Summer 7-11-2017

AN INTEGRATED BIOINFORMATIC/EXPERIMENTAL APPROACH FOR DISCOVERING NOVEL TYPE II POLYKETIDES ENCODED IN ACTINOBACTERIAL GENOMES

Wubin Gao
University of New Mexico

Wubin Gao

Candidate

Chemistry and Chemical Biology

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Jeremy S. Edwards, Chairperson

Charles E. Melançon III, Advisor

Lina Cui

Changjian (Jim) Feng

AN INTEGRATED BIOINFORMATIC/EXPERIMENTAL APPROACH FOR DISCOVERING NOVEL TYPE II POLYKETIDES ENCODED IN ACTINOBACTERIAL GENOMES

by

WUBIN GAO

B.S., Bioengineering, China University of Mining and Technology, Beijing, 2012

DISSERTATION

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
Chemistry

The University of New Mexico
Albuquerque, New Mexico

July 2017

DEDICATION

This dissertation is dedicated to my altruistic parents, Wannian Gao and Saifeng Li, who never stopped encouraging me to learn more and always supported my decisions on study and life. I also dedicate this work to my loving wife Mengqin Cai, who has accompanied me throughout the whole PhD period, as a sincere friend, girlfriend and supportive spouse. In addition, I hope this work could be dedicated to my great country, China, where I grew up and was educated for more than twenty years. Lastly, I want to dedicate this work to Buddha, whose teachings of wisdom guided me to continually improve.

ACKNOWLEDGMENTS

Time flies! It seems yesterday when I started at UNM as a naive freshman who wholeheartedly thirsted for knowledge, research, and foreign culture. However, looking back, immense vivid pictures — laughs, joy, anxiety, struggles, disappointments, persistence, bitter tears, and happiness for the harvest — all emerged in my mind in a moment.

In the summer of 2011, I got my first taste of scientific research by doing my graduation project as undergraduate in Keqian Yang's group. I did not expect that this involvement, an ordinary event as it seemed at that time, would be a turning point in my life. During those two years, I became enthusiastic about doing research. Yang's and other lab members' dedication to their work and the harmonious environment inspired me to pursue a PhD in the United States.

When I submitted my applications to a number of universities in the United States, Dr. Melançon's ideas about doing research on natural product discovery using a combination of computational tools and experimental methods attracted me immediately. The feeling about Dr. Melançon's group was like when I placed my feet into the right shoes. My inner voice said "Yes! This is exactly the group I am looking for!"

In the Melançon group, I have learned much about natural product discovery while receiving a rigorous, professional training. From literature reading and presenting to progress reports, all have been conducive to my growth and development as a scientist. Dr. Melançon always encouraged me to develop and test my own ideas and to work independently. This freedom has forced me to think independently and work hard. In addition, I met several awesome lab members, Yasushi Ogasawara, Benjamin Yackley,

Shijie Huang, Jingxuan He, Han Nguyen, and Xuechen Zhu, who became my good friends and influential teachers.

The things I have learned during my PhD studies have been tremendous and have changed me immeasurably, from my ways of thinking to my habits in handling things. For example, I tend to take some things for granted. “Be rigorous, well organized!” is what I learned from Dr. Melançon. These words immediately came into my mind when I was lazy or was considering skipping some experimental steps. All the things I have learned here will surely continue to influence me for the remainder of my life.

I am also sincerely grateful to my committee members, Dr. Jeremy Edwards, Dr. Lina Cui and Dr. Changjian (Jim) Feng for their insightful suggestions in writing the research proposal and their assistance in completing my research projects.

All those who have helped me and have been influential in my life will remain in my memory. I owe them a great debt for all the kind help they have offered to me. I wish everyone who played an indispensable role in my life much happiness in the future.

AN INTEGRATED BIOINFORMATIC/EXPERIMENTAL APPROACH FOR DISCOVERING NOVEL TYPE II POLYKETIDES ENCODED IN ACTINOBACTERIAL GENOMES

by

Wubin Gao

B.S., Bioengineering, China University of Mining and Technology, Beijing, 2012

Ph.D., Chemistry, University of New Mexico, 2017

ABSTRACT

Discovery of new natural products (NPs) is critical both for diseases treatment and crops protection. Numerous NP biosynthetic gene clusters (BGCs) in sequenced microbial genomes allow identification of new NPs through genome mining. Developing an integrated bioinformatic/experimental approach for discovering novel type II polyketides (PK-IIs) facilitates investigation of this family of NPs in an efficient, systematic way. Here, we developed an approach to analyze ketosynthase α/β (KSα/β) gene sequences to predict PK-II core structures, allowing us to target novel PK-II BGCs either from isolated genomic DNA or genomes from the NCBI databank, and to isolate novel PK-IIs produced by these BGCs.

First, new degenerate primers were designed to amplify the region containing key "fingerprint residues" used as predictive indicators of the KSα/β gene product novelty. This work resulted in identification of several BGCs encoding potentially novel PK-IIs in the genomes of 54 Actinobacteria, including 38 unique environmental strains. Next, complete

PK-II BGCs were obtained through whole genome sequencing of 5 strains of high priority. Through combined core structure prediction and bioinformatic analysis, *Alloactinosynnema sp.* L-07 was chosen for compound isolation. By optimizing fermentation conditions, we purified 3.8 mg of sample to elucidate the structure of this compound, a pentangular PK-II which we named alloactinomicin.

In another route, we bioinformatically identified over 500 PK-II BGCs from the NCBI databank, and selected 28 of these predicted to produce structurally novel PK-IIs for experimental characterization by a combined genetic/metabolic approach. As a proof of concept, the CRISPR/Cas9-based genome editing was utilized to achieve KSα gene inactivation in two *Streptomyces* PK-II BGCs. Comparison of wild-type and mutant metabolite profiles led to identification of new putative PK-IIs and correlation of genotype to chemotype for the PK-II BGC being characterized. Finally, we purified one of the metabolites identified and obtained uv-visible and mass spectral evidences consistent with an angucycline-type PK-II, which we named flavochromycin.

This work demonstrates two routes for discovery of novel PK-IIs using genomics-driven bioinformatic/experimental approaches. Results from both routes have laid the foundation for more targeted and efficient ways to discover novel NPs for drug and agrochemical development.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Apr	Apramycin
BGC	Biosynthetic gene cluster
CIP	Calf Intestinal Alkaline Phosphatase
Cm	Chloramphenicol
COSY	Correlation spectroscopy
CRISPR	Clustered regularly interspaced short palindromic repeats
DMSO	Dimethylsulfoxide
E. coli	*Escherichia coli*
EDTA	Ethylenediamine tetraacetic acid
ESI-MS	Electrospray ionization mass spectrometry
HMBC	Heteronuclear multiple bond coherence
HPLC	High performance liquid chromatography
HSQC	Heteronuclear single quantum correlation
Kan	Kanamycin
LB	Luria-Bertani
Mbp	Mega base pair
MOPS	Morpholinopropane sulphonic acid
NMR	Nuclear magnetic resonance
NP	Natural product
NRPS	Nonribosomal peptide synthetase
OD	Optical density
ORF	Open reading frame

PCR	Polymerase chain reaction
PK-II	Type II polyketide
PKS	Polyketide synthase
RBS	Ribosome binding site
RiPPs	Ribosomally synthesized and post-translationally modified peptides
SDS	Sodium dodecyl sulfate
TSB	Tryptic soy broth
UV	Ultraviolet

Chapter 1. Background and Significance

1. Introduction

Natural products are ubiquitous chemical compounds or substances with diverse sources in nature, including bacteria, fungi, plants, animals, and even human beings. In the field of microbial natural product discovery, this definition is often restricted to secondary metabolites, which are not essential for survival but provide a wide range of ecological functions. These include quorum sensing molecules (such as oligo-peptides used by Gram-positive bacteria) that increase their concentration to coordinate virulence, sporulation, and antibiotic production based on the bacterial population density[1], communication molecules that regulate the symbiotic relationship between microbes and other organisms (such as plants and higher animals), agents that transport nutrients (such as siderophores chelating iron), and competitive weapons that are used against other bacteria[2].

While the true natural roles of many secondary metabolites are elusive or unknown, they have exhibited diverse biological activities that are often employed by pharmaceutical industry, food industry and agricultural sector (Figure 1-1). Several consecutive reviews have proven that such bioactive compounds had a prominent influence in drug discovery and development process for the treatment of human diseases[3]. To note, in the past 30 years from 1981 to 2010, natural products or their semisynthetic derivatives comprise 33% (358/1073) of all small-molecule approved therapeutic agents; and comprise a marked 74% (77/104) of antibacterial new chemical entities (NCEs) and 43% (43/99) of anticancer NCEs[3]. Natural products are also widely used in food industry as food additives and in the agricultural sector as pesticides/agrochemicals[4]. Perhaps one-third of major crops would

be lost due to the lack of effective pesticides[4]. In addition, natural products are useful small-molecule probes in advancing the understanding of biological systems, gaining important insights into the biological pathways, and evaluating therapeutic targets[5] (Figure 1-1). A well-known example is the discovery of serine/threonine protein kinase mTOR (mammalian target of rapamycin). It is found to be the homolog of the yeast TOR gene product, which was first identified via molecular genetic studies of rapamycin resistant mutants of *Saccharomyces cerevisiae*[6]. This immunosuppressants, rapamycin, also helped to recognize mTOR as a core component in signaling pathways regulating different cellular processes, such as protein synthesis and cell proliferation[7].

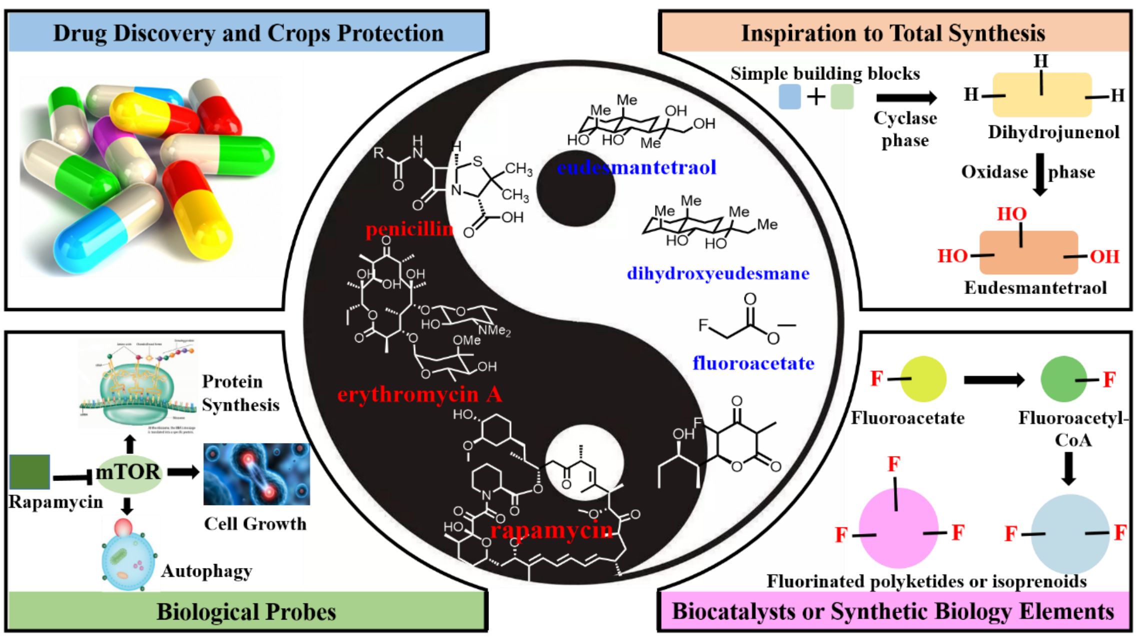

Figure 1-1. Applications of natural products related to their bioactivity and structural complexity and diversity. Natural products or their semisynthetic derivatives are clinically useful drugs or agrochemicals. They are useful biological probes; mTOR stands for mammalian target of rapamycin; rapamycin helped to establish the role of mTOR in signaling pathways such as protein synthesis and cell growth. The complexity and diversity of natural product scaffolds and functional groups invigorated chemists in designing innovative strategies for total synthesis such as site-selective oxidation. The structurally novel features usually indicate new enzymology, which has offered new opportunities for synthetic biology to make unnatural compounds such as novel fluorinated polyketides.

For the complexity and diversity of natural product scaffolds and functional groups, they have and will continue to invigorate chemists in revolutionizing chemical analytical technologies and designing innovative strategies for total synthesis (Figure 1-1)[8,9]. Although synthetic methodology has achieved great advances and almost any molecule could be a reasonable synthetic target, it is still very challenging in constructing chemical libraries with the architectural and functional group complexity of natural products and undergoing site-selective oxidations[8,9]. Furthermore, the novel structural features usually indicate new enzymology, which has contributed to our gradual realization of the enzyme diversity and has offered new opportunities for synthetic biology to make unnatural compounds, as well as novel biocatalysts exploited by biotechnology companies (Figure 1-1). Indeed, the production of the organofluorine compound fluoroacetate in *Streptomyces cattleya* medium culture leads to the striking discovery of the first fluorinase enzyme, which was exploited to incorporate fluorine into polyketide backbone both in vitro and in vivo, highlighting the prospects for the generation of novel complex fluorinated natural products using synthetic biology[10] (Figure 1-1).

Chemical diversity in nature is based on biological variety. Traditionally, natural products were chemically classified into polyketides, peptides, oligosaccharides, terpenoids, and alkaloids, reflecting the complexity and diversity of natural product scaffolds and functional groups[8]. However, this categorization did not provide any clue to how and why Nature could synthesize a great variety of natural product scaffolds. In the early 1990s, a couple of parallel landmark discoveries that the biosynthesis of specified metabolites such as erythromycin, vancomycin, and rapamycin were being deciphered by cloning the genes that encodes the assembly-line enzymes stimulated the endeavors to

solve this mystery[8]. Moreover, there is the growing appreciation that natural-products-encoding genes are physically clustered in the microbial genomes (the genes for plant natural products usually are not chromosomally adjacent). Consequently, these biosynthetic studies lead to a genetic insights-based classification system as valid as chemistry in categorizing natural products[8].

A physically clustered group of genes in a genome that genetically encodes the enzymes, regulatory proteins and transporters involved in the production, regulation and excretion of a natural product is defined as biosynthetic gene cluster (BGC)[11]. Currently, the typical BGCs identified in a genome could be divided into gene clusters encoding polyketides (PKs), nonribosomal peptides (NRPs), ribosomally biosynthesized and post-translationally modified peptides (RiPPs), NRP-PK hybrids, terpenes, saccharides and alkaloids[11] (Figure 1-2).

Polyketide BGCs encode multi-enzyme complexes, called polyketide synthases (PKSs), involved in the assembly of polyketides via the decarboxylative condensation of one or more (alkyl)malonyl thioester extender units with an acyl thioester starter unit, which is reminiscent of fatty acid biosynthesis[4] (Figuare 1-2a). In bacteria, two different types of PKSs, type I modular PKSs and type II PKSs, were utilized to assemble structurally complex polyketides and simple aromatic polyketides, respectively[4]. Type I modular PKSs consist of a certain number of modules (each of them harbors several catalytic domains with distinct functions) responsible for the incorporation of thioester building blocks and concomitant modifications of β-keto groups on the resulting intermediates[4].

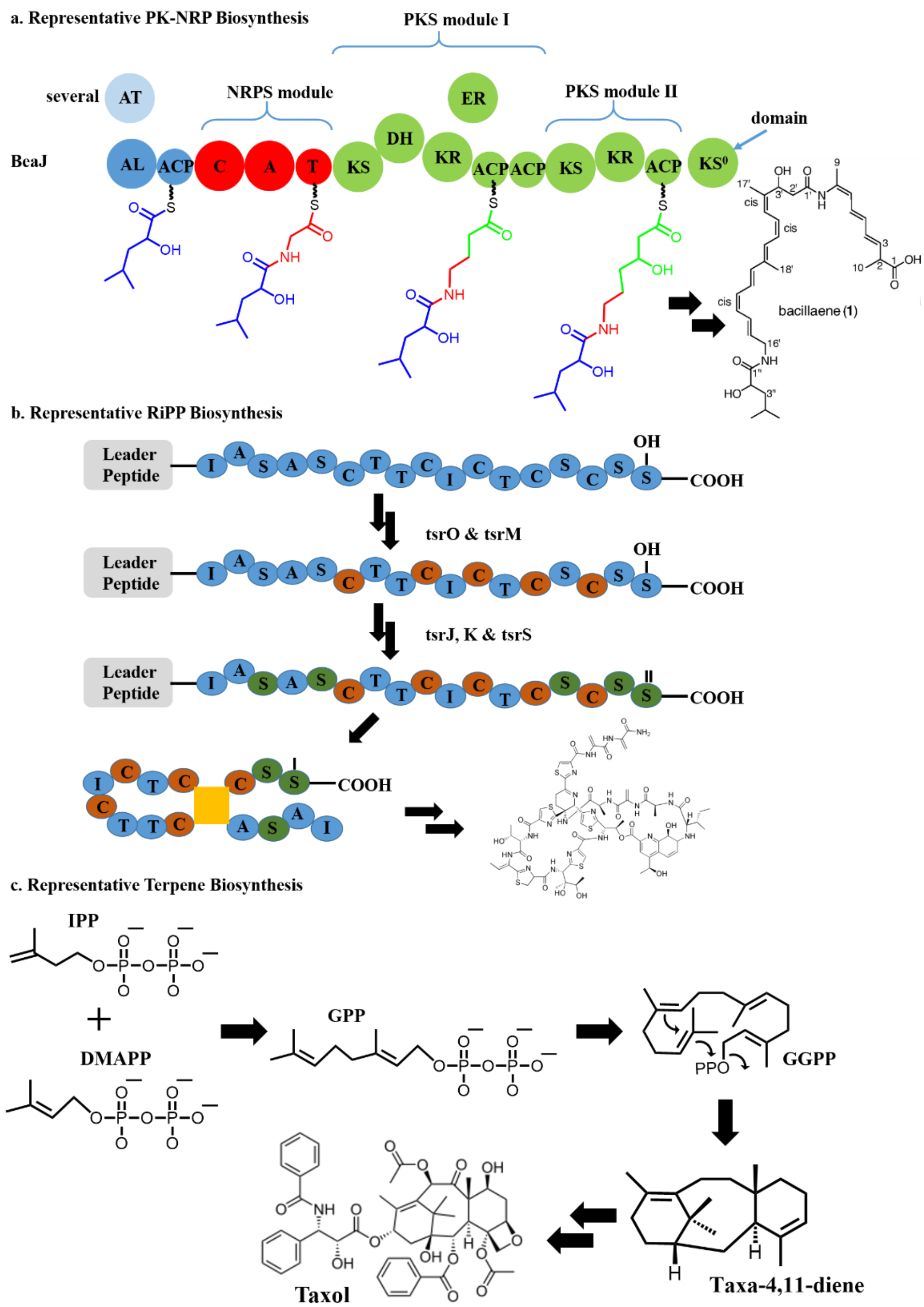

Figure 1-2. Biosynthesis of different types of natural products. a) Bacillaene pathway as an example of PK-NRP hybrid biosynthesis. Incorporated amino acids are shown in red. b) Thiostrepton A pathway as an example of RiPP biosynthesis. c) Taxol pathway as an example of Terpene biosynthesis.

Type II PKSs consist of a minimal set of iteratively used enzymes, two β-ketoacyl synthase subunits (KSα and KSβ) and an acyl carrier protein (ACP), responsible for the chain elongation of a nascent poly-β-keto intermediate, which undergoes specific ketoreduction and cyclization to yield a polyphenol with subsequent elaborations catalyzed by discrete tailoring enzymes[4]. In fungi, however, both simple aromatic and structurally more complex polyketides are encoded by type I iterative PKSs, which consist of a set of catalytic domains within a single iteratively used multi-enzyme complex[4]. NRPs are peptide natural products biosynthesized by nonribosomal peptide synthetases (NRPSs) via a ribosome- and messenger RNA-independent manner in both bacteria and fungi. These modular multi-enzymes catalyzed biosynthesis share characteristics with type I modular PKSs encoded polyketides. Owing to these architectural and mechanistic similarities, some NRPS assembly lines are compatible with PKSs modules to generate NRP-PK hybrids (Figure 1-2a)[4]. RiPPs derive from ribosomally produced peptides which undergo subsequent chemical modification by biosynthetic enzymes. RiPPs could further be divided into subclasses (e.g. lanthipeptides, thiopeptides and lasso peptides) based on the shared structural features and biosynthetic commonality (e.g. precursor peptides and post-translational peptide modifications) (Figuare 1-2b)[12]. Terpenes derive biosynthetically from elongation of a dimethylallyl pyrophosphate (DMAPP) starter unit with a set of isopentenyl pyrophosphate (IPP) extender units[4]. The resulting oligo-isoprenyl diphosphates (e.g. geranyl diphosphate, farnesyl diphosphate) are chemically transformed into a diverse array of terpenes or terpenoids (Figuare 1-2c)[4]. Saccharides are built from one or more simple sugar units (e.g. glucose, fructose) linked with glycosidic bonds that are catalyzed by glycosyltransferases (GTs). Alkaloids derive mostly from amino acid

precursors and their biosynthesis include numerous distinct pathways that are in charged by various enzymes belonging to a wide range of protein families. Recently, impressive progress has been achieved in the elucidation of key alkaloid biosynthetic genes and the understating of alkaloid biosynthesis, especially plant alkaloid metabolism, will assist in developing a more clear classification based on their biosynthesis[13].

The realm of natural product researches contain a broad range of field. A major field is the discovery of novel natural products used as starting points for the development of drugs for the treatment of human diseases. An immense clinical concern is the rapid emergence of drug-resistant pathogens, which shortens the useful lifespan of antibiotics and results in the lack of effective therapeutic compounds[14,15]. In addition, as revealed by the genome or metagenome sequencing projects, much of Nature's treasure trove of small molecules remains to be explored. Thus, even though most of pharmaceutical companies have turned away from natural product discovery efforts, there has been constant interest in the academic institutes to search new compounds[16]. Rather than increased traditional bioactivity-based screening efforts to discover new natural product scaffolds, the academic investigators turn to strategies providing better solutions to the current productivity crisis facing by the scientific community engaged in drug discovery. Notably, many groups use genomics and bioinformatics as predictors of new molecules.

Another large field is deciphering the biosynthetic pathways of these natural products, especially those unique reactions and enzyme functions. The incentive that rational genetic manipulation (e.g. prevalent combinatorial biosynthesis) of these biosynthetic pathways could churn out natural product analog library has resulted in numerous meticulous endeavor in the elucidation of the assembly of specified metabolites

over the past three decades. Addition of various ingenious DNA assembly methods into modern genetic molecular toolkit makes the complicated reconstruction of natural product clusters feasible in the production of novel compounds by synthetic biology. Indeed, synthetic biologists have employed these biosynthetic elements to reconstitute complex polyketide, non-ribosomal peptide pathways in heterologous expression hosts like *Saccharomyces cerevisiae*, *Escherichia coli*, and *Streptomyces coelicolor*[17]. In addition, since the microbes are exposed to various environmental signals, such as nutrition limitation, osmotic pressure and pH shift, the production of natural products are coordinated by delicate global and pathway-specific regulations. There are intensive studies attempting in depicting these complex regulatory networks and insights gained from these efforts has guided the activation of cryptic BGCs.

A primary goal of microbial natural product discovery is the mining of microorganisms for bioactive compounds, which could be developed into antibiotics, cancer chemotherapeutics and pesticides. Another important goal is to discover unique chemical scaffolds, which could expand the chemical space in the repertoire and could be inspirations to enrich the toolbox of chemists.

Traditionally, the routine procedure of bioactivity-based screening begins with the collection of biological samples, microbes recognized as prolific producers, from diverse environments around the world, then undergoes either direct extraction or cultivation and fermentation under standard laboratory conditions prior to extraction[18]. Extracts or their fractions are then screened for a desired bioactivity, and hits identified are subject to single compound isolation, purification and structure elucidation (Figure 1-3a)[18]. Although it is a critical component of the drug development process when the common bioactive

compounds were still being discovered, classic natural product discovery route suffers from the increasing rates of rediscovery of known metabolites over time, resulting in a concomitant low-throughput in discovering new bioactive scaffolds[19]. There are a couple of reasons: first, this traditional pipeline could not provide any information on the structural novelty prior to laborious bioactivity test and compound isolation; second, only a narrow phylogenetic loci (most of them are *Streptomyces* in Actinobacteria) have been investigated, so the biosynthetic potential of neglected microbial strains was excluded from our scope of sight[20].

Nevertheless, meliorated variations of bioactivity-based approach, such as sampling a more diverse environments (e.g. caves, extreme environments, and deep seas) and isolating endosymbiotic bacteria, have mitigated the degree of rediscovery of known compounds, enabling it a still viable route in natural product discovery[16]. For example, investigation of two fungal isolates from a microbial mat in an iron-rich natural spring led to the discovery of six new compounds: clearanols A-E and disulochrin[21]. To extract enough amount of sample for structure elucidation, different growth conditions, mimicking the natural environments or adding environmental signals, are often adopted to ferment organisms from these diverse environments. An alternative strategies based on multispecies interactions also lead to the characterization of novel compounds, such as amychelin, by implementing binary interspecies interaction assays[22]. Apart from these methods, an *in situ* cultivation technique, the iChip, was recently developed to isolate those uncultured bacteria, which are able to grow *in vitro* using specific growth factors[14]. This method assists in identifying a new antibiotic, termed teixobactin, without extant detectable resistance.

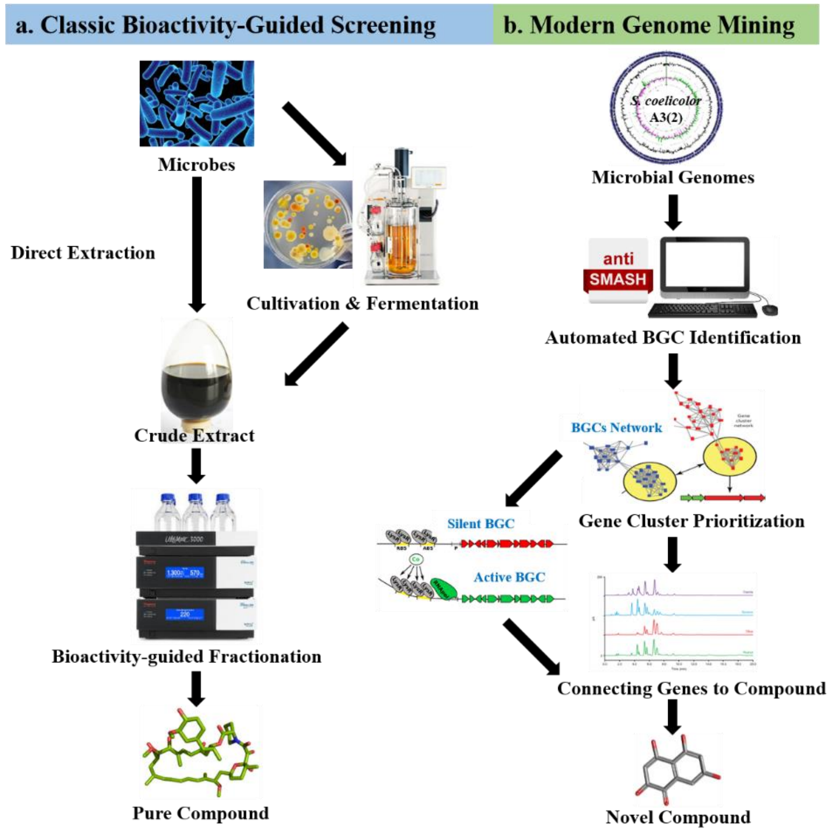

Figure 1-3. Classic bioactivity-guided screening versus genome mining approach. a) The typical workflow of bioactivity-guided screening starts by collecting and purifying microbes, then extracting metabolites from these microbial samples, fractionating crude extract based on bioactivity of interest, finally purifying the "hit" compound. b) The pipeline of genome mining approach starts by sequencing and analyzing microbial genomes, then identifying and prioritizing secondary metabolite gene clusters for compound isolation, finally elucidating the structure of the novel compound.

All the drawbacks of bioactivity-based natural product discovery arouse scientists to investigate alternative strategies to identify novel compounds. One compelling tactic to tackle these drawbacks has arisen from the exponential accumulation of microbial genomes in public databases, which were benefited from more and more sophisticated and cost-effective genome sequencing technologies. These tremendous genomic sequences give an important revelation that microbes have far greater potential to produce natural products than that identified by traditional bioactivity-based screening[23]. Meanwhile, the concept of exploiting numerous microbial genomes obtained from genome sequencing projects

requires the introduction of computational tools in order to comply with the need for organizing and analyzing the genomic data deluge. This genomics- and bioinformatics-guided approach, termed genome mining, holds the promise that transforms the landscape of discovering natural products into a high-throughput pipeline that would yield hundreds and thousands of novel structural small molecules from microbes (Figure 1-3b)[24].

With genome sequence data in hand for a specific strain, it is feasible to rapidly identify numerous natural product gene clusters by recently developed bioinformatic software such as antiSMASH[24]. In order to de-replicate and prioritize the enormous putative BGCs, several research groups have already taken the first attempts, either classifying BGCs into gene cluster families (GCFs) or predicting chemical structures of their products directly from the sequences. Bioinformatic analysis of genome sequences alone is inadequate to achieve the promise of high-throughput genome mining, and experimental methods are also essential to rapidly connect identified BGCs to their products[24]. However, prior to the experimental correlation between BGCs and their products, expression of the gene clusters is required, which is hampered by the fact that 90% BGCs are cryptic under typical laboratory growth conditions owing to their strict regulation at transcription and/or translation levels[24]. Thus, these dormant gene clusters become the bottleneck, and developing methods to activate them is of paramount importance to the realization of the full potential of genome mining[4]. Following activated expression of a BGC, the next challenge is the identification (as described above), purification and structure elucidation of the target compound. Outstanding progress has been seen during the last decade, such as the advent of cryogenically cooled NMR

microprobes. Detailed description of genomics-driven natural product discovery is introduced in Section 4 of this chapter.

Recent decades, the rise and spread of antibiotic resistant strains is a global threat to human health, animals and agricultural products[25]. This evitable problem during the use of antibiotics has caused enormous economic and human cost. For example, the emerging multidrug-resistant bacteria in Europe in 2007 infected 400,000 patients and took away 25,000 humans' lives during the routine surgical procedures[25]. Thus, the discovery of new structural classes of antibiotics is an urgent and essential action needed to tackle this worldwide life-threatening crisis. It is now clear that natural products discovered so far are only the tip of the iceberg in comparison to the whole wealth of bioactive compounds in nature revealed by the complete genome sequences. With such a treasure trove to mine, natural products based drug discovery still matter and need to be more productive to meet our increasing requirements.

Accordingly, the work described in this dissertation is taking the first step toward development of a more sophisticated and efficient approach to address these issues and to systematically identify type II polyketide (PK-II) BGCs responsible for producing novel compounds using genomics- and bioinformatics-guided dereplication and prioritization.

This background information does not intend to cover all the topics in this field with detailed description due to the scope of it, but it has three main focuses: Natural Products Produced by Actinobacteria (Section 2); Type II Polyketide Biosynthesis (Section 3); Genomics-driven Discovery of Natural Products (Section 4). In section 2, it will show the exemplified compounds discovered in Actinobacteria that were developed into valuable therapeutic agents or agricultural pesticides, which explains why Actinobacteria are an

attractive source of bioactive compounds. In section 3, it will give a big picture on the biosynthesis of type II polyketides, which is the solid basis for developing computational tools to connect natural products to the genes encoded them. In section 4, it will show the historical development of genomic-driven natural product discovery and the advantages over traditional bioassay-based screening.

2. Natural Products Produced by Actinobacteria

In taxonomy, the phylum Actinobacteria, also called Actinomycetes, belongs to a group of filamentous, high G+C, Gram-positive bacteria. They exhibit varied morphologies, physiologies, and metabolic properties that adapt themselves to the access to terrestrial soils, the rhizospheres of plant roots, marine sediments and marine sponges. To date, there are more than 140 actinobacterial genera discovered, among which *Streptomyces* is the largest genus with over 900 described species[26,27].

One of the most important feature widely appealed to the scientific community is the prosperity of their secondary metabolism pathways which produce plenty of valuable bioactive metabolites used in medicine, such as antibiotics, antifungals, antivirals, and anticancer agents, and in agriculture, including insecticides, herbicides[26]. Among filamentous Actinobacteria, the genus *Streptomyces* are responsible for the majority of known bioactive compounds, although other actinobacterial genera, such as *Saccharopolyspora*, *Amycolatopsis*, *Micromonospora* and *Actinoplanes*, are also found to be the producer of bioactive compounds[26]. Of the 22,500 microbial metabolites described, about 45% (10,100) are compounds isolated from actinobacterial fermentation. In the extant antibiotics, about 70% are of actinobacterial origin[26]. A possible explanation is that

antibiotics produced by Actinobacteria have been evolved since ~1 billion years ago to compete with other microbes and inhibit the target enzymes, macromolecules or macromolecular structures[28]. Therefore, Actinobacteria received a laudatory name as "natural drug synthesis factory".

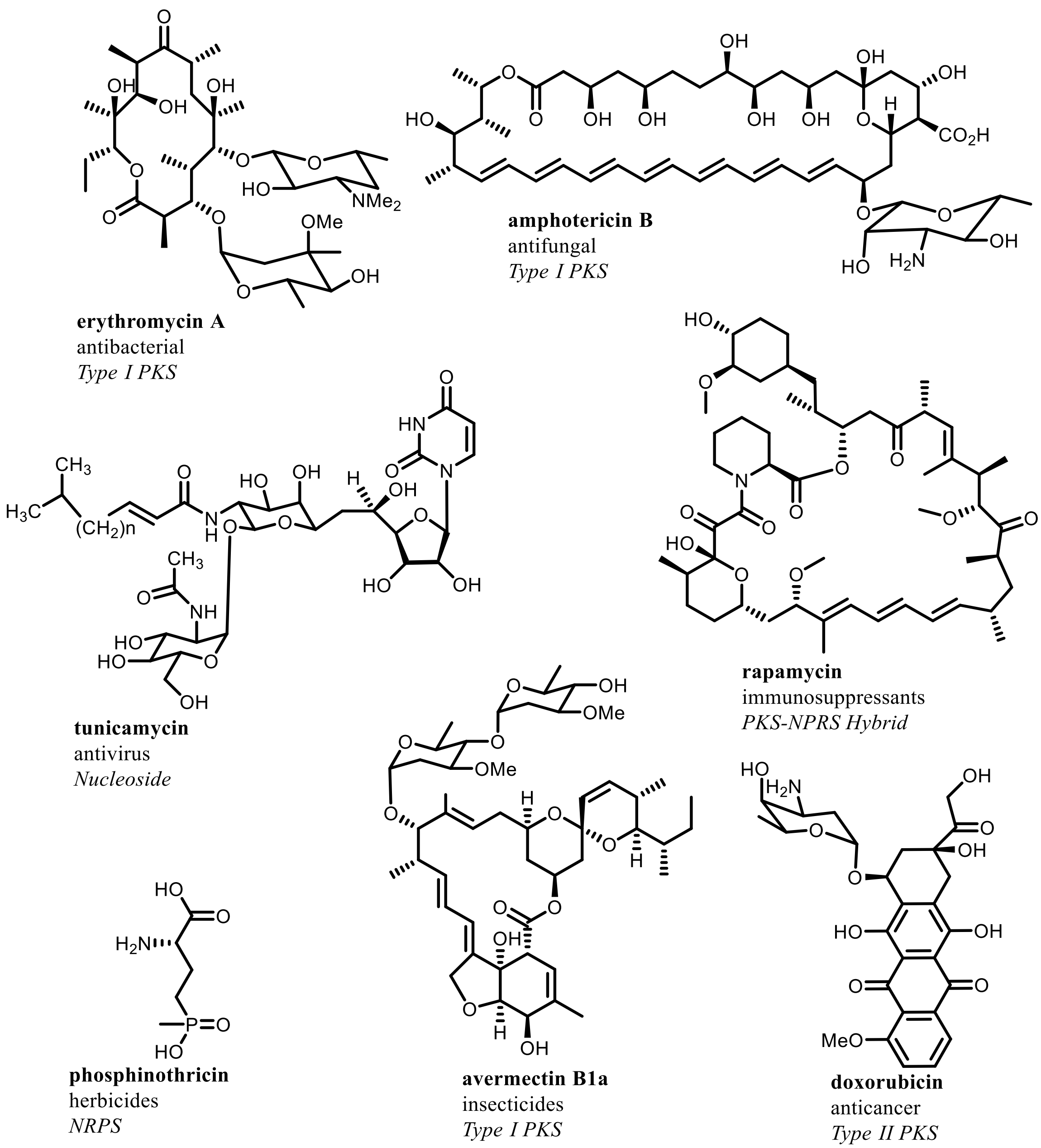

Figure 1-4. Representatives of bioactive natural products of actinobacterial origin. The application, chemical classification and biosynthetic origin for each compound are shown below its name.

Examples of bioactive metabolites used in medicine and agriculture that are produced by Actinobacteria include antibacterials (e.g., erythromycin A), antifungals (e.g., amphotericin B), antivirus (e.g. tunicamycin), immunosuppressants (e.g., rapamycin), anticancer agents (e.g., doxorubicin), insecticides (e.g., avermectin B1a) and herbicides (e.g., phosphinothricin) (Figure 1-4)[26,27,29]. Although Actinobacteria strains are such a rich producer of bioactive metabolites, it becomes increasingly difficult to screen out novel drug leads from them. For example, in the 1940s, the streptomycin was found in about 1% of random soil Actinobacteria screened, whereas vancomycin and daptomycin were discovered at much lower frequencies of 10^{-5} and 10^{-7} in the 1950s and in the late 1980s, respectively[28]. One promising solution is the enrichments and selections for uncommon terrestrial and marine Actinobacteria, which will produce new chemical diversity[28]. Indeed, the novel obligate marine Actinobacteria *Salinispora tropica* produce a β-lactone-γ-lactam salinosporamide A discovered in 2003[28].

The large genome of Actinobacteria devotes 5-10% of their coding capacity to the production of secondary metabolites, providing a solid basis for their natural product biosynthetic diversity[27,28]. In particular, the genome of *Streptomyces* (8-10 Mbp) is usually larger than other bacteria, which is consistent with the linear trend between the BGCs and genome size that is conformed by most bacterial species[30]. For example, the genome of the model Actinobacteria, *Streptomyces coelicolor* A3(2), has 8.66 million base pairs (Mbp) of DNA and harbors at least 17 BGCs for the production of chemically distinct classes of specialized metabolites[29]. In comparison, the genome of Escherichia coli K-12 (4.6 Mbp) is much smaller and contains less BGCs[27] (Figure 1-5). Hence, the genome mining of

various Actinobacteria will discover thousands of novel bioactive metabolites that are not realized by the conventional natural product isolation efforts[27].

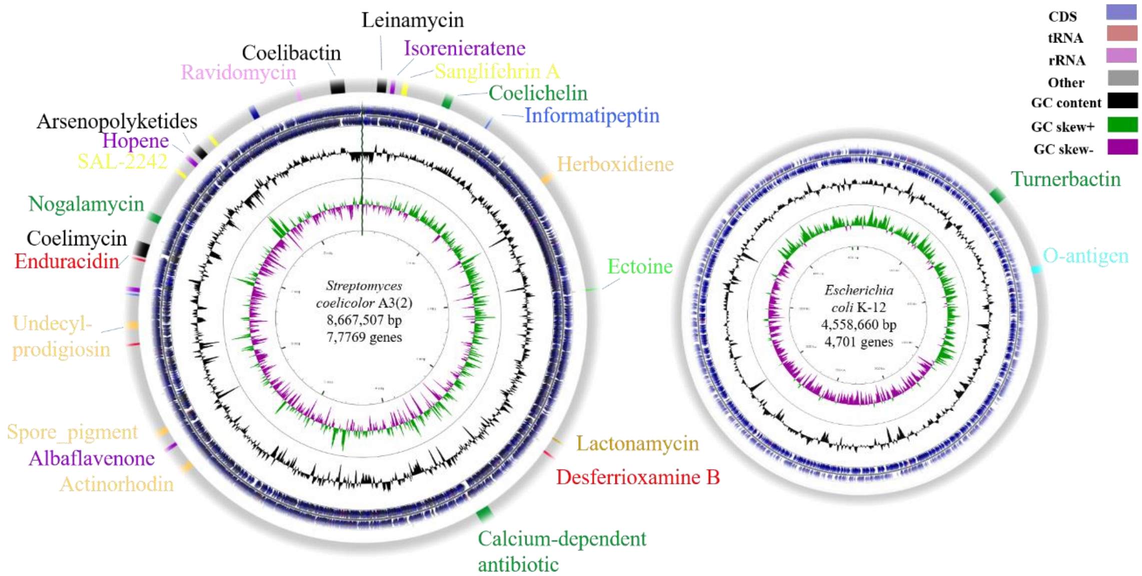

Figure 1-5. Circular representation of the genomes of *S. coelicolor* and *E. coli* K-12. The outer circles show the distribution of secondary metabolite gene clusters on the chromosome. BGCs with identified products are labeled accordingly.

3. Biosynthesis of Type II Polyketides

Type II polyketides (PK-IIs), whose core structures feature a variety of planar aromatic fused rings, comprise an important and structurally diverse family of bacterial natural products and display a wide range of biological activities, such as anti-bacterial, anti-fungal, anti-vial, anti-cancer (Figure 1-6). The antibiotics actinorhodin, a blue pigment, is the best-studied type II polyketide to date. Tetracyclines are important antibiotics with broad-spectrum used to treat a number of infections in both human and animal over several decades. Mithramycin is in clinical use for several cancer therapies. Pradimicin is an antifungal compound that also exhibits antiviral activity. Doxorubicin is commonly used for the treatment of cancers of breast and ovaries, as well as soft tissue sarcomas and aggressive lymphomas.

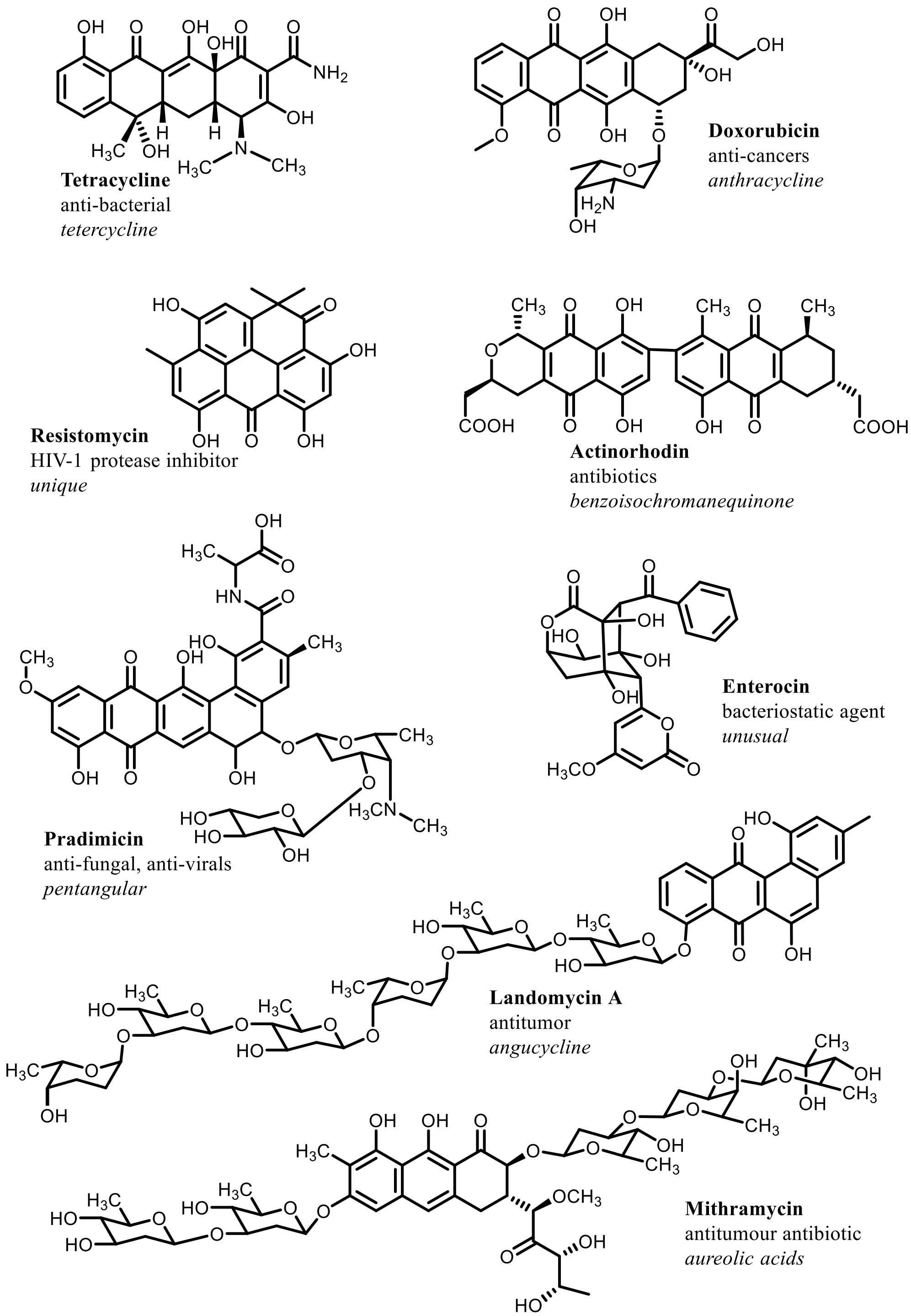

Figure 1-6. PK-IIs with various structures and broad spectrum of bioactivity. The bioactivity and subclass for each compound are shown below its name.

Landomycin A possesses strong antitumor activities, in particular against prostate cancer cell lines, which depend on the extended sugar chain. Resistomycin exhibits a

variety of potential therapeutic properties, including inhibition of HIV-1 protease and RNA polymerase. The bacteriostatic agent enterocin is a heavily rearranged aromatic molecule with a benzoyl-derived starter unit[31].

Based on the structural commonalities, the folding patterns and their polyphenolic ring modification enzymes, they could be subdivided into anthracyclines, angucyclines, aureolic acids, tetracyclines, tetracenomycins, benzoisochromanequinones, pentangular polyphenols[31] (Figure 1-6). The core structure of anthracyclines is tetracyclic ring constituted of quinone-hydroquinone moieties in ring C and B, a methoxy substituent at C-4 in ring D and a short side chain at C-9[31]. Representative compounds are doxorubicin and daunorubicin. The core structure of aureolic acids is a tricyclic ring, consisting of two aliphatic side chains at C-3 and C-7. Representative aureolic acids are chromomycin and mithramycin. The core structure of tetracyclines is also a tetracyclic ring but with a different folding patterns from anthracyclines. Representative compounds are oxytetracycline, chlorotetracycline and tigecycline. The tetracenomycins also have a tetracyclic core structure. Representative tetracenomycins are tetracenomycin and elloramycin. Angucyclines have an angular tetracyclic ring system with benz[a]anthracene/tetraphene-derived core. Representative compounds are landomycins, gilvocarcins and jadomycins. The core structure of typical pentangular polyketides is an angular polyphenolic ring system with a benzo[a]tetracene skeleton. Representative compounds are pradimicins and benastatins. The core structure of benzoisochromanequinone is tricyclic. Representative compounds are actinorhodin, medermycin and griseusin A. Except for above subclasses, there are several rare polyphenolic polyketides with unusual ring systems, such as resistomycin with

naphthanthrene, chartreusin with pentacyclic bislactone, and clostrubin with benzo[a]tetraphene (Figure 1-7)[20,32]. These structurally complex compounds possessing a broad-spectrum of bioactivity are biosynthesized by type II polyketide synthases (PKSs).

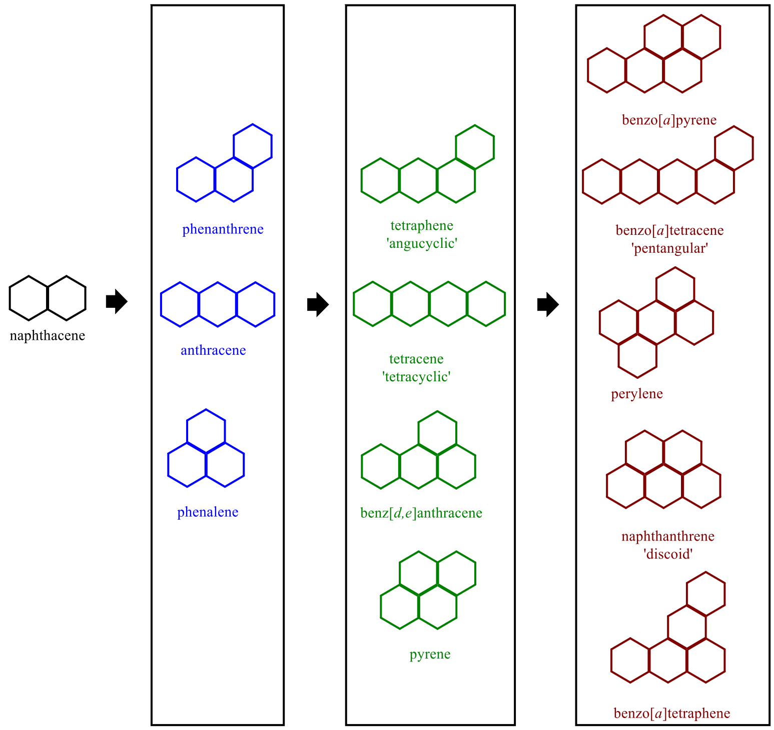

Figure 1-7. Structure-based phylogeny of fundamental di-to pentacyclic ring systems identified in aromatic polyketide scaffolds. Shared faces are highlighted in bold. This figure is modified from ref.32.

Below is a brief introduction to the formation of these structurally diverse aromatic polyketides. Basically, the biosynthesis of type II polyketide involves four phases (Figure 1-8): priming, extension, cyclization, and tailoring. Priming occurs through transfer of an acyl starter unit from corresponding acyl-coenzyme A (CoA) to an acyl carrier protein, which is subsequently transferred to the active site of a ketosynthase (KS). While PKSs commonly recruit acetate as starter unit, various alternative non-acetate units, such as propionate, (iso)butyrate, malonamate and benzoate, are employed in distinct type II polyketide biosynthetic pathways (Figure 1-9)[31]. Additional starter units, such as 2-

methylbutyryl-CoA and anthraniloyl-CoA, have been discovered recently[33]. There are three different priming mechanisms proposed: one is the acetate priming via decarboxylation of malonate; another two are non-acetate priming via either a KS3/ACPp or an AT/ADE. These non-acetate starter units in PK-II priming were typically incorporated by either a type III ketosynthase (KS3) and editing acyltransferase (AT) or stand-alone adenylation (ADE) domain together with an additional priming acyl carrier protein (ACP_p)[31].

