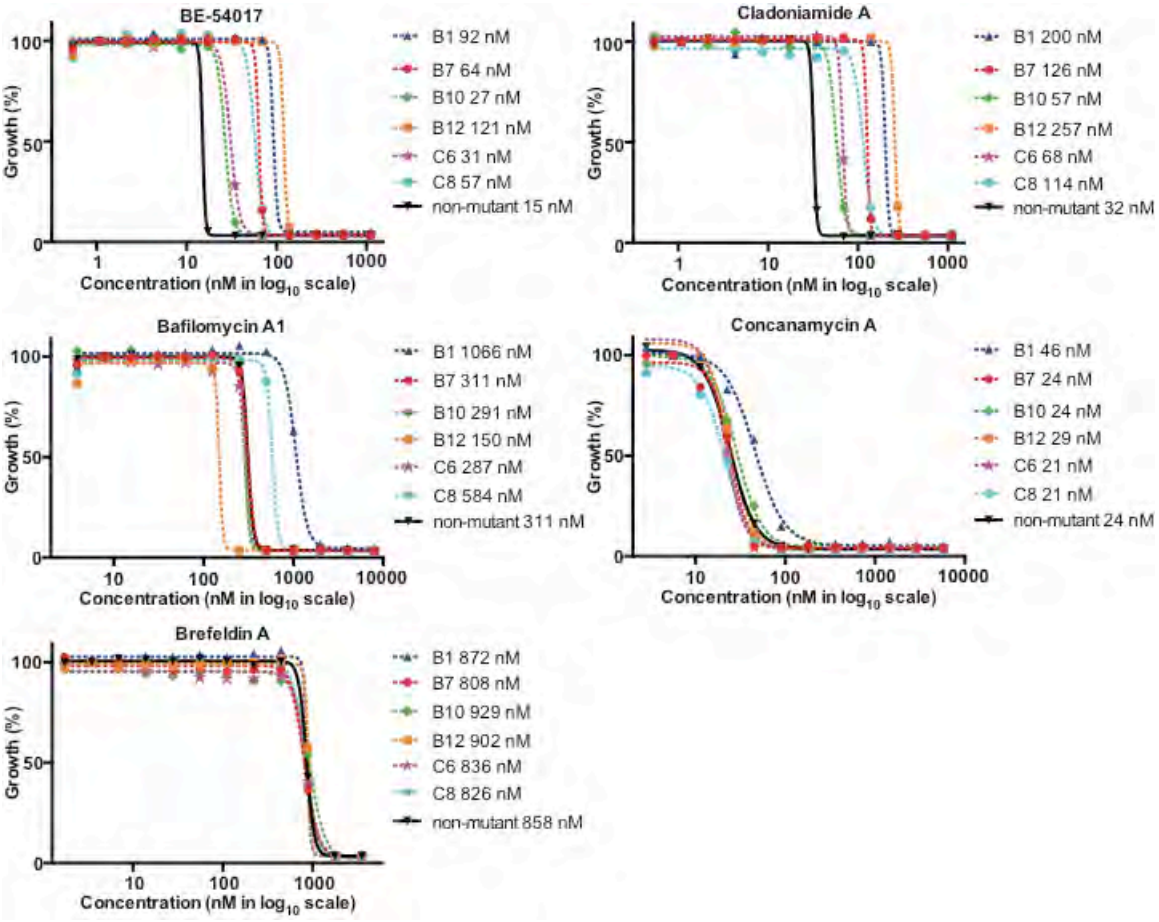
Kinase Target	Borreogmycin A @ 10 μM
CaMKIIβ(h)	51
CaMKIIγ(h)	76
CaMKIδ(h)	108
CaMKIIδ(h)	26
CaMKIV(h)	91
PhKγ2(h)	98

C



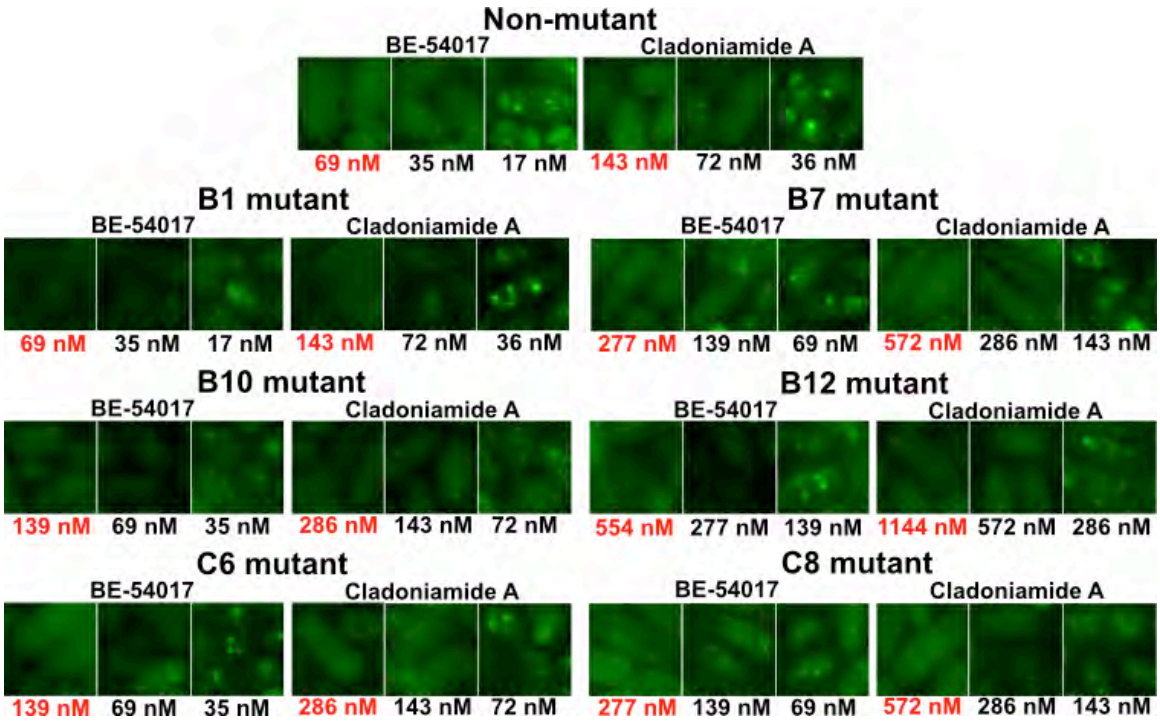
A and **B**) KinaseProfiler results against 59 diverse disease-relevant kinases (**A**) and CaMKI-related kinases (**B)** (Millipore). Numbers represent percentage of kinase target activity in the presence of 10 μM of borrgeomycin A (**9**) or B (**10**) normalized aganst a no-compound control. **C**) IC_{50}Profiler result for compound **1** against CaMKIIδ (Millipore). For IC_{50} determination the % residual activity data was curve fitted using Graphpad Prism. Error bar represents standard deviation where n=2.

Appendix 5: Whole-cell cytotoxicity dose response curves



Whole-cell cytotoxicity dose response curves of compounds tested in the target identificaiton study in resistant mutant and un-mutagenized MDR-sup *S. pombe* strains. The IC_{50} values determined from these dose response curves are noted on the legend of each graph.

Appendix 6: Fluorescent visualization of acidified vacuoles



Fluorescent images of quinacrine-stained resistant mutants or un-mutagenized MDR-sup *S. pombe* upon incubation with either BE-54017 or cladoniamide A. The minimal inhibitory concentration (MIC) (red) for *in vivo* V-ATPase activity was defined as the minimal concentration of compound at which no fluorescent puncta was observed in >95% of the cells.

Appendix 7: CPA synthase amplicons from TD biodiversity study

Chromopyrrolic acid (CPA) synthase amplicon sequences. Number of reads populating the amplicon sequences (NMCC_#) at 95% identity are tabulated by soil sample (NS1-NS20), with the consensus sequences listed below.

Amplicons	NS1	NS2	NS3	NS4	NS5	NS6	NS7	NS8	NS9	NS10	NS11	NS12	NS13	NS14	NS15	NS16	NS17	NS18	NS19	NS20
NMCC_72					3			3	2	327	44	34					185	17	1	11
NMCC_29					17					6		1			8		298			14
NMCC_4												1							318	
NMCC_350	231		1	4	1	32		6	1	3	12					4				1
NMCC_39	228		1	6		29		6		8	7	2				5				4
NMCC_1	136		85	1	61			1										8		
NMCC_8							2		5	260	5	4					10	1		
NMCC_30	18			156																
NMCC_14								162				2								
NMCC_11	96		2	13	2	27	1	3	1	6	6	2				2				2
NMCC_12	68		4			3	30													
NMCC_24			70	5	5		7				1			10						
NMCC_27					85															
NMCC_6											6		7	19	44		6			
NMCC_18			79																	
NMCC_15																				68
NMCC_100									58											
NMCC_241					52															
NMCC_10	50																			
NMCC_206	48																			
NMCC_28			39	6																
NMCC_166														1	43					
NMCC_89	1		37				6													
NMCC_22																	2		41	
NMCC_204											8			34						
NMCC_278								40												
NMCC_44																	38			
NMCC_268					1										1		34	1		1
NMCC_104												37								
NMCC_81			36																	
NMCC_5							1										34			

NMCC_72:

GCGTACTCGGTGCCGACGTACGCGGCGGCCGAGGCCTTCGTCCACCGCGGTC
TGTGGACCCCCGAGCAGCTGCGGGTCGCCCTGGGTGACGGCGGAGAGACCCT
CGACGGCGGCGTCCGGGGAAAGCTGGTGGAGATCGCGCGCGAGGAGATGAT
CCATTTCCTCCTCGTCAACAACATCCTCATGGCTCTCGGTGAGCCCTTCTGCG
TCCCCGCGCTGGACTTCGGCGCGCTCGGCACCGACCTGCCGGTGCCCCTGGA
CCTGTGCCTCGAAGGGCTCGACATCGCCACCGTGGCGCGGTTCATCGCGATC
GAGCAGCCCGCGGTGGGCACGCCGGAGGTACGGCGACCCGACCTGCCCACC
ACCACCGGCGCGGCGGGTGCGGGCCGGTACGAGACGTTGAGCGAGATGTAC
GCGCGGATCCGTCAGGGGCTGCAGGACGTCCCG

NMCC_29:

GCGTACTCGGTGCCGACGTACGGGGCGGCCGAGGACTTCGTCCACCGAGGTC
TGTGGACCCCCGAGCAGCTGCGGGTCACCGTGGGCGACGGCGGAGAGGCCCT
CGACGGCGGCGTCCGGGGAAAGCTGGTGGAGATCGCGCGCGAGGAGATGAT
CCATTTCCTCCTCGTCAACAACATCCTCATGGCGCTGGGCGAGCCCTTCTGCG
TGCCCGCACTGGACTTCGGCGCGCTCGGCTCCGACCTGCCGGTGCCCCTCGAC
CTCTGCCTCGAAGGGCTCGACATCGGCAGCGTGGCGCGGTTCATCGCGATCG
AGCAGCCCGCGGTGGGCACGCCGGAGGTGCGGCGACCCGACCTGCCCACCG
GCACCGACGCGGCGAGCGCCGGCCGGTACGAGACGTTGAGCGAGATGTACG
CGCGGATCCGCCAGGGCTTGGCGGACGTCCCGGA

