NATURAL PRODUCT DISCOVERY AND ENGINEERING AT THE PROTEIN, PATHWAY, AND GENOME SCALES

BY

RYAN ELLIS COBB

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Chemical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor Huimin Zhao, Chair
Professor William W. Metcalf
Professor Satish K. Nair
Professor Christopher V. Rao

Abstract

For centuries, mankind has turned to natural sources for cures, remedies, and ways to improve its quality of life. It was not until the discovery of the antibiotic penicillin in 1928, however, that natural products – small molecule secondary metabolites produced by a variety of organisms for a variety of purposes – were truly appreciated as the source of these beneficial properties. In the ensuing decades, research in the field of natural products boomed, and a number of discoveries were made that revolutionized global health. In the 1970s and 1980s, however, these discoveries started to become fewer and farther between, prompting pharmaceutical companies to turn away from natural products research in favor of synthetic chemistry approaches to drug discovery. Nevertheless, despite diminishing returns from traditional natural product discovery methods, modern genomics has in fact revealed that vast numbers of natural product gene clusters exist across all domains of life, far in excess of the number of known natural products and far surpassing any previous predictions. Thus, natural product discovery efforts to date have only scratched the surface of Nature's true capabilities.

In this work, we sought to leverage modern techniques for natural product discovery at the protein, pathway, and genome scales to tap into this biosynthetic potential. The past several years have seen the development of many tools for the manipulation of DNA with ease and precision far exceeding the standards set by traditional methods. These methods, combined with the ever-increasing number of putative natural product pathways revealed by genome sequencing efforts, open the door for new platforms of natural product discovery and engineering. At the protein level, we demonstrate the synthesis of novel derivatives of the antimalarial natural product FR-900098 by leveraging the substrate promiscuity of the native biosynthetic machinery.

Through structural and biochemical characterization of the *N*-acetyltransferase FrbF and the corresponding target for FR-900098 inhibition, Dxr from *Plasmodium falciparum*, we show that the novel FR-900098 derivatives can serve as more potent inhibitors. A platform for their biosynthesis is also established in *Escherichia coli*.

We additionally demonstrate a genome-mining platform for fungal polyketide synthases via the one-step assembly of expression-ready plasmids by homologous recombination in yeast. Through the evaluation of previously uncharacterized endogenous promoters from *Saccharomyces cerevisiae*, we demonstrate the heterologous production of polyketides from the dimorphic fungus *Talaromyces marneffei*. Extension of this approach to the pathway level is also demonstrated to assemble both a 6-gene resorcylic acid lactone cluster and a 13-gene phosphonic acid biosynthetic cluster. The latter case, in which all 13-genes from a phosphonic acid non-producing strain were placed under individual strong promoters, enabled the production of novel phosphonic acid compounds when heterologously expressed in *Streptomyces lividans*.

At the genome scale, we developed a versatile system from genome engineering in a broad range of *Streptomyces* species. As the genus *Streptomyces* is by far the most important producer of pharmaceutical natural products to date, tools to facilitate their genetic engineering are of significant interest to aid discovery and improve production. Here, we show that the CRISPR/Cas9 system of *S. pyogenes* can be reconstituted in *S. lividans* for high-efficiency, multiplex editing of genomic loci. We show that deletions ranging from 20 bp to 31 kb can be successfully introduced in one step, and extend this approach to additional *Streptomyces* strains.

To my parents,
for all of their love and support

Acknowledgements

"We don't learn. We just get older, and we know."
Leslie Caron, "Lili" (1953, Metro-Goldwyn-Mayer Studios)

First and foremost, I thank my advisor, Prof. Huimin Zhao, for granting me the opportunity to perform graduate research in his lab. His patience and trust in my judgment enabled me to pursue my goals to the best of my ability. Further, I am grateful for the many opportunities he presented me to develop my love of writing. I also thank my committee, Profs. Bill Metcalf, Satish Nair, and Chris Rao, for the valuable feedback they have given me, along with Profs. Wilfred van der Donk and Neil Kelleher of the MMG theme. For technical assistance, I am grateful to Dr. Lucas Li and Dr. Alex Ulanov of the Roy J. Carver Metabolomics Center; Dr. Jiangtao Gao, Dr. Brad Evans and Dr. Jaeheon Lee of the IGB MMG theme; Dr. Furong Sun of the SCS Mass Spectrometry Facility; and Dr. Xudong Guan of the IGB Core Facilities. For financial support, I am grateful for the NIH Chemistry-Biology Interface Training Grant and the SCS Harry Drickamer Fellowship.

And finally, to all the rest: this process is long and sometimes arduous, with many twists, turns, false starts, and dead ends along the way. Certainly we are different in leaving than we were upon entering, but by the end it can be easy to lose sight of the journey as chronology is sacrificed for the benefit of the narrative. Over time, that which has become known– detailed in the pages that follow– becomes the story, supplanting the learning process from which it came. Nevertheless, know that I am indebted to all of those who have helped to shape my journey and, in doing so, myself. Over the years, I have learned so much from so many that any attempt to list them all would surely fall short of its mark, to the chagrin of the author and the disappointment of those omitted. Thus will I acquiesce to brevity with a simple "thank you."

Table of Contents

CHAPTER 1. Natural product discovery and engineering[1]

1.1 Natural Products: Invaluable Natural Resources

For centuries, mankind has turned to Nature for the means by which to improve quality of life. Microbial fermentation, for example, has been extensively utilized for the production of beer, bread, and wine, with the oldest known fermentations occurring at least 9000 years ago [1]. In the area of human health, natural sources have long been known to possess therapeutic properties. Plants such as *Artemisia annua* (sweet wormwood) and *Salix alba* (white willow) are referenced in texts from thousands of years ago for their healing potential. It was not, however, until Alexander Fleming's 1928 discovery of the antibacterial penicillin that small molecule secondary metabolites, known simply as natural products, were fully appreciated as the effectors of these beneficial properties [2]. In the years that followed, natural products saw intense research interest as scientists searched and screened to find the next "miracle drug." Evidencing this prodigious effort is the CRC *Dictionary of Natural Products*, which currently lists 170,000 unique natural products [3].

At present, natural products continue to see extensive pharmaceutical use. In the period from 1981–2010, approximately half of all small molecule NCEs (new chemical entities) were natural products, natural product derivatives, natural product mimics, or synthetic analogues of natural

[1]Portions of this chapter are adapted with permission from the following:
Cobb, R.E., Sun, N., and Zhao, H. (2013) Directed evolution as a powerful synthetic biology tool. *Methods*, **60**, 81-90.
Cobb, R.E., Chao, R., and Zhao, H. (2013) Directed evolution: past, present, and future. *AIChE J*, **59**, 1432-1440.
Cobb, R.E., Luo, Y., Freestone, T. and Zhao, H. (2013) Drug discovery and development via synthetic biology. In Zhao, H. (ed.) *Synthetic Biology: Tools and Applications*. Elsevier, New York, pp. 183-206.
Cobb, R.E., Ning, J.C., and Zhao, H. (2014) DNA assembly techniques for next-generation combinatorial biosynthesis of natural products. *J Ind Microbiol Biotechnol*, **41**, 469-477.
Luo, Y., Cobb, R.E., and Zhao, H. (2014) Recent advances in natural product discovery. *Curr Opin Biotechnol*, **30**, 230–237.

product pharmacophores [4]. Despite this success, pharmaceutical companies in the 1990s began to turn away from natural product discovery, which required significant investments of time and resources, toward synthetic chemical libraries, which could be synthesized in large numbers through combinatorial approaches [5]. While such libraries can be more easily prepared and screened, they tend to feature limited chemical complexity compared to the diversity of natural products.

For pharmaceutical companies, a significant liability in natural product discovery efforts was the increasing rate of rediscovery of known compounds. This finding raised concerns over the overall diversity of compounds found in Nature, and whether this supply was reaching exhaustion. In the academic field, however, these fears were abated as genome sequencing began to take off. Across all domains of life, a major finding in sequenced genomes was the presence of a large number of putative secondary metabolite pathways for which the corresponding natural product was not known [6]. Amazingly, such "cryptic" clusters were found even in well-studied microbial species, suggesting a silent majority of natural products waiting to be discovered with the right tools or the right conditions. When this observation is coupled with the notion that the vast majority (> 99 %) of microorganisms remain uncultivable in the laboratory setting [7], it becomes apparent that there still exists a great wealth of as yet undiscovered natural products for the treatment of any number of human diseases. The grand challenge, then, is to develop the tools and methodologies needed to find these compounds and develop them into the next generation of pharmaceuticals.

1.2 Methods for Natural Product Discovery

1.2.1 Pathway-independent approaches

1.2.1.1 Environmental sampling

Life is manifested in many unique and varied forms to suit a broad range of physical and ecological environments. This diversity can be traced down to the chemical level as organisms have grown and evolved to maximize their own fitness in the face of varied stresses and competitors. It follows, then, that diverse sampling in ecological space will also lead to diverse sampling in chemical space.

Early on, soil sampling was the most common means to access microbial diversity, and many notable successes were achieved. The antibiotic kanamycin, for example, was identified from a *Streptomyces kanamyceticus* strain isolated from a Japanese flower bed in 1957 (Figure 1.1) [8]. Remarkably, development of the compound into a widely available commercial drug took less than a year– a feat irreproducible with modern regulatory standards. In the soil outside of an Italian castle, another *Streptomyces* strain was identified that produced the anthracycline chemotherapeutic daunorubicin [9]. The antibiotic chloramphenicol (originally chloromycetin) was discovered in 1947 from a Venezuelan soil isolate [10]. Notably, the same compound was concomitantly discovered in a soil isolate from Urbana, Illinois [11], highlighting the (now common) observation that physical distance is no guarantee of uniqueness. In fact, identical natural product gene clusters can be found from strains all over the world [12], making rediscovery of known compounds a significant concern.

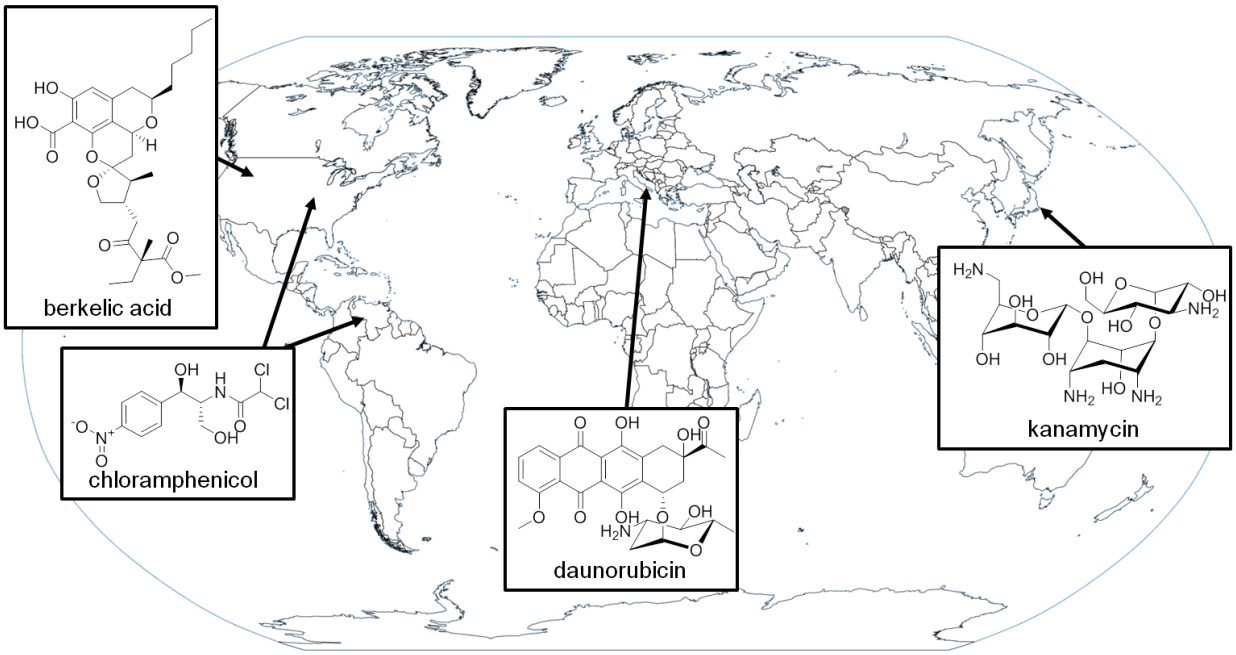


Figure 1.1: Representative natural products from global sampling efforts.

Outside of "traditional" soil sampling, a more recent trend has been to explore more extreme environments in search of greater chemical diversity. For example, sampling from a salt field led the discovery of a halophilic actinomycete that produces three novel polyketides with anticancer properties [13]. In the acidic, metal-contaminated Berkeley Pit copper mine, a fungus was identified that produced the novel anticancer spiroketal berkelic acid [14]. Recently, sampling from a microbial mat in an iron-rich spring led to the characterization of six novel fungal natural products: clearanols A-E and disulochrin [15]. While the native biological functions of compounds isolated from extremophiles are generally unknown, they likely play a role in helping the native producers cope with harsh environmental stresses.

To access even greater natural product diversity, sampling from marine environments has become a popular approach. Compared to natural products of terrestrial origin, marine natural

products exhibit a greater diversity of molecular scaffolds and have shown a higher incidence of clinically relevant bioactivity [16]. The jamaicamides, discovered from marine cyanobacterium *Lyngbya majuscula*, provide an excellent example of this rich chemical diversity; jamaicamide A, for example, features such uncommon moieties as a vinyl chloride, an alkynyl bromide, a β-methoxy eneone system, and a pyrrolinone ring (Figure 1.2) [17]. Recently, anthracimycin was discovered from a marine actinomycete and found to possess activity against the significant human pathogens *Bacillus anthracis* and methicillin-resistant *Staphylococcus aureus* (MRSA) [18].

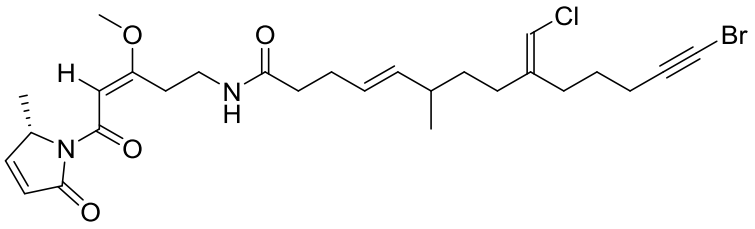


Figure 1.2: Structure of jamaicamide A.

The great diversity of life in marine environments has also led to particularly unique interspecies relationships. Endosymbiotic bacteria, for example, represent a rich source of natural products [19], as many higher organisms rely on them to supplement their own defense systems. In a family of bivalve mollusks known as shipworms, boronated natural products with antibiotic activity were recently discovered from the cultivated symbiotic bacterium *Teredinibacter turnerae* [20]. Similarly, diverse pyrone polyketides have been discovered from *Nocardiopsis alba* CR167, an actinomycete endosymbiont of the cone snails *Conus rolani* and *Conus tribblei* [20]. Of course, symbiosis is not a unique property of marine environments. For example, several agriculturally relevant insects have also proven to be sources of antibiotic natural

products [21]. Nevertheless, marine environments currently seem to be the most promising for natural product discovery overall, with discovery rates increasing every year. In 2012 alone, over 1200 new chemical entities were discovered from marine environments [22].

1.2.1.2 Culture manipulation

While laboratory cultivation of microorganisms is typically conducted under well-defined conditions (often with an abundance of nutrients), in Nature growing conditions vary widely in time and space. As a result, an organism must be able to adapt to survive. At the metabolic level, this entails dynamic regulation of anabolic and catabolic pathways to maximize efficiency and competitive advantage. To culture an organism under a single set of conditions, then, is to observe only one dimension of its full biosynthetic potential.

To mimic in a laboratory setting the diverse environmental stresses an organism may experience in nature, screening of different growth conditions is often employed. This process is generally termed the "one strain-many compounds" (OSMAC) approach [23]. As an early example, the fungus *Aspergillus ochraceus* DSM7428 was initially known to produce only one secondary metabolite. However, by varying the culture conditions and culture vessels, 15 additional compounds were discovered. A recent example of this approach is the utilization of chemostat fermentation of *Aspergillus nidulans* to manipulate specific growth rate as well as carbon, nitrogen, and phosphorous levels [24]. Transcriptomic and metabolomic analyses revealed production of many polyketides under different conditions, including one novel compound. Besides systematic control of nutrient levels, more unorthodox culture techniques have also been employed for natural product discovery. For example, fermentation medium containing

Cheerios breakfast cereal was found to promote growth and secondary metabolism of several isolated fungal strains, including the production of a novel diarylcyclopentendione metabolite and a novel biphenyl metabolite from *Preussia typharum* (Sacc.) Cain [25]. Also, an *in vivo* culture method to elicit the production of typically silent glidobactin/luminmycin metabolites from *Photorhabdus asymbiotica* was demonstrated by injecting the bacterium into live crickets and extracting the carcasses [26]. It should be noted, however, that the four compounds observed, including two new derivatives, were first identified in defined liquid medium culture.

An alternate technique to introduce external stresses in laboratory culture is to challenge a target organism with other co-cultured species. This method has been demonstrated for many biotechnological applications [27,28]. For example, cultivation of the marine fungus *Emericella* sp. with the actinomycete *Salinispora arenicola* activated production of the antibiotic emericellamides by the fungus [29]. As a recent example, a novel prenylated polyketide was discovered from the fungal pathogen *Aspergillus fumigatus* only when cultured with the actinomycete bacterium *Streptomyces rapamycinicus* [30]. A particularly interesting example is the co-culture of the bacteria *Rhodococcus fascians*, not known to produce antibiotics, and *Streptomyces padanus*, an actinomycin producer [31]. Following co-culture, a surviving *R. fascians* mutant was discovered that produced novel aminoglycoside antibiotics rhodostreptomycin A and B. Interestingly, antibiotic production was correlated with the acquisition of DNA from the *S. padanus* strain, but the compounds produced did not match known *Streptomyces* metabolites.

1.2.1.3 Global genetic and epigenetic manipulation

While the approaches described in the previous section manipulate an organism's external cues to elicit a desired response, a complimentary approach would be to manipulate an organism's internal cues. A proven means by which to accomplish this goal is to introduce mutations to housekeeping genes. In actinomycetes, RNA polymerase and ribosomal protein S12 have been popular targets through the isolation of mutant strains resistant to the antibiotics rifampicin and streptomycin, respectively [32-34]. For example, the Ochi group screened 1068 soil-isolated actinomycetes, and found that 6 % of non-*Streptomyces* strains and 43 % of *Streptomyces* strains produced antibiotics only after isolation of spontaneous resistance mutations in *rpoB* (RNA polymerase β-subunit) and/or *rpsL* (ribosomal protein S12). Other targets for this so-called "ribosome engineering" approach include *rsmG*, encoding 16S ribosomal RNA methyltransferase, the mutation of which activated the transcription of several cryptic secondary metabolite genes [35]. Notably, this approach is not limited to actinomycetes. In *Bacillus subtilis*, an RNA polymerase mutation activated production of an aminosugar antibiotic [36].

Regulatory proteins can also serve as targets for the induction of secondary metabolism. For example, over-expression of the Crp transcription factor, a cyclic AMP receptor protein, was found to activate production of known and unknown antibiotics in multiple *Streptomyces* species [37]. In fungi, nuclear protein LaeA has been found to be a global regulator of secondary metabolism [38]. While many secondary metabolite clusters have been found to be positively regulated by LaeA, discovery of a novel compound via LaeA overexpression has not yet been achieved [39]. An alternative global strategy in fungi has been the knockout of kinases, known to play important roles in regulation and signal transduction. In *Aspergillus nidulans*, screening

of a genome-wide kinase knockout library led to identification of the novel prenylated isoindolinone aspernidine A, along with its corresponding gene cluster [40].

Outside of genetic modification, epigenetic modification has been demonstrated as a means to induce the expression of otherwise silent gene clusters in eukaryotes [41]. In a pioneering study in this area, the Cichewicz group treated fungi with the histone deacetylase (HDAC) inhibitor suberoylanilide hydroxamic acid (SAHA) and the DNA methyltransferase inhibitor 5-azacytidine, eliciting the production of novel oxylipins, perylenequinones, and polyketides. This technique has similarly been applied to discover novel meroterpenes from *Penicillium citreonigrum* [42], novel fusaric acid derivatives from plant endophytic fungi [43], and a novel coumarin from *Pestalotiopsis crassiuscula* [44], among other examples. Curiously, treatment with HDAC inhibitors has also been evaluated in *Streptomyces* strains [45]. Though lacking histones, *Streptomyces* species do possess HDAC-like proteins. Treatment with various HDAC inhibitors was found to stimulate some known antibiotic pathways in *Streptomyces coelicolor*, though no novel compounds were discovered.

1.2.2 Pathway-specific approaches

1.2.2.1 Bioinformatic tools to identify natural product gene clusters

Remarkable advances in DNA sequencing technologies have made available the complete genome sequences for thousands of organisms [46]. Encoded in this data are the templates for countless proteins which catalyze a myriad of chemical transformations, many of which are involved in secondary metabolic pathways. The ability to identify these genetic blueprints for natural product biosynthesis has inspired a paradigm shift in natural product discovery, in which

identification of an interesting gene cluster precedes discovery of its cognate natural product. Given the sheer volume of data that has been generated, *in silico* tools are absolutely necessary to identify, annotate, and prioritize potentially interesting genes and pathways.

The earliest gene identification algorithms were developed mainly for the analysis of shorter DNA sequences in which the exact coding sequence of a protein was ambiguous. These methods were reasonably simplistic, but provided fairly accurate predictions of "coding" versus "non-coding" sequences [47]. Subsequent prediction algorithms employed more sophisticated approaches to achieve better results. For example, the GeneMark (initially GENMARK) program of Borodovsky and McIninch combined non-homogeneous Markov chain models with Bayesian decision making for coding sequence prediction [48]. This program also introduced simultaneous analysis of both DNA strands as a method to improve accuracy. As the sequencing of entire genomes became realized, the need for reliable gene prediction was underscored. To improve the GeneMark program for entire bacterial genomes, a hidden Markov model framework was implemented, as well as recognition of ribosome binding site sequences [49]. Further improvements came with the implication of self training for new prokaryotic genome sequences [50] and expansion to eukaryotic and viral systems [51,52]. GLIMMER represents a complementary tool for gene identification that was built on interpolated Markov models [53,54]. This tool has similarly been adapted to eukaryotic DNA [55,56] as well as endosymbiont and metagenome DNA [57,58].

Identification of coding sequences is an important first step, but sequence information alone is not enough to describe a protein's utility. Bench-top experiments both *in vitro* and *in vivo* are, of

course, the best way to determine protein function. However, the vast success in DNA sequencing and coding sequence identification has provided such a wealth of putative protein targets that laboratory characterization of them all is simply not feasible. Fortunately, if two proteins have similar primary sequences, it is quite likely that they will also share similar functions. As a result, sequence alignment and homology analysis based on proteins of known function has proved vital to the accurate prediction of protein function from sequence data alone.

The earliest exercises in protein homology comparisons were carried out to evaluate evolutionary relationships rather than to predict function [59-61]. Nevertheless, these algorithms provided the groundwork upon which subsequent protein alignment tools would be built. In 1985, Lipman and Pearson noted the increasing number of protein sequences made available at the time, and that functions could be inferred by comparison to other characterized proteins. As a result, they developed the FASTP algorithm for rapid *in silico* comparison of a query sequence to a protein sequence database [62]. This was followed by FASTA, which featured improved sensitivity, and LFASTA, which allowed for analyses of local similarity [63,64]. Other tools followed, such as MSA and CLUSTAL W, for the high sensitivity alignment of smaller sets of proteins [65,66].

Currently, state-of-the-art sequence database search tools employ more sophisticated algorithms to reduce computation time and increase sensitivity to weak similarities. For example, tools such as SAM [67] and HMMER [68] employ hidden Markov models to identify sequence homology. Perhaps the most widely used search tool is the Basic Local Alignment Search Tool, or BLAST, first presented by Altschul and coworkers in 1990 [69]. At its inception, BLAST offered high sensitivity database searching at speeds much faster than any previous algorithm, and proved

amenable to mathematical and statistical analysis. Subsequent versions, such as gapped BLAST and position-specific iterative (PSI) BLAST, have further improved computation time and sensitivity to weak, but still biologically relevant, similarity [70]. Functional prediction via BLAST is further enhanced through coupling with the Conserved Domain Database (CDD) [71], which integrates data from sources such as Pfam [72] and SMART [73] to identify regions of the query sequence with evolutionarily conserved functions, such as binding a metal ion or cofactor. BLAST results can also be coupled to tools such as GCView [74], which enable analysis of the genomic context of search results to facilitate more accurate functional prediction.

Predictive algorithms have also expanded beyond individual coding sequence queries to a variety of other targets. For example, IsoRankN enables the alignment of entire protein-protein interaction networks for the prediction of functional orthologs across species [75]. Tools such as PromPredict [76], ConTra [77], and RSAT [78], among others, focus not on protein sequences, but on the sequences of regulatory regions such as promoters and transcription factor binding sites.

With accurate prediction of the functions of encoded proteins, it has become possible to identify gene clusters for particular classes of natural products based upon the presence of characteristic enzymes. In particular, large, multi-domain polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs) have been especially amenable to this approach. In the past two decades, a number of computational tools have been developed not only for the identification of natural product gene clusters from DNA sequence data, but also the prediction of their corresponding products. Some of the earliest efforts toward *in silico* prediction of NRPS

products focused on the specificity of adenylation domains. In 1997, de Crécy-Lagard and coworkers examined 55 adenylation domain sequences to devise rules for specificity prediction, but found that they could only come up with good predictions in 43 % of cases [79]. Two years later, however, analysis of the crystal structure of the adenylation domain PheA involved in gramicidin S biosynthesis enabled two groups to provide much more accurate specificity predictions. Stachelhaus and coworkers identified 10 specificity-conferring residues, allowing 86 % accuracy in specificity prediction [80]. Challis and coworkers took a very similar approach, identifying an 8 residue signature sequence [81]. More recently, a sophisticated prediction algorithm based on transductive support vector machines have been devised that also incorporates the physico-chemical properties of the residues lining the active site to narrow down its predictions [82]. This tool is available online via NRPSpredictor2 [83]. For the analysis and comparison of the nonribosomal peptide products themselves, the database NORINE was established, which currently contains over 1100 compounds [84].

For the analysis of PKSs, the earliest computational tools focused on the identification of their constituent domains [85,86], as well as of the linker regions between them [87]. The purpose here was not only to aid product prediction, but also facilitate the reconstitution of individual domains and the combinatorial biosynthesis of "unnatural" natural products via domain swapping. Identification of the specificity-determining residues of acyltransferase domains was again enabled by crystallographic data, allowing prediction of malonyl-CoA or methylmalonyl-CoA specificity [85,86]. More recent tools have afforded more accurate predictions and increased functionality. For example, ASMPKS allows the input of entire genome sequences for analysis [88]. As with NRPSs, support vector machines have been applied to type III PKSs to

afford improved prediction accuracy [89]. Finally, SBSPKS allows PKS domains and modules to be modeled and docked to better predict and engineer inter-subunit contacts, as well as to predict the order of substrate channeling in gene clusters with multiple PKS open reading frames [90].

Currently, the bulk of the state-of-the-art computational tools for analysis of secondary metabolites focus on concomitant analysis of both PKS and NRPS genes. These tools, such as NRPS-PKS [91], ClustScan [92], CLUSEAN [93], NP.Searcher [94], and the PKS-NRPS Analysis Web-site [95] continue to improve upon the accuracy and functionality of their predecessors while providing user-friendly interfaces and increased computational power. Nevertheless, other classes of secondary metabolites have seen increased attention in recent years as well. For example, the genome mining tool BAGEL2 focuses exclusively on bacteriocins, which are ribosomally-synthesized antimicrobial peptides from bacteria [96]. BAGEL2 considers conserved domains, physical properties and genomic context to identify putative bacteriocins, which can easily be missed by other genome annotation tools due to their short size (<100 amino acids). Another example is the SMURF program, which can be used not only to scan fungal genome sequences for gene clusters producing polyketides and nonribosomal peptides, but also for indole alkaloids and terpenes by identifying prenyltransferases and terpene cyclases, respectively [97]. Finally, a new program called antiSMASH promises the greatest versatility of any program to date, expanding beyond just polyketides and nonribosomal peptides to include such classes of compounds as β-lactams, lantibiotics, and siderophores among others [98]. The identification and analysis of such diverse secondary metabolism genes will surely be of great benefit to the synthetic biologist for the design and engineering of new pathways. Of

course, it must be stated that despite their increasing utility, *in silico* tools will never completely replace laboratory experiments for the analysis and understanding of secondary metabolite genes. Nevertheless, such tools clearly provide valuable support to natural products research.

1.2.2.2 Manipulation of native strains

The modern bioinformatic tools described above have made trivial the transformation of genomic sequence data to a set of putative natural product gene clusters. The rate limiting step, then, lies in drawing a connection between a putative gene cluster and its corresponding natural product. Unfortunately, direct *in vitro* demonstration of catalytic function for every step in a putative pathway, though possible [99], is inherently difficult. Identifying suitable hosts and conditions for the expression and purification of every enzyme in a pathway is a non-trivial problem, even with the most advanced expression optimization algorithms. Further, many gene clusters of interest are very large and/or include complex, multi-domain megasynthases that are particularly difficult to express and purify in a non-native host. Coupling these facts with insufficient means for predicting substrate and cofactor identity *a priori*, *in vitro* characterization is rendered an impractical first-pass approach for identifying the product of an uncharacterized gene cluster. As a result, genetic manipulation of the native strain to elicit expression of a target pathway is a much more viable approach.

Any target pathway with unknown product(s) identified via genome mining falls into one of two categories: 1) the product has already been discovered, but the link between the cluster and the product has not been established; or 2) the product has not yet been discovered. For any pathway falling into the first category and some falling into the second, deletion of key genes coupled

with metabolomic profiling can be a successful approach to pair pathways with products (Figure 1.3). For example, PKS deletion in *S. coelicolor* led to the discovery of germicidins produced by this strain [100]. In fungi, systematic deletion of all PKS genes was performed in *Aspergillus nidulans*, identifying the gene responsible for synthesis of the previously identified austinol family of meroterpenoids [101]. Similarly, single deletion of several NRPS genes in *A. nidulans* was performed, leading to the discovery of novel emericellamide products [102]. In the latter example, several of the NRPS deletion mutants did not exhibit any differences in their metabolite profiles, highlighting the limits of this approach.

Despite some successes in deletion-based pathway characterization, the fact remains that this method is ineffective for natural product pathways that are functionally "silent"– i.e., those that yield no product under laboratory cultivation conditions, or those whose products are generated at levels below the practical limit of detection. In these cases, activation of the gene cluster of interest is necessary to elucidate the corresponding product. Conceptually, the simplest way to accomplish this goal is perhaps to locate and over-express a pathway-specific regulatory protein, which could activate all of the genes in the pathway of interest (Figure 1.3). This has been demonstrated very successfully in *A. nidulans*, where cloning of pathway-specific regulators under the inducible alcohol dehydrogenase *alcA* promoter has led to the discovery of the polyketide asperfuranone [103,104] and the aspyridones [105]. Similarly, expression of a pathway-specific regulator under the *ermE** promoter activated production of a family of glycosylated macrolides in *Streptomyces ambofaciens* [106]. Notably, this approach necessitates only supplementation of the regulatory gene *in trans*, and thus is generally applicable to any strain for which rudimentary genetic tools are available.

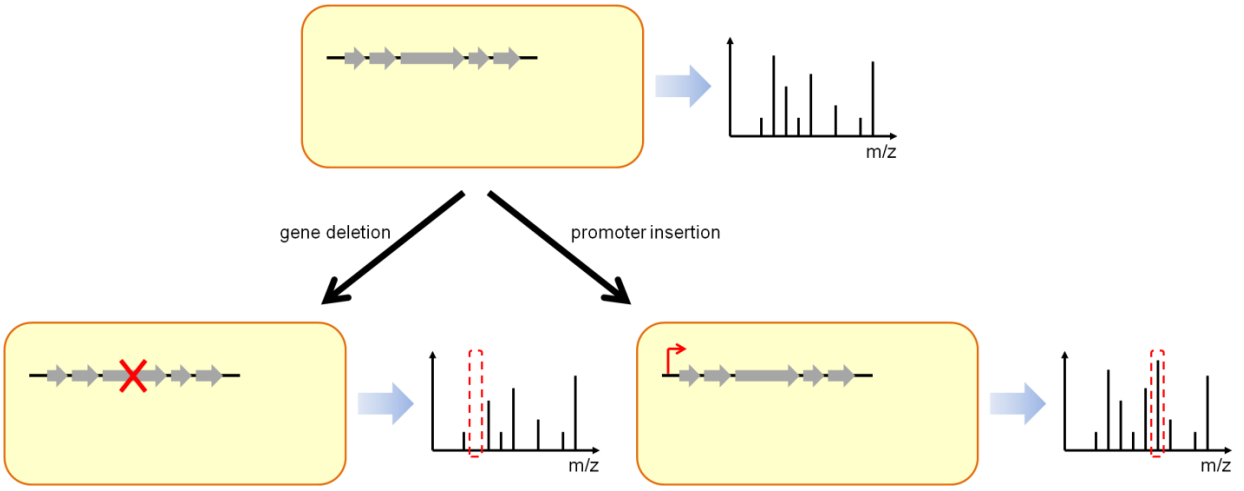


Figure 1.3: Coupled with gene deletion (left) or promoter insertion (right), metabolomic profiling enables identification of the product of a target gene cluster.

Besides regulators, strong promoters can also be introduced in front of pathway genes of interest for a more direct activation strategy. For example, knock-in of the *ermE** promoter in front of key genes in *Streptomyces albus* activated production of the blue pigment indigoidine as well as the two novel polycyclic tetramic acid macrolactams [107]. In *Streptomyces avermitilis*, cloning of a terpene synthase under an endogenous strong promoter and re-introduction to a genome-minimized strain led to discovery of a novel sesquiterpene, avermitilol [108].

1.2.2.3 Heterologous expression

While the methods described above have enabled discovery of many novel natural products, they all entail attempts to mitigate, manipulate, or override native regulatory mechanisms, which at best behave like a black box (defined input/unpredictable output), and at worst like a black hole (defined input/no output). An alternative approach is to remove the pathway from its native context completely via heterologous expression, and thus liberate it from intrinsic regulation [109-111]. Heterologous expression has been demonstrated for many native natural product

gene clusters [112-115] as a means of confirming their boundaries and necessary elements. However, by and large a wild-type cryptic natural product gene cluster will remain cryptic in a heterologous host either via similarly strict regulation (if the heterologous host is closely related to the native strain) or an inability to recognize the native promoters (if the heterologous host and the native strain are more distant).

To rectify a cryptic natural product gene cluster with its heterologous host, replacement of native promoters with well-characterized constitutive or inducible promoters is a common approach. On the single gene level, for example, expression of a cryptic terpene cyclase from *Arabidopsis thaliana* in *S. cerevisiae* resulted in discovery of a novel triterpenoid, thalianol [116]. Similarly, a silent myxobacterial type III PKS was found to produce flaviolin when expressed in various *Pseudomonas* hosts [117]. This approach has also been extended to multi-gene clusters. For example, two natural product gene clusters from *Photorhabdus luminescens* were cloned under the tetracycline-inducible *tetO* promoter in *E. coli* to enable discovery of luminmycin and luminmide natural products [118]. Similarly, the 56 kb epothilone cluster from *Sorangium cellulosum* was cloned under the *act*I promoter to achieve heterologous expression in *S. coelicolor* [119]. Notably, the latter example was not the purpose of discovery, but to transfer the cluster to a more suitable host for fermentation. Pushing this approach to its limit, replacement of every promoter in a cryptic five-gene polycyclic tetramic acid macrolactam pathway from *Streptomyces griseus* led to the discovery of multiple novel compounds when expressed in *Streptomyces lividans* [120].

Reconstruction of pathways for heterologous expression also enables the beneficial manipulation of native regulatory elements. For example, heterologous expression of the thiocoraline pathway from marine *Micromonospora* strains in *S. lividans* or *S. albus* was only possible after replacing the native promoter of the activator *tioA* with the strong *ermE* promoter [121]. On the other end of the spectrum, discovery of the daptomycin-like lipopeptide taromycin via heterologous expression in *S. coelicolor* required deletion of pathway repressor *tar20* [122], while activation of a silent spectinabilin cluster from *Streptomyces orinoci* required deletion of repressor *norD* [123]. Here, replacement of all promoters in the pathway further improved heterologous production levels.

Heterologous expression in a well-studied host further enables metabolic fine-tuning to optimize product titer by increasing flux into the pathway of interest and eliminating competing reactions. In *E. coli*, engineering of precursor pathways enabled expression of the polyketide 6-deoxyerythronolide B [124]. *Streptomyces* strains have also been significantly engineered for heterologous expression. In *S. coelicolor*, native antibiotic clusters have been deleted or inactivated to create a host strain with a cleaner metabolite background [125]. With additional activating mutations in *rpoB* and *rpsL*, such a strain has been shown to improve titers of heterologous natural products [126]. In *S. avermitilis*, a large-scale deletion approach was employed to remove over 1.4 Mb of the chromosome, eliminating endogenous natural product clusters [127]. The resulting host was shown to produce higher titers of some heterologous natural product pathways [128,129]. As a result, such hosts should benefit natural product discovery efforts.

1.3 Enabling Technologies for Natural Product Engineering

Current and future efforts in natural product discovery and engineering will depend upon new technologies to enable sophisticated manipulation of natural product gene clusters, the proteins they encode, and the organisms that produce them. At the DNA level, modern assembly methods allow flexible, reliable, high-throughput construction of natural product gene clusters. At the protein level, direction evolution represents a powerful tool to engineer new pathways and new products. At the organism level, genome-scale engineering techniques enable engineering of native and heterologous strains to optimize natural product biosynthesis.

1.3.1 Modern DNA assembly methods

In the mid-1970s, cloning methods based on restriction digestion and ligation revolutionized biology by enabling unprecedented manipulation of the very building blocks that define biological systems [130]. While the contributions of such techniques to the field cannot be overestimated, the fact remains that for the manipulation of large numbers of DNA sequences of interest (i.e., multi-gene pathways or even entire chromosomes or genomes), traditional methods are tedious, time-consuming, and too sequence-specific to generalize. As a result, a lengthy series of sub-cloning steps is typically required *en route* to the desired pathway constructs, and even then assemblies of significant size and complexity can be prone to failure. In recent years, however, a number of revolutionary DNA manipulation techniques have been developed, transforming arduous constructions into relatively routine tasks.

Modern DNA assembly techniques can broadly be classified into two groups: those based on homology and those based on digestion/ligation (Figure 1.4). Homology-based methods require

neighboring DNA fragments to share identical sequences, such that splicing can occur either by annealing and extension of the homologous ends *in vitro* or by homologous recombination *in vivo*. Perhaps the most prominent *in vitro* technique is the one-pot isothermal assembly pioneered by Gibson and coworkers, colloquially known as "Gibson assembly" [131]. In this process, DNA fragments with homologous termini are spliced via three enzymatic reactions. First, T5 exonuclease catalyzes "chew-back" (single strand degradation) of the 5' ends of each fragment. This exposes their complementary single-stranded 3' ends, which anneal to each other in the desired order to form the target construct with single-stranded gaps. Phusion polymerase then fills in the gaps, and *Taq* ligase seals the nicks to produce the intact final product, which can subsequently be used to transform a host of choice. A variety of previously developed *in vitro* assembly techniques present variations on this theme, including sequence and ligase independent cloning (SLIC), which utilizes T4 DNA polymerase for both 3' chew-back and partial gap-filling, but requires addition of a single deoxynucleotide to switch between the two functions [132]; polymerase incomplete primer extension (PIPE) cloning, which relies on incomplete primer extension during PCR of each fragment to leave single-stranded 3' ends [133]; and uracil-specific excision reagent (USER) cloning, which utilizes uracil-containing primers and a uracil-specific glycosylase and endonuclease to generate defined single-stranded 3' ends [134]. Note that besides Gibson assembly, none of the above methods employ a ligase enzyme, instead requiring nick-sealing to occur *in vivo* following transformation into the desired host. Further, both SLIC and PIPE cloning also require additional gap-filling *in vivo* to generate the nicked target construct.

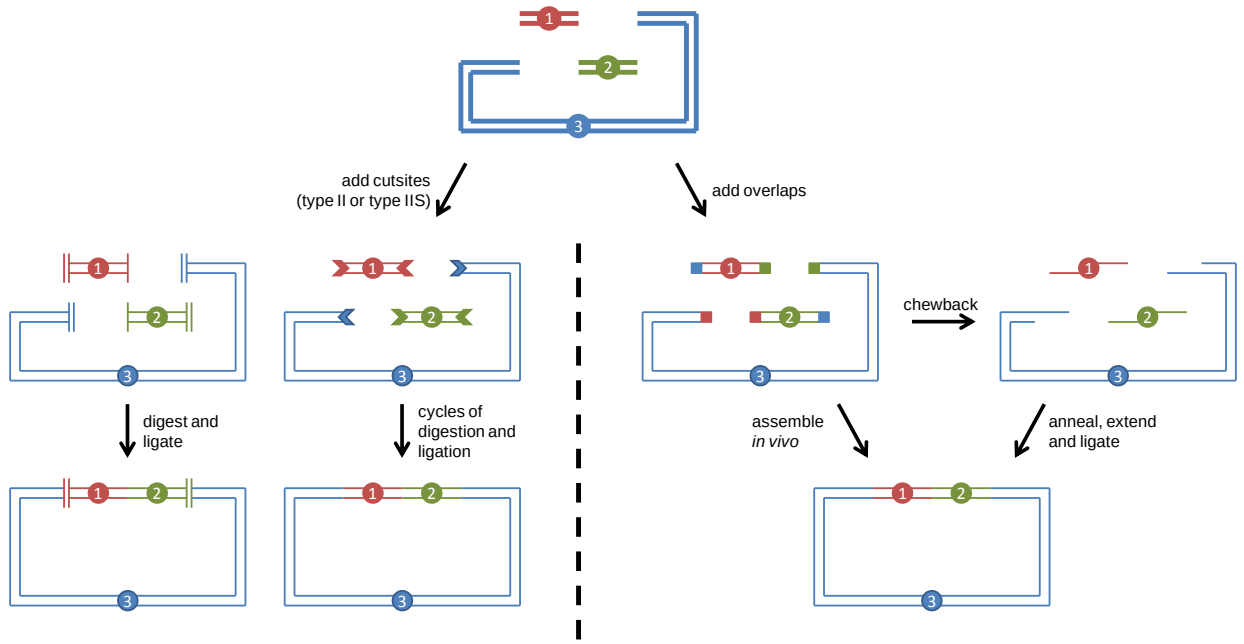


Figure 1.4: Representative schematics of a three-piece DNA assembly protocol via digestion/ligation-based (left) or homology-based (right) methods.

Published shortly after Gibson assembly, circular polymerase extension cloning (CPEC) presents an alternative to the "chew back and anneal" strategy. Starting from a set of DNA fragments with homologous ends, this method instead relies on cycles of heating to denature the duplex fragments, cooling to anneal neighboring strands at their overlapping ends, and polymerase-mediated extension to generate the concatenated duplex. After several cycles, the nicked target construct is formed, which can be sealed *in vivo* [135]. Site-specific recombination-based tandem assembly (SSRTA), on the other hand, employs the *Streptomyces* phage φBT1 integrase to splice neighboring fragments *in vitro* [136]. This method requires each fragment to be flanked by a set of orthogonal recombination sites, and consequently leaves interstitial scar sequences. While this is a clear disadvantage compared to the other techniques described, the absence of a polymerase extension step and the high specificity of the φBT1 integrase for its cognate

recognition sequences make this method attractive by avoiding the introduction of mutations or off-target recombination events.

An alternative to *in vitro* assembly is to allow fragment splicing by native cellular homologous recombination machinery. A key example of this is the DNA Assembler method, which relies on *Saccharomyces cerevisiae* to assemble DNA fragments with terminal overlap sequences [137]. In this approach, the assembly host is simultaneously transformed with individual fragments containing homologous ends. The target construct generated *in vivo* via homologous recombination can then be selected via an incorporated selection marker. Additional *in vivo* assembly methods include the Red recombination system, in which homologous recombination in *E. coli* is enhanced through expression of the Redαβ proteins from the lambda prophage or RecET proteins from the Rac prophage [138]; mating-assisted genetically integrated cloning (MAGIC), which employs bacterial conjugation to transfer a donor plasmid to the assembly host strain containing a receiver plasmid, a homing endonuclease (to generate linear fragments), and inducible lambda recombinases (to facilitate homologous recombination) [139]; transformation-associated recombination (TAR) cloning, which can be used to clone large portions of genomic DNA via simultaneous transformation of *S. cerevisiae* with genomic DNA containing the target sequence and a receiver vector [140]; and RecET-mediated "direct" cloning, which is similar to TAR cloning but carried out in *E. coli* with inducible expression of the RecET recombinases and requires prior digestion of the genomic DNA to liberate the target sequence as a linear fragment [118]. An *ex vivo* recombination-based method named for its seamless ligation cloning extract (SLiCE) has also been recently described, which utilizes *E. coli* extracts rather than whole cells to catalyze fragment assembly [141].