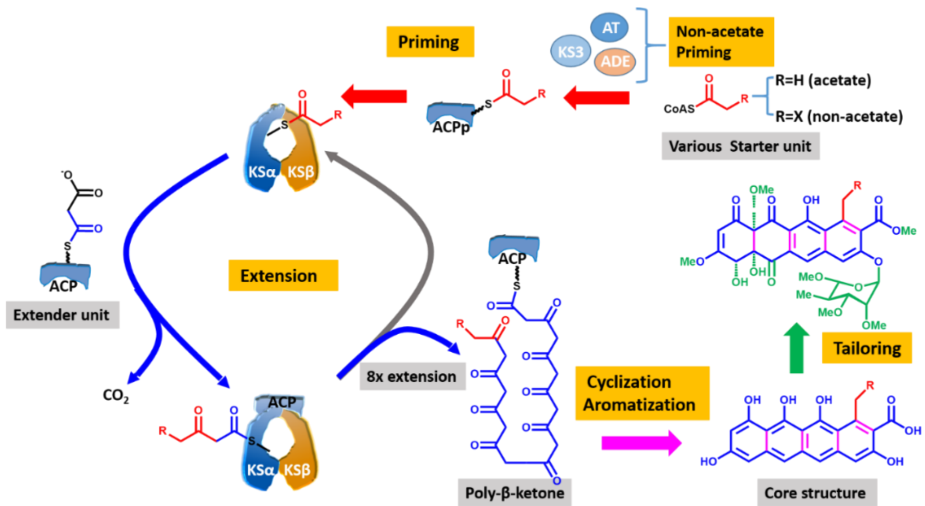

Figure 1-8. Overview of PK-II biosynthesis with the priming and extension phases shown in details. Basically, the biosynthesis of PK-II could be divided into four phases: priming, which charges an acyl starter unit; extension, which elongates via iterative Claisen-like decarboxylation to form a poly-β-ketone chain; cyclization, which cyclizes with a specific folding pattern to form the planar aromatic core structure; tailoring, which chemically modifies the core structure.

During extension, the minimal PKS complex (including ketosynthase α/β (KSα/β) and ACP) catalyzes a certain number of extension cycles (ranging from 7 to 12) via iterative Claisen-like decarboxylation using malonyl-CoA as extender units and forms a nascent poly-β-keto chain of defined length, with 16 to 30-carbon long chains found so far (Figure 1-9)[34]. The crystal structure of the actinorhodin KSα/β shows that KSα and KSβ form a heterodimer akin to the homodimeric ketosynthases from bacterial fatty acid

synthases and that the chain length of nascent poly-β-ketone could be dictated by the size and shape of the amphipathic tunnel at the heterodimer interface[35]. While the KSα subunit catalyzes C-C bond formations in chain elongation, the KSβ subunit is proposed to be the dominant element in controlling chain length[36]. Investigation of type II polyketide assembly is notoriously challenging, since it is catalyzed by a multi-enzyme complex instead of single enzymes. Thus, ingenious studies are required to offer more insights into this biosynthetic machinery. Nevertheless, the repetitive employment of KSα/β heterodimer provides us an entry point for the identification of PK-II BGCs either from sequenced genomes in NCBI databank or environmental Actinobacteria using degenerate primer-based PCR screening and next-generation sequencing.

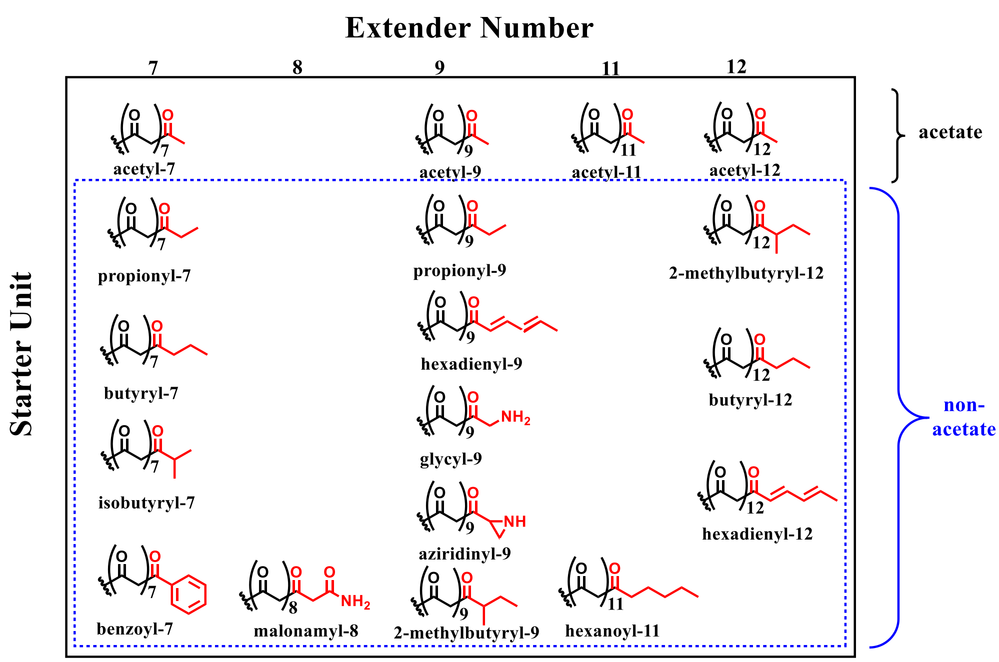

Figure 1-9. Various acetate/non-acetate starter units and extender number identified in type II polyketide biosynthesis. There are 11 different starter units and 5 different extender numbers, which give a combination of 18 different poly-b-ketone products. The non-acetate priming units are shown in the blue box.

Subsequently, the intermediate poly-β-ketone products serve as the substrate for cyclization and intermediate tailoring steps, converting to the planar aromatic core

structures. Intriguingly, a number of distinct folding patterns, which are defined by an optional C-9 ketoreductase (C9KR) and 2 to 4 different aromatase/cyclase (AroCyc) and cyclase (Cyc) enzymes, are recognized in various PK-IIs cyclizations that afford a variety of aromatic polyketide core structures (Figure 1-6)[34]. The C-9 ketoreductase could transform the keto group into a secondary alcohol, inducing a bend of the poly-β-ketone intermediate, which is placed in an orientation for a favored first cyclization (an aldol condensation)[31]. Without C9KR in the PKS set, it will result in wrong cyclized polyketides. For instance, in the absence of the *act* KR, the minimal PKS will produce a majority of incorrectly folded octaketide SEK4b instead of the correct shunt product SEK4[37]. In addition to C9KR, cyclases and aromatases are essential in directing nascent polyketide intermediate into particular folded rings. Ten cyclases groups involved in various PK-Iis cyclizations are currently identified based on phylogenetic analysis[38]. A comprehensive sequence-function correlation reveals that the ring topology correlates well with the types of cyclases. While most type II PKS pathways generally perform an individual, sequential cyclization, the resistomycin pathway is suggested to form a multi-enzyme system working in a concerted action of all components in order to shape the unusual discoid ring structure[38]. Based on the first ring cyclization, the substrate poly-β-ketone, and amino acid sequences, Aro/Cyc could also be classified into C7-C12 or C9-C14, reducing or non-reducing, and mono-domain or di-domain, respectively[33]. Following the cyclase-mediated C-C bond formations is the dehydration catalyzed by aromatases.

Finally, these aromatic core structures are then elaborately modified by a series of polyketide tailoring enzymes, such as methyltransferases (MTs), oxidoreductases, glycosyltransferases (GTs), and other enzymes[34]. These post-PKS modifications polish the

aromatic polyketide cores with structural diversities that are often potent therapeutic contributors. MTs generally use cofactor S-adenosyl-L-methinonine (SAM) and transfer their activated methyl moiety to the target sites in type II polyketide scaffolds[31]. Oxidation, catalyzed by either monooxygenases or dioxygenases, is a crucial contributor to the vast diversity of type II polyketide structures. Oxygenases catalyze a wide range of reactions, including hydroxylation, epoxidation, quinone formation, and oxidative rearrangement. These enzymes could be classified into different types, e.g. anthrone-type oxygenases, flavin-dependent oxygenases and cytochrome-dependent P450 monooxygenases[31]. GTs are responsible for the attachment of deoxysugars to their aglycones as pivotal structural elements with significant influence on their biological activities. Many GTs are demonstrated to be capable of acting on a broad range of substrates. This substrate promiscuity associated with the glycosyltransferation reversibility allows the combinatorial biosynthesists to yield numerous glycosylated aromatic polyketide analogs[39].

4. Genomics-driven Discovery of Natural Products

The process of genomics-driven natural product discovery could roughly be divided into four phases. First, it obtains (meta)genome sequences either from microorganisms of interest or from environmental samples. Then, computational tools will analyze the sequenced (meta)genomic data and identify and prioritize BGCs. Third, it genetically manipulate or activate the target BGCs using molecular biology or synthetic biology. Finally, novel natural products are isolated from culture extracts, followed by structural characterization using a combination of analytical chemistry techniques[4,23] (Figure 1-10).

In the past two decades, bacterial genome sequencing experienced three technological revolutions: whole-genome shotgun sequencing, high-throughput sequencing and single-molecule long-read sequencing[40]. The initiation of bacterial genome-sequencing revolution was in the early 1990s when the first bacterial genome sequence, from *Haemophilus influenza*, pioneered the complete genome sequencing[40].

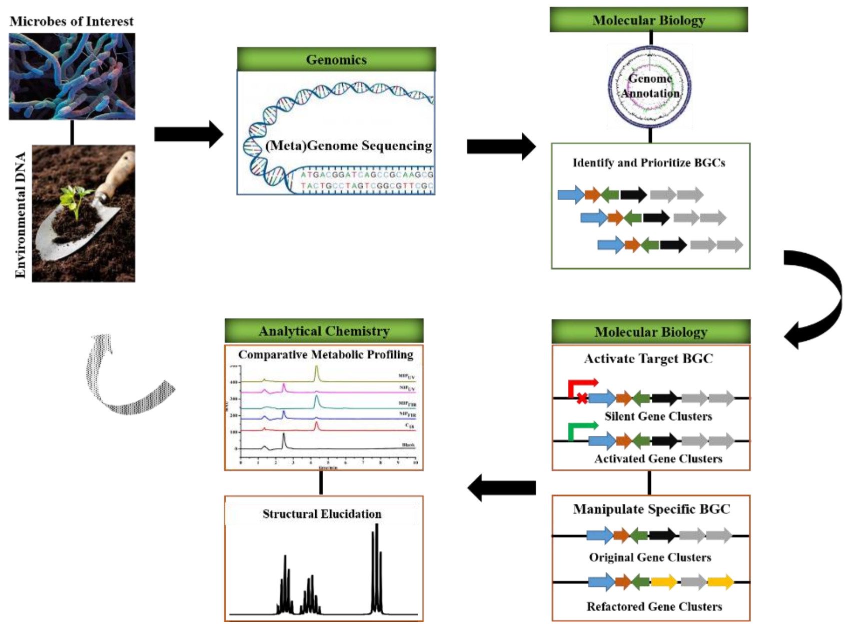

Figure 1-10. The workflow of genomics-driven natural product discovery. As shown in this schematic illustration, genome mining approach begins with identification of a microbe of interest or an environmental DNA sample. Following the (meta)genome sequencing, computational tools are used to analyze the (meta)genomic sequences, identify and prioritize BGCs. Silent BGCs are activated using a variety of genetic molecular methods. Specific BGC could also be refactored in heterologous hosts. Finally, novel natural products are identified in culture extracts via metabolic profiling, followed by compound purification and structural elucidation using advanced techniques in analytical chemistry.

In the first phase of genome sequencing, dominant efforts focused on sequencing genomes from pathogens, model organisms and extremophiles. During this period, bioinformatics-driven analysis of these genomic data, such as comparative genomics and functional genomics, has been introduced to exploit and achieve their value, and sequence-based analysis have delivered unprecedented insights into our understanding of the biology, diversity and evolution of bacteria. For instance, the genome of the model

antibiotic-producing Actinobacteria *Streptomyces coelicolor* A3 (2) sequenced fifteen years ago has revealed 16 additional biosynthetic gene clusters for potential production of novel compounds[29,41], demonstrating that the availability of a complete genome sequence has significantly advanced the identification of biosynthetic pathways encoding bioactive metabolites. In 2005, the launch of the 454 GS20 represented the arrival of high-throughput sequencing, also known as next-generation sequencing (NGS). By 2012, the emergence of benchtop sequencing platforms has a far-reaching impact on the rate of genomic data acquisition at a reduced cost[40]. From this time on, hundreds and thousands of microbial genomes have been sequenced and this trend has radically altered the landscape of genome mining. However, the massive increase in throughput came at the sacrifice of read length, which caused problems in sequence assembly when the genome contains regions of high sequence similarity, such as the modular NRPS and PKS encoding genes.

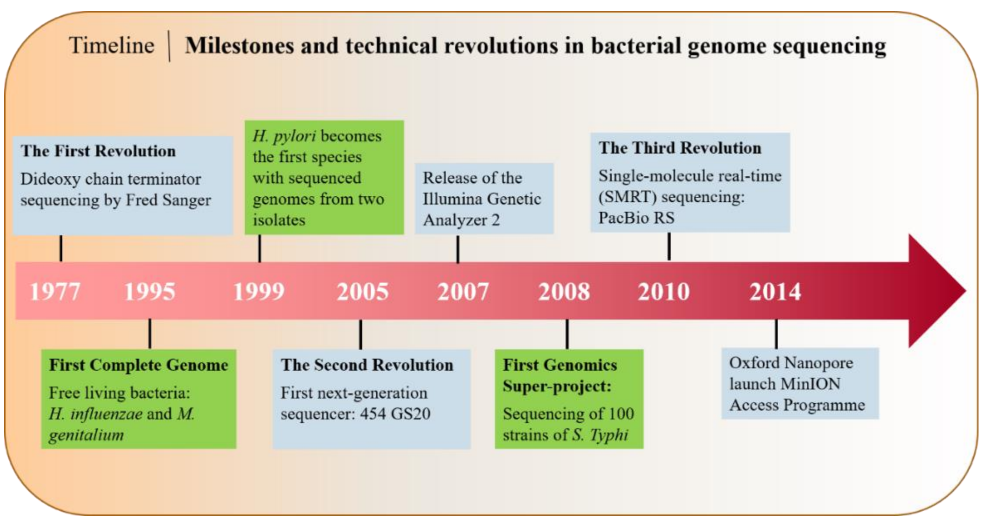

Figure 1-11. Milestones and technical revolutions in the first two decades of microbial genome sequencing. Key to the colored boxes: technical revolutions (blue); milestones of sequencing microbial genome (green).

With the advent of third revolution, these shortcomings were easily overcame. The single-molecule real-time (SMRT) sequencing technology, developed by Pacific

Biosciences, is the first widely used long-read technology to generate reads spanning large-scale repeats. Nanopore sequencing offers an innovative alternative to deliver high-quality genome-scale assemblies in microbial genome sequencing[40] (Figure 1-11). The ongoing technological revolution in microbial genome sequencing is most likely to continuously generate exponentially increasing amounts of genomic data in a more rapid, low cost and low error rates manner[42].

The arrival of cost-effective high-throughput sequencing has led to two important revelations with respect to natural product discovery: one is that microbes have far greater potential to produce structurally diverse metabolites[30]; another one is that the sequenced genomes possess a wealth of cryptic BGCs, which represent an untapped source for discovering novel specialized metabolites[4]. These revelations provided the impetus for the development of computational tools for identifying and prioritizing the BGCs that are most likely to deliver novel metabolites.

Currently, tools for computational identification of BGCs in genome sequences can be grouped into two categories[24] (Figure 1-12). The basic principle behind the first category is based on manually curated lists of query sequences, applying profile hidden Markov models (HMMs) generated from multiple sequence alignments to identify signature genes or protein domains that are highly specific for certain classes of secondary metabolite BGCs. The representative software tools devised to date include antiSMASH, ClustScan, SMURF and np.searcher. It is now common for researchers to use these tools to perform a quick and reliable detection of the BGC repertoire on the sequenced genome from a specific bacterium with high confidence. For instance, the widely used antiSMASH has been continuously updated to enable the identification of more than 20 classes of specialized

metabolite BGCs[24]. Although these routinely used tools are highly specific for the detection of known BGC classes, an inherent shortcoming is that they cannot recognize unknown types of gene clusters. Thus, the second category emerged to detect new classes of BGCs, which are of top priority because they have high chances for encoding new molecules with entirely novel scaffolds[24]. Three recently published strategies take the first attempts to solve this challenge. The first strategy, exemplified by ClusterFinder, is based on global patterns of broad gene families encoded in secondary metabolite pathways. The behind algorithm of ClusterFinder is capable of identifying gene clusters because a biosynthetic pathway is rich of Pfam domains with high frequency of presence in the known BGCs. Indeed, ClusterFinder identified a large family of BGCs encoding aryl polyenes that are unrecognized by previous signature genes based approaches. The second strategy is EvoMining that detects the presence of distant paralogs of primary metabolic enzymes in genomes. These paralog enzymes are subsequently used as indicators in identifying new types of BGCs. The third strategy is based on large-scale comparative genomic alignment that detects largely syntenic regions within the non-core genomes of a taxon. This motif-independent algorithm is applied to the prediction of BGCs in sequenced genomes of *Asperigillus* species. Following the identification of tens of thousands of putative BGCs, prioritizing this large number of BGCs is of paramount importance as well as challenging. Usually, enormous BGCs are classified into gene cluster families (GCFs) based on overall similarity of gene content between BGCs and sequence identity of biosynthetic genes. For instance, phylogenetic analysis of the ketosynthase and condensation domains from the polyketide and nonribosomal peptide BGCs. Hitherto, efforts toward prioritization of BGCs have limited resolution, and the development of creative algorithms are required. A

further step to BGC prioritization is predicting chemical structures directly from genome sequences, which is a daunting challenge. Currently, software tools such as NP.searcher and antiSMASH are only able to predict the constituent monomer of a polyketide or nonribosomal peptide based on the substrate specificities of NRPS adenylation domains and PKS acyltransferase domains[24].

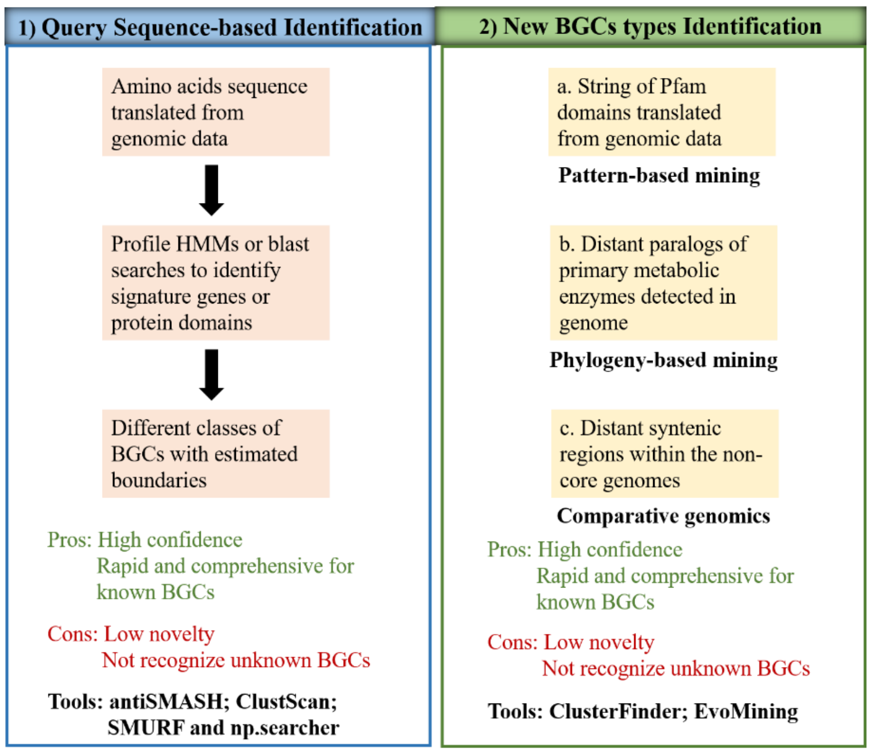

Figure 1-12. Computational tools for the genomic identification of BGCs. 1) The query sequence-based BGC identification uses profile HMMs or BLAST searches to identify genes or protein domains that are conserved in certain classes of BGCs. 2) There are at least three approaches developed for the identification of new BGC types. The principal behind each approach are listed above their names. At the bottom of the figure, advantages and disadvantages are given, as well as corresponding bioinformatic tools/software.

The ongoing development of sophisticated computational tools to the identification and prioritization of BGCs will further narrow down the ever-increasing number of BGCs to a number that is feasible to the subsequent extraction of target compounds they encoded. However, most natural product BGCs identified in the genome are expressed poorly or completely silent under standard laboratory fermentation cultures. Hence, approaches to

awaken these silent BGCs are of critical importance to address the problem and finally achieve the full potential of genomics-driven natural product discovery (Figure1-13).

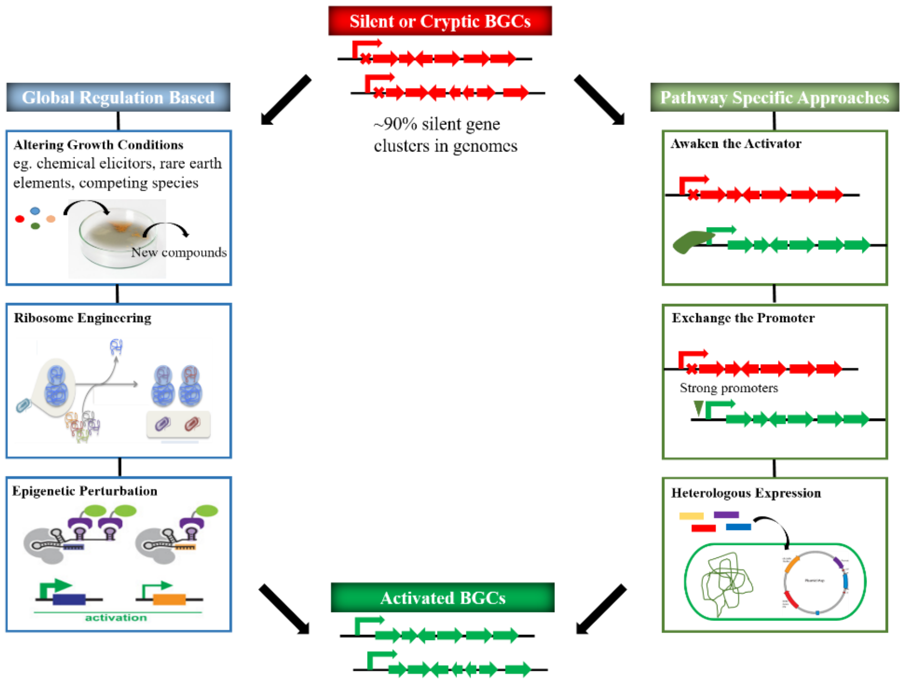

Figure 1-13. Various strategies for the activation of silent BGCs. These strategies could roughly be divided into two groups: global regulation-based approaches and pathway specific activation approaches. Examples of each approach are illustrated. Details of each approach are described in the context.

One strategy is based on global regulations of secondary metabolites BGCs, including altering growth conditions, manipulating global regulators, engineering the transcription and translation machinery and perturbing epigenetics[4]. These methods have advantages that they are readily accomplished, high throughput and do not require knowledge about the regulation of a specific BGC. Among these approaches, perhaps the simplest way is growing the microorganisms in a range of different culture conditions, including different carbon/nitrogen sources, addition of chemical elicitors, addition of an aqueous soil extract, introduction of rare earth elements and competing species. Indeed, these methods enabled discovery of several remarkable compounds such as polythioamide. However, the drawbacks of this strategy are also involved: BGCs are randomly activated

and are not predictable; nonspecific activation of multiple BGCs complicates the compound identification[4]. Thus, this strategy is more applicable to efforts that dedicate to the discovery of novel metabolites, even in phylo-genetically distinct organisms.

Alternatively, strategies aiming to the specific activation of a particular silent BGC, including manipulating pathway-specific regulators, exchanging the natural promoters and heterologously expressing a BGC[4,43] were developed. Compared with the global regulation-based strategy, pathway-specific strategy offers highly precise and controlled activation of a pathway of interest, and simplifies the identification of metabolic compounds[4]. Among these methods, heterologous expression maybe the most promising way that could be generalized to achieve the full chemical space recognized by cryptic BGCs in microbial genomes, especially when the native producer is slow growing and/or genetically intractable. The heterologous expression of microbial gene clusters was often coupled with genetic manipulation, refactoring, or synthetic biology, such as increasing the expression of pathway precursors, inserting self-resistance genes, or incorporating stronger promoters[43]. Since an enormous number of natural products are discovered from Actinobacteria, there is a focus on development of genetically well-studied actinobacterial strains into versatile heterologous hosts, particularly *Streptomyces* due to their inherent capability to produce varied secondary metabolites[44,45]. A big challenge for heterologous expression is the difficulty of cloning and manipulating large gene clusters, often as long as over 100 kb, in an effective manner. Several solutions have been developed to address this challenge, including Gibson isothermal assembly, homologous recombination-based approaches (e.g. TAR system, Red/ET system), ΦBT1 integrase-mediated site-specific recombination (i.e. IR system) and Bacterial Artificial Chromosomes (BAC) vectors (e.g.

pSBAC vector system, pTARa vector system)[43,46]. Recently, a new cloning method, based on the *in vitro* activity of RNA-guided Cas9 nuclease to cleave target genome segment and subsequently use of Gibson assembly to ligate into cloning vector, can effectively clone long bacterial genomic sequences of up to 100 kb in a single step[47]. The versatility of this Cas9-assited targeting of chromosome segments (CATCH) method was exemplified by cloning large gene clusters from various bacteria, such as a 78-kb bacillaene-producing *psk* gene cluster from *Bacillus subtilis*, a 36-kb jadomycin-producing *jad* gene cluster from *Streptomyces venezuelae* and a 83-kb calcium-dependent antibiotic-producing *cda* gene cluster from *Streptomyces coelicolor*[47,48].

More recently, an inventive strategy developed for the targeted activation of a specific silent BGC, termed reporter-guided mutant selection, combines genome-scale random mutagenesis with a pathway-specific reporter-guided mutant selection. This strategy exploits the global mutagenesis to introduce genetic diversity, while enabling the pathway-specific detection of mutants in which transcription of target BGC is activated[49].

After the activation of a target BGC, the next key step is to identify, purify and structurally characterize the metabolites of interest using a combination of analytical chemistry techniques[24]. Normally, the correlation between genes and products has most been achieved via gene knockout followed by comparative metabolomics profiling, and it is currently more efficient with the development of genetic manipulation techniques such as the prevalent CRISPR/Cas9-based genome editing. Apart from BGCs inactivation, advances in mass spectrometry led to the emergence of peptidogenomics and glycogenomics, which use tandem mass spectrometry to identify fragments (peptide fragments and glycosyl groups) and link them to corresponding biosynthetic genes in the

BGCs. These mass spectrometry-guided methods facilitate fast identification and characterization of peptide natural products and glycosylated molecules, respectively[50,51]. After targeting the compound of interest, a crude extract is pre-fractioned to simply its components by flash chromatographic methods such as silica gel, sephadex LH-20 and XAD-7 resin. Subsequent chromatography (e.g. routinely used semi-preparative HPLC) on existing fractions is used to isolate each compound. Finally, the HRMS, together with NMR, will solve the complete structures of the compounds. An impressive advancement in this area is the development of small volume probes coupled with cryogenically cooled preamplifier electronics[52]. With the arrival of sensitive microcyroprobe NMR, it became possible to solve the structure of a novel metabolite at nanomole scale[52,53]. For example, the structure of phorbaside F was solved using 7.5 μg sample and the structure determination of hemi-phorboxazole A was achieved at 16.5 μg[52]. One difficulty of structural elucidation frequently associated with natural products is the high degree of stereochemistry exhibited by their scaffolds[53]. Circular dichroism (CD) is a complement in this difficult task of acquisition of complete stereostructures of complex natural products[52]. Taken together, applications of modern innovations in analytical spectroscopy, particularly NMR, allow structure determination for natural products at nanomole level, which expand the chemical diversity within a single organism and broadens the scope of natural product discovery to rare environmental samples[52].

5. Summary and Thesis Statement

Natural products are an important source of drugs used in the treatment of human diseases and play an essential role in the protection of crops. The development of modern

molecular biological techniques associated with the sophisticated genome sequencing techniques have caused dramatic changes in the way bioactive or structurally novel natural products being discovered. The genomics- and bioinformatics-guide approaches, termed genome mining, were emerged into the natural product discovery pipeline, holding the promise that it would transform natural product discovery into a high-throughput pipeline. One prominent advantage of genome mining was that the computational tools allowed us to rapidly identify and annotate the natural product biosynthetic gene clusters in the sequenced genomes and facilitated the characterization of the novel natural products by predicting and prioritizing the structures from the sequences.

Owing to the urgent needs for new drugs, natural product discovery became an important research field. The work described in this dissertation focused on two genome mining routes for discovering novel PK-IIs, a structurally diverse family of natural products, either from the genomic DNA isolated from environmental Actinobacteria or from deposited genomes in the NCBI databank. The first route started by isolating Actinobacteria of interest from environmental samples collected from extreme environment such as caves. Then the genomic DNA was extracted from purified strains and probed by a new pair of degenerate primer that would amplify a 1.2kb region from the highly conserved KSα/β genes. Next, the sequences of these KSα/β amplicons were obtained by Sanger sequencing, followed by automated extraction and analysis of key "fingerprint residues" using our computational tools. These fingerprint residues were identified by two different predictive models: one was based on principal component analysis of representive descriptors of amino acids and poly-β-ketone structures; another one was based on machine learning algorithm in scoring the predictive ability of amino

acid positions. The fingerprint residues obtained from the more accurate model strongly correlated with the structure of poly-β-ketone (Chapter2). Then the genome of strains predicted to produce novel PK-IIs was sequenced using third generation genome sequencing technology (Chapter2). After obtaining the whole genome of strains of interest, complete type II PKS gene clusters were identified from the sequenced genome. Finally, the bioinformatics identified gene clusters were subjected to compound isolation and structural elucidation (Chapter3). This work resulted in characterization of a novel pentangular PK-II from *Alloactinosynnema sp.* L-07, which we named alloactinomicin.

The second route started by identifying the complete PK-II gene clusters from deposited genomes in the NCBI databank. Through *Dynamite*, a bioinformatic software developed by our group, we bioinformatically identified about 530 currently unstudied PK-II BGCs, among which about 10% were estimated to encode structurally novel PK-II compounds. After the identification of these PK-II BGCs, further bioinformatic analysis was carried out to prioritize them for compound isolation based on the sequence characteristics. Based on the criteria of structural novelty including non-acetate priming, novel KSα/β fingerprint residues, and atypical cyclization enzymes, we selected 28 PK-II BGCs predicted to produce structurally novel PK-II compounds for experimental characterization (Chapter4). Next, we utilized the CRISPR/Cas9 system-based genome editing to inactivate KSα genes from two *Streptomyces* PK-II BGCs and identified new putative PK-II compounds by comparing wild-type and mutant metabolite profiles (Chapter4). This work resulted in isolation of a novel angucycline-type PK-II from *S. flavochromogenes*, which we named flavochromycin.

The work described in this dissertation provided a global insight into the biosynthetic potential of PK-IIs in bacteria from diverse phylogenetic loci. Second, the genome mining approaches reported in this dissertation allowed us to efficiently de-replicate and prioritize the PK-II BGCs to target structurally novel compounds in a systematic manner. Third, the fermentation condition optimization and CRISPR/Cas9 system-based PK-II BGC inactivation would be used in discovering more PK-IIs in the future. Forth, two novel PK-II compounds with structural novelty offered new insights into the biosynthesis of PK-IIs. Finally, the novel features of isolated compounds offered new biosynthetic elements, which would in turn be used to mine databases to discover untapped bacterial producers of compounds with novel features.

6. References

1. Miller, M. B., & Bassler, B. L. (2001). Quorum sensing in bacteria. *Annual Reviews in Microbiology*, *55*(1), 165-199.
2. Demain, A. L., & Fang, A. (2000). The natural functions of secondary metabolites. In *History of Modern Biotechnology I* (pp. 1-39). Springer Berlin Heidelberg.
3. Newman, D. J., & Cragg, G. M. (2012). Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of natural products*, *75*(3), 311-335.
4. Rutledge, P. J., & Challis, G. L. (2015). Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nature Reviews Microbiology*, *13*(8), 509-523.

5. Schreiber, S. L., Kotz, J. D., Li, M., Aubé, J., Austin, C. P., Reed, J. C., ... & Alexander, B. R. (2015). Advancing biological understanding and therapeutics discovery with small-molecule probes. *Cell*, *161*(6), 1252-1265.

6. Sabatini, D. M., Erdjument-Bromage, H., Lui, M., Tempst, P., & Snyder, S. H. (1994). RAFT1: a mammalian protein that binds to FKBP12 in a rapamycin-dependent fashion and is homologous to yeast TORs. *Cell*, *78*(1), 35-43.

7. Lipton, J. O., & Sahin, M. (2014). The neurology of mTOR. *Neuron*, *84*(2), 275-291.

8. Walsh, C. T., & Fischbach, M. A. (2010). Natural products version 2.0: connecting genes to molecules. *Journal of the American Chemical Society*, *132*(8), 2469-2493.

9. Davies, H. M. (2009). Organic chemistry: Synthetic lessons from nature. *Nature*, *459*(7248), 786-787.

10. Walker, M. C., Thuronyi, B. W., Charkoudian, L. K., Lowry, B., Khosla, C., & Chang, M. C. (2013). Expanding the fluorine chemistry of living systems using engineered polyketide synthase pathways. *Science*, *341*(6150), 1089-1094.

11. Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., ... & Cruz-Morales, P. (2015). Minimum information about a biosynthetic gene cluster. *Nature chemical biology*, *11*(9), 625-631.

12. Arnison, P. G., Bibb, M. J., Bierbaum, G., Bowers, A. A., Bugni, T. S., Bulaj, G., ... & Cotter, P. D. (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Natural product reports*, *30*(1), 108-160.

13. Ziegler, J., & Facchini, P. J. (2008). Alkaloid biosynthesis: metabolism and trafficking. *Annu. Rev. Plant Biol.*, *59*, 735-769.

14. Ling, L. L., Schneider, T., Peoples, A. J., Spoering, A. L., Engels, I., Conlon, B. P., ... & Jones, M. (2015). A new antibiotic kills pathogens without detectable resistance. *Nature*, *517*(7535), 455-459.

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

16. Luo, Y., Cobb, R. E., & Zhao, H. (2014). Recent advances in natural product discovery. *Current opinion in biotechnology*, *30*, 230-237.

17. Gomez-Escribano, J. P., & Bibb, M. J. (2014). Heterologous expression of natural product biosynthetic gene clusters in Streptomyces coelicolor: from genome mining to manipulation of biosynthetic pathways. *Journal of industrial microbiology & biotechnology*, *41*(2), 425-431.

18. Katz, L., & Baltz, R. H. (2016). Natural product discovery: past, present, and future. *Journal of industrial microbiology & biotechnology*, *43*(2-3), 155-176.

19. Baltz, R. H. (2006). Marcel Faber Roundtable: is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration?. *Journal of Industrial Microbiology and Biotechnology*, *33*(7), 507-513.

20. Pidot, S., Ishida, K., Cyrulies, M., & Hertweck, C. (2014). Discovery of clostrubin, an exceptional polyphenolic polyketide antibiotic from a strictly anaerobic bacterium. *Angewandte chemie*, *126*(30), 7990-7993.

21. Gerea, A. L., Branscum, K. M., King, J. B., You, J., Powell, D. R., Miller, A. N., ... & Cichewicz, R. H. (2012). Secondary metabolites produced by fungi derived from a microbial mat encountered in an iron-rich natural spring. *Tetrahedron letters*, *53*(32), 4202-4205.

22. Seyedsayamdost, M. R., Traxler, M. F., Clardy, J., & Kolter, R. (2012). Old meets new: using interspecies interactions to detect secondary metabolite production in actinomycetes. *Methods in enzymology*, *517*, 89.

23. Winter, J. M., Behnken, S., & Hertweck, C. (2011). Genomics-inspired discovery of natural products. *Current opinion in chemical biology*, *15*(1), 22-31.

24. Medema, M. H., & Fischbach, M. A. (2015). Computational approaches to natural product discovery. *Nature chemical biology*, *11*(9), 639-648.

25. Bush, K., Courvalin, P., Dantas, G., Davies, J., Eisenstein, B., Huovinen, P., ... & Lerner, S. A. (2011). Tackling antibiotic resistance. *Nature Reviews Microbiology*, *9*(12), 894-896.

26. Solecka, J., Zajko, J., Postek, M., & Rajnisz, A. (2012). Biologically active secondary metabolites from Actinomycetes. *Open Life Sciences*, *7*(3), 373-390.

27. Nett, M., Ikeda, H., & Moore, B. S. (2009). Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Natural product reports*, *26*(11), 1362-1384.

28. Baltz, R. H. (2008). Renaissance in antibacterial discovery from actinomycetes. *Current opinion in pharmacology*, *8*(5), 557-563.

29. Challis, G. L. (2014). Exploitation of the Streptomyces coelicolor A3 (2) genome sequence for discovery of new natural products and biosynthetic pathways. *Journal of industrial microbiology & biotechnology*, *41*(2), 219-232.

30. Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Brown, L. C. W., Mavrommatis, K., ... & Birren, B. W. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, *158*(2), 412-421.

31. Hertweck, C., Luzhetskyy, A., Rebets, Y., & Bechthold, A. (2007). Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork. *Natural product reports*, *24*(1), 162-190.

32. Huang, X., He, J., Niu, X., Menzel, K. D., Dahse, H. M., Grabley, S., ... & Hertweck, C. (2008). Benzopyrenomycin, a Cytotoxic Bacterial Polyketide Metabolite with a Benzo [a] pyrene‐Type Carbocyclic Ring System. *Angewandte Chemie International Edition*, *47*(21), 3995-3998.

33. Zhang, Z., Pan, H. X., & Tang, G. L. (2017). New insights into bacterial type II polyketide biosynthesis. *F1000Research*, *6*.

34. Ogasawara, Y., Yackley, B. J., Greenberg, J. A., Rogelj, S., & Melançon III, C. E. (2015). Expanding our understanding of sequence-function relationships of type II polyketide biosynthetic gene clusters: bioinformatics-guided identification of Frankiamicin A from Frankia sp. EAN1pec. *PloS one*, *10*(4), e0121505.

35. Keatinge-Clay, A. T., Maltby, D. A., Medzihradszky, K. F., Khosla, C., & Stroud, R. M. (2004). An antibiotic factory caught in action. *Nature structural & molecular biology*, *11*(9), 888-893.

36. Tang, Y., Tsai, S. C., & Khosla, C. (2003). Polyketide chain length control by chain length factor. *Journal of the American Chemical Society*, *125*(42), 12708-12709.

37. Fu, H., Hopwood, D. A., & Khosla, C. (1994). Engineered biosynthesis of novel polyketides: evidence for temporal, but not regiospecific, control of cyclization of an aromatic polyketide precursor. *Chemistry & biology*, *1*(4), 205-210.

38. Fritzsche, K., Ishida, K., & Hertweck, C. (2008). Orchestration of discoid polyketide cyclization in the resistomycin pathway. *Journal of the American Chemical Society*, *130*(26), 8307-8316.

39. Griffith, B. R., Langenhan, J. M., & Thorson, J. S. (2005). 'Sweetening'natural products via glycorandomization. *Current opinion in biotechnology*, *16*(6), 622-630.

40. Loman, N. J., & Pallen, M. J. (2015). Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology*.

41. Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., ... & Bateman, A. (2002). Complete genome sequence of the model actinomycete Streptomyces coelicolor A3 (2). *Nature*, *417*(6885), 141-147.

42. Land, M., Hauser, L., Jun, S. R., Nookaew, I., Leuze, M. R., Ahn, T. H., ... & Poudel, S. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*, *15*(2), 141-161.

43. Ongley, S. E., Bian, X., Neilan, B. A., & Müller, R. (2013). Recent advances in the heterologous expression of microbial natural product biosynthetic pathways. *Natural product reports*, *30*(8), 1121-1138.

44. Baltz, R. H. (2010). Streptomyces and Saccharopolyspora hosts for heterologous expression of secondary metabolite gene clusters. *Journal of industrial microbiology & biotechnology*, *37*(8), 759-772.

45. Baltz, R. H. (2016). Genetic manipulation of secondary metabolite biosynthesis for improved production in Streptomyces and other actinomycetes. *Journal of industrial microbiology & biotechnology*, *43*(2-3), 343-370.

46. Nah, H. J., Pyeon, H. R., Kang, S. H., Choi, S. S., & Kim, E. S. (2017). Cloning and Heterologous Expression of a Large-sized Natural Product Biosynthetic Gene Cluster in Streptomyces Species. *Frontiers in Microbiology*, *8*.

47. Jiang, W., Zhao, X., Gabrieli, T., Lou, C., Ebenstein, Y., & Zhu, T. F. (2015). Cas9-Assisted Targeting of CHromosome segments CATCH enables one-step targeted cloning of large gene clusters. *Nature communications*, *6*.

48. Jiang, W., & Zhu, T. F. (2016). Targeted isolation and cloning of 100-kb microbial genomic sequences by Cas9-assisted targeting of chromosome segments. *Nature protocols*, *11*(5), 960-975.

49. Guo, F., Xiang, S., Li, L., Wang, B., Rajasärkkä, J., Gröndahl-Yli-Hannuksela, K., ... & Yang, K. (2015). Targeted activation of silent natural product biosynthesis pathways by reporter-guided mutant selection. *Metabolic engineering*, *28*, 134-142.

50. Kersten, R. D., Yang, Y. L., Xu, Y., Cimermancic, P., Nam, S. J., Fenical, W., ... & Dorrestein, P. C. (2011). A mass spectrometry–guided genome mining approach for natural product peptidogenomics. *Nature chemical biology*, *7*(11), 794-802.

51. Kersten, R. D., Ziemert, N., Gonzalez, D. J., Duggan, B. M., Nizet, V., Dorrestein, P. C., & Moore, B. S. (2013). Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proceedings of the National Academy of Sciences*, *110*(47), E4407-E4416.

52. Molinski, T. F. (2010). Microscale methodology for structure elucidation of natural products. *Current opinion in biotechnology*, *21*(6), 819-826.

53. Harvey, A. L., Edrada-Ebel, R., & Quinn, R. J. (2015). The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery*, *14*(2), 111-129.

Chapter 2. Automated KSα/β Amplicon-based Identification and Chemotyping of Type II Polyketide BGCs

1. Introduction

Much of the structural diversity of type II polyketides (PK-IIs), such as the antibacterial tetracycline and the anticancer agent doxorubicin, is the result of sequence divergence in the ketosynthase α and β subunits (KSα/β) of type II polyketide synthases (PKSs). Each KSα/β enzyme pair selects a specific starter unit (either acetate or non-acetate) and determines the number of extension cycles employed to form a poly-β-ketone product of defined length ranging from 16 to 30 carbons. This product serves as the substrate for subsequent cyclization and tailoring steps, resulting in the final product with a broad scaffold diversity and functional group complexity as detailed in Chapter 1. Although the gene contents encoding these diverse PK-IIs have displayed a substantial difference, they contained, almost in all cases, the highly conserved KSα/β genes with an identical organization that they were always adjacent and co-directional in various PK-II biosynthetic gene clusters (BGCs) identified in the phylum Actinobacteria to date. Therefore, several degenerate primer pairs (Figure 2-1) were designed for the amplification of a partial region of the highly conserved KSα/β genes through polymerase chain reaction (PCR) either from the genomes of isolated Actinobacteria or environmental DNA libraries.

The Seow KSβ degenerate primers (Figure 2-1) aimed at acquiring full-length KSβ genes from a panel of soil Actinobacteria and uncharacterized *Streptomyces* strains, which were employed to construct heterologous KSα/β pairs to investigate how the size of the poly-β-ketone product of the minimal PKS is determined[1]. The Metsa-Ketela KSα degenerate primers (Figure 2-1) were designed for rapid preliminary classification of

unidentified bacterial strains carrying type II PKSs. Based on the phylogenetic analysis of amino acid sequences of amplified KSα fragments, spore pigments and antibiotics were obviously distributed in separate clades, and the antibiotics clade could further be divided into separate branches[2].

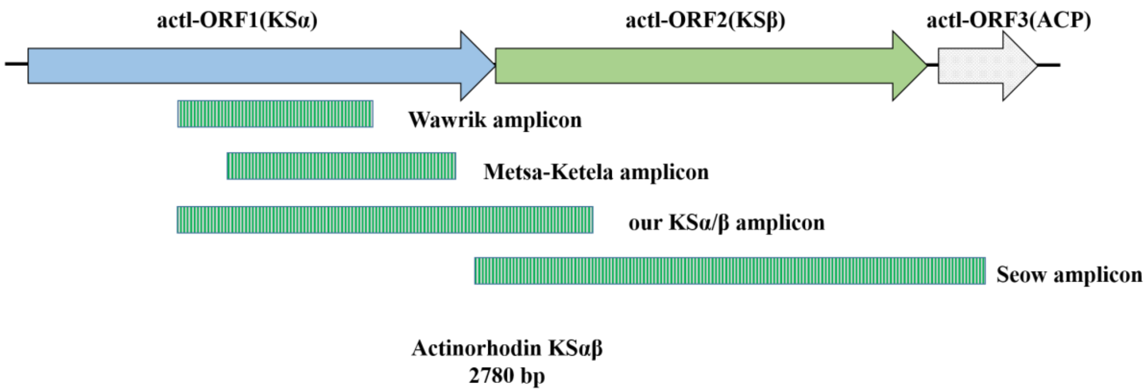

Figure 2-1. Minimal PKS genes of actinorhodin gene cluster and various KSα/β amplicons. The Seow primers amplify the full-length of KSβ gene, while the Metsa-Ketela primer pair and Wawrik primer pair amplify partial KSα gene. The new degenerate primers developed in this study amplify approximately 730 bp of KSα and 430 bp of KSβ genes.

Comparative analysis between the KSα and 16S rRNA gene phylogenetic trees of various *Streptomyces* species revealed that the genus *Streptomyces* was capable of encoding diverse type II PKS gene clusters and the traditional marker like 16S rRNA gene was an inadequate predictor for the types of PK-IIs produced because of the potential horizontal transfer of PK-II biosynthetic genes within *Streptomyces*[3]. The Wawrik KSα degenerate primers (Figure 2-1) were designed for the direct identification of soil samples containing novel and unique type II PKS pathways based on terminal restriction fragment length polymorphism (TRFLP) analysis of both 16S rRNA gene fragments and KSα amplicons[4]. Recently, the Seow KSβ degenerate primers were extended to mine soil metagenomic libraries and the phylogenetic analysis of environmental DNA (eDNA)-derived KSβ amplicons revealed a high percentage of sequences with low similarity to KSβ

from known type II PKS gene clusters[5], suggesting that a large portion of PK-II functional diversity remains to be explored. Indeed, this KSβ amplicon phylogeny-guided mining of eDNA libraries led to the discovery of several novel PK-IIs[6,7,8], in particular pentangular polyphenols such as erdacin[7], calixanthomycin A, and arenimycins[8].

Although these degenerate primer pairs were used for identification of PK-II BGCs, none of them could provide an accurate and detailed prediction for the structure of KSα/β product from sequence. However, a reliable prediction for the structure of poly-β-ketone product is of important significance in the discovery of novel PK-II compounds because the novelty of KSα/β products is a critical indicator for the novelty of final aromatic polyketides, and allows efficient de-replication and prioritization of type II PKS gene clusters for further experimental investigation including structure characterization and bioactivity profiling.

In order to devise a rapid way to predict the structure of KSα/β product, extensive analysis on the crystal structure of the actinorhodin KSα/β subunits[9] and multiple sequence comparison of numerous KSα/β with different poly-β-ketone chain lengths were carried out to investigate whether there was a correlation between the KSα/β structure and/or sequence, and their small molecule product. The 2.0-Å structure of the actinorhodin KSα/β obtained from *Streptomyces coelicolor* was the only one solved in type II PKS systems, and it provided unprecedented insight into the roles of KSα/β subunits. This structural study has shown that polyketides were elongated inside an amphipathic tunnel at the heterodimer interface and led to the proposal that the chain length was regulated by size and shape of the KSα/β active site[9]. Further analysis based on X-ray crystal structure of actinorhodin KSα/β and sequence comparison of several KSα/β with different poly-β-ketone chain

lengths inspired us that the identities of the amino acids lining the active site cavity or in close proximity to it should co-vary with the chemical structure of the poly-β-ketone chain. Indeed, a package in *Dynamite*[10] developed by our group was capable of predicting the structure of poly-β-ketone product from KSα/β amino acids sequence motifs.

Given the successful development of the predictive model in *Dynamite* for KSα/β products, same type of product structure prediction could also be achieved by analyzing the highly predictive amino acid positions of an appropriately positioned KSα/β amplicon (details were described below). This idea inspired us to conceive an approach/route to efficiently identify bacteria that have the genetic capacity to produce potentially novel PK-IIs, but whose genomes have not yet been sequenced.

Here, a combined bioinformatic/experimental route, which we termed "chemotyping", was developed for the rapid identification of KSα/β gene sequences and for the prediction of their product specificities (Figure 2-2). This route involves amplifying a 1.2 kb amplicon comprising portions of both KSα and KSβ genes through PCR using a newly designed degenerate primer pair, sequencing the KSα/β amplicons, and automated extraction and analysis of a 25 amino acid "fingerprint" found to strongly co-vary with product structure in a 78-membered training set of studied KSα/β proteins.

To test the feasibility of this KSα/β amplicon-based identification and chemotyping approach, 16 bacteria were selected from laboratory culture collection as initial test strains, which contained positive and negative controls whose genome information were available. After the validation of the accuracy and reliability of this approach, it was extended to additional unique Actinobacteria from 102 environmental samples — 59 soil samples were collected from Lechuguilla Caverns and 43 samples from Hawaiian lava tubes. While these

environmental sites have been extensively investigated for their microbial diversity, very limited works were carried out in probing secondary metabolites produced by them. Thus, these caves remain to be an underexploited reservoir of bioactive secondary metabolites, since they are extreme environments for the development of rare Actinobacteria.