NMCC_4:

GCGTACTCGGTGCCGACCTACGGGGCGGCCGAGACCTTCGTCCAGCGAGGTC
TGTGGACCCCCGAGCAGCTGCGGGTCACCCTTGGCGACGGCGGCGAGGCCCT
CGACGGCGGCGTCCGCGGAAAGCTGGTGGAGATCGCGCGCGAGGAGATGAT
CCATTTCCTCCTCGTCAACAACATCCTCATGGCGGTGGGGGAGCCCTTCTACG
TCCCGCGCTGGACTTCGGCGCGCTCGGCGCCGACCTGCCAGTGCCTCTCGACC
TCTGCCTCGAAGGGCTCGACACCGCCACCGTGGCGCGGTTCATCGCGATCGA
GCAGCCCGCGACCGAGACGCCGGTGGTCCGGCGACCCGACCTGCCCCCCACC
ACCGCACCGGGTGTCGGTACGAGACGTTGAGCGAGATGTACGCGCGGATCCG
CCGGGGCCTGGCGGACGTGCCGGATCTCTTC

NMCC_350:

GCGTACTCCGTGCCGACCTACGGCGTCGGCGCGGCGTGGGTGCGCGAGGGCC
GGTGGACCGCGGAACACCTCGAACTCGCCTGCGGCGACGGCGGGCGGACGC
TGCGCACCGGCATCCGGGGCGGCCTGCTCGGGGTCGCCCGCGAGGAGATGGC
GCACTACCTGGTCGTCAACAACGTCCTCATGGCCGTGGGCGAGCCGTTCCAC
GTGCCCGAGGTGGACTTCGCCACGATCAACGCCCGCATGCCGCTGCCGGTGG
ACTTCGCGCTCGAACCGCTGCACCTGGGCAGCGTGCAGCGGTTCATCGCCAT
CGAGGAACCCGACGACGGCCGGCCGGGCGGCGGGCCGTACCGGTCGCTGAG

CGAGCTGTACGGCGCCATCCGCGACGGCCTGACCCGGGTGCCCGACCTGTTC
GCGTCGAGCGCGGCAGGGGCGCGGCGAGCACCGA

NMCC_39:
GCGTACTCCGTGCCGACCTACGGCGTCGGCGCGGCGTGGGTGCGCGAGGGCC
GGTGGACCGCGGAACACCTCGAACTCGCCTGCGGCGACGGCGGGCGGACGC
TGCGCACCGGCATCCGGGGCGGCCTGCTCGGGGTCGCCCGCGAGGAGATGGC
GCACTACCTGGTCGTCAACAACGTCCTCATGGCCGTGGGCGAGCCGTTCCAC
GTGCCCGAGGTGGACTTCGCCACGATCAACGCCCGCATGCCGCTGCCGGTGG
ACTTCGCGCTCGAACCGCTGCACCTGGGCAGCGTGCAGCGGTTCATCGCCAT
CGAGGAACCCGACGACGGCCGGCCGGGCGGCGGGCCGTACCGGTCGCTGAG
CGAGCTGTACGGCGCCATCCGCGACGGCCTGACCCGGGTGCCCGACCTGTTC
GCGGTCGAGCGCGGCAGGGGCGGCGGCGAGCACCGC

NMCC_1:
GCGTACTCGGTGCCGACCTACGGCGTGGGCGAGGCGTGGGTGCGCGGGGCCG
GTGGACCGCGGAGCAGCTCGAACTCGCCTGCGGCGACGGCGGGCGGACGCT
GCACACCGGCATCCGGGGCGGCCTGCTCGGCGTCGCCCGCGAGGAGATGGTC
CACTACCTGGTCGTCAACAACATCCTCATGGCCGTCGGCGAACCGTTCCTGGT
GCCGCGGGTGGACTTCGCCACGATCAACGCGGACCTGCCGCTGCCGGTGGAC
TTCGCGCTCGAACCGCTGCACCTCGGCAGCGCGCAGCGGTTCGCCGCCATCG
AGGAGCCGGGCGACGGCGGCCCGGGCACCGGGCCGTACCGGTCGCTGAGCG
AGCTGTACGGCTCCATCCGCGACGGCCTGTCCCGGGTGCCGGACCTGTTCGC
GGTCGAGCGCGGCCGGGGCGGCGGCGAGCACCG

NMCC_8:
GCGTACTCGGTGCCGACGTACGCGGCGGCCGAGGCCTTCGTCCACCGCGGTC
TGTGGACCCCCGAGCAACTGCAGGTCACCCTGGGTGACGGCGGAGAGGCCCT
CGACGGCGGCGTCCGGGGAAAGCTGGTGGAGATCGCGCGCGAGGAGATGAT
CCATTTCCTCCTCGTCAACAACATCCTCATGGCTCTGGGCGAGCCCTTCTGCG
TCCCCGCGCTGGACTTCGGCGCGCTCGGCAGCGACCTGCCGGTGCCCCTGGA

CCTGTGCCTGGAAGGGCTCGACATCGCCACCGTGGCGCGGTTCATCGCGATC
GAGCAGCCCGCGGTGGGCACACCGGAGGTGCGGCGACCCGACCTGCCCCCC
ACACCAGCCCGGCGGGTGCGGGCCGGTACGAGACGTTGAGCGAGATGTACG
CGCGGATCCGCCAGGGGCTGCAGGACGTCCCGGAT

NMCC_30:
CTGTGGTCCATCCCCACCTACTCGGCGGGCGCCCAGTACGTCAGGCGGGGGA
GTGGACCCCCGAACAGCTCCGGCTGATGTGCGGCGCGGGGCCGCACCGCATC
GACGGAGGCGTCCGGCAGCGCCTGTTCGAGGTGGCCAGGGAGGAGATGATC
CATTTCCTGCTGATCAACAACATCGTCATGGCGACCGGGCAGCCGTTCCACCT
TCCGGCGATCGACTTCGGCACGGTGAACAACGAGCTGCCGGTGCCGCTCGAC
TTCTGCCTGGAGCCCTTCGGACGCGGCGCCCTGGAGCGGTTCATCGCGCTGG
AGCGGCCCTACGACCTGGTCAGGGACATCGCCGGGAACGACGCGCCGGCCG
GCGGCGTGCCCGAGGGGCGCGCCCCGTACGGTACGGCTCGCTGAGCGAGCTG
TACTCGGCCATCCGCGAAGCCGTCAGGCCATCC

NMCC_14:
GCCTACTCGGTACCCACCCACGGCGCCGGCGCGGAGTACGTACGCCGGGGCC
TGTGGACGCCCGACCAACTGCGGCTCGCGTGCGGTGACGGCGGGGAGACCCT
CGACGAGGGCATCCGCAGCATGCTGCTGACCATCGCCCGCGAGGAGATGATC
CACTTCCTCCTCGTCAACAACATCCTCATGGCGGTGGGCGAACCCTTCCACGC
GCCCCGGATCGACTTCGGCACCGTCAACCGACAGCTGGCCGTCCCGCTGGAC
TTCGCCCTGGAGCGCCTGGGGCCCGGCAGCGTGGAGCGGTTCGTACAGATCG
AACGCCCCGAGGACCTCGTCGACGAGGTACGGCGCGGCGACGCTCCGGCGCC
CCCGCCGCGTACGACGAGCGGCACCCGTACGCCTCGCTGAGCGAGCTGTACG
CGGACATCCGGGAAGGGCTGGAGAGCATCCCG

NMCC_11:
GCGTACTCCGTGCCGACCTACGGCGTCGGCGCGGCGTGGGTGCGCGAGGGCC
GGTGGACCGCGGAACACCTCGAACTCGCCTGCGGCGACGGCGGGCGGACGC
TGCGCACCGGCATCCGGGGCGGCCTGCTCGGGGTCGCCCGCGAGGAGATGGC

GCACTACCTGGTCGTCAACAACGTCCTCATGGCCGTGGGCGAGCCGTTCCAC
GTGCCCGAGGTGGACTTCGCCACGATCAACGCCCGCATGCCGCTGCCGGTGG
ACTTCGCGCTCGAACCGCTGCACCTGGGCAGCGTGCAGCGGTTCATCGCCAT
CGAGGAACCCGACGACGGCCGGCCGGGCGGCGGGCCGTACCGGTCGCTGAG
CGAGCTGTACGGCGCCATCCGCGACGGCCTGACCCGGGTGCCCGACCTGTTC
GCGGTCGAGCGCGCAGGGGCGCGGCGAGCACCGA

NMCC_12:
GCGTACTCGGTGCCGGCCTACGGGGCGGGGGAGGAGTACGTCCGGCGCGGG
CTGTGGACCCCCGAACAGCTGCGGCTCGCCTGTGGGGACGGCGGCCGGACCC
GCGACGGGGGCATCCGCGGCACGCTGCTCGGCATCGCCCGCGAGGAAATGAT
CCACTTCCTGATCGTCAACAACATCATCATGGCGATGGGCGAACCGTTCTAC
GTCCCCGACGTCGACTTCGGCACGATCAACAACACCCTGCCCGTGCCGCTGG
ACTTCGCCCTGGAGCCCCTCGGCGTGGGCAGCGTGCAGCGGTTCATCGCGAT
CGAACGGCCGGAGGACCAGGTCGGCGAGCTCCACCCCCCGGTCCCGGTTCGG
TTCCCCGCCGGCCGCCGAGCACCCTTACGCCTCGCTCAGCGAGCTGTACGGC
GACATCCGCGAGGCCTGCAACGGGTCCCCGACGTC

NMCC_24:
CTGTGGTCCATCCCCACCCACTCGGCCGGGACCGAGTACGTGCGGCGAGGCG
AGTGGACGCCCGGGCAGCTCCGGCTCATGTGCGGCGAGGGCCCGCACAGCCT
CGACGGCGGCGTCCGGCAGGACCTGTTCGCCGTCGCCCGCGAGGAAATGATC
CACTTCCTGCTGATCAACAACATCATCATGGCGACGGGTCAGCCGTTCCACCT
GCCCCGGATCGACTTCGGCACGGTGAACGGCGAGCTCCCGTCCCGTCGACCT
GTGCCTGGAGCCCTTCGGGCGGGGCAGCCTGCAACGGTTCGCCGCTCTGGAA
CGGCCCTACGACCTGGTCCGCGACCTGGCCGCCGACCGGCCACCGGTGACCG
CGCGCGACCCGTACCCCTACGGTTCGCTCAGCGAGCTGTACGGGGCCATCCG
CCAGGCGATCCAGGACATCCCGGACGTCTTCC