Although numerous powerful homology-based assembly methods have been developed, there are still non-trivial limitations to their general utility. Among these is the necessity to avoid multiple fragments with similar homologous ends in the assembly design, as this can lead to incorrect pairing of fragments not intended to be neighbors. Such concerns become significant in clusters containing repeated similar elements, such as the domains of a modular PKS or NRPS or the exogenous promoters and terminators used in pathway refactoring. Thus, there still exists a need for assembly techniques that do not rely on homologous recombination. Of course, the classic restriction digestion/ligation method is one such technique which generates only short single-stranded overhangs at specific sites. As noted above, however, this method has limited applicability for rapid combinatorial assembly. To facilitate and streamline its application, the concept of BioBricks (and subsequent variants, including BglBricks) has been proposed [142]. BioBrick assembly can be seen as the standardization of traditional cloning techniques. The BioBrick assembly standard dictates the restriction enzyme recognition sequences that should be positioned at the 5' and 3' ends of the assembly fragments. Utilization of two restriction enzymes with different recognition sequences but identical single-stranded overhangs (e.g., *Xba*I and *Spe*I) renders the assembly of two fragments an idempotent operation. In other words, correct ligation of two fragments abolishes the recognition sites between them while retaining those at the 5' and 3' termini of the product fragment. Thus, the product fragment can be employed in subsequent assemblies under the same standardized conditions, eliminating the need to identify new restriction enzymes for each fragment in the target construct. Nevertheless, this restriction enzyme-recycling approach necessitates a stepwise rather than simultaneous assembly scheme as only two fragments can be joined per round of assembly. As a result, construction of large secondary metabolite gene clusters by this approach can still be time-consuming.

To reconcile the assembly of several fragments with a convenient enzyme-recycling methodology, a restriction enzyme that can recognize only a single defined sequence but generate many different single-stranded overhangs is needed. Fortuitously, both of these properties are manifested in Type IIS restriction endonucleases, which can bind only to a specific recognition site but cut indiscriminately at a prescribed distance from this site. Thus, by incorporating Type IIS restriction sites at the termini of each fragment, user-defined overhangs can be generated such that simultaneous assembly of multiple fragments in the desired configuration can be achieved. This technique, initially proposed by Engler and coworkers in 2008, is termed Golden Gate assembly [143].

Note that modern digestion/ligation-based cloning techniques still carry with them a major limitation of traditional cloning; namely, the necessity to remove all DNA recognition sites of the selected restriction endonuclease within the fragments to be assembled. To obviate this requirement, Chen and coworkers recently presented a method to limit restriction endonuclease digestion only to the desired terminal sequences [144]. Their method, methylation-assisted tailorable ends rational (MASTER) ligation, expands the utility of Golden Gate assembly via utilization of *Msp*JI, a Type IIS restriction endonuclease containing 5-methylcytosine in its recognition sequence. Through incorporation of 5-methylcytosine in the primers used to amplify each fragment for assembly, digestion only occurs at the desired terminal locations and not within the fragments where only unmodified cytosines are present.

Along with the development of more advanced DNA manipulation techniques, more advanced *in silico* tools for the design of assembly schemes have been created. When given a list of parts to

assemble, for example, the j5 program of Hillson and colleagues can output a set of PCR primers to amplify each of the parts with suitable overhangs (i.e., restriction sites or homologous overhangs) for any of a variety of popular assembly methods [145]. This program can also recommend direct synthesis or PCR amplification of parts based on cost analysis, and will suggest a hierarchical assembly scheme in cases where there are apparent barriers to one-step assembly (e.g., repeated use of a single part). Using a suitable language such as PR-PR [146,147], assembly protocols can then be sent to a liquid handling device for automation purposes. Moving a step further, the Raven software suite applies multiple design algorithms to optimize assembly strategies for human or robotic implementation [148]. Notably, it can also learn from past assembly results to improve future predictions.

1.3.2 Directed evolution

Directed evolution is a versatile laboratory process for the creation of biological entities with desired properties. Though the term had been occasionally applied in past decades to describe adaptive evolution experiments, directed evolution in the modern sense began to take root in the 1990s. In broad terms, directed evolution can be defined as an iterative two-step process involving first the generation of a library of variants of a biological entity of interest, and second the screening of this library in a high-throughput fashion to identify those mutants that exhibit better properties, such as higher activity or selectivity. The best mutants from each round then serve as the templates for the subsequent rounds of diversification and selection, and the process is repeated until the desired level of improvement is attained (Figure 1.5). As compared to rational design, which had been pioneered several years prior [149], directed evolution on the protein level provides the distinct advantage of requiring no knowledge of the protein structure

or of the effects of specific amino acid substitutions, which are very difficult to predict *a priori* [150].

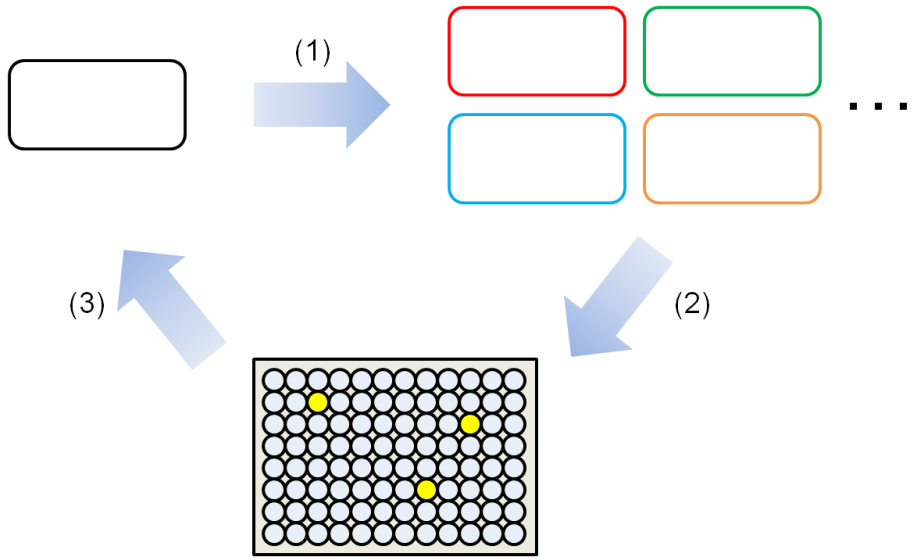


Figure 1.5: Direct evolution involves iterative cycles of diversity generation (1) and screening or selection (2), with the top hits of the previous round serving as the basis of the next round (3).

As modular biopolymers build from twenty canonical building blocks, proteins are highly evolvable through mutations in their amino acid sequences, which are directly encoded via their cognate genes. Because of this, proteins have by far been the most common targets for directed evolution experiments, where a host of techniques have been developed to manipulate individual gene sequences for the creation of diverse protein libraries. From the reductionist perspective that has dominated much of the biological sciences in the modern era, proteins provide defined, discrete elements that can readily be isolated and studied under highly controlled conditions, making them ideal experimental subjects. In the context of natural products research, the evolution of pathway enzymes can serve not only as a way to de-bottleneck rate-limiting steps

for improved overall productivity, but also as a means to generate "unnatural" derivatives by manipulating substrate specificity.

1.3.2.1 Random evolution strategies

Given the aforementioned properties of proteins, it is no wonder that the earliest modern directed evolution experiments focused almost exclusively on protein targets. As the *a priori* prediction of the effect of mutations on a given protein is often difficult, the earliest techniques focused simply on random mutagenesis. A landmark example in this field is the evolution of subtilisin E, a serine protease useful in several industrial applications, for increased activity in dimethylformamide [151]. In this pioneering study, random mutations were introduced to the subtilisin E gene using an error-prone PCR amplification strategy. After three sequential rounds of mutagenesis and screening, a mutant was identified with six additional point mutations that exhibited 256-fold higher activity in 60 % dimethylformamide. This effort clearly demonstrated the power of a sequential, evolutionary protein engineering strategy to identify multiple cooperative mutations for vast protein improvement. In contrast, previous protein engineering efforts typically employed parallel single rounds of selection, identifying individual mutations that were by no means guaranteed to combine beneficially.

While error-prone mutagenesis is an effective means to introduce gradual changes to a protein, in natural evolution this is supplemented by an additional mechanism: recombination. Recombination allows larger pieces of similar genes to be exchanged, significantly diversifying the resultant pool of variants. Thus, to harness this power for directed evolution experiments, techniques were developed to mimic natural recombination *in vitro*. A key example of this is the

DNA shuffling method developed in 1994 by Willem Stemmer of the Affymax Research Institute. In this method, a library of similar genes (e.g., homologs from different organisms or mutants identified from previous engineering efforts) is PCR amplified and digested into smaller fragments with DNase I. Following subsequent isolation of these fragments, a primer-free PCR-like assembly step recombines the fragments from diverse parent genes into new chimeric hybrids, which can then be cloned into an expression vector and screened [152]. As an example of the power of this approach, a β-lactamase was evolved to improve the resistance of its host *Escherichia coli* strain to the antibiotic cefotaxime [153]. After three cycles of shuffling and two cycles of backcrossing (to remove non-essential mutations), a mutant was identified that increased the minimum inhibitory concentration (MIC) of the host by 32,000-fold, compared to the 16-fold increase observed when non-recombinogenic methods were employed.

In subsequent years, numerous refined recombination-based directed evolution strategies were developed [154]. One such example is the staggered extension process (StEP) for *in vitro* recombination [155]. Similar to DNA shuffling, this approach generates chimeric progenies from a set of parent genes. However, in StEP, the full-length recombined genes are synthesized in the presence of the parent genes without the intermediate step of generating and purifying short fragments. This is accomplished by sequential annealing of the nascent polynucleotide to different templates with abbreviated extension times, allowing only a small portion of the gene to be filled in before dissociating and annealing to a new template. Using a StEP-based approach, subtilisin E was evolved to exhibit thermostability equal to that of thermitase, a thermophilic homolog from the extremophile *Thermoactinomyces vulgaris* [156].

While error-prone mutagenesis and recombination approaches both seek to develop improved protein functionalities via random variation of existing scaffolds, a more extreme approach to random evolution of proteins is to start from completely random sequences. This tactic has been employed by Keefe and Szostak to identify novel ATP-binding proteins from a library of completely random 80-mer polypeptides [157]. From a library of 6×10^{12} proteins, each covalently bound to its cognate mRNA, they employed 18 rounds of selection on ATP-agarose to identify four with ATP-binding functionality. Of course, such a system is dependent upon both the ability to synthesize very large libraries and, most importantly, screen them in a high-throughput manner, which may be difficult for certain target properties. Nevertheless, this work serves as an impressive example of the power of a random evolutionary approach.

1.3.2.2 Targeted evolution strategies

While random evolution strategies have a proven record of success, the solution space they explore for any given protein is large – too large, in fact, for many screening methods to feasibly allow sufficient library coverage. However, it is apparent that with any random mutagenesis strategy, the vast majority of the mutants generated will exhibit no improvement or even inferior performance relative to the parent protein. As a result, a number of techniques have been developed to leverage the substantial amount of protein structural data available for the design of smaller, targeted libraries enriched in variants most likely to exhibit improved properties. Two examples of this are structure-based combinatorial protein engineering (SCOPE) and the SCHEMA algorithm, both presented in 2002 [158,159]. These strategies employ protein structural data to first identify discrete units of protein secondary structure. By biasing a recombination-based evolution strategy such that recombination occurs only between these units

and not within them, the chances of the resulting chimeric proteins maintaining a correct folding pattern, and thus functionality, increase. Additionally, SCHEMA and SCOPE carry the benefit of enabling shuffling of parent proteins with low sequence identity, provided that they share similar folds. Besides these techniques, a number of such methods have been developed [160,161].

While SCHEMA and SCOPE focus on structural elements of a target protein, other approaches focus instead on the functional elements. To evolve a nuclear hormone receptor to bind a new ligand, for example, Chockalingam and coworkers employed stepwise saturation mutagenesis of active site residues to target only those that are expected to play a role in contacting the ligand [162]. Similarly, Reetz and coworkers utilized an iterative saturation mutagenesis approach to increase the thermostability of a lipase from *Bacillus subtilis* by targeting those residues that showcased the highest degrees of thermal motion based on X-ray data [163]. In multi-domain proteins, each domain can be independently evolved in the context of the holoenzyme. As an example of this approach, each domain of cytochrome $P450_{BM3}$ from *Bacillus megaterium* was separately evolved using a combination of random, saturation, and site-directed mutagenesis. When the beneficial mutations in each domain were combined, the resulting protein was able to hydroxylate propane, a nonnative substrate, with native-like coupling efficiency of cofactor utilization [164].

When structural information for a particular target protein is unavailable, computational modeling can be employed to guide directed evolution experiments. A particularly impressive example of this is the evolution of a transaminase for the industrial synthesis of the antidiabetic

drug sitagliptin by researchers at Codexis and Merck [165]. Motivated by the desire to replace a rhodium-catalyzed enamine hydrogenation (necessitating high pressure and extra purification steps) with a more efficient enzymatic process, they first sought a transaminase with activity towards the prositagliptin ketone. While a particular enzyme with the desired stereospecificity was identified, it had no activity toward the desired substrate. As a result, a homology model of the transaminase was built *in silico* to guide reconstruction of the active site. Applying multiple rounds of targeted mutagenesis, detectable activity toward the prositagliptin ketone was detected. Subsequently, the enzyme was evolved to function under process conditions, yielding a practical biocatalyst for the industrial process.

Beyond evolving a given protein scaffold for activity with a new substrate, multiple directed evolution efforts in recent years have set the ambitious goal of evolving the scaffold for completely novel activities. In 2006, Park and coworkers presented a method to do so called simultaneous incorporation and adjustment of functional elements (SIAFE) [166]. In this approach, functional elements (including active site loops involved in catalysis and in substrate binding) are systematically and combinatorially incorporated into a chosen template sequence. As proof of concept, they evolved β-lactamase activity from an αβ/βα metallohydrolase scaffold through deletion, insertion and remodeling of active site functional elements. At the end of the process, although the kinetic properties of the new β-lactamase protein were inferior to those of natural β-lactamases, it no longer exhibited its former metallohydrolase activity. For this experiment, the choice of functional elements was guided by the structural and mechanistic data available for known β-lactamases of the αβ/βα superfamily; it would likely be significantly more difficult, therefore, to evolve a completely novel function using this approach. Nevertheless, this

has been achieved by Röthlisberger and coworkers to create an enzyme capable of catalyzing the Kemp elimination reaction, for which no known natural enzymes exist [167]. To accomplish this goal, a computational design strategy was employed, coupled with directed evolution. First, an ideal active site for the desired Kemp elimination enzyme was designed *in silico* based on quantum mechanical transition state calculations. Next, further computational analysis was used to identify the protein scaffolds that could best support the designed active site. Eventually, 59 different designs were experimentally characterized, of which eight exhibited the desired function at a detectable level. After seven rounds of directed evolution, comprising both random mutagenesis and shuffling, an increase of greater than 200-fold was observed in k_{cat}/K_m for the best mutant, yielding an overall rate enhancement of 1.18×10^6 relative to the uncatalyzed reaction. Another example of coupled *in silico* design and directed evolution was recently provided by Karanicolas and coworkers, who engineered a heterodimerization interface between two unrelated proteins [168]. While the *in silico* design alone yielded a pair with a measured K_d of 130 nM, subsequent directed evolution decreased this value almost 1000-fold to 180 pM. These two examples clearly illustrate the synergistic relationship between computational tools and directed evolution, a trend that is likely to continue to develop in the years to come.

1.3.3 Genome engineering

1.3.3.1 Genome evolution

Natural product biosynthesis is most often thought of in terms of either biosynthetic pathways or their constituent enzymes, as these elements directly participate in the conversion of substrates to products. Nevertheless, it is important to remember that natural product biosynthesis occurs in the context of the cellular milieu, and thus cannot be truly decoupled from the rest of cellular

metabolism. As a result, techniques that focus on engineering of microorganisms at the genome scale can be particularly useful for natural products research to create ideal production strains.

Classic strain development processes involve the random introduction of mutations to an organism in a stressed environment, relying either on intrinsic mutation rate or supplementation of mutagenizing chemicals or radiation to generate an improvement genotype (and thus phenotype) [169]. Modern processes employ new methods to exploit the diversity generated or define the specific changes introduced. One such technique is genome shuffling, a process that takes advantage of whole-genome recombination between members of a diverse population to create complex, multi-parent progeny with significantly improved fitness in every successive generation [170]. This technique has been applied to increase production of (2*S*,3*R*)-hydroxycitric acid (HCA) in the native producer *Streptomyces* sp. U121, the biosynthetic pathway of which is not well characterized [171]. Typically, genome shuffling experiments proceed in two major steps: creation/identification of the parent strains to be shuffled, and subsequent recombination (via protoplast fusion) and selection to identify the desired progeny. In this study, the parent strains were created through random mutagenesis using nitrosoguanidine (NTG). From this initial round of mutagenesis, four strains were identified with increased HCA production and used for genome shuffling. Multiple rounds of shuffling led to a five- to six-fold improvement in product titer.

Genome shuffling can be applied in a variety of different hosts. For example, application in the strict anaerobe *Clostridium diolis* DSM 15410 led to improved 1,3-propanediol (1,3-PD) production [172]. In this study, NTG-based chemical mutagenesis was employed to generate

two sets of parental strains: those improved in glycerol (substrate) tolerance, and those improved in 1,3-PD (product) tolerance. Four rounds of genome shuffling in the presence of 1,3-PD and the side products acetic and butyric acid were carried out, yielding a strain with 80% greater volumetric production of 1,3-PD. Further application of genome shuffling in *Propionibacterium shermanii* and *Penicillium decumbens* JU-A10 have been explored as well, resulting in mutants with increased production of vitamin B12 and cellulase, respectively [173,174].

As a complement to genome engineering approaches based on random mutagenesis, the Church group developed a method for large-scale, parallel modification of a genome at many targeted sites [175]. Termed multiplex automated genome engineering, or MAGE, this technique makes use of the bacteriophage λ-Red ssDNA-binding protein β to achieve allelic replacement of desired gene loci with synthetic oligos. One cycle of MAGE includes transformation of cells with the synthetic oligos, recovery and growth to mid-log phase, and heat shock to induce β protein expression and recombination. With successive cycles, increasing variation can be introduced into the host genome (Figure 1.6). Depending on the oligo design, this technique can be used to introduce mutations, insertions, or deletions, all in a simultaneous fashion. As proof of concept, 24 different genes were simultaneously targeted to induce a lycopene overproduction phenotype in *E. coli*. Twenty of these genes had been reported to increase lycopene production, and so oligos with degenerate RBS sequences were designed to enhance translation efficiency. The remaining four genes came from alternate pathways and were targeted for inactivation via oligos containing nonsense mutations. After 5-35 cycles of MAGE, up to five-fold greater lycopene production was observed when compared to the ancestral strain, yielding titers of ~9000 ppm.

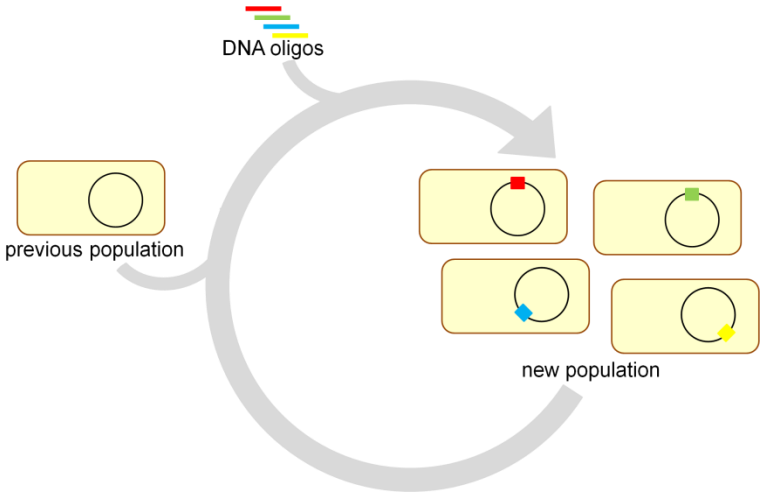


Figure 1.6: MAGE consists of iterative cycles of oligo-mediated mutagenesis.

The decreasing cost of DNA oligomer synthesis has been an important factor driving genome engineering efforts. Microarray technology has enabled parallel synthesis of oligos to target every gene in the genome of *E. coli*, for example, facilitating rapid and comprehensive library assembly. When coupled with the MAGE process, microarray technology enabled construction of a genome-wide T7 promoter insertion library [176]. The trackable multiplex recombineering (TRMR) method similarly utilizes microarray-derived oligos to target the full genome of *E. coli*, and has been applied to introduce both up- and down-regulation [177].

While such techniques are primarily limited to *E. coli*, expansion to other hosts could greatly accelerate industrial strain development. For example, the MAGE approach was recently expanded to *S. cerevisiae* strains through a similar process called yeast oligo-mediated genome engineering, or YOGE [178]. Additionally, to evolve complex phenotypes in *S. cerevisiae*, genome-wide reduction-of-function libraries can be created by leveraging a heterologous RNA-interference (RNAi) system [179]. This RNAi-assisted genome evolution method, known as RAGE, entails iterative cycles of RNAi-mediated gene knockdown with a genome-wide library

and screening or selection for a desired phenotype. Beneficial knockdown cassettes can then be integrated to the genome to generate improved basis strains for the next round of screening or selection. This approach has been successfully applied to increase *S. cerevisiae* tolerance to fermentation inhibitors acetic acid [179] and furfural [180], underscoring its potential as an industrial strain development platform.

1.3.3.2 Genome tailoring

As opposed to the broad-scale evolutionary approaches describe above, precise and rapid manipulation of specific loci in a genome of interest can also be highly beneficial from a strain engineering perspective. Such manipulations can include gene deletions, for example to remove competing pathways; site-specific mutagenesis, to introduce beneficial mutations *in situ*; or gene replacement, to diversify pathways or evaluate homologs. While recombination between the intact genome and a supplied DNA "editing template" can be harnessed to make such changes in a variety of organisms, the fact remains that even in the most proficient organisms, recombination rates are quite low. As a result, laborious screening for the desired mutation is often required.

In the cell, DNA damage is known to activate repair machinery. As a result, introduction of a double-strand break (DSB) at a genomic locus of interest can be used stimulate the homologous recombination machinery, facilitating exchange between the chromosome and the editing template. Site-specific introduction of a DSB, however, requires a highly-specific nuclease that will not cleave similar sites present at other locations in the genome. To meet these demands, researchers have looked to exploit native DNA-binding proteins. The most obvious first choices

in this area were homing endonucleases, restriction enzymes with large recognition sequences [181]. While some successes were achieved in the engineering of such enzymes [182,183], it became apparent that their recognition and cleavage domains did not exhibit the requisite modularity for facile recognition site alteration.

Moving forward, greater modularity was achieved by creating artificial endonucleases with programmable DNA binding domains fused to non-specific nuclease domains [181]. The first wave of these chimeric endonucleases utilized zinc finger (ZF) DNA recognition domains, each of which can recognize a specific 3 nucleotide sequence [184]. By combining multiple ZF domains in a single polypeptide, longer recognition sequences could be built. Further, fusion of the ZF domains to a nuclease domain that cleaves only upon dimerization, such as FokI, enables creation of ZF nucleases (ZFNs) that introduce a DSB only when two ZFNs are present, increasing specificity and decreasing off-target effects. However, over time it was realized that design of functional ZFNs can be difficult, as a *de novo* designed ZFN pairs often fail to perform as expected [185].

The second wave of chimeric endonucleases replaced the ZF domains with transcription activator-like effector (TALE) DNA binding domains. TALE domains recognize nucleic acids on a 1:1 basis based on a two amino acid specificity code [186], offering greater overall modularity to facilitate customization. As a result, TALE nucleases (TALENs) have been applied for genome editing purposes in a number of different hosts [187]. Notably, the highly repetitive architecture of a TALE binding domain complicates their assembly by traditional

cloning methods, though modern cloning techniques can circumvent these difficulties given an appropriate library of building blocks [188,189].

Though TALENs offer great versatility in single-locus genome editing, multi-locus targeting can become cumbersome as a new pair of TALEN genes must be synthesized and introduced to the desired host for every site of interest. Recently, an alternative approach based on the clustered regularly interspaced repeats (CRISPR)/CRISPR-associated (Cas) proteins of *Streptococcus pyogenes* has been pioneered [190]. In the native host, the CRISPR/Cas system serves as an immune system against foreign nucleic acids. Short fragments of the invading DNA are integrated into the genome of the host in CRISPR array, which can then be transcribed to facilitate cleavage of the foreign DNA by Cas9, an RNA-guided endonuclease, upon subsequent invasion [191]. For genome editing purposes, Cas9 and a guide RNA sequence can be heterologously expressed in any host of interest to induce site-specific cleavage. As DNA recognition is determined only by the short RNA sequence, adapting Cas9 to any desired locus simply requires the expression of a new guide sequence. As a result, this system offers unprecedented flexibility for multiplexing [192,193] and library creation [194].

1.4 Project Overview

Now more than ever, the raw materials for natural product pathway assembly (i.e., genes from sequenced genomes) are plentiful. Couple this with an abundance of new techniques for DNA manipulation and the field appears poised for a new age of natural product discovery. Gene clusters can be constructed, de-constructed, rearranged, repurposed, and repackaged with unprecedented ease, opening the door to rapid, facile platforms for natural product discovery,

characterization, and engineering. Throughout my graduate career, I have sought to leverage these new technologies for the discovery and engineering of natural products at the protein, pathway, and genome scales.

Chapter 2 describes the synthesis of new analogs of the natural product FR-900098 by discovering and exploiting the promiscuity of the native biosynthetic machinery. Previous studies have shown that FR-900098, a phosphonic acid produced by *Streptomyces rubellomurinus*, can serve as a potent anti-malarial by inhibiting the 1-deoxy-D-xylulose-5-phosphate reductoisomerase (Dxr) enzyme of malaria parasite *Plasmodium falciparum*. Here we characterized the *N*-acetyltransferase FrbF from the FR-900098 biosynthetic pathway, determining its catalytic mechanism and scope of acyl-coenzyme A substrates. Guided by our co-crystal structure of Dxr with phosphonic acid inhibitors, we enzymatically synthesized a novel *N*-propionyl derivative, FR-900098P, and demonstrated its superior Dxr inhibition *in vitro*. Further, we developed an *in vivo* platform for FR-900098P biosynthesis by leveraging the previously established *E. coli* expression system with metabolic engineering and mutasynthetic approaches.

Chapter 3 outlines efforts towards the discovery of fungal polyketides via heterologous expression in *S. cerevisiae*. The pathogenic, thermal dimorphic fungus *Talaromyces marneffei* was selected as a source of uncharacterized PKS genes, as a total of 25 were identified in its sequenced genome. Yeast homologous recombination was demonstrated as a means to assemble entire pathways from individual gene expression cassettes, leading to the discovery of a resorcylic acid lactone halogenase. Additionally, a helper plasmid was designed for expression,

detection, and purification of PKSs from yeast culture. Following one-step assembly of 11 PKS expression cassettes, soluble protein expression was detected by Western blot. We then characterized novel strong endogenous promoters to improve protein expression in *S. cerevisiae* and demonstrated their superior performance in the expression of a PKS responsible for production of fungal pigments YWA1 and citreoisocoumarin.

Chapter 4 details the discovery of novel phosphonic acid compounds from a cryptic gene cluster discovered in a *Streptomyces* sp. soil isolate. Bioinformatic analysis showed that the cluster contained three genes related to the FR-900098 biosynthetic pathway along with several other genes that diverge from this pathway. Functionality of the early steps in the cryptic pathway was demonstrated via complementation of FR-900098 biosynthesis in *E. coli*, although no phosphonic acids were detected from the native strain. Utilizing a two-step assembly strategy based on yeast homologous recombination, each of thirteen genes in the pathway was placed under a strong promoter on a single plasmid for heterologous expression in *S. lividans*. As a result, several phosphonic acids were detected, and NMR-characterization revealed novel structures not previously reported in the literature. Additionally, by expressing a phosphoenolpyruvate phosphomutase alone, we shed new light on the native metabolism of phosphonates in *S. lividans*, revealing an endogenous pathway to form 2-aminoethyl phosphonate from phosphonopyruvate.

Chapter 5 presents a high-efficiency genome editing method for *Streptomyces* species using an engineered CRISPR/Cas system. In the past two years, CRISPR/Cas systems have been one of the hottest topics in biological research, with impressive results demonstrated in a great variety

of organisms. In *Streptomyces* species, a reconstituted CRISPR/Cas technology would enable manipulation of natural product gene clusters– such as site-specific mutation, gene deletion, promoter knock-in, or full cluster deletion– with unprecedented ease. To accomplish this goal, a versatile base plasmid was designed and assembled with a codon-optimized *cas9* gene and the necessary RNA components under the control of strong actinomycete promoters. By targeting different loci in the genome of *S. lividans* 66, we were able to show high efficiency, multiplexed introduction of defined deletions ranging from 20 bp to 31 kb. Further, we demonstrated high efficiency genome editing in two additional species, *S. albus* J1074 and *Streptomyces viridochromogenes* DSM40736, using the same base plasmid. Thus, our designed system represents a powerful new tool for the manipulation of natural product gene clusters in a broad range of *Streptomyces* strains.

1.5 References

1. McGovern, P.E., Zhang, J., Tang, J., Zhang, Z., Hall, G.R., Moreau, R.A., Nunez, A., Butrym, E.D., Richards, M.P., Wang, C.S. *et al.* (2004) Fermented beverages of pre- and proto-historic China. *Proc Natl Acad Sci U S A*, **101**, 17593-17598.
2. Fleming, A. (2001) On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of *B. influenzae*. 1929. *Bull World Health Organ*, **79**, 780-790.
3. *The Dictionary of Natural Products*. (2014) Chapman & Hall / CRC, http://dnp.chemnetbase.com/.
4. Newman, D.J. and Cragg, G.M. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod*, **75**, 311-335.

5. Li, J.W. and Vederas, J.C. (2009) Drug discovery and natural products: end of an era or an endless frontier? *Science*, **325**, 161-165.

6. Chiang, Y.M., Chang, S.L., Oakley, B.R. and Wang, C.C. (2011) Recent advances in awakening silent biosynthetic gene clusters and linking orphan clusters to natural products in microorganisms. *Curr Opin Chem Biol*, **15**, 137-143.

7. Kaeberlein, T., Lewis, K. and Epstein, S.S. (2002) Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment. *Science*, **296**, 1127-1129.

8. (1958) Kanamycin stops Staph. *Chem Eng News*, **36**, 24.

9. Otten, S.L., Stutzman-Engwall, K.J. and Hutchinson, C.R. (1990) Cloning and expression of daunorubicin biosynthesis genes from *Streptomyces peucetius* and *S. peucetius subsp. caesius*. *J Bacteriol*, **172**, 3427-3434.

10. Ehrlich, J., Bartz, Q.R., Smith, R.M., Joslyn, D.A. and Burkholder, P.R. (1947) Chloromycetin, a new antibiotic from a soil actinomycete. *Science*, **106**, 417.

11. Gottlieb, D., Bhattacharyya, P.K., Anderson, H.W. and Carter, H.E. (1948) Some properties of an antibiotic obtained from a species of *Streptomyces*. *J Bacteriol*, **55**, 409-417.

12. Baltz, R. (2007) Antimicrobials from actinomycetes: back to the future. *Microbe*, **2**, 125-131.

13. Zhao, L.-X., Huang, S.-X., Tang, S.-K., Jiang, C.-L., Duan, Y., Beutler, J.A., Henrich, C.J., McMahon, J.B., Schmid, T., Blees, J.S. *et al.* (2011) Actinopolysporins A–C and tubercidin as a Pdcd4 stabilizer from the halophilic actinomycete *Actinopolyspora erythraea* YIM 90600. *J Nat Prod*, **74**, 1990-1995.

14. Stierle, A.A., Stierle, D.B. and Kelly, K. (2006) Berkelic acid, a novel spiroketal with selective anticancer activity from an acid mine waste fungal extremophile. *J Org Chem*, **71**, 5357-5360.

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

16. Montaser, R. and Luesch, H. (2011) Marine natural products: a new wave of drugs? *Future Med Chem*, **3**, 1475-1489.

17. Edwards, D.J., Marquez, B.L., Nogle, L.M., McPhail, K., Goeger, D.E., Roberts, M.A. and Gerwick, W.H. (2004) Structure and biosynthesis of the jamaicamides, new mixed polyketide-peptide neurotoxins from the marine cyanobacterium *Lyngbya majuscula*. *Chem Biol*, **11**, 817-833.

18. Jang, K.H., Nam, S.J., Locke, J.B., Kauffman, C.A., Beatty, D.S., Paul, L.A. and Fenical, W. (2013) Anthracimycin, a potent anthrax antibiotic from a marine-derived actinomycete. *Angew Chem Int Ed Engl*, **52**, 7822-7824.

19. Schmidt, E.W., Donia, M.S., McIntosh, J.A., Fricke, W.F. and Ravel, J. (2012) Origin and variation of tunicate secondary metabolites. *J Nat Prod*, **75**, 295-304.

20. Elshahawi, S.I., Trindade-Silva, A.E., Hanora, A., Han, A.W., Flores, M.S., Vizzoni, V., Schrago, C.G., Soares, C.A., Concepcion, G.P., Distel, D.L. *et al.* (2013) Boronated tartrolon antibiotic produced by symbiotic cellulose-degrading bacteria in shipworm gills. *Proc Natl Acad Sci U S A*, **110**, E295-304.

21. Ramadhar, T.R., Beemelmanns, C., Currie, C.R. and Clardy, J. (2014) Bacterial symbionts in agricultural systems provide a strategic source for antibiotic discovery. *J Antibiot (Tokyo)*, **67**, 53-58.

22. Blunt, J.W., Copp, B.R., Keyzers, R.A., Munro, M.H. and Prinsep, M.R. (2014) Marine natural products. *Nat Prod Rep*, **31**, 160-258.

23. Bode, H.B., Bethe, B., Hofs, R. and Zeeck, A. (2002) Big effects from small changes: possible ways to explore nature's chemical diversity. *Chembiochem*, **3**, 619-627.

24. Sarkar, A., Funk, A.N., Scherlach, K., Horn, F., Schroeckh, V., Chankhamjon, P., Westermann, M., Roth, M., Brakhage, A.A., Hertweck, C. *et al.* (2012) Differential expression of silent polyketide biosynthesis gene clusters in chemostat cultures of *Aspergillus nidulans*. *J Biotechnol*, **160**, 64-71.

25. Du, L., King, J.B., Morrow, B.H., Shen, J.K., Miller, A.N. and Cichewicz, R.H. (2012) Diarylcyclopentendione metabolite obtained from a *Preussia typharum* isolate procured using an unconventional cultivation approach. *J Nat Prod*, **75**, 1819-1823.

26. Theodore, C.M., King, J.B., You, J. and Cichewicz, R.H. (2012) Production of Cytotoxic Glidobactins/Luminmycins by Photorhabdus asymbiotica in Liquid Media and Live Crickets. *J Nat Prod*, **75**, 2007-2011.

27. Seyedsayamdost, M.R., Traxler, M.F., Clardy, J. and Kolter, R. (2012) Old meets new: using interspecies interactions to detect secondary metabolite production in actinomycetes. *Methods Enzymol*, **517**, 89-109.

28. Bader, J., Mast-Gerlach, E., Popovic, M.K., Bajpai, R. and Stahl, U. (2010) Relevance of microbial coculture fermentations in biotechnology. *J Appl Microbiol*, **109**, 371-387.

29. Oh, D.C., Kauffman, C.A., Jensen, P.R. and Fenical, W. (2007) Induced production of emericellamides A and B from the marine-derived fungus *Emericella* sp. in competing co-culture. *J Nat Prod*, **70**, 515-520.

30. Konig, C.C., Scherlach, K., Schroeckh, V., Horn, F., Nietzsche, S., Brakhage, A.A. and Hertweck, C. (2013) Bacterium induces cryptic meroterpenoid pathway in the pathogenic fungus *Aspergillus fumigatus*. *Chembiochem*, **14**, 938-942.

31. Kurosawa, K., Ghiviriga, I., Sambandan, T.G., Lessard, P.A., Barbara, J.E., Rha, C. and Sinskey, A.J. (2008) Rhodostreptomycins, antibiotics biosynthesized following horizontal gene transfer from *Streptomyces padanus* to *Rhodococcus fascians*. *J Am Chem Soc*, **130**, 1126-1127.

32. Hosaka, T., Ohnishi-Kameyama, M., Muramatsu, H., Murakami, K., Tsurumi, Y., Kodani, S., Yoshida, M., Fujie, A. and Ochi, K. (2009) Antibacterial discovery in actinomycetes strains with mutations in RNA polymerase or ribosomal protein S12. *Nat Biotechnol*, **27**, 462-464.

33. Tanaka, Y., Kasahara, K., Hirose, Y., Murakami, K., Kugimiya, R. and Ochi, K. (2013) Activation and products of the cryptic secondary metabolite biosynthetic gene clusters by rifampin resistance (*rpoB*) mutations in actinomycetes. *J Bacteriol*, **195**, 2959-2970.

34. Fu, P., Jamison, M., La, S. and MacMillan, J.B. (2014) Inducamides A-C, chlorinated alkaloids from an RNA polymerase mutant strain of *Streptomyces* sp. *Org Lett*, **16**, 5656-5659.

35. Tanaka, Y., Tokuyama, S. and Ochi, K. (2009) Activation of secondary metabolite-biosynthetic gene clusters by generating *rsmG* mutations in *Streptomyces griseus*. *J Antibiot (Tokyo)*, **62**, 669-673.

36. Inaoka, T., Takahashi, K., Yada, H., Yoshida, M. and Ochi, K. (2004) RNA polymerase mutation activates the production of a dormant antibiotic 3,3'-neotrehalosadiamine via an autoinduction mechanism in *Bacillus subtilis*. *J Biol Chem*, **279**, 3885-3892.

37. Gao, C., Hindra, Mulder, D., Yin, C. and Elliot, M.A. (2012) Crp is a global regulator of antibiotic production in *Streptomyces*. *mBio*, **3**, e00407-00412.

38. Jain, S. and Keller, N. (2013) Insights to fungal biology through LaeA sleuthing. *Fungal Biol Rev*, **27**, 51-59.

39. Brakhage, A.A. and Schroeckh, V. (2011) Fungal secondary metabolites - strategies to activate silent gene clusters. *Fungal Genet Biol*, **48**, 15-22.

40. Yaegashi, J., Praseuth, M.B., Tyan, S.W., Sanchez, J.F., Entwistle, R., Chiang, Y.M., Oakley, B.R. and Wang, C.C. (2013) Molecular genetic characterization of the biosynthesis cluster of a prenylated isoindolinone alkaloid aspernidine A in *Aspergillus nidulans*. *Org Lett*, **15**, 2862-2865.

41. Williams, R.B., Henrikson, J.C., Hoover, A.R., Lee, A.E. and Cichewicz, R.H. (2008) Epigenetic remodeling of the fungal secondary metabolome. *Org Biomol Chem*, **6**, 1895-1897.

42. Wang, X., Sena Filho, J.G., Hoover, A.R., King, J.B., Ellis, T.K., Powell, D.R. and Cichewicz, R.H. (2010) Chemical epigenetics alters the secondary metabolite composition of guttate excreted by an Atlantic-forest-soil-derived *Penicillium citreonigrum*. *J Nat Prod*, **73**, 942-948.

43. Chen, H.-J., Awakawa, T., Sun, J.-Y., Wakimoto, T. and Abe, I. (2013) Epigenetic modifier-induced biosynthesis of novel fusaric acid derivatives in endophytic fungi from *Datura stramonium L*. *Nat Prod Bioprospect*, **3**, 20-23.

44. Yang, X.L., Huang, L. and Ruan, X.L. (2014) Epigenetic modifiers alter the secondary metabolite composition of a plant endophytic fungus, *Pestalotiopsis crassiuscula* obtained from the leaves of *Fragaria chiloensis*. *J Asian Nat Prod Res*, **16**, 412-417.

45. Moore, J.M., Bradshaw, E., Seipke, R.F., Hutchings, M.I. and McArthur, M. (2012) Chapter eighteen - use and discovery of chemical elicitors that stimulate biosynthetic gene clusters in *Streptomyces* bacteria. In Hopwood, D. A. (ed.), *Methods Enzymol*. Academic Press, Vol. 517, pp. 367-385.

46. Reddy, T.B., Thomas, A.D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E.A. and Kyrpides, N.C. (2014) The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res*, doi: 10.1093/nar/gku1950.

47. Fickett, J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*, **10**, 5303-5318.

48. Borodovsky, M. and McIninch, J. (1993) GENMARK: Parallel gene recognition for both DNA strands. *Compu Chem*, **17**, 123-133.

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

50. Besemer, J., Lomsadze, A. and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*, **29**, 2607-2618.

51. Besemer, J. and Borodovsky, M. (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res*, **33**, W451-454.

52. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O. and Borodovsky, M. (2008) Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Res*, **18**, 1979-1990.

53. Delcher, A.L., Harmon, D., Kasif, S., White, O. and Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*, **27**, 4636-4641.

54. Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*, **26**, 544-548.

55. Majoros, W.H., Pertea, M. and Salzberg, S.L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878-2879.

56. Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J. and Tettelin, H. (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics*, **59**, 24-31.

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

58. Kelley, D.R., Liu, B., Delcher, A.L., Pop, M. and Salzberg, S.L. (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res*, **40**, e9.

59. Fitch, W.M. (1966) An improved method of testing for evolutionary homology. *J Mol Biol*, **16**, 9-16.

60. Sellers, P.H. (1974) On the theory and computation of evolutionary distances. *SIAM J Appl Math*, **26**, 787-793.

61. Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J Mol Biol*, **162**, 705-708.

62. Lipman, D.J. and Pearson, W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435-1441.

63. Pearson, W.R. (1994) Using the FASTA program to search protein and DNA sequence databases. *Methods Mol Biol*, **24**, 307-331.

64. Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, **85**, 2444-2448.

65. Lipman, D.J., Altschul, S.F. and Kececioglu, J.D. (1989) A tool for multiple sequence alignment. *Proc Natl Acad Sci U S A*, **86**, 4412-4415.

66. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673-4680.

67. Karplus, K., Barrett, C. and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846-856.

68. Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755-763.

69. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.

70. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.

71. Marchler-Bauer, A., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M. *et al.* (2009) CDD:

specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res*, **37**, D205-210.

72. Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res*, **38**, D211-222.

73. Letunic, I., Doerks, T. and Bork, P. (2012) SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res*, **40**, D302-D305.

74. Grin, I. and Linke, D. (2011) GCView: the genomic context viewer for protein homology searches. *Nucleic Acids Res*, **39**, W353-356.

75. Liao, C.S., Lu, K., Baym, M., Singh, R. and Berger, B. (2009) IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, **25**, i253-258.

76. Rangannan, V. and Bansal, M. (2010) High-quality annotation of promoter regions for 913 bacterial genomes. *Bioinformatics*, **26**, 3043-3050.

77. Broos, S., Hulpiau, P., Galle, J., Hooghe, B., Van Roy, F. and De Bleser, P. (2011) ConTra v2: a tool to identify transcription factor binding sites across species, update 2011. *Nucleic Acids Res*, **39**, W74-78.

78. Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D. and van Helden, J. (2011) RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res*, **39**, W86-91.

79. de Crecy-Lagard, V., Saurin, W., Thibaut, D., Gil, P., Naudin, L., Crouzet, J. and Blanc, V. (1997) Streptogramin B biosynthesis in *Streptomyces pristinaespiralis* and *Streptomyces virginiae*: molecular characterization of the last structural peptide synthetase gene. *Antimicrob Agents Chemother*, **41**, 1904-1909.

80. Stachelhaus, T., Mootz, H.D. and Marahiel, M.A. (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol*, **6**, 493-505.

81. Challis, G.L., Ravel, J. and Townsend, C.A. (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem Biol*, **7**, 211-224.

82. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. and Huson, D.H. (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res*, **33**, 5799-5808.

83. Rottig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C. and Kohlbacher, O. (2011) NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res*, **39**, W362-367.

84. Caboche, S., Pupin, M., Leclere, V., Fontaine, A., Jacques, P. and Kucherov, G. (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Res*, **36**, D326-331.

85. Yadav, G., Gokhale, R.S. and Mohanty, D. (2003) Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. *J Mol Biol*, **328**, 335-363.

86. Yadav, G., Gokhale, R.S. and Mohanty, D. (2003) SEARCHPKS: A program for detection and analysis of polyketide synthase domains. *Nucleic Acids Res*, **31**, 3654-3658.

87. Udwary, D.W., Merski, M. and Townsend, C.A. (2002) A method for prediction of the locations of linker regions within large multifunctional proteins, and application to a type I polyketide synthase. *J Mol Biol*, **323**, 585-598.