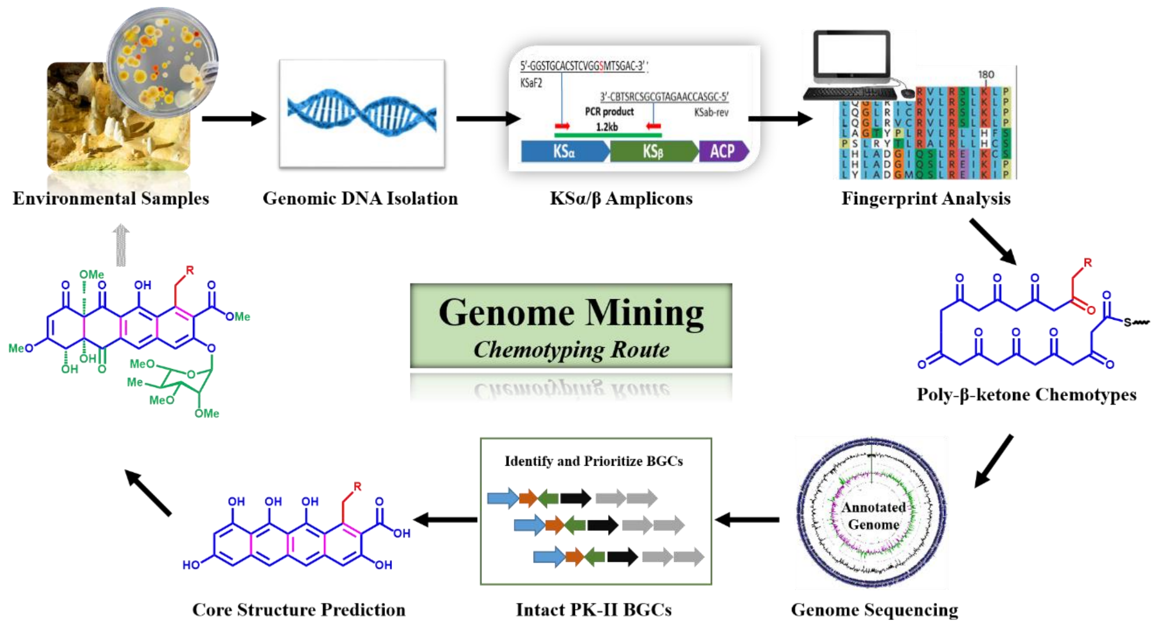

Figure 2-2. The workflow of genome mining approach developed in this study. It starts by isolating genomic DNA from Actinobacteria of interest, then sequencing the KSα/β amplicons obtained from PCR, chemotyping the poly-β-ketone product by analyzing the predictive fingerprint residues in KSα/β amplicon sequences, obtaining the complete PK-II BGCs from sequenced genomes that encode novel KSα/β products, and prioritizing the gene clusters for subsequent compound purification and characterization based on the novelty of predicted core structures.

Two fingerprint analysis models were developed to predict the structure of poly-β-ketone products. One model was based on principle component analysis (PCA) of an array of molecular descriptors[11] representing the structure of amino acids and poly-β-ketone chains. Another model was developed using machine learning algorithm, specifically Metropolis-Hastings sampling[12]. Based on the more accurate predictive model developed using machine learning algorithm, 39 unique KSα/β sequences from 54 environmental and culture collection Actinobacteria were chemotyped and prioritized based on poly-β-ketone

product novelty. Five actinobacterial strains that were rare PK-II producer and/or encoded novel KSα/β products were selected for whole genome sequencing using PacBio RSII third generation single molecule real time sequencing (SMRT) platform[13]. This chemotyping route not only circumvented the reinvestigation of previously identified PK-IIs, but also alleviated the biases introduced by traditional bioactivity-based screening, allowing efficient identification of full repertoire of type II PKS gene clusters, including those cryptic and novel ones.

2. Experimental Materials and Methods

General. The majority of chemical reagents and supplements including media components were purchased from Sigma-Aldrich (St. Louis, MO), VWR (Atlanta, GA), or Fisher Scientific (Pittsburgh, PA) and were used without further purification such as D-(+)-Glucose and Type I Soybean Flour. Exceptions were Difco Luria-Bertani (LB) medium , Miller (Becton-Dickinson (BD), Franklin Lakes, NJ), Tryptic Soy Broth (TSB) medium (Cole Parmer, Vernon Hills, IL), Kanamycin Sulfate (Genlantis, San Diego, CA), Lysozyme (MP Biomedicals, Santa Ana, CA), and Proteinase K (Invitrogen, Carlsbad, CA). Pure water for media and solution preparation was distilled water further purified by Barnstead/Thermolyne HN Ultrapure system. Genomic DNA for molecular cloning was isolated by Microbial DNA Isolation Kit (MO BIO Laboratories, Inc). PCR primers and oligonucleotides were synthesized by Integrated DNA Technologies (abbreviated IDT, Coralville, IA). Restriction enzymes, OneTaq Hot Start DNA polymerase, Phusion Hot Start Flex DNA polymerase, T4 DNA ligase and Calf Intestinal Alkaline Phosphatase (CIP) used for routine molecular cloning were products of New England Biolabs (Ipswich,

MA), except for AmpliTaq Gold DNA polymerase, Phusion High-Fidelity DNA polymerase (Thermo Scientific, Waltham, MA), and GoTaq DNA polymerase (Promega, Madison, WI). DNA marker (1kb plus DNA ladder) was product of Thermo Scientific. All PCR products and restriction enzymes-digested DNA fragments were purified using the DNA Clean & Concentrator Kit or Gel DNA Recovery Kit (both from Zymo Research, Irvine, CA). Plasmids were extracted using the QIAprep Spin Miniprep Kit (Qiagen, Valencia, CA). Routine DNA sequencing service was provided by GENEWIZ (South Plainfield, NJ), while 96-well plate Single Pass Sequencing was performed at Beckman Coulter Genomics (Danvers, MA). Plasmid and DNA sequence analysis was conducted by Vector NTI Advance 11.5 (Life Technologies, Carlsbad, CA). All restriction enzyme digestion were accomplished according to the manufactures' instructions, while ligation was carried out in a 10 μL concentrated reaction. A typical 10 μL reaction mixture was prepared by adding 1 μL of 10x T4 ligase buffer, 1 μL of T4 DNA ligase, 25-30 ng of vector, corresponding amount of inserted DNA fragment (the amount was determined by calculating the molar ratio between vector and inserted fragment, which usually was 1:10), certain amount of H_2O to fill up to 10 μL. The reaction mixture was mixed thoroughly and incubated at 16 °C in a water bath overnight. Agarose gel electrophoresis was often performed with 1% agarose gel. Other general molecular cloning experiments not mentioned here were accomplished using standard molecular cloning methods, protocols unless with specified description.

Plasmids and Vectors. The amplified DNA fragments, including KSα and KSα/β amplicons, were cloned into *Stu*I-digested vector pCR-Blunt (Figure 2-3), which was acquired from Zero Blunt PCR Cloning Kit purchased from Life Technologies (Carlsbad,

CA). This vector contains a lethal gene *ccd*B, whose protein product is toxic in the absence of CcdA protein since CcdB interferes with DNA gyrase, resulting in cell death. This provides positive selection that the expression of *ccd*B gene is disrupted by a correctly inserted DNA fragment. The plasmid pCR-Blunt also harbors a kanamycin resistance gene for positive selection of correct transformants, two *Eco*RI recognition sites for digestion verification of inserted fragments, and two universal primers (M13-Rev and M13-Fwd(-20)) for subsequent Sanger sequencing analysis (Figure 2-3).

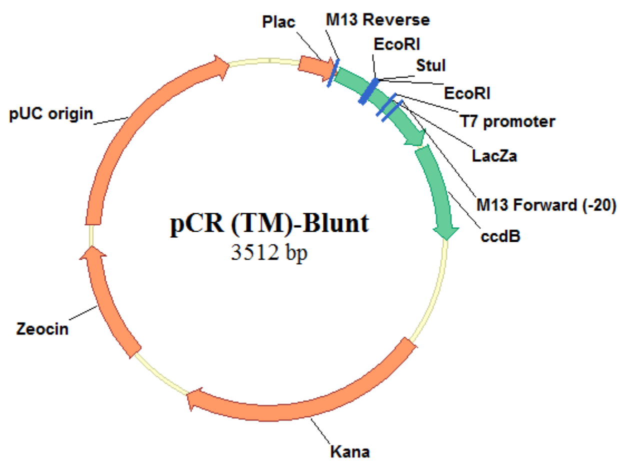

Figure 2-3. The map of vector pCR-Blunt used in this study. It contains a *Stu*I recognition site for insertion of blunt-end DNA fragments.

Bacterial Strains. *E. coli* DH5α was used as a host for routine plasmid construction and reproduction. Laboratory (Lab) strains were gifts from Prof. Hung-wen Liu (University of Texas at Austin), while other 38 unique Actinobacteria were isolated by our group, in collaboration with Dr. Diana Northup from UNM Department of Biology, from 102 soil samples collected from Lechuguilla Caverns, NM and Hawaiian lava tubes.

Instrumentation. The pH values were determined by a CORNING pH meter model 130 purchased from Corning Glass Works (Medfield, MA). OD_{600} values were measured on a CO8000 Cell Density Meter from Denville Scientific Inc. (Holliston, MA). PCR

reactions were performed on a Bio-Rad S1000 thermal cycler and DNA gel was imaged using a molecular imager Gel Doc XR+ (Bio-Rad, Hercules, CA). DNA concentration was quantified by a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA). The concentration of high quality genomic DNA for genome sequencing was also quantified by a Qubit 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA). Agarose gel electrophoresis was carried out on a mini-sub cell GT apparatus from Bio-Rad (Hercules, CA). Centrifugation was performed using either an Avanti J-E centrifuge from Beckman-Coulter (Arlington Heights, IL) for large volumes, or using an Eppendorf 5415C microcentrifuge from Brinkmann Instruments, Inc (Westbury, NY) for small volumes. Bacteria were incubated in either an Innova 42 or Innova 26 incubator/shaker (New Brunswick Scientific, Edison, NJ).

Bacteria Cultivation. *E. coli* DH5α containing pCR-Blunt plasmid was grown in 3 mL of LB broth supplemented with kanamycin (final conc. 50 μg/mL) in a sterile 15 mL conical tube at 37 °C, 250 rpm overnight (15-20 h). *E. coli* DH5α transformants were selected by overnight growth (15-20 h) on LB agar plates containing the kanamycin (final conc. 50 μg/mL) at 37 °C. Actinobacterial strains were routinely maintained in 5 mL of GYM (sometimes in MYM or TSB) liquid medium and allowed to grow at 28 °C or 30 °C, 200-250 rpm, for 3-7 days in 25×150 mm glass culture tubes with about fifteen 4 mm glass beads as needed for preventing clumpy cells. These actinobacterial strains were also grown on GYM agar plates (a few of them were grown on MYM agar plates as well) and incubated at 28 °C or 30 °C for the purpose of strain purification and spore suspension preparation. GYM medium recipe was acquired from DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen, Germany) and their ingredients for 1 liter medium

consisted of 4.0 g of Glucose, 4.0 g of Yeast extract, 10.0 g of Malt extract, 2.0 g of $CaCO_3$ and 12 g of Agar powder. The pH of GYM medium was adjusted to 7.2 before adding agar, and $CaCO_3$ was deleted if liquid medium was used. MYM medium[14] was prepared by the addition of 4 g of maltose, 4 g of yeast extract and 10 g of malt extract into 1 L of deionized water, adjustment of the pH to 7.0, and supplement of 15 g of agar if agar plates were made.

Purification of Environmental Actinobacterial Species. Actinobacteria used in this work were preliminarily isolated from environmental samples collected by previous students in our lab from Lechuguilla Caverns and Hawaiian lava tubes. Further purification and identity confirmation were performed as follows: if an organism has shown a single colony morphology on GYM plate, then it was considered as pure; if an organism has shown a single colony morphology, but had a minor contaminant, then a single colony was re-streaked on GYM plate to get pure; if an organism has shown 2 or more different colony morphologies, then one colony of each morphology was picked up and grown up in liquid GYM medium for genomic DNA isolation. After isolation of genome, the 16S rRNA gene was amplified using universal primers (as described below) and subjected to sequencing analysis to verify their true identity.

Preparation of Spore Suspensions and Frozen Mycelia for Actinobacteria. The pure actinobacterial strains were stored as spore suspensions or mycelia in 20% glycerol in -80 °C freezer. Colonies of actinobacterial species were first made by streaking out on GYM plates and allowed to grow for up to three weeks at 28 °C or 30 °C. To prepare spore suspensions, 3 mL of sterile 20% aqueous glycerol was added to each plate, and the spores were gently scraped off the surface of the plate with sterile Q-tips. The resulting spore suspensions were stored in -80 °C freezer.

To prepare frozen mycelia of actinobacterial strains, a 10 to 50 μL aliquot of spore suspension or a small piece of agar with single colony was inoculated into 5 mL of GYM or TSB or MYM liquid media in 25×150 mm glass culture tubes with about fifteen 4 mm glass beads as needed for separating chunky cells and allowed to grow for 2 to 7 days at 28 °C or 30 °C in a 250 rpm rotary shaker. Cell culture was transferred into 1.5 mL microcentrifuge tubes and centrifuged at room temperature for 1 min at 8,000 *g*. The supernatant was removed and cell pellet was resuspended in 1 mL of sterile 20% aqueous glycerol. The mycelia was keep in -80 °C freezer or -20 °C refrigerator.

Production of Photos for Strains on Agar Plates. The pure Actinobacteria were grown on GYM agar plates for one to three weeks so that their spores were observable, but a few of them did not sporulate. The photos were taken from the front of agar plates with lid off using a black plastic-covered cardboard as background so that the colonies and spores were clearly seen. The photos were produced by a Nikon D750 digital camera, then cropped into 2 x 2 inches square, and merged together in software Photoshop.

Pretreatment of Microbial Cells for Genomic DNA Isolation. In the case that the actinobacterial species formed clumpy cells, a lysozyme pretreatment step was taken to complement the Microbial DNA Isolation Kit (MoBio Laboratories, Inc) for genomic DNA isolation. This pretreatment was carried out by mixing clumpy cells with 1 mL of 25 mM TES solution supplemented with 20 mg/mL lysozyme, incubating at 37 ℃ rotary shaker for 1 h, mixing cells thoroughly and violently by pipette or on a vortex, and passing cells through a 0.45 mm needle (Becton Dickinson). The TES solution was prepared by the addition of 260 μL of 0.96 M Tris-HCl (pH 8.0) and 20 μL of 500 mM EDTA (pH 8.0)

into 9.72 mL of sterile 10.3% sucrose, and sterilization using Corning 250 mL disposable vacuum 0.22 μm filter/storage bottle system (VWR).

General PCR Procedure. Although each DNA fragment was amplified by slightly different PCR procedure, a general one for primer design and PCR amplification of DNA is described here. Two oligonucleotide primers complementary to the template sequence were located at each end of amplified region with varied length, and the restriction enzyme sites used for plasmid construction were added. The rules followed by primer design: first, the primer pair was highly specific to the sequence of amplified region to prevent off-target; second, the melting temperatures (Tm) of the primers were as close to each other as possible to ensure efficient annealing; third, each primer was GC-rich at the 3' end, avoiding base A or continuous identical bases such as GGG or CCC; finally, each primer was check by IDT oligo Analyzer 3.1 to ensure no severe hairpin, self-dimer, and heterodimer occurred. Additionally, if restriction enzyme sites were incorporated into primers, a four-base variable sequence was added to the 5' end of each primer. The PCR amplification of DNA was carried out in 0.5 mL thin-walled microcentrifuge tubes. The applied amount of templates, primers, DNA polymerase, dNTPs, PCR buffer, and DMSO were recommended by the corresponding manufacturers' brochures. The PCR reaction mixture was mixed thoroughly and split into 18-25 μL aliquots in each tube. The PCR program for Phusion DNA polymerase was started with 2 min of denaturation at 98 °C, followed by 30 cycles of incubation at 98 °C for 10 s (denaturation), appropriate annealing temperature (determined by the melting temperatures of primer pair) for 20s, and 72 °C for appropriate time defined by the length of amplified DNA fragments (extension). The program was ended with a 10 min extension step at 72 °C and a constant temperature step

at 4 °C. To note, the extension temperature for OneTaq Hot Start DNA polymerase was 68 °C instead of 72 °C.

Amplification and Phylogenetic Analysis of 16S rRNA Genes. The genomic DNA of purified strains was extracted using Microbial DNA Isolation Kit (MoBio Laboratories, Inc) by following the manufacturer's protocol, and the 16S rRNA gene was amplified with universal primers 8F[15] and 1492R[16] (Table 2-1) using general PCR procedure as described above, which yielded nearly complete 16S rRNA gene sequence (~1,400bp). The OneTaq Hot Start DNA polymerase and annealing temperature 45°C were used in the PCR amplification of 16S rRNA gene. Then the purified PCR products were sent for Sanger sequencing using the same primers, 8F and 1492R.

Primer Name	Sequence (5'-3')	Description
8F	AGAGTTTGATCCTGGCTCAG	amplification of a ~1,400 bp 16 rRNA genes from all environmental samples
1492R	GGTTACCTTGTTACGACTT	

Table 2-1. List of Primers used for PCR amplification of 16S rRNA genes. These primers are the most common primer pair used in phylogenetic studies.

The sequencing results were checked for base calling quality, and the 16S rRNA gene sequences were obtained by assembling the forward and reverse amplicons together by ContigExpress (a software package of VectorNTI). The orientation of 16S rRNA gene sequence was corrected, and the duplicates were removed. After acquisition of correct 16S rRNA amplicons for each unique species, all the sequences were searched for closest relatives in NCBI GenBank database using basic local alignment search tool (BLAST)[17]. Only the 16S rRNA sequences of Actinobacteria were used for the construction of phylogenetic tree displayed in Results and Discussion. The 16S rRNA sequences of strains in Lab Collection were downloaded from NCBI GenBank. If no 16S rRNA sequence for a specific strain was found, the one from the closest species was used. The phylogenetic tree

of 16S rRNA genes was built by three steps: first, multiple sequence alignment of all 16S rRNA sequences was carried out by Clustal Omega[18]; second, an approximately-maximum-likelihood dendrogram based on above alignments of nucleotide sequences was outputted by the software FastTree 2[19]; third, the circular dendrogram was displayed, annotated, and exported as high-resolution figure using the Interactive Tree of Life (iTOL) web interface[20]. Bootstrap analysis was used to evaluate the tree topology. For clear visualization, only the designated names, bootstrap values, and color-coded suborders were shown in the high-resolution graph, while the detailed information pertaining to each strain were compiled into the table 2-6 in Results and Discussion.

Degenerate Primers Design. In order to identify most of type II PKS gene clusters from isolated genomic DNA, degenerate primers should be designed at the highly conserved region of KSα/β genes and annealed on various KSα/β sequences with least bias in PCR. To this end, the protein sequences of 78 actinobacteria type II KSαβ genes (regarded as training set, see table 2-5 below) known to encode various poly-β-ketone products were obtained from NCBI GenBank and performed multiple sequences alignment using Clustal Omega[18]. The highly conserved regions of KSα/β proteins were recognized from this alignment, and degenerate primers were designed based on the sequences in these regions. A degenerate primer pair, KSaF2 and KSaR1 (Figure 2-4a), for the amplification of a KSα fragment were first designed to identify the Actinobacteria that encode PK-II BGCs. The sequence of KSaF2 was designed based on a conserved region at the position near active site of KSα subunit, while KSaR1 was designed based on the conserved sequences at the C-termini of KSα subunit. This degenerate primer pair would amplify a ~550 bp DNA fragment (i.e. about 180 amino acids would be amplified) from the KSα

gene, corresponding to amino acid positions between 168 of actI-ORF1 (KSα) and 352 of actI-ORF1 (KSα) from the actinorhodin pathway from *Streptomyces coelicolor* A3(2) (Figure 2-4a). Later, new degenerate primers, KSab-fwd and KSab-rev (Figure 2-4b), were designed for both PK-II BGC identification and KSα/β amplicon chemotyping. The sequence of KSab-fwd was also designed based on a conserved region at the position near active site of KSα subunit, but a degenerated base was changed from S to I within KSaF2 primer, which would be less stringent. In parallel, the reverse primer KSab-rev was designed based on the conserved sequences in the middle of KSβ subunit. A ~1.2 kb fragment located in the middle of KSα/β genes, approximately 730 bp of KSα and 430 bp of KSβ genes, corresponding to amino acid positions between 168 of actI-ORF1 (KSα) and 140 of actI-ORF2 (KSβ), was amplified by this newly designed degenerate primer pair (Figure 2-4a).

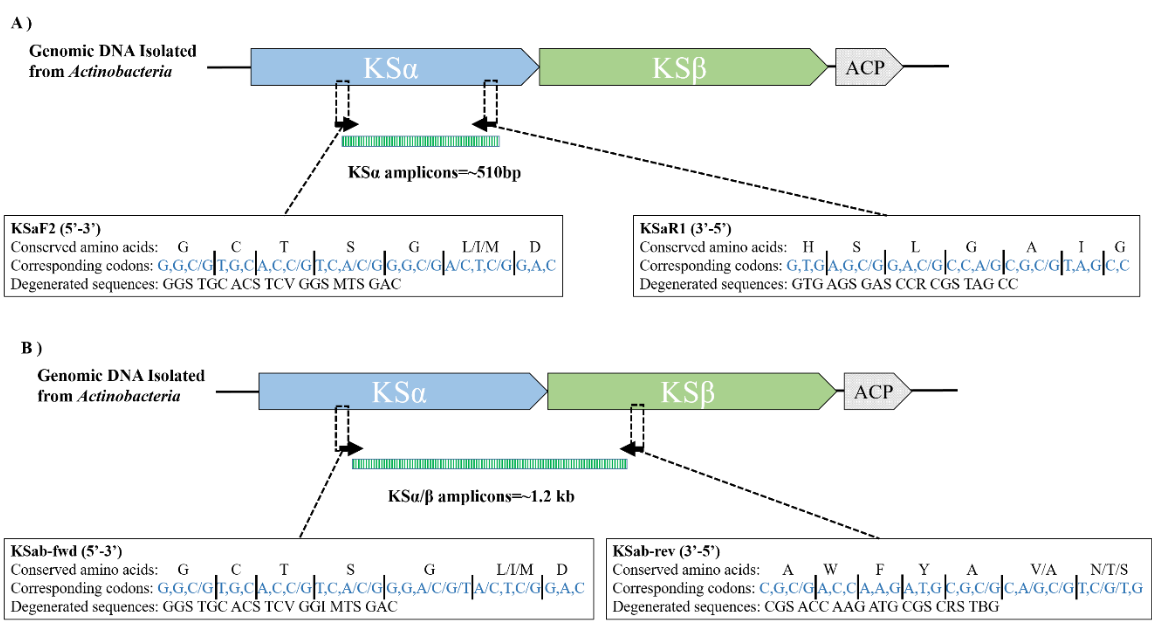

Figure 2-4. Degenerated primer sequences and their annealing sites in minimal PKS clusters. KSα represents β-ketoacyl synthase alpha subunit, while KSβ represents β-ketoacyl synthase beta subunit. ACP represents acyl carrier protein.

The criteria for the design of the degenerate primers were set as follows: the oligonucleotide sequence is able to accommodate all possible codons of all amino acid residues at that site; the region rich in amino acids that only have one or two possible codons are preferred; degeneracy at the 3' end of the primer need to be avoided; no more than 20% of each primer is degenerated nucleotides; the amplified fragments should contain the key residues for subsequent fingerprint analysis. The best primer pair was back-translated from the amino acid sequences by using preferred codon usage in Actinobacteria.

Amplification of KSα and KSα/β Amplicons. The KSα amplicons were first amplified with primers, KSaF2 and KSaR1, using AmpliTaq Gold DNA polymerase. The genomic DNA of *Kibdelosporangium sp.* MJ126-NF4 was used as positive control since it was known to harbor four PK-II BGCs in sequenced genome. Due to the products from the first PCR had a 3' –dA tail and very low amount of DNA, second PCR using the same primer pair and Phusion DNA polymerase was carried out to obtain enough amount of KSα amplicon fragments with blunt ends. The template for second PCR was obtained by cutting and recovering the band with correct size (~510 bp) on agarose gel using Gel DNA Recovery Kit. The PCR product of second PCR was also purified by running agarose gel electrophoresis and gel band recovery. Then the DNA fragment was ligated into *Stu*I-digested vector pCR-Blunt, which was transferred into *E. coli* DH5α. The transformants were selected on LB agar plates containing 50 μg/mL kanamycin. Five single colonies for each strain were picked for subsequent plasmid extraction.

To effectively amplify the KSα/β amplicons, above protocol was modified as follows. First, the OneTaq Hot Start DNA polymerase was used to substitute AmpliTaq Gold DNA polymerase because it was suitable for PCR with GC-rich templates. Second,

various amount of template, different annealing temperature (51 ℃ to 60 ℃), and different primer combination (KSaF2 versus KSab-fwd paired with KSab-rev) were tested. The 5 ng of template per reaction, 51 ℃ annealing temperature, and primer pair (KSaF2 and KSab-rev) were determined to be the optimal PCR condition (Table 2-2). Third, the correct transformants were directly subjected to colony PCR screening, which could be scalable and high-throughput compared with the way using plasmids extraction and restriction enzyme digestion.

First PCR	**Amount (µL)**
ddH_2O	18
5x GC Reaction Buffer	5
dNTPs 2.5 mM each	0.5
50 µM KSaF2	0.25
50 µM KSab-rev	0.25
5 ng/µL genomic DNA	1
OneTaq DNA polymerase	0.125
In total	**25 µL**

First PCR Program	
94 ℃	2 min
94 ℃	30 s
51 ℃	40 s
68 ℃	1 min 30 s
30 cycles	
68 ℃	10 min
4 ℃	forever

Second PCR	**Amount (µL)**
ddH_2O	36
5x Phusion Buffer	10
dNTPs 2.5 mM each	1
50 µM KSaF2	0.5
50 µM KSab-rev	0.5
2 ng/µL genomic DNA	1
DMSO	1.5
Phusion DNA polymerase	0.5
In total	**50 µL**

Second PCR Program	
98 ℃	1 min
98 ℃	20 s
51 ℃	35 s
72 ℃	1 min 30 s
30 cycles	
72 ℃	10 min
4 ℃	forever

Table 2-2. The optimal condition for first PCR. Although this optimal condition for first PCR was able to amplify most KSα/β amplicons from various strains, slight adjustments (e.g. template amount, annealing temperature) were required for few strains to obtain better results such as larger amount and less non-specific bands.

Plasmid Mini-Preparation. Complementary to the QIAprep Spin Miniprep Kit, plasmids were also isolated using an alkaline lysis protocol[21] as follows. Four mL of overnight (18 to 20 h) culture of *E. coli* DH5α was collected and centrifuged at 12,000 *g* for 1 min. The cell pellets were resuspended in 250 µL of Buffer 1 consisting of 50 mM Tris-HCl, 10 mM EDTA and 100 µg/mL RNase, adjusting pH to 8.0. Then 250 µL of

Buffer 2 (0.2 M NaOH and 1% SDS) was used to lyse the cells by inverting the tube 6-8 times. To lower down the pH, 350 μL of ice-cold Buffer 3 (3 M potassium acetate, pH 5.5) was added into the solution and the tube was inverted 6-8 times. Then it was centrifuged at 13,000 *g* for 5 min, and the supernatant was transferred to a clean 1.5 mL microcentrifuge tube. Equal volume of isopropanol (400 μL) was added to precipitate the nuclei acids by briefly mixing it on a vortex. After the DNA precipitation, it was centrifuged at 13,000 *g* for 5 min to take DNA pellet down to the bottom, and the supernatant was discarded carefully. The DNA pellet was washed twice by 500 μL of pure ethanol and dried in the hood for 15 min. Finally, the plasmid was dissolved in 40 μL of nuclease free water and ready for use.

KSα and KSα/β Amplicons Sequencing. Prior to DNA sequencing, the plasmids with correctly inserted KSα amplicons were subjected to digestion verification using restriction enzyme *Eco*RI, giving an band for the backbone (about 3,500 bp) and an band for KSα fragment (about 510 bp) on agarose gel. Plasmids showed correct *Eco*RI digestion pattern were sent for Sanger sequencing at GeneWiz.

Primer Name	Sequence (5'-3')	Description
XW-fwd	TACACTTTATGCTTCCGGCTCG	used in colony PCR screening to amplify a ~1,550 bp fragment from single colonies on transfromation plates
XW-rev2	CGGGCCTCTTCGCTATTACG	

Table 2-3. List of Primers used for colony PCR screening. The DNA fragments obtained using this primer pair contain the priming sites for subsequent Sanger sequencing.

To sequence the KSα/β amplicons, an alternative way was adopted to accommodate the scalable and high-throughput requirements. Instead of plasmids isolation and subsequent *Eco*RI digestion verification, two primers, XW-fwd and XW-rev2 (Table 2-3), were designed to undergo colony PCR to rapidly screen out the single colonies with correct inserts. This primer pair was located on the vector: without inserts, it would give a ~400

bp band; with correct inserts, it would give a ~1.55 kb band. The colony PCR was carried out using GoTaq DNA polymerase at annealing temperature 62 ℃ in an 18 μL reaction solution. Next, the PCR products showing correct inserts were subjected to following 96-well plate sequencing using internal sequencing primers, M13-fwd (-20) and M13-rev.

96-well Plate High-throughput Sequencing. To meet the requirement for sequencing a large number of KSα/β amplicons, Single Pass Sequencing service from Beckman Coulter was used. This high-throughput Sanger sequencing platform involved purification of colony PCR products using SPRI technology, DNA sequencing using BigDye Terminator v3.1, and sequence delineation on an ABI PRISM 3730xI with base calling and data compilation. After estimation of the concentration of colony PCR products (about 35 ng/μL) by column purification using the DNA Clean & Concentrator Kit, 12 μL of PCR product of each KSα/β amplicon was diluted into 23 μL with nuclease free water and sent for sequencing in AB-0800 96-well full skirted plates with AB-0784 flat 8-strip caps (Thermo Scientific). The sequencing primers were universal primers, M13-fwd and M13-rev, provided by Beckman Coulter.

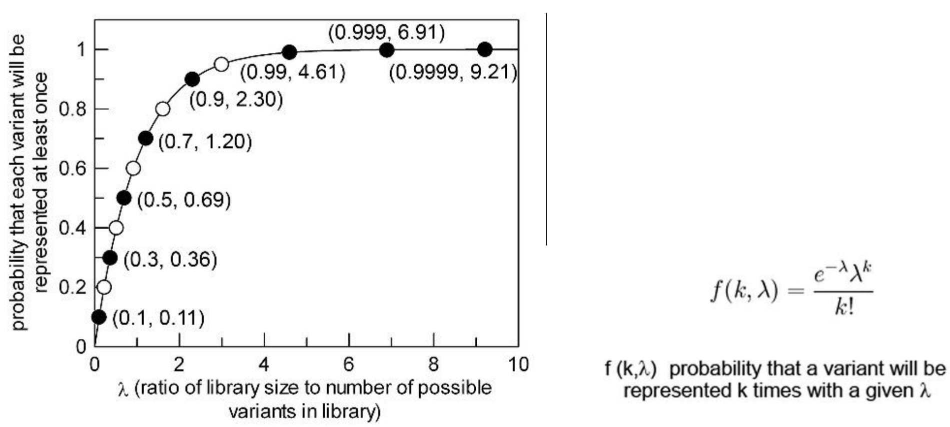

Figure 2-5. Statistical model for determination of the number of colonies for sequencing. For example, if a strain only contains two PK-II BGCs, 4.61 (λ) times 2 colonies, round to 10, need to be sequenced to reach a confidence of 99%. This model was obtained from Dr. Melançon’s personal manuscript.

Most of KSα/β amplicons were sequenced using this high-throughput pipeline. Since numerous actinobacterial species have been shown to harbor more than one PK-II BGCs, a number of clones should be sent for sequencing analysis. Based on a statistical model (Figure 2-5), 5 to 25 colonies from different strains were subjected to this high-throughput sequencing analysis to obtain all possible KSα/β amplicons in each species.

KSα/β amplicon name	Primer sequence (5'-3')		DNA length (bp)	Probed strains
KS2	Forward	GAGGGAGTGGCGTACCGACACAGTG	388	Lab01, Lab05, Lab11, Lab12
	Reverse	CCATGTCGAACTCGGGCAGTTCACTGG		
KS4	Forward	GGAGGGATCGGCGGACGTGATGATC	736	Lab02
	Reverse	GTCCTTCGGTTCGCTCATCGTGTGTCC		
KS5	Forward	GTCACTGCTGGCGATCGAGCACAACG	481	Lab02
	Reverse	CGTAGTCGGGCAGGGTGTTCTCGTCG		
KS7	Forward	GGAGTGAGCCGATGAAACCGTCCACAG	313	Lab05, L-34
	Reverse	CTCATGTCGTAGCCGTCGAGCTGGTC		
KS8	Forward	GATGATCTTCCGGCGGCTCAGGGAG	356	Lab05, Lab11
	Reverse	GACGAGCTCGCCATGATCACCGAGAC		
KS13	Forward	GAGGTAGCGGCATGAGCCAGGCGAC	346	Lab08, Lab10
	Reverse	CTCGAAACCGCCCTGGGAGTTGGAG		
KS20	Forward	GTGGCTTCCAGTCGGCCATGGTGTTCC	402	Lab12
	Reverse	GAACTGGCGAGGGTGACCGCGTACC		
KS25	Forward	CTGGAGGACCTGGAGTCGGCCAAGG	569	L-07
	Reverse	CCACGTGTTGGTGTCTTCGGTGAGAGC		
KS28	Forward	CGTGCGTACGGTCGTCACCGGTATC	269	L-14, L-32;
	Reverse	GGATCGATGCCCGCGTCCTTGATCG		
KS32	Forward	CCTGGACGAATACTGGCGCACCACC	283	L-03, L-06, L-22, L-26, L-35, L-36, L-39
	Reverse	CAAACTCGAAGCCGCCCGAACTGCTC		
KS34	Forward	CCGAAAGGAAGATCGCATGAGCGCCTC	381	L-34, L-38, L-83
	Reverse	CACAGCCGCTGGAGTTCGTTCTGGC		
KS37	Forward	CTACGCCACCAGGGCCAACGCCTATC	465	L-03, L-06, L-22, L-26, L-35, L-36
	Reverse	CTCCCTTCGGGGTGGTGAGCACCATG		
KS57	Forward	CGTCGGGTACGCCTATCACGCGGTC	748	L-72
	Reverse	CGCTCATCGCAGCCCTCCTTCACTGC		
KS58	Forward	GCGAGGGAAGTGCGGATGTGATGGTC	476	L-72
	Reverse	CGATTTGACCGAGCTGATCGGCACGTC		
KS59	Forward	CAACGACGAGCCGGAACCGTAAAGTGG	634	L-72
	Reverse	GAAGGCGCGAACTCGTCCTCATGCC		
BE7585A	Forward	CACCGGGTCCCGATCAGCTCCGTG	608	Lab04
	Reverse	GAACGACGGCAGCTCGTCCGGGTTC		
Oviedo	Forward	GAGAGGAGCGTGGCATGACCGGCAC	324	Lab02, 1st PCR, 2nd PCR product
	Reverse	CCATGTCGAACTCGGGCAGCTCCTGC		

Table 2-4. List of primers used for identification of true host for specific KSα/β amplicons. These primers were used to identify and confirm the true host of several KSα/β amplicons.

Verification of True Hosts for KSα/β Amplicons. Owing to the cross-contamination during the shipping of 96-well plates and the issue of duplicate strains, identical KSα/β amplicons were obtained from different strains, which was very unlikely. To resolve this issue and identify the true hosts for these KSα/β amplicons, a number of specific primers (Table 2-4) were designed based on the sequences of these KSα/β amplicons. Then each primer pair was used to probe the genomic DNA of strains containing the identical KSα/β amplicon. The PCR were carried out using general PCR condition as described above.

Preparation of E. coli Competent Cells and Transformation. Competent cells of *E. coli* DH5α were made by the commonly used rubidium chloride (RbCl) method[22]. This protocol began with formation of *E. coli* DH5α single colonies by streaking out on a LB agar plate. Then a single fresh colony was inoculated into 3 mL of the LB liquid medium, which grew overnight (17-20 h) at 37 °C, 250 rpm. A 500 μL of the overnight culture was transferred to 50 mL of LB liquid culture in a 250 mL Erlenmeyer flask, which was placed in 37 °C, 250 rpm incubator. When reaching an OD_{600} of approximately 0.4 to 0.6 (about 2 to 2.5 h long), all the cell culture was transferred to a pre-chilled sterile conical tube and incubated on ice for 30 min. After centrifugation (pre-chilled for 5 min) at 4 °C, 3,000 *g* for 5 min, the supernatant was discarded by gentle decant, and the cell pellet was resuspended in one-third of the original culture volume (e.g. 17 mL if 50 mL original culture) of ice-cold RF1 solution (100 mM RbCl, 15% glycerol, 50 mM $MnCl_2$, 30 mM potassium acetate, 10 mM $CaCl_2$, pH 5.8, sterilized by filter equipped with vacuum to draw the solution through a 0.22 μm membrane). After incubation on ice for 15 min, the cell suspension was centrifuged at 3,000 *g*, 4 °C for 5 min, and the resulting cell pellet was resuspended in two twenty-fifths of the original culture volume (e.g. 4 mL if 50 mL original

culture) of ice-cold RF2 solution (10 mM RbCl, 10 mM MOPS, 75 mM $CaCl_2$, 15% glycerol, pH 6.8, sterilized by filter equipped with vacuum through a 0.22 μm membrane). The cells were split into 100 μL aliquots in pre-chilled microcentrifuge tubes and kept frozen at -80 °C. Each tube of 100 μL *E. coli* competent cells was used for one transformation, in which 1 to 20 ng of plasmids or 5 μL of ligation products (DNA <50 ng) were added. The resulting mixture was placed on ice for 30 min, followed by heat shock at 42 °C in a water bath for 90 s, and then placed on ice again for 2 min. One mL of liquid LB was added into the *E. coli* mixture to allow the antibiotic genes being expressed in 37 °C shaker for 1 h of incubation. After 1 h of growth, the *E. coli* mixture was centrifuged at 8,000 *g* for 1 min, and the supernatant was removed to retain ~150 μL of LB. The cells at the bottom of tube were resuspended by the remaining LB, and spread on LB agar plate with corresponding antibiotics for selection of correct transformants.

High Quality Genomic DNA Extraction for Genome Sequencing. Twenty-five mL cell culture of target strains with high density was prepared (about 1 gram wet weight after centrifugation). Cell culture was centrifuged at 4,000 *g*, room temperature for 5 min, and the supernatant was discarded, while the cells were resuspended in 10 mL of 50 mM Tris-HCl, pH 8.0. This resulting cell suspension was transferred to a 40 mL tissue grinder for thoroughly homogenization to ensure that the clumpy cells were well dispersed. Then cell pellets were obtained by centrifugation (4,000 *g*, room temperature for 5 min) and split into 200 mg aliquots in each 1.5 mL microcentrifuge tube. The cell pellets of each aliquots were transferred to a porcelain mortar, treated by liquid nitrogen, and grinded by porcelain pestle. This liquid nitrogen treatment and grinding step was very critical for the final yield and usually repeated twice to achieve adequate cell breakage (looks like cell powder). If

final yield of genomic DNA was low, the cell would be increased up to 450 mg, and rounds of liquid nitrogen treatment and grinding would be increased as well. The resulting cells were washed off from the mortar using about 900 μL of prepared solution 1 (10 % sucrose, 50 mM Tris-HCl, pH 8.0, 10 mM EDTA-2Na) and split into two portions of 450 μL cell solution in 1.5 mL microcentrifuge tubes. The cells were further lysed by 50 μL of freshly prepared lysozyme solution (3 mg lysozyme powder dissolved in solution 1) at 37 °C, 250 rpm, for 1 h, followed by adding 10 μL of Proteinase K (1 mg powder dissolved in solution 1) and 50 μL of 10% SDS, which was incubated at 55 °C in a water bath for 1 h. After this critical digestion step, 500 μL of PCl (Phenol: Chloroform: Isoamyl Alcohol in the ratio 25: 24: 1, which is saturated with 10 mM Tris, pH 8.0, and 1 mM EDTA in 100 mL) was applied to extraction of DNA and RNA by inverting tubes violently for 10 times and mixing them on a vortex for 30 s. Each sample was centrifuged at 14,000 rpm for 5 min, and the viscous supernatant was carefully transferred to new tubes. This PCl treatment and centrifugation process was repeated. Next, 500 μL of chloroform was added to the supernatant to remove the residual PCl by inverting tubes violently for 10 times and mixing them on a vortex for 30 s, followed by centrifugation (14,000 rpm, 5 min). The supernatant was carefully transferred into new tubes, and 700 μL of 4 °C pre-chilled isopropanol was employed to each sample to precipitate the genomic DNA, which was centrifuged to the bottom of the tube at 3,000 *g* for 10 s. The supernatant was removed, and 1 mL of 4 °C pre-chilled 70% ethanol was added to wash the genomic DNA by gently inverting 10 times. The genomic DNA from two tubes were centrifuged at 3,000 *g* for 10 s and combined together, while the supernatant was discarded. Next, the genomic DNA was dried by opening cap of tube and placing in 37 °C incubator for 15 min. Then 500 μL of TE buffer

(10 mM Tris-HCl, 1 mM EDTA-2Na) containing 0.1 mg/mL of RNase A was added into the tube and hold at 4 °C overnight to remove the RNA from the genomic DNA sample.

After overnight RNase A treatment, the genomic DNA was further purified by one round of PCl and chloroform processing, isopropanol precipitation, 70% ethanol wash, and drying at 37 °C as described above. Finally, the high quality genomic DNA was dissolved in 100 μL of Tris-HCl (10 mM, pH 8.0) and stored at 4 °C.

The concentration of isolated genomic DNA adopting this protocol were quantified using both NanoDrop spectrophotometer and Qubit fluorometer. Qubit fluorometer usually gave more accurate measurement because the concentration of the genomic DNA sample was reported by a fluorescent dye that emits a signal only when bound to double-stranded DNA molecules, which differentiate them from single-stranded RNA and minimizes the interference of contaminants in the sample, including degraded DNA or RNA. Then the samples were ready for downstream genome sequencing.

Bioinformatics Analysis. The *Dynamite* developed by Ben Yackley of our group used a series of protein sequences as query to identify many conserved protein families in type II PKS gene clusters in NCBI protein databank using the Blastp algorithm. Summaries of the features (including fingerprint analysis, specific combination of immediate tailoring enzymes) of all gene clusters were outputted as text files that could be easily viewed and further analyzed by user.

The KSα/β amplicon sequences obtained from sequencing were first checked for base calling quality, and those ambiguous bases were corrected by carefully examining their sequencing chromatograph. Each sequence was trimmed by removing the degenerate primer sequences.

Identified Type II Polyketides	Producer Species	KSα GI# NCBI Protein NO	KSβ GI#	starter unit	extender units
A-74528	Streptomyces sp. SANK 61196	296046088	296046089	hexadienyl	12
aclacinomycin/aclarubicin	Streptomyces galilaeus	7800665	7800666	propionyl	9
aclacinomycin/aclarubicin	Streptomyces galilaeus	16945714	16945715	propionyl	9
actinorhodin	Streptomyces coelicolor A3(2)	21223458	21223459	acetyl	7
alnumycin	Streptomyces sp. CM020	209863916	209863917	butyryl	7
aranciamycin	Streptomyces echinatus	118722503	118722502	acetyl	9
arimetamycin	uncultured bacterium	558613787	558613788	acetyl	9
arixanthomycin	uncultured bacterium	613432370	613432371	acetyl	12
azicemicin	Kibdelosporangium sp. MJ126-NF4	282801740	282801741	aziridinyl	9
BE-7585A	Amycolatopsis orientalis subsp. vinearia	298256334	298256335	acetyl	9
benastatin	Streptomyces sp. A2991200	169402965	169402966	hexanoyl	11
calixanthomycin	uncultured bacterium	745698432	745698431	acetyl	12
chartreusin	Streptomyces chartreusis	68146474	68146475	acetyl	9
chattamycin	Streptomyces chattanoogensis	700746519	700746520	acetyl	9
chelocardin	Amycolatopsis sulphurea	568402357	568402358	malonamyl	8
chlortetracycline	Streptomyces aureofaciens	338776764	338776763	malonamyl	8
chromomycin	Streptomyces griseus subsp. griseus	40644834	40644833	acetyl	9
chrysomycin	Streptomyces albaduncus	266631088	266631089	propionyl	9
cinerubin	Streptomyces sp. SPB74	197695599	197695598	propionyl	9
cosmomycin (partial)	Streptomyces olindensis	640937564	640937565	propionyl	9
dactylocycline	Dactylosporangium sp. SC14051	408451285	408451286	malonamyl	8
daunorubicin/doxorubicin/ daunomycin	Streptomyces peucetius	532245	532246	propionyl	9
daunorubicin/doxorubicin/ daunomycin (partial)	Streptomyces sp.	516109	516110	propionyl	9
elloramycin	Streptomyces olivaceus	15848282	15848283	acetyl	9
enterocin	Streptomyces maritimus	8926190	8926191	benzoyl	7
erdacin	uncultured soil bacterium V167	261497157	261497158	acetyl	7
fasamycin (AZ154)	uncultured bacterium	343479100	343479099	acetyl	12
FD-594	Streptomyces sp. TA-0256	316997093	316997094	butyryl	12
fluostatin	uncultured bacterium BAC AB649/1850	332380592	332380591	acetyl	9
frankiamicin	Frankia sp. EAN1pec	158109628	158109629	acetyl	11
fredericamycin	Streptomyces griseus	33327096	33327097	hexadienyl	12
frenolicin	Streptomyces roseofulvus	3170577	3170578	acetyl/butyryl	7
gilvocarcin	Streptomyces griseoflavus	32140283	32140284	propionyl	9
granaticin	Streptomyces violaceoruber	4218564	4218565	acetyl	7
granaticins	Streptomyces vietnamensis	308445212	308445213	acetyl	7
grincamycin	Streptomyces lusitanus	514389165	514389166	acetyl	9
griseorhodin	Streptomyces sp. JP95	21039488	21039489	acetyl	12
griseorhodin	Streptomyces sp. CN48+	662748189	662748190	acetyl	12
griseusin (partial)	Streptomyces griseus	581665	581666	acetyl	9
hatomarubigin	Streptomyces sp. 2238-SVT4	296178419	296178421	acetyl	9
hedamycin	Streptomyces griseoruber	32492544	32492543	hexadienyl	9
jadomycin	Streptomyces venezuelae ISP5230	510722	510723	acetyl	9
kinamycin	Streptomyces murayamaensis	29469233	29469234	acetyl	9
kosinostatin	Micromonospora sp. TP-A0468	387134545	387134546	acetyl	9
lactonamycin	Streptomyces rishiriensis	161367388	161367389	glycyl	9
lactonamycin (partial)	Streptomyces sanglieri	161367423	161367424	glycyl	9
landomycin	Streptomyces cyanogenus	4240405	4240406	acetyl	9
lomaiviticin	Salinispora pacifica strain DPJ-0016	634794954	634794955	propionyl	9
lomaiviticin	Salinispora pacifica strain DPJ-0019	573017216	573017217	propionyl	9
lysolipin	Streptomyces tendae	154623217	154623216	acetyl	12
medermycin	Streptomyces sp. AM-7161	32469270	32469271	acetyl	7
mithramycin	Streptomyces argillaceus	927517	927518	acetyl	9

Table 2-5 (cont.)

naphthocyclinone (partial)	Streptomyces arenae	4416222	4416223	acetyl	7
nivetetracyclates	Streptomyces sp. Ls2151	556715479	556715477	propionyl	9
nogalamycin	Streptomyces nogalater	2916812	2916813	acetyl	9
oviedomycin	Streptomyces antibioticus	46237518	46237519	acetyl	9
oxytetracycline	Streptomyces rimosus	73621271	73621272	malonamyl	8
PD 116740	Streptomyces sp. WP 4669	29469252	29469253	acetyl	9
polyketomycin	Streptomyces diastatochromogenes	224812396	224812397	acetyl	9
pradimicin(partial)	Actinomadura hibisca	120431566	120431567	acetyl	11
pradimicin	Actinomadura hibisca	2580442	2580443	acetyl	11
R1128	Streptomyces sp. R1128	11096114	11096113	acetyl; propionyl; isobutyryl; butyryl	7
ravidomycin	Streptomyces ravidus	268322287	268322286	propionyl	9
resistomycin	Streptomyces resistomycificus	45259316	45259317	acetyl	9
rubromycin	Streptomyces collinus	9944994	9944995	acetyl	12
saquayamycin/galtamycin	Micromonospora sp. Tu 6368	227121321	227121322	acetyl	9
Sch 47554	Streptomyces sp. SCC 2136	88319793	88319792	acetyl	9
SF2575	Streptomyces sp. SF2575	292659136	292659137	malonamyl	8
simocyclinone	Streptomyces antibioticus	12744820	12744821	acetyl	9
steffimycin	Streptomyces steffisburgensis	84619196	84619195	acetyl	9
tetarimycin	uncultured bacterium	426272821	426272820	acetyl	9
tetracenomycin	Streptomyces glaucescens	153496	153497	acetyl	9
TLN-05220/TLN-05223	Micromonospora echinospora subsp. challisensis	283484105	283484106	2-methyl butyryl	12
urdamycin	Streptomyces fradiae	809105	809106	acetyl	9
WhiE spore pigment	Streptomyces coelicolor A3(2)	21223681	21223680	acetyl	11
WhiE spore pigment(partial)	Streptomyces coelicolor A3(2)	5139588	5139589	acetyl	11
X26	uncultured bacterium	343479142	343479141	acetyl	9
xantholipin	Streptomyces flavogriseus	292386134	292386133	acetyl	12

Table 2-5. The 78-membered training set used for KSα/β amplicon chemotyping. These type II PKS gene clusters were identified by *Dynamite*. GI number is the protein ID in NCBI database.

The forward and reverse fragments were assembled into contigs, which were exported into FASTA file for subsequent fingerprint analysis. The KSα/β amplicon sequences for the 78 training set members were obtained by first retrieving the full-length KSα and KSβ protein sequences from NCBI databank based on their GI number as shown in Table 2-5, then by trimming them into KSα/β amplicons guided by the amino acid sequences of primers, KSaF2 and KSab-rev.

The phylogenetic tree analysis of KSα/β amplicons was similar to that of 16S rRNA as described above, but the amino acid sequences of KSα/β amplicons were used, and the

outgroup was a pseudo-dimer sequence of ketosynthase I (FabB) from the *E. coli* fatty acid biosynthetic pathway. The web-interface program developed by Yasushi Ogasawara of our group was designed for automated fingerprint analysis. The input was the nucleotide sequence of assembled KSα/β amplicon, which was translated into amino acid sequence using all six possible reading frame. These amino acid sequences were blast against KSAo and KSBo to identify the correct reading frame to output KSα and KSβ amino acid sequences, respectively. KSAo and KSBo are sequences identified to give largest matches to other KSα/β genes in multiple sequence alignment of 650 KSα/β gene sequences and regarded as the center of all KSα and KSβ sequences. The resulting KSα and KSβ amino acid sequences were aligned with actI-ORF1 (KSα) and actI-ORF2 (KSβ) of actinorhodin gene cluster from *Streptomyces* coelicolor A3(2) because the crystal structure of actinorhodin KSα/β is the only one solved. Based on the alignment, the fingerprint residues at corresponding positions of actI-ORF1 (KSα) and actI-ORF2 (KSβ) were identified, as well as the closest match to 78 actinobacteria type II KSα/β protein sequences (training set members in Table 2-5) whose poly-β-ketone products were known. The similarity between the closest match and query fingerprint residues was calculated using BLOSUM62[23].