NMCC_27:

GCGTACTCGGTGCCGACGTACGGGGTGGCGGAGGCTTTTGTCCGGCGCGGTC
TGTGGACGCCCGAGCAACTGCAGGTCACCCTGGGCGACGGTGGCGAGGCGCT
CGACAGCGGCGTGCGGGGGAAGCTGATCGAGATCGCGCGCGAGGAGATGAT
CCATTTCCTCCTCGTCAACAACATCCTCATGGCCGTCGGCGAGCCCTTCTGCG
CCCCCGTACTGGACTTCGGGACGCTCGGCGACGATCTCCCGATCCCGCTGGA
CCTCTCCCTCGAGGGGCTCGACATCGGCAGCGTGCAGCGTTTCATCGCGATC
GAGCAGCCCGCGGTGGCCACGCCCGAGGTGAGGCGAGCCGACCTGCCGGTC
ACCACATCCGGACGCGGGTCCGGGCCGCTACGAGACGTTGAGCGAGATGTAC
GCGCGCATCCGGCCAGGGGCCTGGCAGGGACGGTC

NMCC_6:
GCGTACTCCATCCCGACCCACGGCGCCGGGGTGGAGCACGTCCGCCGCGGCC
TGTGGACCCCGCAGCAGCTGGAGCTGGCGTGCGGTGACGGCGGCGCCACCAC
CGCCGGCGGGCTGCGCGGCATGCTGCTCGGCGTGGCCCGCGAGGAGATGATC
CACTTCCTGCTGGTCAACAACATCATCATGGCGATGGGGGAGCCGTTCCACG
TGCCGGTGGTGGACTTCGGCACCGTCAACACCACCCTGCCGGTGCCGCTGGA
CGTCAGCCTGGAGGCCCTCAACCTCGGCAGCGTCCAACGGTTCATCGCCATC
GAGCGGCCGGACTGCGAGGTCGGCGAGCTGCGCCGGGCGCCGGGCACCGGA
GACCCGGCTCGCCGCCGACCCGGCCCGGTCGCCGCGCTGCTACGGCACCCGT
CAGCGAGCTGTACGCCGAGAGTCCGGGAGGGCC

NMCC_18:
GCGTACTCCATCCCGACCTACGGCGCCGGTGTGGAGTTGGTGCGGCAGGGCC
GGTGGACCACCGAACAGCTCGAGCTGGCCTGCGGCGACGGCGGCCAGACCCT
GGACACCGGGATGCGGGGCGGCCTGCTCGGCGTCGCCCGCGAGGAGATGGT
CCACTACCTTGTCGTCAACAACATCCTCATGGCCGTCGGCGAGCCGTTCTTCG
TGCCTCAGGTGGACTTCGGCACCATCAACGTCTCGCTGCCGGTACCGGTGGA
CTTCGCCCTCGAGCCGCTCGGCCTCGGCAGCGTGCAGCGGTTCGTCACCATCG
AGCAGCCGAGCGACGTCCCACCGGGCTCCGGACCGTACCGGTCGTTGAGCGA
GCTGTACGGCGCCATCCGCGACGGACTGTCCCGCGTGCCGGACCTGTTCATG
GTCGAACCCGGCAGGGGCGCGGGGAACAC

NMCC_15:

CTGTGGTCCATCCCCACCCACTCGGCCGGGACCGAGCTCGTGCGGCGCGGGG
AGTGGACGCCCGGTCAGCTCCGGCTCATGTGCGGCGATGGCCCGCAGAGCCT
CGACGGCGGCATCCGGCATGCCCTGTTCGCCGTCGCCCGTGAGGAAATGATC
CACTTCCTGCTGATCAACAACATCATCATGGCGACCGGCCAGCCGTTCCACCT
GCCCCGAATCGACTTCGGAACGGTGAACACCGAGCTCCCCGTTCTCGTCGAC
CTGTGCCTGGAACCCTTCGGGCGGGGCAGCCTGCAGCGGTTCGCCGCACTGG
AACGCCCCCACGACACGGTCCGCGGCCTCGCAGAGGACCACCCGTCGGTGAC
CGCGTGCCACCCGTACCCCTACGGTTCGCTCAGCGAGCTGTACGGGGCCATC
CGCCAGGCCGTTCAGGACATCCCGGATGTCCT

NMCC_100:

ATGTGGTCGCTGCCGACTCACCGTATGGGGGCCGCGCTGGTCCACCAGGGCG
AGTGGACCGAGGACCAGTACACGCTGGTCTGCGGGCGCGGGCCGGCGACGG
CTGACGGGGGGATCCGGGGCGCTCTGTTCGGGGTGGCCCGAGAGGAGATGAT
TCATTTCCTCGTCATCAACAACATCATCATGGCCACGGGTCAGCCCTTCCACG
TGCCCGACATCAACTTCTCCTCGCTCAACGCGCAGATCGACCTGCCGATGGA
CTTCTGCCTGGAGCGGTTCGGCCTGTCCTCGCTGAGCAGGTTCGTCGAGTTTG
AAAGCCGTTTTCGCTCACCGTCGAGCCAGCGCCGGCACCGGGAGGCCCGCAG
CGAAAGGGTGCCCGTTACGGGTCGCTGAGCGAGCTCTACGCGTCGATCCGGG
ACGGCCTCGCCCGGTTCCCGAGGGCCTTCCT

NMCC_241:

GCGTACTCGGTGCCGACGTACGGGGTGGCGGAGGCTTTCGTCCGGCGCGGTC
TGTGGACGCCCGAGCAACTGCAGGTCACCCTGGGCGACGGTGGCGAGGCGCT
CGACAGCGGCGTGCGGGGGAAGCTGATCGAGATCGCGCGCGAGGAGATGAT
CCATTTCCTCCTCGTCAACAACATCCTCATGGCCGTCGGCGAGCCCTTCTGCG
CCCCCGTACTGGACTTCGGGACGCTCGGCGACGATCTCCCGATCCCGCTGGA
CCTCTCCCTCGAGGGGCTTGACATCGGCAGCGTGCAGCGTTTCATCGCGATCG
AGCAGCCCGCGGTGGCCACGCCCGAGGTGAGGCGAGCCGACCTGCCGGTCA

CCACATCCGACGCGGGTCCGGGCCGCTACGAGACGTTGAGCGAGATGTACGC
GCGCATCCGCCAGGGCCTGCAGGAGGTCCCG

NMCC_10:
GCCTACTCGATCCCCACGTACGGCGCCGGCCAGGAGCACGTGCGCCGTGGGC
TGTGGACGCCCGAACAGCTCGCGCTGGTCTGCGGCGACGGCGGGGAGACCAC
CGCCGGCGGGATCCGTGGCACGTTGCTGTCCGTGGCGCGGGAAGAGATGATC
CACTTTTTGTTGATCAATAACGTGATCATGGCGATGGGTGAGCCGTTCTTCGT
GCCGACCGTCGACTTCGGCACCATCAACAACACCCTGCCACTCCCCCTCGAC
CTGGCGTTGGAGCAGTTCGGCATCGGCAGCGTCCAGCGGTTCATCGCCATCG
AACGACCCCACGCCCAGGACGGCGAGATCCAGATCGGCACCGACCAGGGCA
GCGGCGGGCCGCTCGCCGAACCGACCCACACCTACAGCTCGCTCAGTGGCCT
CTACGCCGACATCCGGGAAGGACTGCAGCGGG

NMCC_206:
GCCTACTCGATCCCCACGTACGGCGCCGGCCAGGAGCACGTGCGCCGTGGGC
TGTGGACGCCCGAACAGCTCGCGCTGGTCTGCGGCGACGGCGGGGAGACCAC
CGCCGGCGGGATCCGTGGCACGTTGCTGTCCGTGGCGCGGGAAGAGATGATC
CACTTTTTGTTGATCAATAACGTGATCATGGCGATGGGTGAGCCGTTCTTCGT
GCCGACCGTCGACTTCGGCACCATCAACAACACCCTGCCACTCCCCCTCGAC
CTGGCGTTGGAGCAGTTCGGCATCGGCAGCGTCCAGCGGTTCATCGCCATCG
AACGACCCCACGCCCAGGACGGCGAGATCCAGATCGGCACCGACCAGGGCA
GCGGCGGGCCGCTCGCCGAACCGACCCACACCTACAGCTCGGCTCAGTGGCC
CTCTACGCCGACATCCGGGAAGGACCTGCG

NMCC_28:
GCGTACTCCATCCCGACCTACGGCGCCGGCGTGGAGCAGGTACGGCGGGGCC
GATGGACCACCCAACAGCTAGAGCTGGCCTGCGGCGACGGCGGCCAGACCCT
GCACACCGGGATCCGGGGCGGCCTGCTCGGCGTCGCCCGCGAGGAGATGATC
CACTTCCTGGTCGTCAACAACATCCTCATGGCCGTCGGCGAACCGTTCTTCGT
GCCCCGAGTGGACTTCGGCACCCTCAACGCCGAACTTCCGCTCCCGGTGGAC

TTCGCCCTCGAACCGCTACACCTTGGCAGCGTGCAGCGGTTCATCACCATCGA
GCGACCCAGCGACATCGCACCGAGCACCGGACCGTACCGGTCGCTGAGCGA
GCTGTACGGCGCCATCCGCGACGGCCTGTCCCGCGTACCGGACCTGTTCATG
GTCGAACCCGGCAGAGGCGGCGGGGAACACCA

NMCC_166:
GCGTACTCCGTCCCGACCCACGGTGCCGGGGTGGAGCACGTCCGCCGCGGCC
TGTGGACCCCGCACCAGCTGGAACTCGCCTGCGGTGATGGCGGCGCCACCAC
CGCCGGCGGGCTGCGCGGCATGCTGCTCGGCGTGGCCCGCGAGGAGATGATC
CACTTCCTGCTGGTCAACAACATCATCATGGCGATGGGGGAGCCGTTCCACG
TGCCGGTGGTGGACTTCGGCACCGTCAACACCACCCTGCCGGTGCCGCTGGA
CGTCAGTCTGGAGGCCCTCAACCTCGGCAGCGTCCAACGGTTCATCGCCATC
GAGCGGCCGGACTGCGAGGTCGGCGAGCTGCGCCGGGCGCCGGGCACCGGA
GAACCGCTCGCCCACCGACCCGGCCGGTCGCCGCGCTGCTACGGCACCGTCA
GCGAGCTGTACGCCGAGACGTCCGGGAGGGCCT