88. Tae, H., Kong, E.B. and Park, K. (2007) ASMPKS: an analysis system for modular polyketide synthases. *BMC Bioinformatics*, **8**, 327.

89. Mallika, V., Sivakumar, K.C., Jaichand, S. and Soniya, E.V. (2010) Kernel based machine learning algorithm for the efficient prediction of type III polyketide synthase family of proteins. *J Integr Bioinform*, **7**, 143.

90. Anand, S., Prasad, M.V., Yadav, G., Kumar, N., Shehara, J., Ansari, M.Z. and Mohanty, D. (2010) SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic Acids Res*, **38**, W487-496.

91. Ansari, M.Z., Yadav, G., Gokhale, R.S. and Mohanty, D. (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res*, **32**, W405-413.

92. Starcevic, A., Zucko, J., Simunkovic, J., Long, P.F., Cullum, J. and Hranueli, D. (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures. *Nucleic Acids Res*, **36**, 6882-6892.

93. Weber, T., Rausch, C., Lopez, P., Hoof, I., Gaykova, V., Huson, D.H. and Wohlleben, W. (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J Biotechnol*, **140**, 13-17.

94. Li, M.H., Ung, P.M., Zajkowski, J., Garneau-Tsodikova, S. and Sherman, D.H. (2009) Automated genome mining for natural products. *BMC Bioinformatics*, **10**, 185.

95. Bachmann, B.O. and Ravel, J. (2009) Methods for *in silico* prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol*, **458**, 181-217.

96. de Jong, A., van Heel, A.J., Kok, J. and Kuipers, O.P. (2011) BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Res*, **38**, W647-651.

97. Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H. and Fedorova, N.D. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet Biol*, **47**, 736-741.

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

99. McClerren, A.L., Cooper, L.E., Quan, C., Thomas, P.M., Kelleher, N.L. and van der Donk, W.A. (2006) Discovery and *in vitro* biosynthesis of haloduracin, a two-component lantibiotic. *Proc Natl Acad Sci U S A*, **103**, 17243-17248.

100. Song, L., Barona-Gomez, F., Corre, C., Xiang, L., Udwary, D.W., Austin, M.B., Noel, J.P., Moore, B.S. and Challis, G.L. (2006) Type III polyketide synthase beta-ketoacyl-ACP starter unit and ethylmalonyl-CoA extender unit selectivity discovered by *Streptomyces coelicolor* genome mining. *J Am Chem Soc*, **128**, 14754-14755.

101. Nielsen, M.L., Nielsen, J.B., Rank, C., Klejnstrup, M.L., Holm, D.K., Brogaard, K.H., Hansen, B.G., Frisvad, J.C., Larsen, T.O. and Mortensen, U.H. (2011) A genome-wide polyketide synthase deletion library uncovers novel genetic links to polyketides and meroterpenoids in *Aspergillus nidulans*. *FEMS Microbiol Lett*, **321**, 157-166.

102. Chiang, Y.M., Szewczyk, E., Nayak, T., Davidson, A.D., Sanchez, J.F., Lo, H.C., Ho, W.Y., Simityan, H., Kuo, E., Praseuth, A. *et al.* (2008) Molecular genetic mining of the

Aspergillus secondary metabolome: discovery of the emericellamide biosynthetic pathway. *Chem Biol*, **15**, 527-532.

103. Chiang, Y.M., Szewczyk, E., Davidson, A.D., Keller, N., Oakley, B.R. and Wang, C.C. (2009) A gene cluster containing two fungal polyketide synthases encodes the biosynthetic pathway for a polyketide, asperfuranone, in *Aspergillus nidulans*. *J Am Chem Soc*, **131**, 2965-2970.

104. Bergmann, S., Funk, A.N., Scherlach, K., Schroeckh, V., Shelest, E., Horn, U., Hertweck, C. and Brakhage, A.A. (2010) Activation of a silent fungal polyketide biosynthesis pathway through regulatory cross talk with a cryptic nonribosomal peptide synthetase gene cluster. *Appl Environ Microbiol*, **76**, 8143-8149.

105. Bergmann, S., Schumann, J., Scherlach, K., Lange, C., Brakhage, A.A. and Hertweck, C. (2007) Genomics-driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. *Nat Chem Biol*, **3**, 213-217.

106. Laureti, L., Song, L., Huang, S., Corre, C., Leblond, P., Challis, G.L. and Aigle, B. (2011) Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in *Streptomyces ambofaciens*. *Proc Natl Acad Sci U S A*, **108**, 6258-6263.

107. Olano, C., Garcia, I., Gonzalez, A., Rodriguez, M., Rozas, D., Rubio, J., Sanchez-Hidalgo, M., Brana, A.F., Mendez, C. and Salas, J.A. (2014) Activation and identification of five clusters for secondary metabolites in *Streptomyces albus* J1074. *Microb Biotechnol*, **7**, 242-256.

108. Chou, W.K., Fanizza, I., Uchiyama, T., Komatsu, M., Ikeda, H. and Cane, D.E. (2010) Genome mining in *Streptomyces avermitilis*: cloning and characterization of SAV_76, the synthase for a new sesquiterpene, avermitilol. *J Am Chem Soc*, **132**, 8850-8851.

109. Baltz, R.H. (2010) *Streptomyces* and *Saccharopolyspora* hosts for heterologous expression of secondary metabolite gene clusters. *J Ind Microbiol Biotechnol*, **37**, 759-772.

110. Zhang, H., Boghigian, B.A., Armando, J. and Pfeifer, B.A. (2011) Methods and options for the heterologous production of complex natural products. *Nat Prod Rep*, **28**, 125-151.

111. Ongley, S.E., Bian, X., Neilan, B.A. and Muller, R. (2013) Recent advances in the heterologous expression of microbial natural product biosynthetic pathways. *Nat Prod Rep*, **30**, 1121-1138.

112. Johannes, T.W., DeSieno, M.A., Griffin, B.M., Thomas, P.M., Kelleher, N.L., Metcalf, W.W. and Zhao, H. (2010) Deciphering the late biosynthetic steps of antimalarial compound FR-900098. *Chem Biol*, **17**, 57-64.

113. Juguet, M., Lautru, S., Francou, F.X., Nezbedova, S., Leblond, P., Gondry, M. and Pernodet, J.L. (2009) An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in *Streptomyces ambofaciens*. *Chem Biol*, **16**, 421-431.

114. Woodyer, R.D., Shao, Z., Thomas, P.M., Kelleher, N.L., Blodgett, J.A., Metcalf, W.W., van der Donk, W.A. and Zhao, H. (2006) Heterologous production of fosfomycin and identification of the minimal biosynthetic gene cluster. *Chem Biol*, **13**, 1171-1182.

115. Kaysser, L., Bernhardt, P., Nam, S.J., Loesgen, S., Ruby, J.G., Skewes-Cox, P., Jensen, P.R., Fenical, W. and Moore, B.S. (2012) Merochlorins A-D, cyclic meroterpenoid

antibiotics biosynthesized in divergent pathways with vanadium-dependent chloroperoxidases. *J Am Chem Soc*, **134**, 11988-11991.

116. Fazio, G.C., Xu, R. and Matsuda, S.P. (2004) Genome mining to identify new plant triterpenoids. *J Am Chem Soc*, **126**, 5678-5679.

117. Gross, F., Luniak, N., Perlova, O., Gaitatzis, N., Jenke-Kodama, H., Gerth, K., Gottschalk, D., Dittmann, E. and Muller, R. (2006) Bacterial type III polyketide synthases: phylogenetic analysis and potential for the production of novel secondary metabolites by heterologous expression in pseudomonads. *Arch Microbiol*, **185**, 28-38.

118. Fu, J., Bian, X., Hu, S., Wang, H., Huang, F., Seibert, P.M., Plaza, A., Xia, L., Muller, R., Stewart, A.F. *et al.* (2012) Full-length RecE enhances linear-linear homologous recombination and facilitates direct cloning for bioprospecting. *Nat Biotechnol*, **30**, 440-446.

119. Tang, L., Shah, S., Chung, L., Carney, J., Katz, L., Khosla, C. and Julien, B. (2000) Cloning and heterologous expression of the epothilone gene cluster. *Science*, **287**, 640-642.

120. Luo, Y., Huang, H., Liang, J., Wang, M., Lu, L., Shao, Z., Cobb, R.E. and Zhao, H. (2013) Activation and characterization of a cryptic polycyclic tetramate macrolactam biosynthetic gene cluster. *Nat Commun*, **4**, 2894.

121. Lombo, F., Velasco, A., Castro, A., de la Calle, F., Brana, A.F., Sanchez-Puelles, J.M., Mendez, C. and Salas, J.A. (2006) Deciphering the biosynthesis pathway of the antitumor thiocoraline from a marine actinomycete and its expression in two *Streptomyces* species. *Chembiochem*, **7**, 366-376.

122. Yamanaka, K., Reynolds, K.A., Kersten, R.D., Ryan, K.S., Gonzalez, D.J., Nizet, V., Dorrestein, P.C. and Moore, B.S. (2014) Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc Natl Acad Sci U S A*, **111**, 1957-1962.

123. Shao, Z., Rao, G., Li, C., Abil, Z., Luo, Y. and Zhao, H. (2013) Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold. *ACS Synth Biol*, **2**, 662-669.

124. Pfeifer, B.A., Admiraal, S.J., Gramajo, H., Cane, D.E. and Khosla, C. (2001) Biosynthesis of complex polyketides in a metabolically engineered strain of *E. coli*. *Science*, **291**, 1790-1792.

125. McDaniel, R., Ebert-Khosla, S., Hopwood, D.A. and Khosla, C. (1993) Engineered biosynthesis of novel polyketides. *Science*, **262**, 1546-1550.

126. Gomez-Escribano, J.P. and Bibb, M.J. (2011) Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. *Microb Biotechnol*, **4**, 207-215.

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

128. Ikeda, H., Kazuo, S.Y. and Omura, S. (2014) Genome mining of the *Streptomyces avermitilis* genome and development of genome-minimized hosts for heterologous expression of biosynthetic gene clusters. *J Ind Microbiol Biotechnol*, **41**, 233-250.

129. Komatsu, M., Komatsu, K., Koiwai, H., Yamada, Y., Kozone, I., Izumikawa, M., Hashimoto, J., Takagi, M., Omura, S., Shin-ya, K. *et al.* (2013) Engineered *Streptomyces*

avermitilis host for heterologous expression of biosynthetic gene cluster for secondary metabolites. *ACS Synth Biol*, **2**, 384-396.

130. Nathans, D. and Smith, H.O. (1975) Restriction endonucleases in the analysis and restructuring of DNA molecules. *Annu Rev Biochem*, **44**, 273-293.

131. Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd and Smith, H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*, **6**, 343-345.

132. Li, M.Z. and Elledge, S.J. (2012) SLIC: a method for sequence- and ligation-independent cloning. *Methods Mol Biol*, **852**, 51-59.

133. Klock, H.E., Koesema, E.J., Knuth, M.W. and Lesley, S.A. (2008) Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. *Proteins*, **71**, 982-994.

134. Bitinaite, J., Rubino, M., Varma, K.H., Schildkraut, I., Vaisvila, R. and Vaiskunaite, R. (2007) USER friendly DNA engineering and cloning method by uracil excision. *Nucleic Acids Res*, **35**, 1992-2002.

135. Quan, J. and Tian, J. (2011) Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. *Nat Protoc*, **6**, 242-251.

136. Zhang, L., Zhao, G. and Ding, X. (2011) Tandem assembly of the epothilone biosynthetic gene cluster by in vitro site-specific recombination. *Sci. Rep.*, **1**, 141.

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

138. Zhang, Y., Buchholz, F., Muyrers, J.P. and Stewart, A.F. (1998) A new logic for DNA engineering using recombination in *Escherichia coli*. *Nat Genet*, **20**, 123-128.

139. Li, M.Z. and Elledge, S.J. (2005) MAGIC, an in vivo genetic method for the rapid construction of recombinant DNA molecules. *Nat Genet*, **37**, 311-319.

140. Ma, H., Kunes, S., Schatz, P.J. and Botstein, D. (1987) Plasmid construction by homologous recombination in yeast. *Gene*, **58**, 201-216.

141. Zhang, Y., Werling, U. and Edelmann, W. (2012) SLiCE: a novel bacterial cell extract-based DNA cloning method. *Nucleic Acids Res*, **40**, e55.

142. Shetty, R.P., Endy, D. and Knight, T.F., Jr. (2008) Engineering BioBrick vectors from BioBrick parts. *J Biol Eng*, **2**, 5.

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

144. Chen, W.H., Qin, Z.J., Wang, J. and Zhao, G.P. (2013) The MASTER (methylation-assisted tailorable ends rational) ligation method for seamless DNA assembly. *Nucleic Acids Res*, **41**, e93.

145. Hillson, N.J., Rosengarten, R.D. and Keasling, J.D. (2012) j5 DNA assembly design automation software. *ACS Synth Biol*, **1**, 14-21.

146. Linshiz, G., Stawski, N., Poust, S., Bi, C., Keasling, J.D. and Hillson, N.J. (2013) PaR-PaR laboratory automation platform. *ACS Synth Biol*, **2**, 216-222.

147. Linshiz, G., Stawski, N., Goyal, G., Bi, C., Poust, S., Sharma, M., Mutalik, V., Keasling, J.D. and Hillson, N.J. (2014) PR-PR: cross-platform laboratory automation system. *ACS Synth Biol*, **3**, 515-524.

148. Appleton, E., Tao, J., Haddock, T. and Densmore, D. (2014) Interactive assembly algorithms for molecular cloning. *Nat Methods*, **11**, 657-662.

149. Brannigan, J.A. and Wilkinson, A.J. (2002) Protein engineering 20 years on. *Nat Rev Mol Cell Biol*, **3**, 964-970.

150. Tobin, M.B., Gustafsson, C. and Huisman, G.W. (2000) Directed evolution: the 'rational' basis for 'irrational' design. *Curr Opin Struct Biol*, **10**, 421-427.

151. Chen, K. and Arnold, F.H. (1993) Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc Natl Acad Sci U S A*, **90**, 5618-5622.

152. Stemmer, W.P. (1994) DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution. *Proc Natl Acad Sci U S A*, **91**, 10747-10751.

153. Stemmer, W.P. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature*, **370**, 389-391.

154. Rubin-Pitel, S., Cho, C.M.-H., Chen, W. and Zhao, H. (2006) Directed evolution tools in bioproduct and bioprocess development. In Yang, S.-T. (ed.), *Bioprocessing for Value-Added Products from Renewable Resources: New Technologies and Applications*. Elsevier, New York, pp. 49-72.

155. Zhao, H., Giver, L., Shao, Z., Affholter, J.A. and Arnold, F.H. (1998) Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat Biotechnol*, **16**, 258-261.

156. Zhao, H. and Arnold, F.H. (1999) Directed evolution converts subtilisin E into a functional equivalent of thermitase. *Protein Eng*, **12**, 47-53.

157. Keefe, A.D. and Szostak, J.W. (2001) Functional proteins from a random-sequence library. *Nature*, **410**, 715-718.

158. O'Maille, P.E., Bakhtina, M. and Tsai, M.-D. (2002) Structure-based combinatorial protein engineering (SCOPE). *J Mol Biol*, **321**, 677-691.

159. Voigt, C.A., Martinez, C., Wang, Z.-G., Mayo, S.L. and Arnold, F.H. (2002) Protein building blocks preserved by recombination. *Nat Struct Mol Biol*, **9**, 553-558.

160. Dalby, P.A. (2011) Strategy and success for the directed evolution of enzymes. *Curr Opin Struct Biol*, **21**, 473-480.

161. Cobb, R.E., Si, T. and Zhao, H. (2012) Directed evolution: an evolving and enabling synthetic biology tool. *Curr Opin Chem Biol*, **16**, 285-291.

162. Chockalingam, K., Chen, Z., Katzenellenbogen, J.A. and Zhao, H. (2005) Directed evolution of specific receptor–ligand pairs for use in the creation of gene switches. *Proc Natl Acad Sci U S A*, **102**, 5691-5696.

163. Reetz, M.T., Carballeira, J.D. and Vogel, A. (2006) Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew Chem Int Ed Engl*, **45**, 7745-7751.

164. Fasan, R., Chen, M.M., Crook, N.C. and Arnold, F.H. (2007) Engineered alkane-hydroxylating cytochrome P450BM3 exhibiting nativelike catalytic properties. *Angew Chem*, **119**, 8566-8570.

165. Savile, C.K., Janey, J.M., Mundorff, E.C., Moore, J.C., Tam, S., Jarvis, W.R., Colbeck, J.C., Krebber, A., Fleitz, F.J., Brands, J. *et al.* (2010) Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science*, **329**, 305-309.

166. Park, H.-S., Nam, S.-H., Lee, J.K., Yoon, C.N., Mannervik, B., Benkovic, S.J. and Kim, H.-S. (2006) Design and evolution of new catalytic activity with an existing protein scaffold. *Science*, **311**, 535-538.

167. Röthlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J.L., Althoff, E.A., Zanghellini, A., Dym, O. *et al.* (2008) Kemp elimination catalysts by computational enzyme design. *Nature*, **453**, 190-195.

168. Karanicolas, J., Corn, J.E., Chen, I., Joachimiak, L.A., Dym, O., Peck, S.H., Albeck, S., Unger, T., Hu, W., Liu, G. *et al.* (2011) A *de novo* protein binding pair by computational design and directed evolution. *Mol Cell*, **42**, 250-260.

169. Crook, N. and Alper, H.S. (2012) Classical Strain Improvement. In Patnaik, R. (ed.), *Engineering Complex Phenotypes in Industrial Strains*. John Wiley & Sons, Inc., pp. 1-33.

170. Zhang, Y.X., Perry, K., Vinci, V.A., Powell, K., Stemmer, W.P. and del Cardayre, S.B. (2002) Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature*, **415**, 644-646.

171. Hida, H., Yamada, T. and Yamada, Y. (2007) Genome shuffling of *Streptomyces* sp. U121 for improved production of hydroxycitric acid. *Appl Microbiol Biotechnol*, **73**, 1387-1393.

172. Otte, B., Grunwaldt, E., Mahmoud, O. and Jennewein, S. (2009) Genome shuffling in *Clostridium diolis* DSM 15410 for improved 1,3-propanediol production. *Appl Environ Microbiol*, **75**, 7610-7616.

173. Zhang, Y., Liu, J.Z., Huang, J.S. and Mao, Z.W. (2010) Genome shuffling of *Propionibacterium shermanii* for improving vitamin B12 production and comparative proteome analysis. *J Biotechnol*, **148**, 139-143.

174. Cheng, Y., Song, X., Qin, Y. and Qu, Y. (2009) Genome shuffling improves production of cellulase by *Penicillium decumbens* JU-A10. *J Appl Microbiol*, **107**, 1837-1846.

175. Wang, H.H., Isaacs, F.J., Carr, P.A., Sun, Z.Z., Xu, G., Forest, C.R. and Church, G.M. (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, **460**, 894-898.

176. Bonde, M.T., Kosuri, S., Genee, H.J., Sarup-Lytzen, K., Church, G.M., Sommer, M.O.A. and Wang, H.H. (2014) Direct mutagenesis of thousands of genomic targets using microarray-derived oligonucleotides. *ACS Synth Biol*, doi: 10.1021/sb5001565.

177. Warner, J.R., Reeder, P.J., Karimpour-Fard, A., Woodruff, L.B.A. and Gill, R.T. (2010) Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. *Nat Biotechnol*, **28**, 856-862.

178. DiCarlo, J.E., Conley, A.J., Penttila, M., Jantti, J., Wang, H.H. and Church, G.M. (2013) Yeast oligo-mediated genome engineering (YOGE). *ACS Synth Biol*, **2**, 741-749.

179. Si, T., Luo, Y., Bao, Z. and Zhao, H. (2014) RNAi-assisted genome evolution in *Saccharomyces cerevisiae* for complex phenotype engineering. *ACS Synth Biol*, doi: 10.1021/sb500074a.

180. Xiao, H. and Zhao, H. (2014) Genome-wide RNAi screen reveals the E3 SUMO-protein ligase gene *SIZ1* as a novel determinant of furfural tolerance in *Saccharomyces cerevisiae*. *Biotechnol Biofuels*, **7**, 78.

181. Sun, N., Abil, Z. and Zhao, H. (2012) Recent advances in targeted genome engineering in mammalian systems. *Biotechnol J*, **7**, 1074-1087.

182. Chen, Z., Wen, F., Sun, N. and Zhao, H. (2009) Directed evolution of homing endonuclease I-SceI with altered sequence specificity. *Protein Eng Des Sel*, **22**, 249-256.

183. Ashworth, J., Havranek, J.J., Duarte, C.M., Sussman, D., Monnat, R.J., Jr., Stoddard, B.L. and Baker, D. (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, **441**, 656-659.

184. Porteus, M.H. and Carroll, D. (2005) Gene targeting using zinc finger nucleases. *Nat Biotechnol*, **23**, 967-973.

185. Ramirez, C.L., Foley, J.E., Wright, D.A., Muller-Lerch, F., Rahman, S.H., Cornu, T.I., Winfrey, R.J., Sander, J.D., Fu, F., Townsend, J.A. *et al.* (2008) Unexpected failure rates for modular assembly of engineered zinc fingers. *Nat Methods*, **5**, 374-375.

186. Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A. and Bonas, U. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**, 1509-1512.

187. Joung, J.K. and Sander, J.D. (2013) TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol*, **14**, 49-55.

188. Cermak, T., Doyle, E.L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J.A., Somia, N.V., Bogdanove, A.J. and Voytas, D.F. (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res*, **39**, e82.

189. Liang, J., Chao, R., Abil, Z., Bao, Z. and Zhao, H. (2014) FairyTALE: a high-throughput TAL effector synthesis platform. *ACS Synth Biol*, **3**, 67-73.

190. Sander, J.D. and Joung, J.K. (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol*, **32**, 347-355.

191. Mali, P., Esvelt, K.M. and Church, G.M. (2013) Cas9 as a versatile tool for engineering biology. *Nat Methods*, **10**, 957-963.

192. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. and Church, G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823-826.

193. Bao, Z., Xiao, H., Liang, J., Zhang, L., Xiong, X., Sun, N., Si, T. and Zhao, H. (2014) Homology-integrated CRISPR-Cas (HI-CRISPR) system for one-step multigene disruption in *Saccharomyces cerevisiae*. *ACS Synth Biol*, doi: 10.1021/sb500255k.

194. Zhou, Y., Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y. and Wei, W. (2014) High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature*, **509**, 487-491.

CHAPTER 2. Characterization and engineering of FrbF for the synthesis of novel FR-900098 analogs

2.1 Introduction

2.1.1 The ongoing malaria threat

With 207 million cases and 627,000 deaths estimated in 2012 alone, malaria remains one of the most significant threats to global health today [1]. Despite past and current control efforts, nearly half of the world's population still remains at risk. The disease is caused by any of four different species of *Plasmodia* – *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, and *Plasmodium ovale* – all of which are transmitted to human hosts through mosquitoes of the genus *Anopheles*. Of these four parasite species, *Plasmodium falciparum* and *Plasmodium vivax* are the most common [1]. Children under five years of age are particularly susceptible to the disease, accounting for 76 % of the reported deaths in 2012.

Artemisinin, a sesquiterpene lactone derived from the plant *Artemisia annua* (Figure 2.1), currently serves as the basis for frontline malaria treatments. Known as *qing hao* in China, the plant has a long history of therapeutic application [2]. Artemisinin itself was first isolated from *Artemisia annua* in 1972, and subsequent assays against *Plasmodium falciparum* demonstrated significant potency relative to established antimalarials chloroquine and mefloquine. Trials in mice also showed good efficacy in treating *Plasmodium berghei*, and a 1979 human trial cured 2099 patients, although moderate recrudescence was observed [3].

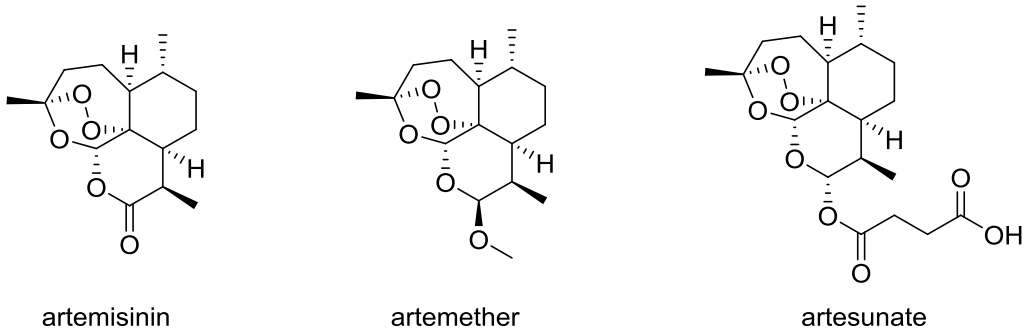


Figure 2.1: Artemisinin and representative clinically-relevant derivatives.

Although the native producer has served as the primary source of artemisinin, cultivation is time and resource intensive and cannot keep up with the significant demand [4]. To solve this problem, total synthesis of the compound was attempted by several research groups in the 1980s and 1990s. Although multiple total syntheses were achieved, none was suitable for commercialization [5-8]. As vast improvements were made in recombinant DNA technology around the turn of the century, biosynthesis became a much more attractive alternative. Efforts toward the biosynthesis of artemisinin were spearheaded by researchers at UC Berkeley, and further developed by the start-up company Amyris. In 2003, synthesis of the artemisinin precursor amorphadiene was achieved in *Escherichia coli* via an engineered mevalonate pathway from *Saccharomyces cerevisiae* and a synthetic amorpha-4,11-diene synthase gene [9]. In 2006, artemisinic acid titers of up to 100 mg L^{-1} were achieved in an engineered strain of *S. cerevisiae*, paving the way to commercially relevant biosynthesis in this organism [10]. In 2013, titers of 25 g L^{-1} were achieved in an engineered *S. cerevisiae* strain, which was subsequently licensed to pharmaceutical company Sanofi [11]. An efficient chemical conversion of artemisinic acid to artemisinin was also demonstrated, allowing further derivatization to more pharmacologically favorable forms of the drug (such as artesunate or artemether) through established chemical

processes. Industrial-scale production of 50 to 60 tons of artemisinin per year is expected in 2014, which is equivalent to between 80 and 150 million malaria treatments [12].

Despite these great successes in the biosynthetic production of artemisinin, artemisinin itself does not present a complete solution to the ongoing malaria problem. The World Health Organization strongly discourages the use of artemisinin or its derivatives as solo therapeutics, as this can provide selective pressure for the development of resistance in the *Plasmodium* parasite. Evidence of artemisinin resistance has already been observed in field isolates from Cambodia, Myanmar, Thailand, and Vietnam, calling for urgent containment measures to stop its spread [1,13-15]. Rather, artemisinin-based drugs are to be used only in combination with complementary antimalarial drugs as artemisinin combination therapies (ACTs). However, as many other antimalarials have seen decades of monotherapeutic application already, resistance can be a significant concern with their usage as well. Thus, there is still a need for new antimalarial drugs, not only to replace the old but to supplement each other in novel combination therapies as the fight against malaria continues.

2.1.2 Antimalarial phosphonic acids fosmidomycin and FR-900098

2.1.2.1 Initial discovery and studies

Two natural products of particular interest in the antimalarial field are the phosphonic acids fosmidomycin (originally called FR-31564) and FR-900098 (Figure 2.2). These compounds were first isolated in the late 1970s by the Fujisawa Pharmaceutical Co., Ltd. from *Streptomyces lavendulae* and *Streptomyces rubellomurinus* sp. nov., respectively [16,17]. They were discovered in a screening program of soil isolates against the nocardicin C-supersensitive mutant

of *Pseudomonas aeruginosa* NCTC 10490, designed to identify novel inhibitors of cell wall biosynthesis. The chemical structures of both compounds were solved at the time of isolation. They were found to differ only in their *N*-acyl groups, with fosmidomycin containing a formyl group and FR-900098 containing an acetyl moiety. Antibiotic spectra for these compounds were determined against a variety of microorganisms in a serial dilution agar plate assay, and both were found to be much more effective against Gram-negative bacteria than Gram-positive. In all cases tested, fosmidomycin was found to be at least as effective as FR-900098, often exhibiting a significantly lower MIC. As a result, further development efforts and clinical trials proceeded with fosmidomycin in the 1980s [18-22]. It was found to have good pharmacokinetic properties and was initially targeted as a treatment of urinary tract infections. Though initial results showed promise, the drug was found to be poorly effective against recurrent infections, and so further development was halted [23].

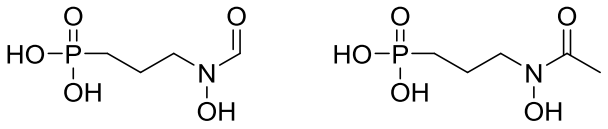


Figure 2.2: Phosphonic acids fosmidomycin and FR-900098.

2.1.2.2 Renewed interest and clinical trials

Following its initial discovery, little subsequent research interest was expressed in FR-900098, as it was thought to be the inferior of the two compounds from an antibiotic standpoint. However, interest in both compounds was rekindled in 1993 with the elucidation of the non-mevalonate pathway for the biosynthesis of isoprenoid precursors [24]. As isoprenoid synthesis was previously thought to proceed solely through the mevalonate pathway, the discovery of this

alternate pathway shed light on novel targets for herbicides and antibiotics. In 1989, it had already been demonstrated that fosmidomycin derived its antibiotic activity through inhibition of isoprenoid synthesis [25]. Kuzuyama and co-workers demonstrated this inhibition *in vitro* with purified deoxyxylulose-5-phosphate reductoisomerase (Dxr) from the non-mevalonate pathway in *E. coli*, observing mixed-mode inhibition with a K_i value of 38 nM for fosmidomycin [26]. The full resurgence of fosmidomycin and FR-900098 was not underway until 1999, however, when the Jomaa research group first identified the two compounds as potential antimalarials [27]. In this study, bioinformatic analysis was employed to identify a possible Dxr enzyme in the recently sequenced genome of *Plasmodium falciparum*. Fosmidomycin and FR-900098 were both assayed against three strains of *P. falciparum*. While both compounds exhibited low IC_{50} values (≤370 nM), FR-900098 proved to be roughly twice as effective as fosmidomycin. This also proved true in *in vitro* inhibition assays against the purified recombinant target enzyme. Finally, animal studies were conducted in a mouse model with *P. vinckei*. Again, FR-900098 proved to be the most effective antimalarial drug, exhibiting <1 % parasitemia with a 2 mg/kg dosage.

Likely owing to the difficulty in preparing large quantities of FR-900098, antimalarial clinical trials in humans instead proceeded with fosmidomycin. Two initial trials evaluated the efficacy of fosmidomycin: first in 27 adults in Gabon, and then in 10 adults each from Gabon and Thailand [28,29]. In both trials, rapid initial clearance of the parasite was observed; however, recrudescence was found to be a significant concern, as several subjects had re-acquired the disease when evaluated at two or four weeks after starting treatment. These results highlight the limits of fosmidomycin monotherapy, again pointing to the benefits of combination therapies for

malaria treatment. Subsequently, fosmidomycin-clindamycin was evaluated in children of various age groups [30-32]. Initial studies focused on children aged 7 – 14 and found no evidence of recrudescence on day 14 when a multi-day treatment regimen was applied. Further, comparison with 5-day regimens of either fosmidomycin alone or clindamycin alone clearly demonstrated the advantage of the combination therapy. Extension of the subject age range showed similar cure rates in children as young as 3, but a significant decrease in children aged 1 – 2. An ACT consisting of fosmidomycin and the artemisinin derivative artesunate was also evaluated in children aged 6 – 15, revealing 100 % 14-day cure rates with dosing regimens of ≥2 days. Follow-up at 28 days revealed ≥90 % cure rates with dosing regimens of ≥3 days [33].

2.1.2.3 Chemical synthesis of analogs and derivatives

Although fosmidomycin and FR-900098 were demonstrated to be potent antibiotic and antimalarial compounds, the fact remains that they are not optimized for use as human therapeutics. As a result, several efforts have been directed toward the synthesis of a variety of analogs and derivatives, both in a search for greater efficacy and to facilitate structure-activity relationship (SAR) studies. The earliest of these was carried out in an effort to develop improved antibiotics by researchers at the Fujisawa Chemical Co., Ltd. [34]. Through their synthetic efforts, they were able to generate four different types of compounds: fosmidomycin and FR-900098 with an additional hydroxyl group on the β-carbon; 1-*trans*-propenyl analogs of both compounds; ethylene analogs of both compounds, containing only a two-carbon spacer; and methylphosphinic acids of both compounds. These derivatives were assayed against a broad spectrum of microbes for antibiotic behavior. Although some of the derivatives showed minor improvements against particular species, fosmidomycin was again found to be the most potent

overall antibiotic. As a result, interest in synthesis of derivatives and analogs remained minimal for the next several years. One exception, however, is the work of Öhler and Kanzler, who in 1995 published an alternate synthetic route to the 1-*trans*-propenyl analogs [35]. Through subsequent hydrogenation, the corresponding propylphosphonates could be generated for further derivatization. Nevertheless, the authors' were interested in these compounds solely from a synthetic perspective, and as a result biological assays were not pursued.

Following the discovery of fosmidomycin and FR-900098 as potent antimalarials and subsequent clinical trials, interest in derivative synthesis boomed. Among the goals of these efforts was improvement of the 20 – 40 % bioavailability observed for fosmidomycin [36]. To this end, multiple syntheses of ester prodrugs were reported (Figure 2.3). These derivatives were designed to increase uptake by masking the charged phosphonate moiety until the compounds enter the bloodstream, at which point ubiquitous esterases release the active drug. In 2001, Reichenberg and co-workers synthesized three diaryl ester prodrugs of FR-900098: the diphenyl ester, the bis-(2-methyl-phenyl) ester, and the bis-(4-methoxy-phenyl) ester [37]. Of these, oral administration of the diphenyl ester was found to be more effective than oral FR-900098 in the *P. vinckei* mouse model, while oral administration of the bis-(4-methoxy-phenyl) ester was most effective, performing comparably to intraperitoneally dosed FR-900098. However, concern over the toxicity of the phenol product of ester bond cleavage led to the exploration of alternate alkyl masking groups. Synthesis of a variety of acyloxyalkyl prodrugs and evaluation in the *P. vinckei* mouse model revealed the pivaloyloxymethyl, acetyloxyethyl, and propionyloxyethyl esters to be more effective than FR-900098. Additionally, evaluation of plasma concentrations of FR-900098 after dosage confirmed the improved bioavailability of the acetyloxyethyl ester [38]. In

a follow-up study, alkoxycarbonyloxy esters were also synthesized and evaluated [39]. Both the methoxycarbonyloxyethyl and ethoxycarbonyloxyethyl esters were found to be more effective than FR-900098. Single ester prodrugs were also synthesized and evaluated, as these crystalline compounds would lend themselves more easily to capsule formulation than the oily diester compounds. However, no improvement in efficacy over FR-900098 was observed with these compounds.

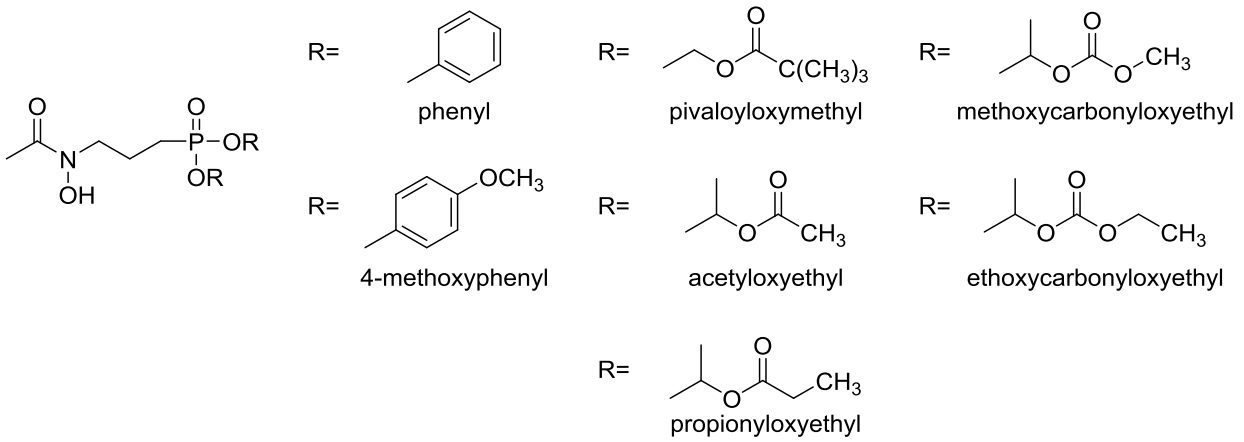


Figure 2.3: Ester protecting groups found to increase the antimalarial efficacy of FR-900098.

The next area of derivatization to be explored in depth was substitution on the α-carbon (Figure 2.4), with several examples published within the last decade [40-45]. The van Calenbergh group, for example, explored a variety of substituted phenyl groups in the α-position and assayed them both against purified *E. coli* Dxr and two *P. falciparum* strains. Curiously, none of the derivatives proved to be better inhibitors of *E. coli* Dxr than fosmidomycin or FR-900098, but three of the compounds (one fosmidomycin derivative and two FR-900098 derivatives, each featuring chlorine substituents on the phenyl moiety) did show significant improvement in IC_{50} against one or both of the *P. falciparum* strains. The Kurz group similarly tested a series of

arylmethyl fosmidomycin and FR-900098 derivatives against *P. falciparum*, also including pivaloyloxymethyl ester protecting groups on the phosphonate. Although they were able to identify one compound (containing a 3,4-dichlorobenzyl substituent) with improved activity when compared to the fosmidomycin prodrug, none of their analogs could outperform the FR-900098 prodrug. Finally, a different study by Kurz and co-workers looked at various alkyl substitutions at the α-position of the pivaloyloxymethyl ester prodrugs. It was found that ethyl, propyl, isopropyl, and dimethyl substitutions significantly decreased antimalarial efficacy, while methyl substitution produced a compound of comparable antimalarial activity to the FR-900098 prodrug. Some substituted phenyl derivatives were also included in this study, with results similar to those described above. More recently, the van Calenbergh group examined heteroatom derivatization at the α-position. Halogenation of the FR-900098 α-carbon with fluorine or chlorine was analyzed by growth inhibition of *P. falciparum* and in the *P. berghei* mouse model. Both compounds performed comparably to FR-900098 in *P. falciparum* inhibition, and the fluorinated derivative performed better in the *in vivo* trial. Addition of an amine or hydroxyl group to the α-carbon of FR-900098 was also explored to introduce ether, amide, urea, and 1,4-triazole functionalities. However, none of these derivatives outperformed FR-900098 *in vitro* or *in vivo*.

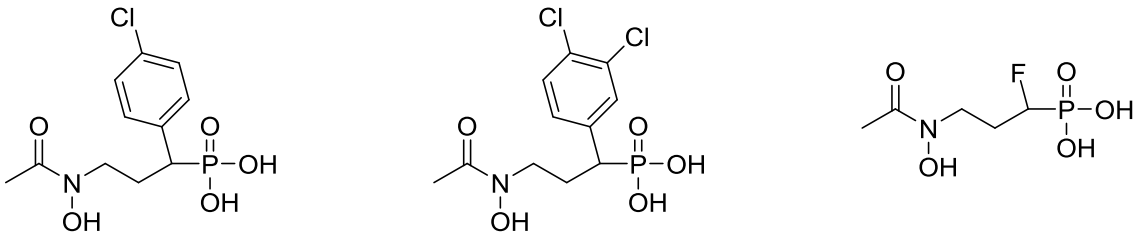


Figure 2.4: Substitutions at the α position with significantly improved antimalarial activity compared to FR-900098.

Beyond just α-substituted carbon spacers, cyclized and unsaturated derivatives have also been synthesized and analyzed for antimalarial efficacy. Another study from the van Calenbergh group looked at cyclopropyl fosmidomycin and FR-900098 derivatives, restricting rotational freedom of the spacer by bridging the α- and β-carbons [46]. Unfortunately, none of the derivatives synthesized could outperform FR-900098 in inhibition of *E. coli* Dxr or *P. falciparum* growth. Similarly, incorporation of the three-carbon spacer into a cyclopentane ring failed to match the inhibition levels of fosmidomycin and FR-900098 toward *E. coli* Dxr [47]. Finally, synthesis of α,β-unsaturated derivatives with a variety of α-aryl substitutions was carried out, but the resulting compounds were all significantly worse inhibitors of *E. coli* Dxr [48].

Among the most recent attempts at the synthesis of novel derivatives are the β- and γ-oxa isosteres and the so-called "reverse analogs," replacing the N-terminal moiety with a reversed hydroxamic acid group (Figure 2.5). Of the β- and γ-oxa isosteres synthesized, γ-oxa substitution was found to significantly reduce activity, while β-oxa substitution of FR-900098 or the N-methyl hydroxamate reverse analog yielded slightly better inhibitors of *P. falciparum* growth [49]. Alternatively, the Kurz group synthesized the α-phenyl-substituted hydroxamate and N-methyl hydroxamate FR-900098 reverse analogs, revealing the latter to be a particularly potent inhibitor of *P. falciparum* and its recombinant Dxr enzyme [50]. In a follow-up study, the 3,4-difluorophenyl α- substituted hydroxamate and N-methyl hydroxamate FR-900098 reverse analogs were synthesized, yielding further improvement in IC_{50} measured against *P. falciparum* [51]. The α-fluorinated N-methyl hydroxamate reverse analog has also been found to have high antimalarial activity when assayed in the *P. berghei* mouse model, further illustrating the promise of reverse analogs as more potent antimalarials [44].

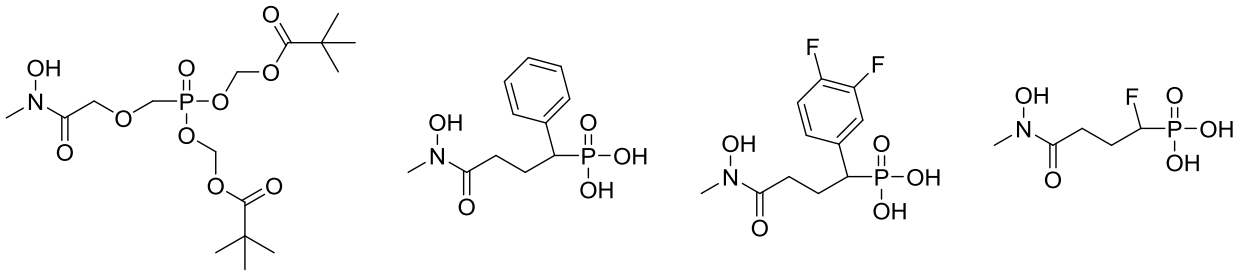


Figure 2.5: Potent reverse analogs of FR-900098, including the β-oxa reverse analog with pivaloyloxymethyl ester protecting groups and three α-substituted compounds.

2.1.2.4 Heterologous expression and characterization

With renewed interest in FR-900098 came renewed interest in its biosynthesis, both to elucidate the potentially unique biochemical transformations involved and to facilitate industrially relevant production through protein and metabolic engineering efforts. Although few phosphonic acid gene clusters had previously been characterized, a common initial step in their biosyntheses was noted to be conversion of phosphoenolpyruvate (PEP) to phosphonopyruvate through the activity of a PEP mutase (PEPM) gene [52]. As a result, identification of the FR-900098 biosynthetic cluster from the native producer *S. rubellomurinus* began with PCR-amplification of the PEPM gene from a genomic DNA fosmid library using degenerate primers. One hit from this screening process was subsequently integrated into the chromosome of the heterologous host *Streptomyces lividans*, enabling production of FR-900098. Sequencing and deletion analysis led to the identification of eight open reading frames (ORFs), designated *frbA – frbH*, which were found to be necessary for FR-900098 production; two additional ORFs, *frbI* and *frbJ*, immediately downstream of the minimal cluster but not necessary for FR-900098 production; and one final ORF, *dxrB*, which appeared to be encode a Dxr homolog. Of the eight requisite ORFs, *frbD* was found to encode the PEPM originally identified in the fosmid library screening, catalyzing the first step of FR-900098 biosynthesis (Figure 2.6). As no phosphonopyruvate decarboxylase gene

was identified to catalyze the typical second step of known phosphonic acid biosyntheses, it was proposed that the early steps of FR-900098 biosynthesis parallel the TCA cycle, catalyzed by FrbA, FrbB, FrbC, and FrbE. *In vitro* assays with purified FrbD and FrbC clearly demonstrated the formation of 2-phosphonomethylmalate, confirming FrbC to be phosphonomethylmalate synthase. FrbA and either FrbB or FrbE (or both) were predicted to catalyze the subsequent TCA-analogous steps, yielding 2-oxo-4-phosphonobutyrate.