3. Results and Discussion

Recovery of Actinobacterial Strains from Environmental Samples and Culture Collections. In order to test the feasibility of the conceived approach that would identify and prioritize PK-II BGCs through amplifying the defined KSα/β amplicons from isolated genomes using degenerate primers, automatically extracting and analyzing the highly predictive fingerprint residues, and predicting the poly-β-ketone structures, a batch of 16

unique Actinobacteria strains from laboratory culture collections were chosen as initial test set. Five strains from this batch were used as positive controls since their sequenced genomes or chromosomal segments are known to harbor type II PKS gene clusters, while 3 strains would be negative controls that no PK-II gene clusters are present in their genomes.

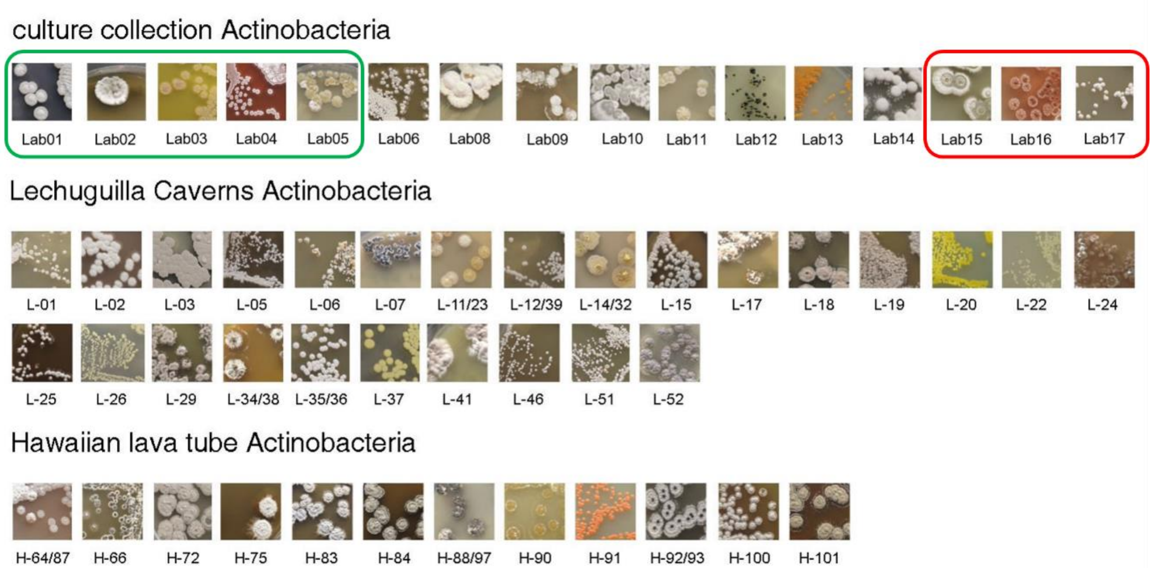

Figure 2-6. Images of Actinobacteria analyzed in this study. These strains shown were pure and grown on GYM agar plates for one to two weeks. Green box represents positive controls, while red box represents negative controls. The various color displayed by these Actinobacteria suggested that PK-II BGCs in their genomes might be transcriptionally active.

To expand the scope of this study, our group, in collaboration with Dr. Diana Northup from UNM Department of Biology, isolated and purified 38 additional unique Actinobacteria from 102 environmental samples — 59 soil samples were collected from Lechuguilla Caverns and 43 samples from Hawaiian lava tubes (Figure 2-6). These environmental strains were recovered by previous undergraduates of our group and further purified and confirmed by several rounds of cultivation on GYM plates, genomic DNA

isolation, and 16S rRNA genes sequencing until they displayed a single 16S rRNA genotype.

Lechuguilla Cave (Carlsbad Caverns National Park, New Mexico) is the fourth-longest cave in the United States and well-known for its unusual geology, rare formations, and pristine condition. Rare sulfur-, iron, and manganese-oxidizing bacteria from this extremely nutrient-limited environment were found to be critical to those geological formations such as corrosion residues[24,25]. In addition, extensive cave-wide bacterial and fungal colonization, including *Aspergillus sp.* fungi, *Devosia sp.*, and *Sphingopyxis sp.* bacteria, was revealed in this extreme environment[25,26]. Most intriguingly, a single bacteria isolated from this cave exhibited remarkable resistance diversity (specifically resistant to 26 of 40 clinically used antibiotics) and new resistance mechanisms, which indicated the opportunities of discovery of novel antibiotics[27]. Lava tubes, a type of lave caves, form when a fluid lava flow cools and thickens but the molten lava flowing underneath drains out and develops a cave. Diverse microbial communities, including various novel actinobacteria, were found to inhabit lave caves on the Big Island of Hawaii, in particular the microbial mats[28,29]. Caves are extreme environments for the development of rare actinobacteria and remain to be an underexploited reservoir of bioactive secondary metabolites, which are promising sources of lead compounds of potential pharmaceutical relevance[30, 31].

To analyze the bacterial community composition of these environmental samples, the genomic DNA was extracted from those purified strains, and their 16S rRNA genes were PCR amplified, sequenced, and analyzed (Table 2-6).

Culture Collection Bacteria						
Lab01	Streptomyces coelicolor A3(2)					
Lab02	Streptomyces antibioticus ATCC 11891					
Lab03	Streptomyces peucetius ATCC 29050					
Lab04	Amycolatopsis orientalis subsp. vinearia BA-07585					
Lab05	Kibdelosporangium sp. MJ126-NF4					
Lab06/07	Streptomyces ficellus NRRL 8067/DSM 930					
Lab08	Streptomyces lavendulae NRRL 2564					
Lab09	Streptomyces venezuelae ATCC 15439					
Lab10	Streptomyces narbonensis ATCC 19790					
Lab11	Actinomadura kijaniata SCC 1256 (ATCC 31588)					
Lab12	Micromonospora olivaterospora NRRL 8178					
Lab13	Micromonospora megalomicea subsp. nigra ATCC 27598					
Lab14	Streptomyces mobaraensis DSM 40847					
Lab15	Streptomyces griseus IFO 13350					
Lab16	Saccharopolyspora erythraea NRRL 2338					
Lab17	Saccharopolyspora spinosa NRRL 18395					
Designated name	**New name in our collection**	**16S rRNA gene Blastn results**				**Amplicon size (bp)**
Lechuguilla Caverns Bacteria		Hit (NCBI blastn)	coverage	identity	accession	
L-01	Nocardia sp. L-01	Nocardia fluminea strain 173590	100%	99%	EU593589.1	1374
L-02	Saccharothrix sp. L-02	Saccharothrix_sp._LM	100%	99%	AF328678.1	1371
L-03	Streptomyces sp. L-03	Streptomyces sp. XKE25	100%	99%	KP872949.1	1380
L-04	Bacillus sp. L-04	Bacillus cereus strain HYM75	100%	99%	KT982232.1	1346
L-05	Nocardia sp. L-05	Nocardia alba strain YIM 30243	100%	99%	NR_025726.1	1371
L-06	Nocardia sp. L-06	Nocardia asteroides strain N18	100%	99%	KT003509.1	1384
L-07	Alloactinosynnema sp. L-07	Alloactinosynnema sp. L-07	100%	100%	LN850107.1	1381
L-09	Staphylococcus sp. L-09	Staphylococcus hominis strain CSY17	100%	100%	KM091706.1	1417
L-10	Staphylococcus sp. L-10	Staphylococcus hominis strain R14	100%	99%	KM017979.1	1418
L-11/23	Nocardiopsis sp. L-11	Nocardiopsis aegyptia strain TGT-R2	100%	99%	KR476435.1	1403
L-12/39	Nocardia sp. L-12	Nocardia sp. QLS42	100%	99%	JQ838093.1	1348
L-13	Ralstonia sp. L-13	Ralstonia sp. SK1	100%	99%	DQ026295.1	1400
L-14/32	Streptomyces sp. L-14	Streptomyces sp. FXJ3.004	100%	99%	JN683659.1	1390
L-15	Streptomyces sp. L-15	Streptomyces zaomyceticus strain 174505	100%	99%	EU593738.1	1392
L-16	Bacillus sp. L-16	Bacillus cereus strain HYM75	100%	99%	KT982232.1	1399
L-17	Streptomyces sp. L-17	Streptomyces wedmorensis strain HBUM173193	100%	99%	FJ486459.1	1384
L-18	Streptomyces sp. L-18	Streptomyces sp. 37	100%	99%	KJ888155.1	1397
L-19	Streptomyces sp. L-19	Streptomyces zaomyceticus strain 174487	100%	100%	EU593685.1	1391
L-20	Streptomyces sp. L-20	Streptomyces sp. N22	100%	99%	KJ648180.1	1390
L-21	Paenibacillus sp. L-21	Paenibacillus xylanilyticus strain BAB-1610	99%	99%	KF535141.1	1398
L-22	Micrococcus sp. L-22	Micrococcus sp. 13-33-7	100%	99%	KM886194.1	1371
L-24	Streptomyces sp. L-24	Streptomyces venezuelae ATCC 15439	100%	99%	LN881739.1	1391
L-25	Nocardia sp. L-25	Nocardia jejuensis strain T20-4	100%	99%	KJ571094.1	1383
L-26	Cellulomonas sp. L-26	Cellulomonas sp. strain 1533	100%	99%	Y09658.1	1392
L-27	Paenibacillus sp. L-27	Paenibacillus sp. 6495m-C2	100%	99%	AJ509004.1	1427
L-28	Bacillus sp. L-28	Bacillus aryabhattai	100%	99%	LN890215.1	1427
L-29	Streptomyces sp. L-29	Streptomyces zaomyceticus strain 174487	100%	100%	EU593685.1	1392
L-31	Bacillus sp. L-31	Bacillus altitudinis strain Bacteria VII	100%	99%	KT427442.1	1420
L-33	Paenibacillus sp. L-33	Paenibacillus sp. 6495m-C2	100%	99%	AJ509004.1	1428
L-34/38	Streptomyces sp. L-34	Streptomyces sp. AR2	100%	99%	EF491601.1	1392
L-35/36	Nocardia sp. L-35	Nocardia asteroides strain N18	100%	99%	KT003509.1	1390
L-37	Kocuria sp. L-37	Kocuria palustris strain MU14/1	100%	99%	CP012507.1	1403
L-40	Bacillus sp. L-40	Bacillus subtilis strain E1-3	100%	99%	KJ958215.1	1383
L-41	Streptomyces sp. L-41	Streptomyces sp. HBUM190111	100%	99%	KR906464.1	1390
L-42	Bacillus sp. L-42	Bacillus thuringiensis strain Xmb014	100%	99%	KT986144.1	1429
L-43	Bacillus sp. L-43	Bacillus thuringiensis strain ZLynn500-22	100%	99%	KY316414.1	1432
L-44	Stenotrophomonas sp. L-44	Stenotrophomonas sp. NR17	100%	99%	JN082748.1	1415
L-45	Bacillus sp. L-45	Bacillus sp. SG2-6	100%	99%	KP992138.1	1406
L-46	Cellulosimicrobium sp. L-46	Cellulosimicrobium cellulans strain DSM 43879	100%	99%	NR_119095.1	1395
L-47	Bacillus sp. L-47	Bacillus thuringiensis strain BH29	100%	99%	KY910254.1	1423
L-48	Bacillus sp. L-48	Bacillus thuringiensis strain Xmb014	100%	99%	KT986144.1	1427
L-49	Bacillus sp. L-49	Bacillus cereus strain WJB94	100%	99%	KU877653.1	1423
L-50	Bacillus sp. L-50	Bacillus thuringiensis strain ZLynn800-5	100%	100%	KY316426.1	1385

Table 2-6 (cont.)

L-51	Nocardia sp. L-51	Nocardia sp. 193538	100%	100%	KU982631.1	1374
L-52	Streptomyces sp. L-52	Streptomyces werraensis strain RB1-22	100%	99%	LC128333.1	1390
L-53	Bacillus sp. L-53	Bacillus cereus strain ML208	100%	99%	KC692193.1	1423
L-54	Bacillus sp. L-54	Bacillus sp. strain FJAT-25710	100%	99%	KY949539.1	1415
L-55	Paenibacillus sp. L-55	Paenibacillus sp. FSL H7-0357	100%	99%	CP009241.1	1427
L-56	Bacillus sp. L-56	Bacillus sp. Dma11	100%	99%	JQ977497.1	1414
L-57	Bacillus sp. L-57	Bacillus sp. JCM 28842 strain T7822-2-1b	100%	100%	LC150680.1	1419
L-58	Pseudomonas sp. L-58	Pseudomonas sp. YAnl_w1	100%	99%	KU851251.1	1403
Hawiian Lava Tube Bacteria		Hit (NCBI blastn)	coverage	identity	accession	
H-62	Bacillus sp. H-62	Bacillus toyonensis strain ZLynn1000-37	100%	100%	KY316458.1	1383
H-64/87	Lentzea sp. H-64	Lentzea violacea strain 173535	100%	99%	EU570364.1	1370
H-65	Acidobacteria sp. H-65	Uncultured Acidobacteria bacterium	100%	99%	AM902634.1	1339
H-66	Streptomyces sp. H-66	Streptomyces scabiei strain OR9T	100%	99%	AB894410.1	1387
H-67	Bacillus sp. H-67	Bacillus sp. L25	100%	99%	KR007003.1	1430
H-72	Streptomyces sp. H-72	Streptomyces sp. E5N91	100%	100%	KX279540.1	1394
H-74	Staphylococcus sp. H-74	Staphylococcus epidermidis strain B13	100%	100%	MF083080.1	1422
H-75	Streptomyces sp. H-75	Streptomyces chilikensis strain RC 1830	100%	100%	NR_118246.1	1398
H-77	Bacillus sp. H-77	Bacillus mojavensis strain LMB3G81	100%	100%	MF040286.1	1422
H-79	Bacillus sp. H-79	Bacillus subtilis subsp. subtilis strain MER_89	100%	99%	KT719664.1	1421
H-80	Bacillus sp. H-80	Bacillus subtilis	100%	99%	EU256502.1	1432
H-81	Bacillus sp. H-81	Bacillus subtilis strain TLO3	100%	100%	CP021169.1	1423
H-83	Streptomyces sp. H-83	Streptomyces cinnamocastaneus strain HBUM173422	99%	99%	EU841658.1	1378
H-84	Streptomyces sp. H-84	Streptomyces sp. VTT E-042677	100%	99%	EF564804.1	1391
H-85	Bacillus sp. H-85	Bacillus licheniformis strain WJB11	100%	99%	KU877628.1	1431
H-86	Bacillus sp. H-86	Bacillus subtilis strain JPR4	100%	100%	KM083800.1	1424
H-88/97	Streptomyces sp. H-88	Streptomyces sp. E3N208	100%	99%	KX279588.1	1392
H-90	Streptomyces sp. H-90	Streptomyces fradiae strain NIOT-Cu-51	100%	99%	KJ575069.1	1396
H-91	Williamsia sp. H-91	Williamsia serinedens strain IMMIB W-9660 R variant	100%	99%	FN673550.1	1387
H-92/93	Streptomyces sp. H-92	Streptomyces sp. QZGY-A34	100%	100%	JQ812091.1	1394
H-94	Streptomyces sp. H-94	Streptomyces werraensis strain RB1-22	99%	99%	LC128333.1	1395
H-96	Staphylococcus sp. H-96	Staphylococcus hominis subsp. novobiosepticus strain IIF3SW-P1	100%	99%	KY218858.1	1418
H-99	Bacillus sp. H-99	Bacillus sp. strain WC5	100%	100%	JN975953.1	1415
H-100	Streptomyces sp. H-100	Streptomyces sp. strain HMU101	100%	99%	KU058411.1	1400
H-101	Streptomyces sp. H-101	Streptomyces sp. BAB5	100%	99%	JF799913.1	1396
H-102	Staphylococcus sp. H-102	Staphylococcus hominis	100%	99%	HG941661.1	1432

Table 2-6. Summary of all strains containing 16S rRNA information. Each strain has a designated name for the convenience of referring to them. Actinobacteria are colored green, while non-actinobacterial strains are colored red. The duplicate species have been removed.

From the 102 probed strains, 38 were unique Actinobacteria after removing the redundant species. An obvious observation is that most non-actinobacterial strains are *Bacillus* species, suggesting the need for designing more specific primers for amplification of 16S rRNA genes of Actinobacteria to prevent this interference in future studies.

Previously, the phylogenetic distribution of PK-II BGCs in bacteria was analyzed by *Dynamite* in our group and revealed that PK-II BGCs were prevalent within six

suborders of phylum Actinobacteria, specifically in *Pseudonocardineae*, *Micromonosporineae*, *Streptosporangineae*, *Catenulisporineae*, *Frankineae*, and *Streptomycineae*, which were referred to as phylogenetic "hotspots" for PK-IIs production (Table 2-7).

suborder	# genomes	# PK-II positive genomes	% PK-II positive genomes	# genera	# genera with genomes	% genera with >0 genomes	# PK-II positive genera with >0 genomes	% PK-II positive genera with >0 genomes
Micromonosporineae	55	10	18.2%	24	6	25.0%	4	66.7%
Streptomycineae	109	67	61.5%	5	2	40.0%	2	100.0%
Catenulisporineae	2	1	50.0%	2	1	50.0%	1	100.0%
Streptosporangineae	27	15	55.6%	26	8	30.8%	6	75.0%
Pseudonocardineae	38	16	42.1%	23	11	47.8%	7	63.6%
Frankineae	17	11	64.7%	12	7	58.3%	3	42.9%
Micrococcineae	57	2	3.5%	30	17	56.7%	2	11.8%
Corynebacterineae	513	1	0.2%	17	12	70.6%	1	8.3%
sum/average	818	123	37.0%	139	64	47.4%	26	58.5%

Table 2-7. Phylogenetic distribution analysis of PK-II BGCs in phylum Actinobacteria. Six phylogenetic hotspot suborders were found to harbor type II PKS gene clusters with different percentages. The figure was acquired from Dr. Melançon's personal manuscript.

The phylogenetic tree analysis of 16S rRNA (Figure 2-7) revealed that the Actinobacteria collection in this study comprised 57% species from the hotspot suborder *Streptomycineae* and only 22% species represented other suborders. According to a previous report[32], six actinobacterial orders, *Streptomycales*, *Frankiales*, *Micromononsporales*, *Pseudonocardales*, *Streptosporangiales*, and *Corynebacteriales*, were abundant in type II PKS gene clusters. Analysis based on orders revealed that the majority of strains from the Actinobacteria collection in this study belonged to prolific orders, while only 7% belonged to non-prolific orders. Furthermore, studies have shown that the diversity and uniqueness of Actinobacteria communities were well correlated with

the novelty of PK-II BGCs[4]. Taken together, above phylogenetic analysis strongly indicates that the strains in our Actinobacteria collection are promising producer of structurally novel PK-IIs.

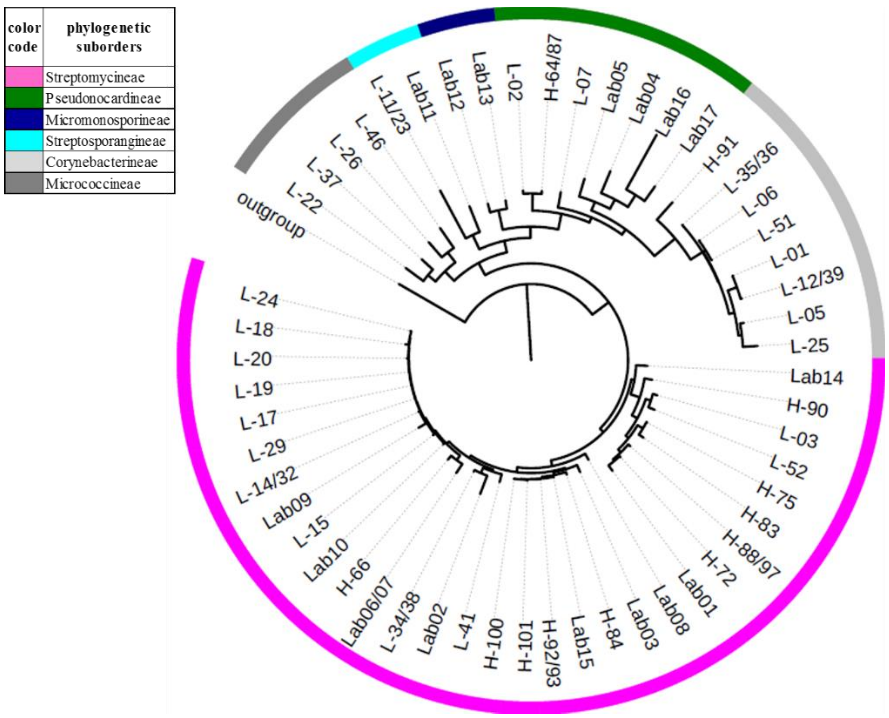

Figure 2-7. Phylogenetic diversity analysis of all unique actinobacteria. This phylogenetic tree only contains the designated names for the sake of clarity. The full strain name are summarized in Table 2-6. The colored rings show the suborders they belonging to.

KSα/β Amplicon-based Identification of PK-II BGCs. In order to efficiently identify the bacteria that have the genetic potential to produce PK-IIs from above 54 environmental Actinobacteria whose genomes have not yet been sequenced, degenerate primers were designed to amplify a DNA fragment (amplicon) of the KSα gene from the genomic DNA isolated from pure-cultured Actinobacteria. Consensus sequences identified by multiple sequence alignment of various KSα genes were used to design two degenerate oligonucleotide primers, KSaF2 and KSaR1, for the PCR amplification of KSα amplicons.

The data set comprises the proteins encoded by the KSα genes from 78-membered training set whose KSα/β product were identified. The sizes of amplified DNA fragments are expected to be approximately 510 bp, encoding about 170 amino acids.

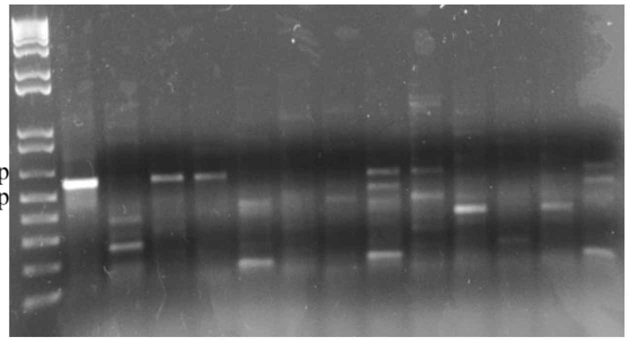

Figure 2-8. PCR amplification of KSα amplicons. KSα was amplified by degenerate primers KSaF2 and KSaR1. From left to right, wells are 1 kb plus DNA ladder, PCR products from positive control (Lab05), L-01, L-02, L-03, L-05, L-06, L-07, L-11, L-15, L-17, L-18, L-19 and L-23.

Using this degenerate primer pair, we were only able to identify a few strains (33%, 4 out of 12, Figure 2-8) from the environmental samples harboring PK-II BGCs, which may result from the primer bias. However, using the degenerate primes for KSα/β amplicons (described below), we identified 3 more PK-II BGC positive strains that were missed by KSα amplicons identification.

To improve the efficiency of identification of PK-II BGCs, new degenerate primers were needed to be designed. Furthermore, based on the idea that the identities of amino acids at specific positions within each KSα/β enzyme should co-vary with the chemical structure of the poly β-ketone produced, and should be predictive of its structure, several algorithms (detailed below) were developed as a function within *Dynamite* to identify amino acid positions within the 72 training set KSα/β sequences that were most highly predictive of product structure for the 18 poly β-ketone products collectively made by these enzymes. Given our ability to accurately predict poly-β-ketone product structure

(chemotype) from KSα/β amino acid sequence motifs using *Dynamite,* we envisioned that the highly predictive positions of an appropriately positioned KSα/β amplicon could also give accurate prediction of KSα/β product chemotypes, encompassing 11 different starter units and 16- to 26-carbon chain lengths. To this end, we designed a new degenerate primer pair, KSab-fwd and KSab-rev, which would amplify an approximately 1.2 kb amplicon encoding about 400 amino acids that contained portions of both the KSα and KSβ genes. In principle, this method could be used to clone the full-length KSα/β genes, but it would create stronger bias or less degeneracy. Compared with existing degenerate primers, the new degenerate primer pair designed in this work has two advantages: first, it covers the region that enabled the development of most detailed predictive models for KSα/β product structure from sequence; second, this amplified region containing the key fingerprint residues allows more accurate prediction than KSα/β genes or partial amplicons based phylogenetic analysis.

To effectively amplify every possible PK-II KSα/β genes from the extracted genomic DNA, the initial PCR conditions were tested with five positive strains and three negative strains chosen from Lab culture collections. The five positive strains are Lab01, Lab02, Lab03, Lab04 and Lab05, and contain 2, 3, 1, 1 and 4 PK-II BGCs, respectively. PCR conditions were optimized as follows: switching AmpliTaq Gold DNA polymerase to OneTaq Hot Start DNA polymerase (data not shown) since the latter one is more suitable for GC-rich amplicons with improved performance and yield, testing annealing temperature gradient from 51 to 60 ℃ (Figure 2-9a), changing the amount of template (Figure 2-9b), trying different degenerate primer pairing (Figure 2-9c), adding OneTaq High GC enhancer or DMSO (data not shown). After extensive PCR condition

investigation, annealing temperature at 51 ℃, 5 ng of template per 25 μL reaction, KSaF2 paired with KSab-rev, and without addition of DMSO were determined to be the optimal PCR condition (Figure 2-9).

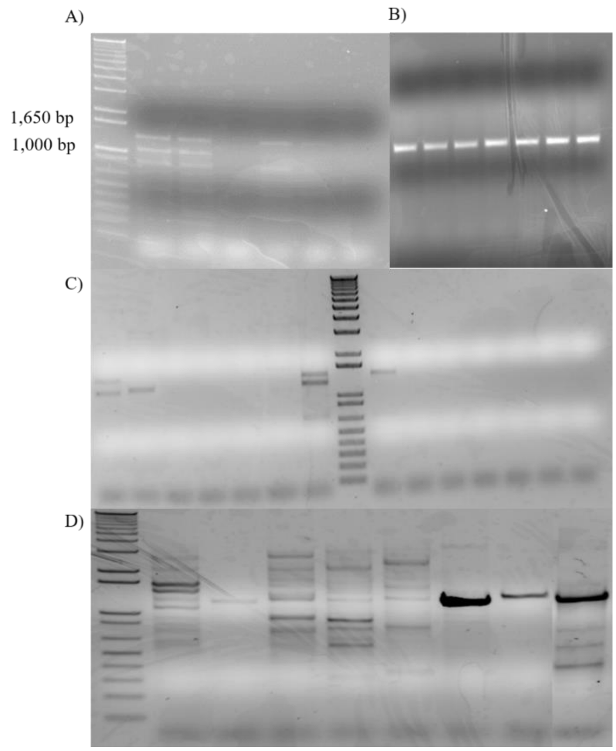

Figure 2-9. PCR condition optimization for the amplification of KSα/β amplicons. A) Investigation of the gradient of annealing temperature using the genomic DNA of Lab05 as template. From left to right, wells are 1 kb plus DNA ladder, annealing temperature 51 ℃, 53.1 ℃, 54.6 ℃, 56.6 ℃, 58.3 ℃ and 60 ℃, respectively. B) Investigation of the amount of template. Lab05 genome was used. From left to right, wells are 1 ng, 2 ng, 5 ng, 7 ng, 10 ng, 15 ng and 20 ng per 25 μL reaction, respectively. C) Comparison of primer KSaF2 with KSab-fwd paired with KSab-rev. On the left side of marker is KSaF2, while the right side is KSab-fwd. From left to right, template for each well are genomic DNA of L-12, L-14, L-20, L-22, L-46, L-51 and L-52, respectively. D) Various performance displayed using genomic DNA of different strains as template. From left to right, wells are genomic DNA of Lab04, Lab02, Lab16, Lab11, Lab12, Lab14, Lab03 and Lab05, respectively. All above PCR were carried out using OneTaq Hot Start DNA polymerase.

Although genomic DNA from different strains used as templates in amplification of KSα/β amplicons have shown distinct efficiency (Figure 2-9d), all positive controls were identified, and subsequent KSα/β amplicons sequencing analysis showed that 9 out of 11 KSα/β amplicons were cloned. Owing to the low amount of amplified DNA and the 3’-dA tail produced by the OneTaq Hot Start DNA polymerase, two rounds of PCR were applied

to increase the amount of PCR product and blunt-end the PCR fragments using Phusion DNA polymerase.

All of the amplified KSα/β fragments were cloned into vector pCR-Blunt for subsequent Sanger sequencing. Since numerous actinobacterial species have been shown to harbor more than one PK-II BGCs, several clones (range from 5 to 25) from different strains were subjected to high-throughput sequencing analysis. After sequencing, their sizes were determined to be about 1,200 bp, corresponding to the downstream halves of KSα subunit genes and upstream halves of KSβ subunit genes. In each sequences, the KSα and KSβ are linked by a stop-start codon overlap typical of KSα/β gene pair junctions. Then sequenced clones were assembled into unique KSα/β amplicons based on multiple sequence alignment. Results of this initial study (Figure 2-10) indicates that the new primer pair amplified 9 of 11 (82%) known amplicons from positive control organisms, and the two missing are from Lab02 (harbor 3 PK-II BGCs) and Lab05 (harbor 4 PK-II BGCs), respectively. In the case of Lab02, the missing one could be detected in the first PCR and second PCR products using specific primers, Oviedo, indicating that this KSα/β amplicon is cloned by the degenerate primers. In a special case, this new degenerate primers could identify 3 PK-II BGCs from the genome of Lab05, *Kibdelosporangium sp.* MJ126-NF4, which possesses 4 type II PKS gene clusters known from its sequenced genome. Using this primer pair, we also identified 8 new amplicons from these experimental strains. Therefore, this newly designed degenerate primer pair represents an efficient tool to identify bacteria harboring gene clusters that encode PK-IIs (Figure 2-10).

After verifying the specificity and accuracy of the protocol using several control actinobacterial strains known to harbor specific PK-II BGCs, the same approach was then applied to probe the genomic DNA isolated from environmental samples.

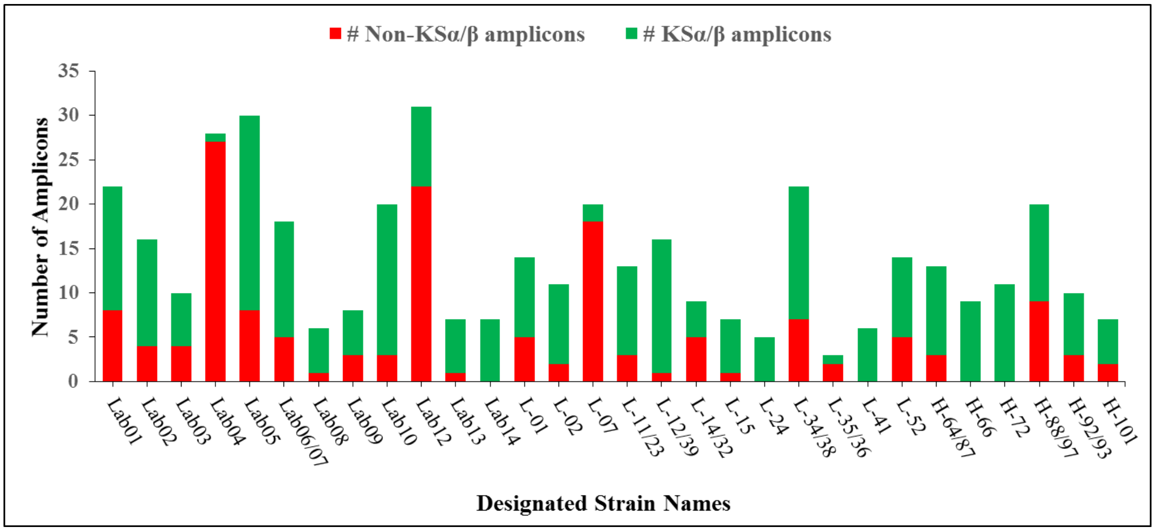

Figure 2-10. Statistics of KSα/β amplicon sequencing. Green color denotes the number of sequenced colonies that are KSα/β amplicons based on subsequent bioinformatic analysis, while red color denotes non-KSα/β amplicons. This statistics demonstrated that the new degenerate primers displayed different efficiency in cloning different KSα/β amplicons.

Overall, 30 out of 54 (56%) strains were identified through KSα/β amplicon sequencing as being positive for harboring PK-II BGCs, and most of KSα/β amplicons were efficiently cloned, except for two from Lab04 and L-07 (Figure 2-10). Among these KSα/β amplicon positive organisms, 19 out of 30 (63%) are *Streptomyces* species. This result is consistent with our prediction based on the phylogenetic analysis of our strains. Again, it provides additional support to that there is a strong correlation between phylogeny and PK-II BGCs abundance and diversity. In conclusion, our analysis of the abundance and distribution of PK-II BGCs in the data set, combined with the KSα/β amplicon amplification, provides an efficient, scalable route for the selection and prioritization of actinobacterial strains for future PK-II natural product discovery endeavor.

Phylogenetic Analysis of KSα/β Amplicons. The nucleotide sequences from sequenced KSα/β amplicons were translated into protein sequences, which were trimmed at the degenerate primer sites, aligned, and visualized in a phylogenetic tree (Figure 2-11).

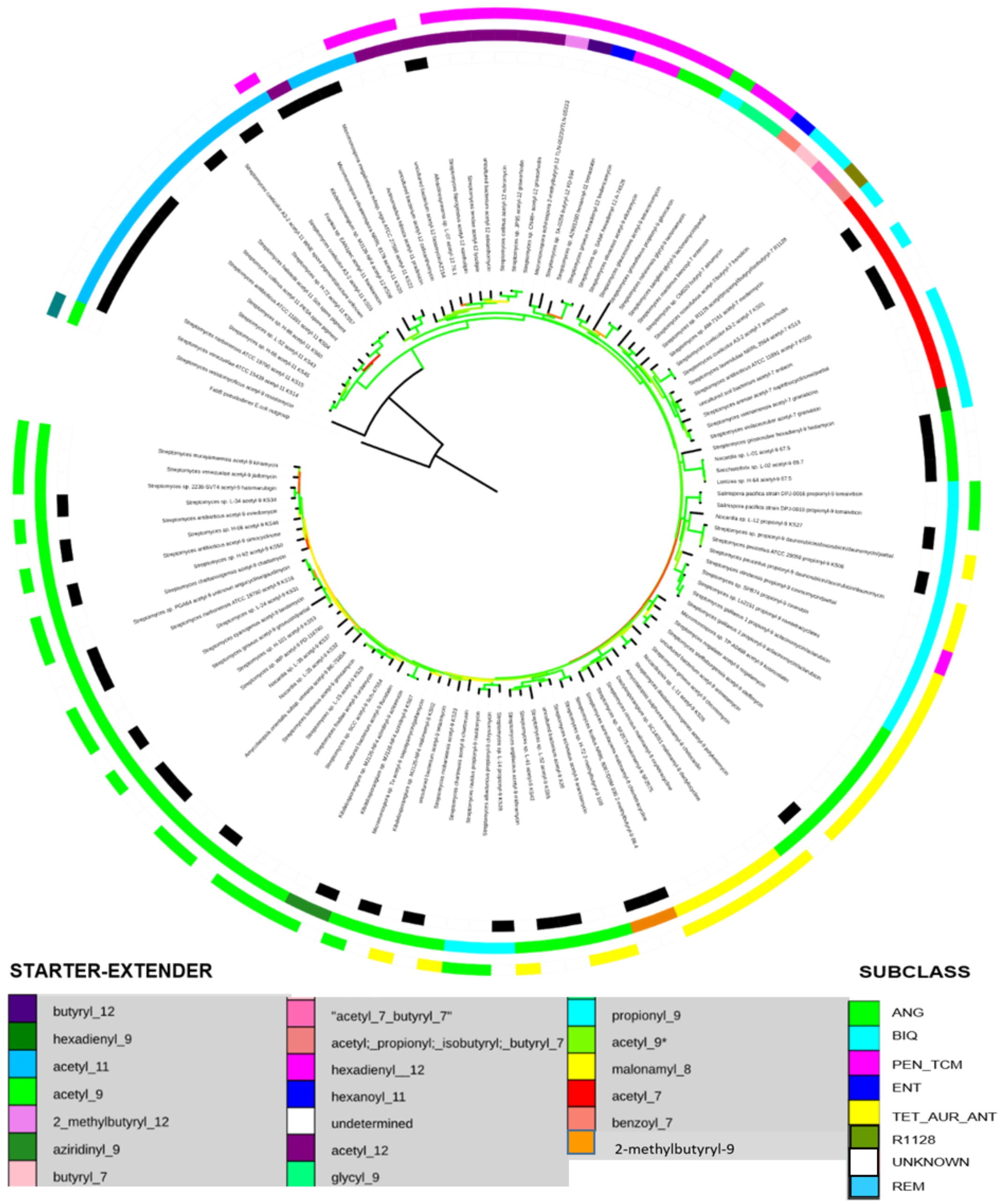

Figure 2-11. Phylogenetic tree analysis of KSα/β amplicons. From outside to center: subclasses for training set members with characterized PK-II structures, starter-extender units, and KSα/β amplicons identified in this work are denoted by black bars (see bottom figure legend). Subclass abbreviations: ANG-angucycline; BIQ-benzoisochromanequinone; PEN_TCM-pentangular or tetracenomycin; ENT-enterocin; TET-tetracycline; AUR-aureolic acid; ANT- anthracyclines; R1128-R1128; Unknown-subclass undetermined; REM-resistomycin. FabB_pseudo-dimer from *E. coli* fatty acid pathway was used as outgroup.

This phylogenetic tree analysis was expanded by adding all 78 known KSα/β sequences (the training set) from different type II PKS pathways. As reported by other studies, the sequences of the putative spore pigment polypeptides obviously separated from those of the antibiotics, and the sequence identity within the spore pigment polypeptide group was very high, ranging from 85% to 94%. Given the structural variety of PK-II scaffolds, the sequences of antibiotics group have expected diversity, which could further be divided into subgroups. The results of the dendrogram show an interesting correlation between the polyketide products and the starter units and extender units used by the corresponding KSα/β (Figure 2-11). This analysis supports the critical role of KSα/β in defining the starter unit and chain length during PK-II biosynthesis.

Computational Prediction of Poly-β-ketone Chemotypes. With the KSα/β amplicon sequences in hand, fingerprint analysis of each unique KSα/β amplicon could be carried out to give prediction of poly-β-ketone chemotypes (i.e. molecular structures encoded by genes) and corresponding similarity scores. Previously, the crystal structure of the actinorhodin KSα/β revealed that polyketides are elongated inside an amphipathic tunnel at the heterodimer interface and has led to the proposal that the chain length are regulated by size and shape of the KSα/β active site (Figure 2-12)[9]. Our analysis based on X-ray crystal structure of actinorhodin KSα/β and sequence comparison of several KSα/β with different poly-β-ketone chain lengths gave us the idea that the identities of the amino acids lining the active site cavity or in close proximity to it should co-vary with the chemical structure of the poly-β-ketone chain.

Based on this idea, two fingerprint residues models for the prediction the KSα/β product chemotypes were developed by Benjamin J. Yackley of our group. The model

developed using the principal component analysis (PCA) approach was based on X-ray crystal structure of actinorhodin KSα/β and multiple sequence comparison of KSα/β with different poly-β-ketone chain lengths. Through PCA, the amino acids lining the active site cavity or in close proximity were revealed to well correlate with the chemical structure of the poly-β-ketone chain. Twenty-five most highly covariant positions were chosen from this model and employed to predict the poly-β-ketone product structure of each test set member based on the amino acid similarity (as determined by the BLOSUM62[23] scoring matrix) between a test set KSα/β and each of the training set KSα/β. In most cases, application of this 25 amino acid model to the test set members resulted in a clear statistical preference for a single predicted poly-β-ketone product, suggesting that the model was valid. However, in some cases, particularly when two poly-β-ketone products are structurally highly similar, the model gave an ambiguous result.

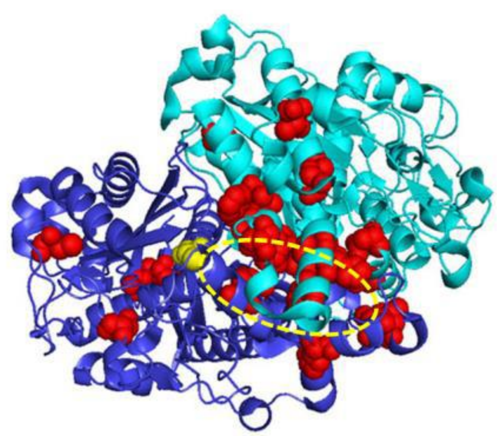

Figure 2-12. The X-ray crystal structure of the actinorhodin KSα/β. The 25 predictive fingerprint residues are shown in red. The active site Cysteine is shown in yellow. The active site pocket is shown with a yellow dashed oval. Most of the fingerprint residue positions are near the active site.

To overcome this problem, an alternative model for predicting poly-β-ketone chemotypes from KSα/β sequence was developed based on machine learning algorithm,

specifically Metropolis-Hastings sampling. Each trial started by randomly choosing an amino acid position in the training set sequence alignment as initial predictive model and scoring its ability to differentiate a correct from incorrect poly-β-ketone product prediction for each training set sequence. If an amino acid position showed an ability (above a threshold value) to differentiate a correct from an incorrect product prediction, that position was retained in the model; and if it showed no discriminating ability, then it was discarded. Then another one amino acids was randomly added to the model, and its predictive ability was scored, and so on. This process was run for 100,000 steps of Meteropolis-Hastings sampling (finished at UNM's supercomputer center) to explore the space of all possible model, obtaining a highest-scoring model (a set of predictive amino acid positions) at the end of sampling process. After 100 independent search trials, each amino acid position was ranked by occurrence in all 100 models, and 25 most frequently occurring positions (determined by computationally testing the predictive ability of different number of these positions) were selected as the final model for prediction.

A 78-membered training set was used to test the accuracy of two different models in the prediction of KSα/β products. After careful testing and extensive comparison, the machine learning based model gave more accurate prediction for a single predicted poly-β-ketone product, even in cases where the products are structurally highly similar. As expected, most of the positions in both models cluster near the active site in the actinorhodin KSα/β structure, providing additional support for their validity. Finally, the 25 fingerprint residues from machine learning based model were employed to predict the starter unit and extender units.

strain	amplicon	# detect	PCA results	% match	ML results	% match	Fingerprints motif from PCA model
Lab01	KS01	8	acetyl-7	100	acetyl-7	100	RTPVISFAAHSAGCLSGFTHFWSVE
	KS03	2	acetyl-11	100	acetyl-11	100	HSPTISFGAHSAGCAFGGGQLWGPQ
Lab02	KS04	2	acetyl-11	95.9	acetyl-11	95.9	HSPTISFGAHSAGCQFGGGQLWGPQ
	KS05	10	acetyl-7	100	acetyl-7	91.9	QTPVISFAAHSAGCLSGFTHFWSVE
Lab03	KS06	5	propionyl-9	100	propionyl-9	100	EAPTISFGAHTAGLLMGFGQLWSAM
Lab04	KS101	1	acetyl-9	100	acetyl-9	100	EAPTISFGAHSAGCRMGFGQLWSAQ
Lab05	KS02	20	acetyl-9	100	acetyl-9	93.2	EAPTISFGAHSAGCLMGFGQLWSAQ
	KS07	6	aziridinyl-9	100	aziridinyl-9	100	EAPTISFGAHSAGCLMSFGELWSAQ
	KS08	5	acetyl-11	80	acetyl-11	73.4	QSPTISFGGHSAGCTWSNGQLWGAQ
Lab06/07	KS11	12	2-methyl butyryl-9	85.7	2-methyl butyryl-9	88.4	DAPTICLGAHSAGLLMGVGQLWSAQ
Lab08	KS13	6	acetyl-7	100	acetyl-7	97.3	RTPVISFAAHSAGCLSGFTHFWSVE
Lab09	KS14	5	acetyl-11	95.9	acetyl-11	96.8	HSPTISFGAHSAGCRFGGGQLWGPQ
Lab10	KS15	7	acetyl-11	95.9	acetyl-11	96.8	HSPTISFGAHSAGCRFGGGQLWGPQ
	KS16	9	acetyl-9	100	acetyl-9	98	EAPTISFGAHSAGCRMGFGQLWSAQ
Lab12	KS20	8	acetyl-12	78.9	acetyl-11	76.8	QSPTISFAGHSAGCAWANGQLWGAQ
Lab13	KS22	6	acetyl-12	77.6	acetyl-11	75	QSPTISFAGHSAGCRWANGQLWGAQ
Lab14	KS23	7	acetyl-9	92.8	acetyl-9	89.1	GAPTISFGAHSAGCRMGFGELWSAQ
L-01	KS54	9	hexanoyl-11	87	acetyl-9	67.5	EAPTICLGAHSAGCLMAYGQLWSAQ
L-02	KS24	9	hexanoyl-11	87	acetyl-9	69.7	EAPTICLGAHSAGCLMAYGQLWSAQ
L-07	KS25	8	acetyl-12	100	acetyl-12	76.1	QAPTISFGAHSAGCLWATGQMYGAQ
L-11/23	KS26	10	acetyl-9	100	acetyl-9	93	RAPTISFGAHSAGCLSGFGHLWSAQ
L-12/39	KS27	15	propionyl-9	91.2	propionyl-9	70.5	EAPTICFGAHTAGLLVGFGQLWSAM
L-14/32	KS28	8	propionyl-9	92.5	propionyl-9	90.2	EAPTISFGGHSAGLLMGYGQLWSAQ
L-15	KS29	6	acetyl-9	100	acetyl-9	94.4	EAPTISFGAHSAGCRMGFGQLWSAQ
L-24	KS31	5	acetyl-9	100	acetyl-9	98	EAPTISFGAHSAGCRMGFGQLWSAQ
L-34/38	KS34	17	acetyl-9	100	acetyl-9	96.6	EAPTISFGAHSAGCRMGFGQLWSAQ
L-35/36	KS32	15	acetyl-9	100	acetyl-9	91	EAPTISFGAHSAGCLMGFGQLWSAQ
	KS37	1	acetyl-9	92.5	acetyl-9	85.4	QAPTISFGAHSAGCRVGFAQLWSAQ
L-41	KS42	6	acetyl-9	97.8	acetyl-9	90.1	EAPTISFGAHSAGCLVGFGQLWSAQ
L-52	KS43	7	acetyl-11	97.9	acetyl-11	96.1	HSPTISFGAHSAGCSFGGGQLWGPQ
	KS55	3	acetyl-9	93.4	acetyl-9	84	EAPTISFGAHSAGCLVGFGQLWGAQ
H-64/87	KS44	10	hexanoyl-11	87	acetyl-9	67.5	EAPTICLGAHSAGCLMAYGQLWSAQ
H-66	KS45	5	acetyl-11	95.9	acetyl-11	96.8	HSPTISFGAHSAGCRFGGGQLWGPQ
	KS46	5	acetyl-9	100	acetyl-9	98	EAPTISFGAHSAGCRMGFGQLWSAQ
H-72	KS57	5	acetyl-11	100	acetyl-11	98.1	HSPTISFGAHSAGCAFGGGQLWGPQ
	KS58	5	2-methyl butyryl-9	85.6	2-methyl butyryl-9	100	EAPTICLGAHSAGLLMGVGQLWSAQ
H-88/97	KS60	11	acetyl-11	95.9	acetyl-11	94.8	HSPTISFGAHSAGCRFGGGQLWGPQ
H-92/93	KS50	7	acetyl-9	95.6	acetyl-9	95.8	EAPTISFGAHSAGCTMGFGQLWSAQ
H-101	KS53	5	acetyl-9	100	malonamyl-8	95.9	EAPTISFGAHSAGCLMGFGQLWSAQ

Table 2-8. Summary of all 39 unique KSα/β amplicons and their predicted poly-β-ketone products. Light green represents positive controls with identified products, while potentially novel poly-β-ketones are highlighted in dark green. Number of every KSα/β amplicons detected are shown after its designated name.

To facilitate KSα/β fingerprint analysis, a web-based program was developed by Yasushi Ogasawara of our group, which could automatically find out the fingerprint residues and score their similarity to the members of training set via BLOSUM62 scoring

matrix. Fingerprints with low similarity scores indicates the query KSα/β enzymes produce potential novel poly-β-ketone products. Using this automated fingerprint analysis tool, ~79% (31) were >90% matches to KSα/β sequences from known PK-IIs, while the remaining ~21% (8) showed relative low similarity scores to known sequences, which suggests their KSα/β products likely possesses different starter units or chain length from training set members (Table 2-8). Consistent with our prediction that phylogenetic diversity correlates with biosynthetic uniqueness, these 8 atypical KSα/β sequences are from 7 different genera, and only 1 is from genus *Streptomyces*. Therefore, this chemotyping analysis could provide a guide for prioritizing PK-II gene clusters for further experimental characterization, including whole genome sequencing (detailed below) and compound detection and isolation.

Whole Genome Sequencing of Strains Harboring Novel Poly-β-ketone. To achieve the full potential of the bioinformatics-guided PK-II discovery approach, several actinobacterial genomes from organisms with KSα/β amplicons identified with low similarity to characterized examples were sequenced to obtain complete sequences of the corresponding PK-II BGCs. From among those that fit this criterion, additional priority was given to organisms from rare genera that have few or no genome sequenced representatives, with the goal of eventually having at least one sequenced genome representing each genus in each type II polyketide phylogenetic hotspot.

Five strains meeting above criteria were subjected to PacBio RSII third generation single molecule real time sequencing (SMRT). To meet the requirements of this long reads sequencing platform, genomic DNA extraction method was developed to isolated high quality genomes from these strains (Figure 2-13). Our initial sequencing target is

Kibdelosporangium sp. MJ126-NF4, a rare Actinobacterium that produces the unusual aziridine ring-containing PK-II azicemicin. Chemotyping analysis identified two unique KSα/β amplicons in addition to the azicemicin amplicon, including a predicted pentangular subclass member with low similarity to known examples (Table 2-8). This genome was first sequenced (conducted by Yasushi Ogasawara of our group) using the Ion Torrent sequencing platform in collaboration with Dr. Jeremy Edwards of UNM Department of Chemistry and Chemical Biology; and produced a ~500 contig draft genome. Subsequent sequencing using PacBio technology and hybrid Illumina/PacBio assembly and polishing produced an 11.75 megabase pair (Mbp), 28 contig, 21 scaffold high quality draft genome[33]. In this work, in collaboration with Dr. Jeremy Edwards and the National Center for Genome Resources (NCGR), the *Kibdelosoprangium sp.* MJ126-NF4 genome sequence was obtained as a complete, single 12.1 Mbp contig.