NMCC_89:
GCGTACTCGGTGCCGGCCTACGGGGCGGGGGAGGAGTACGTCCGGCGCGGG
CTGTGGACCCCCGAGCAGCTGCGGCTCGCCTGTGGGGACGGCGGCCGGACCC
GCGACGGGGGCATCCGCGGCACGCTGCTCGGCATCGCCCGCGAGGAGATGAT
CCACTTCCTGATCGTCAACAACATCATCATGGCGATGGGCGAACCGTTCCAC
GTCCCCGACGTCGACTTCGGCACGATCAACAACACCCTGCCCGTGCCGCTGG
ACTTCGCCCTGGAGCCCCTCGGTGTGGGCAGCGTGCAGCGGTTCATCGCGAT
CGAACGGCCGGAGGACCAGATCGGCGAGCTCCAGCCGCCGTCCCGGTACGG
CTCCCCGCGGCCGCCGAGCACCAGTACGCCTCGCTCAGCGAGCTGTACGGCG
ACATCCCGCGAGGGCCGCCAACCGGGTCCCCG

NMCC_22:
GGCTATTCGGTCCCCACGCACGGCACGGGGTTCGCGTACGTGCGCAGCGGGG
TGTGGACCCCGCGTCAGCTCCGGCTCGCGTGCGGGGACGGCGGGGAGACCCT
CGCCAAGGGGGTCAGGGACAGCCTCTTCGACGTGGCCCGCGAGGAGATGAT

GCACTTCCTCGTCGTCAACAACATCCTGATGGCCATGGGCGAGCCCTTCCACG
TCCCGGAGATCGACTTCGGCACGCCGGGCCGGCTGCCGCTGCCGCTCGACTT
CGCCCTCGAACCGCTGCACCTGGGAAGCCTCCAGCGGTTCATCGCCATCGAG
CGGCCCGAGCGCCTCGCGGGCGGCACGGGAGCCGTGGACGGGCCCGGCCCG
TTCGGTTCGCCGAGCGAGCTGTACGCGGGGATCAGGGAGGGCCTGACCCGGG
TCCCCGACCCTGTTCCCCGTGGACCGGGGGCGG

NMCC_204:
CTGTGGTCCATCCCCACCCACTCGGCCGGGATCGAGTACGTGCGGCGACGCG
AGTGGACGCCCGGGCAGCTCCGGCTCATGTGCGGCGAGGGCCCGCACAGCCT
CGACGGCGGCGTCCGGCAGGACCTGTTCGCCGTCGCCCGCGAGGAAATGATC
CACTTCCTGCTGATAAACAACATCATCATGGCGACGGGTCAGCCGTTCCACCT
GCCCCGGATCGACTTCGGCACAGTGAACGGCGAGCTCCCGTCCCATCGACCT
GTGCCTGGAGCCCTTCGGGCGGGCAGCCTGCAACGGTTCGCCGCGCTGGAAC
GGCCCTACGATCTGGTCCGCGACCTCGCCGCCGACCGGCCACCGGTGACCGC
GTGCGACCCGTATCCCTACGGTTCGCTCAGCGAGCTGTACAAGGCCATCCGC
CAGGCGATCCAGGGACATCCCGGACGT

NMCC_278:
GCCTACTCGGTACCCACCCACGGCGCCGGCGCGGAGTACGTACGCCGGGGCC
TGTGGACGCCCGACCAACTGCGGCTCGCGTGCGGTGACGGCGGGGAGACCCT
CGACGAGGGCATCCGCAGCATGCTGCTGACCATCGCCCGCGAGGAGATGATC
CACTTCCTCCTCGTCAACAACATCCTCATGGCGGTGGGCGAACCCTTCCACGC
GCCCCGGATCGACTTCGGCACCGTCAACCGACAGCTGGCCGTCCCGCTGGAC
TTCGCCCTGGAGCGCCTGGGGCCCGGCAGCGTGGAGCGGTTCGTACAGATCG
AACGCCCCGAGGACCTCGTCGACGAGGTACGGCGCGGCGACGCTCCGGCGCC
CCCGCCGCGTACGACGAGCGGCACCCGTACGCCTCGCTGAGCGAGCTGTACG
CGGACAGTCCCGGGGAAGGGC

NMCC_44:

GCCTGGTCGATCCCGACGGCCGGCGCCGCCGCCGAGTTCGTCCGCCGCGGCG
AATGGACGCCGGAGCAGCTGCGGCTGGCGTGCGGCGACGGCGGACCCACCC
TCGACTACGGGATGCGCGGCACGCTGCTCAACGTGGCCCGCGAGGAAATGAT
CCACTTCCTGGTCATCAACAACATCATCACCGCGACCGGCGACGCGTTCCAC
GTGCCGGCGATCGACTTCGGCACCCTCAACGAGCAGCTGCCGGTGCCGCTGG
ACTTCAGCCTGGAGGCGTTCGGGCTGGGGCCGCTGCAACGCTTCATCGCCAT
CGAACGGCCGGACGACCAGACCGTCGAGTTCGCCGGGACCGACACGCTGCTC
GACCGGGGCGACGCGCTGTACCCGTACGGCTCGCTCAGCGAGCTCTACGCGG
CCATCCCGCGAGGGGCCATTCCCAGCCGGGGT

NMCC_268:
GCGTACTCGGTGCCGACGTACGGGGCGGCCGAGGACTTCGTCCACCGAGGTC
TGTGGACCCCCGAGCAGCTGCGGGTCACCGTGGGCGACGGCGGAGAGGCCCT
CGACGGCGGCGTCCGGGGAAAGCTGGTGGAGATCGCGCGCGAGGAGATGAT
CCATTTCCTCCTCGTCAACAACATCCTCATGGCGCTGGGCGAGCCCTTCTGCG
TGCCCGCACTGGACTTCGGCGCGCTCGGCTCCGACCTGCCGGTGCCCCTCGAC
CTCTGCCTCGAAGGGCTCGACATCGGCAGCGTGGCGCGGTTCATCGCGATCG
AGCAGCCCGCGGTGGGCACGCCGGAGGTGCGGCGACCCGGACCTGGCCCAC
CGGCACCGACGCGGCGAGCGCCGGCCGGTACGAGACGTTGAGCGAGATGTA
CGCGCGGATCCGGCCAGGGCTTGGGGG

NMCC_104:
GCGTACTCCATCCCCACCCACGGGGCGGGCAGCCGGTACGTCGCGTCGGGGG
AGTGGAGCGAGGACCAGTTCCGGCTCGCCTGCGGTGACGGGGCCCAGACGTT
GAGCAACGGCATCCGCGGCAGCCTGCTGAACGTGGCACGCGAGGAGATGAT
GCACTTCCTCGCCATCAACAACATCATCATGGCCGCGGGCAGGCCGTTCTGC
CTGCCGCGCCTGGACTTCGGCGAGATCAACAACCAGATCCCGGTGCCGATCG
AGTTCGCGCTCGAACCGTTCGGGCTGGGCAGCGTGCAGCGGTTCGTCGAGCT
GGAGAAGCCGCACGACCTGATCATGGACGTGGCCGGGACCGAACCAGCCGG
CAAGCGGGAGAACGGCGAGCCCTACCGGTACGGTCGCTCAGCGAGCTGTAC
GAGGCCATCCGCGAGGGCGTCCGAGCGGGTGCCGG

NMCC_81:

GCCTACTCGGTGCCGACCTACGGCGCCGGGCTGGCACTCGTGCGGCGCGGCC
TGTGGACGCCGGAGCAGCTCGACCTCGCCTGCGGCGACGGCGGGCGGACGC
GGCACACCGGGGTTCGCGGCGCCCTGCTCGGCGTCGCGCGCGAGGAGATGAT
CCACTTCCTGGTCGTCAACAACATCCTGATGGCCATCGGCGAGCCGTTCTTCA
GCCCGGCCGTCGACTTCGGCACCGTCAACACCGAACTGCCGGTGCCGGTGGA
CTTCGCGCTCGAACCGCTGAACGTCGGCAGCGTGCAGCGGTTCGTCGCCCTG
GAGCGGCCCGCCGACCCGGAGCAGGTGGCCGGCCCCTACCGGTCGCTCAGCG
AGCTGTACGGCGCCATCCGCGAAGGCTTGCGGCGGTACCCGACCTGTTCCTG
GTGCGGCCGGGCCGGGGCGGCGGCGAGCACCGAC

NMCC_5:

GCCTACTCGATACCGACGTACGGGTCGGGGCTGGCGTACGTCCGGCGCGGGC
TGTGGACGCCCGAGCAGCTGGCACTCGCCTGCGGCGACGGCGGTGCGACGCT
GGCCAACGGGATGCGGGGCAGCCTGCTGTCCGTCGCCCGCGAGGAGATGATC
CATTTCCTGGTCATCAACAACGTCATCATGGCGATGGGCGAGGCGTTCACCG
TCCCGGCCATCGACTTCGGCACCGTCAACGGCTCGTTGCCGGTCCCGGTGGA
CTTCGCCCTGGAACGTTTCGGGGTCGGTAGCGTGCAACGGTTCATCGCGATC
GAACAGCCCGAGTCCCTCGTCGGCGAGGTACCCCGGGAGTTCGGCAGCGGGT
CCGGTACGTCAGAGCCGACACCCACCTACCGGTCCCTCAGCGAGCTGTACGG
CGCGATCCGGGAGGGCTGCAGCGGGTGCCCGACT

References

1. Williams DH, Stone MJ, Hauck PR, & Rahman SK (1989) Why are secondary metabolites (natural products) biosynthesized? *J Nat Prod* 52(6):1189-1208.
2. Demain AL & Fang A (2000) The natural functions of secondary metabolites. *Adv Biochem Eng Biotechnol* 69:1-39.
3. Bassler BL (1999) How bacteria talk to each other: regulation of gene expression by quorum sensing. *Curr Opin Microbiol* 2(6):582-587.
4. Aminov RI (2009) The role of antibiotics and antibiotic resistance in nature. *Environ Microbiol* 11(12):2970-2988.
5. Pietra F (2002) *Biodiversity and Natural Product Diversity* (Elsevier).
6. Newman DJ & Cragg GM (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod* 75(3):311-335.
7. Zhang L & Demain AL (2007) *Natural Products: Drug Discovery and Therapeutic Medicine* (Springer).
8. Liu J, *et al*. (1991) Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes. *Cell* 66(4):807-815.
9. Fenteany G, *et al*. (1995) Inhibition of proteasome activities and subunit-specific amino-terminal threonine modification by lactacystin. *Science* 268(5211):726-731.
10. Buskirk AR & Liu DR (2005) Creating small-molecule-dependent switches to modulate biological functions. *Chem Biol* 12(2):151-161.
11. Koehn FE (2013) *Natural products and cancer drug discovery* (Humana Press).

12. Leclercq R (2009) Epidemiological and resistance issues in multidrug-resistant staphylococci and enterococci. *Clin Microbiol Infect* 15(3):224-231.

13. Siegel R, Ma J, Zou Z, & Jemal A (2014) Cancer statistics, 2014. *CA Cancer J Clin* 64(1):9-29.

14. Torsvik V, Ovreas L, & Thingstad TF (2002) Prokaryotic diversity--magnitude, dynamics, and controlling factors. *Science* 296(5570):1064-1066.