To decipher the downstream steps in FR-900098 biosynthesis (Figure 2.6), the entire gene cluster was transferred to *E. coli*. The eight requisite ORFs and *dxrB*, necessary for self-resistance, were placed on three compatible plasmids, each under the control of an IPTG-inducible T7 promoter. The result was a strain capable of reaching FR-900098 titers of 6.3 mg/L in shake-flask cultures in LB media [53]. Through feeding studies, *in silico* analysis and *in vitro* enzymatic assays, it was determined that FrbH catalyzed both the PLP-dependent decarboxylation of 2-amino-4-phosphonobutyrate (derived from 2-oxo-4-phosphonobutyrate via endogenous transamination) and nucleotide transfer to yield CMP-5'-3-aminopropylphosphonate (CMP-5'-3APn). Further *in vitro* studies showed that CMP-5'-FR-900098 was formed via hydroxylation and subsequent acetylation of CMP-5'-3APn, catalyzed by FrbG and FrbF, respectively. Formation of the final product FR-900098 could be achieved through hydrolysis of the nucleotide catalyzed by FrbI, or by ubiquitous nucleotide hydrolases within the host. Finally, a possible function for FrbJ was demonstrated *in vitro* in the conversion of FR-900098 to FR-33289, a derivative first reported by Okuhara and co-workers at the same time that FR-900098 was originally identified [17,54]. However, *in vivo* production of this compound was not detected.

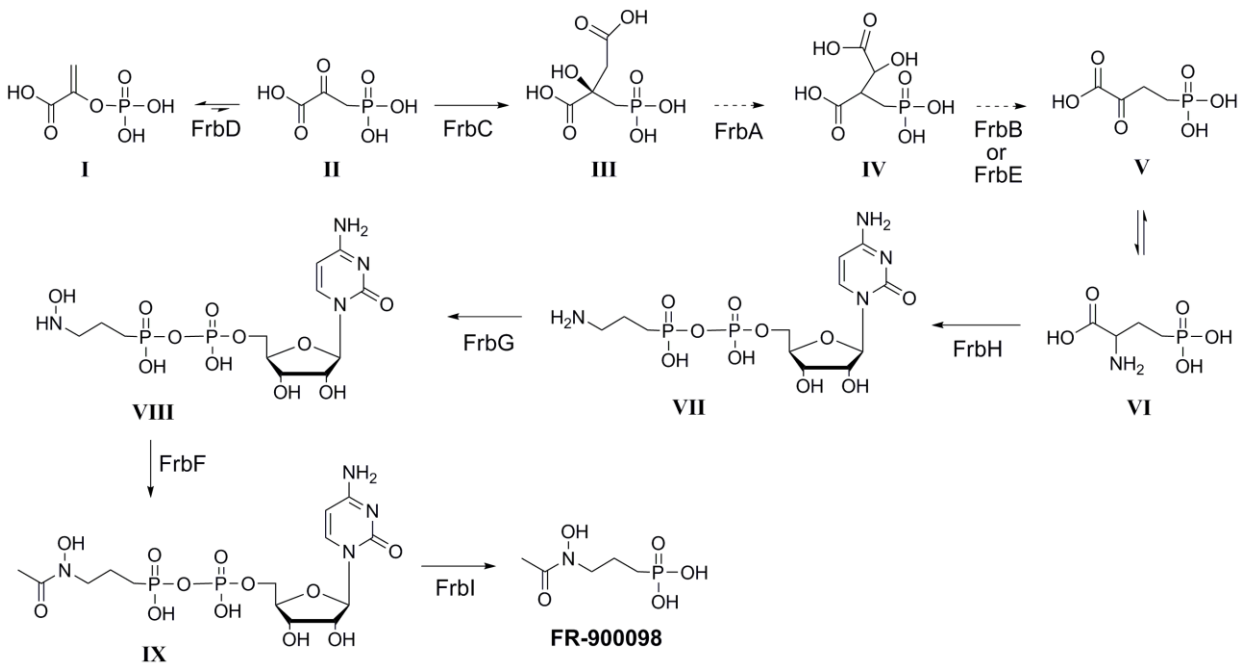


Figure 2.6: The FR-900098 biosynthetic pathway, adapted from ref. 53.

2.1.2.5 Biosynthetic derivatization

To complement the chemical derivatization efforts outlined above, the focus of the work described here is to apply a biosynthetic approach to the synthesis of novel FR-900098 derivatives. To this end, the *N*-acetyltransferase FrbF was selected as the target for kinetic and structural characterization as well as directed evolution for multiple reasons. First of all, FrbF catalyzes the penultimate step of FR-900098 biosynthesis, followed only by nucleotide cleavage to give the final product. As a result, none of the upstream pathway intermediates will be altered in the context of a mutated FrbF, indicating that all of the upstream enzymes will still be able to function normally. Mutating an enzyme further upstream in the pathway, in contrast, would yield a series of altered pathway intermediates. Given the high substrate specificity observed for some of the pathway enzymes, such as the bifunctional nucleotide transferase/decarboxylase FrbH, this could significantly slow or completely halt the biosynthesis of any novel derivatives.

The only enzyme downstream of FrbF, the nucleotide hydrolase FrbI, exhibits very broad substrate specificity, and thus would likely be unaffected by the introduction of any FrbF mutant. Further, even in the absence of FrbI, ubiquitous nucleotide hydrolases within the cell can also serve to catalyze the final step. Second, with the exception of a study by Gießmann and coworkers in which several bulky aromatic moieties were examined [55], variation in the *N*-acyl substituent remains largely unexplored. Moreover, it has already been shown that variation at this end of the molecule is tolerated and can lead to more potent Dxr inhibitors, as exemplified by the *N*-methyl hydroxamate "reverse analogs" described above. Third, although FrbF demonstrates activity analogous to enzymes of the well-studied GCN5-like *N*-acetyltransferase (GNAT) superfamily, FrbF shares no sequence similarity with any members thereof. Rather, FrbF shares sequence similarity with a number of uncharacterized, putative aminoglycoside *N*(3')-acetyltransferases. As a result, characterization of FrbF could grant useful biochemical insights into a new class of *N*-acetyltransferases, and could facilitate further studies of new uncharacterized secondary metabolite pathways.

Here, it is our goal to characterize the catalytic activity of FrbF with its native substrates. Additionally, thorough analysis of the active site will be applied to elucidate the catalytic mechanism of this enzyme. Substrate specificity of the wild type enzyme will be then explored with respect to its acyl-donor to synthesize novel FR-900098 derivatives *in vitro* and *in vivo*. To evaluate new FR-900098 derivatives as antimalarials, *in vitro* inhibition assays will be conducted against the heterologously expressed and purified Dxr enzyme from *P. falciparum*.

2.2 Results and Discussion

2.2.1 Elucidation of the FrbF catalytic mechanism

In order to elucidate the catalytic mechanism of FrbF, the residues located within 6 Å of the acyl moiety of the bound acetyl-CoA were analyzed in the x-ray crystal structure solved by Dr. Brian Bae of Prof. Satish Nair's group at the University of Illinois, Urbana-Champaign. Based on literature precedent, one could expect the reaction mechanism to proceed either via direct attack of the amine or hydroxylamine substrate on the carbonyl group of acetyl-CoA, as has been observed for the majority of GNAT superfamily members, or via an acyl-enzyme intermediate, as demonstrated in the yeast histone acetyltransferase Esa1 and in multiple mycobacterial enzymes [56,57]. Of the residues present in the active site, Glu-187, Ser-188, Thr-190, and Glu-231 were identified as possible general acids, while only His-193 was identified as a potential general base. Alanine mutants at each of these five positions were created, and their relative activities were measured in comparison to the wild type enzyme (Figure 2.7).

The results indicate that the Glu-187→Ala and Glu231→Ala mutants demonstrate activities similar to the wild-type enzyme, confirming that these two residues do not play a role in catalysis and ruling out their participation in a mechanism that requires formation of a covalent acyl intermediate with an active site carboxylate. The Ser-188→Ala mutant exhibited a modest decrease in activity, as anticipated based on its interaction with the phosphopantetheine arm of the acetyl-CoA substrate. In contrast, the activity of the His-193→Ala mutant was less than 1 % of that of the wild-type, and no activity could be observed for the Thr-190→Ala mutant. These results strongly suggest the involvement of Thr-190 and His-193 in acid/base chemistry during acetyl-group transfer. These results are consistent with the mechanistic route proposed in Figure

2.8. Following the binding of the phosphonic acid substrate acetyl-CoA at the active site, His-193 functions as a general base to facilitate the attack of the amine or hydroxylamine on the electrophilic carbonyl of the acetyl group. Protonation of the resultant tetrahedral intermediate results in its collapse to yield the hydroxamate and CoA. Alternatively, it is worth noting that His-193 could instead serve to stabilize the tetrahedral intermediate, as the N-O distance varies from 2.9 Å to 4.2 Å between the two pairs of FrbF dimers.

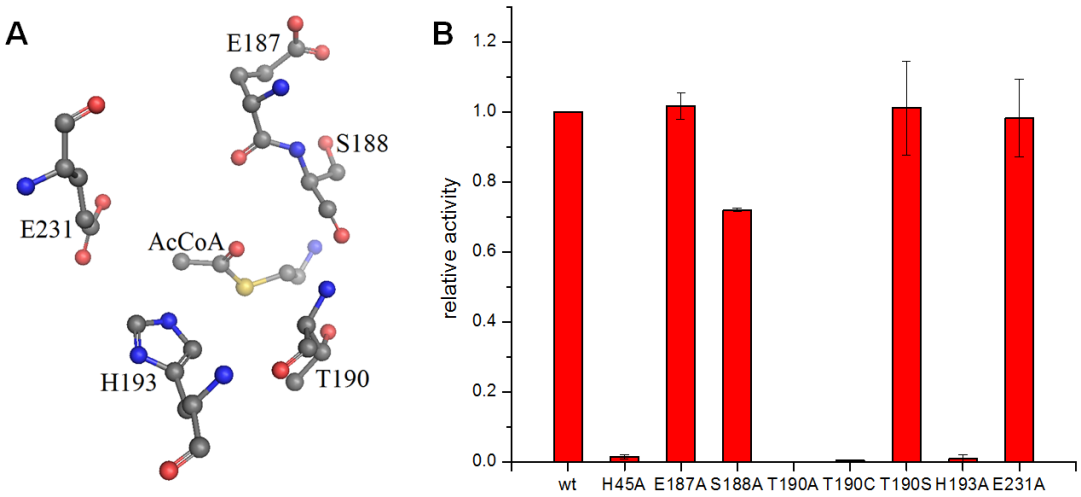


Figure 2.7: Mechanistic characterization of FrbF. (a) FrbF residues near the bound acetyl-CoA substrate. (b) Relative activities of FrbF active site mutants.

Further analysis was carried out at the Thr-190 position through mutations to both Ser and Cys, as well as mutation of His-45 to Ala. His-45 is the nearest residue to Thr-190, and so we speculated that this residue might serve to activate Thr-190 for a novel reaction mechanism involving attack of Thr-190 on the carbonyl of the acetyl group to form an acyl-enzyme intermediate. The His-45→Ala mutation reduced the relative activity of FrbF to 1.5% of the wild type, supporting the catalytic relevance of this residue. However, high-resolution mass spectrometry provided no evidence of an acyl-enzyme intermediate (data not shown). Further

analysis of the crystal structure suggests that His-45 can participate in hydrogen bonding interactions both with a carbonyl group of the acetyl-CoA phosphopantetheine arm and with the backbone amide nitrogen of either Thr-190 or Ser-191, suggesting that this residue is critical for the positioning of acetyl-CoA and the catalytically requisite Thr-190 residue. Further, the Thr-190→Cys mutation, which would more closely reflect the active site of NAT enzymes that utilize an acyl-enzyme intermediate, lowered the activity dramatically, discrediting the prospect of a novel, Thr-mediated ping pong mechanism. Finally, the Thr-190→Ser mutation had no detectable effect on the relative activity.

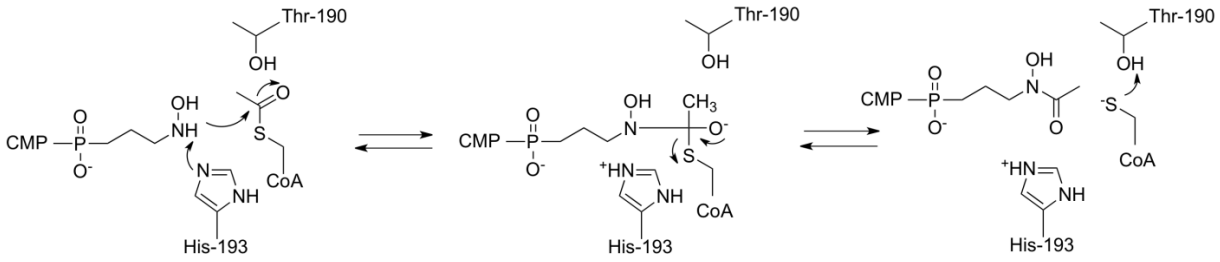


Figure 2.8: The proposed mechanism for the *N*-acetylation of CMP-5'-H3APn catalyzed by FrbF.

2.2.2 Kinetics characterization with phosphonate substrates

2.2.2.1 CMP-5'-3APn synthesis and kinetics

Although the FR-900098 biosynthetic pathway has been definitively demonstrated to proceed via *N*-hydroxylation followed by *N*-acetylation [53], FrbF is nevertheless able to acetylate both the primary amine substrate CMP-5'-3APn and the hydroxylamine substrate CMP-5'-H3APn (compounds **VII** and **VIII**, respectively, in Figure 2.6). As a result, CMP-5'-3APn was enzymatically synthesized using FrbH and purified for kinetic assays. Kinetic parameters with CMP-5'-3APn have previously been reported using a continuous assay with 5,5'-dithiobis-(2-

nitrobenzoic acid) (DTNB), but numerous efforts to replicate these results under identical conditions proved unsuccessful. This is most likely due to the limited sensitivity of DTNB-based quantitation, especially given the low activity level of FrbF. Further, DTNB can react with the protein itself, increasing the noise level.

As a result, an HPLC-based assay, in which the CMP-conjugated product can be directly assayed at a UV absorbance wavelength of 254 nm, was instead employed. Surprisingly, linear dependence on substrate concentration was observed far beyond the K_m value of 35 μM previously reported, ultimately yielding a K_m value of 391 μM with a k_{cat} value of 2.042 min^{-1}. Given that CMP-5'-3APn is not the native substrate for FrbF, it is not surprising to see this relatively low level of activity. Further evaluation of the enzyme with the His_6 tag at the *C*-terminal position or with the *N*-terminal His_6 tag removed via thrombin cleavage did not show significantly altered activity (data not shown).

2.2.2.2 CMP-5'-H3APn assays

The native substrate for FR-900098 synthesis, CMP-5'-H3APn, can be synthesized *in vitro* using the purified FrbH and FrbG enzymes. However, numerous efforts to observe FrbF catalysis with the purified reaction product revealed no reproducible activity. Subsequent LC-MS analysis of the purified substrate revealed a dominant signal corresponding to a parent ion m/z of 457 in negative mode, corresponding to the nitroso derivative CMP-5'-NO3APn.

Two schemes were proposed for the oxidation of the hydroxylamine to the nitroso. First, the hydroxylamine could become a surrogate substrate for FrbG and pass through a second catalytic

cycle, yielding the nitroso after spontaneous dehydration. Second, the hydroxylamine could simply undergo a non-enzymatic oxidation, either through reaction with oxygen or through a disproportionation reaction with another hydroxylamine molecule. To study the first possibility, the FrbG reaction was monitored over time by LC-MS, and the signals for the substrate and both products were followed with time. In this case, we would expect to see an initial formation of the hydroxylamine, followed by consumption of this compound and a lagging increase in formation of the nitroso. However, concomitant formation of both compounds was observed, even at the earliest time points assayed, indicating that the second possibility may be at play.

To further study the second possibility, two methods were pursued. First, the FrbG reaction was run for a specified amount of time, after which the enzyme was removed by diafiltration. Subsequently, the filtered reaction mixture was incubated at a constant temperature and periodically analyzed by LC-MS. At 30 °C, rapid loss of the hydroxylamine compound was observed, with half of the sample depleted in ~90 min. Even at 5 °C, the m/z = 459 signal was completely lost from the LC-MS trace following overnight incubation. To corroborate this observation, attempts were made to purify the hydroxylamine from FrbG reaction mixtures, but this could not be achieved. Instead, the nitroso compound was purified and subsequently reduced back to the hydroxylamine under anaerobic conditions using samarium diiodide. Following anaerobic incubation for several hours, little reformation of the nitroso was observed by LC-MS. Once exposed to air, however, rapid reformation occurred, with the hydroxylamine signal almost completely lost after 30 min. Thus, it can be concluded that the conversion of the hydroxylamine to the nitroso is a rapid non-enzymatic process, as has been observed previously for other hydroxylamine compounds [58].

To assess the kinetic behavior of FrbF with this substrate, *in situ* generation via concomitant FrbG catalysis was attempted. Interestingly, incubation of equimolar amounts of FrbF and FrbG showed no significant difference in FrbF activity, both via HPLC analysis and DTNB endpoint assay. This is likely due to the fact that even though FrbG can supply the true hydroxylamine substrate for FrbF, both enzymes can also compete for the primary amine substrate, and some of the hydroxylamine can be lost to the non-reactive nitroso. Increasing the FrbG concentration, however, led to an increase in observed FrbF activity, up to a 10:1 molar ratio of FrbG to FrbF. Beyond this ratio, FrbF activity did not increase further. Given the same initial concentration of FrbF, the observed activity level of FrbF increased 6.7-fold when supplemented with a 10-fold molar excess of FrbG. This indicates that FrbF does in fact show greater activity with CMP-5'-H3APn than CMP-5'-3APn, which correlates with its understood role in FR-900098 biosynthesis.

2.2.3 Kinetics characterization with coenzyme A substrates

2.2.3.1 Acetyl-CoA kinetic characterization

For the biosynthesis of FR-900098, FrbF installs an acetyl moiety on the hydroxylamine substrate by transferring the respective functional group from acetyl-CoA. To determine the kinetic parameters with respect to this substrate, we employed the HPLC-based assay described above. In this case, the CMP-5'-3APn surrogate substrate was used to ensure that the concentration of this substrate remained constant from trial to trial. Kinetic characterization with acetyl-CoA revealed a 20-fold lower K_m value than that observed for the CMP-5'-3APn substrate. This is likely due to the significant number of favorable contacts that the enzyme can form with this rather large substrate molecule in its binding cleft. The k_{obs} with acetyl-CoA is

approximately 55 % of the k_{cat} observed for CMP-5'-3APn. This is a reasonable observation given that the acetyl-CoA kinetic parameters were measured with the CMP-5'-3APn concentration near its K_m value, at which point half of the maximum rate would be expected.

2.2.3.2 Evaluation of alternate acyl-CoAs

Here our goal was to determine the scope of acyl-donor substrates that FrbF will accept for the synthesis of novel *N*-acylated FR-900098 derivatives. As a result, seven alternate acyl-CoAs were selected for relative activity assays, including propionyl-, isobutyryl-, β-hydroxybutyryl-, acetoacetyl-, malonyl-, succinyl-, and glutaryl-CoA (Figure 2.9). These substrates were selected to gradually expand the size of the acyl substituent, such that only small changes to the FR-900098 scaffold will be made. Additionally, these acyl-CoA substrates are biologically relevant, and their overproduction could potentially be engineered in future host strains for the overproduction of specific FR-900098 derivatives.

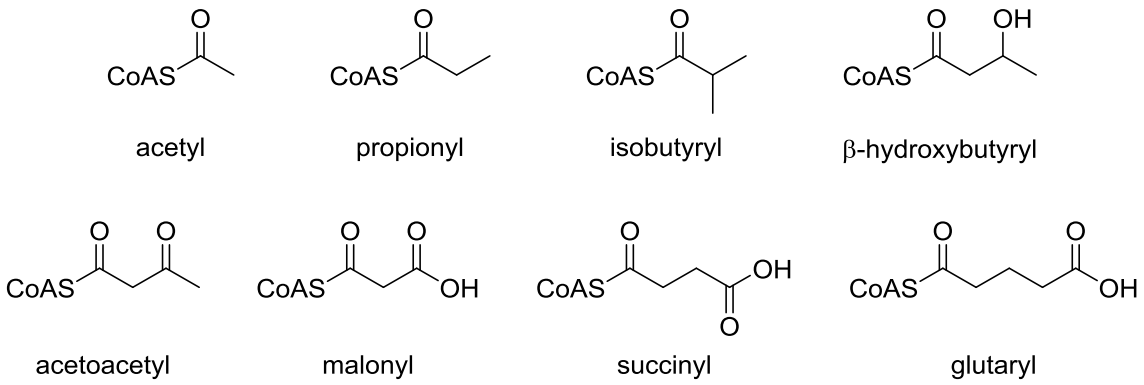


Figure 2.9: Side chains of alternate acyl-CoA substrates.

Activities with alternate acyl-CoA substrates were measured via an HPLC-MS assay (Table 2.1). Initial analyses showed no detectable activity by UV absorbance or MS for glutaryl-CoA, which

is not surprising given that this was the largest acyl substituent evaluated. Surprising, however, was the absence of detectable activity for β-hydroxybutyryl-CoA, given that activity could be detected for the similar acetoacetyl- and malonyl-CoA substrates. This suggests that the second carbonyl moiety not present in β-hydroxybutyryl-CoA may be necessary for favorable interactions with the active site to support catalysis. Of the other substrates evaluated, both isobutyryl- and succinyl-CoA showed activity by MS analysis, as indicated by the presence of the expected m/z parent ion with the corresponding m/z = 322 daughter ion due to loss of CMP. However, no detectable UV peak was observed for either of these two products, preventing accurate quantitation of relative activity toward these novel substrates. Propionyl-, malonyl-, and acetoacetyl-CoA all showed activity both by UV absorbance and MS analysis, and so activity levels were measured for each of these substrates relative to acetyl-CoA. Additionally, kinetic parameters were determined for propionyl-CoA, which exhibited the highest relative activity. This is not surprising, given that the propionyl moiety is the most similar to the acetyl moiety.

Interestingly, the K_m value obtained for propionyl-CoA is approximately the same as that obtained for acetyl-CoA. This likely indicates that binding of the acyl-CoA substrate relies predominantly on interactions with the CoA backbone itself, and not with the acyl functionality. The ~12-fold drop in k_{obs} for this substrate relative to acetyl-CoA, then, is most likely caused by the increased steric hindrance introduced by the bulkier acyl chain, which could potentially block the amine or hydroxylamine of the substrate from attacking the thioester carbonyl. This is further evinced by the decreasing activity observed with the bulkier malonyl-CoA (2.7 % relative

activity) and acetoacetyl-CoA (0.4 % relative activity), and the lack of quantifiable activity observed when the isobutyryl-CoA or succinyl-CoA substrates are utilized.

Table 2.1: Acyl-CoA activity assays (mean ± standard deviation of three replicates)

substrate	K_m (μM)	k_{cat} (min^{-1})	relative activity
acetyl-CoA	19.5 ± 0.4	1.12 ± 0.03[a]	1.0
propionyl-CoA	23.6 ± 4.2	0.089 ± 0.024[a]	0.080 ± 0.022
malonyl-CoA	ND	ND	0.027 ± 0.006
acetoacetyl-CoA	ND	ND	0.0044 ± 0.0013
isobutyryl-CoA	ND	ND	MS
succinyl-CoA	ND	ND	MS

[a]: measured at 400μM CMP-5'-3APn
ND: not determined
MS: below UV detection limit; product observed by MS/MS only

2.2.4 *In vitro* synthesis of FR-900098 derivatives

To demonstrate *in vitro* synthesis of FR-900098 derivatives, the non-native propionyl-CoA substrate was chosen, as it showed the highest relative activity among all acyl-CoAs evaluated. Synthesis of the propionyl-FR-900098 derivative (named FR-900098P) was attempted through sequential reactions with the purified FrbH, FrbG, and FrbF enzymes. After each step, the enzyme was removed via diafiltration and products were isolated via HPLC fractionation. However, this procedure resulted in almost undetectable amounts of the desired CMP-conjugated FR-900098 derivative, again due to the non-enzymatic oxidation of the hydroxylamine intermediate, CMP-5'-H3APn, to its nitroso derivative. In an attempt to prevent this undesired side reaction, the FrbG and FrbF reactions were instead performed simultaneously, such that the hydroxylamine product of FrbG could immediately be acylated by FrbF. This was found to increase the amount of CMP-FR-900098P generated, such that it became easily detectable by HPLC (Figure 2.10a). Following purification of this product, CMP cleavage was performed by addition of FrbI. The FrbI reaction was performed separately to avoid non-specific cleavage of

CMP from other pathway intermediates [53]. The identity of the newly synthesized product was confirmed by tandem MS/MS (Figure 2.10b). In positive mode, FR-900098 undergoes a characteristic 198→138 fragmentation, indicating loss of both the *N*-acetyl and *N*-hydroxyl moieties. As expected, we observed the same characteristic fragmentation (212→138) for FR-900098P.

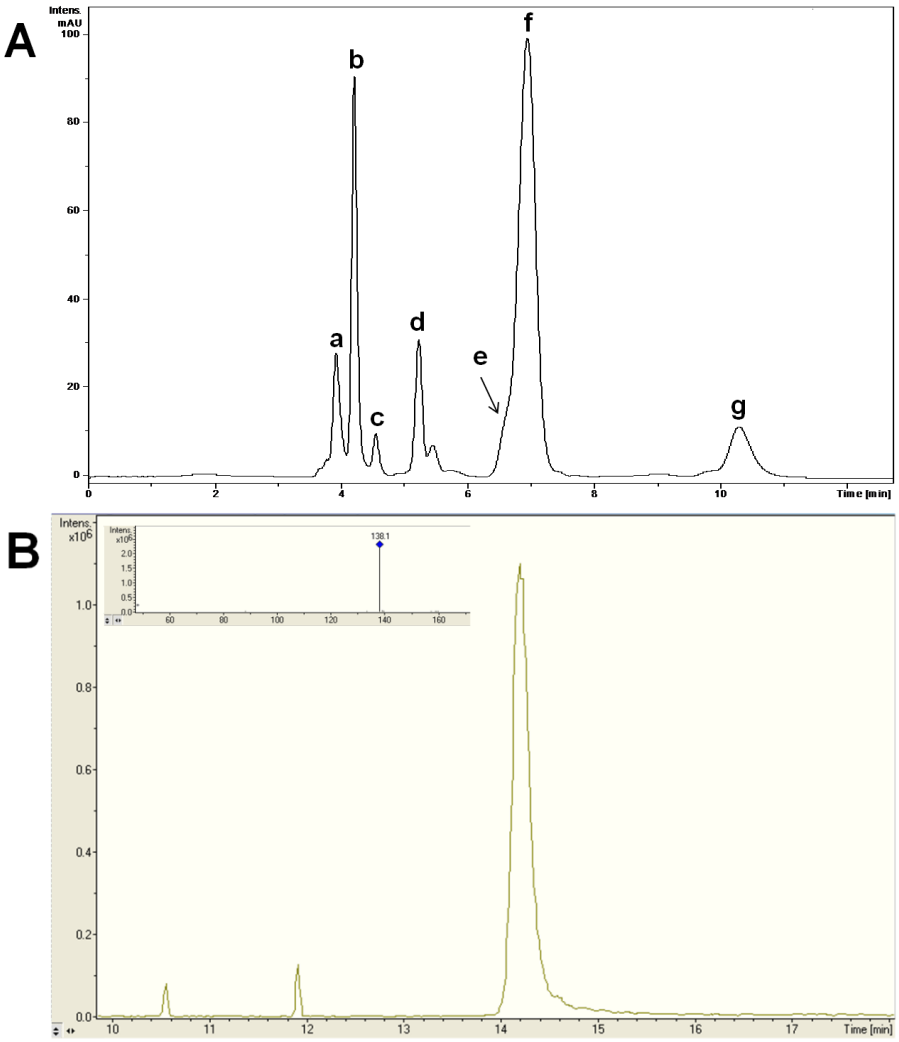


Figure 2.10: Enzymatic synthesis of FR-900098P. (a) Co-incubation of FrbG and FrbF following the FrbH reaction yields a mix of UV active compounds, including: **a**, CDP; **b**, CMP; **c**, CMP-5'-NO3APn; **d**, CMP-5'-3APn; **e**, CMP-5'-P3APn; **f**, $NADP^+$; and **g**, CMP-5'-FR-900098P. (b) MS trace showing synthesized FR-900098P (m/z = 212^+) with the expected m/z = 138^+ daughter ion.

To extend the scope of FR-900098 analogs synthesized, analogous enzymatic syntheses were performed with malonyl-, acetoacetyl, and *n*-butyryl-CoA. The resulting CMP-conjugated FR-900098 derivatives were all observable by HPLC-MS (Figure 2.11), and could be cleaved by FrbI to yield the corresponding FR-900098 derivative. However, low activity of FrbF with the non-native substrates relative to the rapid non-enzymatic oxidation of CMP-5'-H3APn led to very poor estimated yields, preventing further analysis with these compounds.

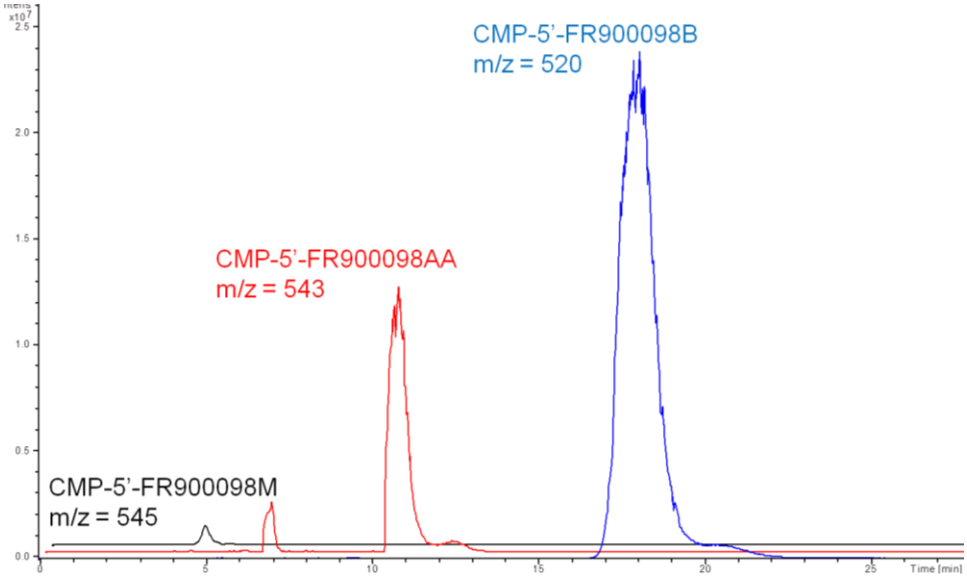


Figure 2.11: Synthesis of FR-900098 analogs. Feeding of malonyl-CoA, acetoacetyl-CoA, and *n*-butyryl-CoA to FrbF enables synthesis of the corresponding CMP-conjugated FR-900098 analogs. Extracted ion chromatograms are shown for the malonyl (black, top trace), acetoacetyl (red, middle trace), and *n*-butyryl (blue, bottom trace) products.

2.2.5 PfDxr expression and characterization

2.2.5.1 Recombinant expression in E. coli

To develop a means to evaluate the antimalarial potential of novel FR-900098 derivatives, we elected to heterologously express and purify the PfDxr target enzyme in *E. coli*. Expression of

the full-length protein (including the coding sequence for a 72 amino acid leader peptide) proved unsuccessful, and the tag was removed to express only the core enzyme [27]. Following optimization of expression conditions, including expression temperature, IPTG concentration, and total culture volume, the pure PfDxr protein was successfully purified, as confirmed by SDS-PAGE (Figure 2.12a). To demonstrate activity of the pure protein, an assay with the native substrates NADPH and DXP was carried out, demonstrating reduction in absorbance at 340 nm indicative of NADPH oxidation (Figure 2.12b). To prove that this activity was due to the PfDxr enzyme itself and not a contaminating protein, the assay was repeated in the presence of fosmidomycin, demonstrating nearly complete inhibition.

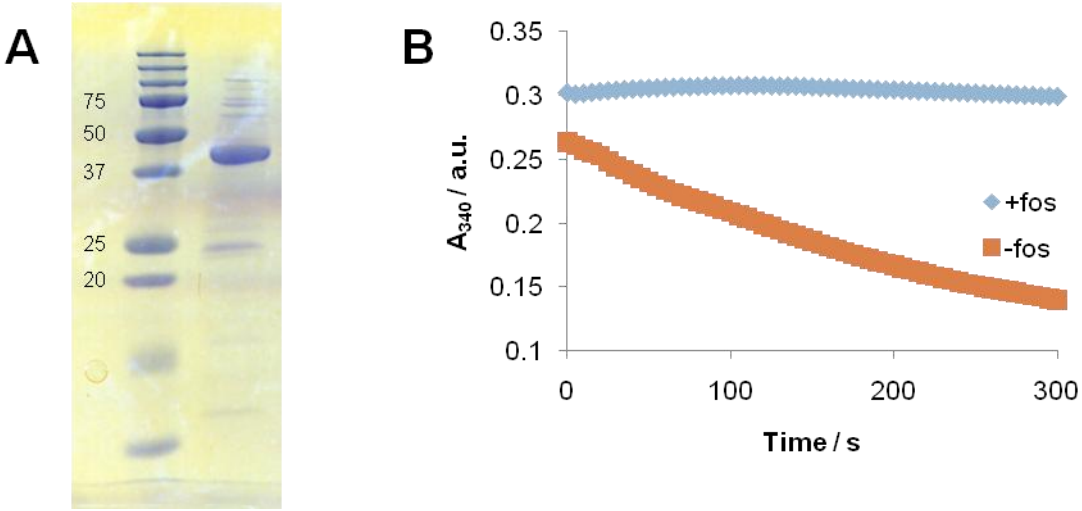


Figure 2.12: (a) Purity of the recombinant PfDxr enzyme was confirmed via SDS-PAGE (expected MW = 47 kDa), while (b) activity was confirmed by tracking NADPH oxidation in the presence or absence of fosmidomycin.

2.2.5.2 Kinetic characterization

Based on the literature precedent in the study of Dxr enzymes from other organisms, we elected to measure the kinetics of PfDxr in the presence of Mn^{2+}, as the greatest or near-greatest activities have typically been observed with this divalent cation [59-61]. Only NADPH was

evaluated as the co-factor, as no Dxr protein has been identified with preference toward NADH. Activity measurements by cuvette in a total volume of 300 µL proved to be reproducible, but it was found that k_{cat} values measured with the same PfDxr stock decreased with time. In contrast, the calculated K_m values did not change significantly with time, indicating that the enzyme loses activity in solution, even when kept on ice. As a result, a fresh aliquot of PfDxr (stored at -80 °C in 15 % glycerol) was used for all subsequently performed assays to try to minimize loss of activity.

Kinetic analysis of PfDxr revealed a K_m value of 52 ± 13 µM and a k_{cat} value of 2.0 ± 0.5 s^{-1}. Interestingly, the K_m value obtained for PfDxr is significantly lower than the value reported for DXP reductoisomerases from *E. coli* Dxr or *Zymomonas mobilis* in the presence of the same divalent cation. This suggests that the PfDxr enzyme is particularly well suited to bind DXP, which may have implications for the binding of inhibitors as well. The observed k_{cat} value is approximately twice that of EcDxr, but not as great as those obtained for *Z. mobilis* (13 s^{-1}) or *Pseudomonas aeruginosa* (8 s^{-1}) [60,62]. Overall, these kinetic parameters are most similar to those obtained for *Mycobacterium tuberculosis* Dxr with Mg^{2+} (K_m = 42 µM, k_{cat} = 2.1 s^{-1}) [61].

2.2.5.3 Structure elucidation with fosmidomycin and FR-900098

The X-ray crystal structure of PfDxr in complex with fosmidomycin or FR-900098, along with the co-factor NADPH and the divalent cation Mn^{2+}, was solved by Dr. Brian Bae and Dr. Zhi Li of Prof. Satish Nair's group at the University of Illinois, Urbana-Champaign. In both structures, the hydroxamate moiety of the inhibitor coordinates to the Mn^{2+} ion via its two oxygen atoms. Notably, the *N*-acyl moiety inserts into a hydrophobic binding pocket formed by Trp-296, Met-

298, Met-360, and the nicotinamide ring of the cofactor. The presence of this pocket justifies the increased potency of FR-900098 relative to fosmidomycin as a PfDxr inhibitor, and suggests that further hydrophobic extension of the *N*-acyl side chain could provide even better binding to the active site.

Comparison of our PfDxr structure to the EcDxr co-crystal structure with the co-factor and fosmidomycin reveals key differences that suggest increased potency of the inhibitor toward PfDxr. For example, PfDxr possesses a narrower substrate-binding cavity, with movement of Met-298 toward the *N*-acyl moiety to form the aforementioned hydrophobic pocket. Concomitantly, H-293 moves into a position in which it can hydrogen bond with the phosphonate end of the inhibitor. The equivalent residues in EcDxr (His-208 and Met-213) are not positioned well for either of these interactions. An additional difference is seen at the position of Ser-269 in the PfDxr structure, which can also hydrogen bond with the phosphonate. In EcDxr, Gly-184 occupies this space and thus cannot hydrogen bond.

To evaluate these interactions biochemically, the Ser-269→Ala and Met-298→Ala mutants were constructed. Following purification, both mutants were assayed for activity. While the S269A mutant was capable of turning over NADPH and was inhibited by fosmidomycin, the M298A mutant was inactive. A repeated attempt at purification of the M298A mutant again yielded the inactive enzyme. This likely indicates that this active site residue is catalytically requisite, and might be involved in maintaining structural integrity of the active site or positioning the substrate for catalysis. Inhibition assays were carried out with the active S269A mutant, and an increase in the K_i value for fosmidomycin from 7.5 nM to 9.8 nM was observed. This demonstrates that this

residue's interaction with the inhibitor is partly responsible for the superior inhibition of PfDxr by fosmidomycin relative to EcDxr.

2.2.6 PfDxr inhibition studies

Having demonstrated synthesis of FR-900098P *in vitro*, it was our goal to determine the potential of this compound as an antimalarial drug candidate. Thus, we carried out inhibition studies with the purified PfDxr target enzyme. For comparison, inhibition constants for both fosmidomycin and FR-900098 were first determined (Table 2.2). To calculate these values, rate measurements were made for multiple substrate concentrations at multiple concentrations of each inhibitor. Double-reciprocal plots for each inhibitor showed intersection at the vertical axis, indicating competitive inhibition made manifest by an increase in K_m only. From the measured K_m values, the inhibition constants for fosmidomycin and FR-900098 were determined to be 7.5 nM and 3.7 nM, respectively. These values correlate very well with the *in vivo* IC_{50} values measured against both the HB3 strain (350 nM for fosmidomycin, 170 nM for FR-900098) and A2 strain (370 nM for fosmidomycin, 170 nM for FR-900098) of *P. falciparum* [27]. Both values are significantly lower than the 38 nM inhibition constant reported for EcDxr, and are more comparable to the values reported for *Synechocystis sp.* PCC6803 (4 nM for fosmidomycin, 0.9 nM for FR-900098) [63,64]. After we had carried out these measurements, Behrendt and co-workers published very similar PfDxr inhibition constants of 8.4 nM for fosmidomycin and 2.6 nM for FR-900098, which further justify the validity of our inhibition assay and the results derived therefrom [50].

Preparation of pure FR-900098P for inhibition studies was performed from the *in vitro* reactions. The amount of product generated was quantified by weighing the lyophilized product, and was further checked via measurement of the LC-MS extracted ion peak area and by comparison to the concentration of the CMP-conjugated intermediate prior to treatment with the nucleotide hydrolase FrbI, which could be quantified via measurement of the UV absorbance peak area at 254 nm. Kinetic measurements at multiple inhibitor concentrations yielded a sub-nanomolar inhibition constant of 0.92 nM, indicating a significantly more potent inhibitor than fosmidomycin and FR-900098. This result correlates very well with the observation of a hydrophobic binding pocket in the PfDxr crystal structures.

Table 2.2: PfDxr inhibition constants (mean ± standard deviation of three replicates)

Inhibitor	fosmidomycin	FR-900098	FR-900098P
K_i**, nM**	7.5 ± 2.0	3.7 ± 0.9	0.92 ± 0.19

2.2.7 An *in vivo* platform for FR-900098P biosynthesis

Given the improved potency of FR-900098P as a PfDxr inhibitor, we sought to develop an *in vivo* biosynthetic platform, as *in vitro* enzymatic synthesis with purified enzymes is not suitable for large-scale production given the cost of protein purification and of the substrate materials. Similarly, chemical synthesis often requires expensive, hazardous, or non-renewable raw materials, and can also require harsh operating conditions. Microbial biosyntheses, in contrast, can be performed using renewable biomass as a primary feedstock, and require only very mild operating conditions. Additionally, biosynthetic machinery can access complicated reactions and specific stereochemistry that would otherwise be very difficult or even impossible to achieve by synthetic means [65].

To develop a biosynthetic chassis for FR-900098P production, we built upon an *E. coli* FR-900098 production platform previously developed in the Zhao lab [53]. We first attempted to identify FR-900098P from the FR-900098 producing strain directly, given that *E. coli* can naturally produce propionyl-CoA and that the wild type FrbF enzyme can accept propionyl-CoA as a substrate. However, no FR-900098P was detected, likely indicating that intracellular propionyl-CoA concentrations are too low to compete with acetyl-CoA in FrbF-catalyzed acyl transfer. To overcome this problem, we investigated a mutasynthetic approach by which sodium propionate was fed to the culture at the time of induction. Once taken up by the cell, propionate can be converted to propionyl-CoA via the propionyl-CoA synthetase gene *prpE*, which can then be utilized by FrbF. As shown in Figure 2.13a, this approach proved to be successful, and FR-900098P was detected in the cell culture broth by MS/MS. FR-900098, however, was still the dominant phosphonate product, and FR-900098P represented only ~5 – 8 % of the total phosphonate yield.

In an attempt to further increase FR-900098P production *in vivo* using the native FR-900098 biosynthetic machinery, a metabolically engineered *E. coli* host was investigated. In general, to increase flux through a secondary metabolite pathway, a driving force needs to be created. One method by which to accomplish this goal is to engineer the host's metabolism to over-produce a key substrate for a rate-limiting step. In the case of FR-900098P, it is likely that the FrbF-catalyzed acylation is slow, given the relatively low activity observed with the non-native substrate. To overcome this, then, propionyl-CoA concentrations in the cell can be increased via two modifications. First, over-expression of *prpE* can increase the rate at which propionyl-CoA is synthesized. Second, deletion of the *prpRBCD* operon prevents propionyl-CoA from being

consumed in primary metabolism via conversion to succinate and pyruvate. Both of these modifications have been reported in the literature and successfully utilized in the heterologous production of secondary metabolites [66]. As a result, we obtained this engineered strain, called BAP1, from the Blaine Pfeifer lab at Tufts University. To engineer the BAP1 strain for FR-900098P production, it was simultaneously transformed with the three compatible plasmids containing the FR-900098 pathway via electroporation. Following selection on LB media with three antibiotics, a colony was isolated containing all three plasmids. Production of FR-900098P in this host was analyzed under identical growth conditions as the BL21 host. Relative to cell culture density, the FR-900098P titer increased 3.5-fold in the BAP1 strain as compared to the BL21 strain (Figure 2.13b). However, it was observed that the BAP1 strain did not reach as high a density as the BL21 strain, resulting in an overall increase in endpoint FR-900098P concentration of only ~50 %. This is likely due to the increased metabolic burden placed on the BAP1 strain by over-expressing *prpE* under a T7 promoter while simultaneously limiting its ability to utilize the exogenous propionate for central metabolism. As observed in the BL21 strain, the BAP1 strain also produced FR-900098 as its primary phosphonate product. However, the relative amount of FR-900098P was significantly greater in BAP1, comprising ~22 – 25 % of the total phosphonate yield.

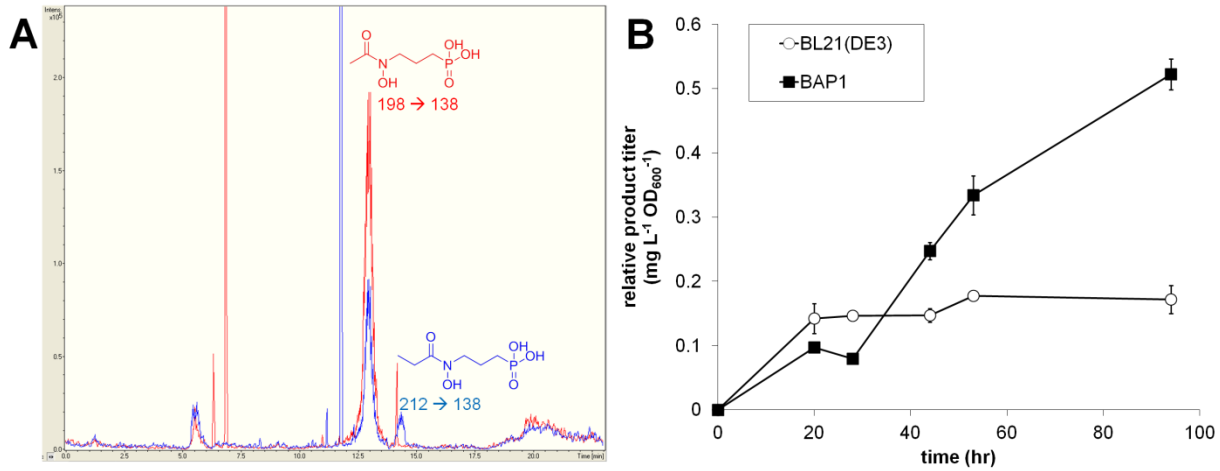


Figure 2.13: *In vivo* biosynthesis of FR-900098P. (a) MS trace showing phosphonates produced (FR-900098 = 198^{+}, FR-900098P = 212^{+}) in the presence (blue, lower trace) or absence (red, upper trace) of exogenous sodium propionate. (b) Comparison of relative FR-900098P concentrations produced in BL21 and BAP1.