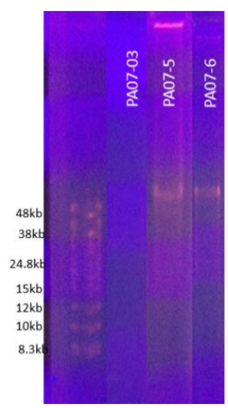

Figure 2-13. Pulse field gel analysis of isolated genomic DNA of *Alloactinosynnema sp.* L-07. As shown on gel, they are high quality genomic DNA that majority of them are larger than 50 kb fragments with less smear DNA.

Importantly, comparison of the sequence of the 50.4 kb azicemicin BGC assembled with our data and the previously reported sequence determined by Sanger sequencing showed they were identical, confirming the quality of the assembly. Analysis of this large

genome revealed four of the type II polyketide class, including an additional predicted pentangular subclass member with low similarity to known examples that was missed during KSα/β amplicon chemotyping.

species	size (Mbp)	contigs	scaffold	coverage	GC%	gene	CDS (%)	secondary metabolite clusters	PK-II gene clusters
Kibdelosporangium sp. MJ126-NF4	12.1	1	1	117x	67.83%	11,392	11,318 (99.4%)	54	4
Alloactinosynnema sp. L-07	7.4	1	1	97x	69.59%	7,400	7,284 (98.4%)	32	1
Lentzea sp. H-64	8.6	1	1	105x	68.61%	8,630	8,552 (99.1%)	26	1
Streptomyces ficellus DSM930	7.1	1	1	100x	72.58%	6,713	6,229 (98.7%)	29	1
Streptomyces venezuelae ATCC 15439	9	1	1	163x	71.74%	8,775	8,682 (98.9%)	34	2

Table 2-9. Summary of information of genomes sequenced. They all have only one contig, high GC content, and a number of secondary metabolite gene clusters. The CDS (coding sequence) percentages were calculated as the number of CDS divided by total number of genes. The secondary metabolite clusters were identified by antiSMASH[34].

We also extracted the genome of *Alloactinosynnema sp.* L-07 (Figure 2-13), an organism we isolated from Lechuguilla Caverns, and sent for genome sequencing using the same technology. This organism harbors a KSα/β amplicon predicted to be from a gene cluster responsible for producing a rare 26 carbon pentangular type II polyketide most similar to the potent antibacterial/anticancer/antifibrotic compound xantholipin. The *Alloactinosynnema sp.* L-07 genome sequence has been deposited in the NCBI databank. Spurred on by the extremely high quality of the genomes obtained using PacBio RSII technology, the genomes of three additional organisms (*Lentzea sp*. H-64 from the Hawaiian lava tubes, *Streptomyces ficellus* DSM930 and *Streptomyces venezuelae* ATCC15439 (conducted by Jingxuan He of our group)), all of which harbor KSα/β amplicons predictive of novel chemotypes, were sequenced in collaboration with NCGR. All the sequenced genomes were assembled by Hierarchical Genome Assembly Process

(HGAP.2) workflow using the Celera Assembler and annotated using Rapid Annotation using Subsystem Technology (RAST)[35] pipeline with GLIMMER 3[36] (Table 2-9).

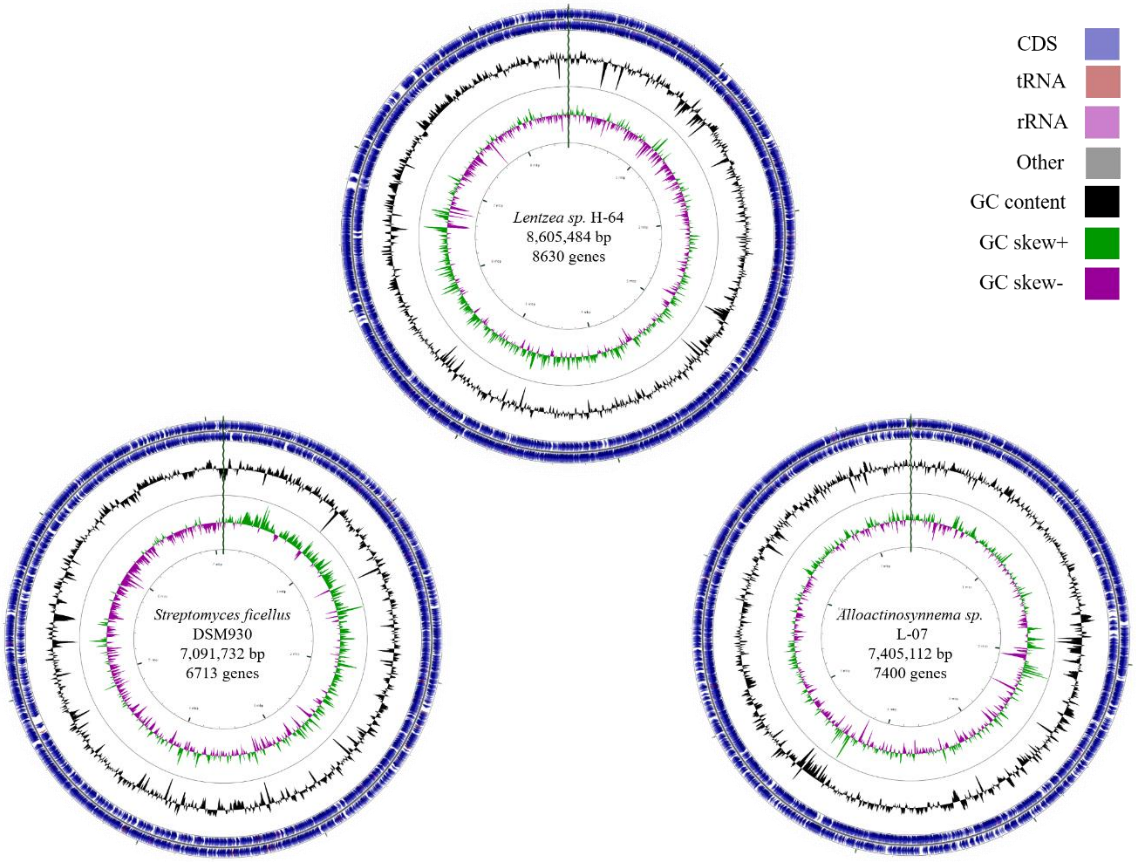

Figure 2-14. Circular maps of genomes of *Alloactinosynnema sp.* L-07, *Lentzea sp.* H-64 and *Streptomyces ficellus* DSM930. From the outside to the center: coding sequences (CDSs) on forward strand and on reverse strand, GC content, and GC skew. These maps were generated using the CGView[37].

Using this genome sequencing platform, we have obtained complete, single contig assemblies of all five actinobacterial genomes, including 3 organisms (*Alloactinosynnema sp*. L-07, *Lentzea sp.* H-64, and *Streptomyces ficellus* DSM930) identified by KSα/β amplicon sequencing (Figure 2-14). *Alloactinosynnema sp*. L-07 is the first genome sequenced in its genus, while *Lentzea sp.* H-64 and *Streptomyces ficellus* DSM930 are potential producer of novel poly-β-ketone products based on fingerprint analysis results.

4. Conclusions

The KSα/β amplicon-based identification and chemotyping approach described in this work allowed rapid identification and prioritization of type II PKS gene clusters from actinobacterial genomes. Using this approach, 8 strains containing atypical KSα/β sequences from 7 different genera were identified for subsequent compound characterization and bioactivity profiling. This strongly indicates a strong correlation between the phylogenetic diversity and the uniqueness of secondary metabolites produced.

This work reported here also provides further support for the notion that some microbial community are indeed gifted for encoding PK-II BGCs with high abundance and diversity. Based on KSα/β amplicon amplification and sequencing, 30 out of 54 strains in our chosen Actinobacteria collection were identified as PK-II producers, among which 23% possesses more than one type II PKS gene clusters. Using the newly designed degenerate primer pair for amplification of partial KSα/β, we identified 10 out 11 positive control PK-II BGCs, demonstrating an effective tool in identification of bacteria harboring PK-II BGCs.

The predictive model developed for chemotyping KSα/β products was a rapid way for in silico dereplication and prioritization of type II PKS pathways for further experimental investigation. The accuracy of several algorithms to identify predictive amino acid positions within the 78 training set KSα/β with known product structures were tested, and a final set of 25 highly predictive amino acid positions (fingerprints) was generated from machine learning-based (specifically Metropolis-Hastings sampling) model. With this fingerprint analysis model, 11 different starter units (either acetate or non-acetate units) and 5 different cycles of extender units (ranging from 7 to 12) were easily determined based

on the KSα/β amplicon sequences, and similarity matching to the 78 members of training set was also given as a predictor of the novelty of poly-β-ketone product structure.

Guided by KSα/β amplicon chemotyping, 5 actinobacterial strains were selected for genome sequencing using third generation sequencing technology PacBio RSII system, among which *Alloactinosynnema sp*. L-07 representes the first genome sequenced in its genus. The genome for each strain sequenced by this genome sequencing technology was assembled into only 1 contig, which proves that it is a powerful platform for acquisition of small genomes of high quality and will facilitate genome mining in natural product discovery field in the future.

5. References

1. Seow, K. T., Meurer, G. U. I. D. O., Gerlitz, M. A. R. T. I. N., Wendt-Pienkowski, E. V. E. L. Y. N., Hutchinson, C. R., & Davies, J. U. L. I. A. N. (1997). A study of iterative type II polyketide synthases, using bacterial genes cloned from soil DNA: a means to access and use genes from uncultured microorganisms. *Journal of bacteriology*, *179*(23), 7360-7368.
2. Metsä-Ketelä, M., Salo, V., Halo, L., Hautala, A., Hakala, J., Mäntsälä, P., & Ylihonko, K. (1999). An efficient approach for screening minimal PKS genes from Streptomyces. *FEMS Microbiology Letters, 180*(1), 1-6.
3. Metsä-Ketelä, M., Halo, L., Munukka, E., Hakala, J., Mäntsälä, P., & Ylihonko, K. (2002). Molecular evolution of aromatic polyketides and comparative sequence analysis of polyketide ketosynthase and 16S ribosomal DNA genes from various Streptomyces species. *Applied and Environmental Microbiology*, *68*(9), 4472-4479.

4. Wawrik, B., Kerkhof, L., Zylstra, G. J., & Kukor, J. J. (2005). Identification of unique type II polyketide synthase genes in soil. *Applied and environmental microbiology*, *71*(5), 2232-2238.

5. Feng, Z., Kallifidas, D., & Brady, S. F. (2011). Functional analysis of environmental DNA-derived type II polyketide synthases reveals structurally diverse secondary metabolites. *Proceedings of the National Academy of Sciences, 108*(31), 12629-12634.

6. Feng, Z., Kim, J. H., & Brady, S. F. (2010). Fluostatins produced by the heterologous expression of a TAR reassembled environmental DNA derived type II PKS gene cluster. *Journal of the American Chemical Society*, *132*(34), 11902-11903.

7. King, R. W., Bauer, J. D., & Brady, S. F. (2009). An Environmental DNA-Derived Type II Polyketide Biosynthetic Pathway Encodes the Biosynthesis of the Pentacyclic Polyketide Erdacin. *Angewandte Chemie International Edition*, *48*(34), 6257-6261.

8. Kang, H. S., & Brady, S. F. (2014). Mining soil metagenomes to better understand the evolution of natural product structural diversity: pentangular polyphenols as a case study. *Journal of the American Chemical Society*, *136*(52), 18111-18119.

9. Keatinge-Clay, A. T., Maltby, D. A., Medzihradszky, K. F., Khosla, C., & Stroud, R. M. (2004). An antibiotic factory caught in action. *Nature structural & molecular biology*, *11*(9), 888-893.

10. Ogasawara, Y., Yackley, B. J., Greenberg, J. A., Rogelj, S., & Melançon III, C. E. (2015). Expanding our understanding of sequence-function relationships of type II polyketide biosynthetic gene clusters: bioinformatics-guided identification of Frankiamicin A from Frankia sp. EAN1pec. *PloS one*, *10*(4), e0121505.

11. Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry, 32*(7), 1466-1474.

12. Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician, 49*(4), 327-335.

13. Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome biology, 14*(6), 405.

14. Fan, K., Pan, G., Peng, X., Zheng, J., Gao, W., Wang, J., ... & Yang, K. (2012). Identification of JadG as the B ring opening oxygenase catalyzing the oxidative CC bond cleavage reaction in jadomycin biosynthesis. *Chemistry & biology, 19*(11), 1381-1390.

15. Eden, P. A., Schmidt, T. M., BLAKEMORE, R. P., & Pace, N. R. (1991). Phylogenetic analysis of Aquaspirillum magnetotacticum using polymerase chain reaction-amplified 16S rRNA-specific DNA. *International Journal of Systematic and Evolutionary Microbiology, 41*(2), 324-325.

16. Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of bacteriology,* 173(2), 697-703.

17. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic acids research, 36*(suppl 2), W5-W9.

18. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., ... & Thompson, J. D. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology, 7*(1), 539.

19. Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one, 5*(3), e9490.

20. Letunic, I., & Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research, 44*(W1), W242-W245.

21. Feliciello, I., & Chinali, G. (1993). A modified alkaline lysis method for the preparation of highly purified plasmid DNA from Escherichia coli. *Analytical biochemistry*, *212*(2), 394-401.

22. Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989). *Molecular cloning: a laboratory manual* (No. Ed. 2). Cold spring harbor laboratory press.

23. Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915-10919.

24. E. Northup, Kathleen H. Lavoie, D. (2001). Geomicrobiology of caves: a review. *Geomicrobiology journal*, *18*(3), 199-222.

25. Cunningham, K. I., Northup, D. E., Pollastro, R. M., Wright, W. G., & LaRock, E. J. (1995). Bacteria, fungi and biokarst in Lechuguilla Cave, Carlsbad Caverns National Park, New Mexico. *Environmental Geology, 25*(1), 2-8.

26. Gan, H. Y., Gan, H. M., Tarasco, A. M., Busairi, N. I., Barton, H. A., Hudson, A. O., & Savka, M. A. (2014). Whole-genome sequences of five oligotrophic bacteria isolated from deep within Lechuguilla Cave, New Mexico. *Genome announcements*, *2*(6), e01133-14.

27. Pawlowski, A. C., Wang, W., Koteva, K., Barton, H. A., McArthur, A. G., & Wright, G. D. (2016). A diverse intrinsic antibiotic resistome from a cave bacterium. *Nature Communications, 7*.

28. Northup, D. E., Melim, L. A., Spilde, M. N., Hathaway, J. J. M., Garcia, M. G., Moya, M., ... & Riquelme, C. (2011). Lava cave microbial communities within mats and secondary mineral deposits: implications for life detection on other planets. *Astrobiology*, *11*(7), 601-618.

29. Hathaway, J. J. M., Garcia, M. G., Balasch, M. M., Spilde, M. N., Stone, F. D., Dapkevicius, M. D. L. N., ... & Northup, D. E. (2014). Comparison of bacterial diversity in Azorean and Hawai'ian lava cave microbial mats. *Geomicrobiology journal*, *31*(3), 205-220.

30. Riquelme, C., Dapkevicius, M. D. L. E., Miller, A. Z., Charlop-Powers, Z., Brady, S., Mason, C., & Cheeptham, N. (2017). Biotechnological potential of Actinobacteria from Canadian and Azorean volcanic caves. *Applied Microbiology and Biotechnology*, *101*(2), 843-857.

31. Tiwari, K., & Gupta, R. K. (2012). Rare actinomycetes: a potential storehouse for novel antibiotics. *Critical reviews in biotechnology*, *32*(2), 108-132.

32. Doroghazi, J. R., Albright, J. C., Goering, A. W., Ju, K. S., Haines, R. R., Tchalukov, K. A., ... & Metcalf, W. W. (2014). A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nature chemical biology*, *10*(11), 963-968.

33. Ogasawara, Y., Torrez-Martinez, N., Aragon, A. D., Yackley, B. J., Weber, J. A., Sundararajan, A., ... & Melançon, C. E. (2015). High-quality draft genome sequence of Actinobacterium Kibdelosporangium sp. MJ126-NF4, producer of Type II

polyketide azicemicins, using Illumina and PacBio technologies. *Genome announcements*, *3*(2), e00114-15.

34. Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., ... & Breitling, R. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic acids research*, *43*(W1), W237-W243.

35. Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., ... & Meyer, F. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, *9*(1), 75.

36. Delcher, A. L., Bratke, K. A., Powers, E. C., & Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, *23*(6), 673-679.

37. Stothard, P., & Wishart, D. S. (2005). Circular genome visualization and exploration using CGView. *Bioinformatics*, *21*(4), 537-539.

Chapter 3. Genomics/Bioinformatics-guided Discovery of Alloactinomicin from *Alloactinosynnema sp.* L-07

1. Introduction

As discussed in Chapter 1, even after extensive screening using conventional bioactivity-guided natural product discovery approach, 90% of the secondary metabolic potential of the Actinobacteria remains to be an untapped reservoir of natural products, not to mention the strains not intensively studied[1]. To unlock the treasure trove of chemical diversity hid in actinobacterial genomes, computational tools were developed to identify the full complement of secondary metabolite biosynthetic gene clusters (BGCs) encoded in sequenced genomes, in particular those cryptic gene clusters[2]. Compared with traditional bioactivity-guided screening approaches, one prominent advantage of genome mining approach is its ability to computationally identify BGCs and predict the scaffold or partial structure of their small-molecule products from the genomic data, which excludes the risk of rediscovery of known metabolites by in silico de-replication and facilitates targeting the gene clusters of interest from immense biosynthetic pathways for subsequent experimental characterization[2]. Although hundreds of small molecules were discovered over the last decade using genome mining approach, only limited researches were carried out in mining the genomes of rare Actinobacteria that were isolated from special environments such as cave, mangrove, and plant roots[3,4,5]. However, the sequenced genomes of diverse Actinobacteria have revealed that the genetic potential to produce secondary metabolites, including bacterial type II polyketides (PK-IIs), was widespread, and extended into many unexploited or under-exploited bacterial genera, which often contained genetic regions for the production of compounds with unexpected bioactivities or unprecedented scaffolds[3,4,5].

For instance, a novel type of polythioamide antibiotic closthioamide and a PK-II clostrubin with unusual structures were discovered from the obligate anaerobic bacteria, *Clostridium cellulolyticum* and *Clostridium beijerinckii*, respectively, through bioinformatics-guided mining of their sequenced genomes[6,7]. These outstanding studies strongly indicated that biologically active molecules with novel structural features could be unearthed with great possibilities from these poorly investigated bacterial species. *Alloactinosynnema sp*. L-07 is also such a rare Actinobacterium isolated from a soil sample collected from Lechuguilla Caverns, New Mexico, and its genome represents the first one sequenced in the genus *Alloactinosynnema*.

Alloactinosynnema is a relative new genus, proposed in 2010, that belongs to the suborder *Pseudonocardiaceae* of Actinobacteria[8]. Its name was proposed based on the observations that it displayed a similar morphology to *Actinosynnema* but chemotaxonomically distinct features and different phylogenetic positions (Figure 3-1). To our best knowledge, only two other *Alloactinosynnema* species, *A. album* gen. nov., sp. nov. and *A. iranicum* sp. nov., were reported in the literature to date[8,9]. They were described to contain a type III cell-wall composition (meso-diaminopimelic acid along with galactose and ribose as typical whole-cell sugar pattern), type II phospholipid pattern, and MK-9(H_4) as the major menaquinone[8,9].

In the course of KSα/β product chemotyping project described in Chapter 2, we identified *Alloactinosynnema* sp. L-07 that encoded a potential novel PK-II based on the KSα/β amplicon fingerprint analysis and sequenced its whole genome. To provide solid evidence for KSα/β amplicon chemotyping approach, here we reported the bioinformatic analysis of complete type II PKS gene cluster identified in the sequenced genome of

Alloactinosynnema sp. L-07 and the discovery of a novel pentangular polyphenols, termed alloactinomicin, as described in Results and Discussion.

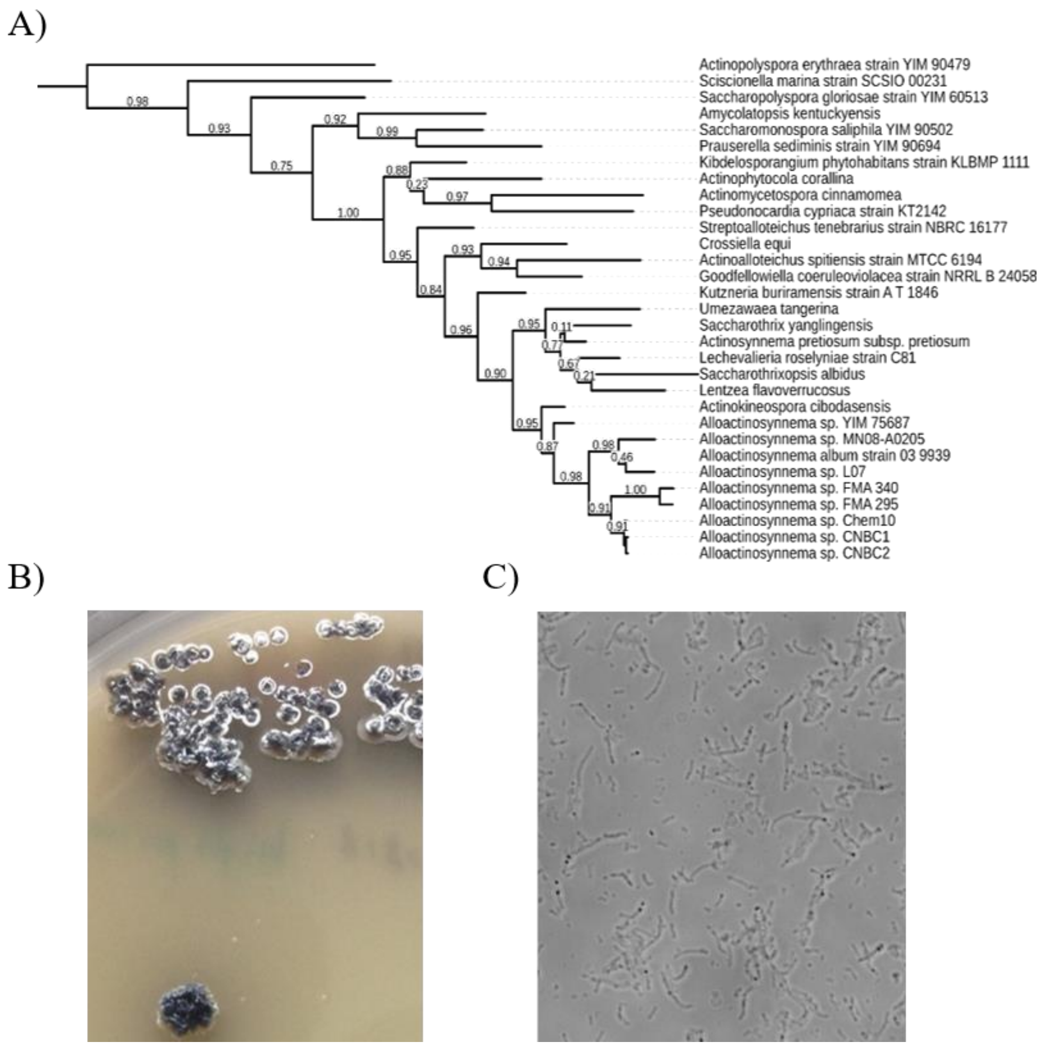

Figure 3-1. Strain information related to *Alloactinosynnema sp.* L-07. A) Phylogenetic tree of select Actinobacteria showing the position of *Alloactinosynnema sp.* L-07. B) Images of *Alloactinosynnema sp.* L-07 colonies. C) Micrograph of dispersed single cells of *Alloactinosynnema sp.* L-07. Bar, 10 μm.

In order to assign accurate functions for each gene of the PK-II BGC in *Alloactinosynnema* sp. L-07, a global comparative analysis of a training set of 17 studied bacterial pentangular gene clusters and their corresponding PK-II products and the phylogenetic analysis of homologous proteins, were carried out with particular focus on

aromatase/cyclase enzymes, C-11/C-19 ketoreductase, and flavin-dependent oxygenase. To predict the structure of final product of this PK-II BGC, further correlation analysis between the gene and certain structural features was performed according to the biosynthetic logic. This comprehensive bioinformatic analysis indicated that the structure of final product was a pentangular polyphenol with core structure featuring a quinone moiety and a lactam ring F.

As part of the structure characterization effort, we enhanced the titer of alloactinomicin through optimization of production media and fermentation conditions, as well as isolation, featuring a XAD-7 resin chromatography step, affording 3.8 mg of alloactinomicin at the end. This pentangular polyphenols were structurally elucidated as a 26-carbon benzo[a]tetracene quinone skeleton using high-resolution mass spectrometry (HRMS) and nuclear magnetic resonance (NMR). Results from structure elucidation provided experimental support for above bioinformatic analysis and structural prediction. Also, it connected the gene cluster to the molecules they produce, which offered new insights into gene cluster sequence/function relationship within the class of pentangular polyketides. This genomics/bioinformatics-guided approach provided a paradigm for the discovery of novel PK-II compounds in the future.

2. Experimental Materials and Methods

General. Most materials and methods used for the work in this chapter were identical to those described in the Experimental section of Chapter 2. Amberlite XAD-7 HP resin was purchased from Acros Organics (Thermo Fisher Scientific). Soybean flour type I was product of Sigma, and soluble starch was product from Acros Organics. NMR

tube was 5 mm Thin Wall Sample Tube Wilmad-528 purchased from Wilmad-LabGlass (Vineland, NJ). ^{1}H and ^{13}C NMR, DEPT and 2D NMR spectra were measured on a Bruker Avance III 500 MHz instrument. The chemical shift values (δ) were given in parts per million (ppm), and coupling constants (J) in hertz (Hz). ^{1}H and ^{13}C chemical shifts were referenced to the solvent residual peaks for DMSO-d_6 at δH 2.50 ppm and δC 39.5 ppm, respectively. All the organic solvents were purchased from Sigma-Aldrich (St. Louis, MO) or VWR (Radnor, PA) and used without further purification.

Bacterial Strains. *Alloactinosynnema sp*. L-07 is an environmental strain isolated from Lechuguilla Caverns, NM as described in Chapter 2. The spore suspension and frozen mycelia of *Alloactinosynnema* sp. L-07 were prepared as described in Chapter 2.

Instrumentation. The pH value determination, centrifugation, and strain cultivation were performed using identical equipment to those described in the Experimental section of Chapter 2. Micrograph imaging of strains was performed on Zeiss Axio Observer D1 inverted microscope (Jena, Germany). Solvents were removed using Buchi (New Castle, DE) R-215 Rotavapor equipped with Buchi V-850 Vacuum Controller, B-491 Heating Bath, and Welch DryFast Ultra Diaphragm 2032 Pump (Mt. Prospect, IL). Residual solvents were removed using Welch DuoSeal Vaccum 1402 Pump. HPLC analysis was performed on a Dionex Ultimate 3000 system equipped with a photo diode array (PDA) detector and the specific C18 column described in particular procedures. Analytical C18 HPLC column (Kromasil 100-5-C18, 5 µm, 3.0 x 150 mm) was product of AkzoNobel from Bohus, Sweden, while the another analytical C18 column (Thermo Scientific ODS Hypersil, 5 µm, 3.0 x 150 mm) was purchased from Thermo Scientific (Waltham, MA). Semi-preparative C18 HPLC column (Kromasil 100-5-C18, 5 µm, 10 x

250 mm) was also purchased from AkzoNobel. The water used for HPLC was obtained from Milli-Q Ultrapure Water Systems (MilliporeSigma, Darmstadt, Germany) and further purified by Phenomenex filter (Phenomenex, Torrance, CA) with 0.2 μM membrane (Sartorius Stedim Biotech, Goettingen, Germany). High resolution mass spectrometry data were obtained using a Waters LCT Premier ESI-TOF mass spectrometer housed in the Mass Spectrometry and Proteomics Core Facility in the Department of Chemistry and Chemical Biology at the University of New Mexico. NMR spectra were obtained using Bruker Avance III 500 spectrometers housed in the NMR Core Facility in the Depertement of Chemistry and Chemical Biology at the University of New Mexico.

Bacteria Cultivation. *Alloactinosynnema sp*. L-07 was maintained in MYM or GYM Streptomyces Medium (GYM) containing following ingredient: 0.4% w/v glucose, 0.4% w/v yeast extract, 1% w/v malt extract, 0.2% w/v $CaCO_3$, and 1.2% w/v agar. The pH of GYM medium was adjusted to 7.2 before autoclaving. The $CaCO_3$ and agar powder were removed if the liquid medium was used. The MYM agar medium was prepared by the addition of 4 g of maltose, 4 g of yeast extract, 10 g of malt extract and 15 g of agar into 1 L of deionized water, adjustment to pH 7.0.

Micrograph Imaging of Alloactinosynnema sp. L-07. The morphology of spore chains and structures was examined using 14-day old *Alloactinosynnema* sp. L-07 from GYM agar plate. The spores were scratched off the surface using pre-wet Q-tip and spread on glass slide, which was then covered with cover glass. The images were produced under 10x ocular lens (eyepiece) and 63x oil immersion lens on Zeiss Axio Observer D1 inverted microscope.

16S rDNA Amplification and Phylogenetic Analysis. The protocol used in this work was the same as Chapter 2. The 16S rRNA gene sequences of related species were retrieved from NCBI GenBank.

Preparation of XAD-7 Resin. This pretreatment of purchased XAD-7 resin is able to remove the salts or residues in it. It started by weighing out 100 g of XAD-7 resin, which was placed into a 500 mL bottle. Then 150 mL of 10% aqueous methanol (MeOH) was added to wash the resin by stirring for 10 min. The solvent was decanted, and 200 mL of water was used to wash the resin. The water was decanted, and another 200 mL water was used to wash the resin, which was repeated once. After that, 100 mL of MeOH was added to wash the resin once and 200 mL of water was used to wash the resin three time to remove the MeOH. Prior to adding to the fermentation culture, washed XAD-7 resin was resuspended in 150 mL of water and autoclaved for 20 min.

Optimization of Fermentation Conditions. For the preliminary fermentation, a small piece of MYM agar with fully grown colony was inoculated into 25 mL of GYM in a 125 mL flask or 50 mL of GYM in a 250 mL flask and grown at 30 ℃, 250 rpm in a rotary shaker for 7 days. Then 12-15 mL of above seed culture was inoculated into 500 mL of GYM liquid medium in the 2 L flasks and incubated at 30 ℃, 250 rpm in a rotary shaker for 7 days. The effects of different amount of GYM medium (50 mL, 200 mL, 500 mL, 600 mL and 1 L) in several flask sizes (250 mL, 500 mL, 1 L, 2 L) and the rotary speed (e.g. 50 rpm, 80 rpm, 100 rpm, 150 rpm, 200 rpm, 250 rpm) on the production medium were investigated.

To find out the optimal production media, a panel of media containing different carbon and nitrogen source was tested (Table 3-1). It started by inoculating cells of

Alloactinosynnema sp. L-07 into 25 mL of GYM liquid medium in a 125 mL flask with glass beads, which was grown at 30 ℃, 250 rpm in a rotary shaker for 7 days, since it showed better growth than 50 mL or 100 mL of GYM seed culture with or without glass beads. The 25 mL GYM seed culture was centrifuged at 3,000 rpm for 5 min and the cell pellets were resuspended in 4 mL 1/4 x Ringers' solution (0.45 g of NaCl, 2.3 mg of KCl and 14 mg of $CaCl_2 \cdot 2H_2O$ into 250 mL of deionized water, pH 7.0, autoclaved for 30 min). Each 50 mL culture of different production media was inoculated with 0.4 mL of such resuspended seed culture and incubated at 30 ℃, 200 rpm in a rotary shaker for 7 days.

Name	**Ingredients**	**Reference**
CGS	Cane molasses (2%) Glucose (0.5%) Soluble starch (3%) Pharmamedia (2%) pH 7.0	10
CPX	Soluble starch (1%) Glucose (1%) Bacto Peptone type 3 (1%) Glycerol (2%) Yeast extract (0.3%) $(NH_4)_2SO_4$ (0.2%) $CoCO_3$ (0.2%) pH 7.4	11
GMC	Glucose (1%) Millet meal (2%) Cottonseed meal (2%) MOPS (2%) pH 7.0	10
GOT	Glycerol (6%) Oat meal (1.5%) Tomato paste (0.5%) $CaCO_3$ (0.3%) pH 7.0	10
GYM	Glucose (0.4%) Yeast extract (0.4%) Malt extract (1%) pH 7.2	
NSG	Glucose (10 g/L) Soluble Starch (20 g/L) Yeast Extract (5 g/L) N-Z Amine type A (5 g/L) $CaCO_3$ (1 g/L) pH 7.3	
OM	Oat meal (3%) Meat extract (1%) pH 7.5	12
R5	Sucrose (103 g) casamino acids (0.1 g) glucose (10 g) yeast extract (5 g) TES (5.73 g) trace elements (2 mL) K_2SO_4 (0.25 g) $MgCl_2 \cdot 6H_2O$ (10.12 g) pH 7.2 before use,adding 20% L-proline/glutamate (15 mL) 2% $NaNO_3$ (15 mL) 0.5% KH_2PO_4 (10 mL) 1M $CaCl_2$ (20 mL) 1N NaOH (7 mL)	
SG	Soy bean flour (4%), Glucose (1%), $CaCO_3$ (0.25%) pH 7.0	this study
SGM	Soy bean flour (2%) Glycerol (4%) MES (1.95%) pH 6.8	13
SM	Soybean meal (2%) D-Mannitol (2%) $CaCO_3$ (0.2%)	14
SS	Soy bean flour (3%) Soluble starch (4%) $CaCO_3$ (0.25%) pH 7.0	12
SS+1	Soy bean flour (6%) Soluble starch (1.5%) $CaCO_3$ (0.25%) pH 7.1	this study
SYD	Soluble starch (6%) Dry yeast (1%) β-cyclodextrin (1%) $CaCO_3$ (0.2%) pH 6.8	15
TSB	Dextrose (0.25%) Casein peptone (1.7%) Soy peptone (0.3%) NaCl (0.5%) K_2HPO_4 (0.25%) pH 7.3	
TSB+2	TSB (1.5%) $CaSO_4$ (1.5%) Yeast extract (1.1%) Glucose (5%) Trace elements solution (0.2%)	12
Notes	The recipe for GYM and NSG media was from Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ), while the TSB was from manufacturer	
	Pharmamedia is an economical, finely ground, yellow flour made from the embryo of cottonseed, thus cottonseed powder was used	
	$CoCO_3$ was substitued by $CoCl_2$; Meat extract was replaced by Beef extract; enzymatic hydrolyzate of Casein is equivalent to N-Z Amine Type A	
	MOPS stands for morpholinepropanesulfonic acid, MES for 2-(N-morpholino)ethanesulfonic acid	
	Oat meal was grinded and sifted to obtain the powder	
	Trace Elements Solution: $ZnCl_2$ (40 mg/L) $CuCl_2 \cdot 2H_2O$ (10 mg/L) $MnCl_2 \cdot 4H_2O$ (10 mg/L) $FeCl_3 \cdot 6H_2O$ (200 mg/L) $Na_2B_4O_7 \cdot 10H_2O$ (10 mg/L) $(NH_4)_2Mo_7O_{24} \cdot 4H_2O$(10 mg/L)	
	$CaCO_3$ may increased the dissolved oxygen and viscosity, and may change pH in medium	

Table 3-1. Production media tested on *Alloactinosynnema sp.* L-07. These media are widely used for production of various actinobacterial secondary metabolites.

From the results of above media investigation, SS media was the one yielding highest amount of alloactinomicin. Based on this result, SS+1 medium with different carbon/nitrogen source ratio was tested, which was reported to have positive effect on the production of secondary metabolite[16]. In parallel, because soybean flour was reported to be superior to yeast extract as nitrogen source[16], the carbon source glucose was compared with the soluble starch at different fermentation conditions: two rotary shaker speed (80 rpm vs 200 rpm), with or without adding XAD-7 resin.

To examine the optimal ratio of the selected carbon and nitrogen sources, three different concentrations of soybean flour (10, 15 and 20 g/L) and three different concentrations of soluble starch (40, 60 and 80 g/L) were investigated. In parallel, the effect of addition of inorganic salts (2 g/L $(NH_4)_2SO_4$, 2 g/L NaCl and 0.5 g/L K_2HPO_4) into the production media was also studied. This time, 100 μL of 20% glycerol mycelia suspension was used for making 25 mL seed culture. After 7 days of growth, the seed culture was centrifuged at 4,000 *g* for 5 min, and the cell pellets were resuspended in 4 mL of fresh GYM liquid medium. Then 0.8 mL of above seed culture was inoculated into 100 mL of GYM larger seed culture. After 2 days of growth, the 100 mL seed culture was centrifuged at 4,000 *g* for 5 min, and the cell pellets were resuspended in 20 mL of fresh GYM liquid medium. The 0.4 mL of above seed culture was used to inoculate each 50 mL of production medium. Meanwhile, 5 mL of XAD-7 resin was added into each production medium, which was incubated at 30 °C, 80 rpm in a rotary shaker for 7 days. To further improve the yield of alloactinomicin, the above optimized condition was scaled up into 600 mL of SS and SS+1 media in a 2 L flask.

All these fermentation cultures of each batch were grown in parallel for performing the comparison. The metabolites were extracted and prepared for HPLC analysis as described below. The integration of peak areas at 460 nm were plotted versus time as shown in Results and Discussion section. The above optimized condition was scaled up into 600 mL medium in a 2 liter flask, but failed to produce any colored compounds.

Fermentation of Alloactinosynnema sp. L-07. After extensive optimization of fermentation conditions, a final version was used for the production of alloactinomicin. This final version of fermentation condition used 50 mL of the optimal medium SS+3 (soy bean flour (1.5%), soluble starch (4%), $CaCO_3$ (0.25%) pH 7.0) with addition of 5 mL of Amberlite XAD-7 resin and fermented at 28℃, 80 rpm for 7 days. For the preparation of seed culture, 100 μL of 20% glycerol mycelia stock of *Alloactinosynnema* sp. L-07 was inoculated into 25 mL of TSB or GYM liquid medium in a 125 mL Erlenmeyer flask with glass beads, which was incubated at 30 °C, 250 rpm in a rotary incubator. After 4 days of incubation, the seed culture was centrifuged at 4,000 *g* for 5 min, and the supernatant was discarded while the cell pellets were resuspended into 3 mL fresh GYM liquid medium. To inoculate large scale fermentation culture, the seed culture was first scaled up by inoculating 4% (v/v) into a 100 mL of TSB or GYM liquid culture in 500 mL flask with glass beads and grown in 30 °C, 250 rpm shaker. After 2 days of growth, the 100 mL seed culture was centrifuged at 4,000 *g* for 5 min, and the supernatant was discarded while the cell pellets were resuspended into 20 mL of fresh GYM liquid medium. Each 250 mL flask, 30 flasks in total, containing 50 mL of SS+3 medium was inoculated with 0.5 mL of above resuspended seed culture and 5 mL of autoclaved XAD-7 resin. These large scale fermentation cultures were maintained at 28 °C, 80 rpm for 7 days.

Extraction of Metabolites. Two methods were employed to extract the metabolites from fermentation cultures depending on whether XAD-7 resin was added or not. Without adding XAD-7 resin into the fermentation, the culture broth was harvested by centrifugation at 4,000 *g* for 5 min to remove the cells. The supernatant portion was adjusted to pH 3.0 to enhance the efficiency of extraction of PK-IIs since they often contain carboxyl group in the structure. Then the acidified supernatant was extracted with equal amount of ethyl acetate (EtOAc) twice and separated in a 4 L cone-shaped separatory funnel. The EtOAc phase was subjected to rotary evaporator to remove the solvents, and residual solvents were removed using high vacuum to yield the dry crude extract.

With XAD-7 resin added into the fermentation, the following steps were carried out to extract the metabolites. The XAD-7 resin was collected by decanting the unbound supernatant and washed with deionized water multiple times to remove the cells. The residual water was pipette out as much as possible. Then XAD-7 resin was soaked in appropriate amount of MeOH and stirred for 30 min. The above step was repeated twice to extract the metabolites as thoroughly as possible. The MeOH organic phase was filtered through a filter paper (Whatman) and subjected to rotary evaporator to remove the solvents, and residual solvents were removed using high vacuum to yield the dry crude extract.

HPLC Analysis of Alloactinosynnema sp. L-07 Metabolites. Samples used in HPLC analysis were prepared by a 1 mL sample of the growing GYM culture, subjecting the sample to centrifugation to remove cells and insoluble media components. The resulting supernatant was adjusted to pH 3.0 with concentrated HCl and extracted with an equal volume of EtOAc. After the solvent was removed by rotary evaporation, the crude extract was dissolved in 40 μL of 50% aqueous acetonitrile. Ten μL of sample was subjected to

HPLC analysis. To analyze the samples from fermentation with addition of Amberlite XAD-7 resin, the metabolites were extracted as described above. A little bit of sample was taken from the dry crude extract and used for HPLC analysis.

All samples prepared as above were analyzed on analytical C18 column (3 x 150 mm). The mobile phase consisted of solvent A (ultrapure H_2O containing 0.1% formic acid) and solvent B (HPLC grade acetonitrile containing 0.1% formic acid) was utilized at a flow rate of 0.6 mL min^{-1}. A multiple step gradient program (10% to 55% solvent B over 4 min, 55% to 75% B over 15 min, 75% to 98% B over 2 min, 98% B for 2 min, back to 10% B in 2 min) was employed to detect alloactinomicin both at 254 nm and 460 nm.

Purification of Alloactinomicin. The crude extract (~1.5 g methanol soluble compounds) obtained from 3 L large scale culture was fractionated on a silica gel column (35 g silica powder) using step-gradient elution: MeOH/$CHCl_3$ (0:100, 1:99, 2:98, 5:95, 10:90 and 100:0). These eluate were analyzed by HPLC on analytical C18 column and combined into 5 fractions. Then the fraction (306 mg) containing our targeted compounds was further purified by two successive rounds of HPLC on semi-preparative C18 column. A multiple step program (25% to 80% B over 3 min, 80% to 98% B over 10 min, 98% B for 1 min, back to 25% B in 1 min) was used first to roughly purify the targeted fraction. Then, a second multiple step program was used as follows: 25% to 55% B over 3 min, 55% to 80% B over 20 min, 80% to 98% B over 2 min, hold at 98% B for 3 min, then back to 25% B in 1 min. Solvent was removed by rotary evaporation and was dried under high vacuum overnight, yielding 3.8 mg of an orange solid. The purified alloactinomicin was subjected to HRMS and NMR analysis.

Structure Elucidation of Alloactinomicin. Mass spectrometric data for alloactinomicin was taken by high-resolution mass spectrometry (HRMS). The 3.8 mg of alloactinomicin sample was dissolved in 600 μL of newly purchased DMSO-d_6 solvent. Normal 1D and 2D NMR experiments (^{1}H, ^{13}C, DEPT, COSY, HSQC and HMBC) with standard parameters were carried out on a 500MHz Bruker Avance III 500 spectrometers. An additional openSW proton spectrum from 0 to 17 ppm was carried out, while an openSW carbon spectrum from 0 to 240 ppm was also performed. To solve the uncertain part of structure, the HMBC and HSQC 2D spectra with special pulse programs (hmbcetgpl3nd and hsqcetgpsp.2, respectively) were also taken by doubling both TDs. The NMR data were visualized and inspected in software MestReNova 9.0.1 by following the user manual, and manually curated into an Excel spreadsheet.

Bioinformatic Analysis. The *Dynamite* developed by Ben Yackley of our group used a series of protein sequences as query to identify a number of conserved protein families in PK-II BGCs in NCBI protein databank using the Blastp algorithm. Summaries of the sequence characteristics (including fingerprint analysis, specific combination of immediate tailoring enzymes) of all gene clusters were outputted as text files that could be easily viewed and further analyzed by user.

The secondary metabolite gene clusters were analyzed by the widely used antiSMASH web interface (http://antismash.secondarymetabolites.org/)[17]. The individual gene functions were annotated based on the Blastp search in NCBI and results from phylogenetic trees. The sequences of all pentangular PK-II BGCs were retrieved from NCBI database by *Dynamite*. The protein sequences for each genes were retrieved from

NCBI database using corresponding GI numbers, and phylogenetic trees was generated as described in Chapter 2.

3. Results and Discussion

Description of Alloactinosynnema sp. L-07. *Alloactinosynnema sp*. L-07 was an environmental strain isolated from Lechuguilla Caverns, NM as described in Chapter 2. This newly isolated Actinobacteria displayed slow growth on GYM agar plate with dark purple hyphae and white spores observed (Figure 3-1). These morphological features resembled the characteristics of member species from the genus *Alloactinosynnema*. The 16S rRNA gene sequence showed highest similarity to *Alloactinosynnema alba* strain 03-9939 in NCBI Blast search and clustered together with the members of the genus *Alloactinosynnema* in the phylogenetic tree (Figure 3-1). The G+C content of genomic DNA was determined to be 69.59% from genome sequencing results, which is similar to the other *Alloactinosynnema* species (Figure 3-2). Based on all these features displayed by this newly isolated strain, we proposed that it is a novel rare Actinobacteria belonging to the genus *Alloactinosynnema* and should be named as *Alloactinosynnema* sp. L-07.

Genome Mining of Secondary Metabolite BGCs. Recently, in the course of KSα/β product chemotyping, the whole genome of *Alloactinosynnema* sp. L-07 was sequenced using PacBio RSII platform as detailed in Chapter 2. *Alloactinosynnema* sp. L-07 had a linear chromosome with genome size of 7.4 Mbp and 6,875 genes (Figure 3-2). Because the genome of *Alloactinosynnema* sp. L-07 was assembled as one contig, the annotation of all secondary metabolite BGCs was greatly facilitated, especially those contain repetitive sequences like type I modular PKS or NRPS. An analysis performed by antiSMASH[17]

using the standard cluster rule based approach resulted in the revelation of a range of different natural product classes including the identified type II PKS gene cluster, named the *allo* cluster (Figure 3-3), in *Alloactinosynnema* sp. L-07 genome.

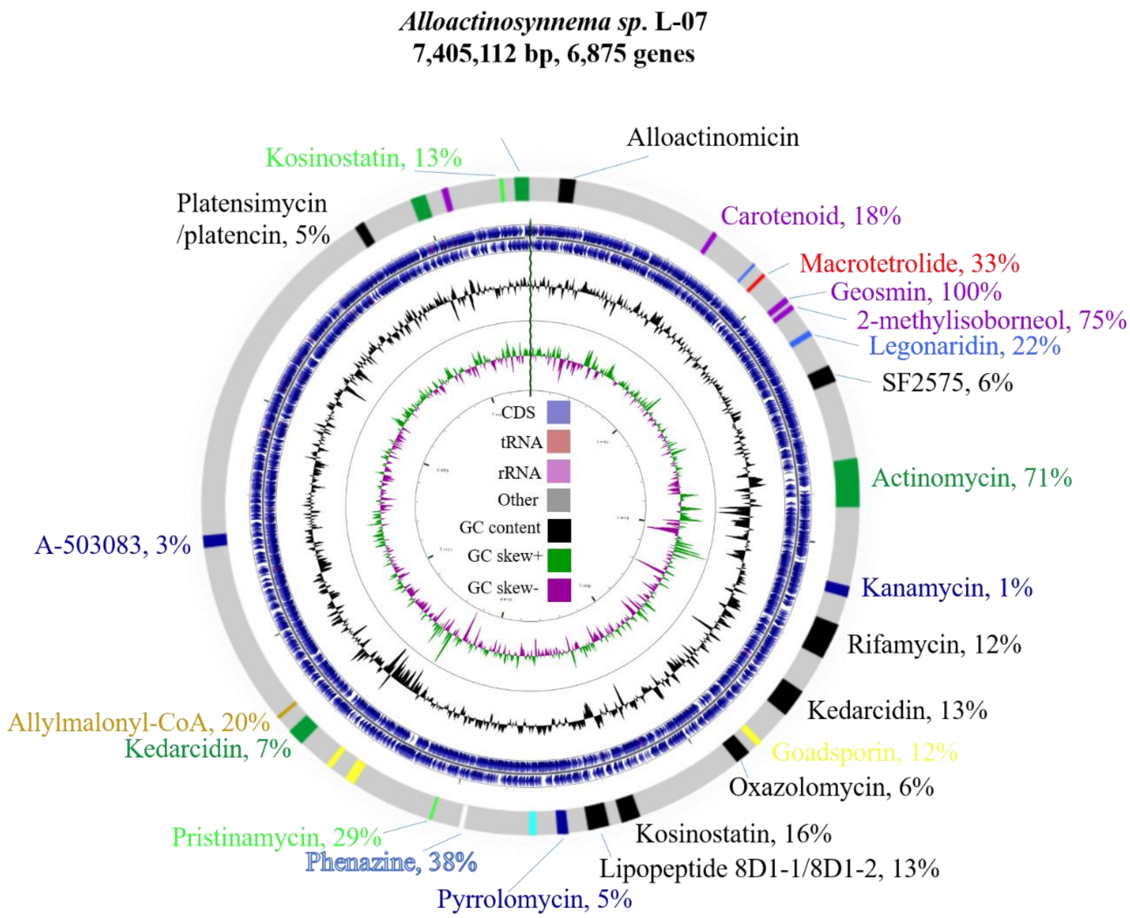

Figure 3-2. Circular map of the genome of *Alloactinosynnema sp.* L-07. The inner rings show a normalized plot of GC skew, while the middle rings show a normalized plot of GC content and coding sequences (CDSs) on both strands. The outer circles show the distribution of secondary metabolite gene clusters annotated by antiSMASH. The predicted product of each BGC is color coded and labeled with similarity percentage. This map was generated by CGView[18] and DNAplotter[19].

About 14% of *Alloactinosynnema* sp. L-07 genome is dedicated to secondary metabolism, which is comparable compared to the model Actinobacterium *Streptomyces coelicolor* A3(2). In total, 30 putative secondary metabolite biosynthetic gene clusters were identified and they were distributed uniformly across the chromosome (Figure 3-2). Three

clusters of *Alloactinosynnema* sp. L-07 encode the biosynthesis of PKS compounds, while four clusters encode the biosynthesis of nonribosomal peptides and four clusters encode the biosynthesis of hybrid PKS/NRPS compounds. In *Alloactinosynnema* sp. L-07, 14 gene clusters have no significant similarities to characterized pathways and either code for analogs of known metabolites, whose biosynthetic pathways were not yet linked or potentially novel natural products that have not been discovered before. In the following sections, detailed bioinformatic analysis of the PK-II BGC, the *allo* cluster, as well as the compound isolation and structural characterization of the PK-II were described.