15. Tulp M & Bohlin L (2005) Rediscovery of known natural compounds: nuisance or goldmine? *Bioorg Med Chem* 13(17):5274-5282.

16. Torsvik V, Goksoyr J, & Daae FL (1990) High diversity in DNA of soil bacteria. *Appl Environ Microbiol* 56(3):782-787.

17. Rappe MS & Giovannoni SJ (2003) The uncultured microbial majority. *Annu Rev Microbiol* 57:369-394.

18. Achtman M & Wagner M (2008) Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol* 6(6):431-440.

19. Hugenholtz P, Goebel BM, & Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* 180(18):4765-4774.

20. 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(10):R245-249.

21. Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* 68(4):669-685.

22. Henne A, Daniel R, Schmitz RA, & Gottschalk G (1999) Construction of environmental DNA libraries in Escherichia coli and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *Appl Environ Microbiol* 65(9):3901-3907.

23. Fischbach MA, Walsh CT, & Clardy J (2008) The evolution of gene collectives: How natural selection drives chemical innovation. *Proc Natl Acad Sci U S A* 105(12):4601-4608.

24. Iqbal HA, Feng Z, & Brady SF (2012) Biocatalysts and small molecule products from metagenomic studies. *Curr Opin Chem Biol* 16(1-2):109-116.

25. 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(51):12903-12904.

26. Gillespie DE*, et al*. (2002) Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Applied and Environmental Microbiology* 68(9):4301-4306.

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

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

29. Gabor EM, Alkema WB, & Janssen DB (2004) Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol* 6(9):879-886.

30. 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(5):1633-1641.

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

32. Bagg A & Neilands JB (1987) Molecular mechanism of regulation of siderophore-mediated iron assimilation. *Microbiol Rev* 51(4):509-518.

33. Bentley SD, *et al*. (2002) Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* 417(6885):141-147.

34. Wilkinson B & Micklefield J (2007) Mining and engineering natural-product biosynthetic pathways. *Nat Chem Biol* 3(7):379-386.

35. Zerikly M & Challis GL (2009) Strategies for the discovery of new natural products by genome mining. *Chembiochem* 10(4):625-633.

36. Venter JC, *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304(5667):66-74.

37. Roesch LF, *et al*. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* 1(4):283-290.

38. Alberts B, Wilson JH, & Hunt T (2008) *Molecular biology of the cell* (Garland Science, New York) 5th Ed pp xxxiii, 1601, 1690 p.

39. Cane DE, Walsh CT, & Khosla C (1998) Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science* 282(5386):63-68.

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

41. 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. *J Am Chem Soc* 132(44):15661-15670.

42. 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(31):12629-12634.

43. Owen JG, *et al*. (2013) Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proceedings of the National Academy of Sciences of the United States of America* 110(29):11797-11802.

44. Arnison PG, *et al*. (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep* 30(1):108-160.

45. Ryan KS & Drennan CL (2009) Divergent pathways in the biosynthesis of bisindole natural products. *Chem Biol* 16(4):351-364.

46. Sanchez C, Mendez C, & Salas JA (2006) Indolocarbazole natural products: occurrence, biosynthesis, and biological activity. *Nat Prod Rep* 23(6):1007-1045.

47. Nakano H & Omura S (2009) Chemical biology of natural indolocarbazole products: 30 years since the discovery of staurosporine. *J Antibiot (Tokyo)* 62(1):17-26.

48. Prade L, *et al*. (1997) Staurosporine-induced conformational changes of cAMP-dependent protein kinase catalytic subunit explain inhibitory potential. *Structure* 5(12):1627-1637.

49. Staker BL, *et al*. (2005) Structures of three classes of anticancer agents bound to the human topoisomerase I-DNA covalent complex. *J Med Chem* 48(7):2336-2345.

50. Butler MS (2005) Natural products to drugs: natural product derived compounds in clinical trials. *Nat Prod Rep* 22(2):162-195.

51. Saif MW & Diasio RB (2005) Edotecarin: a novel topoisomerase I inhibitor. *Clin Colorectal Cancer* 5(1):27-36.

52. Schwandt A, *et al*. (2012) Phase-II trial of rebeccamycin analog, a dual topoisomerase-I and -II inhibitor, in relapsed "sensitive" small cell lung cancer. *J Thorac Oncol* 7(4):751-754.

53. Sausville EA, *et al*. (2001) Phase I trial of 72-hour continuous infusion UCN-01 in patients with refractory neoplasms. *J Clin Oncol* 19(8):2319-2333.

54. Fischer T, *et al*. (2010) Phase IIB trial of oral Midostaurin (PKC412), the FMS-like tyrosine kinase 3 receptor (FLT3) and multi-targeted kinase inhibitor, in patients with acute myeloid leukemia and high-risk myelodysplastic syndrome with either wild-type or mutated FLT3. *J Clin Oncol* 28(28):4339-4345.

55. Knapper S, *et al*. (2006) A phase 2 trial of the FLT3 inhibitor lestaurtinib (CEP701) as first-line treatment for older patients with acute myeloid leukemia not considered fit for intensive chemotherapy. *Blood* 108(10):3262-3270.

56. Onaka H, Taniguchi S, Igarashi Y, & Furumai T (2002) Cloning of the staurosporine biosynthetic gene cluster from Streptomyces sp. TP-A0274 and its heterologous expression in Streptomyces lividans. *Journal of Antibiotics* 55(12):1063-1071.

57. Sanchez C, *et al*. (2002) The biosynthetic gene cluster for the antitumor rebeccamycin: characterization and generation of indolocarbazole derivatives. *Chem Biol* 9(4):519-531.

58. Chiu HT, *et al*. (2009) Molecular cloning, sequence analysis and functional characterization of the gene cluster for biosynthesis of K-252a and its analogs. *Mol Biosyst* 5(10):1180-1191.

59. Kim SY, *et al*. (2007) Genetic organization of the biosynthetic gene cluster for the indolocarbazole K-252a in Nonomuraea longicatena JCM 11136. *Appl Microbiol Biotechnol* 75(5):1119-1126.

60. Gao Q, Zhang C, Blanchard S, & Thorson JS (2006) Deciphering indolocarbazole and enediyne aminodideoxypentose biosynthesis through comparative genomics: insights from the AT2433 biosynthetic locus. *Chem Biol* 13(7):733-743.

61. Pemberton JM, Vincent KM, & Penfold RJ (1991) Cloning and Heterologous Expression of the Violacein Biosynthesis Gene-Cluster from Chromobacterium-Violaceum. *Current Microbiology* 22(6):355-358.

62. Polz MF, Alm EJ, & Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet* 29(3):170-175.

63. Stone MJ & Williams DH (1992) On the evolution of functional secondary metabolites (natural products). *Mol Microbiol* 6(1):29-34.

64. Jenke-Kodama H, Muller R, & Dittmann E (2008) Evolutionary mechanisms underlying secondary metabolite diversity. *Prog Drug Res* 65:119, 121-140.

65. Fischbach MA & Walsh CT (2006) Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chem Rev* 106(8):3468-3496.

66. Amoutzias GD, Van de Peer Y, & Mossialos D (2008) Evolution and taxonomic distribution of nonribosomal peptide and polyketide synthases. *Future Microbiol* 3(3):361-370.

67. Jenke-Kodama H, Sandmann A, Muller R, & Dittmann E (2005) Evolutionary implications of bacterial polyketide synthases. *Mol Biol Evol* 22(10):2027-2039.

68. Speck K & Magauer T (2013) The chemistry of isoindole natural products. *Beilstein J Org Chem* 9:2048-2078.

69. Bergman J, Janosik T, & Wahlström N (2001) *Indolocarbazoles* (Academic Press).

70. Onaka H (2009) Biosynthesis of indolocarbazole and goadsporin, two different heterocyclic antibiotics produced by actinomycetes. *Biosci Biotechnol Biochem* 73(10):2149-2155.

71. Hoshino T (2011) Violacein and related tryptophan metabolites produced by Chromobacterium violaceum: biosynthetic mechanism and pathway for construction of violacein core. *Appl Microbiol Biotechnol* 91(6):1463-1475.

72. Khayatt BI, Overmars L, Siezen RJ, & Francke C (2013) Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PLoS One* 8(4):e62136.

73. Rausch C, Hoof I, Weber T, Wohlleben W, & Huson DH (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol Biol* 7:78.

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

75. Gurgui C & Piel J (2010) Metagenomic approaches to identify and isolate bioactive natural products from microbiota of marine sponges. *Methods Mol Biol* 668:247-264.

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

77. Sommer MO, Dantas G, & Church GM (2009) Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* 325(5944):1128-1131.

78. Hildebrand M, *et al*. (2004) bryA: an unusual modular polyketide synthase gene from the uncultivated bacterial symbiont of the marine bryozoan Bugula neritina. *Chem Biol* 11(11):1543-1552.

79. Whitman WB, Coleman DC, & Wiebe WJ (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* 95(12):6578-6583.

80. Page RD (2012) Space, time, form: viewing the Tree of Life. *Trends Ecol Evol* 27(2):113-120.

81. Fierer N & Jackson RB (2006) The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci U S A* 103(3):626-631.

82. Kang HS & Brady SF (2013) Arimetamycin A: improving clinically relevant families of natural products through sequence-guided screening of soil metagenomes. *Angew Chem Int Ed Engl* 52(42):11063-11067.

83. Woese CR & Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* 74(11):5088-5090.

84. Chang FY, Ternei MA, Calle PY, & Brady SF (2013) Discovery and synthetic refactoring of tryptophan dimer gene clusters from the environment. *J Am Chem Soc* 135(47):17906-17912.

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

86. Ouyang Y, *et al*. (2010) Isolation of high molecular weight DNA from marine sponge bacteria for BAC library construction. *Mar Biotechnol (NY)* 12(3):318-325.

87. Duran N, Antonio RV, Haun M, & Pilli RA (1994) Biosynthesis of a trypanocide by Chromobacterium violaceum. *World J Microbiol Biotechnol* 10(6):686-690.

88. 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(45):17273-17277.

89. Balibar CJ & Walsh CT (2006) In vitro biosynthesis of violacein from L-tryptophan by the enzymes VioA-E from Chromobacterium violaceum. *Biochemistry* 45(51):15444-15457.

90. Ferreira CV, *et al*. (2004) Molecular mechanism of violacein-mediated human leukemia cell death. *Blood* 104(5):1459-1464.

91. McClean KH, *et al*. (1997) Quorum sensing and Chromobacterium violaceum: exploitation of violacein production and inhibition for the detection of N-acylhomoserine lactones. *Microbiology* 143 (Pt 12):3703-3711.

92. Stauff DL & Bassler BL (2011) Quorum sensing in Chromobacterium violaceum: DNA recognition and gene regulation by the CviR receptor. *J Bacteriol* 193(15):3871-3878.