2.2.8 Directed evolution of FrbF

2.2.8.1 Screening method

While we were able to demonstrate FR-900098P synthesis *in vitro* and *in vivo*, it is clear that wild type FrbF is not optimized for utilization of non-native acyl-donors. As a result, *in vitro* synthetic yields are poor, and *in vivo* the native acetyl-CoA substrate out-competes the desired non-native substrate. As a result, it would be desirable to identify mutants of FrbF that demonstrate increased activity toward non-native acyl-CoA substrates.

To achieve this goal, we developed a screening method to detect FrbF activity in cell lysates in a high-throughput manner. This method relies on 5,5'-dithiobis-(2-nitrobenzoic acid), known as DTNB or Ellman's reagent, to detect the free coenzyme A product of the acyltransfer reaction [67]. Following a FrbF catalytic cycle, coenzyme A is generated, which possesses a free sulfhydryl terminus. DTNB can participate in a rapid, non-enzymatic disulfide exchange with

this product, liberating 2-nitro-5-thiobenzoic acid (TNB). TNB has a large extinction coefficient at a wavelength of 412 nm, and as a result its formation can readily be measured. The overall flowchart of the FrbF screening procedure is outlined in Figure 2.14.

To quantify FrbF activity, cells were grown and lysed in microtiter plates, and lysate was subsequently split into two assay plates for initial and final quantification with DTNB. To maximize signal-to-noise ratio, a 60 minute assay was developed in which the native hydroxylamine substrate CMP-5'-H3APn was generated *in situ* via addition of FrbG to the assay mixture. With this method, variation in output signal was found to be very low for wild type FrbF with acetyl-CoA or propionyl-CoA (~6 % relative to the mean), with a sufficient signal-to-noise level of ~4 – 7 depending on the substrate.

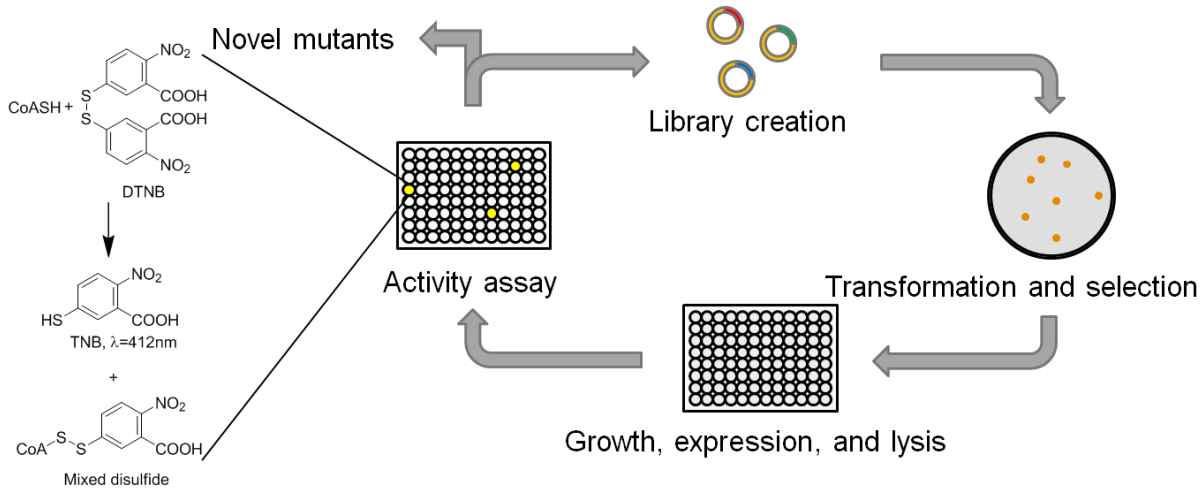


Figure 2.14: Flowchart of screening method for detection of FrbF activity toward novel substrates. A library of FrbF mutant constructs is used to transform *E. coli* BL21(DE3). Individual colonies are grown, induced, and lysed in a microtiter plate. Following addition of lysate to the FrbF assay mixture, TNB formation is monitored at λ = 412 nm to detect FrbF activity.

2.2.8.2 Library creation by saturation mutagenesis

To limit the size of the library of FrbF mutants that would need to be screened, we first selected a site-directed mutagenesis approach to target only those residues that might be actively involved in blocking the incorporation of a larger acyl-CoA substrate. Having determined the crystal structure of FrbF in complex with acetyl-CoA, we selected for saturation mutagenesis those residues located within 6 Å of the substrate acetyl group, with the exception of Thr-190 and His-193, which were found to be catalytically requisite. The resulting sites targeted for saturation mutagenesis were Phe-80, Glu-187, Ser-188, Asn-189, Glu-231, and Phe-234. Degenerate primers were designed at each of these sites such that a saturation mutagenesis library could be constructed via megaprimer PCR. Following library transformation, multiple colonies from the Phe-80, Glu-231, and Phe-234 libraries were chosen and sequenced to confirm mutation only at the desired site. The results of the sequencing indicated a different codon in each construct successfully sequenced with no off-target mutations.

For each of the saturation mutagenesis libraries generated, one microtiter plate was screened, representing ~5-fold coverage of the sequence space. In all of the six libraries, few colonies significantly exceeded the signal output of the control strain expressing wild type FrbF. Re-screening of top hits showed at best modest improvement relative to the wild type control. However, sequencing revealed all of the top hits to contain the wild type *frbF* sequence. SDS-PAGE analysis revealed that the best hits produced a greater output signal not due to mutation of the protein, but due to higher protein expression levels.

2.2.8.3 Library creation by error-prone PCR

To expand the sequence space of FrbF mutants available for screening, error-prone PCR was employed. Using *Taq* polymerase, a biased nucleotide mixture, and various concentrations of $MnCl_2$, mutation rates ranging from 1.7 amino acids per protein to 3.4 amino acids per protein were achieved, as determined by sequencing ten clones each from three different libraries. A library with an average mutation rate of 2.0 amino acids per protein was chosen for screening. From this library, ~1000 mutants were screened. Interestingly, many mutants in this library were found to generate only background-level signal. Of the few hits obtained, the best contained single amino acid changes of Asp-98→Gly and Arg-250→Glu, but the improvement in activity was found to be minimal (< 15 %) when rescreened. Despite the limited improvement achieved, the developed screening method could potentially be useful for screening of FrbF with other substrates, for the improvement of other properties of FrbF, or for the evolution of other *N*-acetyltransferases.

2.3 Conclusions and Outlook

In this study, we biochemically characterized the *N*-acetyltransferase FrbF from the FR-900098 biosynthetic cluster. Kinetic parameters were determined with acetyl-CoA, the native acyl-donor, and CMP-5'-3APn, a surrogate phosphonate substrate. Additionally, we identified a rapid, non-enzymatic oxidation of the native hydroxylamine substrate, CMP-5'-H3APn, to its nitroso derivative. Although this prevented the measurement of kinetic parameters with this substrate, we nevertheless demonstrated via *in situ* generation of CMP-5'-H3APn that this compound is in fact the preferred substrate for FrbF. Further, we demonstrated that FrbF exhibits reasonably high substrate specificity for acetyl-CoA as its acyl-donor, as relative activity

levels fell significantly with increasing substrate size. For propionyl-CoA, this was due predominantly to a decrease in k_{cat}, indicating that it is likely hindrance of attack by the nucleophilic substrate, and not of binding, that lowers activity with novel acyl-donors. To elucidate the roles of key active site residues in catalysis, we employed a site-directed mutagenesis approach.

Using the knowledge gained here, we synthesized the propionyl derivative of FR-900098, called FR-900098P, both through *in vitro* enzymatic reactions and through *in vivo* transformations. For the latter approach, a metabolically engineered strain was employed along with a mutasynthetic strategy to boost production of the desired FR-900098 analog. To assess the antimalarial potential of FR-900098P, we expressed and characterized DXP reductoisomerase, its target enzyme from *P. falciparum*. Inhibition constants were determined for fosmidomycin, FR-900098, and FR-900098P, revealing the latter to be the superior inhibitor. Finally, a colorimetric screening method was developed to identify FrbF mutants with improved activity.

Moving forward, chemical modification of FR-900098P could be employed to generate a more potent drug candidate for *in vivo* testing. Addition of ester groups to mask the phosphonic acid moiety, as has been demonstrated with fosmidomycin and FR-900098, would be expected to yield similar improvements in efficacy. Beyond FR-900098P, the screening assay developed here can be employed to pursue other FR-900098 derivatives with other acyl substituents, as detection relies only on the formation of the free coenzyme A product. The best FrbF mutants can then be integrated into the FR-900098 gene cluster to improve heterologous production.

Additionally, FR-900098 pathway optimization efforts are actively being pursued in the Zhao lab. The current heterologous FR-900098 pathway in *E. coli* utilizes a strong T7 promoter for every pathway gene, which introduces a heavy metabolic burden on the cell. To alleviate this burden, the expression levels of each gene need to be balanced. Further metabolic engineering is also being investigated to increase overall flux through the FR-900098 synthetic pathway. The results of both of these approaches will be readily applicable to the production of novel FR-900098 derivatives. Finally, characterization of FrbF, a representative of a new class of *N*-acetyltransferases, may facilitate future studies of previously uncharacterized members of this family, which could lead to the discovery of new and interesting secondary metabolite gene clusters.

2.4 Materials and Methods

2.4.1 Strains, media and reagents

The *Escherichia coli* cloning strain DH5α, expression strain BL21(DE3) and the expression vector pET28a were obtained from EMD Biosciences (San Diego, CA). Antibiotics and isopropyl-β-D-1-thiogalactopyranoside (IPTG) were purchased from Gold Biotechnology (St. Louis, MO). All chemicals, including reaction buffers and the reagents samarium diiodide, cytidine triphosphate (CTP), 2-amino-4-phosphonobutyrate (2APn), and all acyl-CoAs, were purchased from Sigma-Aldrich (St. Louis, MO) except for the co-factor NADPH, which was purchased from Applichem (Darmstadt, Germany), and 1-deoxy-D-xylulose 5-phosphate (DXP), which was purchased from Echelon Biosciences (Salt Lake City, UT). Organic solvents were purchased from Thermo-Fisher Scientific (Pittsburgh, PA). All PCR reagents and restriction enzymes were purchased from New England Biolabs (Ipswich, MA), and DNA Miniprep and

Gel Purification Kits were purchased from Qiagen (Valencia, CA). All primers were synthesized by Integrated DNA Technologies (Coralville, IA). Talon Cobalt immobilized metal affinity chromatography (IMAC) resin was purchased from Clontech Laboratories (Mountain View, CA). All other materials were purchased from Thermo-Fisher Scientific (Pittsburgh, PA).

2.4.2 Protein expression and purification

The *E. coli* BL21(DE3) strains containing pET28a-FrbF, pET28a-FrbG, and pET28a-FrbH were previously prepared via ligation of the respective ORFs between the *Nde*I and *Hin*dIII restriction sites. The complete *dxr* gene from *Plasmodium falciparum* was codon-optimized for expression in *E. coli* by DNA 2.0 (Menlo Park, CA). The core enzyme sequence (without nucleotides 1-216, corresponding to the 72 amino acid leader peptide) was PCR amplified with added *Nde*I and *Hind*III restriction sites and cloned into the respective sites of vector pET28a, introducing an *N*-terminal His_6 tag to the protein. *E. coli* BL21(DE3) was transformed with the resulting construct by electroporation. Cells were grown in Terrific Broth (TB) media supplemented with kanamycin (50 μg/mL) at 37 °C to an OD_{600} of ~0.8, after which induction was carried out by addition of 0.3 mM IPTG (for *frb* constructs) or 0.5 mM IPTG (for *dxr* construct) at 25 °C. In the case of FrbH, IPTG induction was accompanied by addition of 50 μg/mL pyridoxal-5'-phosphate (PLP). After 18 hr, the cells were harvested by centrifugation at 7500 rpm for 15 min and resuspended in 20 mM Tris-HCl (pH 7.65), 0.5 M NaCl, and 15 % glycerol supplemented with 1 mg/mL lysozyme. After a freeze-thaw cycle at -80 °C, the cell suspension was sonicated to ensure sufficient lysis. The lysate was clarified multiple times by centrifugation at 15000 rpm for 15 min, after which the His_6-tagged proteins were purified by affinity chromatography on TALON Superflow Co^{2+} resin coupled to fast-performance liquid chromatography. The eluted

proteins were washed three times in 50 mM HEPES (pH 7.25), concentrated, and stored in 15 % glycerol at -80 °C.

2.4.3 Generation of mutants

Mutants of the FrbF and PfDxr enzymes were generated by the megaprimer PCR method. Briefly, primers were designed at the site of interest containing the desired mutation such that a larger primer (a "megaprimer") containing the mutation could be generated by PCR, using the regular forward or reverse primer as the second primer (whichever yielded the shortest PCR product). Next, PCR amplification of the full-length mutant gene was carried out using the megaprimer at one end of the gene and the regular primer at the opposite end. Finally, the full length gene containing the mutation was digested with *Nde*I and *Hind*III and ligated into the similarly digested pET28a vector backbone. The ligation product was then used to transform the *E. coli* cloning strain DH5α. Transformants were picked and grown in 5 mL LB + 50 μg/mL kanamycin. Their plasmids were isolated and sequenced by the UIUC Core Sequencing Facility (Urbana, IL) or ACGT, Inc. (Wheeling, IL), and Sequencher software (Gene Codes Corporation, Ann Arbor, MI) was used to confirm the presence of only the desired mutation. The correct plasmids were then used to transform the expression strain *E. coli* BL21(DE3), and frozen stocks of the resultant strains were stored in 15 % glycerol at -80 °C.

2.4.4 Preparation of CMP-5'-3APn

CMP-5'-3APn for kinetic and relative activity assays was synthesized enzymatically using purified N-His_6-FrbH. Synthesis was carried out in 50 mM HEPES buffer (pH 7.25) with 1.5 mM CTP, 3 mM 2APn, 10 mM $MgCl_2$, and ~10 μM FrbH at 30 °C for 4 hr. Following removal

of FrbH with a 10 kDa cutoff Amicon centrifugal filter unit (Millipore, Billerica, MA), the product was purified by fractionation on an Agilent 1100 Series HPLC (Agilent, Palo Alto, CA) using an Alltech Prevail C18 reverse phase column (Grace Davison Discovery Sciences, Deerfield, IL) with isocratic flow of 15 mM ammonium formate buffer. Fractions containing the product were pooled, lyophilized, and resuspended in double-distilled water. The concentration of the purified CMP-5'-3APn product was determined by UV absorbance at 254 nm.

2.4.5 CMP-5'-H3APn oxidation assays

Enzymatic synthesis of CMP-5'-H3APn was performed using purified N-His_6-FrbG. Reactions were carried out in 50 mM HEPES buffer (pH 7.25) with 500 µM CMP-5'-3APn, 500 µM NADPH, and ~30 µM FrbG at 30 °C for 15 min. Following removal of FrbG with a 10 kDa cutoff Amicon centrifugal filter unit, the product mixture was incubated at 30 °C. Samples were collected at specified time points and kept on dry ice until analysis using an Agilent XCT ion-trap MSD mass spectrometer at the Roy J. Carver Metabolomics Center (University of Illinois, Urbana, IL). Chemical synthesis of CMP-5'-H3APn was performed using the oxidation product CMP-5'-NO3APn, which was purified from FrbG reaction mixtures via HPLC fractionation. Reduction of the nitroso to the hydroxylamine was performed anaerobically following the method of Kende and Mendoza [68]. In an anaerobic chamber, the lyophilized nitroso compound was resuspended in a mixture of tetrahydrofuran (THF) and methanol (95%/5% v/v). To this solution was added a 4-fold molar excess of samarium diiodide in THF. Following agitation for 3 min at room temperature, the reaction was acidified with 0.1 % trifluoroacetic acid and extracted with ethyl acetate. The product solution was then analyzed by LC-MS both before and after aerobic exposure.

2.4.6 FrbF kinetic assays

All activity assays were performed with N-His_6-tagged FrbF in 50 mM HEPES buffer (pH 7.25) at 30 °C. For determination of kinetic parameters, the concentration of one substrate was maintained at 400 µM, while the other substrate concentration was varied from 0 – 1000 µM. Samples were collected over a 7 minute time course for product quantitation and determination of initial rate. Reactions were initiated with the addition of 1.3 µM FrbF and quenched with 1 % trifluoroacetic acid. All assays were performed in triplicate. Kinetic parameters were fit to the data using nonlinear least-squares regression in OriginPro 8 (OriginLab Corporation, Northampton, MA). Relative activity assays with alternate CoA substrates were performed with 400 µM CMP-5'-3APn and 200 µM of the CoA compound. Reactions were initiated with 15 µM FrbF, and samples were collected over a 60 minute time course. For substrates for which no activity was detected in this time, the assay time was extended up to 12 hr. For all assays, products were detected by UV absorbance at 254 nm on an Agilent 1100 Series HPLC (Agilent, Palo Alto, CA), with retention times confirmed using an Agilent XCT ion-trap MSD mass spectrometer at the Roy J. Carver Metabolomics Center (University of Illinois, Urbana, IL). For all LC-MS analyses with alternate CoA substrates, product identities were confirmed based on detection of the expected m/z parent ion with a corresponding m/z = 322^- (CMP) daughter ion in the MS/MS profile.

2.4.7 FrbF relative activity assays

Relative activity assays with purified N-His_6-tagged FrbF mutants were performed with 400 µM CMP-5'-3APn and 200 µM acetyl-CoA over a 7 minute time course. For those mutants that

exhibited significantly reduced activity, the assay time was extended up to 60 minutes. Samples were collected and analyzed as described above.

2.4.8 PfDxr kinetic assays

All kinetic assays were carried out with N-His_6-tagged PfDxr in 100 mM Tris-HCl buffer (pH 7.5). Determination of kinetic parameters was carried out in a total volume of 300 µL with 1 mM $MnCl_2$, 0.3 mM NADPH, and 50 nM enzyme while the DXP concentration was varied from 0 – 250 µM. Oxidation of NADPH was continuously monitored via absorbance at 340 nm in a Nanodrop 2000c spectrophotometer (Thermo Scientific, Wilmington, DE) at 37 °C for a maximum of 5 min. All assays were performed in triplicate. Michaelis-Menten kinetic parameters were fit to the data using OriginPro 8 software (OriginLab Corporation, Northampton, MA). Due to the instability of the PfDxr enzyme, a fresh aliquot was used for each independent assay.

2.4.9 PfDxr inhibition assays

All inhibition assays were carried out with N-His_6-tagged PfDxr in 100 mM Tris-HCl buffer (pH 7.5). To determine inhibition constants for fosmidomycin, FR-900098, and FR-900098P, kinetic assays were performed at multiple concentrations of each inhibitor ranging from 0 – 10 nM. At each inhibitor concentration, K_m values were determined, with reactions carried out at 37 °C in a total volume of 300 µL with 1 mM $MnCl_2$, 0.3 mM NADPH, and 50 nM enzyme while the DXP concentration was varied from 0 – 250 µM. From the measured K_m values, K_i values were determined using the equation for competitive inhibition. This procedure was repeated in triplicate for each inhibitor.

2.4.10 *In vitro* FR-900098P synthesis

Synthesis of FR-900098P *in vitro* was carried out with the purified N-His_6-tagged FrbF, FrbG, FrbH, and FrbI enzymes and commercially available substrates. First, cytidine monophosphate-5'-3-aminopropylphosphonate (CMP-5'-3APn) was synthesized in 50 mM HEPES buffer, pH 7.25, containing 1.5 mM cytidine triphosphate (CTP), 3 mM 2-amino-4-phosphonobutyrate (2APn), 10 mM $MgCl_2$, and ~10 µM FrbH with incubation at 30 °C for 4 hours, after which nearly complete conversion was observed. Following removal of FrbH with a 10 kDa cutoff filter, CMP-5'-FR-900098P was synthesized by addition of 1.5 mM NADPH, 1.5 mM propionyl-CoA, ~10 µM FrbF, and ~15 µM FrbG with incubation at 30 °C for 4 hours. At this point, side products such as CMP-5'-2-amino-4-phosphonobutyrate (CMP-5'-2APn), CMP-5'-N-propionyl-3-aminopropylphosphonate (CMP-5'-P3APn), CMP, CDP, and $NADP^+$ were removed via fractionation on an Agilent 1100 Series HPLC using an Alltech Prevail C18 column with isocratic 15 mM ammonium formate elution. The resulting CMP-5'-FR-900098P was lyophilized and resuspended in double-distilled H_2O. Finally, CMP cleavage was carried out in 10 mM Tris-HCl buffer, pH 7.5, by addition of ~10 µM FrbI with incubation at 30 °C for 4 hours. The final product titer was estimated from the concentration of CMP-5'-FR-900098P (measured by UV absorbance at 254 nm) and confirmed by dry weight of the final lyophilized product. Correct identity and purity of the final product was determined using an Agilent XCT ion-trap MSD mass spectrometer, where the anticipated m/z = 212^+ parent ion and m/z = 138^+ daughter ion were observed.

2.4.11 *In vivo* FR-900098P synthesis

Synthesis of FR-900098P *in vivo* was carried out in two different *E. coli* expression strains. The first was a BL21(DE3) strain transformed with three compatible plasmids containing the entire FR-900098 biosynthetic pathway except for the nucleotide hydrolase FrbI. This strain was previously prepared and described by Johannes and DeSieno, *et al*. [53]. The second FR-900098P expression strain was prepared from the metabolically engineered BAP1 strain provided by Dr. Blaine Pfeifer, which has been genetically modified to block conversion of propionyl-CoA to succinate and pyruvate for central metabolism [66]. This strain was transformed with the same three-plasmid FR-900098 production system. Successful transformation was confirmed by diagnostic PCR amplification of a pathway gene from each plasmid. Both strains were cultured in 50 mL LB medium supplemented with kanamycin (50 µg/mL), ampicillin (100 µg/mL), and chloramphenicol (25 µg/mL) at 37 °C to an OD_{600} of ~0.8. At this point, expression of the FR-900098 pathway enzymes was induced at 30 °C by addition of 0.5 mM IPTG, as well as 20 mM sodium propionate to increase propionyl-CoA synthesis. Samples of culture supernatant were collected at multiple time points up to 94 hr and clarified via centrifugation at 13,200 rpm for 10 min. Analysis of clarified samples was carried out on an Agilent XCT ion-trap MSD mass spectrometer with a 100 x 4.6 mm Synergi 4µ Fusion-RP 80A column (Phenomenex, Torrance, CA) using 0.1 % formic acid in H_2O (mobile phase A) and acetonitrile (mobile phase B) with the following elution program: 0 % B to 40 % B over 10 min; 40 % B to 100 % B over 3 min; and 100 % B for an additional 5 min. Products were identified in positive mode by their expected m/z values (198 for FR-900098, 212 for FR-900098P) with the corresponding m/z = 138 daughter ion.

2.4.12 FrbF saturation mutagenesis

The FrbF crystal structure was visualized using Molecular Operating Environment software (Chemical Computing Group, Montreal, Canada). Residues within 6 Å of the acetyl-CoA acyl group were selected as saturation mutagenesis candidates, with the exception of those identified as catalytically requisite. Saturation mutagenesis was carried out using primers containing a randomized codon at the site of interest, and the mutant FrbF constructs were generated using the megaprimer PCR method described above. Libraries of FrbF mutants were digested with the restriction enzymes NdeI and HindIII and ligated into the respective sites of the correspondingly digested pET28a vector. Following transformation of BL21(DE3) and selection on LB plates containing kanamycin (50 µg/mL), plasmids were isolated from multiple colonies in each library and sequenced to confirm diversity at the targeted amino acid position.

2.4.13 FrbF error-prone PCR

Error-prone PCR was performed in mutagenic buffer consisting of 7 mM $MgCl_2$, 50 mM KCl, 10 mM Tris, and 0.01 % w/v gelatin, adjusted to pH 8.3 with HCl. A biased nucleotide mixture comprised of 1 mM dCTP, 1 mM dTTP, 0.2 mM dATP, and 0.2 mM dGTP (final concentrations) was used in all PCR reactions, along with forward and reverse primers (0.5 µM each), the pET28a-*frbF* template (50 ng), *Taq* ligase (1 µL), and $MnCl_2$ ranging from 0.05 to 0.15 mM. Reactions were cycled 24 times, with a 58 °C annealing step and a 2 minute extension time in each cycle. The subsequent PCR product was gel-purified, digested with *Nde*I and *Hind*III, and ligated into vector pET28a. Ten clones from each library were picked for sequencing to estimate the frequency of mutation in the *frbF* gene.

2.4.14 FrbF screening method

Screening of FrbF mutant libraries was performed in a high-throughput microtiter plate assay. Using sterile toothpicks, individual colonies were picked from library transformation plates to the wells of a 96-well plate containing 100 µL of LB medium supplemented with kanamycin (50 µg/mL) per well. Following overnight growth to saturation at 37 °C, 5 µL from each well was used to inoculated 100 µL of fresh LB + kanamycin in a separate plate. This plate was then incubated at 37 °C until an average OD_{600} of ~1.2 was reached, at which point 100 µL of LB + kanamycin supplemented with 0.6 mM IPTG was added to each well, yielding a final concentration of 0.3 mM IPTG. Induced plates were incubated at 30 °C overnight to allow FrbF overexpression. Afterward, the cells were pelleted by centrifugation at 4000 rpm for 10 min and resuspended in 100 µL of 50 mM HEPES buffer (pH 7.25) supplemented with 1 mg/mL lysozyme. Following incubation at room temperature for 30 min, the lysis plate was subjected to a freeze-thaw cycle at -80 °C. Immediately before the assay was performed, the plate was thawed and the cell lysate clarified via centrifugation at 4000 rpm for 10 min at 4 °C. To perform the assay, 30 µL of lysate from each well was added to two separate 96-well plates. To the first plate was added 50 µL of 10 mM DTNB and 70 µL of 50 mM HEPES buffer (pH 7.25), and the absorbance at 412 nm was measured immediately. To the second plate was added 70 µL of the assay mixture, which contained (final concentrations): 400 µM CMP-5'-3APn, 400 µM NADPH, 300 µM of the alternate CoA substrate, and 4.8 µM FrbG in 50 mM HEPES buffer. The second plate was incubated at 30 °C for 60 min, after which 50 µL of 10 mM DTNB was added to each well and the absorbance at 412 nm was measured. The difference between the two measured absorbance values was taken as an estimate of the total amount of product generated.

2.5 References

1. World Health Organization. (2013) World Malaria Report 2013.
2. Klayman, D.L. (1985) Qinghaosu (artemisinin): an antimalarial drug from China. *Science*, **228**, 1049-1055.
3. Qinghaosu Antimalarial Coordinating Research Group. (1979) *Chin Med J*, **92**.
4. White, N.J. (2008) Qinghaosu (artemisinin): the price of success. *Science*, **320**, 330-334.
5. Avery, M.A., Jennings-White, C. and Chong, W.K.M. (1987) The total synthesis of (+)-artemisinin and (+)-9-desmethylartemisinin. *Tetrahedron Lett*, **28**, 4629-4632.
6. Avery, M.A., Chong, W.K.M. and Jennings-White, C. (1992) Stereoselective total synthesis of (+)-artemisinin, the antimalarial constituent of *Artemisia annua L*. *J Am Chem Soc*, **114**, 974-979.
7. Schmid, G. and Hofheinz, W. (1983) Total synthesis of qinghaosu. *J Am Chem Soc*, **105**, 624-625.
8. Zhou, W.S. (1986) Total synthesis of arteannuin (qinghaosu) and related compounds. *Pure Appl Chem*, **58**.
9. Martin, V.J., Pitera, D.J., Withers, S.T., Newman, J.D. and Keasling, J.D. (2003) Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat Biotechnol*, **21**, 796-802.
10. Ro, D.K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J. *et al.* (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, **440**, 940-943.

11. Paddon, C.J., Westfall, P.J., Pitera, D.J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M.D., Tai, A., Main, A., Eng, D. *et al.* (2013) High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, **496**, 528-532.

12. Sanofi. (2013) Sanofi and PATH announce the launch of large-scale production of semisynthetic artemisinin againts malaria. *Press Release*, April 11, 2013.

13. Jambou, R., Legrand, E., Niang, M., Khim, N., Lim, P., Volney, B., Ekala, M.T., Bouchier, C., Esterre, P., Fandeur, T. *et al.* (2005) Resistance of *Plasmodium falciparum* field isolates to *in-vitro* artemether and point mutations of the SERCA-type PfATPase6. *Lancet*, **366**, 1960-1963.

14. Noedl, H., Se, Y., Schaecher, K., Smith, B.L., Socheat, D. and Fukuda, M.M. (2008) Evidence of artemisinin-resistant malaria in western Cambodia. *N Engl J Med*, **359**, 2619-2620.

15. Dondorp, A.M., Nosten, F., Yi, P., Das, D., Phyo, A.P., Tarning, J., Lwin, K.M., Ariey, F., Hanpithakpong, W., Lee, S.J. *et al.* (2009) Artemisinin resistance in *Plasmodium falciparum* malaria. *N Engl J Med*, **361**, 455-467.

16. Kamiya, T., Hemmi, K., Takeno, H. and Hashimoto, M. (1980) Studies on phosphonic acid antibiotics. I. Structure and synthesis of 3-(N-acetyl-N-hydroxyamino)propylphosphonic acid (FR-900098) and its N-formyl analogue (FR-31564). *Tetrahedron Lett*, **21**, 95-98.

17. Okuhara, M., Kuroda, Y., Goto, T., Okamoto, M., Terano, H., Kohsaka, M., Aoki, H. and Imanaka, H. (1980) Studies on new phosphonic acid antibiotics. III. Isolation and characterization of FR-31564, FR-32863 and FR-33289. *J Antibiot (Tokyo)*, **33**, 24-28.

18. Murakawa, T., Sakamoto, H., Fukada, S., Konishi, T. and Nishida, M. (1982) Pharmacokinetics of fosmidomycin, a new phosphonic acid antibiotic. *Antimicrob Agents Chemother*, **21**, 224-230.

19. Neuman, M. (1984) Recent developments in the field of phosphonic acid antibiotics. *J Antimicrob Chemother*, **14**, 309-311.

20. Kuemmerle, H.P., Murakawa, T., Sakamoto, H., Sato, N., Konishi, T. and De Santis, F. (1985) Fosmidomycin, a new phosphonic acid antibiotic. Part II: 1. Human pharmacokinetics. 2. Preliminary early phase IIa clinical studies. *Int J Clin Pharmacol Ther Toxicol*, **23**, 521-528.

21. Kuemmerle, H.P., Murakawa, T., Soneoka, K. and Konishi, T. (1985) Fosmidomycin: a new phosphonic acid antibiotic. Part I: Phase I tolerance studies. *Int J Clin Pharmacol Ther Toxicol*, **23**, 515-520.

22. Kuemmerle, H.P., Murakawa, T. and De Santis, F. (1987) Pharmacokinetic evaluation of fosmidomycin, a new phosphonic acid antibiotic. *Chemioterapia*, **6**, 113-119.

23. Cheoymang, A., Hudchinton, D., Kioy, D. and Na-Bangchang, K. (2007) Bioassay for determination of fosmidomycin in plasma and urine: application for pharmacokinetic dose optimisation. *J Microbiol Methods*, **69**, 65-69.

24. Rohmer, M., Knani, M., Simonin, P., Sutter, B. and Sahm, H. (1993) Isoprenoid biosynthesis in bacteria: a novel pathway for the early steps leading to isopentenyl diphosphate. *Biochem J*, **295 (Pt 2)**, 517-524.

25. Shigi, Y. (1989) Inhibition of bacterial isoprenoid synthesis by fosmidomycin, a phosphonic acid-containing antibiotic. *J Antimicrob Chemother*, **24**, 131-145.

26. Kuzuyama, T., Shimizu, T., Takahashi, S. and Seto, H. (1998) Fosmidomycin, a specific inhibitor of 1-deoxy-D-xylulose-5-phosphate reductoisomerase in the nonmevalonate pathway for terpenoid biosynthesis. *Tetrahedron Lett*, **39**, 7913-7916.

27. Jomaa, H., Wiesner, J., Sanderbrand, S., Altincicek, B., Weidemeyer, C., Hintz, M., Turbachova, I., Eberl, M., Zeidler, J., Lichtenthaler, H.K. *et al.* (1999) Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science*, **285**, 1573-1576.

28. Missinou, M.A., Borrmann, S., Schindler, A., Issifou, S., Adegnika, A.A., Matsiegui, P.B., Binder, R., Lell, B., Wiesner, J., Baranek, T. *et al.* (2002) Fosmidomycin for malaria. *Lancet*, **360**, 1941-1942.

29. Lell, B., Ruangweerayut, R., Wiesner, J., Missinou, M.A., Schindler, A., Baranek, T., Hintz, M., Hutchinson, D., Jomaa, H. and Kremsner, P.G. (2003) Fosmidomycin, a novel chemotherapeutic agent for malaria. *Antimicrob Agents Chemother*, **47**, 735-738.

30. Borrmann, S., Adegnika, A.A., Matsiegui, P.B., Issifou, S., Schindler, A., Mawili-Mboumba, D.P., Baranek, T., Wiesner, J., Jomaa, H. and Kremsner, P.G. (2004) Fosmidomycin-clindamycin for *Plasmodium falciparum* infections in African children. *J Infect Dis*, **189**, 901-908.

31. Borrmann, S., Issifou, S., Esser, G., Adegnika, A.A., Ramharter, M., Matsiegui, P.B., Oyakhirome, S., Mawili-Mboumba, D.P., Missinou, M.A., Kun, J.F. *et al.* (2004) Fosmidomycin-clindamycin for the treatment of *Plasmodium falciparum* malaria. *J Infect Dis*, **190**, 1534-1540.

32. Borrmann, S., Lundgren, I., Oyakhirome, S., Impouma, B., Matsiegui, P.B., Adegnika, A.A., Issifou, S., Kun, J.F., Hutchinson, D., Wiesner, J. *et al.* (2006) Fosmidomycin plus

clindamycin for treatment of pediatric patients aged 1 to 14 years with *Plasmodium falciparum* malaria. *Antimicrob Agents Chemother*, **50**, 2713-2718.

33. Borrmann, S., Adegnika, A.A., Moussavou, F., Oyakhirome, S., Esser, G., Matsiegui, P.B., Ramharter, M., Lundgren, I., Kombila, M., Issifou, S. *et al.* (2005) Short-course regimens of artesunate-fosmidomycin in treatment of uncomplicated *Plasmodium falciparum* malaria. *Antimicrob Agents Chemother*, **49**, 3749-3754.

34. Hemmi, K., Takeno, H., Hashimoto, M. and Kamiya, T. (1982) Studies on phosphonic acid antibiotics. IV. Synthesis and antibacterial activity of analogs of 3-(N-acetyl-N-hydroxyamino)-propylphosphonic acid (FR-900098). *Chem Pharm Bull (Tokyo)*, **30**, 111-118.

35. Öhler, E. and Kanzler, S. (1995) Regioselective palladium(0) catalyzed amination of carbonates of allylic α-hydroxyphosphonates with hydroxylamine derivatives: a convenient route to phosphonic acids related to the antibiotic fosmidomycin. *Synthesis*, **5**, 539-543.

36. Wiesner, J., Borrmann, S. and Jomaa, H. (2003) Fosmidomycin for the treatment of malaria. *Parasitol Res*, **90 Suppl 2**, S71-76.

37. Reichenberg, A., Wiesner, J., Weidemeyer, C., Dreiseidler, E., Sanderbrand, S., Altincicek, B., Beck, E., Schlitzer, M. and Jomaa, H. (2001) Diaryl ester prodrugs of FR900098 with improved *in vivo* antimalarial activity. *Bioorg Med Chem Lett*, **11**, 833-835.

38. Ortmann, R., Wiesner, J., Reichenberg, A., Henschker, D., Beck, E., Jomaa, H. and Schlitzer, M. (2003) Acyloxyalkyl ester prodrugs of FR900098 with improved *in vivo* anti-malarial activity. *Bioorg Med Chem Lett*, **13**, 2163-2166.

39. Ortmann, R., Wiesner, J., Reichenberg, A., Henschker, D., Beck, E., Jomaa, H. and Schlitzer, M. (2005) Alkoxycarbonyloxyethyl ester prodrugs of FR900098 with improved *in vivo* antimalarial activity. *Arch Pharm (Weinheim)*, **338**, 305-314.

40. Kurz, T., Schluter, K., Kaula, U., Bergmann, B., Walter, R.D. and Geffken, D. (2006) Synthesis and antimalarial activity of chain substituted pivaloyloxymethyl ester analogues of fosmidomycin and FR900098. *Bioorg Med Chem*, **14**, 5121-5135.

41. Schluter, K., Walter, R.D., Bergmann, B. and Kurz, T. (2006) Arylmethyl substituted derivatives of fosmidomycin: synthesis and antimalarial activity. *Eur J Med Chem*, **41**, 1385-1397.

42. Haemers, T., Wiesner, J., Van Poecke, S., Goeman, J., Henschker, D., Beck, E., Jomaa, H. and Van Calenbergh, S. (2006) Synthesis of alpha-substituted fosmidomycin analogues as highly potent *Plasmodium falciparum* growth inhibitors. *Bioorg Med Chem Lett*, **16**, 1888-1891.

43. Devreux, V., Wiesner, J., Jomaa, H., Rozenski, J., Van der Eycken, J. and Van Calenbergh, S. (2007) Divergent strategy for the synthesis of alpha-aryl-substituted fosmidomycin analogues. *J Org Chem*, **72**, 3783-3789.

44. Verbrugghen, T., Cos, P., Maes, L. and Van Calenbergh, S. (2010) Synthesis and evaluation of alpha-halogenated analogues of 3-(acetylhydroxyamino)propylphosphonic acid (FR900098) as antimalarials. *J Med Chem*, **53**, 5342-5346.

45. Verbrugghen, T., Vandurm, P., Pouyez, J., Maes, L., Wouters, J. and Van Calenbergh, S. (2013) Alpha-heteroatom derivatized analogues of 3-(acetylhydroxyamino)propyl phosphonic acid (FR900098) as antimalarials. *J Med Chem*, **56**, 376-380.

46. Devreux, V., Wiesner, J., Goeman, J.L., Van der Eycken, J., Jomaa, H. and Van Calenbergh, S. (2006) Synthesis and biological evaluation of cyclopropyl analogues of fosmidomycin as potent *Plasmodium falciparum* growth inhibitors. *J Med Chem*, **49**, 2656-2660.

47. Haemers, T., Wiesner, J., Busson, R., Jomaa, H. and Van Calenbergh, S. (2006) Synthesis of alpha-aryl-substituted and conformationally restricted fosmidomycin analogues as promising antimalarials. *European J Org Chem*, 3856-3863.

48. Devreux, V., Wiesner, J., Jomaa, H., Van der Eycken, J. and Van Calenbergh, S. (2007) Synthesis and evaluation of alpha,beta-unsaturated alpha-aryl-substituted fosmidomycin analogues as DXR inhibitors. *Bioorg Med Chem Lett*, **17**, 4920-4923.

49. Haemers, T., Wiesner, J., Giessmann, D., Verbrugghen, T., Hillaert, U., Ortmann, R., Jomaa, H., Link, A., Schlitzer, M. and Van Calenbergh, S. (2008) Synthesis of beta- and gamma-oxa isosteres of fosmidomycin and FR900098 as antimalarial candidates. *Bioorg Med Chem*, **16**, 3361-3371.

50. Behrendt, C.T., Kunfermann, A., Illarionova, V., Matheeussen, A., Grawert, T., Groll, M., Rohdich, F., Bacher, A., Eisenreich, W., Fischer, M. *et al.* (2010) Synthesis and antiplasmodial activity of highly active reverse analogues of the antimalarial drug candidate fosmidomycin. *ChemMedChem*, **5**, 1673-1676.

51. Behrendt, C.T., Kunfermann, A., Illarionova, V., Matheeussen, A., Pein, M.K., Grawert, T., Kaiser, J., Bacher, A., Eisenreich, W., Illarionov, B. *et al.* (2011) Reverse fosmidomycin derivatives against the antimalarial drug target IspC (Dxr). *J Med Chem*, **54**, 6796-6802.

52. Eliot, A.C., Griffin, B.M., Thomas, P.M., Johannes, T.W., Kelleher, N.L., Zhao, H. and Metcalf, W.W. (2008) Cloning, expression, and biochemical characterization of *Streptomyces rubellomurinus* genes required for biosynthesis of antimalarial compound FR900098. *Chem Biol*, **15**, 765-770.

53. Johannes, T.W., DeSieno, M.A., Griffin, B.M., Thomas, P.M., Kelleher, N.L., Metcalf, W.W. and Zhao, H. (2010) Deciphering the late biosynthetic steps of antimalarial compound FR-900098. *Chem Biol*, **17**, 57-64.

54. DeSieno, M.A., van der Donk, W.A. and Zhao, H. (2011) Characterization and application of the Fe(II) and alpha-ketoglutarate dependent hydroxylase FrbJ. *Chem Commun (Camb)*, **47**, 10025-10027.

55. Gießmann, D., Heidler, P., Haemers, T., Van Calenbergh, S., Reichenberg, A., Jomaa, H., Weidemeyer, C., Sanderbrand, S., Wiesner, J. and Link, A. (2008) Towards new antimalarial drugs: synthesis of non-hydrolyzable phosphate mimics as feed for a predictive QSAR study on 1-deoxy-D-xylulose-5-phosphate reductoisomerase inhibitors. *Chem Biodivers*, **5**, 643-656.

56. Vetting, M.W., LP, S.d.C., Yu, M., Hegde, S.S., Magnet, S., Roderick, S.L. and Blanchard, J.S. (2005) Structure and functions of the GNAT superfamily of acetyltransferases. *Arch Biochem Biophys*, **433**, 212-226.

57. Sim, E., Sandy, J., Evangelopoulos, D., Fullam, E., Bhakta, S., Westwood, I., Krylova, A., Lack, N. and Noble, M. (2008) Arylamine N-acetyltransferases in mycobacteria. *Curr Drug Metab*, **9**, 510-519.

58. Li, N., Korboukh, V.K., Krebs, C. and Bollinger, J.M., Jr. (2010) Four-electron oxidation of p-hydroxylaminobenzoate to p-nitrobenzoate by a peroxodiferric complex in AurF from *Streptomyces thioluteus*. *Proc Natl Acad Sci U S A*, **107**, 15722-15727.

59. Takahashi, S., Kuzuyama, T., Watanabe, H. and Seto, H. (1998) A 1-deoxy-D-xylulose 5-phosphate reductoisomerase catalyzing the formation of 2-C-methyl-D-erythritol 4-phosphate in an alternative nonmevalonate pathway for terpenoid biosynthesis. *Proc Natl Acad Sci U S A*, **95**, 9879-9884.

60. Grolle, S., Bringer-Meyer, S. and Sahm, H. (2000) Isolation of the *dxr* gene of *Zymomonas mobilis* and characterization of the 1-deoxy-D-xylulose 5-phosphate reductoisomerase. *FEMS Microbiol Lett*, **191**, 131-137.

61. Argyrou, A. and Blanchard, J.S. (2004) Kinetic and chemical mechanism of *Mycobacterium tuberculosis* 1-deoxy-D-xylulose-5-phosphate isomeroreductase. *Biochemistry*, **43**, 4375-4384.

62. Altincicek, B., Hintz, M., Sanderbrand, S., Wiesner, J., Beck, E. and Jomaa, H. (2000) Tools for discovery of inhibitors of the 1-deoxy-D-xylulose 5-phosphate (DXP) synthase and DXP reductoisomerase: an approach with enzymes from the pathogenic bacterium Pseudomonas aeruginosa. *FEMS Microbiol Lett*, **190**, 329-333.

63. Kuzuyama, T., Shimizu, T., Takahashi, S. and Seto, H. (1998) Fosmidomycin, a specific inhibitor of 1-deoxy-D-xylulose 5-phosphate reductoisomerase in the nonmevalonate pathway for terpenoid biosynthesis. *Tetrahedron Lett*, **39**, 7913-7916.