Bioinformatic Analysis of PK-II Gene Cluster. In previous work, our lab has developed a bioinformatics software package, *Dynamite*, which globally identifies and annotates all PK-II BGCs currently deposited in the NCBI databank. After identifying ~600 putative PK-II BGCs, a comparative analysis of genes within these clusters was carried out to identify and classify them into subclasses based on the sequence characteristics. Fourteen functionally distinct types of aromatase/cyclase (AroCyc) and cyclase (Cyc) enzymes belonging to 7 unrelated protein families were recognized, and the presence of specific sets of AroCyc/Cyc types within a PK-II BGC were correlated with the cyclization/folding pattern undergone by the poly-β-ketone product to form a particular core structure. A number of tailoring enzymes were also identified and classified into subclasses, such as ketoreductase (KR), monooxygenase (ABM, MOX, FOX) and dehydrogenase (DHyII) (Table 3-2).

Dynamite analysis of the proteins encoded by genes adjacent to KSα/β genes of PK-II BGC in the *allo* cluster revealed several other proteins: an acyl carrier protein (ACP), three aromatase/cyclases (AroCycN2, Cyc1, Cyc2), two antibiotic biosynthesis

monooxygenases (ABMph, ABMi), two ketoreductase (KRpen11, KRpen19), and two putative monooxygenases (MOXpen), as well as several proteins with homology to those involved in post-modification of core structure and regulation of gene expression (Table 3-3, Figure 3-3). This preliminary analysis of each putative protein in the *allo* cluster indicates the final product structure encoded by this PK-II BGC is an acetate primed at least 26-carbon pentangular polyketide.

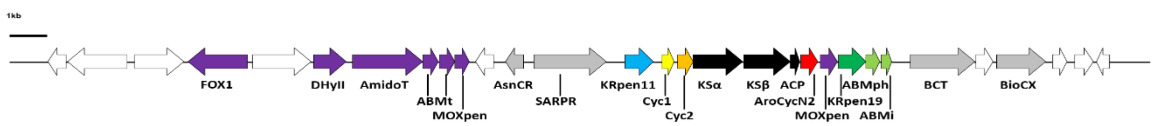

Figure 3-3. The gene organization of *Alloactinosynnema sp.* L-07 PK-II gene cluster. Proposed functions for individual open reading frames are shown in Table 3-3.

Comparative Analysis of Closely Related Pentangular PK-II Gene Clusters. To gain deeper insight into the structure of the pentangular PK-II encoded by the *allo* gene cluster, further comparative analysis was carried out by careful and detailed comparison with existing identified pentangular polyketide gene clusters (Table 3-2, Figure 3-4). These pentangular PK-II BGCs include the *llp* cluster encoding lysolipin from *Streptomyces tendae*[20], the *xan* cluster encoding xantholipin from *Streptomyces flavogreiseus*[21], the *pnx* cluster encoding FD-594 from *Streptomyces sp.* TA-0256[22], the PK-II gene cluster encoding arenimycins A&B from *Salinispora arenicola* CNB-527[23], the *arn* cluster encoding arenimycins C&D from uncultured bacterium BAC-AB1442/1414/561[24], the *arx* cluster encoding arixanthomycin from uncultured bacterium[25], the *clx* cluster encoding calixanthomycin from uncultured bacterium[24], the PK-II gene cluster encoding TLN-05220 from *Micromonospora echinospora subsp. Challisensis*[26], the *fdm* cluster encoding fredericamycin from *Streptomyces griseus*[27], the *san* cluster encoding A-74528 from

Streptomyces sp. SANK 61196[28], the *rub* cluster encoding rubromycin from *Streptomyces collinus*[29], the *grh* cluster encoding griseorhodin from both *Streptomyces sp.* JP95[30] and *Streptomyces sp.* CN48+, the PK-II gene cluster encoding fasamycin from uncultured bacterium[31], the *fkm* cluster encoding frankiamicin from *Frankia sp.* EAN1pec[32], the *ben* cluster encoding benastatin from *Streptomyces sp.* A2991200[33], and the *pdm* cluster encoding pradimicin from *Actinomadura hibisca*[34].

Compound	**Species**	**starter-extender**	**KSα**	**KSβ**	**ACP**	**AroCycN2**	**Cyc1**	**Cyc2**	**ABMi**	**ABMph**	**KRpen19**	**KRpen11**
alloactinomicin	Alloactinosynnema sp. L-07	acetyl-12	929020675	929020678	929020680	929020681	929020673	929031750	929020692	929020689	929020687	929020671
			alloA	alloB	alloC	alloC1	alloC3	alloC2	alloO1	alloO2	alloD1	alloD2
			xanF(80%)	llpE(67%)	llpD(52%)	grhT(68%)	llpCIII(73%)	xanC2(72%)	xanO6(56%)	llpOIII(69%)	llpZI(73%)	xanZ4(65%)
lysolipin	Streptomyces tendae	acetyl-12	154623217	154623216	154623215	154623214	154623219	154623218	154623210	154623211	154623212	154623242
			llpF	llpE	llpD	llpCI	llpCIII	llpCII	llpOII	llpOIII	llpZI	llpZIII
xantholipin	Streptomyces flavogriseus	acetyl-12	292386134	292386133	292386141	292386132	292386142	292386135	292386128	292386129	292386130	292386140
			xanF	xanE	xanD	xanC1	xanC3	xanC2	xanO6	xanO7	xanZ3	xanZ4
FD-594	Streptomyces sp. TA-0256	butyryl-12	316997093	316997094	316997095	316997096	316997091	316997092	316997100	316997099	316997098	316997120
			pnxA	pnxB	pnxC	pnxD	pnxK	pnxL	pnxI	pnxH	pnxG	pnxW
arenimycins A & B	Salinispora arenicola CNB-527	acetyl-11	655984133	655984134	655984136	655984127	757713290	655984126	757713315	757713292	655984135	
			CDS34	CDS35	CDS39	CDS26	CDS24	CDS25	CDS27	CDS40	CDS36	
arenimycins C & D	Uncultured bacterium BAC-AB1442/1414/561	acetyl-11	712001773	712001774	712001778	712001763	724052196	724052197	712001764	712001779	712001775	
			arn31	arn32	arn36	arn 21	arn19	arn20	arn22	arn37	arn33	
arixanthomycin	uncultured bacterium	acetyl-12	613432370	613432371	613432372	613432373	613432368	613432369	613432377	613432376	613432375	613432381
			arx16	arx17	arx18	arx19	arx14	arx15	arx23	arx22	arx21	arx27
calixanthomycin	uncultured bacterium	acetyl-12	745698432	745698431	745698430	745698429	745698452	745698451	745698428		745698422	
			clx11	clx10	clx9	clx8	clx31	clx30	clx7		clx1	
TLN-05220	Micromonospora echinospora subsp. Challisensis	2-methylbutyryl-12	283484105	283484106	283484107	283484108	283484103	283484104	283484111	283484110	283484109	283484101
			ORF18	ORF19	ORF20	ORF21	ORF16	ORF17	ORF24	ORF23	ORF22	ORF14
fredericamycin	Streptomyces griseus	hexadienyl-12	33327096	33327097	33327098	33327099	33327094	33327095	33327108	33327107, 33327100	33327106	
			fdmF	fdmG	fdmH	fdmI	fdmD	fdmE	fdmQ	fdmP, fdm J	fdmO	
A-74528	Streptomyces sp. SANK 61196	hexadienyl-12	296046088	296046089	296046090	296046091	296046086	296046087	296046100	296046099, 296046092	296046098	
			sanF	sanG	sanH	sanI	sanD	sanE	sanQ	sanP, sanJ	sanO	
rubromycin	Streptomyces collinus	acetyl-12	9944994	9944995	9944996	22477117	22477116	9944993	22477132	9944999	9944998	
			rubA	rubB	rubC	rubF	rubE	rubD	rubT	rubH	rubG	
griseorhodin (JP95)	Streptomyces sp. JP95	acetyl-12	21039488	21039489	21039490	21039520	21039513	21039517	21039519	21039518	21039520	21039503
			grhA	grhB	grhC	grhT	grhRQ	grhS	grhV	grhU	grhT	grhO10
griseorhodin (CN48+)	Streptomyces sp. CN48+	acetyl-12	662748189	662748190		662748191	662748187	662748188	662748194	662748193	662748191	662748192
			grhA	grhB	not found	grhT	grhRQ	grhS	grhV	grhU	grhT	grhO10
fasamycin	uncultured bacterium	acetyl-12	343479100	343479099	343479098	343479097	343479096	343479101	343479092			
			ORF23	ORF22	ORF21	ORF19	ORF20	ORF24	ORF15			
frankiamicin	Frankia sp. EAN1pec	acetyl-11	158109628	158109629	158109625	158109626	158109624	158109627	158109632	158109631	158109630	
			fkmA	fkmB	fkmC	fkmC1	fkmC3	fkmC2	fkmO2	fkmO1	fkmD	
benastatin	Streptomyces sp. A2991200	hexanoyl-11	169402965	169402966	169402967	169402968	169402963	169402964	169402969	169402968	169402971	
			benA	benB	benC	benH	benE	benD	benJ	benH	benL	
pradimicin	Actinomadura hibisca	acetyl-9	120431566	120431567	120431568	120431569	120431576	117956303	120431574	120431573	120431572	
			pdmA	pdmB	pdmC	pdmD	pdmK	pdmL	pdmI	pdmH	pdmG	

Table 3-2. Comparative analysis of *allo* cluster biosynthetic enzymes and their homologous proteins in pentangular training set clusters. The protein GI number in NCBI is shown above each gene name. The closest match of individual genes in the *allo* cluster are also shown with similarity percentages. To be readable, only genes encoding the core structures are shown in the table.

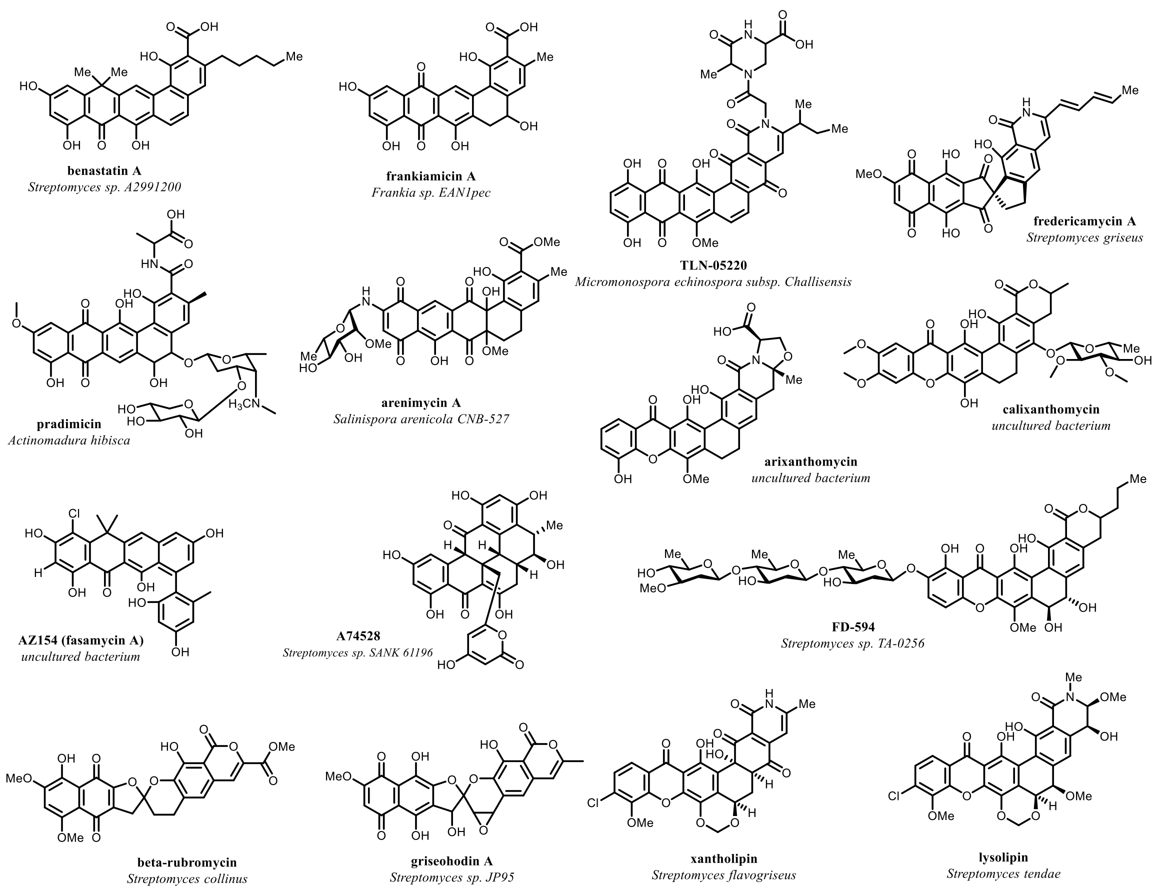

Figure 3-4. The structures of pentangular polyketides encoded by above training set clusters. The compounds are arranged based on their structural similarities. The compound name is given above each strain name.

To facilitate the discussion of genes encoded in the *allo* cluster from *Alloactinosynnema sp.* L-07, each gene was assigned with a systematic name. These designated names, their corresponding GI numbers (protein ID in NCBI database), gene size, closest homologous proteins and proposed functions are summarized in Table 3-3. Most of the genes in the *allo* cluster encode proteins sharing high similarities with enzymes involved in the biosynthetic pathways of lysolipin[20], xantholipin[21] and FD-594[22] (Table 3-

3). Thus, based on the biosynthesis of these three pentangular polyphenols, more accurate and detailed analysis is described below.

ORF	GI#	Size (bp)	Homologues and % Identity			NCBI Annotation	Proposed Function
alloO5	929020652	1563	*llpOV*	*xanO5* (42%)	*pnxO4*	hydroxylase WhiE VIII	FAD-dependent monooxygenase (FOX1)
alloT2	929020655	1485	N/M	N/M	N/M	Di/tripeptide transporter	transporter
alloF	929020657	912	*llpS* (41%)	*xanS1* (51%)	*pnxO7* (49%)	reductase	NAD-dependent dehydrogenase (DHyII)
alloE	929020659	1869	*llpA*	*xanA*	N/M	asparagine synthase	δ-lactam formation (AmidoT)
alloM2	929020661	396	*llpV* (69%)	*xanV* (67%)	*pnxO8* (55%)	amidotransferase	monooxygenase (ABMt)
alloM1	929020663	378	*llpT* (57%)	*xanT* (58%)	*pnxO8* (55%)	WhiE I protein paralog	monooxygenase (ABMt)
alloO4	929020664	456	*llpQ* (58%)	*xanO10* (58%)	*pnxE1* (46%)	monooxygenase GrhM	hydroxylation (MOXpen)
alloG	929020666	528	N/M	N/M	N/M	hypothetical protein	unknown
alloR1	929020668	498	N/M	N/M	*pnxR1* (38%)	MarR family transcriptional regulator	cluster regulation (AsnCR)
alloR2	929020670	1923	*llpRIV* (47%)	N/M	*pnxR2* (39%)	regulatory protein	cluster regulation (SARPR)
alloD2	929020671	741	*llpZIII* (65%)	*xanZ4* (65%)	*pnxG* (51%)	reductase	C-11 ketoreduction (KRpen11)
alloC3	929020673	336	*llpCIII* (73%)	*xanC3* (62%)	*pnxK*	cyclase WhiE VII	D, E ring cyclization (Cyc1)
alloC2	929031750	477	*llpCII* (66%)	*xanC2* (72%)	*pnxL*	cyclase WhiE II	C ring cyclization (Cyc2)
alloA	929020675	1254	*llpF* (79%)	*xanF* (80%)	*pnxA*	β-ketoacyl synthase WhiE-KS paralog	ketosynthase α (KSα)
alloB	929020678	1236	*llpE* (67%)	*xanE* (66%)	*pnxB*	β-ketoacyl synthase WhiE-CLF paralog	ketosynthase β (KSβ)
alloC	929020680	261	*llpD* (52%)	*xanD*	*pnxC* (49%)	acyl carrier protein	acyl carrier protein (ACP)
alloC1	929020681	459	*llpCI* (63%)	*xanC1* (64%)	*pnxD*	aromatase WhiE VI	A, B ring cyclization (AroCycN2)
alloO3	929020685	471	*llpB* (60%)	*xanO8* (67%)	*pnxE2* (46%)	hypothetical protein	monooxygenase (MOXpen)
alloD1	929020687	753	*llpZI* (73%)	*xanZ3* (70%)	*pnxG* (57%)	reductase	C-19 ketoreduction (KRpen19)
alloO2	929020689	369	*llpOIII* (69%)	*xanO7* (61%)	*pnxH* (60%)	monooxygenase	quinone formation (ABMph)
alloO1	929020692	309	*llpOII* (48%)	xanO6 (56%)	*pnxI* (41%)	monooxygenase	quinone formation (ABMi)
alloT1	929031751	171	N/M	N/M	N/M	hypothetical protein	transporter
alloN3	929020694	1719	N/M	*xanB3* (63%)	N/M	acetyl-CoA carboxyl transferase	carboxyl transferase (BCT)
alloN2	929031752	540	N/M	*xanB2* (45%)	N/M	biotin carboxyl carrier protein of acetyl-CoA carboxylase	acetyl-CoA biotin carboxylase, BCCP
alloN1	929020696	1356	N/M	*xanB1* (70%)	N/M	biotin carboxylase of acetyl-CoA carboxylase	acetyl-CoA biotin carboxylase (BioCX)

Table 3-3. Homologous proteins and proposed functions of genes in the *allo* cluster.

The minimal type II PKS, AlloA (KSα), AlloB (KSβ) and AlloC (ACP) proteins, probably act in a concert manner to synthesize the 26-carbon poly-β-ketone intermediate via 12 extension cycles of Claisen-like C-C condensation with an acetate starter unit. Three cyclases in *allo* cluster, AlloC1, AlloC2 and AlloC3, showed high similarities with cyclases LlpCI/XanC1, LlpCII/XanC2, LlpCIII/XanC3, respectively, from the biosynthesis of lysolipin and xantholipin. Homologous proteins of these three cyclases were shown to be essential for the formation of properly cyclized and aromatized core structures, indicating the product of the *allo* cluster belongs to either pentangular or tetracenomycin subclasses. PdmD, homologues of proposed aromatase/cyclase N-terminal domain (AroCycN2) AlloC1, has been demonstrated to involve in the cyclization and aromatization of both the

A and B rings of the nascent poly-β-ketone chain[35]. Thus, AlloC1 is very likely responsible for the closure of A and B rings at C9-C14 and C7-C16. PdmL, homologues of proposed cyclase type 2 (Cyc2) AlloC2, and PdmK, homologoues of proposed cyclase type 1 (Cyc1) AlloC3, were shown to involve in the cyclization and aromatization of C ring, D ring and E ring in its pathway[36]. Prior to cyclization, the C-11 and C-19 ketoreduction on the intermediate poly-β-ketone chain are proposed to be catalyzed by AlloD2, homologue of LlpZIII (65% identity) and XanZ4 (65% identity), and AlloD1, which show high similarity to LlpZI, XanZ3 and is similar to the characterized C-19 ketoreductase (KRpen19) BenL in benastatin pathway[37]. The function of AlloD2 and its homologous protein, LlpZIII and XanZ4, are deduced based on that they cluster in obviously distinct clade from KRpen19 on the phylogenetic tree (Figure 3-5).

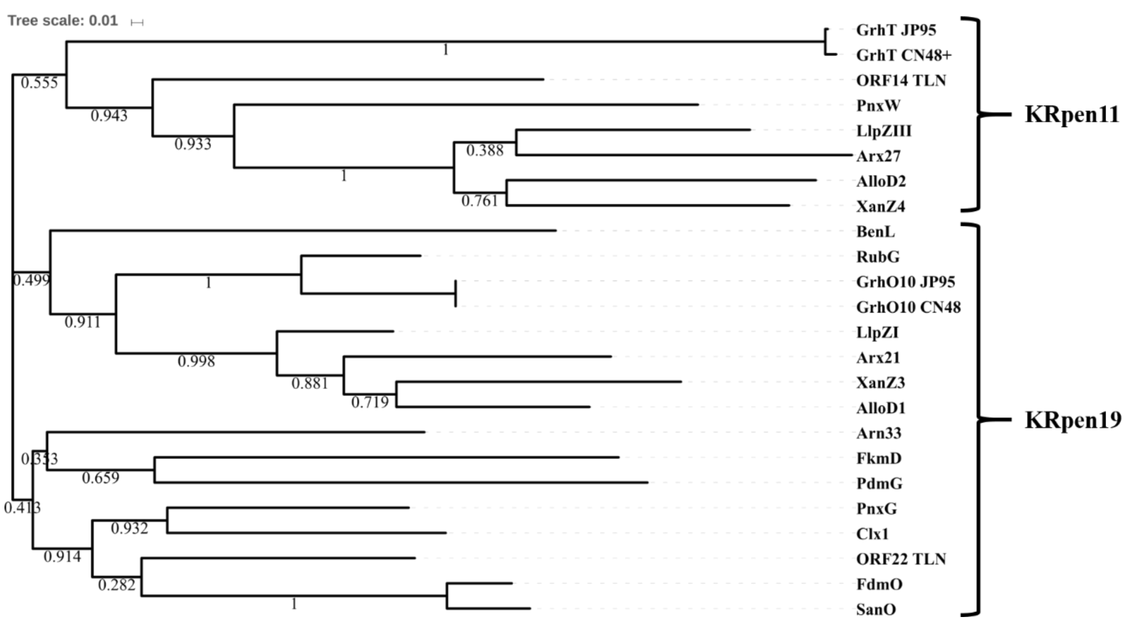

Figure 3-5. Phylogenetic tree of KRpen11 and KRpen19 present in all training set pentangular clusters. KRpen11 display a distinct clade from KRpen19.

Two antibiotic biosynthesis monooxygenase (ABM) superfamily enzymes, AlloO1 (ABMi) and AlloO2 (ABMph), which display closest similarity to XanO6 and LlpOIII,

respectively, are identified to be encoded by adjacent co-directional genes in the cluster. PdmH, homologue of AlloO2, has been proved to hydroxylate B ring to form the quinone moiety and was required for production of fully cyclized pentangular aglycon of pradimicin[36]. Another two pentangular monooxygenases (MOXpen), AlloO3 and AlloO4, may be involved in the hydroxylation of the aromatic rings according to the functions of their corresponding homologous enzymes, FdmM1 and FdmM, which have been shown to catalyze C6 and C8 hydroxylations in fredericamycin[38].

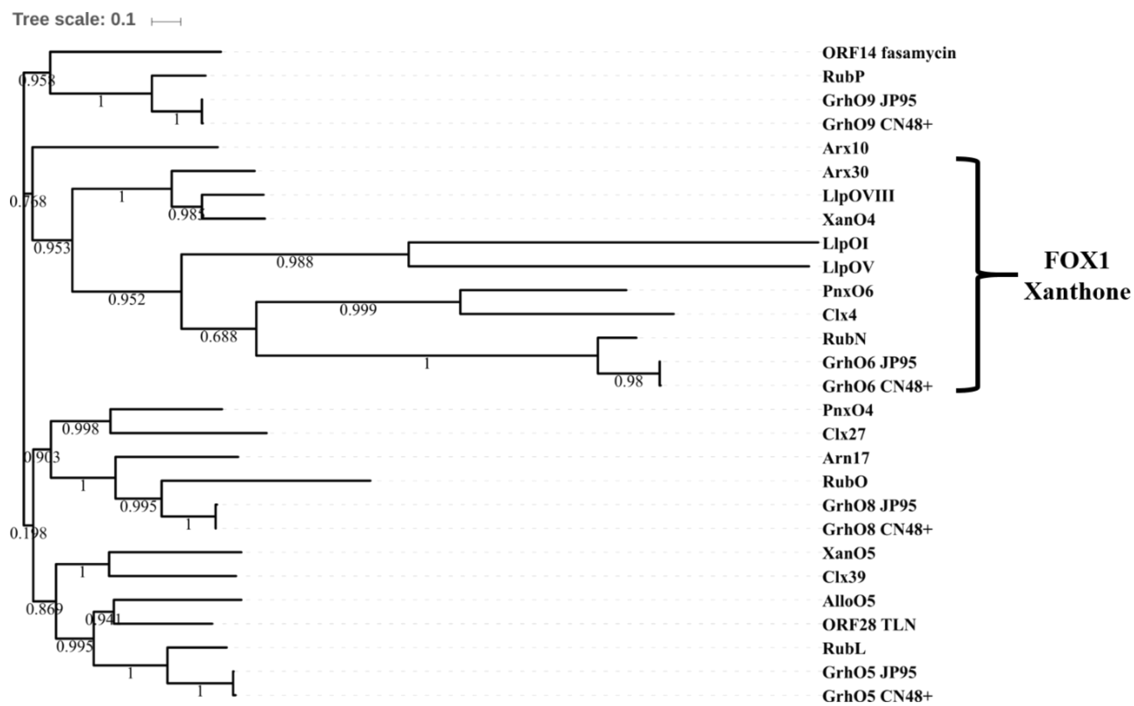

Figure 3-6. Phylogenetic tree of FOX1 present in all training set pentangular clusters. XanO4 and its homologues required for generation of xanthone structure form a distinct clade from other FOX1, and AlloO5 is distant from this clade.

An additional ABM superfamily member, AlloM1, has a closest homologous protein, XanT, in xantholipin biosynthesis, which was suggested to support the formation of the xanthone moiety. However, inactivation of XanT did not abolish the production of the final product[21]. The formation of the xanthone moiety in xantholipin has been shown

to require a flavin-dependent oxygenase type 1 (FOX1), XanO4, which might catalyze a Baeyer-Villiger oxidation, converting the quinone to a lactone. Homologous proteins (Figure 3-6) are also identified in the biosynthesis of other xanthone-containing pentangular polyketides, such as LlpOVIII (lysolipin), PnxO6 (FD-594), Arx30 (arixanthomycin) and Clx4 (calixanthomycin), and in the biosynthesis of rubromycin (RubN) and griseohodin (GrhO6), which have severely rearranged rings. Although the process of xanthone formation remains unclear, this comparative analysis strongly suggests that the C-C bond cleavage at B ring catalyzed by XanO4 and homologues is required for the generation of xanthone structure. One FOX1 protein, AlloO5, is also identified in the *allo* cluster, but it displays a distant relationship to the clade of XanO4 (Figure 3-6). Furthermore, AlloO5 shows 42% identity to XanO5, which has been suggested to catalyze the C4 hydroxylation of the aromatic scaffold. Thus, all these information may indicate no xanthone moiety is present in the final product of the *allo* cluster.

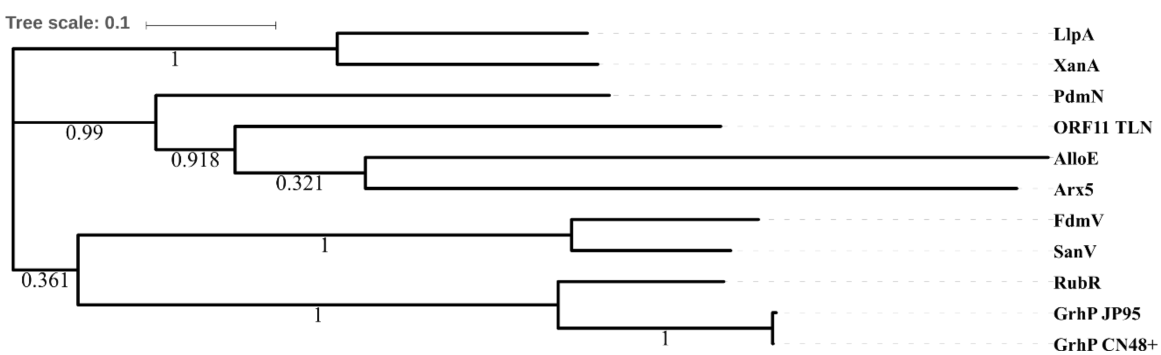

Figure 3-7. Phylogenetic tree of AmidoT present in all training set pentangular clusters. AlloE displays closest similarity to Arx5, whose corresponding pentangular polyketide contains a δ-lactam ring F.

Intriguingly, gene encoding an amidotransferase (AmidoT) that resembles asparagine synthase was identified in the *allo* cluster. Homologous protein, XanA, was characterized to be responsible for the introduction of the nitrogen and the formation of δ-lactam ring F. A phylogenetic inspection of AlloE together with annotated AmidoT from

other pentangular polyphenol clusters (Figure 3-7) revealed that AlloE fell into the subclade containing Arx5, ORF11_TLN, and PdmN. The corresponding pentangular polyketides, arixanthomycin and TLN-05220/05223, have a δ-lactam ring F in their skeletons, while the pradimicin has an amino acid group incorporated. Therefore, the scaffold of the final product from the *allo* cluster most likely contains a characteristic δ-lactam ring F.

The enzymes AlloN1-N3, showing high similarity to XanB1-B3, respectively, are suggested to be responsible for the supplement of malonyl-CoA by undergoing carboxylation of acetyl-CoA. These enzymes were hypothesized to be paralog of the enzymes that supply malonyl-CoA for fatty acid biosynthesis. Homologous proteins are also found in pradimicin biosynthesis, but not identified in the *llp* and *pnx* cluters.

Taken together, we predict the *allo* cluster encodes a pentangular polyphenol with core structure featuring a quinone moiety and a lactam ring F (Figure 3-8). Compared with the structures of reported pentangular polyketides, the predicted structure is obviously different, suggesting a high probability to discover structurally novel PK-IIs from *Alloactinosynnema sp.* L-07.

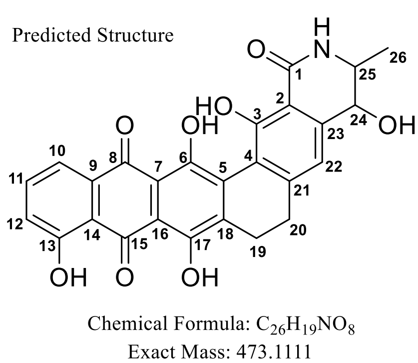

Figure 3-8. Predicted structure of the product of the *allo* cluster based on comparative analysis.

Chromatographic and Spectral Analysis of Alloactinosynnema sp. L-07. In order to isolate the bioinformatically predicted PK-IIs from *Alloactinosynnema sp.* L-07, this actinobacterial strain was grown in a small scale with recommended GYM medium. After the growth of ten days, the medium culture turn into dark purple color (Figure 3-9a), which may indicate the expression of this PK-II BGC in *Alloactinosynnema sp.* L-07, reasoning that pentangular PK-IIs are usually colorful compounds due to the large conjugation system in their scaffolds. The supernatant of medium culture was extracted by ethyl acetate and the expression of the targeted PK-II pathway was further supported by analyzing the extract using analytical HPLC (Peak1 and Peak2 in Figure 3-9b), showing characteristic uv-visible spectrum of pentangular compounds: strong absorption at 240 nm, 302 nm, and 463 nm (Figure 3-9c). These two peaks were collected and analyzed by high-resolution electrospray ionization mass spectrometry (HR-ESI-MS), showing ESI-negative [M-H] calculated for m/z=473.0873; found 473.0875 for the major compound (Figure 3-9d) and ESI-negative [M-H] calculated for m/z=473.0873; found 473.0870 for the minor compound (Figure 3-9f). All these preliminary experimental data supports our prediction that the *allo* gene cluster from *Alloactinosynnema sp.* L-07 produces pentangular compounds (Peak1 and Peak2 in Figure 3-9). The uv-visible spectra of the major and minor compounds (Figure 3-9d,f) closely resembled each other, displaying peaks at ~300 and ~460 nm in the visible region, and have the identical MS, suggesting that they are congeners.

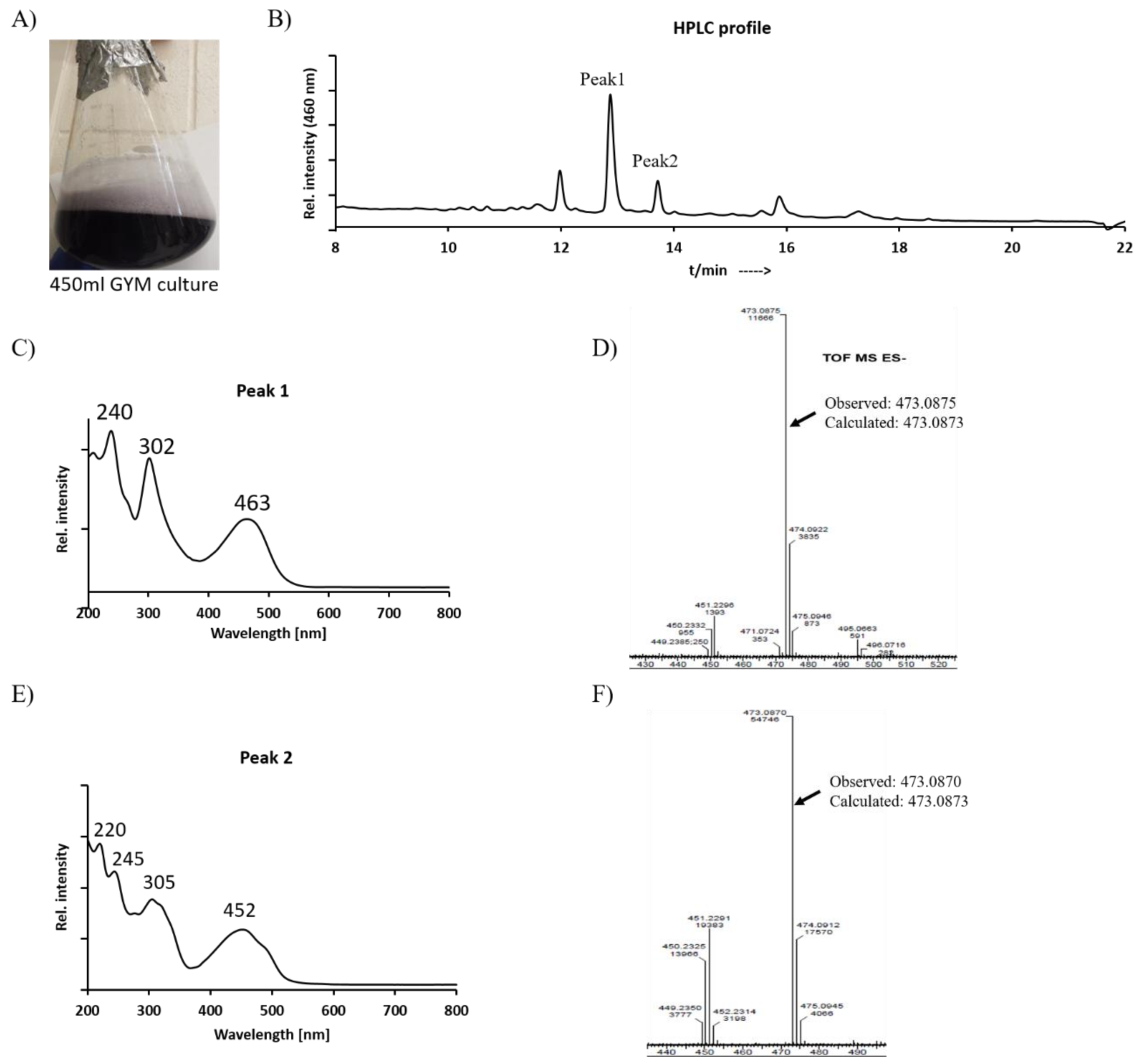

Figure 3-9. UV-visible and mass spectral analysis of *Alloactinosynnema sp.* L-07 metabolites. A) Small volume of GYM medium fermentation, showing the purple culture broth. B) HPLC analysis of extracts from the GYM culture, showing the presence of the major compound (labeled Peak1) and the minor compound (labeled Peak2). C) UV-visible spectrum of the major compound. D) Major peak, negative mode $[M-H]^-$. E) UV-visible spectrum of the minor compound. F) Minor peak, negative mode $[M-H]^-$.

Medium and Fermentation Condition Optimization. After the HPLC and HRMS analysis, it is reasonable to regard the two peaks as pentangular compounds. However, the low titer of target compound from the wild-type strain restrained the subsequent NMR structure elucidation and bioactivity assay. Owing to the lack of experience in isolation of

natural product from complex fermentation culture, this work was only focused on isolation the major compound, termed alloactinomicin.

Because media composition and cultivation conditions are known to have a crucial factors for the natural product production[39], investigation of a number of different media and fermentation conditions is critical and usually the first step to take to improve the yield of the target compound. Thus, a panel of media (Table 3-1 in Experimental section) that contain different carbon/nitrogen sources and facilitate the production of secondary metabolites in Actinobacteria were chosen from literature to test the production of alloactinomicin. The metabolites extracted from each of these media were analyzed by analytical HPLC and the values of peak area at 460 nm were normalized to compare with each other. The comparison analysis based on the values of peak area at 254 nm or peak height gave the same trends. While other media showed no significant yield improvement or lower yield of alloactinomicin compared with that from the original GYM medium (GYM 80 rpm in Figure 3-10), the SS medium showed a larger than 2-fold enhancement on the yield of alloactinomicin (SS 80 rpm in Figure 3-10). Furthermore, SS+1 with different carbon/nitrogen source ratio displayed near 5-fold enhancement (SS+1 80 rpm in Figure 3-10) in alloactinomicin production. When the soluble starch was replaced by glucose, the yield of alloactinomicin was decreased to a lower level (SG vs SS in Figure 3-10). Therefore, these fermentation testing suggest that soluble starch is the preferred carbon source to support the alloactinomicin production, and the SS+1 medium was pursued further.

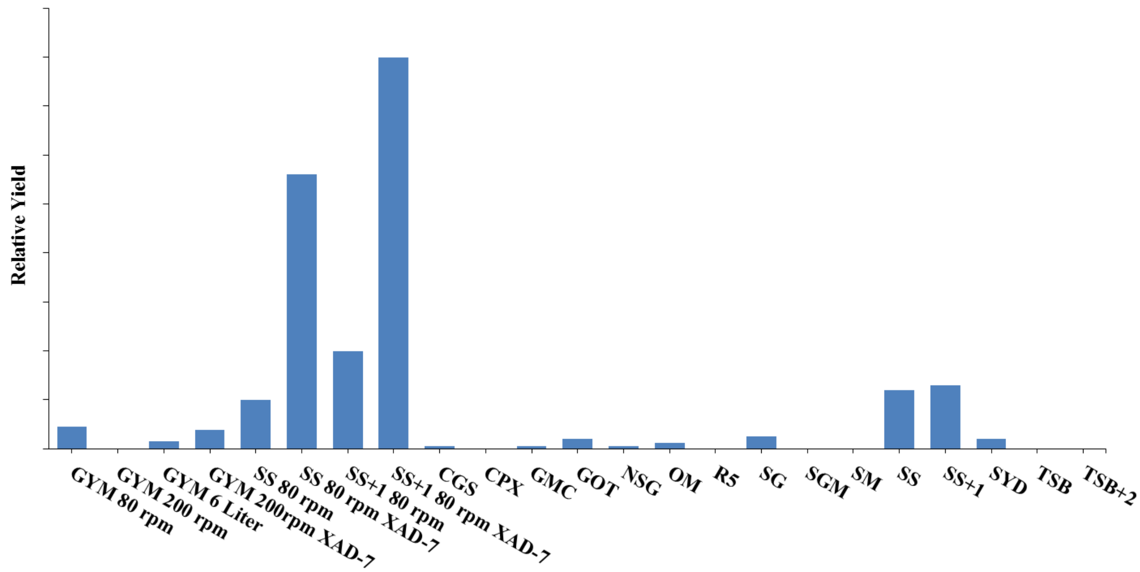

Figure 3-10. Investigation of optimal fermentation condition and production media. These production media were widely used in Actinobacteria fermentation and contain various carbon/nitrogen sources.

Addition of absorptive polymeric resins has been reported to have a profound effect on the production of secondary metabolites and used as a means to consistently enhance the yield of specific low titer compounds[40]. To test this effect on the production of alloactinomicin, Amberlite XAD-7 resin was used for two reasons: one reason is that XAD-7 resin is "moderately polar" and could be used to remove relatively polar compounds from non-aqueous solvents, and to remove non-aromatic compounds from polar solvents; another reason is the expect to increase the yield by binding the targeted compounds, which may eliminate the negative regulation due to the accumulation of final products. Indeed, the addition of 5 mL of autoclaved Amberlite XAD-7 resin into the production media, GYM, SS and SS+1 media, displayed a significant increase of alloactinomicin production (Figure 3-10).

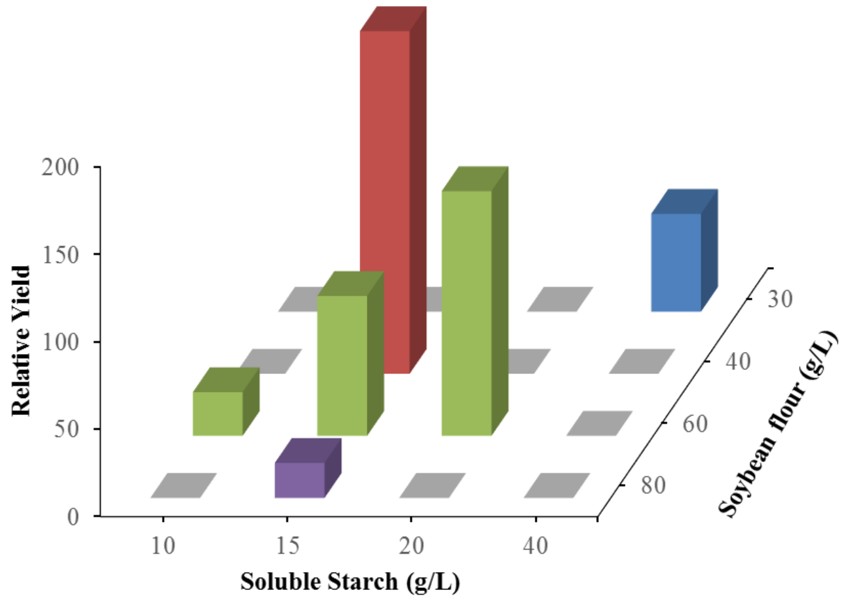

Figure 3-11. Investigation of optimal carbon/nitrogen source ratio. Relative yield of alloactinomicin (peak area at 460 nm) are plotted as a function of soluble starch and soybean flour concentrations. The grey colored ones are data not collected.

According to the above single-factor optimization, the SS+1 medium was determined to be the optimal one. A combination of three soluble starch concentrations and three soybean flour concentrations was carried out to determine the optimal ratio between carbon and nitrogen sources (Figure 3-11). After HPLC analysis and comparison, it turned out that higher amount of soluble starch and lower amount of soybean flour benefited the production of alloactinomicin, because 20 g/L of soluble starch displayed much higher yield than 15 g/L of soluble starch and 40 g/L of soybean flour produce much more alloactinomicin than 60 g/L of soybean flour in the fermentation medium (Figure 3-11).

Inorganic salts and vitamins have been demonstrated to play a crucial role in the secondary metabolite production[16], thus the effect of inorganic salts on the production of alloactinomicin was also simply tested in this work. Two media with the addition of 2 g/L of $(NH_4)_2SO_4$, 2 g/L of NaCl and 0.5 g/L of K_2HPO_4 were tested in parallel with above

carbon/nitrogen ratio investigation. Although these three inorganic salts exhibited positive effect when 15 g/L to 60 g/L carbon/nitrogen ratio was used, they exerted negative effect on alloactinomicin production when 15 g/L to 40 g/L carbon/nitrogen ratio was used (data not shown). Thus, these inorganic salts were not added into the recipe of the final production medium.

Compared with the original GYM medium condition, the final optimal fermentation condition (50 mL of the optimal medium SS+3 fermented at 28℃, 80 rpm for 7 days with addition of 5 mL of Amberlite XAD-7 resin) gained a 44-fold increase on the production of alloactinomicin. Therefore, we next turned to isolation and structure elucidation of alloactinomicin in order to provide further experimental support for above bioinformatic prediction of the core structure chemotype.

Isolation and Structure Elucidation of Alloactinomicin. A large scale (3 liter) fermentation using the above optimal medium and cultivation condition was carried out, affording 3.8 mg of alloactinomicin from the resulting compound absorbed XAD-7 resin by successive chromatographic purification. Alloactinomicin is an orange amorphous solid that is only partially soluble in methanol, but soluble in DMSO.

The molecular formula of alloactinomicin was deduced to be $C_{26}H_{18}O_9$ by HR-ESI-MS (m/z 473.0875 $[M-H]^-$, calculated 473.0873). Proton and Carbon NMR spectral data (Table 3-4 and Figure 3-12) revealed the presence of 16 proton and 22 carbon signals, missing 4 carbon signals. Fifteen carbon signals present in the ^{13}C NMR spectrum have chemical shifts between δ 100 and 170 ppm, which represent carbon atoms on aromatic rings; and two carbonyl resonances were observed at 189.59 and 181.44 ppm, corresponding to the quinone moiety on the skeleton of alloactinomicin.

1-D NMR Data Summary					HMBC
NO.	δC/ppm	C Type	δH/ppm	multiplet (J/Hz)	coupled δH/ppm
1	163.08	C			
2	120.23?				
3	140.41?				
4	120.23	CH	6.66	brs	
4a	140.41?				
5	28.32	CH2	2.75	dd	2.81;
6	19.91	CH2	2.81	dd	2.75;
6a	132.07	C			2.75;2.81;12.51
7	157.85	C			2.81;12.51
7a	113.1	C			12.51;
8	189.59	C			6.59;7.14
8a	109.19	C			6.58;7.14;12.1
9	164.44	C			6.58;12.1
10	107.78	CH	6.59	d (2.37 Hz)	12.1;
11	165.6	C			6.58;7.14
12	108.76	CH	7.14	d (2.37 Hz)	7.14;11.38;
12a	135.41	C			7.14;
13	181.44	C			7.14;
13a	130.23	C			
14	119.57	CH	8.84	s	
14a	140.41	C			2.81;
14b	118.57	C			2.75;
15	40.38?	CH	2.53		
16	48.6?	CH	3.15		
17	28.64	CH3	1.89	brs	
18	171.34	C			
		7-OH	12.51		
		9-OH	12.1		
		11-OH	11.39		
		1-OH	missing		
		18-COOH	missing		

Table 3-4. NMR spectroscopic data (in DMSO-d6) for alloactinomicin. Question mark means that atom is not confidently assigned.

The ^{1}H NMR spectrum has four aromatic proton signals, H-4 (6.66 ppm, broad singlet), H-10 (6.59 ppm, doublet), H-12 (7.14 ppm, doublet) and H-14 (8.84 ppm, singlet). 2D COSY spectrum revealed the connectivity between H-10 and H-12 with coupling constant of 2.37 Hz, suggesting a *meta* relationship. Three hydroxyl protons, OH-7 (12.51

ppm), OH-9 (12.1 ppm) and OH-11 (11.39 ppm), were also identified on the ^{1}H NMR spectrum. Two important proton signals at (2.75 ppm, dd) and (2.81 ppm, dd) were assigned unambiguously by the COSY correlation between H-5 and H-6, and the HMBC correlation between H-5 and C-6 (19.91 ppm) as well as H-6 and C-5 (28.32 ppm), supporting the function of ketoreductase AlloD1 (KRpen19). Single and multiple bond C-H correlations were accomplished by HSQC and HMBC experiments, respectively. On the basis of HRMS, NMR spectroscopic evidence and biosynthetic logic, two tentative structures of alloactinomicin were proposed, which feature a benzo[a]tetracene quinone scaffold (Figure 3-12A) and a benzo[a]tetracene δ-lactone structure (Figure 3-12B), respectively.

A) B)

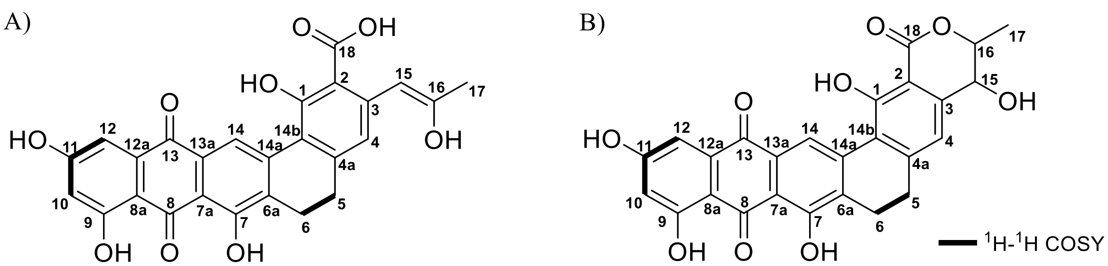

Figure 3-12. The tentative structures of alloactinomicin. 2D NMR correlation ^{1}H-^{1}H COSY are also displayed, while all HMBC correlations are shown in above Table 3-4.

The structure of alloactinomicin together with the fact that *Alloactinosynnema sp.* L-07 harbors only one single PK-II BGC on its sequenced genome strongly support that alloactinomicin is produced by this cluster. Thus, the *Alloactinosynnema sp.* L-07 KSα/β enzymes represent a new 26-carbon poly-β-ketone synthesizing KSα/β as predicted from KSα/β amplicon fingerprint analysis. Furthermore, this result supports that the immediate tailoring enzymes in the cluster collectively function to produce a pentangular core structure. One possible explanation for the discrepancy between the predicted and the

elucidated structure is that the uncharacterized minor congeners contain those chemical modification presented in the predicted structure. Another plausible explanation for this inconsistency is that genes like *allo*D2 (KRpen11) and *allo*E (AmidoT) may be transcriptionally inactive at the fermentation condition used in this work.

Bioactivity of Alloactinomicin. Unlike the vast majority of other type II polyketides discovered to date, which were identified through bioactivity-guided screening, alloactinomicin was discovered through an integrated bioinformatic/experimental approach. Thus, nothing was known about its bioactivity. However, in light of the tentative structures of alloactinomicin, the bioactivity of several compounds with similar structures to alloactinomicin could be found in the literatures. Among the pentangular polyphenols we searched, the tentative structure A of alloactinomicin has the most similar planar structures as WS79089B isolated from *Streptosporangium roseum.* NO. 79089[41]. WS79089B was reported to be a highly selective endothelin coverting enzyme (ECE) inhibitor, but exhibited no antibiotic activity against either gram-negative (e.g. *E. coli*) or gram-positive (e.g. *Staphylococcus aureus* and *Bacillus subtilis*) bacteria, no antifungal activity and no cytotoxic activity against mouse bone marrow cells[41,42]. The tentative structure B of alloactinomicin has the most similar planar structure as hexaricin C discovered from *Streptosporangium sp.* CGMCC 4.7309[1]. Hexaricin C was also reported to possess no significant antibiotic activity against either gram-positive or gram-negative bacterial strains. Therefore, the ECE inhibition activity of alloactinomicin is of top priority to be tested instead of antibiotic or antifungal activities in the future studies.