93. Duran N & Menck CF (2001) Chromobacterium violaceum: a review of pharmacological and industiral perspectives. *Crit Rev Microbiol* 27(3):201-222.

94. Sanchez C, Brana AF, Mendez C, & Salas JA (2006) Reevaluation of the violacein biosynthetic pathway and its relationship to indolocarbazole biosynthesis. *Chembiochem* 7(8):1231-1240.

95. Kang HS & Brady SF (2014) Arixanthomycins A-C: Phylogeny-Guided Discovery of Biologically Active eDNA-Derived Pentangular Polyphenols. *ACS Chem Biol*.

96. Groom K, Bhattacharya A, & Zechel DL (2011) Rebeccamycin and staurosporine biosynthesis: insight into the mechanisms of the flavin-dependent monooxygenases RebC and StaC. *Chembiochem* 12(3):396-400.

97. Goldman PJ, *et al*. (2012) An unusual role for a mobile flavin in StaC-like indolocarbazole biosynthetic enzymes. *Chem Biol* 19(7):855-865.

98. Ongley SE, Bian X, Neilan BA, & Muller R (2013) Recent advances in the heterologous expression of microbial natural product biosynthetic pathways. *Nat Prod Rep* 30(8):1121-1138.

99. Zhang H, Boghigian BA, Armando J, & Pfeifer BA (2011) Methods and options for the heterologous production of complex natural products. *Nat Prod Rep* 28(1):125-151.

100. Fu J, *et al*. (2012) Full-length RecE enhances linear-linear homologous recombination and facilitates direct cloning for bioprospecting. *Nat Biotechnol* 30(5):440-446.

101. Li L, *et al*. (2008) Characterization of the saframycin A gene cluster from Streptomyces lavendulae NRRL 11002 revealing a nonribosomal peptide synthetase system for assembling the unusual tetrapeptidyl skeleton in an iterative manner. *J Bacteriol* 190(1):251-263.

102. McDaniel R, *et al*. (1999) Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel "unnatural" natural products. *Proc Natl Acad Sci U S A* 96(5):1846-1851.

103. Ongley SE, *et al*. (2013) High-titer heterologous production in E. coli of lyngbyatoxin, a protein kinase C activator from an uncultured marine cyanobacterium. *ACS Chem Biol* 8(9):1888-1893.

104. Wenzel SC & Muller R (2005) Recent developments towards the heterologous expression of complex bacterial natural product biosynthetic pathways. *Curr Opin Biotechnol* 16(6):594-606.

105. Eppelmann K, Doekel S, & Marahiel MA (2001) Engineered biosynthesis of the peptide antibiotic bacitracin in the surrogate host Bacillus subtilis. *J Biol Chem* 276(37):34824-34831.

106. Gustafsson C, Govindarajan S, & Minshull J (2004) Codon bias and heterologous protein expression. *Trends Biotechnol* 22(7):346-353.

107. Mutka SC, Carney JR, Liu Y, & Kennedy J (2006) Heterologous production of epothilone C and D in Escherichia coli. *Biochemistry* 45(4):1321-1330.

108. Fierer N, Bradford MA, & Jackson RB (2007) Toward an ecological classification of soil bacteria. *Ecology* 88(6):1354-1364.

109. Craig JW, Chang FY, & Brady SF (2009) Natural products from environmental DNA hosted in Ralstonia metallidurans. *ACS Chem Biol* 4(1):23-28.

110. Biggins JB, Liu X, Feng Z, & Brady SF (2011) Metabolites from the induced expression of cryptic single operons found in the genome of Burkholderia pseudomallei. *J Am Chem Soc* 133(6):1638-1641.

111. Baltz RH (2010) Streptomyces and Saccharopolyspora hosts for heterologous expression of secondary metabolite gene clusters. *J Ind Microbiol Biotechnol* 37(8):759-772.

112. 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(6):2646-2651.

113. Salas AP*, et al*. (2005) Deciphering the late steps in the biosynthesis of the anti-tumour indolocarbazole staurosporine: sugar donor substrate flexibility of the StaG glycosyltransferase. *Mol Microbiol* 58(1):17-27.

114. Sanchez C*, et al*. (2005) Combinatorial biosynthesis of antitumor indolocarbazole compounds. *Proc Natl Acad Sci U S A* 102(2):461-466.

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

116. Biggins JB, Gleber CD, & Brady SF (2011) Acyldepsipeptide HDAC inhibitor production induced in Burkholderia thailandensis. *Org Lett* 13(6):1536-1539.

117. Holden MT*, et al*. (1998) Cryptic carbapenem antibiotic production genes are widespread in Erwinia carotovora: facile trans activation by the carR transcriptional regulator. *Microbiology* 144 (Pt 6):1495-1508.

118. Medema MH, Breitling R, & Takano E (2011) Synthetic biology in Streptomyces bacteria. *Methods Enzymol* 497:485-502.

119. Hertweck C (2009) Hidden biosynthetic treasures brought to light. *Nat Chem Biol* 5(7):450-452.

120. Nett M, Ikeda H, & Moore BS (2009) Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat Prod Rep* 26(11):1362-1384.

121. Luo Y, *et al*. (2013) Activation and characterization of a cryptic polycyclic tetramate macrolactam biosynthetic gene cluster. *Nat Commun* 4:2894.

122. Seyedsayamdost MR (2014) High-throughput platform for the discovery of elicitors of silent bacterial gene clusters. *Proc Natl Acad Sci U S A* 111(20):7266-7271.

123. Fitzgerald JT, Charkoudian LK, Watts KR, & Khosla C (2013) Analysis and refactoring of the A-74528 biosynthetic pathway. *J Am Chem Soc* 135(10):3752-3755.

124. Shao Z, *et al*. (2013) Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold. *ACS Synth Biol* 2(11):662-669.

125. Lim JH, Seo SW, Kim SY, & Jung GY (2013) Refactoring redox cofactor regeneration for high-yield biocatalysis of glucose to butyric acid in Escherichia coli. *Bioresour Technol* 135:568-573.

126. Temme K, Zhao D, & Voigt CA (2012) Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. *Proc Natl Acad Sci U S A* 109(18):7085-7090.

127. Chang FY & Brady SF (2011) Cloning and characterization of an environmental DNA-derived gene cluster that encodes the biosynthesis of the antitumor substance BE-54017. *J Am Chem Soc* 133(26):9996-9999.

128. Chang FY & Brady SF (2013) Discovery of indolotryptoline antiproliferative agents by homology-guided metagenomic screening. *Proc Natl Acad Sci U S A* 110(7):2478-2483.

129. Richards SA & Hollerton JC (2011) *Essential practical NMR for organic chemistry* (John Wiley, Chichester, West Sussex, U.K.) pp x, 216 p.

130. Nakase K*, et al*. (JP 2000178274, 2000) JP 2000178274.

131. Williams DE*, et al*. (2008) Cladoniamides A-G, tryptophan-derived alkaloids produced in culture by Streptomyces uncialis. *Org Lett* 10(16):3501-3504.

132. van Berkel WJ, Kamerbeek NM, & Fraaije MW (2006) Flavoprotein monooxygenases, a diverse class of oxidative biocatalysts. *J Biotechnol* 124(4):670-689.

133. Widersten M, Gurell A, & Lindberg D (2010) Structure-function relationships of epoxide hydrolases and their potential use in biocatalysis. *Biochim Biophys Acta* 1800(3):316-326.

134. Howard-Jones AR & Walsh CT (2006) Staurosporine and rebeccamycin aglycones are assembled by the oxidative action of StaP, StaC, and RebC on chromopyrrolic acid. *J Am Chem Soc* 128(37):12289-12298.

135. Salas JA & Mendez C (2009) Indolocarbazole antitumour compounds by combinatorial biosynthesis. *Curr Opin Chem Biol* 13(2):152-160.

136. Ryan KS (2011) Biosynthetic gene cluster for the cladoniamides, bis-indoles with a rearranged scaffold. *PLoS One* 6(8):e23694.

137. Sanchez C*, et al*. (2005) Combinatorial biosynthesis of antitumor indolocarbazole compounds. *Proc Natl Acad Sci U S A* 102(2):461-466.

138. Hopwood DA (1997) Genetic Contributions to Understanding Polyketide Synthases. *Chem Rev* 97(7):2465-2498.

139. Schwarzer D, Finking R, & Marahiel MA (2003) Nonribosomal peptides: from genes to products. *Nat Prod Rep* 20(3):275-287.

140. Xue Y, Zhao L, Liu HW, & Sherman DH (1998) A gene cluster for macrolide antibiotic biosynthesis in Streptomyces venezuelae: architecture of metabolic diversity. *Proc Natl Acad Sci U S A* 95(21):12111-12116.

141. Wilson MC, Gulder TA, Mahmud T, & Moore BS (2010) Shared biosynthesis of the saliniketals and rifamycins in Salinispora arenicola is controlled by the sare1259-encoded cytochrome P450. *J Am Chem Soc* 132(36):12757-12765.

142. Bush JA, Long BH, Catino JJ, Bradner WT, & Tomita K (1987) Production and biological activity of rebeccamycin, a novel antitumor agent. *J Antibiot (Tokyo)* 40(5):668-678.

143. Trujillo JI, *et al*. (2007) Novel tetrahydro-beta-carboline-1-carboxylic acids as inhibitors of mitogen activated protein kinase-activated protein kinase 2 (MK-2). *Bioorg Med Chem Lett* 17(16):4657-4663.

144. Soni R, *et al*. (2000) Inhibition of cyclin-dependent kinase 4 (Cdk4) by fascaplysin, a marine natural product. *Biochem Biophys Res Commun* 275(3):877-884.

145. Backs J, *et al*. (2009) The delta isoform of CaM kinase II is required for pathological cardiac hypertrophy and remodeling after pressure overload. *Proc Natl Acad Sci U S A* 106(7):2342-2347.

146. Rokhlin OW, *et al*. (2007) Calcium/calmodulin-dependent kinase II plays an important role in prostate cancer cell survival. *Cancer Biol Ther* 6(5):732-742.

147. Hormann A, Chaudhuri B, & Fretz H (2001) DNA binding properties of the marine sponge pigment fascaplysin. *Bioorg Med Chem* 9(4):917-921.

148. Aubry C*, et al*. (2004) New fascaplysin-based CDK4-specific inhibitors: design, synthesis and biological activity. *Chem Commun (Camb)* (15):1696-1697.

149. Du YL, Ding T, & Ryan KS (2013) Biosynthetic O-methylation protects cladoniamides from self-destruction. *Org Lett* 15(10):2538-2541.

150. Du YL, Ding T, Patrick BO, & Ryan KS (2013) Xenocladoniamide F, minimal indolotryptoline from the cladoniamide pathway. *Tetrahedron Letters* 54(41):5635-5638.