64. Woo, Y.H., Fernandes, R.P. and Proteau, P.J. (2006) Evaluation of fosmidomycin analogs as inhibitors of the *Synechocystis* sp. PCC6803 1-deoxy-D-xylulose 5-phosphate reductoisomerase. *Bioorg Med Chem*, **14**, 2375-2385.

65. Dhamankar, H. and Prather, K.L. (2011) Microbial chemical factories: recent advances in pathway engineering for synthesis of value added chemicals. *Curr Opin Struct Biol*, **21**, 488-494.

66. Pfeifer, B.A., Admiraal, S.J., Gramajo, H., Cane, D.E. and Khosla, C. (2001) Biosynthesis of complex polyketides in a metabolically engineered strain of *E. coli*. *Science*, **291**, 1790-1792.

67. Ellman, G.L. (1959) Tissue sulfhydryl groups. *Arch Biochem Biophys*, **82**, 70-77.

68. Kende, A.S. and Mendoza, J.S. (1991) Controlled reduction of nitroalkanes to alkyl hydroxylamines or amines by samarium diiodide. *Tetrahedron Lett*, **32**, 1699-1702.

CHAPTER 3. Cloning and expression of polyketide synthases from *Talaromyces marneffei*

3.1 Introduction

3.1.1 Fungal secondary metabolites

Fungal secondary metabolites are well-known for their diversity of bioactive properties, making them particularly attractive leads for drug development. Following the fortuitous discovery of penicillin from *Penicillium chrysogenum* as a potent antibiotic, research interest in fungal natural products has remained strong for several decades [1]. Many fungal natural products have had profound impacts in the human health sector, including the aforementioned penicillin and the related cephalosporin antibiotics; cyclosporine, widely used as an immunosuppressant as well as a treatment for dry eye syndrome; and lovastatin, a blockbuster cholesterol-lowering drug [2] (Figure 3.1). At the other end of the spectrum, fungal natural products can also have severe adverse health effects, such as the carcinogenic aflatoxins [3] and the nephrotoxic citrinin [4]. Thus, current research on fungal natural products focuses on understanding their biosynthesis and activity as both curative and causative agents.

Fungal natural products belong to diverse classes, including non-ribosomal peptides, assembled from amino acids by nonribosomal peptide synthetases (NRPSs); terpenes, assembled from activated isoprene precursors; and polyketides, assembled from acyl-Coenzyme A building blocks by polyketide synthases (PKSs) [1]. Although their structures vary widely, conserved motifs in some of the key enzymes involved in fungal natural product biosynthesis enable rapid identification of putative gene clusters from genome sequence data. NRPS and PKS

megasynthases, for example, feature well conserved adenylation domains and ketosynthase domains, respectively, while indole alkaloid clusters feature distinct dimethylallyltryptophan synthases (DMATSs) [5]. As a result, bioinformatics tools for the *in silico* analysis of natural product gene clusters are plentiful [6].

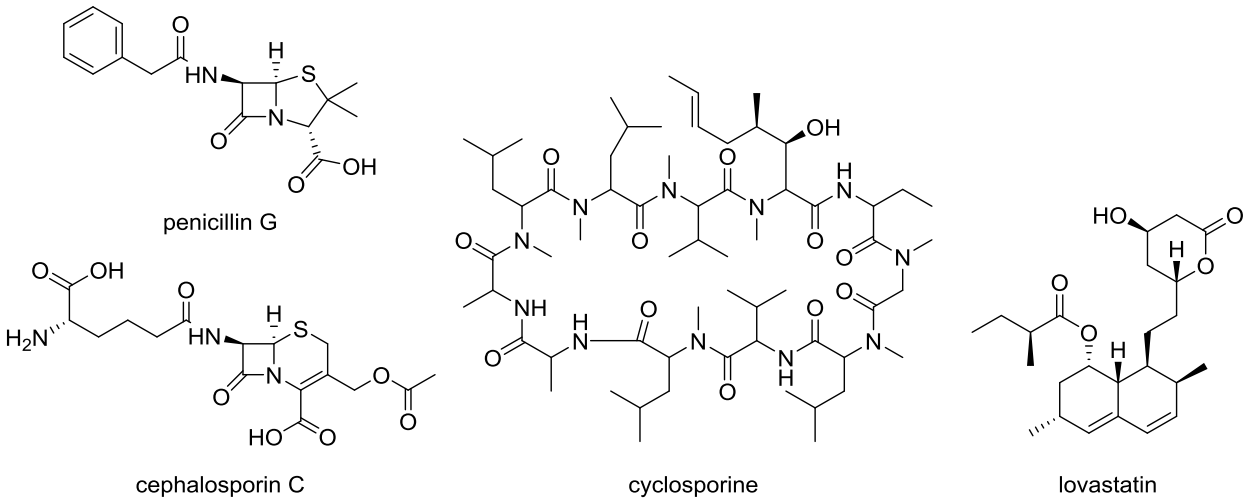


Figure 3.1: Representative fungal natural products with therapeutic properties.

3.1.2 Fungal genome mining

A key finding of genome mining efforts in fungi has been the great abundance of gene clusters, far in excess of any expectations based on previously isolated compounds from fungal strains. This is particularly true in certain fungal genera, such as the *Aspergilli*. *Aspergillus nidulans* [7], *Aspergillus oryzae* and *Aspergillus terreus* [5], for example, each feature more than 50 putative secondary metabolite gene clusters. Of note, the prediction of gene cluster boundaries by bioinformatics alone is difficult given the unknown functions of most genes in a sequenced genome. Recently, it has been demonstrated that microarray profiling of gene expression can greatly aid this process, as genes in the same pathway tend to exhibit similar expression patterns [8].

Regardless of the exact number of gene clusters in a given genome, the large numbers of genes encoding key synthetic enzymes suggests a great diversity of natural products. In *A. nidulans* alone, a total of 83 secondary metabolites have been discovered to date [7], with the vast majority discovered after the 2005 publication of the *A. nidulans* genome sequence [9]. Notably, many gene clusters produce several analogs of the same compound via differential decoration of a core scaffold, increasing the diversity of natural products synthesized even further. As an example, five emericellamide analogs have been observed from one gene cluster [10]. Nevertheless, more than 30 gene clusters in *A. nidulans* have still not been linked to any secondary metabolites, suggesting that there is still much left to uncover.

Outside of the well-studied *A. nidulans*, knowledge of the untapped potential of fungi for natural product synthesis has led to the application of a variety of techniques to elicit the expression of silent gene clusters. For example, epigenetic modifications of *Cladosporium cladosporioides* led to the activation of gene clusters for cladochromes, calphostin B, and oxylipins [11]. Co-culture with a bacterium elicited rhizoxin production from *Rhizopus microspores* [12]. To elicit production of sorbicillamines from marine fungus *Penicillium* sp. F23–2, culture agitation rate was varied [13]. Further examples in recent literature are too numerous to list.

Notably, all of the results described above depend on cultivation or even genetic manipulation of the fungal strain containing silent pathways of interest. However, not all fungi are as amenable to these approaches as, for example, the well-studied *Aspergilli*. Fungi occupying very specific niches in their natural environments may not be suitable for laboratory cultivation. For fungi exhibiting significant pathogenicity, laboratory cultivation may simply be undesirable. Further,

the lack of genetic tools available for most fungal strains limits the number of techniques that can be applied for natural product discovery to only the most non-specific, hit-or-miss approaches. Reconstitution of target gene clusters in genetically tractable model organisms stands as an alternative to cultivation or genetic manipulation of the native host.

The goal of the work presented here is to demonstrate heterologous expression in *S. cerevisiae* as a general strategy for the activation of natural product gene clusters from fungal species. Sequenced fungal genomes will be mined for candidate gene clusters, with a focus on those containing polyketide synthases. The chosen genes will be assembled in strong *S. cerevisiae* expression cassettes, and culture extracts will be analyzed by HPLC-MS/MS for the production of novel compounds. *In vitro* reconstitution will also be attempted for enzymes of interest.

3.2 Results and Discussion

3.2.1 Target identification and analysis

To identify interesting fungal polyketide gene clusters for refactoring, the SMURF program was employed [5]. A total of 27 sequenced genomes were initially studied. The results were first analyzed manually to identify clusters that had reasonably well-defined boundaries; did not contain too many genes, to facilitate efficient assembly; shared some similarity to characterized PKSs, such that product prediction would be feasible; and contained a different set of tailoring genes than any characterized PKS cluster, to increase the chance of discovering a novel product. Eventually, *Talaromyces marneffei* (formerly *Penicillium marneffei*) was chosen for more in-depth analysis, as its genome was relatively well annotated, and it contained a marked abundance of putative PKS genes (25 in total) [14]. Further, at the outset of the project, only one PKS

product, melanin, had been reported from *T. marneffei*. Thus, there exists a large potential space for polyketide discovery from this strain.

Table 3.1: Annotation of the *pks16*/*pks17* cluster from *T. marneffei*

Gene ID	Annotation, Domain Architecture	Predicted Introns (NCBI)
61670 (*pks17*)	HR-PKS, KS-AT-DH-ER-KR-ACP	4
61680	PKS fragment, ACP-TE only	--
61690	*O*-methyltransferase	2
61700	Dithiol-disulfide isomerase/DSBA-like thioredoxin	--
61710	FMO hydroxylase	6
61720 (*pks16*)	NR-PKS, KS-AT-ACP-TE	3
61730	Trp-halogenase	4
61740	Zn_2C_6 transcription factor	--
61750	MFS transporter	--
61760	Zn_2C_6 transcription factor	--

Among all the PKS gene clusters in the sequenced genome, the cluster containing genes *pks16* and *pks17* was first chosen for full-cluster refactoring. The putative gene cluster (Table 3.1) contains two polyketide synthases, one non-reducing (NR) and one highly-reducing (HR), expected to synthesize a resorcylic acid lactone (RAL) scaffold. In characterized gene clusters of this nature, the HR-PKS accepts standard CoA substrates (e.g., malonyl-CoA) to synthesize a reduced product that is subsequently passed to the NR-PKS for further elaboration and cyclization [15]. The majority of the characterized PKSs in this class produce benzenediol lactone (BDL) compounds, which feature a benzenediol moiety fused to a macrocycle of 12- to 14-members [16]. Among these, the RALs have been studied in the most detail [17], although dihydroxyphenylacetic acid lactones (DALs) have also been the subject of investigation [18]. RALs have garnered significant interest due to the diversity of bioactivities observed by the compounds sharing their unique pharmacophore (Figure 3.2).

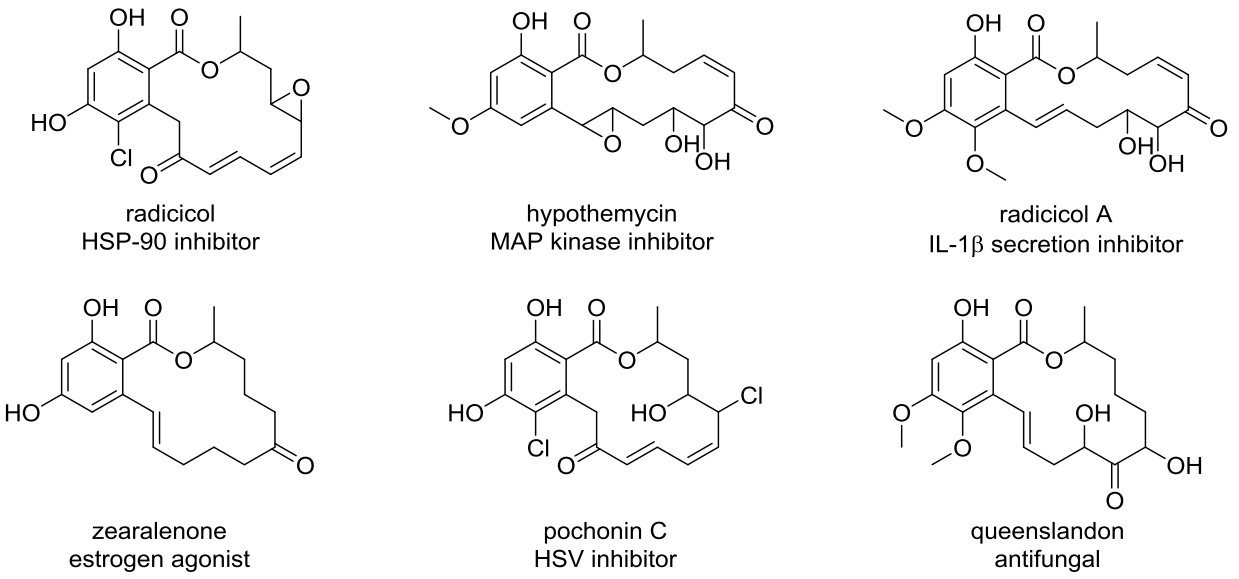


Figure 3.2: Structure and bioactivity of several known resorcylic acid lactones, adapted from ref. 17.

Bioinformatic analysis of the *T. marneffei pks16*/*pks17* cluster revealed relatively high homology (47 % amino acid identity, 64 % similarity) of PKS17 with the highly-reducing PKS of the radicicol gene cluster in *Chaetomium chiversii* [19]. Other characterized PKSs from RAL clusters occupied high positions in the BLAST results, all at ~ 40 % amino acid identity. Similar observations were made for the non-reducing PKS16.

Among the other genes in the cluster are three putative tailoring enzymes. PMAA_061690 was predicted to encode an *O*-methyltransferase. Although an analogous function is found in the hypothemycin gene cluster of *Hypomyces subiculosis* [20], similarity between the two encoded proteins was found to be low (23 % amino acid identity over 63 % query coverage). Gene PMAA_061710 was predicted to encode a flavin-dependent monooxygenase, while PMAA_061730 was predicted to encode a halogenase. Here, significant similarity (64 % at the amino acid level) was found between the putative halogenase and those of the characterized *C. chiversii* and *M. chlamydosporia* radicicol clusters [19,20]. Based on this analysis, the product

of the target cluster is anticipated to be a RAL similar to radicicol or hypothemycin. However, the particular set of tailoring genes identified here does not match those of the characterized RAL compounds, suggesting a new member of this class.

Also included in the cluster is PMAA_061680, which encodes only a phosphopantetheine-binding domain and a thioesterase. However, as PKS16 also includes a thioesterase domain, it is unclear as to whether PMAA_061680 would be necessary for product release. Gene PMAA_061700 encodes a putative dithiol-disulfide isomerase. Similar genes have been found in multiple actinomycete gene clusters and the *Aspergillus clavatus* cytochalasin gene cluster, and are thought to be involved in redox sensing [21-23]. However, their exact role in biosynthesis is unknown. Finally, two putative transcription factors are encoded in the gene cluster, along with a putative major facilitator superfamily transporter.

3.2.2 Refactoring the *pks16*/*pks17* putative resorcylic acid lactone cluster

3.2.2.1 Assembly and initial analysis

To refactor the putative RAL cluster from *T. marneffei* for expression in *Saccharomyces cerevisiae*, an assembly scheme was devised (Figure 3.3). First, individual open reading frames (including both introns and exons) were amplified from the genomic DNA of *T. marneffei*. Next, individual exons were amplified for assembly into single-gene expression cassettes with strong *S. cerevisiae* promoters [24] on a series of helper plasmids. For the PKS genes, assembly of the exons was carried out with the linearized helper plasmid via *in vivo* homologous recombination in yeast. For the tailoring genes, exons were first spliced by overlap-extension PCR, followed by assembly of the coding sequence with the linearized helper plasmid in yeast. Helper plasmids

were designed such that each expression cassette was immediately preceded by the terminator of the previous cassette. Thus, by PCR amplification of the terminator-promoter-gene-terminator cassettes from the five assembled helper plasmids, five linear fragments were generated with long overlaps for yeast *in vivo* assembly of the full five-gene cluster.

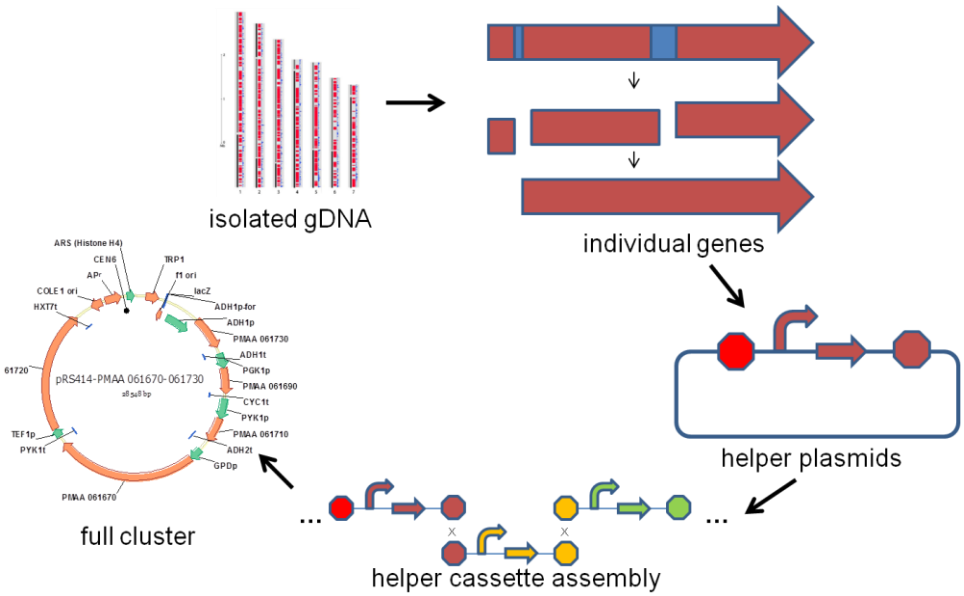


Figure 3.3: Cluster assembly scheme. Exons are amplified from genomic DNA, spliced into complete coding sequences, and assembled into helper plasmids. Single-gene cassettes are then amplified from helper plasmids and assembled into a full cluster plasmid.

Following cluster assembly and confirmation by restriction digestion and diagnostic PCR (data not shown), the full cluster was transformed into *S. cerevisiae* strain HZ848 along with the *Aspergillus nidulans* phosphopantetheinyl transferase gene *npgA* on plasmid pRS416. A negative control strain was also prepared with empty plasmid pRS414 and pRS416-*npgA*. Transformants were grown in both SC and YPAD medium. Culture extracts were analyzed by HPLC-MS for the presence of new UV peaks that eluted around the expected time for a RAL

compound (as assessed by similar analysis of commercially available RAL standards), including one that showed up in both media at 19.5min (Figure 3.4).

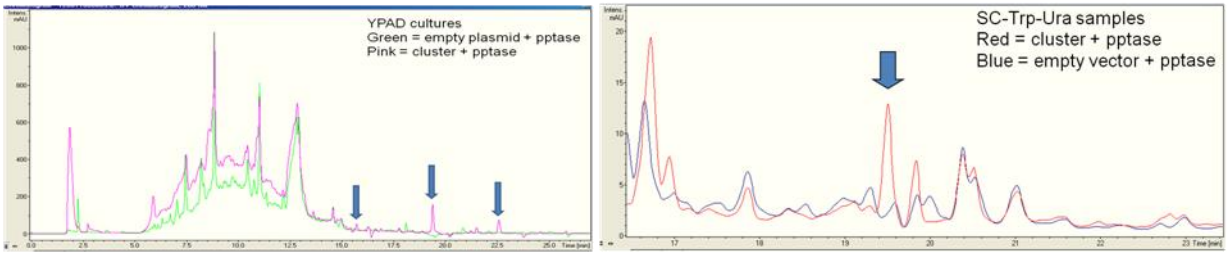


Figure 3.4: UV traces from samples grown in YPAD (left) and SC (right).

Isolation of the 19.5 min peak led to determination of a mass of 399.2 (398.2 in negative node, 400.2 in positive mode) and a chemical formula of $C_{24}H_{34}O_4N$. Notably, the predicted formula did not include an anticipated halogen atom, and instead included an unexpected nitrogen atom. However, although production of the peak of interest was observed in initial replicate cultures of the two-plasmid strain, subsequent replicates no longer showed production. After several troubleshooting measures, including reassembly of the cluster, retransformation of the plasmids, and transfer of the plasmid to a new *S. cerevisiae* host with integrated chromosomally integrated *npgA* (strain BJ5464-NpgA), the new peak could not be recovered. Addition of a sixth gene to the refactored cluster, PMAA_061680 (encoding a thioesterase) also did not lead to reproducible production of a new peak in the UV trace.

3.2.2.2 Confirmation at the transcriptional and translational levels

As a means of confirming the functionality of the refactored plasmid, diagnostic checks were employed at the transcriptional and translational levels. At the transcript level, reverse transcriptase PCR (RT-PCR) was first carried out to qualitatively assess the transcription of each

gene in the cluster, yielding positive results (data not shown). To provide a qualitative assessment, quantitative PCR (qPCR) was next employed at three time points in both the HZ848 and BJ5464-NpgA hosts. Transcription of all five genes was seen in both hosts (Figure 3.5), albeit at relatively low levels comparable to the *alg9* standard.

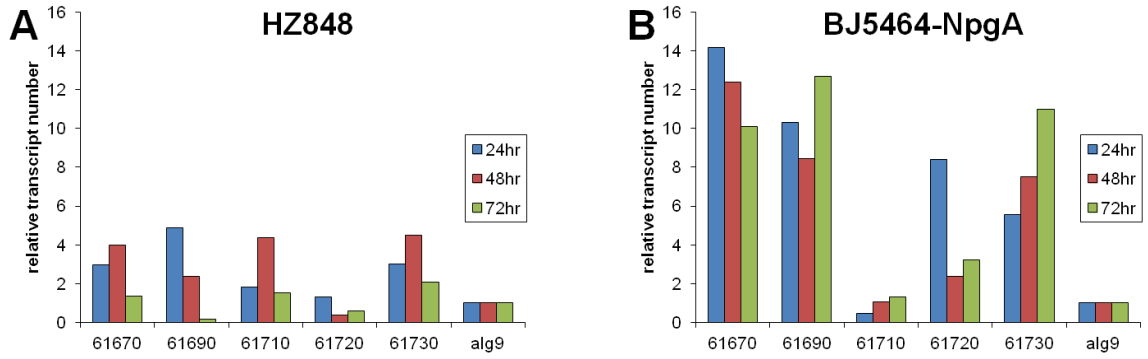


Figure 3.5: Transcript levels of the refactored gene cluster in (a) HZ848 and (b) BJ5464-NpgA.

As transcription of each gene in the cluster was observed by qPCR, protein-level detection was attempted next. For this study, each gene was tagged with an *N*-terminal His_6 tag in its corresponding helper plasmid and assessed individually. However, both Coomassie staining and anti-His Western blotting failed to identify expression. Next, affinity purification of the His-tagged proteins from BJ5464-NpgA lysate was attempted. Following concentration of the collected fractions, the target protein could be observed for both the *O*-methyltransferase and the FMO (Figure 3.6). However, expression of the halogenase and the two PKSs could not be observed, even when switched to a multi-copy plasmid. Notably, the BJ5464-NpgA strain produces two other proteins that appear on an anti-His Western blot. Activity assays with the partially purified *O*-methyltransferase protein were attempted using a zearalenone substrate, but no activity was detected.

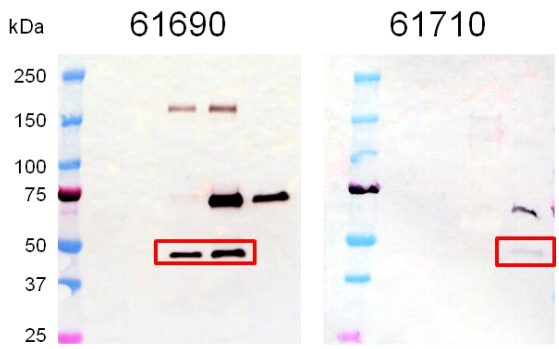


Figure 3.6: Anti-His_6 Western blotting results following expression from single-copy helper plasmids with constitutive promoters and Ni-NTA purification.

Given the poor protein expression results with our panel of constitutive promoters, expression was also tested using the ADH2 promoter. This promoter is repressed in the presence of glucose, but strongly active in its absence, and has previously been used to over-express PKSs in *S. cerevisiae* [15]. As a positive control, we used this promoter to over-express Hpm8 from the hypothemycin cluster, and a band of the correct size was observed by SDS-PAGE and Western blot (data not shown). However, when paired with *pks16* or *pks17* on single or multi-copy plasmids, expression of the target proteins was still not observed.

A disadvantage of Western blotting with large proteins, such as PKSs, is that with only one tag per protein, sensitivity on a mass basis becomes relatively poor as protein size increases. Thus, to increase the sensitivity of detection, the His_6-tags in the ADH2p PKS expression plasmids were replaced with 3xFLAG tags, purported to be 100-fold more sensitive. With this addition, expression of PKS16 was detected in cell lysate (Figure 3.7). However, PKS17 still could not be observed. Thus, it was concluded that *pks17* was not amenable to expression in *S. cerevisiae*.

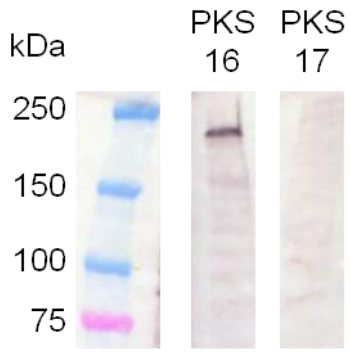


Figure 3.7: Anti-FLAG Western blotting reveals expression of PKS16 only.

3.2.2.3 In vitro assays

At this point, soluble expression in *S. cerevisiae* was found to be the significant obstacle to product formation from this cluster. As an alternative means to demonstrate production, expression of the enzymes in *E. coli* was attempted. Notably, the two PKSs in the hypothemycin gene cluster have been reconstituted *in vitro*, suggesting the viability of this approach [15]. All five of the genes were individually cloned into the pET28a expression vector. While each of the three tailoring proteins could be expressed and purified, the PKS proteins, showed no detectable expression either in the soluble or insoluble protein fractions.

Even though the PKSs could not be expressed in *E. coli* or *S. cerevisiae*, *in vitro* assays were attempted with the purified tailoring enzymes. Here, the commercially available zearalenone or radicicol were used as surrogate substrates, as they are expected to be structurally similar to the product of PKS16 and PKS17. To assist the halogenase, the flavin reductase SsuE from *E. coli* was also cloned and purified. Interestingly, though no activity was observed for the *O*-methyltransferase or the FMO, formation of a product with absorbance at 280 nm was observed when zearalenone was incubated with the halogenase (Figure 3.8). The new peak exhibited an

m/z of 351.4 in negative mode, consistent with the addition of chlorine. Further, the isotope distribution is consistent with that of a chlorinated product. No activity was observed when the halogenase was incubated with zearalenone in the presence of other halogen salts, or when incubated with tryptophan (the substrate of other characterized halogenases RebH [25] and PrnA [26]). Thus, this activity provides some evidence that the true substrate is in fact a zearalenone-like macrolactone. Nevertheless, as other halogenases with this function, significantly higher activity, and broader substrate scope have already been characterized [27,28], this particular halogenase was not investigated further.

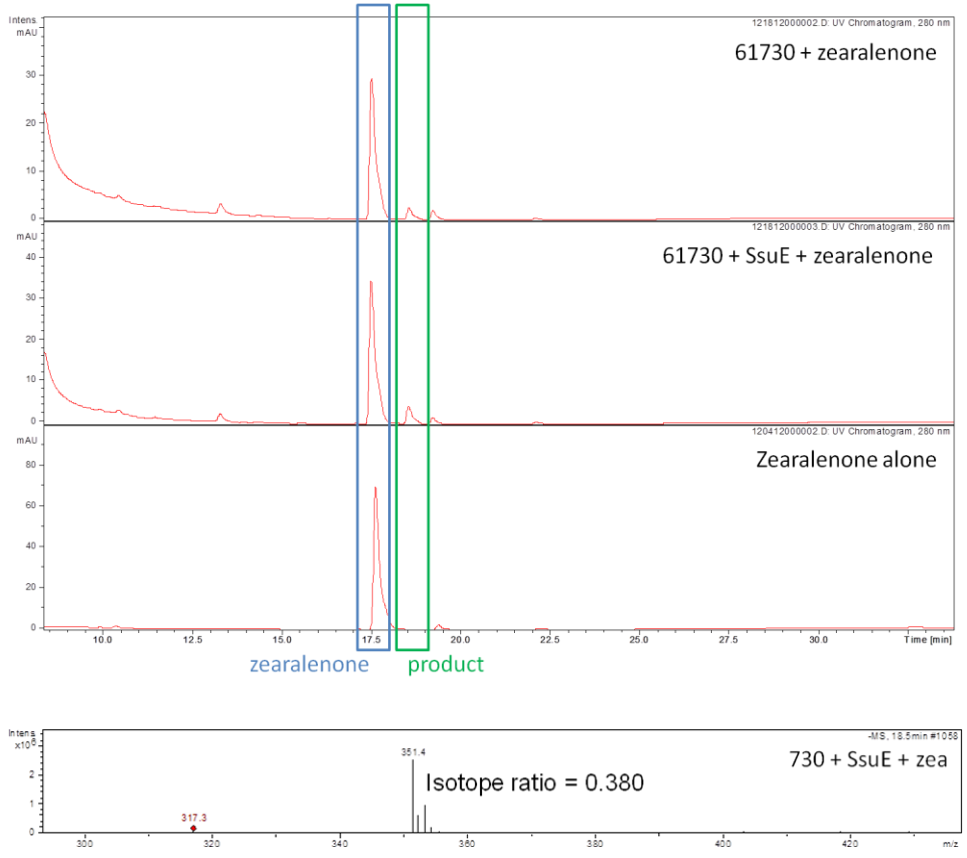


Figure 3.8: 280 nm UV traces show formation of a chloro-zearalenone product (m/z = 351), confirmed by LC-MS.

3.2.3 Expression of a panel of non-reducing PKSs

Given the significant time and labor invested in the single *pks16*/*pks17* gene cluster without detection of a novel product, subsequent studies focused on multiple clusters in parallel. Rather than attempting full cluster refactoring from the outset, expression of only the PKS gene in each cluster was evaluated first. For this study, only the non-reducing PKSs were considered, as they are expected to produce compounds with UV absorbance detectable by HPLC. In total, eleven NR-PKSs were considered, including PKS1, PKS3, PKS4, PKS10, PKS11, PKS12, PKS14, PKS15, PKS18, PKS20, and PKS22 [14]. Additionally, PKS21 was included in the evaluation, as it is located in the neighborhood of PKS20 and is expected to cooperate with PKS20. Notably, PKS4 has previously been shown to be involved in melanin biosynthesis in *T. marneffei*, where knockdown experiments demonstrated loss of pigmentation [14]. Thus, PKS4 can be expected to catalyze the synthesis of YWA1, the precursor of melanin in the fungal melanin pathway, and serves as a positive control.

To evaluate protein expression, the pYAAFH helper plasmid was designed and constructed (Figure 3.9). This plasmid features ADH2p and ADH2t for strong transcription, and *N*-terminal 3xFLAG tag for protein detection, and a *C*-terminal His_6 tag for affinity enrichment/purification. A unique *Age*I recognition site was included between the two tags for assembly of the *pks* coding sequences at this position. A 2μ replication origin was also included to boost copy number.

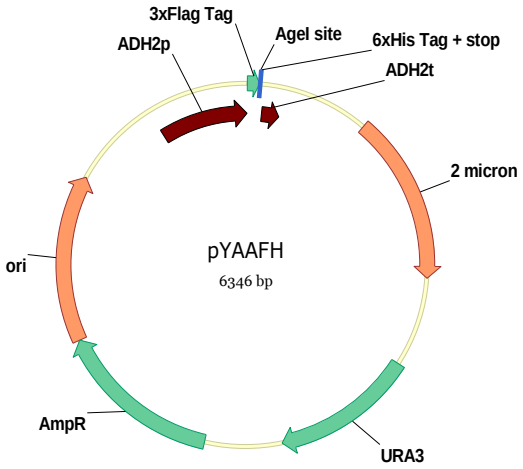


Figure 3.9: The pYAAFH plasmid for PKS expression.

To assemble the PKS expression plasmids, each gene was first amplified from the genomic DNA of *T. marneffei*. Next, the individual exons were amplified from the purified gene. Finally, the purified exons were transformed with the linearized pYAAFH backbone into *S. cerevisiae* HZ848 for assembly. All expression plasmids except for PKS11 could be successfully assembled in this manner, and were subsequently transformed into *S. cerevisiae* expression strain BJ5464-NpgA.

Expression of the PKS proteins from BJ5464-NpgA was evaluated from small cultures by both Coomassie staining and anti-FLAG Western blotting of the cell lysate. Interestingly, although no unique bands were observed for any of the expression strains by Coomassie staining (Figure 3.10a), every strain was found to produce a unique band by Western blot (Figure 3.10b). Although the lack of apparent bands on the SDS-PAGE gels suggested very low expression levels for every PKS evaluated, it was suspected that expression might nevertheless be high enough for activity to be observed. Detection of novel compounds from the PKS expression strains was attempted via HLPC-MS. Each strain, however, was found to produce a nearly

identical UV profile (data not shown). MS profiles were also analyzed, but also did not yield obvious new peaks.

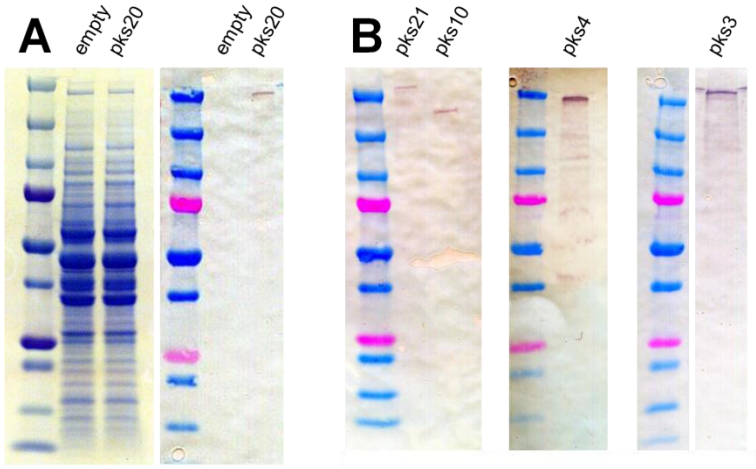


Figure 3.10: (a) Evaluation of PKS20 by Coomassie stain and anti-FLAG Western blot. (b) Representative anti-FLAG Western blot for three additional PKSs.

3.2.4 Characterization and evaluation of PCK1p

3.2.4.1 Characterization of strong S. cerevisiae promoters

As the ADH2p promoter did not yield high expression levels for any of the PKSs evaluated, it was hypothesized that different promoters with strong expression profiles could help to increase protein levels. In collaboration with Jonathan Ning, three such promoters (PCK1p, SIP18p, and SPG4p) were identified from the literature for further analysis [29]. To test protein expression driven by these promoters, each promoter was used to drive the catechol dioxygenase gene *xylE* from a multi-copy plasmid, and turnover of catechol was measured in cell lysate. Two known promoters (ADH2p and TEF1p) were included in this assay for comparison. As expected, the constitutive TEF1p control was found to exhibit the greatest XylE activity at the first time point

(24 hr). However, at all subsequent time points PCK1p yielded the highest XylE activity (Figure 3.11).

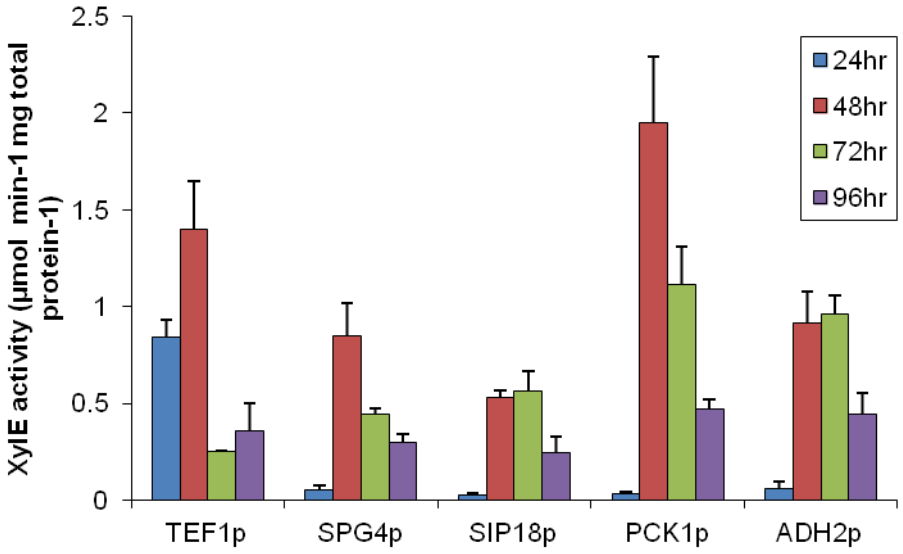


Figure 3.11: XylE activity assay of promoter strength.

3.2.4.2 Expression of pks4 under PCK1p

Given the superior XylE activity observed under the PCK1p promoter, PCK1p was chosen to evaluate its capacity to drive PKS expression. PKS4 was chosen for initial analysis, as it is expected to catalyze the synthesis of the known YWA1 product. Replacement of ADH2p with PCK1p in the YAAFH-*pks4* plasmid led to strong pigmentation in the BJ5464-NpgA production strain (Figure 3.12a). The same phenotype was observed under SPG4p or SIP18p, but not observed in the empty plasmid control or with the ADH2p construct. Interestingly, HPLC analysis of the extracted culture led to the identification of not one, but two significant new peaks in the UV profile eluting at 10.2 min and 11.2 min (Figure 3.12b).

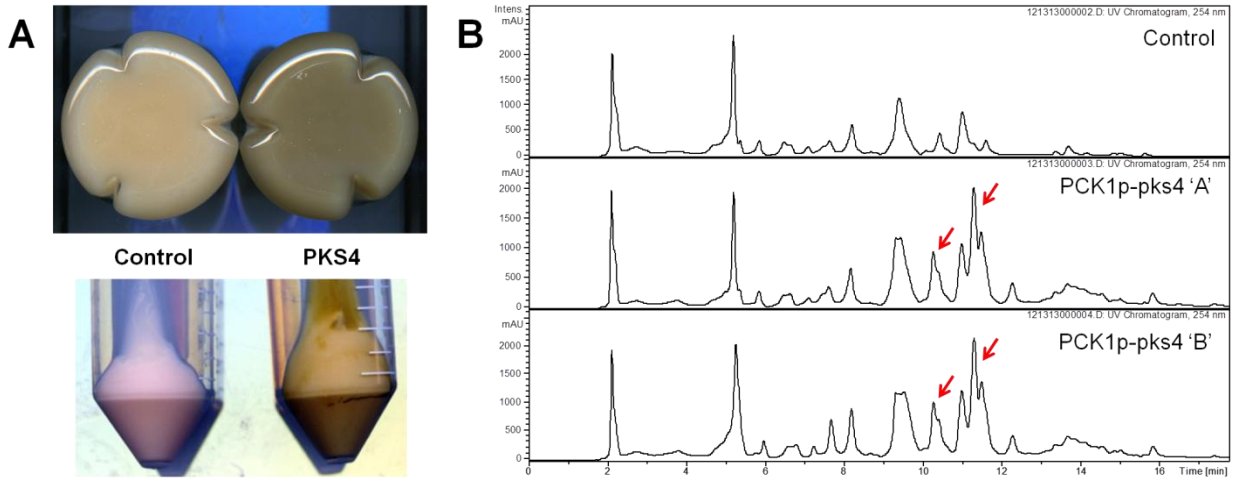


Figure 3.12: (a) PKS4 expression leads to dark pigmentation in *S. cerevisiae* cultures and cell pellets. (b) UV absorbance profiles (254 nm) reveal two new peaks in duplicate PKS4 expression strains (indicated with red arrows).

Notably, PKS4 is a homolog of the PKS encoded by *wA* from *Aspergillus nidulans* (61 % amino acid identity). This enzyme was originally thought to catalyze synthesis of citreoisocoumarin, evidenced by heterologous expression in *Aspergillus oryzae* [30]. However, a year later it was discovered that a single nucleotide had mistakenly been omitted from the originally published *wA* sequence [31], leading to inadvertent truncation of the TE domain in the subsequent cloning and heterologous expression. When the full-length gene was re-cloned and expressed, it was found to in fact synthesize the YWA1 product [32], whereas citreoisocoumarin was a product of incorrect cyclization. MS analysis of the PKS4 extract revealed masses consistent with both YWA1 and citreoisocoumarin coeluting with the new peaks in the UV profile. These peaks were isolated by HPLC fractionation and analyzed by high-resolution MS, where the most probable chemical formulas were predicted to be $C_{14}H_{12}O_6$ for the 10.2 min peak (matching YWA1) and $C_{14}H_{14}O_6$ for the 11.2 min peak (matching citreoisocoumarin). As the biosynthesis of these compounds is well studied, and the heterologous production of YWA1 has recently been

demonstrated in *S. cerevisiae* [33], further analysis of PKS4 and its corresponding gene cluster were not pursued.

3.2.4.3 Assembly of additional non-reducing pks constructs

Having successfully reconstituted PKS4 in *S. cerevisiae*, PCK1p expression plasmids were constructed for the other non-reducing PKSs from *T. marneffei*. In these cases, however, no obvious new pigments were produced, and no new peaks were apparent in the HPLC-MS profiles. Prediction of the products of these enzymes based on bioinformatic analysis provided limited insight; however, for PKS3, similarity with a recently characterized PKS from *Monascus purpureus* suggested a possible product structure (Figure 3.13) [34]. The predicted structure was proposed to be synthesized using one crotonyl-CoA, a metabolite not natively produced in *S. cerevisiae*. Reconstitution of a crotonyl-CoA pathway has previously been demonstrated *en route* to *n*-butanol biosynthesis in *S. cerevisiae* [35].

The crotonyl-CoA pathway, converting acetyl-CoA to crotonyl-CoA via acetoacetyl-CoA and β-hydroxybutyryl-CoA, was integrated to the genome of BJ5464-NpgA. Also integrated to the genome was a heterologous *matB* malonyl-CoA synthetase from *Rhizobium trifolii* in an effort to boost malonyl-CoA levels [36]. To confirm expression of all four integrated genes in the resulting strain (named *S. cerevisiae* BJ5464-THCM), qPCR was conducted (Figure 3.14). While expression levels were low compared to an *alg9* standard, all four gene products were detected. However, expression of PKS3 in this strain did not yield the predicted product.

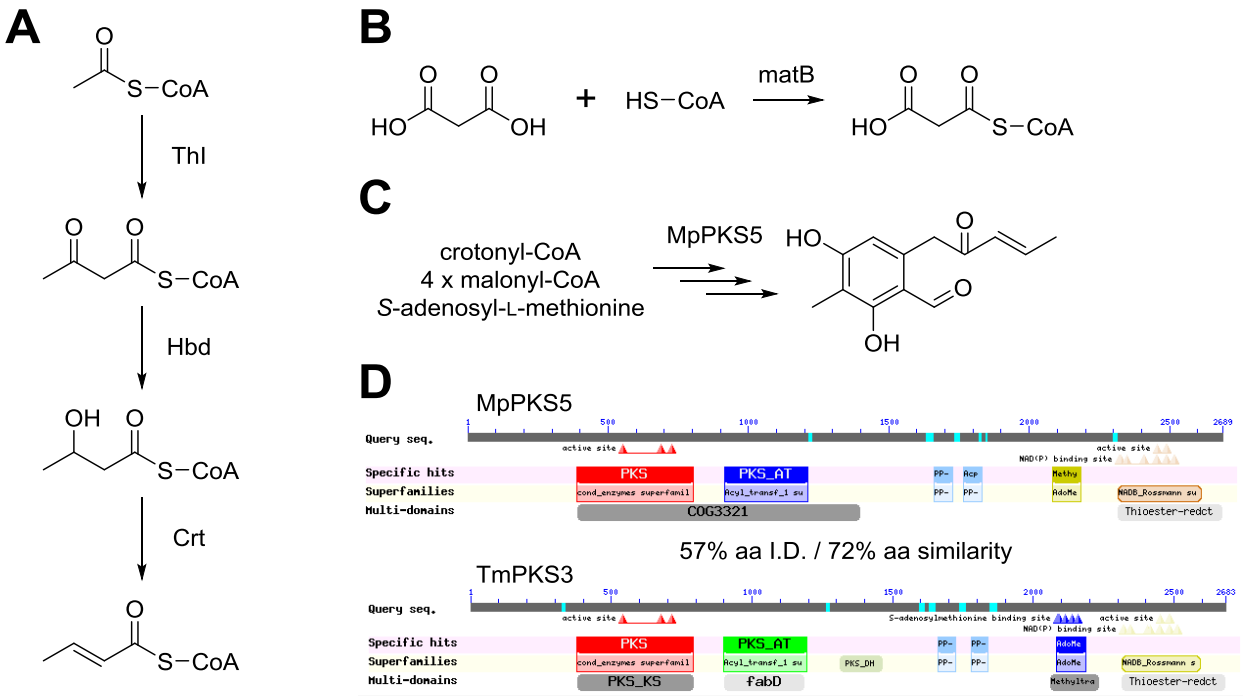


Figure 3.13: (a) The crotonyl-CoA biosynthetic pathway from *Clostridium*. (b) Malonyl-CoA synthesis by MatB from *Rhizobium trifolii*. (c) Proposed synthesis of an azaphilone precursor by MpPKS5. (d) Domain comparison between PKS3 of *T. marneffei* and MpPKS5.