4. Conclusions

The advent of cost-effective bacterial genome sequencing technology, such as the PacBio RSII system used in this work, revealed thousands of unstudied natural product BGCs from diverse and underexploited branches of the phylogenetic tree of microbes. The traditional experimental approaches were insufficient to identify novel compounds from these large amount of gene clusters. In contrast, bioinformatic and comparative genomic analysis-guided approaches, associated with the existing experimentally-derived knowledge, displayed advantages in achieving the value of this large volume of sequence data. This genome mining approach aided in selecting specific gene clusters with atypical sequence attributes for compound isolation and characterization.

In this work, one new pentangular polyketide, named alloactinomicin, was discovered from *Alloactinosynnema sp.* L-07 using in silico prediction of the PK-II BGC from its sequenced genome, and represents the first reported PK-II in the genus *Alloactinosynnema.* The isolation of alloactinomicin for structural characterization was accomplished by extensive production media selection and fermentation condition optimization, including investigation of various carbon/nitrogen source, adjustment of carbon/nitrogen source ratio, and addition of absorptive Amberlite XAD-7. This approach for searching optimal media and fermentation conditions could also be borrowed to increase the titer of other PK-IIs in the future. The characterization of alloactinomicin proves again the power and advantages of genome mining route in natural product discovery efforts. In addition, the results of such genomics/bioinformatics-guided PK-II discovery provide important links between gene clusters and the molecules they produce,

which offer new insights into gene cluster sequence/function relationship within the class of pentangular polyketides.

Based on the global comparative analysis of all identified pentangular PK-II BGCs and the phylogenetic analysis of homologues of individual biosynthetic enzymes, especially C-11/C-19 ketoreductase (KRpen11 or KRpen19), flavin-dependent monooxygenase (FOX1) and aminotransferase (AmidoT), the structure of alloactinomicin was predicted to be a pentangular polyphenols with core structure featuring a quinone moiety and a lactam ring F. Surprisingly, the lactam ring F moiety was not in the characterized structure of alloactinomicin. The discrepancy between the bioinformatically predicted structure and the characterized one prompt us that bioinformatic analysis of genome sequences alone is inadequate to achieve the full potential of genome mining and experimental methods are also essential to the discovery of new natural products. This inconsistency will be an interesting point for isolation and characterization of all minor congeners that may contain those full-fledged decorations presented in the predicted structure. Given the possible transcriptional deficiency of several genes at the fermentation conditions tested in this work, novel analogs are most probably being discovered by activating these tailoring enzymes in future studies.

5. References

1. Tian, J., Chen, H., Guo, Z., Liu, N., Li, J., Huang, Y., ... & Chen, Y. (2016). Discovery of pentangular polyphenols hexaricins A–C from marine Streptosporangium sp. CGMCC 4.7309 by genome mining. *Applied microbiology and biotechnology, 100*(9), 4189-4199.

2. Medema, M. H., & Fischbach, M. A. (2015). Computational approaches to natural product discovery. *Nature chemical biology*, *11*(9), 639-648.

3. Tiwari, K., & Gupta, R. K. (2012). Rare actinomycetes: a potential storehouse for novel antibiotics. *Critical reviews in biotechnology*, *32*(2), 108-132.

4. Azman, A. S., Othman, I., S Velu, S., Chan, K. G., & Lee, L. H. (2015). Mangrove rare actinobacteria: taxonomy, natural compound, and discovery of bioactivity. *Frontiers in microbiology*, *6*, 856.

5. Matsumoto, A., & Takahashi, Y. (2017). Endophytic actinomycetes: promising source of novel bioactive compounds. *The Journal of Antibiotics*.

6. Lincke, T., Behnken, S., Ishida, K., Roth, M., & Hertweck, C. (2010). Closthioamide: an unprecedented polythioamide antibiotic from the strictly anaerobic bacterium Clostridium cellulolyticum. *Angewandte Chemie*, *122*(11), 2055-2057.

7. Pidot, S., Ishida, K., Cyrulies, M., & Hertweck, C. (2014). Discovery of clostrubin, an exceptional polyphenolic polyketide antibiotic from a strictly anaerobic bacterium. *Angewandte chemie*, *126*(30), 7990-7993.

8. Yuan, L. J., Zhang, Y. Q., Yu, L. Y., Liu, H. Y., Guan, Y., Lee, J. C., ... & Zhang, Y. Q. (2010). Alloactinosynnema album gen. nov., sp. nov., a member of the family Actinosynnemataceae isolated from soil. *International journal of systematic and evolutionary microbiology*, *60*(1), 39-43.

9. Nikou, M. M., Ramezani, M., Amoozegar, M. A., Fazeli, S. A. S., Schumann, P., Spröer, C., ... & Ventosa, A. (2014). Alloactinosynnema iranicum sp. nov., a rare actinomycete isolated from a hypersaline wetland, and emended description of the genus

Alloactinosynnema. *International journal of systematic and evolutionary microbiology*, *64*(4), 1173-1179.

10. Tormo, J. R., Garcia, J. B., DeAntonio, M., Feliz, J., Mira, A., Díez, M. T., ... & Pelaez, F. (2003). A method for the selection of production media for actinomycete strains based on their metabolite HPLC profiles. *Journal of Industrial Microbiology and Biotechnology*, *30*(10), 582-588.
11. Ogasawara, Y., & Liu, H. W. (2009). Biosynthetic studies of aziridine formation in azicemicins. *Journal of the American Chemical Society*, *131*(50), 18066-18068.
12. Liao, R., Duan, L., Lei, C., Pan, H., Ding, Y., Zhang, Q., ... & Liu, W. (2009). Thiopeptide biosynthesis featuring ribosomally synthesized precursor peptides and conserved posttranslational modifications. *Chemistry & biology*, *16*(2), 141-147.
13. Qu, X., Jiang, N., Xu, F., Shao, L., Tang, G., Wilkinson, B., & Liu, W. (2011). Cloning, sequencing and characterization of the biosynthetic gene cluster of sanglifehrin A, a potent cyclophilin inhibitor. *Molecular BioSystems*, *7*(3), 852-861.
14. Jia, X. Y., Tian, Z. H., Shao, L., Qu, X. D., Zhao, Q. F., Tang, J., ... & Liu, W. (2006). Genetic characterization of the chlorothricin gene cluster as a model for spirotetronate antibiotic biosynthesis. *Chemistry & biology*, *13*(6), 575-585.
15. Fang, J., Zhang, Y., Huang, L., Jia, X., Zhang, Q., Zhang, X., ... & Liu, W. (2008). Cloning and characterization of the tetrocarcin A gene cluster from Micromonospora chalcea NRRL 11289 reveals a highly conserved strategy for tetronate biosynthesis in spirotetronate antibiotics. *Journal of bacteriology*, *190*(17), 6014-6025.

16. Shi, J., Pan, J., Liu, L., Yang, D., Lu, S., Zhu, X., ... & Huang, Y. (2016). Titer improvement and pilot-scale production of platensimycin from Streptomyces platensis SB12026. *Journal of industrial microbiology & biotechnology, 43*(7), 1027-1035.

17. Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., ... & Breitling, R. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic acids research, 43*(W1), W237-W243.

18. Grant, J. R., & Stothard, P. (2008). The CGView Server: a comparative genomics tool for circular genomes. *Nucleic acids research*, *36*(suppl 2), W181-W184.

19. Carver, T., Thomson, N., Bleasby, A., Berriman, M., & Parkhill, J. (2009). DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics, 25*(1), 119-120.

20. Lopez, P., Hornung, A., Welzel, K., Unsin, C., Wohlleben, W., Weber, T., & Pelzer, S. (2010). Isolation of the lysolipin gene cluster of Streptomyces tendae Tü 4042. *Gene, 461*(1), 5-14.

21. Zhang, W., Wang, L., Kong, L., Wang, T., Chu, Y., Deng, Z., & You, D. (2012). Unveiling the post-PKS redox tailoring steps in biosynthesis of the type II polyketide antitumor antibiotic xantholipin. *Chemistry & biology*, *19*(3), 422-432.

22. Kudo, F., Yonezawa, T., Komatsubara, A., Mizoue, K., & Eguchi, T. (2011). Cloning of the biosynthetic gene cluster for naphthoxanthene antibiotic FD-594 from Streptomyces sp. TA-0256. *The Journal of antibiotics, 64*(1), 123-132.

23. Kersten, R. D., Ziemert, N., Gonzalez, D. J., Duggan, B. M., Nizet, V., Dorrestein, P. C., & Moore, B. S. (2013). Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proceedings of the National Academy of Sciences, 110*(47), E4407-E4416.

24. Kang, H. S., & Brady, S. F. (2014). Mining soil metagenomes to better understand the evolution of natural product structural diversity: pentangular polyphenols as a case study. *Journal of the American Chemical Society, 136*(52), 18111-18119.

25. Kang, H. S., & Brady, S. F. (2014). Arixanthomycins A–C: phylogeny-guided discovery of biologically active eDNA-derived pentangular polyphenols. *ACS chemical biology, 9*(6), 1267-1272.

26. Banskota, A. H., Aouidate, M., Sørensen, D., Ibrahim, A., Piraee, M., Zazopoulos, E., ... & Falardeau, P. (2009). TLN-05220, TLN-05223, new Echinosporamicin-type antibiotics, and proposed revision of the structure of bravomicins. *The Journal of antibiotics, 62*(10), 565-570.

27. Wendt-Pienkowski, E., Huang, Y., Zhang, J., Li, B., Jiang, H., Kwon, H., ... & Shen, B. (2005). Cloning, Sequencing, Analysis, and Heterologous Expression of the Fredericamycin Biosynthetic Gene Cluster from Streptomyces g riseus. *Journal of the American Chemical Society*, *127*(47), 16442-16452.

28. Zaleta-Rivera, K., Charkoudian, L. K., Ridley, C. P., & Khosla, C. (2010). Cloning, sequencing, heterologous expression, and mechanistic analysis of A-74528 biosynthesis. *Journal of the American Chemical Society, 132*(26), 9122-9128.

29. Martin, R., SIERNER, O., Alvarez, M. A., De Clercq, E., Bailey, J. E., & Minas, W. (2001). Collinone, a new recombinant angular polyketide antibiotic made by an engineered Streptomyces strain. *The Journal of antibiotics*, *54*(3), 239-249.

30. Li, A., & Piel, J. (2002). A gene cluster from a marine Streptomyces encoding the biosynthesis of the aromatic spiroketal polyketide griseorhodin A. *Chemistry & biology*, *9*(9), 1017-1026.

31. Feng, Z., Kallifidas, D., & Brady, S. F. (2011). Functional analysis of environmental DNA-derived type II polyketide synthases reveals structurally diverse secondary metabolites. *Proceedings of the National Academy of Sciences, 108*(31), 12629-12634.

32. Ogasawara, Y., Yackley, B. J., Greenberg, J. A., Rogelj, S., & Melançon III, C. E. (2015). Expanding our understanding of sequence-function relationships of type II polyketide biosynthetic gene clusters: bioinformatics-guided identification of Frankiamicin A from Frankia sp. EAN1pec. *PloS one, 10*(4), e0121505.

33. Xu, Z., Schenk, A., & Hertweck, C. (2007). Molecular Analysis of the Benastatin Biosynthetic Pathway and Genetic Engineering of Altered Fatty Acid− Polyketide Hybrids. *Journal of the American Chemical Society, 129*(18), 6022-6030.

34. Kim, B. C., Lee, J. M., Ahn, J. S., & Kim, B. S. (2007). Cloning, sequencing, and characterization of the pradimicin biosynthetic gene cluster of Actinomadura hibisca P157-2. *Journal of microbiology and biotechnology, 17*(5), 830.

35. Lee, T. S., Khosla, C., & Tang, Y. (2005). Engineered biosynthesis of aklanonic acid analogues. *Journal of the American Chemical Society, 127*(35), 12254-12262.

36. Zhan, J., Watanabe, K., & Tang, Y. (2008). Synergistic actions of a monooxygenase and cyclases in aromatic polyketide biosynthesis. *ChemBioChem, 9*(11), 1710-1715.

37. Lackner, G., Schenk, A., Xu, Z., Reinhardt, K., Yunt, Z. S., Piel, J., & Hertweck, C. (2007). Biosynthesis of pentangular polyphenols: deductions from the benastatin and griseorhodin pathways. *Journal of the American Chemical Society, 129*(30), 9306-9312.

38. Chen, Y., Wendt-Pienkoski, E., Rajski, S. R., & Shen, B. (2009). In vivo investigation of the roles of FdmM and FdmM1 in fredericamycin biosynthesis unveiling a new family of oxygenases. *Journal of Biological Chemistry, 284*(37), 24735-24743.

39. Scherlach, K., & Hertweck, C. (2009). Triggering cryptic natural product biosynthesis in microorganisms. *Organic & biomolecular chemistry, 7*(9), 1753-1760.

40. González-Menéndez, V., Asensio, F., Moreno, C., de Pedro, N., Monteiro, M. C., de la Cruz, M., ... & Tormo, J. R. (2014). Assessing the effects of adsorptive polymeric resin additions on fungal secondary metabolite chemical diversity. *Mycology, 5*(3), 179-191.

41. Tsurumi, Y., Ohhata, N., Iwamoto, T., Shigematsu, N., Sakamoto, K., Nishikawa, M., ... & Okuhara, M. (1994). WS79089A, B and C, new endothelin converting enzyme inhibitors isolated from Streptosporangium roseum. No. 79089. *The Journal of antibiotics, 47*(6), 619-630.

42. Tsurumi, Y., Fujie, K., Nishikawa, M., Kiyoto, S., & Okuhara, M. (1995). Biological and pharmacological properties of highly selective new endothelin converting enzyme inhibitor WS79089B isolated from Streptosporangium roseum No. 79089. *The Journal of antibiotics*, *48*(2), 169-174.

Chapter 4. Connecting PK-II BGCs to The Compounds Using CRISPR/Cas9-based KSα Deletion and Comparative Metabolism

1. Introduction

With the development of cost-effective genome sequencing technology, tens of thousands of bacterial genomes have been sequenced and deposited in NCBI GenBank database, and the information from these genome sequences had a profound impact on the enterprise of natural product discovery[1]. First, genome sequencing efforts revealed that the wealth of natural products of microbes, even the most intensively studied *Streptomyces* strains, was far from exhaustion. Second, the majority of unidentified natural product biosynthetic gene clusters (BGCs) were expressed poorly or completely silent under limited fermentation conditions in the laboratory. To rapidly access these cryptic natural product gene clusters, genomics and bioinformatics combined approaches, termed genome mining, have been developed recently.

Connecting gene clusters to the compounds they produce is a key step in genome mining of uncharacterized natural product BGCs, because these links between coding genes and small-molecule products could provide new insights into the biosynthetic understanding of various classes of natural products and the function/structure relationships in these systems. These newly acquired information in return would guide the targeting of gene clusters encoding novel bioactive compounds. Furthermore, the synthesis of analogs via synthetic biology or bioengineering would benefit greatly from these biosynthetic insights. Construction of knock-out mutants of certain biosynthetic gene and subsequent comparative metabolic profiling are a commonly employed strategy to facilitate both the connection and validation of new compounds from gene clusters of

interest, because the knock-out mutants could be used as negative controls in comparable experiments with wild-type strains in various fermentation conditions. In addition, all type II polyketides (PK-IIs) are colorful due to the large conjugated system in their structures, which facilitates the detection process significantly by directly comparing the colors of mutant and wild-type strain cultures. Numerous genetic manipulation methods have been established for a variety of actinobacterial genera, in particular the genus *Streptomyces* owing to that they are the most prolific and best-studied producers of diverse secondary metabolites.

Most recently, gene disruption/deletion and gene cluster inactivation in *Streptomyces* strains were benefited from the application of CRISPR/Cas system based genome editing (Figure 4-1)[2]. This high-efficient multiplex genome editing was achieved by using the type II Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/CRISPR-associated (Cas9) proteins system of *Streptococcus pyogenes* to introduce a double-strand break at the genomic locus of interest, then repairing this gap via homologous recombination[2]. Similar to a bacterial adaptive immune system, the type II CRISPR/Cas9 system (Figure 4-1) first transcribed the genome-targeting sequences (called spacer) of the CRISPR array to CRISPR RNA (crRNA), which then hybridized with an associated *trans*-activating CRISPR RNA (tracrRNA) to form a crRNA-tracrRNA duplex. The Cas9 protein recruited this RNA duplex to form a complex, and identified the target genomic DNA sites, known as protospacer, by a trinucleotide protospacer adjacent motifs (PAMs, NGG in the case of *S. pyogenes*, where N represents any nucleotide), binding to this position if the protospacer sequence was complementary to the spacer sequence in the

crRNA-tracrRNA duplex, which induced a double-strand cleavage by activating the HNH and RuvC-like domains of Cas9 nuclease[3,4].

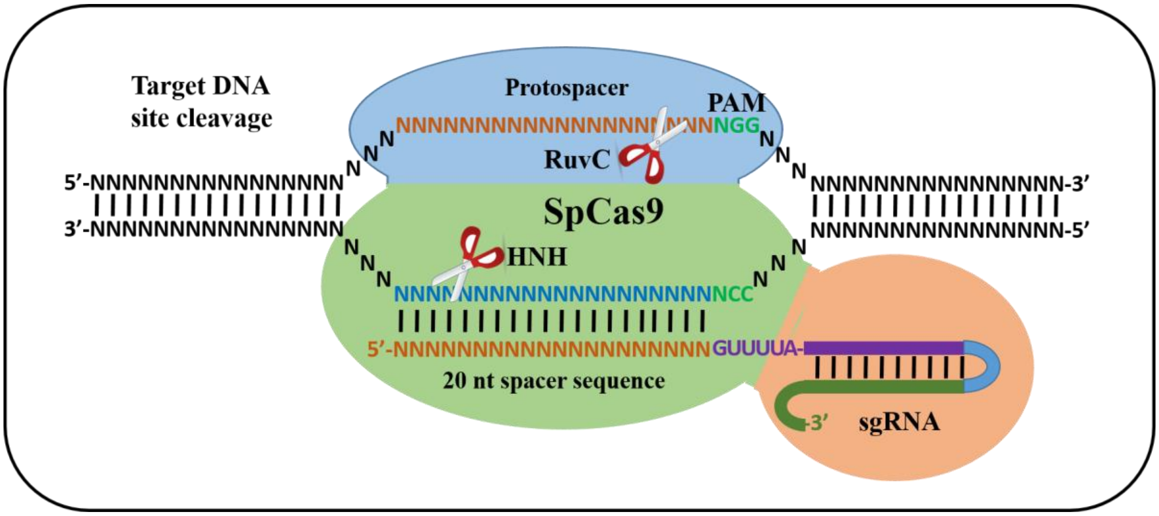

Figure 4-1. Scheme of the CRISPR-Cas9 system based genome editing. The Cas9 HNH and RuvC-like domains cleave the complementary and non-complementary strand of the target sequence, respectively, which is recognized by a single transcript of crRNA-tracrRNA chimera, sgRNA; the PAM is shown in green; the 20 nt spacer sequence is shown in orange; the sgRNA core structure is shown as a sketch.

To further develop this system into a simple, versatile and programmable system for targeted DNA cleavage and genome editing, fusion of the 3' end of crRNA to the 5' end of tracrRNA into a single synthetic guide RNA (sgRNA) transcript has been established, simplifying the process of generation of individual crRNA components[4]. Given its unprecedented modularity, this *S. pyogenes* CRISPR/Cas9 system was successfully transplanted as a genome editing tool into a wide range of organisms spanning all domains of life, including *E. coli*, *Saccharomyces cerevisiae*, and human cell lines. Most recently, this system was extended into *Streptomyces* strains for targeted genome editing associated with homologous recombination by two research groups[2,3]. Compared with classic double-crossover recombination approach, this CRISPR/Cas9 system based genome editing was demonstrated as a much more efficient, precise tool for targeted

chromosomal deletions in *Streptomyces* species. It exhibited unprecedented modularity that targeting any site on the genome only need to insert a target spacer sequence and a homologous recombination template into the CRISPR array/sgRNA construct. Furthermore, this system had its advantage in GC-rich *Streptomyces* genomes that contain remarkably abundant NGG sequences, which facilitated the identification of an optimal target site[2].

We envisioned that it would be appealing if this system could be successfully reconstituted in *Streptomyces* species identified and prioritized by bioinformatic analysis to achieve desired deletion of genes of interest. To this end, bioinformatic analysis of all available PK-II gene clusters using software *Dynamite* developed by our group, together with global phylogenetic analysis of ketosynthase α/β (KSα/β) sequences and correlation analysis between the combination of KSα/β product, C9 ketoreductase, and cyclases and core structure chemotypes revealed a group of unstudied type II polyketide synthase (PKS) gene clusters with evolutionarily divergent KSα/β sequences or atypical aromatase/cyclase (Aro/Cyc) and cyclase (Cyc) sets in bacteria from the genus *Streptomyces*. Of the 28 bioinformatically identified Actinobacteria, 9 strains belong to the genus *Streptomyces*, which are suitable for the CRISPR/Cas9 system based genome editing. To unveil the PK-II compounds encoded by these novel type II PKS gene clusters, we utilized the pCRISPomyces-2 expression system[2], which comprises a sgRNA expression cassette and codon-optimized *cas9* gene from *S. pyogenes*, to create KSα in-frame deletion mutants, and subsequently compared the metabolites of mutant with that of the wild-type strain. To provide additional support for the simplicity of this system, different methods for plasmid

construction were investigated using modern DNA assembly techniques such as overlap-extension PCR[5] and Gibson assembly[6].

Streptomyce venezuelae ATCC 15439 was used as a host to establish this platform in our lab with two concerns: first, this strain is fast growing, which shortens the time of establishment of the system; second, this strain contains an interesting uncharacterized type II PKS gene cluster. Based on preliminary metabolic analysis, three out of 9 *Streptomyces* strains were chosen for KSα in-frame deletion using the CRISPR/Cas9 system to inactivate the corresponding PK-II BGCs. One angucycline-type PK-II, termed flavochromycin, was characterized from *Streptomyces flavochromogenes* NRRL B-2684 (designated as PNP21), demonstrating the feasibility of combining our bioinformatics analysis together with CRISPR/Cas9 system-based PK-II BGC inactivation and subsequent comparative metabolic profiling in discovering novel PK-II compounds. This work provides the first example of a PK-II experimentally characterized by integrating the advantages of bioinformatic analysis and the power of CRISPR/Cas9 system based genome editing.

2. Experimental Materials and Methods

General. Most materials and methods used for molecular cloning work described in this chapter have already been mentioned in the Experimental section of Chapter 2 and Chapter 3. Additionally, apramycin sulfate salt was purchased from Sigma-Aldrich (St. Louis, MO), while chloramphenicol was product of Alfa Aesar (Tewksbury, MA). Gibson master mix was prepared by mixing 6.67 μL of Taq DNA ligase (40 U/μL, purchased from New England Biolabs (NEB), Ipswich, MA), 13.3 μL of 5x isothermal buffer (see below), 0.27 μL of T5 exonuclease (1U/ μL, NEB), 0.83 μL of Phusion Hot Start Flex DNA

polymerase (Thermo Scientific), and nuclease-free H_2O to a total volume of 50 μL. The preparation of 5x isothermal buffer was carried out as follows: 0.75 g of 25% polyethylene glycol 8000 (PEG-8000), 1.5 mL of 500 mM Tris-HCl (pH 7.5), 75 μL of 50 mM $MgCl_2$, 150 μL of 50 mM DTT, 30 μL of 1 mM dATP, 30 μL of 1 mM dTTP, 30 μL of 1 mM dCTP, 30 μL of 1 mM dGTP, 300 μL of 5 mM β-nicotinamide adenine dinucleotide (NAD, purchased from NEB) and nuclease-free H_2O was added to a total volume of 3 mL. *Bbs*I and *Xba*I restriction endonucleases were enzyme products of NEB. CRISPR/Cas off-target analysis on the whole genome sequence was achieved using the functionality of the bioinformatic software Geneious. SPA agar medium was prepared by addition of 1 g of yeast extract (Bacto), 1 g of beef extract, 2 g of tryptone (acumedia), 10 g of D-(+)-glucose, trace amount of $FeSO_4 \cdot 7H_2O$ and 15 g of agar into 1 L deionized water, no pH adjustment. Nalidixic Acid was prepared as follows: 0.3 g of Nalidixic acid was added into 10 mL of 0.3 M NaOH and dissolved completely; the solution was sterilized by filtering through a 0.22 μm syringe filter.

Plasmids and Vectors. The commercial plasmid pCRISPomyces-2 purchased from Addgene (Plasmid #61737) was used for constructing all the KSα knock-out plasmids described below. Plasmid pCRISPomyces-2 (Figure 4-2) comprises two essential components for function of the CRISPR/Cas9 system: a *Spcas9* gene codon optimized for *Streptomyces* expression, encoding the Cas9 nuclease; a synthetic guide RNA (sgRNA-tracr) joined by the crRNA and tracrRNA, which is short RNAs conferring target site specificity and facilitating crRNA recruitment to Cas9 protein. Additional components include the promoter rpsLp(XC) with *Bbs*I recognition sites removed, a wild-type fd terminator, a promoter gapdhp(EL), a oop terminator, a *lacZ* expression cassette flaked by

*Bbs*I recognition sites that are designed for insertion of custom spacer sequence, unique *Xba*I restriction sites for incorporation of homologous recombination template, an *E. coli/Streptomyces* helper fragment containing origin colE1, selection marker aac(3)IV, a temperature-sensitive pSG5 *rep* origin with *Bbs*I site removed, and origin of transfer *ori*T.

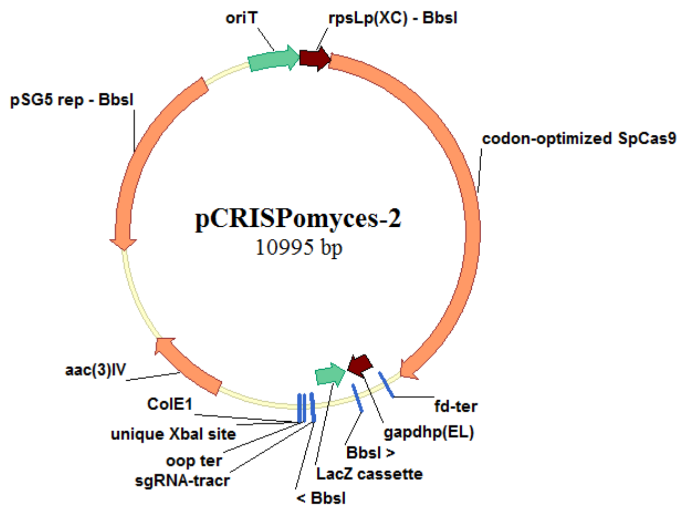

Figure 4-2. The plasmid map of pCRISPomyces-2. This plasmid was designed for targeted genome editing in *Streptomyces* species. Key components include a codon-optimized *cas9* gene from *S. pyogenes*, a single transcript of crRNA-tracrRNA chimera, sgRNA, a *Bbs*I-flanked *lac*Z cassette for insertion of custom spacer sequences, and an *Xba*I site for incorporation of homologous recombination arms.

Bacterial Strains. *E. coli* DH5α used for plasmid construction and maintenance in this Chapter have already been mentioned in the Experimental section of Chapter 2. *E. coli* ET12567/pUZ8002 (ETZ) competent cell was the donor strain used in *E. coli-Streptomyces* conjugal transfer experiments. The 28 bioinformatically identified Actinobacteria were requested from NRRL (Agricultural Research Service Culture Collection, Peoria, IL, USA).

Instrumentation. The pH measurement, PCR reactions, DNA gel imaging, DNA concentration quantification, agarose gel electrophoresis, centrifugation, and bacteria incubation were carried out on identical instruments described in Chapter 2. Rotary

evaporation, HPLC analysis, and compound purification were performed on identical instruments described in Chapter 3. High resolution mass spectrometry (HRMS) data was acquired using the service provided by Mass Facility in the Department of Chemistry at the University of California Riverside.

Bacteria Cultivation. The bacteria cultivation used in this chapter was similar as described in Chapter 2. *E. coli* strains, including *E. coli* DH5α and *E. coli* ETZ, were grown in 2-5 mL of LB broth supplemented with apramycin (final conc. 50 μg/mL) in sterile 15 mL conical tubes at 37 °C, 250 rpm overnight (15-20 h). *E. coli* DH5α transformants was selected on LB agar plates containing the apramycin (final conc. 50 μg/mL) at 37 °C overnight (15-20 h), while *E. coli* ETZ was selected on LB agar plates containing the apramycin (final conc. 50 μg/mL), kanamycin (final conc. 25 μg/mL), and chloramphenicol (final conc. 25 μg/mL) at 37 °C overnight (15-20 h).

Fermentation of PNP21. The wild-type strain of PNP21 (*Streptomyces flavochromogenes* NRRL B-2684) was first grown up in a seed culture by inoculation of 20 μL of 20% glycerol spores stock into 25 mL of GYM liquid medium in a 125 mL Erlenmeyer flask with glass beads, which was incubated at 30 ℃, 250 rpm in a rotary incubator. After 24 h of incubation, the seed culture were centrifuged at 4,000 *g* for 5 min, and the supernatant was discarded while the cell pellets were resuspended into 5 mL of fresh GYM liquid medium. To inoculate large scale fermentation culture, the seed culture was first scaled up by inoculate 4% (v/v) into a 100 mL of GYM liquid culture in 500 mL flask with glass beads and grown in 30 ℃, 250 rpm shaker. After 18 h of growth, the 100 mL seed culture was centrifuged at 4,000 *g* for 5 min, and the supernatant was discarded while the cell pellets were resuspended into 25 mL of fresh GYM liquid medium. After the

re-suspending, 3.5-4 mL of resuspended seed culture was added into a 2.5 L Ultra Yield flask (Thomson, Oceanside, CA) with 500 mL of GYM liquid medium, which was fermented at 27 °C, 200 rpm for 7 days.

Preparation of E. coli Competent Cells. The procedure used to prepare *E. coli* competent cells was the same as described in Chapter 2.

General PCR Conditions. The general PCR conditions used in this chapter was the same as described in the Experimental section of Chapter 2.

Design and Construction of KSα In-frame Deletion Plasmid pCRISPR-ds, pCRISPR-ts, and pCRISPR-sc. All three KSα in-frame deletion plasmids are derivatives of plasmid pCRISPomyces-2 as described above. To construct pCRISPR-ds, pCRISPR-ts, and pCRISPR-sc, the corresponding spacer sequences and homologous recombination arms were sequentially inserted into plasmid pCRISPomyces-2. Because the KSα gene is about 1.2 kb long, so we envisioned that positions localized in the middle of KSα gene would be better target sites than those at the ends. Based on the genome sequence of wild-type *S. venezuelae* ATCC15439, 84 potential spacer sequences on both strands were picked out by eye searching the sequence in the Vector NTI. Then the last 8 to 12 nucleotides (nt) of these tentative spacers were subjected to CRISPR/Cas off-target analysis against its whole genome sequence using the functionality of software Geneious. Additional preferences were given to spacer with purines occupying the last four based at 3’ end, spacer on the non-coding strand, and spacer with G as the last base. Based on these criteria, three top ranking spacers, ds-spacer, ts-spacer and sc-spacer, were chosen to construct plasmids pCRISPR-ds, pCRISPR-ts and pCRISPR-sc, respectively. To incorporate each spacer into pCRISPomyces-2, two complementary oligonucleotides (Table 4-1) containing

the 20 nt spacer sequence and the sticky ends (ACGC on the forward primer and AAAC on the reverse primer) were designed and synthesized by IDT. The forward and reverse primers were mixed together in 1:1 molar ratio with final concentration of 5 µM each in 50 µL solution, which was heated at 95 °C for 5 minutes and slowly cooled down to the 50 °C. After that, the annealed product was diluted by 10-fold and 1 µL of the resulting diluted solution was ligated into 27 ng of *Bbs*I linearized pCRISPomyces-2 vector, which was transferred into *E. coli* DH5α. The intermediate plasmids were replicated in *E. coli* DH5α, extracted, and Sanger sequencing verified.

Primer Name	Sequence (5'-3')	Description
ds-spacer-fwd	**ACGC**CGCGGCCGAACTCATCCGCG	Construction of pCRISPR-ds spacer
ds-spacer-rev	**AAAC**CGCGGATGAGTTCGGCCGCG	
ts-spacer-fwd	**ACGC**GCTGCCCTCGCGGATGAGTT	Construction of pCRISPR-ts spacer
ts-spacer-rev	**AAAC**AACTCATCCGCGAGGGCAGC	
sc-spacer-fwd	**ACGC**ACCGATGTCGTCCTGAGCGT	Construction of pCRISPR-sc spacer
sc-spacer-rev	**AAAC**ACGCTCAGGACGACATCGGT	
KSa-up-fwd	GC*TCTAGA*CCGCGTCCGAGCCCGGACCGGTGCTCGCCG	Amplification of a 1,050 bp upstream arm
KSa-up-rev	**ACTACCGCGG**GGACGCACCCCCGCGCGGCTGCGGAGCG	
KSa-dn-fwd	**GGGTGCGTCC**CCGCGGTAGTGACCGGCATCGGCGTCACCGCG	Amplification of a 1,074 bp downstream arm
Ksa-dn-rev	GC*TCTAGA*CGTACGGGGTCGGCGGATCACGCCGTCGCGC	
cfm-up-fwd	CGACATCTTCATCGCCGGGGTAGGCAGCAGCGTCC	Confirmation of the mutants; if KSa deleted, it give 2,380 bp band
cfm-dn-rev	GTTCCCGCGCAGGACGACGGCGCTGTTGAAGC	
del-seq-rev	GATCCGCGTGGGGTACTGCGTGG	Sequencing verification of the deletion region
del-seq-fwd	GCCTGGACGCCGCTTTCTTCCAG	
pCRISPR-seq1-up	GGTGTGAAACTTCTGTGAATGGC	Sequencing verification of spacer insertion

Table 4-1. Primer list used for constructing pCRISPR-ds, pCRISPR-ts and pCRISPR-sc. The sticky ends are shown in bold and colored green, the overlaps of the two arms are shown in bold and colored orange, and the *Xba*I recognition sites are shown in italic and colored red.

Meanwhile, two approximately 1 kb homologous recombination arms flanking the KSα gene were PCR amplified from the genomic DNA of wild-type *S. venezuelae* ATCC 15439. The upstream arm was obtained by primer pair, KSα-up-fwd and KSα-up-rev, while the downstream arm was obtained by primer pair, KSα-dn-fwd and KSα-dn-rev (Table 4-1). These two DNA fragments, with 20 nt overlaps at the junction of the two arms, were spliced by overlap-extension PCR using primers, KSα-up-fwd and KSα-dn-rev (Table 4-

1). This resulting product, containing *Xba*I recognition sites at both ends, was digested and ligated into the correct spacer-containing intermediate vector, which was digested with *Xba*I restriction enzyme and dephosphorylated with CIP enzyme prior to ligation. Since it is a single restriction enzyme cleavage, either orientation of the homology arms is possible and without effect on subsequent genome editing. The three final plasmids, pCRISPR-ds, pCRISPR-ts and pCRISPR-sc (Figure 4-3), were replicated in *E. coli* DH5α, colony PCR screened using primers, cfm-upF and cfm-dnR (Table 4-1), extracted, and Sanger sequencing verified.

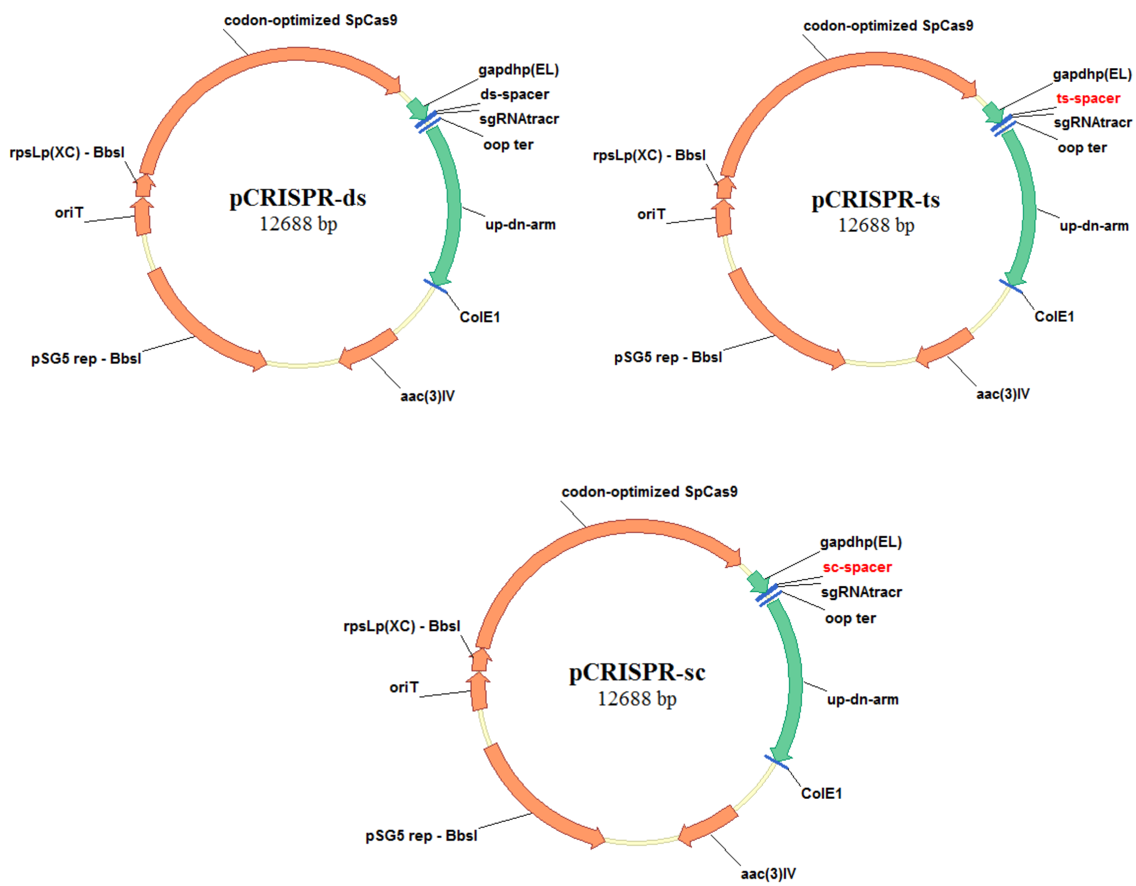

Figure 4-3. Maps of plasmid pCRISPR-ds, pCRISPR-ts and pCRISPR-sc. The components for these three plasmids are identical, except for the spacer sequences, which are highlighted.

Design and Construction of KSα In-frame Deletion Plasmid pCRISPR-PNP19-dual. The plasmid pCRISPR-PNP19-dual contains a dual guide RNA cassettes design with the following configuration: gapdhp(EL)-spacer1-sgRNAtracr-T7term-gapdhp(EL)-spacer2-sgRNAtracr-oopterm. To construct this pCRISPR-PNP19-dual plasmid, two 20 nt spacer sequences (Table 4-2) with NGG at the 3' end were chosen using the same protocol and criteria for designing the ds-spacer, ts-spacer, and sc-spacer as described above. These two spacers were localized at both ends of KSα gene of PNP19 PK-II gene cluster, reasoning that it would facilitate the bridging of the gap via homologous recombination. A synthetic construct containing these two spacers was designed as follows: spacer1 sequence was added at the upstream of sgRNAtracr followed with a T7 terminator, while spacer2 sequence was added at the downstream of gapdhp(EL); *Bbs*I restriction site and protection oligonucleotides were added for the 5' end as 5'-GCTGA**GAAGAC**ATACGC and for the 3' end as GTTTAT**GTCTTC**ACCGG-3', the bold red letters are *Bbs*I recongnition sites; After *Bbs*I digestion, this synthetic construct gave ACGC sticky end at 5' end and CAAA sticky end at 3' end, which were used for ligation of the construct into *Bbs*I digested vector pCRISPomyces-2. The whole synthetic construct is 487 bp long with following configuration: BbsIsite (17 bp)-spacer1 (20 bp)-sgRNAtracr (81 bp)-T7term (47 bp)-gapdhp(EL) (285 bp)-spacer2 (20 bp)-BbsIsite (17 bp). The corresponding sequence is

5'-ctgagaagacatacgcgaacggagcgggatcgaagagttttagagctagaaatagcaagttaaaataaggctagtccgtta
tcaacttgaaaaagtggcaccgagtcggtgctttttttagcataaccccttggggcctctaaacgggtcttgaggggttttttggctgc
tccttcggtcggacgtgcgtctacgggcaccttaccgcagccgtcggctgtgcgacacggacggatcgggcgaactggccgat
gctgggagaagcgcgctgctgtacggcgcgcaccgggtgcggagcccctcggcgagcggtgtgaaacttctgtgaatggcct

gttcggttgcttttttatacggctgccagataaggcttgcagcatctgggcggctaccgctatgatcggggcgttcctgcaattctt

agtgcgagtatctgaaaggggatacgcgcgtaccgaccggccagaaagtttatgtcttcaccgg-3'

This dual-spacer/sgRNA construct was synthesized by IDT and digested with *Bbs*I restriction enzyme followed by ligation into *Bbs*I digested vector pCRISPomyces-2. The resulting intermediate plasmid was replicated in *E. coli* DH5α, colony PCR screened using primers, pCRISPR-seq1-up and pCRISPR-rev (Table 4-2), extracted, and Sanger sequencing verified.

Primer Name	Sequence (5'-3')	Description
PNP19-spacer1	GAACGGAGCGGGATCGAAGA	Construction of the dual-spacer sgRNA
PNP19-spacer2	GCGTACCGACCGGCCAGAAA	
PNP19-upF	GC*TCTAGA*GCGGGATCAGTCCCTCCTCGATGGCGAGCAGG	Amplification of a 1,075 bp upstream arm
PNP19-upR	**AGGCGTCGCA**CCGCCTCGGCAGTGGTGACCGGCCTCGGC	
PNP19-dnF	**GCCGAGGCGG**TGCGACGCCTCCCAGCTGTAGAAGCGGGTGGC	Amplification of a 1,030bp downstream arm
PNP19-dnR	GC*TCTAGA*GCCGCTGACATCTCACGGCTGTGGACCACGCC	
PNP19-cfmF	GCGCTCCTTTCGAAACACGCCCTAGAGCGCACGG	Confirmation of the mutants, if KSα deleted, it give 2,391 bp band
PNP19-cfmR	GATCCGCTGTGACCGCGACACCGGTCACACCTG	
PNP19-seqF	CTCGTCATCCGGTCGGTCTGCG	Sequencing the junction region of arms to verify the plasmid and deletion mutants
PNP19-seq-rev	CACCGCACCCAAGTCGATGACC	
pCRISPR-seq1-up	GGTGTGAAACTTCTGTGAATGGC	Confirmation of dual-spacer construct insertion
pCRISPR-rev	GCCACCTCTGACTTGAGCGTCG	

Table 4-2. Primer list used for constructing pCRISPR-PNP19-dual. The overlaps of the two arms are shown in bold and colored orange, and the *Xba*I recognition sites are shown in italic and colored red.

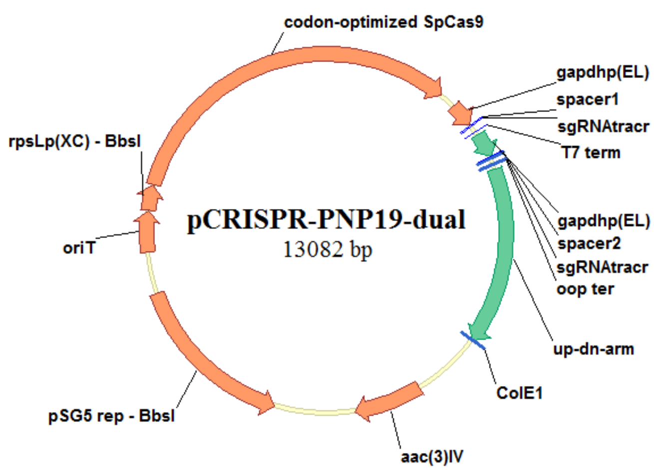

Figure 4-4. Maps of plasmid pCRISPR-PNP19-dual. This plasmid has a dual-spacer sgRNA design that would induce defined cleavage at both ends of the KSα gene.

Meanwhile, two approximately 1 kb homologous recombination arms flanking the KSα gene were PCR amplified from the genomic DNA of wild-type strain PNP19, *Streptomyces anulatus* NRRL B-3362. The upstream arm was obtained by primer pair, PNP19-upF and PNP19-upR, while the downstream arm was obtained by primer pair, PNP19-dnF and PNP19-dnR (Table 4-2). These two DNA fragments, with 20 nt overlaps at the junction of the two arms, were spliced by overlap-extension PCR using primers, PNP19-upF and PNP19-dnR (Table 4-2). This resulting product, containing *Xba*I recognition sites at both ends, was digested and ligated into the correct dual-spacer-containing intermediate vector, which was digested with *Xba*I restriction enzyme and dephosphorylated with CIP enzyme prior to ligation. Since it is a single restriction enzyme cleavage, either orientation of the homology arms is possible and without effect on subsequent genome editing. The final plasmid, pCRISPR-PNP19-dual (Figure 4-4), was replicated in *E. coli* DH5α, colony PCR screened using primers, PNP19-seqF and pCRISPR-rev (Table 4-2), extracted, and Sanger sequencing verified.

Design and Construction of KSα In-frame Deletion Plasmid pCRISPR-PNP20-dual. In the plasmid pCRISP-PNP20-dual, it also harbors two guide RNA cassettes as pCRISPR-PNP19-dual. To construct this pCRISPR-PNP20-dual, two 20 nt spacer sequences with NGG at the 3' end were chosen using the same protocol and criteria for designing the pCRISP-PNP19-dual spacer1 and spacer2 as described above (Table 4-3). These two spacers were localized at both ends of KSα gene of PNP20 PK-II gene cluster. A synthetic construct containing two spacers was designed as pCRISPR-PNP19-dual above. The whole synthetic construct is 487 bp long with following configuration: BbsIsite (17 bp)-spacer1 (20 bp)-sgRNAtracr (81 bp)-T7term (47 bp)-gapdhp(EL) (285 bp)-spacer2

(20 bp)-BbsIsite (17 bp). This construct was PCR amplified from above synthetic PNP19 dual-spacer/sgRNA construct using a primer pair, PNP20-dual-spacerF and PNP20-dual-spacerR (Table 4-3). The purified PCR product was digested with *Bbs*I restriction enzyme and ligated into *Bbs*I digested vector pCRISPomyces-2. The resulting intermediate plasmid was replicated in *E. coli* DH5α, extracted, and Sanger sequencing verified.

Primer Name	Sequence (5'-3')	Description
PNP20-dual-spacerF	GCTGA*GAAGAC*ATACGCgaacgacgcgtcgtcataacGTTTTAGAGCTAGAAATAGCAAG	Construction of the dual-spacer/sgRNA by PCR amplification
PNP20-dual-spacerR	CCGGT*GAAGAC*ATAAACgtcccgaaaggagtgccgcgGCGTATCCCCTTTCAGATAC	
PNP20-upF	GC*TCTAGA*GTATGGGCGTCGGCCTGTACCGCGCCCATCCG	Amplification of a 1,052 bp upstream arm
PNP20-upR	**GCGGTGGCCG**GGGCATCCTCACTCGCTGTCGGCGGGAACC	
PNP20-dnF	**GAGGATGCCC**CGGCCACCGCCACCCGACCCGTGATCACC	Amplification of a 1,067bp downstream arm
PNP20-dnR	GC*TCTAGA*CGCAGGACCAGCAGCGCGGAGGCCAGGTCGAG	
PNP20-cfmF	CCAGCATGGCGGGACAGCGAGGAGAGGAGACC	Confirmation of the mutants, if KSα deleted, it give 2,269 bp
PNP20-cfmR	CTCGCCCCGTCACCAGGTCCACGGGGCAGTCC	
PNP20-seqF	GGACAGGGGCTCACCGTGCTCG	Sequencing the junction region of arms to verify the plasmid and
PNP20-seq-rev	CCACAGTCCGAGGACCTCGTCC	
pCRISPR-seq1-up	GGTGTGAAACTTCTGTGAATGGC	Confirmation of dual-spacer construct insertion
pCRISPR-rev	GCCACCTCTGACTTGAGCGTCG	

Table 4-3. Primer list used for constructing pCRISPR-PNP20-dual. The spacer1 and spacer2 sequences are shown in little letter and colored green, the overlaps of the two arms are shown in bold and colored orange, and the *Bbs*I and *Xba*I recognition sites are shown in italic and colored red.

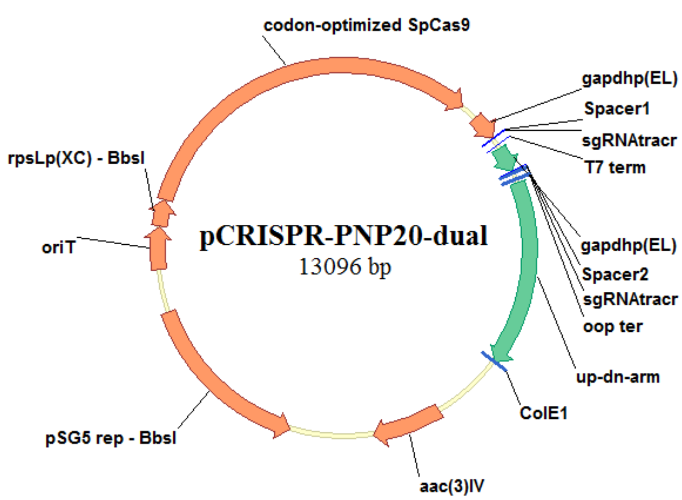

Figure 4-5. Maps of plasmid pCRISPR-PNP20-dual. This plasmid has a dual-spacer/sgRNA design that would induce defined cleavage at both ends of the KSα gene.

Meanwhile, two approximately 1 kb homology arms flanking the KSα gene were PCR amplified from the genomic DNA of wild-type PNP20, *Streptomyces bottropensis* ATCC 25435. The upstream arm was obtained by PCR using primer pair PNP20-upF and PNP20-upR, while the downstream arm was obtained by PCR using primer pair PNP20-dnF and PNP20-dnR (Table 4-3). These two DNA fragments, with 20 nt overlaps at the junction of the two arms, were spliced by overlap-extension PCR using primers PNP20-upF and PNP20-dnR (Table 4-3). This resulting PCR product, containing *Xba*I recognition sites at both ends, was digested and ligated into the correct dual-spacer-containing intermediate vector, which was digested with *Xba*I restriction enzyme and dephosphorylated with CIP enzyme prior to ligation. Since it is a single restriction enzyme cleavage, either orientation of the homology arms is possible and without effect on subsequent genome editing. The final plasmids, pCRISPR-PNP20-dual (Figure 4-5), were replicated in *E. coli* DH5α, colony PCR screened using primers, PNP20-seqF and pCRISPR-rev (Table 4-3), extracted, and Sanger sequencing verified.