151. Howard-Jones AR & Walsh CT (2007) Nonenzymatic oxidative steps accompanying action of the cytochrome P450 enzymes StaP and RebP in the biosynthesis of staurosporine and rebeccamycin. *J Am Chem Soc* 129(36):11016-11017.

152. Asamizu S*, et al*. (2012) Coupling reaction of indolepyruvic acid by StaD and its product: implications for biosynthesis of indolocarbazole and violacein. *Chembiochem* 13(17):2495-2500.

153. Chang FY & Brady SF (2014) Characterization of an environmental DNA-derived gene cluster that encodes the bisindolylmaleimide methylarcyriarubin. *Chembiochem* 15(6):815-821.

154. Howard-Jones AR & Walsh CT (2005) Enzymatic generation of the chromopyrrolic acid scaffold of rebeccamycin by the tandem action of RebO and RebD. *Biochemistry* 44(48):15652-15663.

155. 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(13):1981-1984.

156. Brenner M, Rexhausen H, Steffan B, & Steglich W (1988) Synthesis of arcyriarubin b and related bisindolylmaleimides. *Tetrahedron* 44(10):2887-2892.

157. Wang K & Liu Z (2009) Synthesis of Arcyriarubin A and Arcyriaflavin A via Cross-Coupling of Indolylboronic Acid with Dibromomaleimides. *Synthetic Communications* 40(1):144-150.

158. Asamizu S, Shiro Y, Igarashi Y, Nagano S, & Onaka H (2011) Characterization and functional modification of StaC and RebC, which are involved in the pyrrole oxidation of indolocarbazole biosynthesis. *Biosci Biotechnol Biochem* 75(11):2184-2193.

159. Ryan KS, *et al*. (2007) Crystallographic trapping in the rebeccamycin biosynthetic enzyme RebC. *Proc Natl Acad Sci U S A* 104(39):15311-15316.

160. Ballou DP (2007) Crystallography gets the jump on the enzymologists. *Proc Natl Acad Sci U S A* 104(40):15587-15588.

161. Tsai SC (2012) Babysitting flavin for biosynthesis. *Chem Biol* 19(7):787-788.

162. Barry SM & Challis GL (2013) Mechanism and Catalytic Diversity of Rieske Non-Heme Iron-Dependent Oxygenases. *ACS Catal* 3(10).

163. Hirano S, Asamizu S, Onaka H, Shiro Y, & Nagano S (2008) Crystal structure of VioE, a key player in the construction of the molecular skeleton of violacein. *J Biol Chem* 283(10):6459-6466.

164. Ryan KS, Balibar CJ, Turo KE, Walsh CT, & Drennan CL (2008) The violacein biosynthetic enzyme VioE shares a fold with lipoprotein transporter proteins. *J Biol Chem* 283(10):6467-6475.

165. Gill M & Steglich W (1987) Pigments of fungi (Macromycetes). *Fortschr Chem Org Naturst* 51:1-317.

166. Steglich W, Steffan B, Kopanski L, & Eckhardt G (1980) Fungal Pigments .36. Indole Pigments from the Fruiting Bodies of Slime-Mold Arcyria-Denudata. *Angewandte Chemie. International Edition in English* 19(6):459-460.

167. Zhang J, Yang PL, & Gray NS (2009) Targeting cancer with small molecule kinase inhibitors. *Nat Rev Cancer* 9(1):28-39.

168. Driggers EM, Hale SP, Lee J, & Terrett NK (2008) The exploration of macrocycles for drug discovery--an underexploited structural class. *Nat Rev Drug Discov* 7(7):608-624.

169. Pajak B, Orzechowska S, Gajkowska B, & Orzechowski A (2008) Bisindolylmaleimides in anti-cancer therapy - more than PKC inhibitors. *Adv Med Sci* 53(1):21-31.

170. Anderson PW, McGill JB, & Tuttle KR (2007) Protein kinase C beta inhibition: the promise for treatment of diabetic nephropathy. *Curr Opin Nephrol Hypertens* 16(5):397-402.

171. Katare RG, Zhitian Z, Sodeoka M, & Sasaguri S (2007) Novel bisindolylmaleimide derivative inhibits mitochondrial permeability transition pore and protects the heart from reperfusion injury. *Can J Physiol Pharmacol* 85(10):979-985.

172. Asakai R, Aoyama Y, & Fujimoto T (2002) Bisindolylmaleimide I and V inhibit necrosis induced by oxidative stress in a variety of cells including neurons. *Neurosci Res* 44(3):297-304.

173. Aiello LP, *et al*. (2006) Effect of ruboxistaurin on visual loss in patients with diabetic retinopathy. *Ophthalmology* 113(12):2221-2230.

174. Ma S & Rosen ST (2007) Enzastaurin. *Curr Opin Oncol* 19(6):590-595.

175. Zagouri F, Sergentanis TN, Chrysikos D, Filipits M, & Bartsch R (2012) mTOR inhibitors in breast cancer: a systematic review. *Gynecol Oncol* 127(3):662-672.

176. Teicher BA (2006) Protein kinase C as a therapeutic target. *Clin Cancer Res* 12(18):5336-5345.

177. Brehmer D, Godl K, Zech B, Wissing J, & Daub H (2004) Proteome-wide identification of cellular targets affected by bisindolylmaleimide-type protein kinase C inhibitors. *Mol Cell Proteomics* 3(5):490-500.

178. Davis PD, *et al*. (1992) Inhibitors of protein kinase C. 1. 2,3-Bisarylmaleimides. *J Med Chem* 35(1):177-184.

179. Marmy-Conus N, Hannan KM, & Pearson RB (2002) Ro 31-6045, the inactive analogue of the protein kinase C inhibitor Ro 31-8220, blocks in vivo activation of p70(s6k)/p85(s6k): implications for the analysis of S6K signalling. *FEBS Lett* 519(1-3):135-140.

180. Komander D, *et al*. (2004) Interactions of LY333531 and other bisindolyl maleimide inhibitors with PDK1. *Structure* 12(2):215-226.

181. Gassel M, *et al*. (2004) The protein kinase C inhibitor bisindolyl maleimide 2 binds with reversed orientations to different conformations of protein kinase A. *Journal of Biological Chemistry* 279(22):23679-23690.

182. Toullec D, *et al*. (1991) The Bisindolylmaleimide Gf-109203x Is a Potent and Selective Inhibitor of Protein-Kinase-C. *Journal of Biological Chemistry* 266(24):15771-15781.

183. Zhang C, et al. (2006) RebG- and RebM-catalyzed indolocarbazole diversification. *Chembiochem* (5):795-804.

184. Schenone M, Dancik V, Wagner BK, & Clemons PA (2013) Target identification and mechanism of action in chemical biology and drug discovery. *Nat Chem Biol* 9(4):232-240.

185. Ziegler S, Pries V, Hedberg C, & Waldmann H (2013) Target identification for small bioactive molecules: finding the needle in the haystack. *Angew Chem Int Ed Engl* 52(10):2744-2792.

186. O'Neill AJ & Chopra I (2004) Preclinical evaluation of novel antibacterial agents by microbiological and molecular techniques. *Expert Opin Investig Drugs* 13(8):1045-1063.

187. Delgado MA, Rintoul MR, Farias RN, & Salomon RA (2001) Escherichia coli RNA polymerase is the target of the cyclopeptide antibiotic microcin J25. *J Bacteriol* 183(15):4543-4550.

188. Freiberg C, *et al*. (2004) Identification and characterization of the first class of potent bacterial acetyl-CoA carboxylase inhibitors with antibacterial activity. *J Biol Chem* 279(25):26066-26073.

189. Wacker SA, Houghtaling BR, Elemento O, & Kapoor TM (2012) Using transcriptome sequencing to identify mechanisms of drug action and resistance. *Nat Chem Biol* 8(3):235-237.

190. Ong SE, *et al*. (2009) Identifying the proteins to which small-molecule probes and drugs bind in cells. *Proc Natl Acad Sci U S A* 106(12):4617-4622.

191. Rix U & Superti-Furga G (2009) Target profiling of small molecules by chemical proteomics. *Nat Chem Biol* 5(9):616-624.

192. Goffeau A, *et al*. (1996) Life with 6000 genes. *Science* 274(5287):546, 563-547.

193. Wood V, *et al*. (2002) The genome sequence of Schizosaccharomyces pombe. *Nature* 415(6874):871-880.

194. Wolfger H, Mamnun YM, & Kuchler K (2001) Fungal ABC proteins: pleiotropic drug resistance, stress response and cellular detoxification. *Res Microbiol* 152(3-4):375-389.

195. Kawashima SA, Takemoto A, Nurse P, & Kapoor TM (2012) Analyzing fission yeast multidrug resistance mechanisms to develop a genetically tractable model system for chemical biology. *Chem Biol* 19(7):893-901.

196. Cragg GM, Grothaus PG, & Newman DJ (2009) Impact of natural products on developing new anti-cancer agents. *Chem Rev* 109(7):3012-3043.

197. Boyd MR & Paull KD (1995) Some practical considerations and applications of the national cancer institute in vitro anticancer drug discovery screen. *Drug Development Research* 34(2):91-109.

198. Kimura T, *et al*. (2012) Synthesis and assignment of the absolute configuration of indenotryptoline bisindole alkaloid BE-54017. *Org Lett* 14(17):4418-4421.

199. MacDiarmid CW, Milanick MA, & Eide DJ (2002) Biochemical properties of vacuolar zinc transport systems of Saccharomyces cerevisiae. *J Biol Chem* 277(42):39187-39194.

200. Kawachi M, Kobae Y, Mimura T, & Maeshima M (2008) Deletion of a histidine-rich loop of AtMTP1, a vacuolar Zn(2+)/H(+) antiporter of Arabidopsis thaliana, stimulates the transport activity. *J Biol Chem* 283(13):8374-8383.

201. Kane PM, Yamashiro CT, & Stevens TH (1989) Biochemical characterization of the yeast vacuolar H(+)-ATPase. *J Biol Chem* 264(32):19236-19244.

202. Pongcharoen P*, et al*. (2013) Functional expression of Schizosaccharomyces pombe Vba2p in the vacuolar membrane of Saccharomyces cerevisiae. *Biosci Biotechnol Biochem* 77(9):1988-1990.

203. Iwaki T, Goa T, Tanaka N, & Takegawa K (2004) Characterization of Schizosaccharomyces pombe mutants defective in vacuolar acidification and protein sorting. *Mol Genet Genomics* 271(2):197-207.

204. Liu M, Tarsio M, Charsky CM, & Kane PM (2005) Structural and functional separation of the N- and C-terminal domains of the yeast V-ATPase subunit H. *J Biol Chem* 280(44):36978-36985.

205. Marceau F, Bawolak MT, Bouthillier J, & Morissette G (2009) Vacuolar ATPase-mediated cellular concentration and retention of quinacrine: a model for the distribution of lipophilic cationic drugs to autophagic vacuoles. *Drug Metab Dispos* 37(12):2271-2274.