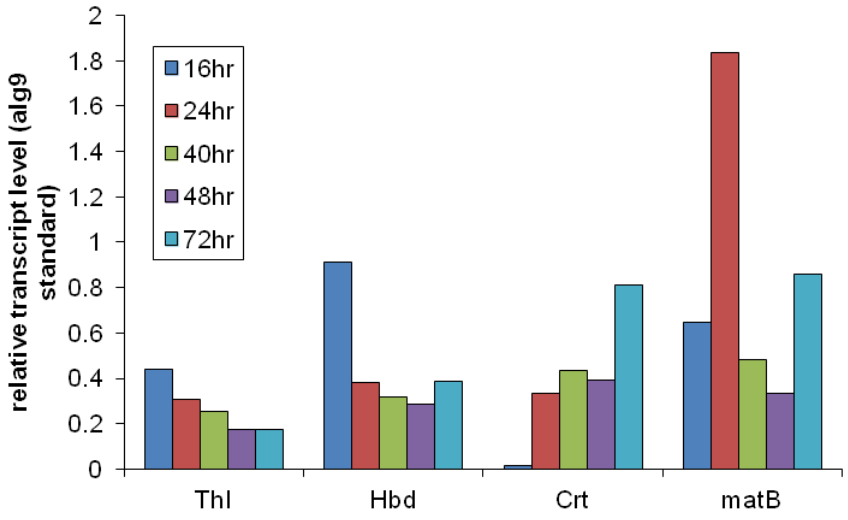


Figure 3.14: Gene expression profiles in *S. cerevisiae* BJ5464-THCM.

In a very recent publication, PKS3 has been shown to be involved in the synthesis of pigments monascorubrin and ankaflavin, as well as citrinin, through gene knockdown experiments in *T. marneffei* [37]. Given this new lead, HPLC-MS data was reanalyzed for the corresponding masses of citrinin and compound 1, both putative products of PKS3. While citrinin was not observed, a unique peak was observed in PKS3 extracts corresponding to the mass of compound 1 (Figure 3.15a). The peak was observed in both pellet and supernatant extracts, and eluted with a very small UV peak (Figure 3.15b). Notably, this peak was not observed in the ADH2p expression strain.

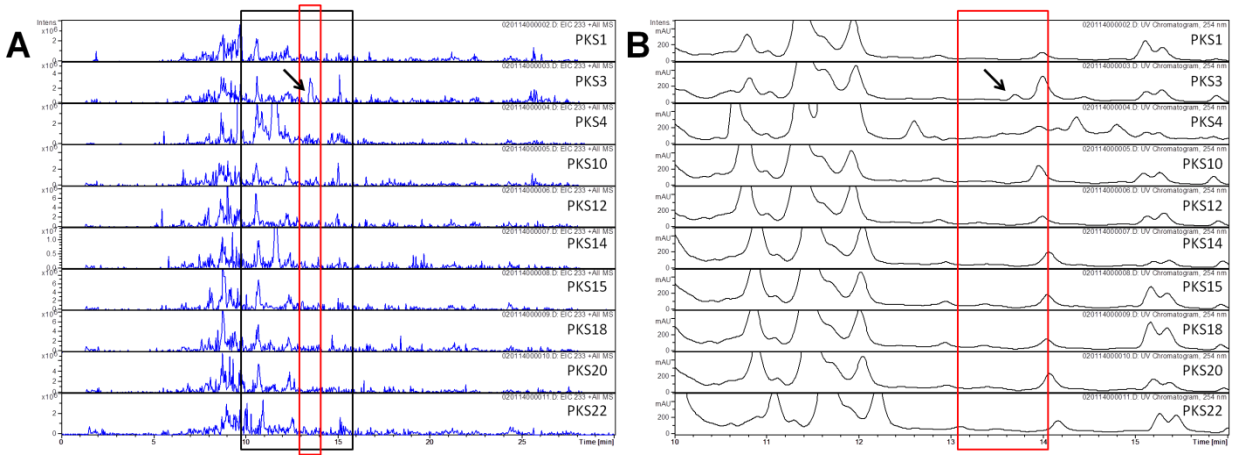


Figure 3.15: (a) MS analysis of all PKS expression strains shows a peak with m/z = 233^{+} at 13 min only in the presence of PKS3 (black arrow). (b) UV traces at 254 nm, zoomed in on the region shown in a black box in (a).

3.3 Conclusions and Outlook

In the study presented here, the polyketide synthases of *T. marneffei* were explored in *S. cerevisiae*. A general helper plasmid was constructed for the expression, detection, and purification of PKSs in *S. cerevisiae*. A set of 14 single PKS expression plasmids were constructed, leading to detectable expression of the encoded proteins by Western blot. However,

polyketide products were not detected. Three strong endogenous promoters from *S. cerevisiae* were characterized and shown to out-perform the commonly used ADH2p to drive expression of PKS4 for production of both YWA1 and citreoisocoumarin. Additionally, the full *pks16*/*pks17* cluster was refactored on a single expression plasmid. Although the expected RAL product was not observed, a halogenase active on a RAL substrate was discovered. Finally, an *S. cerevisiae* expression host was designed to provide the crotonyl-CoA precursor for polyketide synthesis.

Given our detection of soluble protein expression in all cases tested, it is likely that there is some small amount of accompanying activity for at least some of the PKSs. However, this production is likely to be too low to yield an obvious signal in a UV or MS trace. Without knowledge of the exact product of the heterologous gene or cluster, then, the manual detection of novel products amid an extracted metabolome becomes a painstaking, needle-in-a-haystack type of task. More comprehensive analytical tools for metabolic profiling, recently exemplified by the Dorrestein group [38], are needed to supplant manual peak-by-peak analysis.

Here, inherent limitations in the expression of these PKSs in *S. cerevisiae* seem to underlie their low expression levels (and correspondingly, the undetectable product synthesis). Such issues could stem from unfavorable codon bias, incorrect prediction of introns, or other metabolic differences between the native *T. marneffei* strain and *S. cerevisiae*. Notably, this is not a universal result; there is a literature precedent for the expression of PKSs in yeast [39], and during the course of this study a similar effort was undertaken by the Watanabe group with significantly greater success [40]. In our study, however, a valid next step would be to try different fungal hosts for heterologous expression. For example, *Aspergillus nidulans* has

recently been demonstrated as a suitable host for heterologous expression of fungal PKSs [41]. As closer relatives of *T. marneffei* and as native polyketide producers, *Aspergillus* species may provide natural metabolic advantages for heterologous polyketide production.

3.4 Materials and Methods

3.4.1 Strains, media and reagents

The *Escherichia coli* expression strain BL21(DE3) and the expression vector pET28a were obtained from EMD Biosciences (San Diego, CA). *E. coli* cloning strain BW25141 was a gift from Prof. William Metcalf (University of Illinois at Urbana-Champaign). *S. cerevisiae* HZ848 (*MATα, ade2-1, ade3Δ22*, Δ*ura3, his3-11,15, trp1-1, leu2-3,112* and *can1-100*) was used for plasmid assembly and protein expression. *S. cerevisiae* BJ5464-NpgA (*MATα, ura3-52, his3-*Δ*200, leu2-*Δ*1, trp1, pep4*::*HIS3, prb1, Δ1.6R, can1, GAL*) and plasmid pZH126, containing *hpm8* under the ADH2p promoter, were gifts from Prof. Yi Tang (University of California, Los Angeles). *Talaromyces marneffei* genomic DNA was purchased from the American Type Culture Collection (Manassas, VA). *E. coli* strains were grown in LB medium supplemented with the appropriate antibiotic (50 μg/mL kanamycin or 100 μg/mL ampicillin). Yeast strains were grown in appropriate SC dropout media (for plasmid assembly and seed cultures) or YPAD medium (for protein expression cultures). All media components, organic solvents, and consumables were purchased from Thermo-Fisher Scientific (Pittsburgh, PA). Antibiotics and isopropyl-β-D-1-thiogalactopyranoside (IPTG) were purchased from Gold Biotechnology (St. Louis, MO). XAD-1180 resin and resorcylic acid lactone standards were purchased from Sigma-Aldrich (St. Louis, MO). PCR primers were synthesized by Integrated DNA Technologies (Coralville, IA), and PCR reactions were performed in FailSafe PCR PreMix G (Epicentre

Biotechnologies, Madison, WI) with Q5 DNA polymerase (New England Biolabs, Ipswich, MA). All PCR products were purified using the DNA Clean & Concentrator or Zymoclean Gel DNA Recovery Kit (Zymo Research, Irvine, CA). Plasmids were recovered using the QIAprep Spin Miniprep Kit (Qiagen, Valencia, CA). Restriction enzymes and T4 ligase were purchased from New England Biolabs (Ipswich, MA). Talon Cobalt immobilized metal affinity chromatography (IMAC) resin was purchased from Clontech Laboratories (Mountain View, CA). Total RNA was isolated with the RNeasy Mini Kit (Qiagen) and treated with TURBO DNA-free DNase (Life Technologies, Carlsbad, CA). Complimentary DNA was synthesized with the ProtoScript II First Strand cDNA Synthesis Kit (NEB).

3.4.2 Plasmid construction

All single-gene helper plasmids (including *pks16*/*pks17* pathway gene plasmids, pYAAFH expression plasmids, and *xylE* expression plasmids) were assembled via yeast homologous recombination in *S. cerevisiae* HZ848 following previously established protocol [42]. Exons of small genes (< 2 kb) were spliced by overlap-extension PCR [43] prior to yeast assembly, while PKS exons were assembled directly in yeast. Plasmids recovered from yeast assemblies were electroporated into *E. coli* BW25141 for confirmation by diagnostic restriction digestion and sequencing. Multi-gene plasmids were assembled in yeast from the single-gene expression cassettes, PCR amplified from the corresponding helper plasmids with primers annealing to the upstream and downstream terminators. To replace the ADH2p promoter in each pYAAFH expression plasmid with PCK1p, three fragments for each construct were prepared by PCR: the backbone (without promoter), the *pks* gene, and the PCK1p promoter. The three fragments were transformed into *S. cerevisiae* HZ848 for assembly via homologous recombination. Plasmid

pRS405-THCM, for chromosomal integration of *thI*, *hbd*, *crt*, and *matB* to BJ5464-NpgA at the *leu2* locus, was constructed via Gibson isothermal assembly following literature precedent [44]. Single gene expression plasmids for *E. coli* were assembled via ligation of the PCR-amplified gene into the *Nde*I/*Hind*III sites of pET28a (for genes < 2 kb) or Gibson assembly into the *Nde*I/*Hind*III-linearized pET28a backbone (for PKS genes) following literature precedent.

3.4.3 Transformation

E. coli strains BW25141 and BL21(DE3) were transformed by electroporation. *S. cerevisiae* strain HZ848 was transformed by electroporation for all plasmid assemblies, following the procedure reported elsewhere [42]. For transformation of *S. cerevisiae* HZ848 and BJ5464-NpgA with circular plasmids, the LiAc/ssDNA method was employed, as described elsewhere [45]. Chromosomal integration to BJ5464-NpgA was accomplished via electroporation of the *Bst*EII-linearized suicide plasmid.

3.4.4 *S. cerevisiae* culture, extraction, and analysis

Individual colonies of yeast expression strains were picked to 3 mL of the appropriate SC dropout medium in 14 mL tubes and grown at 30 °C, 250 rpm for 1 day. Seed cultures were diluted 100-fold into fresh YPAD medium for protein expression and grown at 30 °C, 250 rpm for 3 days. Culture supernatants were collected by centrifugation and extracted with either XAD-1180 resin (20 g/L) or a 3:1 volume of ethyl acetate. XAD-1180 resin was washed with water and eluted with methanol. Extracts were concentrated to dryness by rotavap, and resuspended in methanol (1 % of the original culture volume). Samples were then analyzed by HPLC-MS/MS at the Roy J. Carver Metabolomics Center, University of Illinois, Urbana-

Champaign, on an Agilent 1100 Series HPLC (Agilent, Palo Alto, CA) coupled to an Agilent XCT ion-trap MSD mass spectrometer with ESI source. Compounds were separated with a Kinetex C18 column (Phenomenex, Torrance, CA) on a 0 – 100 % gradient of 0.1 % aqueous formic acid (solvent A) to acetonitrile (solvent B) over 25 minutes. High-resolution MS analysis was performed by the School of Chemical Sciences Mass Spectrometry Facility, University of Illinois, Urbana-Champaign.

3.4.5 Quantitative PCR

Total RNA was isolated from 1 mL samples taken from 50 mL YPAD cultures at the specified times. The corresponding cDNA was synthesized and used as template for quantitative PCR. Each qPCR reaction contained 10 µL Power SYBR Master Mix (Life Technologies), 0.1 µM of each primer, 6 µL of DNA template (diluted 100-fold from the cDNA synthesis reaction), and water to 20 µL. Serial dilutions of the assembled plasmids were used as templates for standard curves. Reactions were performed on a LightCycler 480 System (Roche, Basel, Switzerland).

3.4.6 Protein expression and purification

E. coli BL21(DE3) pET28a expression strains were grown in Terrific Broth (TB) media supplemented with kanamycin (50 µg/mL) at 37 °C to an OD_{600} of ~0.8, after which induction was carried out by addition of 0.3 mM IPTG at 20 °C. After 18 – 24 hr, the cells were harvested by centrifugation at 7500 rpm for 15 min and resuspended in 20 mM Tris-HCl (pH 7.65), 0.5 M NaCl, and 15 % glycerol supplemented with 1 mg/mL lysozyme. After a freeze-thaw cycle at -80 °C, the cell suspension was sonicated to ensure sufficient lysis. The lysate was clarified multiple times by centrifugation at 15000 rpm for 15 min, after which the His_6-tagged proteins

were purified by affinity chromatography on TALON Superflow Co^{2+} resin coupled to fast-performance liquid chromatography. The eluted proteins were washed three times in 50 mM HEPES (pH 7.25), concentrated, and stored in 15 % glycerol at -80 °C.

3.4.7 SDS-PAGE and Western blotting

S. cerevisiae strains were grown in YPAD medium for all protein expression studies. Cells were pelleted by centrifugation and resuspended in lysis buffer (50 mM NaH_2PO_4, pH 8.0, 0.15 M NaCl, 10 mM imidazole) (20 mL per L of culture) [15]. Cells were lysed via sonication on ice in 5 second pulses with 10 second intervals for a total of 30 minutes. Lysates were clarified by centrifugation at 15,000 rpm and analyzed on 4-20% Min-PROTEAN TGX precast polyacrylamide gels (Bio-Rad, Hercules, CA) stained with SimplyBlue SafeStain (Life Technologies, Carlsbad, CA). For enrichment of His_6-tagged proteins, Ni-NTA agarose (Qiagen) was added to clarified lysate, which was then incubated at 4 °C for 2 hr with gentle agitation. Resin was washed with lysis buffer and eluted with increasing concentrations of imidazole. Western blotting was performed with His-probe primary antibody sc-8036 (Santa Cruz Biotechnology, Dallas, TX) or anti-FLAG primary antibody F3165 (Sigma-Aldrich) and goat anti-mouse–alkaline phosphatase secondary antibody A5153 (Sigma-Aldrich), and visualized with Western Blue stabilized substrate for alkaline phosphatase (Promega, Madison, WI).

3.4.8 *In vitro* enzyme assays

PMAA_061690 *O*-methyltransferase assays were performed in 200 µL of 50 mM HEPES buffer, pH 7.25, with 10 mM $MgSO_4$, 250 µM zearalenone or radicicol, 400 µM SAM, and ca. 10 µg of

enzyme. PMAA_061710 monooxygenase assays were performed in 200 µL of 50 mM HEPES buffer, pH 7.25, with 500 µM zearalenone, 1 mM NAD(P)H, and ca. 10 µg of enzyme. PMAA_061730 halogenase assays were performed in 100 mM Tris-HCl buffer, pH 7.5, with 50 mM NaCl, 1 mM NADPH, 100 µM FAD, 500 µM zearalenone, and up to 50 µM enzyme. Assays with alternate halogen salts were carried out 100 mM KH_2PO_4, pH 7.5. All samples were incubated at 30 °C for 1 – 16 hr before HPLC-MS analysis.

3.4.9 XylE activity assay

For each promoter, three biological replicates and one empty plasmid negative control were grown in 5mL YPAD. Samples (500 µL) were collected at 24, 48, 72, and 96 hr and lysed with YPER according to the manufacturer's suggested protocol. Reactions were performed at 30 °C in a microtiter plate containing 200 µL 1X phosphate buffered saline (Lonza, Walkersville, MD) and 2 mM catechol, initiated with the addition of 1 µL of lysate. Absorbance was monitored at 375 nm for 5 min. Activity was normalized by total protein content, measured by Bradford assay.

3.5 References

1. Hoffmeister, D. and Keller, N.P. (2007) Natural products of filamentous fungi: enzymes, genes, and their regulation. *Nat Prod Rep*, **24**, 393-416.
2. Chiang, Y.M., Lee, K.H., Sanchez, J.F., Keller, N.P. and Wang, C.C. (2009) Unlocking fungal cryptic natural products. *Nat Prod Commun*, **4**, 1505-1510.
3. Minto, R.E. and Townsend, C.A. (1997) Enzymology and molecular biology of aflatoxin biosynthesis. *Chem Rev*, **97**, 2537-2556.

4. Hajjaj, H., Klaebe, A., Loret, M.O., Goma, G., Blanc, P.J. and Francois, J. (1999) Biosynthetic pathway of citrinin in the filamentous fungus *Monascus ruber* as revealed by 13C nuclear magnetic resonance. *Appl Environ Microbiol*, **65**, 311-314.

5. Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H. and Fedorova, N.D. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet Biol*, **47**, 736-741.

6. Weber, T. (2014) *In silico* tools for the analysis of antibiotic biosynthetic pathways. *Int J Med Microbiol*, **304**, 230-235.

7. Yaegashi, J., Oakley, B.R. and Wang, C.C. (2014) Recent advances in genome mining of secondary metabolite biosynthetic gene clusters and the development of heterologous expression systems in *Aspergillus nidulans*. *J Ind Microbiol Biotechnol*, **41**, 433-442.

8. Andersen, M.R., Nielsen, J.B., Klitgaard, A., Petersen, L.M., Zachariasen, M., Hansen, T.J., Blicher, L.H., Gotfredsen, C.H., Larsen, T.O., Nielsen, K.F. *et al.* (2013) Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proc Natl Acad Sci U S A*, **110**, E99-107.

9. Galagan, J.E., Calvo, S.E., Cuomo, C., Ma, L.J., Wortman, J.R., Batzoglou, S., Lee, S.I., Basturkmen, M., Spevak, C.C., Clutterbuck, J. *et al.* (2005) Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature*, **438**, 1105-1115.

10. Chiang, Y.M., Szewczyk, E., Nayak, T., Davidson, A.D., Sanchez, J.F., Lo, H.C., Ho, W.Y., Simityan, H., Kuo, E., Praseuth, A. *et al.* (2008) Molecular genetic mining of the *Aspergillus* secondary metabolome: discovery of the emericellamide biosynthetic pathway. *Chem Biol*, **15**, 527-532.

11. Williams, R.B., Henrikson, J.C., Hoover, A.R., Lee, A.E. and Cichewicz, R.H. (2008) Epigenetic remodeling of the fungal secondary metabolome. *Org Biomol Chem*, **6**, 1895-1897.

12. Partida-Martinez, L.P., Monajembashi, S., Greulich, K.O. and Hertweck, C. (2007) Endosymbiont-dependent host reproduction maintains bacterial-fungal mutualism. *Curr Biol*, **17**, 773-777.

13. Guo, W., Peng, J., Zhu, T., Gu, Q., Keyzers, R.A. and Li, D. (2013) Sorbicillamines A-E, nitrogen-containing sorbicillinoids from the deep-sea-derived fungus *Penicillium* sp. F23-2. *J Nat Prod*, **76**, 2106-2112.

14. Woo, P.C., Tam, E.W., Chong, K.T., Cai, J.J., Tung, E.T., Ngan, A.H., Lau, S.K. and Yuen, K.Y. (2010) High diversity of polyketide synthase genes and the melanin biosynthesis gene cluster in *Penicillium marneffei*. *FEBS J*, **277**, 3750-3758.

15. Zhou, H., Qiao, K., Gao, Z., Meehan, M.J., Li, J.W., Zhao, X., Dorrestein, P.C., Vederas, J.C. and Tang, Y. (2010) Enzymatic synthesis of resorcylic acid lactones by cooperation of fungal iterative polyketide synthases involved in hypothemycin biosynthesis. *J Am Chem Soc*, **132**, 4530-4531.

16. Xu, Y., Zhou, T., Espinosa-Artiles, P., Tang, Y., Zhan, J. and Molnar, I. (2014) Insights into the biosynthesis of 12-membered resorcylic acid lactones from heterologous production in *Saccharomyces cerevisiae*. *ACS Chem Biol*, **9**, 1119-1127.

17. Winssinger, N. and Barluenga, S. (2007) Chemistry and biology of resorcylic acid lactones. *Chem Commun (Camb)*, 22-36.

18. Xu, Y., Espinosa-Artiles, P., Schubert, V., Xu, Y.M., Zhang, W., Lin, M., Gunatilaka, A.A., Sussmuth, R. and Molnar, I. (2013) Characterization of the biosynthetic genes for

10,11-dehydrocurvularin, a heat shock response-modulating anticancer fungal polyketide from *Aspergillus terreus*. *Appl Environ Microbiol*, **79**, 2038-2047.

19. Wang, S., Xu, Y., Maine, E.A., Wijeratne, E.M., Espinosa-Artiles, P., Gunatilaka, A.A. and Molnar, I. (2008) Functional characterization of the biosynthesis of radicicol, an Hsp90 inhibitor resorcylic acid lactone from *Chaetomium chiversii*. *Chem Biol*, **15**, 1328-1338.

20. Reeves, C.D., Hu, Z., Reid, R. and Kealey, J.T. (2008) Genes for the biosynthesis of the fungal polyketides hypothemycin from *Hypomyces subiculosus* and radicicol from *Pochonia chlamydosporia*. *Appl Environ Microbiol*, **74**, 5121-5129.

21. Ichinose, K., Ozawa, M., Itou, K., Kunieda, K. and Ebizuka, Y. (2003) Cloning, sequencing and heterologous expression of the medermycin biosynthetic gene cluster of *Streptomyces* sp. AM-7161: towards comparative analysis of the benzoisochromanequinone gene clusters. *Microbiology*, **149**, 1633-1645.

22. Erb, A., Luzhetskyy, A., Hardter, U. and Bechthold, A. (2009) Cloning and sequencing of the biosynthetic gene cluster for saquayamycin Z and galtamycin B and the elucidation of the assembly of their saccharide chains. *Chembiochem*, **10**, 1392-1401.

23. Qiao, K., Chooi, Y.H. and Tang, Y. (2011) Identification and engineering of the cytochalasin gene cluster from *Aspergillus clavatus* NRRL 1. *Metab Eng*, **13**, 723-732.

24. Sun, J., Shao, Z., Zhao, H., Nair, N., Wen, F., Xu, J.H. and Zhao, H. (2012) Cloning and characterization of a panel of constitutive promoters for applications in pathway engineering in *Saccharomyces cerevisiae*. *Biotechnol Bioeng*, **109**, 2082-2092.

25. Glenn, W.S., Nims, E. and O'Connor, S.E. (2011) Reengineering a tryptophan halogenase to preferentially chlorinate a direct alkaloid precursor. *J Am Chem Soc*, **133**, 19346-19349.

26. Dong, C., Flecks, S., Unversucht, S., Haupt, C., van Pee, K.H. and Naismith, J.H. (2005) Tryptophan 7-halogenase (PrnA) structure suggests a mechanism for regioselective chlorination. *Science*, **309**, 2216-2219.

27. Zeng, J. and Zhan, J. (2010) A novel fungal flavin-dependent halogenase for natural product biosynthesis. *Chembiochem*, **11**, 2119-2123.

28. Zeng, J., Lytle, A.K., Gage, D., Johnson, S.J. and Zhan, J. (2013) Specific chlorination of isoquinolines by a fungal flavin-dependent halogenase. *Bioorg Med Chem Lett*, **23**, 1001-1003.

29. Radonjic, M., Andrau, J.C., Lijnzaad, P., Kemmeren, P., Kockelkorn, T.T., van Leenen, D., van Berkum, N.L. and Holstege, F.C. (2005) Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon *S. cerevisiae* stationary phase exit. *Mol Cell*, **18**, 171-183.

30. Watanabe, A., Ono, Y., Fujii, I., Sankawa, U., Mayorga, M.E., Timberlake, W.E. and Ebizuka, Y. (1998) Product identification of polyketide synthase coded by *Aspergillus nidulans wA* gene. *Tet Lett*, **39**, 7733-7736.

31. Mayorga, M.E. and Timberlake, W.E. (1992) The developmentally regulated *Aspergillus nidulans wA* gene encodes a polypeptide homologous to polyketide and fatty acid synthases. *Mol Gen Genet*, **235**, 205-212.

32. Watanabe, A., Fujii, I., Sankawa, U., Mayorga, M.E., Timberlake, W.E. and Ebizuka, Y. (1999) Re-identification of *Aspergillus nidulans wA* gene to code for a polyketide synthase of naphthopyrone. *Tet Lett*, **40**, 91-94.

33. Rugbjerg, P., Naesby, M., Mortensen, U.H. and Frandsen, R.J. (2013) Reconstruction of the biosynthetic pathway for the core fungal polyketide scaffold rubrofusarin in *Saccharomyces cerevisiae*. *Microb Cell Fact*, **12**, 31.

34. Balakrishnan, B., Karki, S., Chiu, S.H., Kim, H.J., Suh, J.W., Nam, B., Yoon, Y.M., Chen, C.C. and Kwon, H.J. (2013) Genetic localization and in vivo characterization of a *Monascus* azaphilone pigment biosynthetic gene cluster. *Appl Microbiol Biotechnol*, **97**, 6337-6345.

35. Nair, N.U. (2010) PhD Thesis, UIUC, Urbana, IL.

36. An, J.H. and Kim, Y.S. (1998) A gene cluster encoding malonyl-CoA decarboxylase (MatA), malonyl-CoA synthetase (MatB) and a putative dicarboxylate carrier protein (MatC) in *Rhizobium trifolii*--cloning, sequencing, and expression of the enzymes in *Escherichia coli*. *Eur J Biochem*, **257**, 395-402.

37. Woo, P.C., Lam, C.W., Tam, E.W., Lee, K.C., Yung, K.K., Leung, C.K., Sze, K.H., Lau, S.K. and Yuen, K.Y. (2014) The biosynthetic pathway for a thousand-year-old natural food colorant and citrinin in *Penicillium marneffei*. *Sci Rep*, **4**, 6728.

38. Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., Kersten, R.D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J.M. *et al.* (2012) Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci U S A*, **109**, E1743-1752.

39. Tsunematsu, Y., Ishiuchi, K., Hotta, K. and Watanabe, K. (2013) Yeast-based genome mining, production and mechanistic studies of the biosynthesis of fungal polyketide and peptide natural products. *Nat Prod Rep*, **30**, 1139-1149.

40. Ishiuchi, K., Nakazawa, T., Ookuma, T., Sugimoto, S., Sato, M., Tsunematsu, Y., Ishikawa, N., Noguchi, H., Hotta, K., Moriya, H. *et al.* (2012) Establishing a new methodology for genome mining and biosynthesis of polyketides and peptides through yeast molecular genetics. *Chembiochem*, **13**, 846-854.

41. Chiang, Y.M., Oakley, C.E., Ahuja, M., Entwistle, R., Schultz, A., Chang, S.L., Sung, C.T., Wang, C.C. and Oakley, B.R. (2013) An efficient system for heterologous expression of secondary metabolite genes in *Aspergillus nidulans*. *J Am Chem Soc*, **135**, 7720-7731.

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

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

44. Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd and Smith, H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343-345.

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

CHAPTER 4. Activation of a cryptic phosphonic acid gene cluster from *Streptomyces* sp. WM6378

4.1 Introduction

4.1.1 Cryptic natural products: the silent majority

Natural products have historically served as the predominant source of leads for the development of human therapeutics [1]. Despite this fact, since the 1990s major pharmaceutical companies have either largely scaled down or completely abandoned their natural product discovery efforts [2]. Many natural products are produced at relatively low levels in their native hosts, making isolation and characterization difficult. Further, even if a compound with desirable properties is discovered, producing organisms are not always cultivable outside of their native environment, and repeated environmental sampling often does not provide a robust supply of the desired product. Producing strains can even lose their productivity over time [3]. Perhaps most significantly, rediscovery of known compounds is a common occurrence, complicating activity-based screening.

Synthetic chemical libraries, in contrast, provide large numbers of structurally defined compounds in sufficient quantities for high-throughput screening. Nevertheless, it is important to note that quantity is not equivalent to quality. In contrast to natural product libraries, synthetic chemical libraries are comprised primarily of low molecular weight, mostly planar compounds of limited stereochemical complexity [4]. For small molecules drugs to access a greater variety of biological targets, there still exists a need to access the structural diversity afforded only by natural products.

Supply of new natural products to meet this need, then, will require adoption of a new discovery paradigm beyond the classical approach of environmental sampling coupled with activity-guided screening. Thankfully, genome sequencing efforts have revealed that even among the most well-studied, laboratory-friendly microorganisms, a "silent" majority of natural products remains to be discovered. Thus, techniques that allow researchers to "look harder" at previously studied strains can provide new avenues for the discovery of novel natural products [5]. By mining sequenced genomes for key genes in natural product pathways (polyketide synthases, non-ribosomal peptide synthetases, etc.) with bioinformatic tools such as the Secondary Metabolite Unique Regions Finder (SMURF) [6] and the antibiotics & Secondary Metabolite Analysis Shell (antiSMASH) [7], new natural product gene clusters can be identified. A variety of techniques can then be employed in the native host or in a heterologous host to elicit expression of the identified genes and production of the corresponding product [8], as outlined in Chapter 1. In particular, heterologous expression in a tractable host offers the distinct advantage of liberating the genes from native regulation [9], which can be difficult to solve *a priori* in the native host.

4.1.2 Phosphonic acids as drug candidates

One class of natural products underexploited for use as human therapeutics is the phosphonic acids. These compounds are characterized by the presence of a highly stable carbon-phosphorus (C-P) bond, which can withstand enzymatic hydrolysis and harsh treatment with strong acid or base. Of note, these compounds are capable of mimicking phosphate esters and carboxylic acids, two of the most ubiquitous chemical moieties in biologically relevant molecules. As a result, phosphonic acids can serve as analogs of these compounds, and thus as potent inhibitors of a great number of enzymes [10].

In nature, phosphonic acids are found both as small molecules and as constituents of larger macromolecules, including proteins, lipids, and polysaccharides. Once viewed as "niche" natural products produced by only a handful of organisms, more recent analysis has revealed that phosphonic acid biosynthetic gene clusters are in fact relatively common in nature [11]. Although the biosynthesis of phosphonic acid macromolecules remains largely uncharacterized, several small molecule phosphonate pathways have been studied in detail. With only one known exception [10], all phosphonic acid biosyntheses begin with installation of the C-P bond via the conversion of phosphoenol pyruvate (PEP) to phosphonopyruvate (PnPy). The equilibrium of this reversible reaction lies significantly on the side of PEP; thus, a thermodynamically favorable second-step is typically employed to drive the synthesis forward. From this point forward, phosphonic acid biosyntheses diverge significantly. Some pathways employ steps analogous to primary metabolism. The FR-900098 pathway, for example, features three steps analogous to the tri-carboxylic acid (TCA) cycle, while the PTT pathway features two steps analogous to the Embden-Meyerhof-Parnas (EMP) glycolytic pathway. In contrast, other steps feature unique chemical transformations not seen elsewhere [12]. The result is a great diversity of phosphonic acid compounds.

Multiple phosphonic acid (and phosphinic acid, containing a C-P-C bond) natural products with interesting bioactivities have been discovered and characterized (Figure 4.1). Phosphinothricin tripeptide (PTT, also known as bialaphos), for example, is an herbicide produced by multiple strains [13,14]. Fosfomycin, originally called phosphonomycin, was first identified from multiple *Streptomyces* species and later found to be produced by *Pseudomonas syringae* [15,16]. Heterologous expression in *Streptomyces lividans* enabled identification of the minimal gene

cluster from *Streptomyces fradiae* [17]. Currently, fosfomycin is a clinically-approved antibiotic for urinary tract infections and has also been found to be effective against resistant strains of *Staphylococcus aureus*. Another phosphonate antibiotic is dehydrophos, first identified in 1984 [18]. This compound demonstrates broad-spectrum activity, and its biosynthetic cluster has been heterologously expressed and characterized. Other examples of biologically active phosphonic acids include the rhizocticins [19], which have antifungal properties, and FR-900098 and fosmidomycin, which have undergone clinical trials as antimalarials [20].

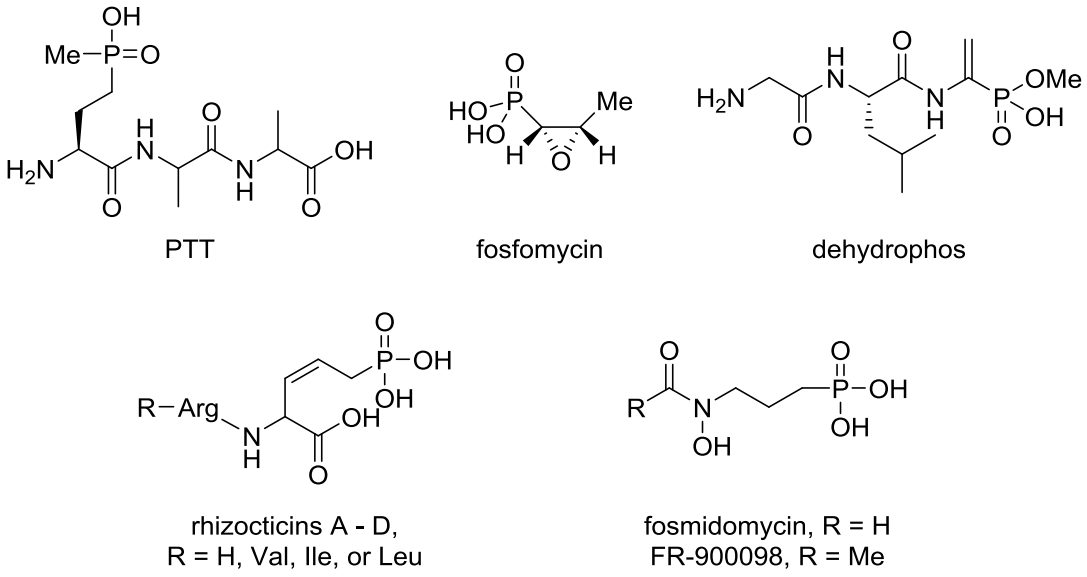


Figure 4.1: Representative bioactive phosphonic/phosphinic acid compounds.

A common property of bioactive phosphonic acid natural products is the pairing of the phosphonate moiety with another biologically relevant feature. For example, fosfomycin employs an epoxide to covalently modify its target enzyme, MurA, leading to irreversible inactivation [10]. FR-900098 and fosmidomycin feature hydroxamic acid moieties to enable chelation of the active site metal in their target enzyme, 1-deoxy-D-xylulose-5-phosphate

reductoisomerase (Dxr), to enhance binding affinity. Finally, a number of phosphonic acids, including PTT, rhizocticin, plumbemycin [21], and fosfazinomycin [22], feature conjugation of the phosphonic acid to a short peptide. This allows the compound to easily pass through the membranes of competitor microbes via peptide uptake transporters, at which point the toxic phosphonic acid payload is hydrolytically released [23].

4.1.3 Phosphonic acid discovery

The discovery of phosphonic acids is aided by the presence of two unique handles. On the genetic level, the necessity of a *pepM* gene allows identification of phosphonic acid gene clusters from both sequenced genomes (via bioinformatic analysis) and genomic DNA libraries (via PCR-based screening) [10]. On the molecular level, phosphonates exhibit chemical shifts far downfield of phosphates in a ^{31}P-NMR spectrum. This signature allows rapid confirmation of phosphonic acid production, in contrast with bioactivity- or HPLC/MS-based screening methods that can be easily confounded by unrelated molecules.

Previous work in the Metcalf lab at the University of Illinois, Urbana-Champaign, has led to the discovery of hundreds of bacterial strains with phosphonic acid biosynthetic potential, as determined via the detection of *pepM* homologs in their genomes. Included in this collection are two nearly identical *Streptomyces* strains (based on 16S ribosomal DNA identity) named *Streptomyces* spp. WM6378 and WM6391. Both strains possess a putative 15-gene phosphonic acid gene cluster that was identified and sequenced from a fosmid library. However, cultivation of the native strains under various conditions did not yield detectable production of phosphonic

acid compounds, as assessed by ^{31}P-NMR. Further studies of the intact cluster in a heterologous *S. lividans* host turned up similar results.

The goal of the work presented here is to apply synthetic biology techniques to decouple the cryptic phosphonic acid cluster from *Streptomyces* sp. WM6378 from native regulation for the purpose of discovering novel phosphonic acid compounds. First, complementation studies with the FR-900098 pathway will be employed to assess the functionality of the three *frb* homolog genes. Next, a refactoring strategy will be employed to replace all of the endogenous promoters in the gene cluster with strong, characterized promoters for heterologous expression in *S. lividans*. 1D and 2D NMR experiments will then be used to identify phosphonic acid compounds in culture extracts based on their characteristic downfield chemical shifts, and potentially interesting compounds will be isolated for further characterization.

4.2 Results and Discussion

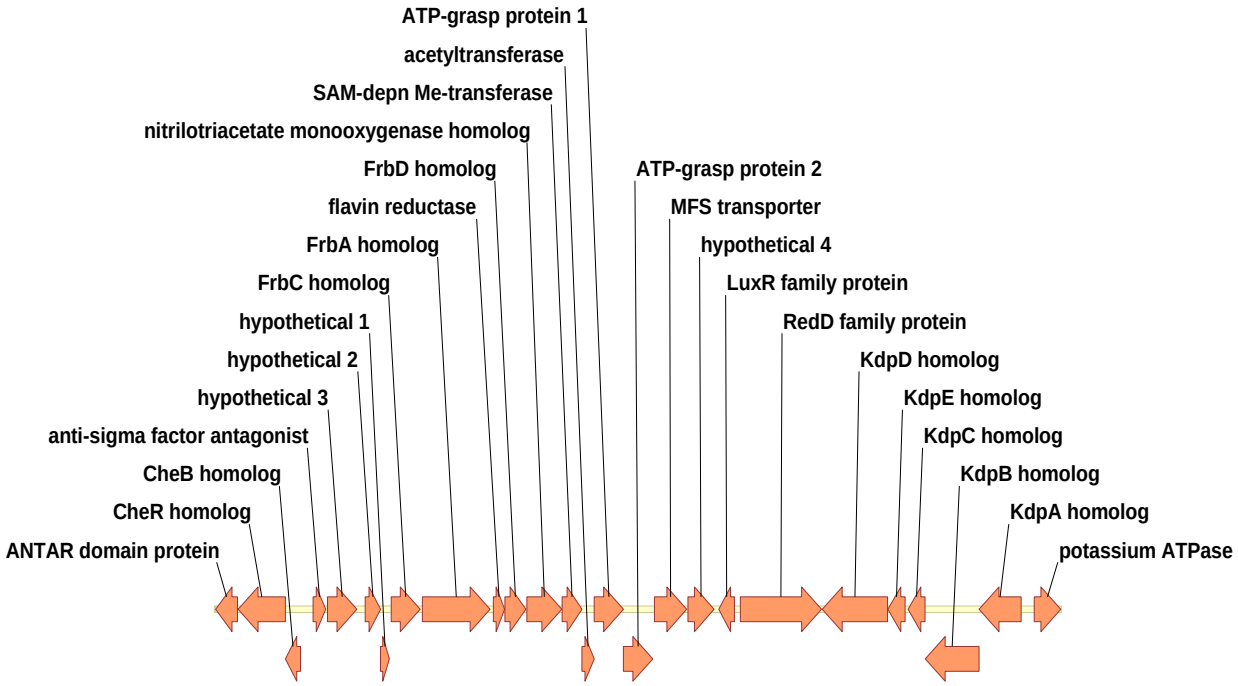
4.2.1 Native strain analyses

4.2.1.1 Bioinformatic analysis

To facilitate prediction of the function of each gene in the cryptic phosphonic acid gene cluster, each gene was annotated by querying the NCBI database via BLAST (Figure 4.2). Three genes were found to have very high sequence similarity (> 70 % identity) with *frbD*, *frbC*, and *frbA* from *Streptomyces rubellomurinus*, which encode the first three transformations of FR-900098 biosynthesis from phosphoenolpyruvate (PEP) to 3-phosphonomethyl malate [24]. Thus, analogous transformations are anticipated in the new pathway. However, no other genes in the cluster share significant sequence similarity with genes from the *frb* cluster, suggesting that the

two pathways diverge at this point. Also found in this cluster are two genes encoding putative members of the ATP-grasp superfamily, sharing weak homology (e-value > 10^{-20}) with *rhiC* of the rhizocticin biosynthetic pathway. These genes suggest the possibility of amino acid conjugation to the phosphonic acid core molecule, a common strategy observed in several phosphonic acid pathways to facilitate cellular uptake.

Additional genes in the cluster are proposed (based on conserved domain analysis) to encode a SAM-dependent methyltransferase, an acetyltransferase, a nitrilotriacetate (NTA) monooxygenase (NTA-Mo) homolog, and a flavin reductase. The remaining four putative biosynthetic genes contain no detectable conserved domains, making functional prediction more difficult. Of these four hypothetical proteins, one of these shares weak homology with some putative acetyltransferases; for the others, no putative functions can be assigned. Also encoded in the cryptic cluster is a transporter of the major facilitator superfamily and putative regulators of the LuxR, RedD and anti-sigma factor antagonist families. Outside of the putative borders of the cluster are *cheB*/*cheR* homlogs related to chemotaxis, and several homologs of *kdp* genes related to potassium transport. Of note, all of the genes in the putative cluster (between the CheB homolog the LuxR family regulator) are transcribed in the same direction. Further, intergenic sequences between the genes encoding the flavin reductase and FrbD homolog, as well as between the genes encoding the NTA-Mo homolog, SAM-dependent methyltransferase, acetyltransferase, ATP-grasp protein 1, and ATP-grasp protein 2, are all less than 10 bp in length, suggesting operonic transcription patterns for these genes.



Name	Abbreviation	Conserved Domains	Top Blast Hits
hypothetical 3	hyp3	none	conserved hypothetical protein [Streptomyces sviceus ATCC 29083], 2e-63, 58%ID
hypothetical 2	hyp2	none	hypothetical protein SCAB_52611 [Streptomyces scabiei 87.22], 9e-53, 75%ID
hypothetical 1	hyp1	none	hypothetical protein SEPMUDRAFT_148397 [Sphaerulina musiva SO2202], 0.084, 48%ID
FrbC homolog	frbC	DRE_TIM_metallolyase	FrbC [Streptomyces rubellomurinus], 0.0, 85%ID
FrbA homolog	frbA	aconitase	FrbA [Streptomyces rubellomurinus], 0.0, 71%ID
flavin reductase	FR	FMN_red	NADPH-dependent FMN reductase family protein [Saccharopolyspora spinosa NRRL 18395], 9e-49, 56%ID
FrbD homolog	frbD	ICL_PEPM	FrbD [Streptomyces rubellomurinus], 3e-139, 75%ID
nitrilotriacetate monooxygenase homolog	FMO	Nitrilotriacetate_monoxygenase	xenobiotic compound monooxygenase YxeK [Pantoea ananatis PA13]; 2e-178, 55%ID
SAM-depn Me-transferase	MT	AdoMeT_Mtase	SAM-dependent methyltransferase [Hahella chejuensis KCTC 2396], 2e-81, 51%ID
acetyltransferase	AT	NAT_SF	hypothetical protein plu1872 [Photorhabdus luminescens subsp. laumondii TTO1], 2e-29, 40%ID
ATP-grasp protein 1	Nik1	PRK02186 (arginosuccinate lyase)	hypothetical protein [Streptomyces nanchangensis], 3e-118, 47%ID; Carbamoyl-phosphate synthase large chain [Erwinia amylovora ATCC BAA-2158], 2e-87
ATP-grasp protein 2	Nik2	PRK02186 (arginosuccinate lyase)	hypothetical protein plu1874 [Photorhabdus luminescens subsp. laumondii TTO1], 9e-83, 42%ID
hypothetical 4	hyp4	none	hypothetical protein plu1876 [Photorhabdus luminescens subsp. laumondii TTO1], 4e-71, 38%ID

Figure 4.2: The annotated cryptic phosphonic acid cluster from *Streptomyces* sp. WM6378.