Design and Construction of KSα In-frame Deletion Plasmid pCRISPR-PNP21. To construct plasmid pCRISPR-PNP21 (Figure 4-6) for KSα in-frame deletion on the genome of PNP21, *Streptomyces flavochromogenes* NRRL B-2684, the corresponding spacer sequences and homologous recombination arms were sequentially inserted into pCRISPomyces-2. The final chosen spacer sequence was localized in the middle of KSα gene and generated using the same protocol and criteria as described in designing pCRISPR-ds spacer. To insert the spacer into pCRISPomyces-2, two complementary oligonucleotides, PNP21-spacerF and PNP21-spacerR (Table 4-4), containing the 20 nt spacer sequence and the sticky ends (ACGC on the forward primer and AAAC on the

reverse primer) were designed and synthesized by IDT. The forward and reverse primers were annealed together and ligated into the *Bbs*I linearized pCRISPomyces-2 vector as described in constructing pCRISPR-ds plasmid. The intermediate plasmid was replicated in *E. coli* DH5α, extracted, and Sanger sequencing verified.

Primer Name	Sequence (5'-3')	Description
PNP21-spacerF	**ACGC**GTGTCCACGGGATGCACCGC	Construction of pCRISPR-PNP21 spacer
PNP21-spacerR	**AAAC**GCGGTGCATCCCGTGGACAC	
PNP21-upF	**CCGGGCGTTTTTTA***TCTAGA*GCCTGTCATCGTCTTCGGCGCGGTCACCGG	Amplification of a 1,027 bp upstream arm
PNP21-upR	**TTTTCTCGCC**GGCAGTCCCGGAGAACGGCGACGGCACGAC	
PNP21-dnF	**CGGGACTGCC**GGCGAGAAAACGGAATCGTCGATGCCGGGGTC	Amplification of a 1,087 bp downstream arm
PNP21-dnR	**TACGGTTCCTGGCC***TCTAGA*GATGAGCAAGGACCCGCGCAACGCACCG	
PNP21-cfmF	GACCGGACGGGGTCGGTCACATTGATCATCGG	Confirmation of the mutants; if KSa deleted, it give 2,349 bp band
PNP21-cfmR	GCTGATCCCTATCCCGAGTACGCGTGGTCACG	
PNP21-seqF	GTACGTCGTAGTCGGCCACCTCG	Verification of KSa deletion
pCRISPR-seq1-up	GGTGTGAAACTTCTGTGAATGGC	Verification of spacer insertion
pCRISPR-rev	GCCACCTCTGACTTGAGCGTCG	

Table 4-4. Primer list used for constructing pCRISPR-PNP21-dual. The sticky ends are shown in bold and colored green, the overlaps of the two arms are shown in bold and colored orange, and the *Xba*I recognition sites are shown in italic and colored red.

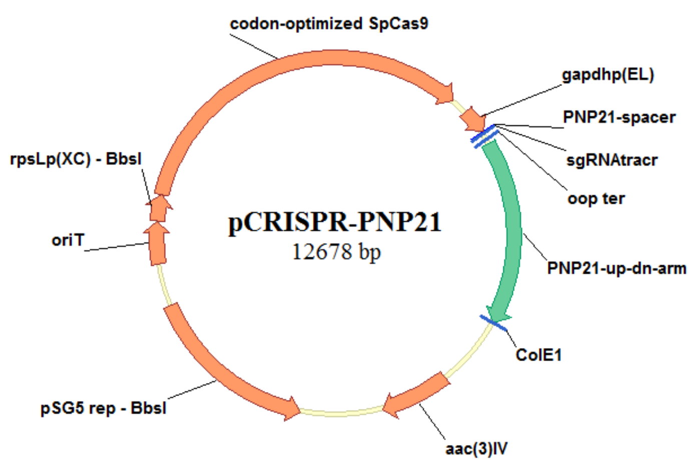

Figure 4-6. Maps of plasmid pCRISPR-PNP21. This plasmid has a single-spacer/sgRNA design that would induce defined cleavage at the middle of the KSα gene.

Meanwhile, two approximately 1 kb homology arms flanking the KSα gene were PCR amplified from the genomic DNA of wild-type PNP21 strain. The upstream arm was PCR amplified by primer pair, PNP21-upF and PNP21-upR, while the downstream arm

was PCR amplified by primer pair, PNP21-dnF and PNP21-dnR (Table 4-4). These two DNA fragments, with 20 nt overlaps at the junction of the two arms and 20 nt overlap sequence at the junction between vector and each arm, were inserted into the spacer-containing intermediate vector by Gibson Assembly. Gibson assembly was carried out as follows: 15 μL of Gibson Master Mix was prepared as described in General section; 5 μL of fragment mixture of each arm, *Xba*I digested and CIP dephosphorylated intermediate vector (molar ratio 10:10:1); this reaction mix was incubated for 1 h at 50 ℃ in a PCR instrument. This resulting product was purified by DNA Clean & Concentrator Kit and 5 μL of it was transferred into competent cell *E. coli* DH5α. The transformants on agar plate were colony PCR screened using primer pair, pCRISPR-seq1-up and pCRISPR-rev and primer pair, PNP21-upF and PNP21-dnR, and correct plasmids were extracted, and Sanger sequencing verified.

Conjugal Transfer of KSα In-frame Deletion Plasmids into Streptomyces Strains. Conjugation of plasmids into *Streptomyces* spores/mycelia was carried out using the modified method described elsewhere[7,8]. Prior to conjugation, the competent cells *E. coli* ET12567/pUZ8002 (abbreviated ETZ) were prepared and the KSα in-frame deletion plasmids constructed above were transferred into *E. coli* ETZ as described in Chapter 2. One single colony of correct transformants was inoculated into 3 mL of LB with Apr (final conc. 50 μg/mL, maintain selection for deletion plasmid), Kan (final conc. 25 μg/mL, maintain selection for pUZ8002), and Cm (final conc. 25 μg/mL, maintain selection for the dam mutation), and grown at 37 ℃, 250 rpm overnight. The overnight culture was inoculated with 1:100 ratio (e.g. 25 μL overnight culture into 25 mL LB) into fresh LB containing the corresponding antibiotics as above. The culture was grown at 37 ℃, 250

rpm for about 3.5 h to reach OD_{600} of 0.4 to 0.6. After grown up, the *E. coli* culture was centrifuged at 4,000 *g* for 5 min, and the supernatant was discarded, while the cell pellets were washed with 1 mL of 2xYT twice to remove the antibiotics and then suspended in 1/10 volume of original culture of 2xYT (e.g. 2.5 mL if original culture is 25 mL). 2xYT liquid medium[8] was prepared by addition of 16 g of Difco Bacto Tryptone, 10 g of Difco Bacto Yeast extract and 5 g of NaCl into 1 L of deionized water, pH adjustment to 7.0. Meanwhile, *Streptomyces* spores or mycelia were scraped off the GYM agar plate and centrifuged at 8,000 *g* for 1 min. The spores or mycelia were washed by 1 mL of 2xYT once and resuspended into 1 mL of 2xYT. The spores need to be treated at 50 ℃ for 10 min and cooled down to room temperature. Then 500 μL aliquot of *E. coli* ETZ and 500 μL aliquot of *Streptomyces* spores or mycelia were mixed and spread on mannitol soya flour (MS) agar plate containing 10 mM $MgCl_2$. MS agar plate[8] was prepared as follows: 20 g of mannitol, 20 g of soybean four and 20 g of agar were added into 1 L of deionized water, autoclaving for 30 min. Following 18-20 h of growth at 30 ℃, the surface of each MS plate was spread with 500 μL of sterile H_2O containing 25 μL of 20 mg/mL Nalidixic acid and 8 μL of 50 mg/mL Apr. The MS plates were incubated at 30 ℃ for 5 to 10 days until the transformants were shown.

Screening of KSα In-frame Deletion Mutants. Once transformants were shown up on the MS agar plate, single colonies were picked up to inoculate into 5 mL of GYM liquid medium supplemented with 50 μg/mL of Apr and grown at 30 ℃, 250 rpm for 3-5 days. Following the cells grow up, 10 μL of cell culture was streaked on SPA agar plates supplemented with 50 μg/mL of Apr. Four single colonies were picked up from each plate and inoculated into 5 mL of GYM liquid medium supplemented with 50 μg/mL of Apr,

and grown at 30 °C, 250 rpm for 3-5 days. 20% glycerol mycelia were prepared prior to PCR screening. The cell pellets were collected by centrifugation and genomic DNA was isolated by Microbial DNA Isolation Kit and used as template for PCR screening. The PCR screening primers binding outside of the homologous recombination arms were employed, and the correct mutants gave a band about 1.2 kb smaller than the wild-type strain. The corresponding primers used for each strain were listed below (Table 4-5).

Primer Name	Sequence (5'-3')	Description
cfm-up-fwd	CGACATCTTCATCGCCGGGGTAGGCAGCAGCGTCC	Confirmation of the *S. ven.* mutants; if KSa deleted, it give 2,380 bp
cfm-dn-rev	GTTCCCGCGCAGGACGACGGCGCTGTTGAAGC	
PNP19-cfmF	GCGCTCCTTTCGAAACACGCCCTAGAGCGCACGG	Confirmation of the PNP19 mutants, if KSα deleted, it give 2,391 bp
PNP19-cfmR	GATCCGCTGTGACCGCGACACCGGTCACACCTG	
PNP20-cfmF	CCAGCATGGCGGGACAGCGAGGAGAGGAGACC	Confirmation of the PNP20 mutants, if KSα deleted, it give 2,269 bp
PNP20-cfmR	CTCGCCCGTCACCAGGTCCACGGGGCAGTCC	
PNP21-cfmF	GACCGGACGGGGTCGGTCACATTGATCATCGG	Confirmation of the PNP21 mutants; if KSa deleted, it give 2,349 bp
PNP21-cfmR	GCTGATCCCTATCCCGAGTACGCGTGGTCACG	

Table 4-5. Primer list used for screening of KSα deletion mutants. These primers were not only used for PCR screening of the transformants on MS plates, but also for verification of the KSα deletion in mutants.

The KSα in-frame deletion mutants were confirmed by Sanger sequencing analysis of PCR amplicons. In order to clear out the temperature-sensitive plasmids from the correct mutants, the 20% glycerol mycelia of KSα deletion strains were inoculated in 5 mL of GYM without Apr, grown at 37 °C and passaged one generation. Fifty μL of the culture were plated on SPA plate, grown at 37 °C for 3-5 days until single colonies were displayed on plate. About 30 single colonies were picked up and pointed onto SPA agar plate supplemented with 50 μg/mL of Apr, then onto SPA plate without Apr and grown at 30 °C for 3-5 days. Several colonies (vary from 20% to 70%) restored the Apr sensitivity, which indicated the successful clearance of KSα in-frame deletion plasmids. These mutants without knock-out plasmids were confirmed on SPA plate supplemented with 50 μg/mL of Apr again and their genomic DNA were isolated, and the KSα regions were PCR amplified

as described above and verified by Sanger sequencing analysis. The sequencing verified mutants were subjected to the subsequent comparative metabolic profiling analysis with wild-type strains.

Comparative Metabolic Profiling of Wild-type Strains and Mutants. To compare the difference of metabolites produced by wild-type strain and mutant, a 25 mL of GYM seed culture of both wild-type strain and mutant were first prepared by growing them at 30 ℃, 250 rpm. After 1-2 days of growth, the seed culture was centrifuged at 4,000 *g* for 5 min, and the supernatant was discarded while the cell pellets were resuspended into 20 mL of fresh GYM liquid medium. In parallel, a panel of pilot production media (50 mL each, Table 4-6) were inoculated with above resuspended seed culture and fermented at 30 ℃, 250 rpm for 7 days.

Name	**Ingredients**	**Reference**
CGS	Cane molasses (2%) Glucose (0.5%) Soluble starch (3%) Pharmamedia (2%) pH 7.0	10
GMC	Glucose (1%) Millet meal (2%) Cottonseed meal (2%) MOPS (2%) pH 7.0	10
GOT	Glycerol (6%) Oat meal (1.5%) Tomato paste (0.5%) $CaCO_3$ (0.3%) pH 7.0	10
GYM	Glucose (0.4%) Yeast extract (0.4%) Malt extract (1%) pH 7.2	
R5	Sucrose (103 g) casamino acids (0.1 g) glucose (10 g) yeast extract (5 g) TES (5.73 g) trace elements (2 mL) K_2SO_4 (0.25 g) $MgCl_2 \cdot 6H_2O$ (10.12 g) pH 7.2 before use, adding 20% L-proline/glutamate (15 mL) 2% $NaNO_3$ (15 mL) 0.5% KH_2PO_4 (10 mL) 1M $CaCl_2$ (20 mL) 1N NaOH (7 mL)	this study
SS+3	Soy bean flour (4%) Soluble starch (1.5%) $CaCO_3$ (0.25%) pH 7.1	this study
TSB	Dextrose (0.25%) Casein peptone (1.7%) Soy peptone (0.3%) NaCl (0.5%) K_2HPO_4 (0.25%) pH 7.3	this study
Notes		
	Pharmamedia is an economical, finely ground, yellow flour made from the embryo of cottonseed, thus cottonseed powder was used	
	MOPS stands for morpholinepropanesulfonic acid, MES for 2-(N-morpholino)ethanesulfonic acid	
	Oat meal was grinded and sifted to obtain the powder	
	Trace Elements Solution: $ZnCl_2$ (40 mg/L) $CuCl_2 \cdot 2H_2O$ (10 mg/L) $MnCl_2 \cdot 4H_2O$ (10 mg/L) $FeCl_3 \cdot 6H_2O$ (200 mg/L) $Na_2B_4O_7 \cdot 10H_2O$ (10 mg/L) $(NH_4)_2Mo_7O_{24} \cdot 4H_2O$ (10 mg/L)	

Table 4-6. Pilot production media used for comparative metabolic profiling.

The metabolites were extracted from 1 mL of fermentation broth as described in Chapter 3. Each crude extracts were dissolved in 40 μL of 50% aqueous acetonitrile. Ten μL of sample was subjected to HPLC analysis using analytical C18 column. The mobile

phase consisted of solvent A (ultrapure H_2O containing 0.1% formic acid) and solvent B (HPLC grade acetonitrile containing 0.1% formic acid) was utilized at a flow rate of 0.6 mL min^{-1}. A multiple step gradient program (25% B for 1 min; 25% to 80% B over 20 min; 80% to 98% B over 2 min; 98% B for 2 min; back to 25% B in 1 min) was employed to compare the metabolic profiles of wild-type strains and mutants at 254 nm.

Extraction of Metabolites. The method for extraction of metabolites from the fermentation culture of wild-type strains and mutants was the same as described in Chapter 3.

Purification of Flavochromycin. The crude extract (~570 mg) obtained from 7 liter large scale fermentation culture was first fractionated by using different solvent, that is 100% $CHCl_3$ (10 mL), 5% MeOH/$CHCl_3$ (10 mL), 10% MeOH/$CHCl_3$ (10 mL), 100% MeOH (3 mL). Most of the flavochromycin were dissolved in 100% $CHCl_3$, thus the portion (~206 mg) was further purified by two successive rounds of HPLC on semi-preparative C18 column. A multiple step program (flow rate: 3.5 mL/min; 25% B for 1 min; 25% to 80% B over 20 min; 80% to 98% B over 2 min; 98% B for 2 min; back to 25% B in 1 min) was used for both rounds of semi-preparative HPLC. Solvent was removed by rotary evaporation and was dried under high vacuum overnight, yielding 2.9 mg of a yellow solid. The purity of isolated compound was checked on analytical HPLC monitored at 254 nm. The purified flavochromycin was subjected to subsequent HRMS and NMR analysis.

Structure Elucidation of Flavochromycin. Mass spectrometric data for alloactinomicin was taken by high-resolution mass spectrometry (HRMS).

Bioinformatic Analysis. The bioinformatic analysis was assisted by the *Dynamite*[9] as described in Chapter 2 and Chapter 3. Each of the chosen gene clusters was further annotated by a software NPgenefinder developed by Yasushi Ogasawara of our group. The software works by applying a whole array of protein queries manually curated from results of *Dynamite* analysis to identify a number of conserved proteins in each gene cluster.

3. Results and Discussion

Bioinformatic Selection of Novel PK-II BGCs. The *Dynamite* globally identified and annotated all PK-II gene clusters whose sequences are available in the NCBI databank. Among ~600 identified putative PK-II BGCs, about 530 clusters were currently unstudied type II PKS gene clusters and about 10% were estimated to encode novel structural PK-II compounds. After the identification of these PK-II BGCs, further analysis were carried out to prioritize them for compound isolation and structure elucidation based on the sequence characteristics. First, KSα/β sequences were correlated with their poly-β-ketone products. Second, different types of aromatase/cyclase (AroCyc) and cyclase (Cyc) enzymes belonging to 7 unrelated protein families were recognized, and the presence of specific sets of AroCyc/Cyc types within a PK-II BGC were correlated with the cyclization/folding pattern of a particular core structure (Figure 4-7). Through *Dynamite* analysis, a number of PK-II BGCs present in sequenced genomes available in the NCBI databank were identified with various combinations of the 5 attributes that were predictive of core structure novelty. These attributes include: KSα/β sequences with either fingerprint similarity scores below the 90% threshold for reliable product prediction and/or that lie in divergent KSα/β phylogenetic clades lacking a training set member; presence of non-acetate priming genes

(KS3a or KS3b or malonyl-CoA acyltransferase (AT) or adenylation domain (ADE)); atypical AroCyc/Cyc sets that render their core structures unpredictable by *Dynamite*; unstudied combinations of AroCyc/Cyc set and predicted KSα/β product; unstudied combinations of AroCyc/Cyc set and non-acetate priming genes. Based on these 5 attributes, a panel of 28 organisms harboring PK-II BGCs with high priority (Table 4-7) were selected for metabolic profiling to identify conditions under which they produce the compounds of interest and/or are transcriptionally active. Strain availability and ease of cultivation were also taken into account when selecting the organisms.

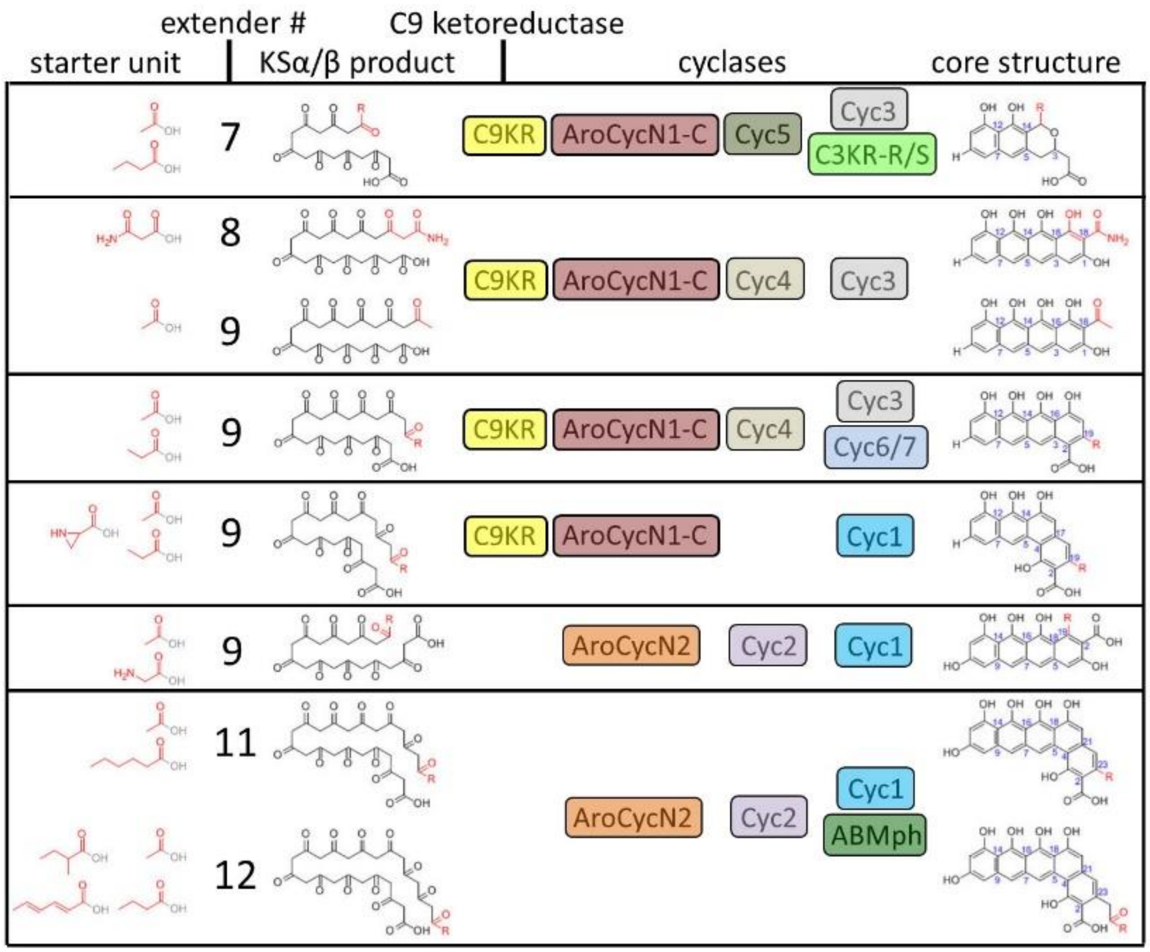

Figure 4-7. Correlation between the combination of KSα/β product, C9 ketoreductase, and cyclases and core structure chemotypes. This figure is acquired from Dr. Melançon's personal manuscript. C9KR-C9 ketoreductase; C3KR-R/S-C3 ketoreductase, R/S stereochemistry ; Cyc1, Cyc2, Cyc3, Cyc4, Cyc5, Cyc6, Cyc7-cyclase type 1, 2, 3, 4, 5, 6, 7; AroCycN1-C-didomain aromatase cyclase N/C terminal domain; AroCycN2-monodomain aromatase cyclase;

Designated Name	Strain's Name	Rational for BGC selection
PNP01	Actinoalloteichus cyanogriseus DSM 43889	PEN; Ac-12 (86.2%);
PNP02	Actinoplanes missouriensis 431	BIQ; 2 KSα/β; N.A.-KS3a; 2 GT;
		BIQ; Ac-7 (93.4%);
PNP03	Actinospica robiniae DSM 44927	spore pigment; peptX;
PNP04	Amycolatopsis azurea DSM 43854	BIQ; Ac-7 (96%);
PNP05	Amycolatopsis decaplanina DSM 44594	TCM; Ac-9 (94.4%); FlvHal
PNP06	Amycolatopsis thermoflava N1165	HED; N.A.-KS3a; KRbiqC3S
PNP07	Catenulispora acidiphila NRRL B-24433	spore pigment
		BIQ; Ac-7 (80.2%); Cyc7
		ANG; Cyc6; FlvHal
PNP08	Cellulomonas flavigena DSM 20109	PEN; no ABMph, ABMi; ABMsp
PNP09	Dactylosporangium aurantiacum NRRL B-8018	dactylocycline
		ANG; Ac-9 (91.3)
PNP10	Kibdelosporangium aridum subsp. largum strain NRRL B-24462	urdamycin
PNP11	Kitasatospora griseola MF730-N6	HED; N.A.-KS3a;
		HED; N.A.-KS3a, KS3b; 2 PK-I modules; KRbiqC3S
PNP12	Kitasatospora phosalacinea NRRL B-16230	spore pigment; PEase;
		R1128; Cyc8; no KR9;
PNP13	Kutzneria albida DSM 43870	PEN; Ac-11 (77%); 2 Cyc1
		TET; 2 Cyc3; 3 GT
PNP14	Lechevalieria aerocolonigenes NRRL B-3298	ANT; Ac-9 (72.4%); N.A.-KS3a;
PNP15	Lentzea albidocapillata NRRL B-24057	TET; Ac-9 (72.4%); N.A.-2 KS3a;
PNP16	Nocardia transvalensis NBRC 15921	PEN; Hex-11 (70.7); N.A.-KS3b; AroCycN1-C;
PNP17	Streptacidiphilus jeojiense NRRL B-24555	BIQ; Ac-9 (62%); N.A.-KS3a;
PNP18	Streptacidiphilus oryzae TH49	BIQ; 2 FlvHal;
PNP19	Streptomyces anulatus NRRL B-3362	ANG; FlvHal;
PNP20	Streptomyces bottropensis ATCC 25435	spore pigment; PEase;
		HED; N.A.-KS3a;
PNP21	Streptomyces flavochromogenes NRRL B-2684	spore pigment; peptX; PEase;
		ANG; Pr-9 (57.8%); N.A.-KS3a; no KR9;
PNP22	Streptomyces fulvissimus DSM 40593	ANG; Cyc7; CycX; FlvHal;
PNP23	Streptomyces monomycini NRRL B-24309	PEN; FOX2b; FOX3;
PNP24	Streptomyces purpeofuscus NRRL ISP-5283	PEN; Ac-11 (74.6%); 2 FlvHal;
PNP25	Streptomyces sp. NRRL F-2664	spore pigment;
		BIQ; no Cyc3; 2 GT;
PNP26	Streptomyces sp. NRRL F-3213	TET; Ac-9 (76%); N.A.-KS3a;
PNP27	Streptomyces varsoviensis NRRL B-3589	TET; Ac-9 (77.4%); N.A.-2 KS3a; no KR9; 4 GT; FOX3
PNP28	Streptosporangium roseum DSM 43021	PEN; FlvHal

Table 4-7. Actinobacteria selected for characterization of novel PK-IIs. These strains were chosen based on comprehensive analysis of PK-II BGCs encoded in their genomes. Abbreviations: PEN-pentangular; BIQ-benzoisochromanequinone; TCM-tetracenomycin; ANT-anthracycline; HED-hedamycin; ANG-angucycline; TET-tetracycline; R1128-R1128; Ac-acetyl; Pr-propionyl; Hex-hexanoyl; N.A.-none acetate primed; KS3a, b-ketosynthase III, subclass a and b; GT-glycosyltransferase; FlvHal-flavin-dependent halogenase; KR9-C9 ketoreductase; KRbiqC3S-BIQ C3 ketoreducatase, S stereochemistry; Cyc1, Cyc3, Cyc6, Cyc7, Cyc8-cyclase type 1, 3, 6, 7, 8; FOX1, 2b, 3-flavin-dependent oxygenase type 1, 2b, 3; ABMph-antibiotic biosynthesis monooxygenase (ABM) superfamily; ABMi-PdmI-like ABM; ABMsp-ABM associated with spore pigment; peptX-peptidase; PEase-putative esterase;

When possible, the predicted novel BGCs were chosen as subgroups of 2 or 3 similar "twin/triplet" BGCs to maximize the chance of identifying at least one transcriptionally active BGC/media combination from each group.

CRISPR/Cas9-guided In-frame Deletion of KSα in S. venezuelae. To establish the CRISPR/Cas9 based genome editing platform in our lab, initial experiment was carried out in the well-studied strain *Streptomyce venezuelae* ATCC 15439. This strain grows very fast and contains two PK-II BGCs, one of them is an interesting uncharacterized gene cluster revealed by the genome sequencing described in Chapter 2. Three 20 bp spacer sequence targeting the KSα gene of this cluster was chosen, because the KSα gene is an essential component in the biosynthesis of PK-II and deletion of it could completely abolish the production of PK-II. To introduce defined in-frame deletion of KSα gene via homologous recombination, two approximately 1kb homology arms flanking at the upstream and downstream of KSα was incorporated into the plasmid. Three knock-out plasmids, pCRISPR-ds, pCRISPR-ts and pCRISPR-sc were transferred into *S. venezuelae* ATCC 15439 via conjugation. As reported in previous study[2], the conjugation efficiency was much lower than no *cas9*-carring plasmids, indicating the inherent toxicity of Cas9 protein.

To screen the correct KSα deleted mutants, the genotypes of several exconjugants were analyzed by isolating the genomic DNA, PCR amplifying the target region, and Sanger sequencing the PCR amplified products. To ensure the DNA fragments were amplified from the genome rather than the knock-out plasmids, the primers, cmf-up-fwd and cmf-dn-rev, were designed at ~100 bp upstream and downstream of the homology arms (Figure 4-8a).

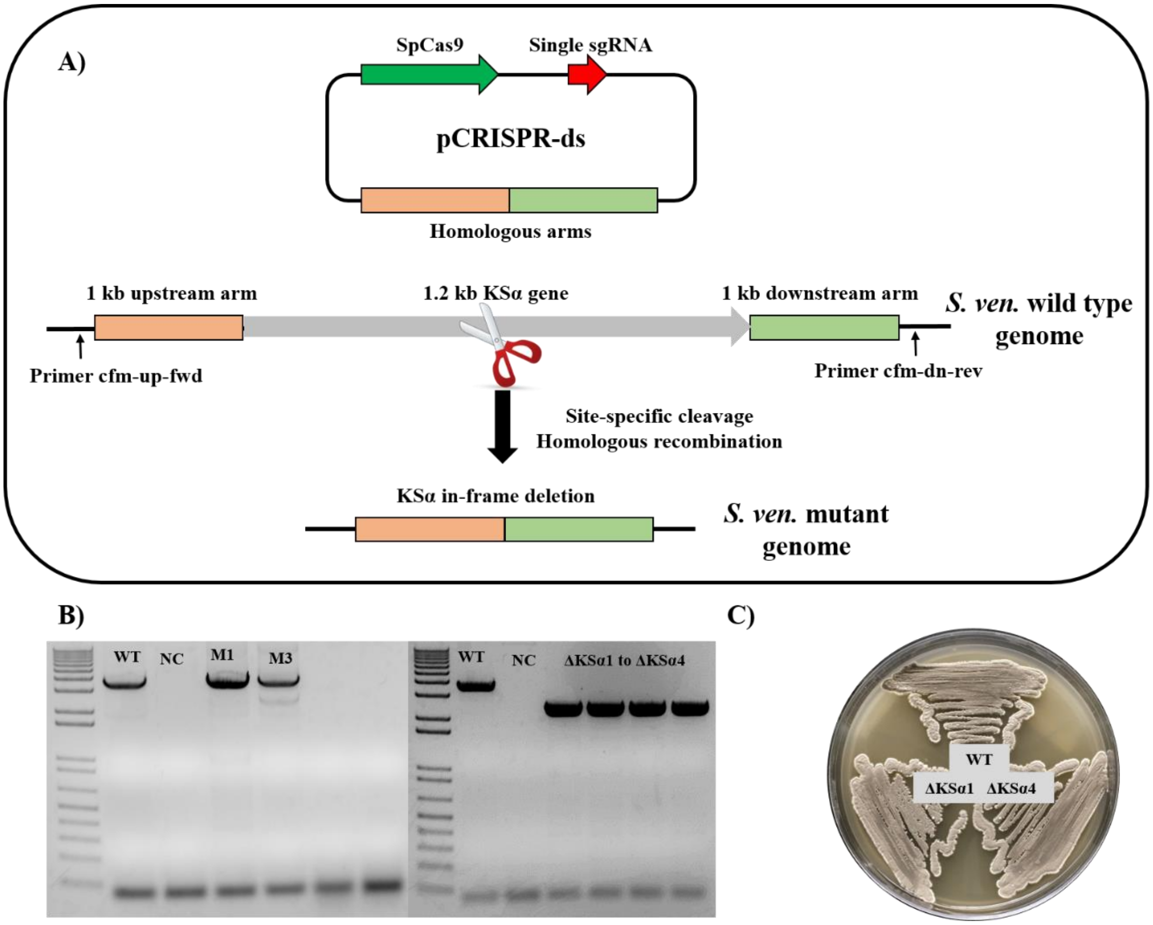

Figure 4-8. Schematic illustration of in-frame deletion of 1.2 kb KSα gene in *S. venezuelae*. A) Single sgRNA transcript guides Cas9 protein to cleave the targeted genomic DNA at the middle of the KSα gene, then two homologous arms bridge the gap via homologous recombination. B) PCR examination of exconjugants (left) and PCR verification of mutants (right) with wild-type (WT) control and knock-out plasmid pCRISPR-ds as negative control (NC). C) Phenotypes of KSα knock-out mutants compared with wild-type control. They were grown on GYM agar plate for 6 days.

Using this single spacer/sgRNA design, only 1 out of 5 exconjugants showed desired KSα deletion, while other 4 strains gave the same genotypes as wild-type strain (Figure 4-8b). After successful elimination of knock-out plasmid out of the exconjugant carrying the desired KSα deletion, the genotypes of four mutants were examined by the primer pair, cmf-up-fwd and cmf-dn-rev, and all were correct mutants (Figure 4-8b). Further confirmation of the mutants were performed by growing them on SPA supplemented with Apr and Sanger sequencing the target region.

Given all PK-II compounds have visible colors, we expected the KSα knock-out mutants would display distinct phenotypes on agar plate compared with wild-type strain. Thus, two mutants were examined with wild-type strain on a range of agar plates, including GYM agar plate (Figure 4-8c). However, after several days of growth, none of them showed phenotypic difference with wild-type control on these agar plates. Then, comparative metabolic analysis of two mutants and wild-type strain were carried out by growing them in a panel of pilot production media in parallel, extracting the metabolites, analyzing the metabolic profiles using HPLC analytical column. Again, no metabolic difference were detected, indicating this type II gene cluster might be transcriptionally silent under these testing fermentation conditions.

CRISPR/Cas9-guided In-frame Deletion of KSα in PNP19 and PNP20. Following the successful construction of KSα in-frame deletion mutants of *S. venezuelae* using CRISPR/Cas9 based genome editing, it would be appealing if this system could be successfully reconstituted in *Streptomyces* species identified by above bioinformatic analysis to achieve desired deletion of genes of interest. Prior to testing this possibility, the 9 *Streptomyces* strains were grown in a panel of pilot production media in parallel and metabolic profiles were examined as described above.

Based on the results of metabolic profiling (data not shown), the PK-II BGCs of 3 strains, *S. anulatus* NRRL B-3362 (PNP19), *S. bottropensis* ATCC 25435 (PNP20), and *S. flavochromogenes* NRRL B-2684 (PNP21) were very likely to be transcriptionally active. Thus these three species were subjected to KSα in-frame deletion using the same protocol as described for *S. venezuelae*. However, only PNP21 obtained intended KSα knock-out mutants, while PNP19 and PNP20 failed to screen out any correct exconjugants,

speculating that this single cleavage strategy is not efficient in PNP19 and PNP20 under the low conjugation efficiency. Thus, similar to the cas9-assisted cloning of large gene clusters[11], a dual-spacer/sgRNA construct strategy was developed in PNP19 (Figure 4-9) and PNP20.

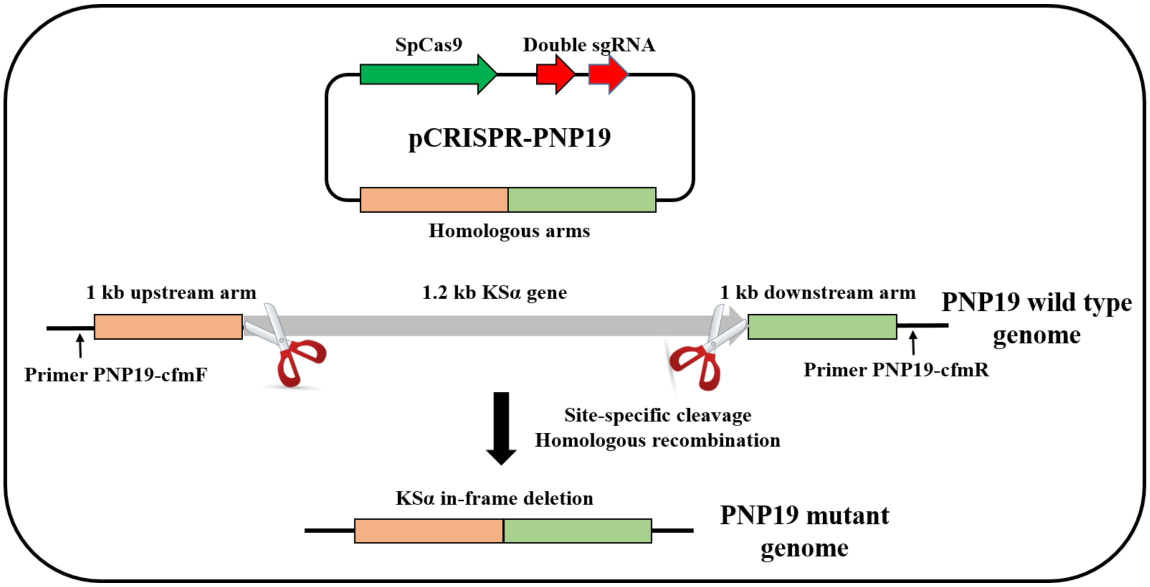

Figure 4-9. Schematic illustration of in-frame deletion of 1.2 kb KSα gene in PNP19. Double sgRNA transcript guides Cas9 protein to cleave the targeted genomic DNA at both ends of the KSα gene, then two homologous arms bridge the gap via homologous recombination.

The dual-spacer/sgRNA construct was supposed to induce simultaneous double-strand break at both ends of KSα genes on the genomes of PNP19 (Figure 4-9) and PNP20. The insertion of two sgRNA cassettes was accomplished by ligation of a DNA fragment containing the first spacer, sgRNA, a terminator, a promoter, and the second spacer into the vector pCRISPomyces-2. Using an analogous conjugation protocol as described for *S. venezuelae*, significantly higher conjugation efficiency was observed, providing preliminary support for the idea that dual-spacer/sgRNA design could afford higher genome editing efficacy. Further genotypic evaluation are needed to be carried out to offer solid evidences.

CRISPR/Cas9-guided In-frame Deletion of KSα in PNP21. As mentioned above, following an analogous single-spacer/sgRNA design, conjugation and genotypic evaluation pipeline as described for *S. venezuelae*, four out four colonies were shown to possessed desired KSα deletion (Figure 4-10a), which was benefited from passing two generations after conjugation. Three mutants were picked and their genotypes (Figure 4-10a) were examined and Apr sensitivity were confirmed as described for *S. venezuelae*.

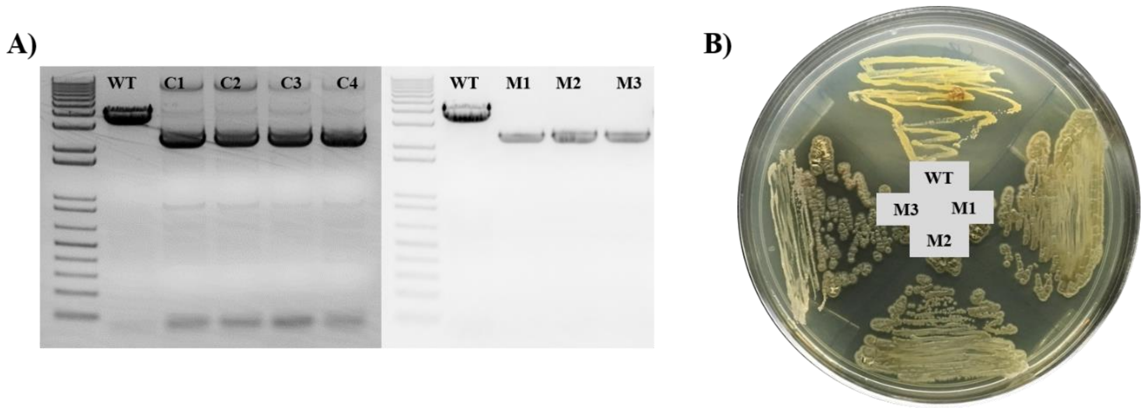

Figure 4-10. Genotypic and phenotypic evaluation of KSα in-frame deletion mutants of PNP21. A) PCR examination of colonies (left) and PCR verification of mutants (right) with wild-type (WT) control. C) Phenotypes of KSα knock-out mutants compared with wild-type control. They were grown on GYM agar plate for a week.

Comparative Metabolic Profiling of Wild-type and Mutants of PNP21. Similarly, these KSα in-frame deletion mutants were examined with wild-type strain on a range of agar plates. After several days of growth, all the mutants exhibited distinct phenotypes compared with wild-type control on GYM agar plate that yellow compounds produced by wild-type strain disappeared in mutants (Figure 4-10b). Then, comparative metabolic analysis of two mutants, M1 and M3, and wild-type strain were conducted by growing them 50 mL of GYM liquid medium in parallel, extracting the metabolites, and analyzing the metabolic profiles using HPLC analytical column. In consistent with the phenotypic difference observed on GYM agar plate, obvious metabolic difference were detected

(Figure 4-11), indicating this type II gene cluster is transcriptionally active. Furthermore, in agreement with above bioinformatic analysis, two peaks displayed characteristic angucycline uv-visible profiles with strong absorption at 232 nm, 282 nm and 406 nm (Figure 4-11), suggesting they might be congeners.

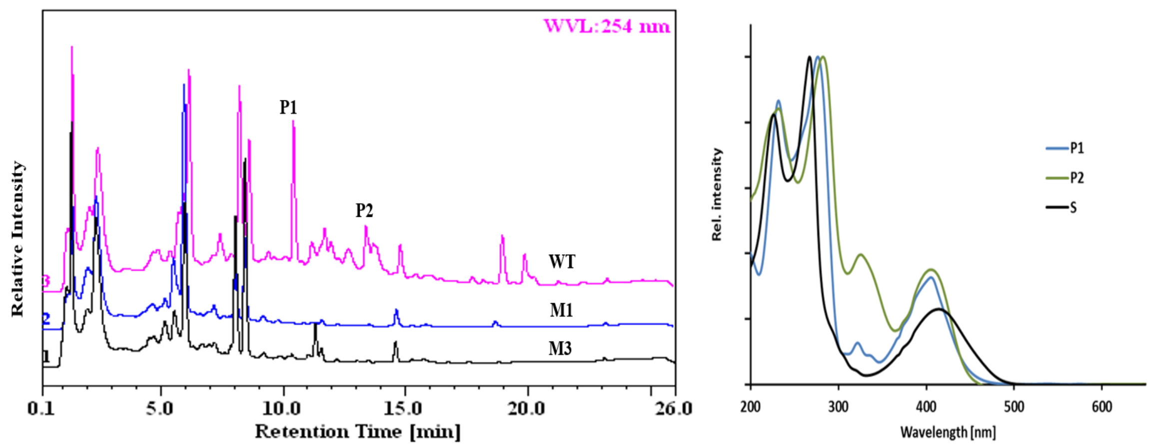

Figure 4-11. HPLC comparative metabolic profiling of KSα in-frame deletion mutants of PNP21. Mutants, M1 and M3, were grown in parallel with wild-type strain (WT) in 50 mL of GYM liquid medium for 7 days. The uv-visible profile of the major peak (P1), minor peak (P2) and standard angucycline-type compound rabelomycin (S) are shown on the right.

PK-II of PNP21 Isolation and Structure Elucidation. In an effort to access the structure of the product of the target PK-II BGC in PNP21, a large scale (7 liter) fermentation using GYM liquid medium was conducted, affording 2.9 mg of the major peak, termed flavochromycin, which was purified by progressive chromatography. Flavochromycin was isolated as a yellow amorphous solid that was soluble in methanol.

The molecular formula of flavochromycin was deduced to be $C_{28}H_{16}O_{12}$ by HR-ESI-MS (m/z 543.2001 $[M-H]^-$, calculated 543.0560) and (m/z 545.1911 $[M+H]^+$, calculated 545.0720) (Figure 4-12). Compared with all identified angucycline-type PK-IIs, flavochromycin has a different molecular formula, indicating it is a new compound. Further

NMR experiments are needed to be conducted to solve the chemical structure of flavochromycin.

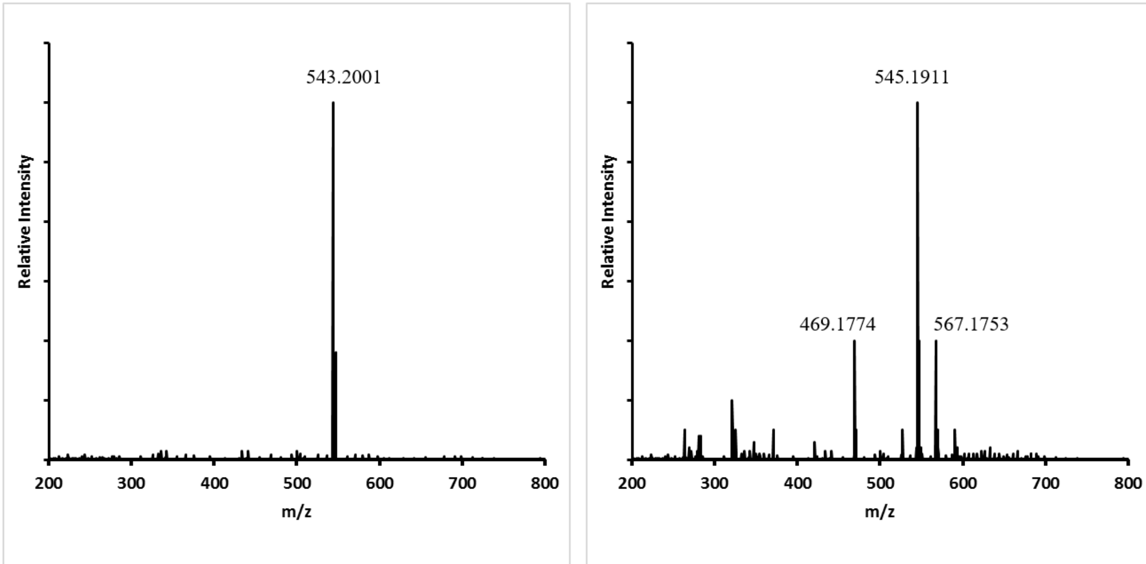

Figure 4-12. Mass spectra of flavochromycin. HR-ESI-MS negative is on the left, while the positive mode is on the right.

4. Conclusions

In this work, the KSα genes of PK-II BGCs of interest from two different *Streptomyces* species, *S. venezuelae* and *S. flavochromogenes*, were deleted using CRISPR/Cas9 system-based genome editing. This work provides additional support for the broad applicability of the type II CRISPR/Cas9 system of *S. pyogenes* in *Streptomyces* species to achieve desired genome editing. Two different strategies for KSα gene inactivation, either single sgRNA cassette or double sgRNA cassette, were tested. The results have shown that the dual sgRNA cassette design gave much higher conjugation efficiency, suggesting potentially higher efficiency in KSα deletion, but further works are needed to provide experimental evidence. In addition, various molecular cloning techniques were used in construction of CRISPR/Cas9 plasmids, including modern DNA

assembly methods such as overlapping PCR and Gibson assembly. This work proves that the CRISPR/Cas9-based genome editing is superior to previous tools for defined genome editing including targeted gene deletion as demonstrated in this work.

This work reported here demonstrates the synergy of our bioinformatic analysis and the CRISPR/Cas9 system-based PK-II BGC inactivation in discovering novel PK-IIs, and provides the first experimentally characterized example of a PK-II produced by a gene cluster containing a member of the divergent group of KSα/β sequences and featuring non-acetate priming. Through *Dynamite*, we bioinformatically identified over 500 PK-II BGCs from the NCBI databank. Based on a range of structural novelty attributes, we selected 28 PK-II BGCs predicted to produce structurally novel PK-II compounds for experimental characterization. KSα inactivation was achieved in *S. flavochromogenes*, and comparison of wild-type and mutant metabolite profiles led to identification of new putative PK-II compounds. Using progressive chromatography, we purified one of the metabolites and obtained uv-visible and mass spectral evidence consistent with an angucycline-type PK-II compound, which we named flavochromycin. Further works would be carried out to solve the structure of this intriguing PK-II compound and to characterize other minor congeners, which may lead to new biosynthetic understanding of PK-II systems.

The application of the CRISPR/Cas9 technology in natural product discovery is not limited to gene cluster inactivation by precise deletion of key biosynthetic genes. Most recently, the CRISPR/Cas9 system has been applied to trigger the expression of multiple silent gene clusters in *Streptomyces* species by promoter knock-in at defined positions. This CRISPR/Cas9 system-based knock-in strategy led to the activation of different types of natural product BGCs in five *Streptomyces* strains and the production of unique

metabolites, including a novel pentangular polyphenol, which indicates the potential to scale up and generalize in activating silent BGCs in the genus *Streptomyces*[12].

5. References

1. Land, M., Hauser, L., Jun, S. R., Nookaew, I., Leuze, M. R., Ahn, T. H., ... & Poudel, S. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*, *15*(2), 141-161.

2. Cobb, R. E., Wang, Y., & Zhao, H. (2014). High-efficiency multiplex genome editing of Streptomyces species using an engineered CRISPR/Cas system. *ACS synthetic biology*, *4*(6), 723-728.

3. Tong, Y., Charusanti, P., Zhang, L., Weber, T., & Lee, S. Y. (2015). CRISPR-Cas9 based engineering of actinomycetal genomes. *ACS synthetic biology*, *4*(9), 1020-1029.

4. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), 816-821.

5. Ho, S. N., Hunt, H. D., Horton, R. M., Pullen, J. K., & Pease, L. R. (1989). Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene*, *77*(1), 51-59.

6. Gibson, D. G., Young, L., Chuang, R. Y., Venter, J. C., Hutchison, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods*, *6*(5), 343-345.

7. Bierman, M., Logan, R., O'brien, K., Seno, E. T., Rao, R. N., & Schoner, B. E. (1992). Plasmid cloning vectors for the conjugal transfer of DNA from Escherichia coli to Streptomyces spp. *Gene*, *116*(1), 43-49.

8. Keiser, T., Bibb, M. J., Buttner, M. J., Chater, K. F., & Hopwood, D. A. (2000). Practical streptomyces genetics. *The John Innes Foundation, Norwich*.

9. Ogasawara, Y., Yackley, B. J., Greenberg, J. A., Rogelj, S., & Melançon III, C. E. (2015). Expanding our understanding of sequence-function relationships of type II polyketide biosynthetic gene clusters: bioinformatics-guided identification of Frankiamicin A from Frankia sp. EAN1pec. *PloS one*, *10*(4), e0121505.

10. Tormo, J. R., Garcia, J. B., DeAntonio, M., Feliz, J., Mira, A., Díez, M. T., ... & Pelaez, F. (2003). A method for the selection of production media for actinomycete strains based on their metabolite HPLC profiles. *Journal of Industrial Microbiology and Biotechnology*, *30*(10), 582-588.

11. Jiang, W., Zhao, X., Gabrieli, T., Lou, C., Ebenstein, Y., & Zhu, T. F. (2015). Cas9-Assisted Targeting of CHromosome segments CATCH enables one-step targeted cloning of large gene clusters. *Nature communications*, *6*.

12. Zhang, M. M., Wong, F. T., Wang, Y., Luo, S., Lim, Y. H., Heng, E., ... & Zhao, H. (2017). CRISPR-Cas9 strategy for activation of silent Streptomyces biosynthetic gene clusters. *Nature Chemical Biology*, *13*(6), 607-609.