206. Jefferies KC, Cipriano DJ, & Forgac M (2008) Function, structure and regulation of the vacuolar (H+)-ATPases. *Arch Biochem Biophys* 476(1):33-42.

207. Benlekbir S, Bueler SA, & Rubinstein JL (2012) Structure of the vacuolar-type ATPase from Saccharomyces cerevisiae at 11-A resolution. *Nat Struct Mol Biol* 19(12):1356-1362.

208. Bowman BJ, McCall ME, Baertsch R, & Bowman EJ (2006) A model for the proteolipid ring and bafilomycin/concanamycin-binding site in the vacuolar ATPase of Neurospora crassa. *J Biol Chem* 281(42):31885-31893.

209. Murata T, Yamato I, Kakinuma Y, Leslie AG, & Walker JE (2005) Structure of the rotor of the V-Type Na+-ATPase from Enterococcus hirae. *Science* 308(5722):654-659.

210. Bockelmann S, *et al*. (2010) Archazolid A binds to the equatorial region of the c-ring of the vacuolar H+-ATPase. *J Biol Chem* 285(49):38304-38314.

211. Huss M & Wieczorek H (2009) Inhibitors of V-ATPases: old and new players. *J Exp Biol* 212(Pt 3):341-346.

212. Perez-Sayans M, Somoza-Martin JM, Barros-Angueira F, Rey JM, & Garcia-Garcia A (2009) V-ATPase inhibitors and implication in cancer treatment. *Cancer Treat Rev* 35(8):707-713.

213. Spugnini EP, Citro G, & Fais S (2010) Proton pump inhibitors as anti vacuolar-ATPases drugs: a novel anticancer strategy. *J Exp Clin Cancer Res* 29:44.

214. Brady SF, Bauer JD, Clarke-Pearson MF, & Daniels R (2007) Natural products from isnA-containing biosynthetic gene clusters recovered from the genomes of cultured and uncultured bacteria. *J Am Chem Soc* 129(40):12102-12103.

215. 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. *J Am Chem Soc* 132(34):11902-11903.

216. Kallifidas D, Kang HS, & Brady SF (2012) Tetarimycin A, an MRSA-active antibiotic identified through induced expression of environmental DNA gene clusters. *J Am Chem Soc* 134(48):19552-19555.

217. Donia MS, Ravel J, & Schmidt EW (2008) A global assembly line for cyanobactins. *Nat Chem Biol* 4(6):341-343.

218. Fisch KM, *et al*. (2009) Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting. *Nat Chem Biol* 5(7):494-501.

219. Reddy BV, *et al*. (2012) Natural product biosynthetic gene diversity in geographically distinct soil microbiomes. *Appl Environ Microbiol* 78(10):3744-3752.

220. Edwards DJ & Gerwick WH (2004) Lyngbyatoxin biosynthesis: sequence of biosynthetic gene cluster and identification of a novel aromatic prenyltransferase. *J Am Chem Soc* 126(37):11432-11433.

221. Mao Y, Varoglu M, & Sherman DH (1999) Molecular characterization and analysis of the biosynthetic gene cluster for the antitumor antibiotic mitomycin C from Streptomyces lavendulae NRRL 2564. *Chem Biol* 6(4):251-263.

222. Kharel MK, *et al*. (2004) Isolation and characterization of the tobramycin biosynthetic gene cluster from Streptomyces tenebrarius. *FEMS Microbiol Lett* 230(2):185-190.

223. Wang HX, *et al*. (2013) PCR screening reveals considerable unexploited biosynthetic potential of ansamycins and a mysterious family of AHBA-containing natural products in actinomycetes. *J Appl Microbiol* 115(1):77-85.

224. Velasquez JE & van der Donk WA (2011) Genome mining for ribosomally synthesized natural products. *Curr Opin Chem Biol* 15(1):11-21.

225. Jensen PR, Mincer TJ, Williams PG, & Fenical W (2005) Marine actinomycete diversity and natural product discovery. *Antonie Van Leeuwenhoek* 87(1):43-48.

226. Zhang W, *et al*. (2012) Spiroindimicins A-D: new bisindole alkaloids from a deep-sea-derived actinomycete. *Org Lett* 14(13):3364-3367.

227. Sternberg N (1990) Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. *Proc Natl Acad Sci U S A* 87(1):103-107.

228. Freel KC, Nam SJ, Fenical W, & Jensen PR (2011) Evolution of secondary metabolite genes in three closely related marine actinomycete species. *Appl Environ Microbiol* 77(20):7261-7270.

229. Selama O, *et al*. (2014) Screening for Genes Coding for Putative Antitumor Compounds, Antimicrobial and Enzymatic Activities from Haloalkalitolerant and Haloalkaliphilic Bacteria Strains of Algerian Sahara Soils. *BioMed Research International* 2014(317524):11.

230. Tulp M & Bohlin L (2005) Rediscovery of known natural compounds: nuisance or goldmine? *Bioorg Med Chem* 13(17):5274-5282.

231. Sakemi S & Sun HH (1991) Nortopsentins A, B, and C. Cytotoxic and antifungal imidazolediylbis[indoles] from the sponge Spongosorites ruetzleri. *The Journal of Organic Chemistry* 56(13):4304-4307.

232. Britton R, de Oliveira JH, Andersen RJ, & Berlinck RG (2001) Granulatimide and 6-bromogranulatimide, minor alkaloids of the Brazilian ascidian Didemnum granulatum. *J Nat Prod* 64(2):254-255.

233. Meragelman KM, *et al*. (2002) Unusual sulfamate indoles and a novel indolo[3,2-a]carbazole from Ancorina sp. *J Org Chem* 67(19):6671-6677.

234. Edenborough MS & Herbert RB (1988) Naturally occurring isocyanides. *Nat Prod Rep* 5(3):229-245.

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

236. Kim JH, *et al*. (2010) Cloning large natural product gene clusters from the environment: piecing environmental DNA gene clusters back together with TAR. *Biopolymers* 93(9):833-844.

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

238. Perlova O, *et al*. (2006) Reconstitution of the myxothiazol biosynthetic gene cluster by Red/ET recombination and heterologous expression in Myxococcus xanthus. *Appl Environ Microbiol* 72(12):7485-7494.

239. Rui Z, *et al*. (2010) Biochemical and genetic insights into asukamycin biosynthesis. *J Biol Chem* 285(32):24915-24924.

240. Yamanaka K, *et al*. (2014) Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc Natl Acad Sci U S A* 111(5):1957-1962.

241. Gibson DG, *et al*. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6(5):343-345.

242. Li MZ & Elledge SJ (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat Methods* 4(3):251-256.

243. Engler C, Kandzia R, & Marillonnet S (2008) A one pot, one step, precision cloning method with high throughput capability. *PLoS One* 3(11):e3647.

244. Du J, Yuan Y, Si T, Lian J, & Zhao H (2012) Customized optimization of metabolic pathways by combinatorial transcriptional engineering. *Nucleic Acids Res* 40(18):e142.

245. Pfleger BF, Pitera DJ, Smolke CD, & Keasling JD (2006) Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat Biotechnol* 24(8):1027-1032.

246. Jones AC, *et al*. (2012) Evaluation of Streptomyces coelicolor A3(2) as a heterologous expression host for the cyanobacterial protein kinase C activator lyngbyatoxin A. *FEBS J* 279(7):1243-1251.

247. Watanabe K, Rude MA, Walsh CT, & Khosla C (2003) Engineered biosynthesis of an ansamycin polyketide precursor in Escherichia coli. *Proc Natl Acad Sci U S A* 100(17):9774-9778.

248. Pitera DJ, Paddon CJ, Newman JD, & Keasling JD (2007) Balancing a heterologous mevalonate pathway for improved isoprenoid production in Escherichia coli. *Metab Eng* 9(2):193-207.

249. Salas JA & Mendez C (2007) Engineering the glycosylation of natural products in actinomycetes. *Trends Microbiol* 15(5):219-232.

250. Terpe K (2006) Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *Appl Microbiol Biotechnol* 72(2):211-222.

251. Zaburannyi N, Rabyk M, Ostash B, Fedorenko V, & Luzhetskyy A (2014) Insights into naturally minimised Streptomyces albus J1074 genome. *Bmc Genomics* 15.

252. Gomez-Escribano JP & Bibb MJ (2011) Engineering Streptomyces coelicolor for heterologous expression of secondary metabolite gene clusters. *Microb Biotechnol* 4(2):207-215.

253. Martin VJ, Pitera DJ, Withers ST, Newman JD, & Keasling JD (2003) Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat Biotechnol* 21(7):796-802.

254. Ro DK, *et al*. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440(7086):940-943.

255. Whicher JR, *et al*. (2013) Cyanobacterial polyketide synthase docking domains: a tool for engineering natural product biosynthesis. *Chem Biol* 20(11):1340-1351.

256. Goss RJ, Shankar S, & Fayad AA (2012) The generation of "unnatural" products: synthetic biology meets synthetic chemistry. *Nat Prod Rep* 29(8):870-889.

257. Mizuoka T, Toume K, Ishibashi M, & Hoshino T (2010) Novel tryptophan metabolites, chromoazepinone A, B and C, produced by a blocked mutant of Chromobacterium violaceum, the biosynthetic implications and the biological activity of chromoazepinone A and B. *Org Biomol Chem* 8(14):3157-3163.

258. Wang Z, *et al*. (2013) A general strategy for the chemoenzymatic synthesis of asymmetrically branched N-glycans. *Science* 341(6144):379-383.

259. Lam KS, Schroeder DR, Veitch JM, Matson JA, & Forenza S (1991) Isolation of a bromo analog of rebeccamycin from Saccharothrix aerocolonigenes. *J Antibiot (Tokyo)* 44(9):934-939.

260. Zhang C, *et al*. (2006) RebG- and RebM-catalyzed indolocarbazole diversification. *Chembiochem* 7(5):795-804.

261. Angiuoli SV, *et al*. (2011) CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* 12:356.

262. Zivadinovic D, Gametchu B, & Watson CS (2005) Membrane estrogen receptor-alpha levels in MCF-7 breast cancer cells predict cAMP and proliferation responses. *Breast Cancer Res* 7(1):R101-112.

263. Hong HJ, Hutchings MI, Hill LM, & Buttner MJ (2005) The role of the novel Fem protein VanK in vancomycin resistance in Streptomyces coelicolor. *J Biol Chem* 280(13):13055-13061.

264. Kupchan SM, Britton RW, Ziegler MF, & Sigel CW (1973) Bruceantin, a new potent antileukemic simaroubolide from Brucea antidysenterica. *J Org Chem* 38(1):178-179.

265. Aoi Y, *et al*. (2014) Dissecting the first and the second meiotic divisions using a marker-less drug-hypersensitive fission yeast. *Cell Cycle* 13(8):1327-1334.

266. Alfa C, Fantes P, Hyams J, McLeod M, & Warbrick E (1993) *Experiments with Fission Yeast* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).