4.2.1.2 Cultivation and RT-PCR

To confirm previous observations from the Metcalf lab, initial experiments were conducted with the native *Streptomyces* sp. WM6378 host. Cultivation of this strain on MYG medium did not yield any peaks in the phosphonic acid range of a ^{31}P-NMR trace, as expected. Further, HPLC-MS/MS analysis did not reveal any significant signal corresponding to 2-phosphonomethyl malic acid or 3-phosphonomethyl malic acid, pathway intermediates expected to be formed by the FrbD, FrbC, and FrbA homologs. To check for pathway expression at the transcript level, reverse-transcriptase PCR (RT-PCR) was performed. Complementary DNA (cDNA) was synthesized from mRNA collected at two time points (2 days and 5 days post-inoculation), and used as template for diagnostic PCR reactions (Table 4.1). Some transcripts were detected only at one time point or the other, while others were not detected at all. Interestingly, transcription of the three *frb* homologs was detected at both time points. Nevertheless, inability to detect the corresponding FR-900098 pathway intermediates called into question the activity of the encoded enzymes, a property that cannot be evaluated at the transcript level.

Table 4.1: RT-PCR detection of transcription in *Streptomyces* sp. WM6378

Gene	2 days	5 days
Anti-sigma factor antagonist	+	+
FrbC	+	+
FrbA	+	+
Flavin reductase	–	+
FrbD	+	+
Monooxygenase	–	+
Methyltransferase	–	–
Acetyltransferase	–	–
ATP-grasp protein 1	+	–
ATP-grasp protein 2	+	–

4.2.2 FR-900098 complementation

In principle, there are three main reasons for a gene cluster to be cryptic. First, the cluster could be completely silenced at the genetic level through inactivating mutation(s). Second, the cluster could be silenced at the transcriptional or translational levels through any number of regulatory pathways. Third, the cluster could in fact be productive, but at a level that falls short of the limit of detectability. To evaluate the possibility of this first means of crypticity, a complementation strategy was devised. Most of the genes in the cluster encode enzymes for which the native substrate cannot be predicted; however, the high similarity of the *frbD*, *frbC*, and *frbA* homologs to their *S. rubellomurinus* counterparts allows for more confident prediction of their probable functions. Thus, by first knocking out individual genes in the FR-900098 pathway and then introducing the corresponding homolog from the WM6378 pathway, production of FR-900098 can serve as an indicator of functional expression of the complemented gene.

Previously, the FR-900098 pathway was reconstituted in *E. coli* on three compatible plasmids bearing *frbABCD*, *frbFGH*, and *frbE*/*dxrB*, respectively [25]. Due to ease of manipulation and cultivation, this system was utilized for the complementation experiments. Each target gene was replaced on the pETDuet-*frbABCD* plasmid and reintroduced to the expression strain along with the other two FR-900098 pathway plasmids. HPLC-MS/MS screening revealed restored production of FR-900098 in both the *frbD* homolog and *frbC* homolog complementation strains (Figure 4.3). Interestingly, only a very small peak above background was observed for *frbA* homolog complementation. While this could be a sign of reduced activity, it could also be an artifact of expression in the non-native *E. coli* host, or evidence that the *frbA* homolog actually catalyzes a different reaction than anticipated. Nevertheless, these studies provided the first

evidence of biosynthetic potential from this gene cluster, inspiring a total cluster refactoring approach.

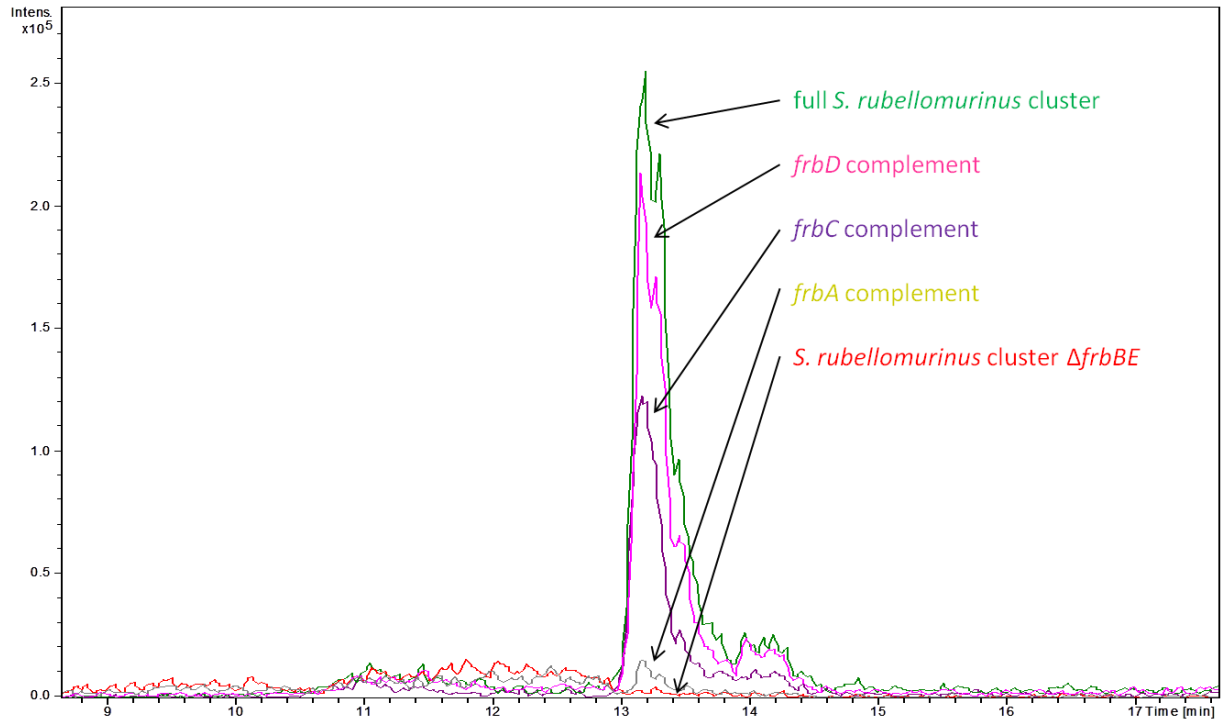


Figure 4.3: Complementation the FR-900098 pathway in *E. coli* with *frb* homologs. EIC traces show m/z = 198→138 fragmentation characteristic of FR-900098.

4.2.3 Refactored plasmid design and assembly

4.2.3.1 Design

Having ruled out inactivating mutation of the *frb* homologs as an explanation for the observed silence of the WM6378 phosphonic acid gene cluster, it was presumed that the cluster was likely rendered silent either by intrinsic repression in the native host or by undetectably low production levels. Either of these problems can be addressed through a full cluster refactoring strategy, in which the cluster is simultaneously removed from native regulation via transfer to a heterologous

host and up-regulated via replacement of native promoters with strong promoters. Due to its high genetic tractability and lack of intrinsic phosphonic acid biosynthetic pathways, *S. lividans* 66 was chosen as the heterologous host.

Initially, all of the structural genes, with the exception of putative regulators and the MFS transporter, were included in the refactored cluster design. In total, assembly of this construct required 29 components: 13 structural genes, 13 promoters, and three helper fragments for plasmid maintenance in *Saccharomyces cerevisiae*, *E. coli*, and *S. lividans*, respectively. Of the panel of actinomycete promoters available [26], the three strongest promoters were paired with the *frb* homolog genes with the goal of pulling flux into the pathway. The next strongest promoters were paired with the two ATP-grasp proteins, the methyltransferase, the monooxygenase, and the acetyltransferase, which seem likely to participate in biosynthesis. Finally, the remaining promoters were paired with the four hypothetical proteins, which may not have an actual function in biosynthesis.

4.2.3.2 Assembly

To reduce the number of fragments necessary for assembly, each gene was first concatenated with its corresponding promoter, reducing the overall number of parts to 16. Assembly of these parts was attempted in one step via yeast homologous recombination. Although numerous plasmids could be recovered, diagnostic digestion repeatedly revealed that many fragments had been left out of the assembled construct. Suspecting the large number of parts as the main reason for failure, a hierarchical assembly strategy was then designed and implemented.

In this strategy, the pathway was divided into two intermediate plasmids containing five and eight gene cassettes, respectively, along with helper fragments for *S. cerevisiae* and *E. coli* (Figure 4.4). One helper fragment in each plasmid was flanked with unique restriction sites to facilitate linearization and assembly into the full-cluster plasmid via a bridging "master" helper fragment in the second assembly step. This strategy proved to be far more successful than the one-step method. For both of the intermediate plasmids, > 70 % assembly fidelity was achieved, as confirmed by diagnostic restriction digestion. Assembly of the final plasmid was also successfully achieved, albeit with significantly lower fidelity.

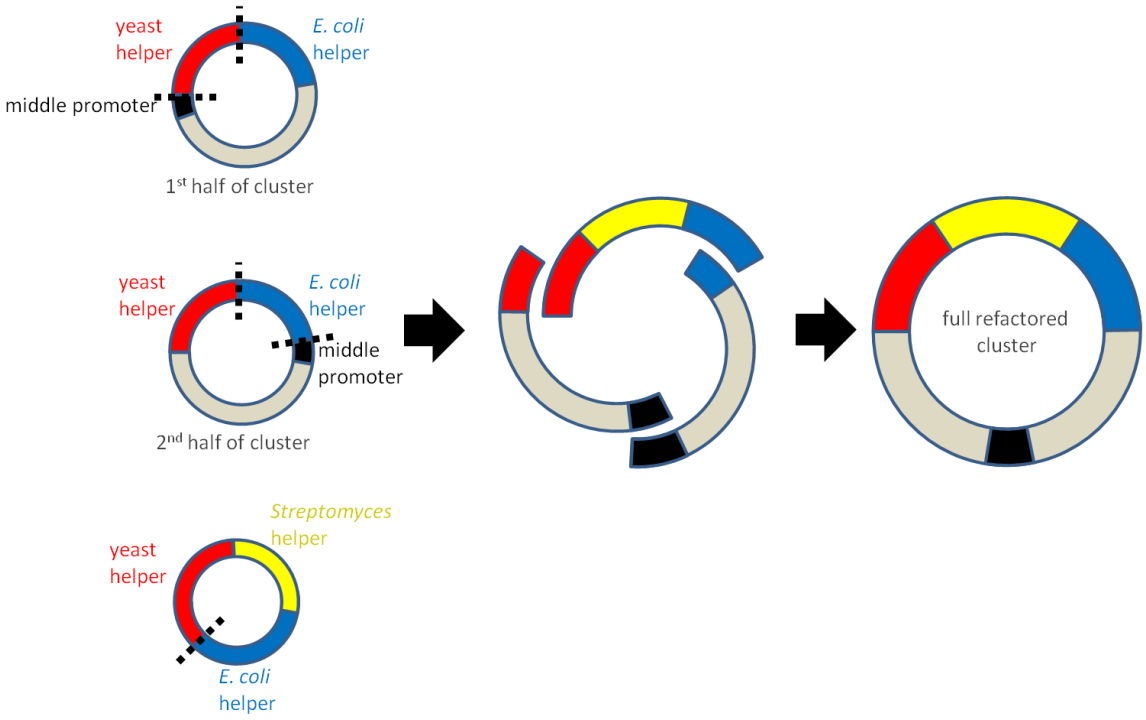


Figure 4.4: Two-step scheme for assembly of the full refactored WM6378 gene cluster. Dashed lines indicate unique restriction sites.

4.2.4 Identification of phosphonic acid products

4.2.4.1 Initial NMR analyses

Following assembly of the full refactored 13-gene cluster, it was integrated to the chromosome of *S. lividans* 66 at the ΦC31 phage attachment site via conjugation from an *E. coli* donor. The resulting strain, named *S. lividans* 3p2.1, was grown in liquid culture to detect the production of phosphonic acids. Following six days of growth, both the dried culture supernatant and cell pellet were extracted with methanol, dried, and resuspended in D_2O. Analysis by ^{31}P-NMR revealed the presence of multiple peaks in the downfield region characteristic of phosphonic acids (Figure 4.5). Analogous cultivation of the native strain, as well as the heterologous *S. lividans* host without the integrated cluster, confirmed no phosphonic acid production in either of these strains.

As further confirmation, a negative control plasmid was assembled in which a single nonsense mutation was introduced into the *frbD* homolog gene (the PEP phosphonomutase) to eliminate the first step of the heterologous pathway. Assembly of this plasmid was analogous to that of the full-cluster plasmid in that an intermediate plasmid bearing the desired mutation was generated first, followed by a three-piece assembly of the final plasmid. The resulting integrated strain, named *S. lividans* 3p2.1-Da, was found not to produce any phosphonic acids, as anticipated (Figure 4.5).

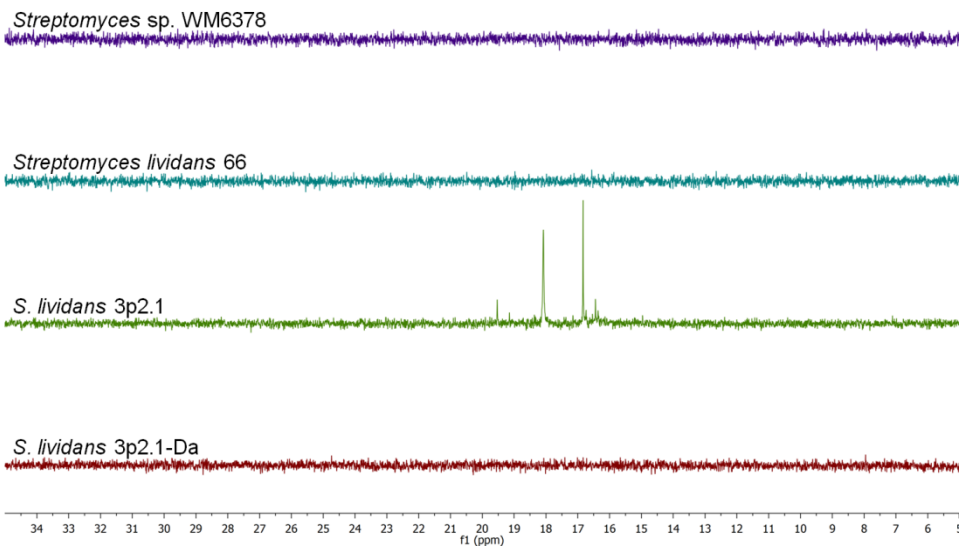


Figure 4.5: ^{31}P-NMR analysis of the full-cluster integrated strain *S. lividans* 3p2.1 with negative controls.

Having observed phosphonate production in the presence of the full refactored cluster, a time-course study was next undertaken to see if any clues could be obtained about the order of the compounds observed in the context of the pathway. Collecting samples every 24 hours over a course of 12 days revealed that the peak at ~17.5 ppm appeared earliest, followed by several smaller peaks (Figure 4.6). As the ~17.5 ppm peak subsequently reduced in size, the signals at ~16.5 ppm and ~18.2 ppm became the dominant peaks in the spectrum, suggesting that these compounds could be closer to the final product of the cluster. Alternatively, such peaks could also correspond to shunt products that accumulate gradually due to off-pathway activities.

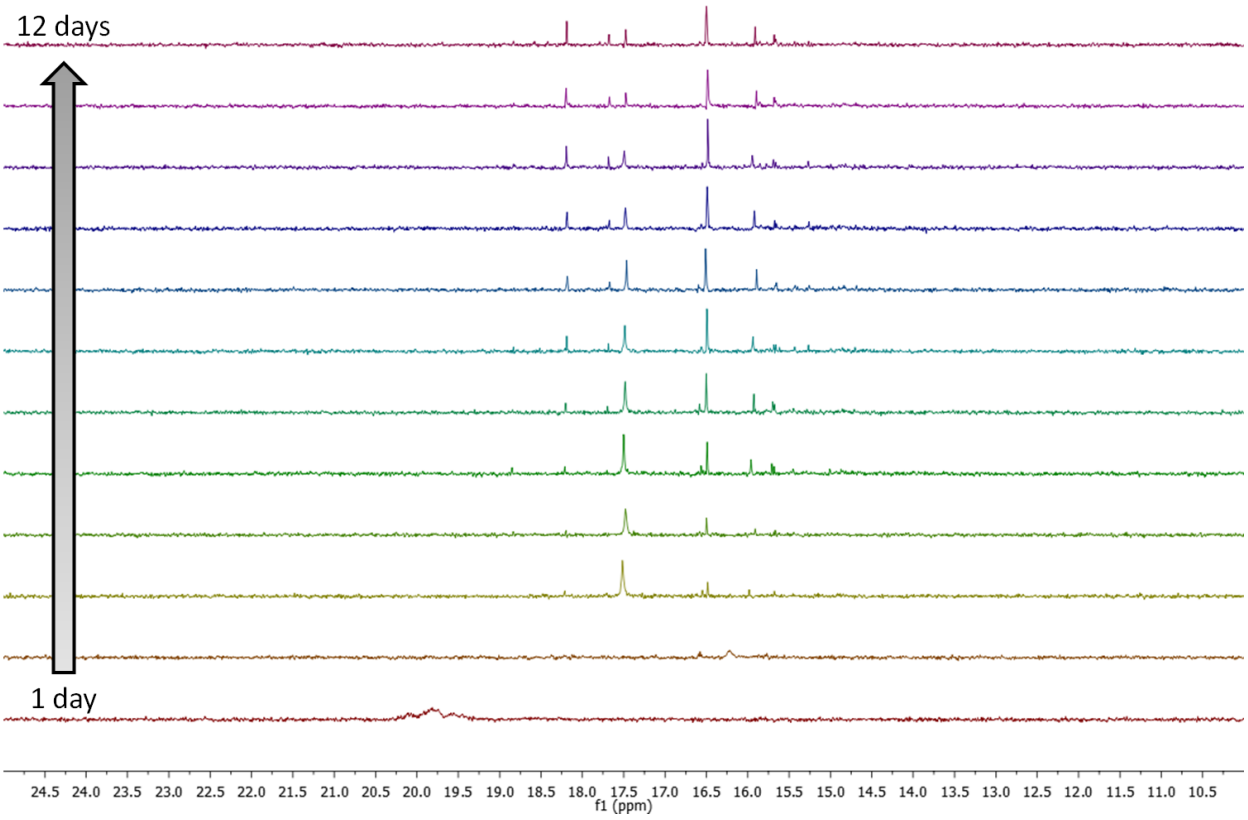


Figure 4.6: ^{31}P-NMR time-course analysis of *S. lividans* 3p2.1.

4.2.4.2 Gene disruption experiments

In a parallel set of experiments, a series of single-gene disruption plasmids was prepared to see if the pathway could be stalled at different intermediate compounds. In each case, an amber stop codon was introduced early in the target gene to silence it at the translational level while minimally perturbing transcription dynamics. Surprisingly, most of the disruptions introduced did not reproducibly affect the phosphonic acid profile, with the only exceptions being of the *frbC* and *frbA* disruption constructs (Figure 4.7). In the *frbA* disruption strain, a peak at ~15.5 ppm was observed, presumably corresponding to the anticipated 2-phosphonomethyl malic acid product. ^{1}H-^{31}P HMBC revealed coupling of this peak to two proton signals at ~1.9 ppm, as would be expected for this compound (data not shown). Further, another strain expressing only

the *frbD*, *frbC*, and *frbA* homologs did not produce the ~15.5 ppm peak, instead yielding a peak at ~19 ppm.

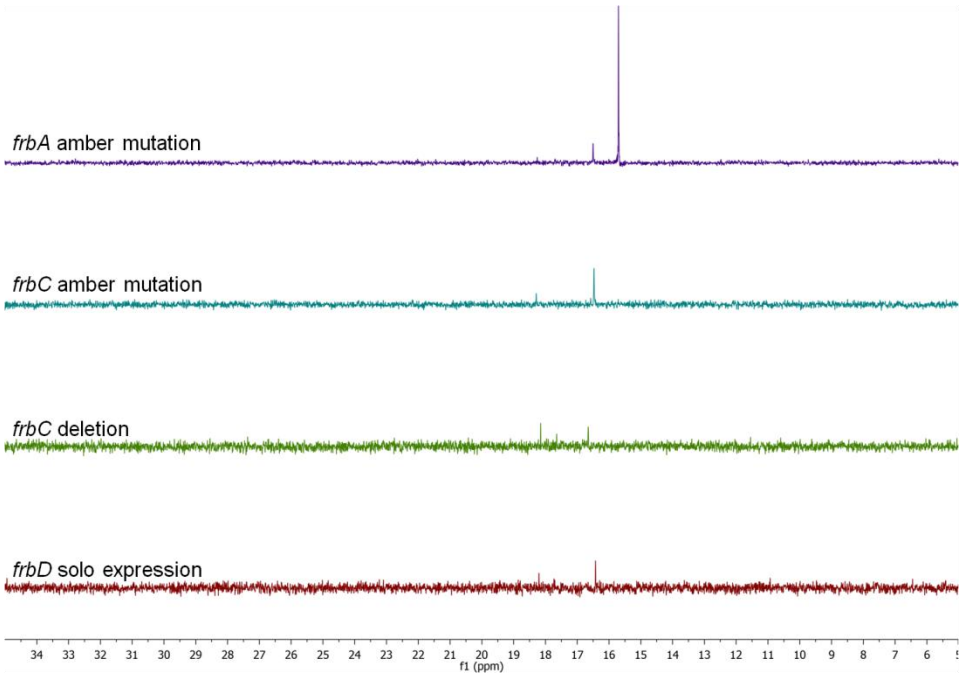


Figure 4.7: Disruption/deletion analysis of early-pathway genes.

Another peak was observed from the *frbA* disruption strain and the *frbDCA* expression strains at ~16.5 ppm. Curiously, it was also seen in the *frbC* disruption strain. As the *frbD*-catalyzed conversion of PEP to phosphonopyruvate is a thermodynamically disfavored process, production of phosphonic acids from the *frbC* disruption strain (presumed to be lacking the subsequent conversion of phosphonopyruvate to 2-phosphonomethyl malic acid) was unexpected. Suspecting that the peak might be due to read-through of the introduced stop codon in *frbC*, an additional strain was created in which the *frbC* homolog and its corresponding promoter were cleanly deleted from the integrated cluster. This strain produced the same ^{31}P-NMR peaks (Figure 4.7 above), confirming their derivation from off-pathway reactions. For further

confirmation, a strain was created in which only the *frbD* homolog expression cassette was integrated to the chromosome in the absence of the other twelve genes. Again, the same peaks were observed in the ^{31}P-NMR spectrum, confirming them as shunt products of endogenous transformations.

4.2.4.3 Structure elucidation

To gain some insight into the structures of the detected products from the full-cluster strain *S. lividans* 3p2.1, ^{1}H-^{31}P HMBC was also employed. As shown in Figure 4.8, the phosphorus peak at ~18 ppm is coupled to proton signals at 2.5 ppm and 5.8 ppm, suggesting a non-functionalized α-carbon and a double bond on the β-carbon, respectively. The peak at ~16.5 ppm is coupled only to a proton signal at 1.8 ppm.

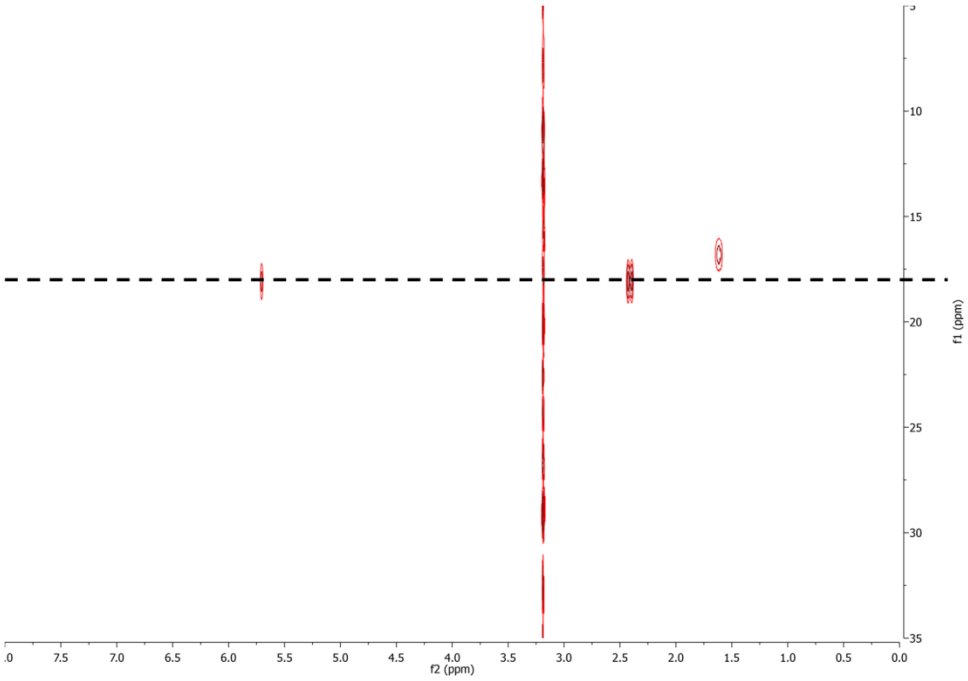


Figure 4.8: ^{1}H-^{31}P HMBC analysis of *S. lividans* 3p2.1.

To obtain enough of the phosphonic acid products for isolation and structural characterization, large-scale cultivation was employed. Through the efforts of Dr. Jiangtao Gao of the Wilfred van der Donk lab at the University of Illinois, Urbana-Champaign, a total of six compounds were isolated from a combined 50 L of culture supernatant (Figure 4.9). The major product was found to be phosphonomethyl fumaric acid (compounds **1** and **6**), which is presumably the product of only three enzymes in the pathway (the *frb* homologs). Surprisingly, 3-phosphonomethyl malic acid, the known product of FrbD, FrbC, and FrbA in the FR-900098 pathway, was not observed. In the FR-900098 pathway, the aconitase-like FrbA converts 2-phosphonomethyl malic acid to 3-phosphonomethyl malic acid in two steps. First, the 2-phosphonomethyl malic acid substrate binds in the active site, undergoes dehydration, and is released in the dehydrated form. In the second step, the dehydrated intermediate binds in the active site in a 180° reversed orientation, undergoes rehydration (at a different position, using the same catalytic residues), and is released as the 3-phosphonomethyl malic acid product. Based on the products obtained here, the WM6378 FrbA homolog appears to bind the substrate in only one orientation, such that the second-step rehydration reaction cannot occur. Although the phosphonomethyl fumaric acid intermediate seen here is presumed to exist in the FR-900098 pathway, its observation has not previously been reported in any natural product pathway.

Small amounts of a 2-phosphonomethyl malic acid methyl ester (**3**) were observed, suggesting activity of the heterologous methyltransferase. However, the other compounds produced by *S. lividans* 3p2.1 appear to be shunt products of the first or second transformations. Of note, compounds **3** and **4** have also not been previously reported in natural product literature. Compound **5** has previously been observed in protozoans of the genus *Tetrahymena*, but not in a

bacterial system [10]. It remains unclear, however, how such a compound could appreciably accumulate given the thermodynamic unfavorability of the first transformation and the reversibility of the presumed transamination second step.

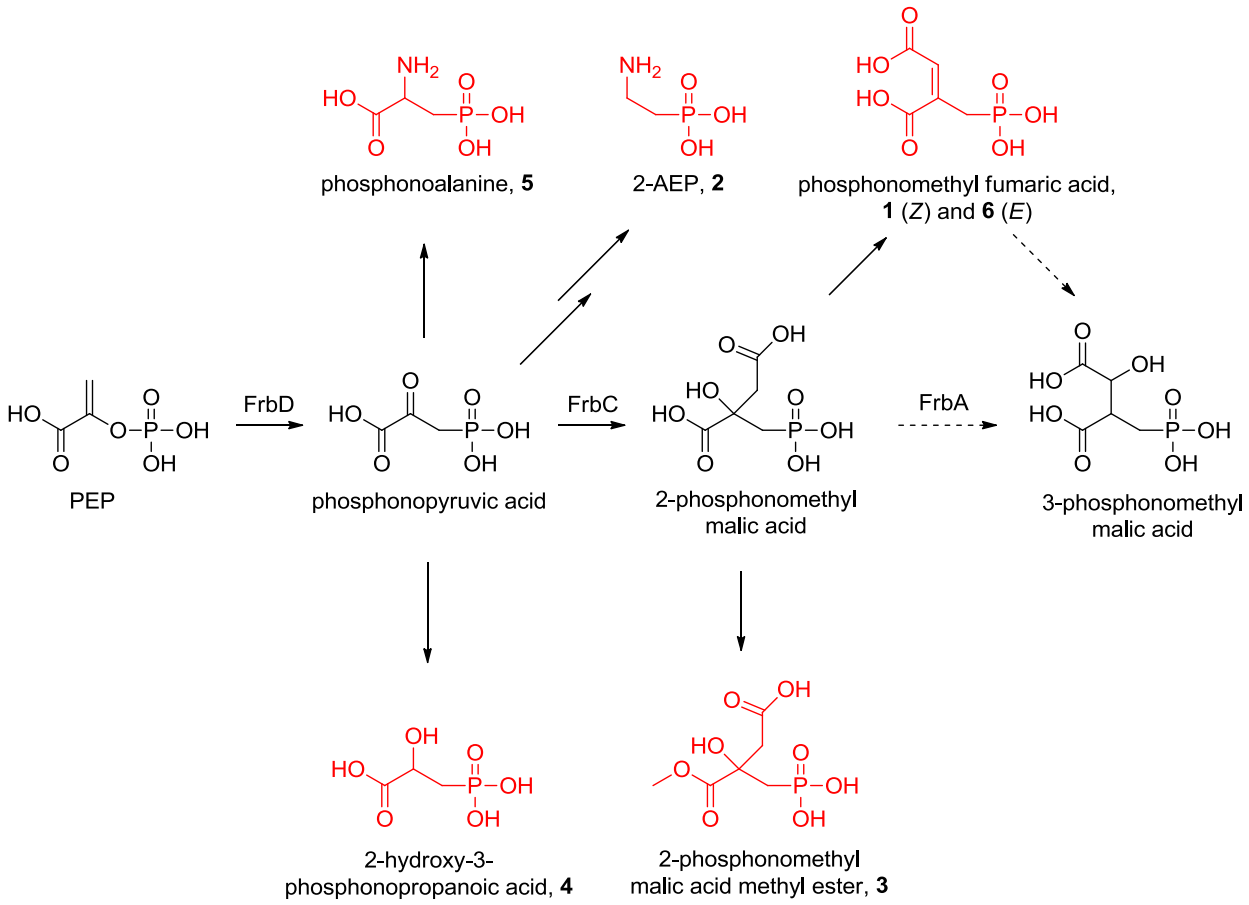


Figure 4.9: Phosphonic acids isolated from *S. lividans* 3p2.1. Isolated compounds are numbered and shown in red. Putative intermediates from the FR-900098 pathway are shown in black.

Based on the shunt products observed from *S. lividans* 3p2.1, it was speculated that the peaks observed in the *frbC* deletion strain and *frbD* solo-expression strain (described in Section 4.2.4.2) could correspond to 2-aminoethyl phosphonic acid (2-AEP), phosphonoalanine, or 2-hydroxy-3-phosphonopropanoic acid. Spiking of the commercially available 2-AEP and phosphonoalanine

compounds into the extract from the *frbC* deletion strain confirmed the ~16.5 ppm peak to be 2-AEP (Figure 4.10), while the small downstream peak (though unconfirmed by spiking) presumably corresponds to 2-hydroxy-3-phosphonopropanoic acid. Thus, it was discovered that *S. lividans* 66 possesses the native ability to metabolize phosphonopyruvate to 2-AEP, presumably through ubiquitous decarboxylation and transamination. This discovery could serve to inform future metabolic engineering studies to optimize phosphonic acid production in *S. lividans*.

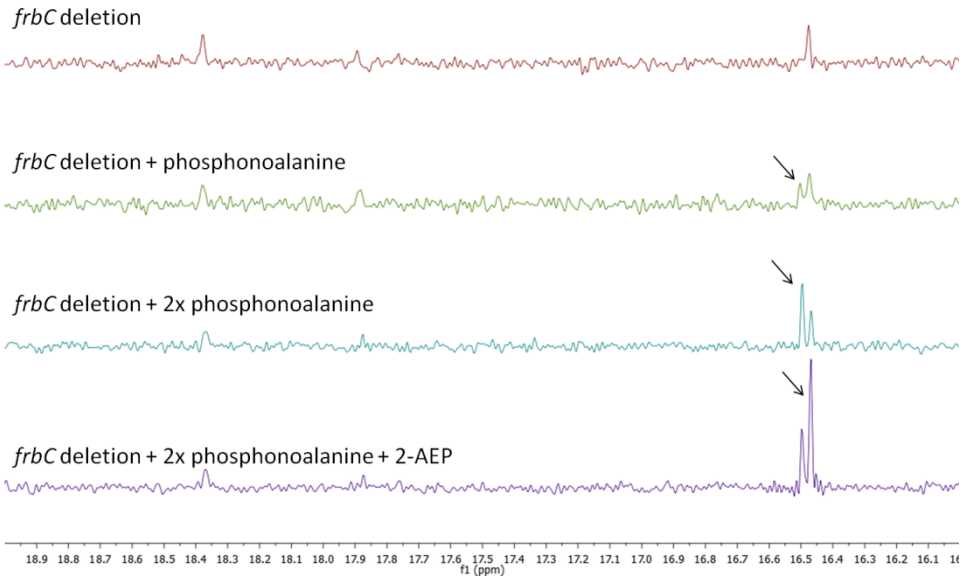


Figure 4.10: Spiking of known phosphonic acids into the *frbC* deletion strain extract. Arrow indicates the position of the spiked compound.

4.2.5 FMO studies and identification of 2-(hydroxy(phosphonomethyl)) fumaric acid

4.2.5.1 E. coli expression

The previous pathway refactoring effort left many genes without observed functions, including the two ATP-grasp domain proteins, the NTA-Mo homolog and flavin reductase, and two

putative acetyltransferases (including hypothetical protein 3). Thus, subsequent efforts focused on the further analysis of these genes and their encoded proteins. Following the structural elucidation of compound **1**, its similarity with nitrilotriacetic acid was noted. It was then hypothesized that the next step in the pathway could be a hydroxylation analogous to that catalyzed by NTA-Mo (Figure 4.11) [27].

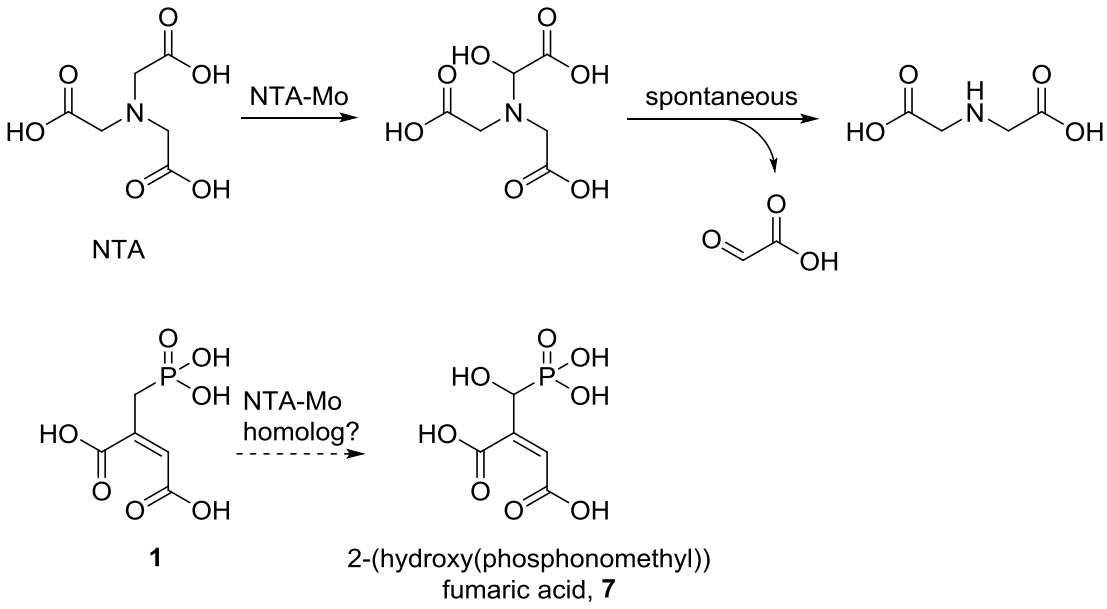


Figure 4.11: The NTA-Mo catalyzed hydroxylation of NTA (top) with the proposed analogous hydroxylation of **1**.

To investigate this possibility, the NTA-Mo homolog gene and the flavin reductase presumed to be its partner were cloned into *E. coli* expression vectors for *in vitro* assays. While the flavin reductase could be expressed and purified, it was found that the expression NTA-Mo homolog yielded a protein of the incorrect size, as determined by SDS-PAGE (Figure 4.12a). Sequencing of the cloned gene revealed an unintended nonsense mutation, which when corrected enabled synthesis of the full-length protein (Figure 4.12b). However, while oxidation of NADH (but not NADPH) was observed by the flavin reductase in the presence of flavin adenine dinucleotide or

flavin mononucleotide (data not shown), coupled *in vitro* assays with the NTA-Mo homolog and the purified compound **1** did not yield any detectable oxidation of the substrate. Subsequently, the two genes were cloned into a single vector for co-expression in an *E. coli* strain containing the *phnCDE* phosphonate uptake transporter under inducible T7 promoters. Feeding of **1** with concomitant induction of protein expression also did not yield any detectable turnover of the substrate.

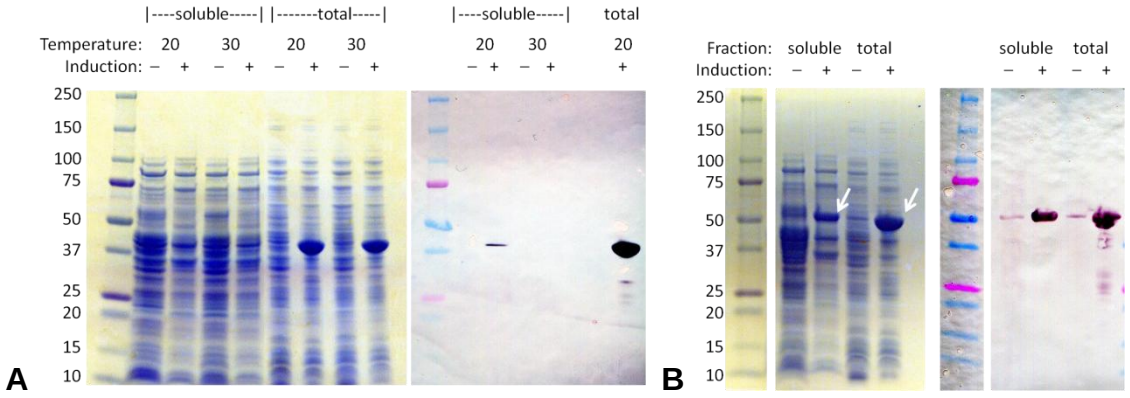


Figure 4.12: SDS-PAGE and Western blot analysis of NTA-Mo homolog expression (a) before and (b) after correction of a nonsense mutation. White arrow indicates the full-length protein (51 kDa).

4.2.5.2 S. lividans expression

Having detected a mutation in the cloned NTA-Mo homolog gene for *E. coli* expression, the sequence was also checked in the 3p2.1 strain. When the same nonsense mutation was found to be present, the full-cluster plasmid was reassembled and re-integrated to *S. lividans* 66 to yield strain *S. lividans* 3pf. ^{31}P-NMR analysis of this strain revealed the production of a new peak with a chemical shift at ~13.2 ppm (Figure 4.13). ^{1}H-^{31}P HMBC revealed two proton signals coupled with the phosphorus: one at 6.2 ppm, indicative of a double-bond, and one at 4.4 ppm, suggesting the presence of the expected hydroxyl group on the α-carbon. Isolation of the new

compound confirmed the production of 2-(hydroxy(phosphonomethyl)) fumaric acid (**7**). This structure has never been previously reported in natural product or synthetic chemistry literature.

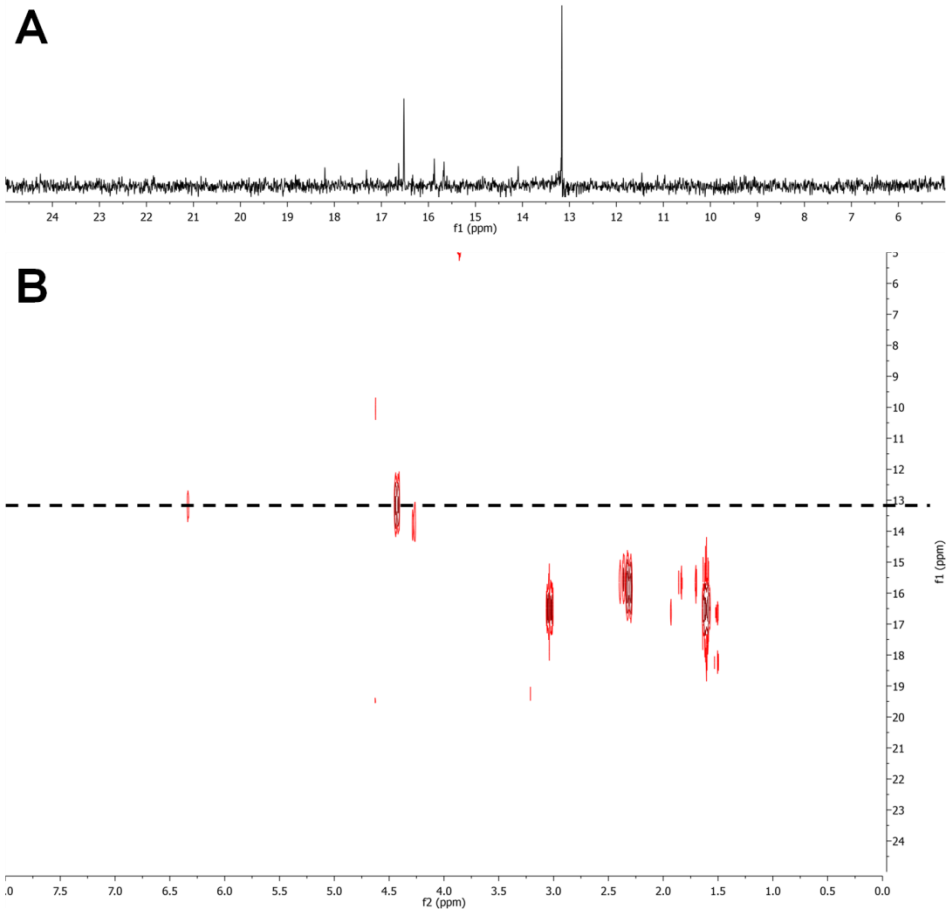


Figure 4.13: ^{31}P-NMR (a) and ^{1}H-^{31}P HMBC (b) analysis after correction of the NTA-Mo homolog gene.

4.2.6 Further refactoring iterations

While repair of the NTA-Mo homolog gene resulted in the discovery of one additional compound, it still did not result in any observable activity for the two ATP-grasp domain proteins and the two putative acetyltransferases. As a result, multiple strategies were adopted to try to increase expression of these genes. First, the entire refactored 13-gene cluster was moved to an episomal plasmid with a pIJ101 origin, boosting the copy number from a single integrated

copy to 50 – 300 copies [28]. The resulting exconjugants, however, showed significantly reduced growth rate, perhaps owing to the metabolic burden of thirteen strong expression cassettes. ^{31}P-NMR analysis did not reveal any phosphonic acid production (data not shown), again suggesting unfavorable effects on overall metabolism.

In the next iteration, only the four genes with unobserved activities were cloned to a pIJ101-based episomal plasmid (plasmid 4GS). Their promoters were replaced with the four strongest promoters from the actinomycete library [26]. This plasmid was then transferred to strain 3pf. Although the resulting strain was able to produce phosphonic acids (data not shown), no new peaks were observed compared to strain 3pf, as confirmed by spiking of samples from the two strains. Further, overall phosphonate production appeared to be lower in the new strain than in strain 3pf.

In an effort to minimize the metabolic burden introduced by so many strong expression cassettes, a new host strain was generated in which only the six genes for which activities were observed (the three *frb* homologs, the NTA-Mo homolog and its putative flavin reductase, and the methyltransferase) were integrated to the chromosome. As anticipated, the resulting strain (named *S. lividans* 6378-6g) retained the ability to produce compound **7**. Supplementation of this strain with individual gene expression plasmids to identify new peaks is ongoing.

4.3 Conclusions and Outlook

Through heterologous expression with complete promoter refactoring, we have successfully demonstrated the production of phosphonic acid compounds from an otherwise silent phosphonic

acid gene cluster. Several compounds were isolated and characterized via one- and two-dimensional NMR analyses. Included among the identified products are two compounds, **3** and **7**, that have never been previously reported in natural product or synthetic chemistry literature. Additionally, we have demonstrated that *Streptomyces lividans* 66, a strain devoid of phosphonic acid biosynthetic pathways, is capable of producing multiple phosphonic acid compounds when supplied with only a PEP mutase gene.

Notably, no activities were observed for multiple genes in the refactored cluster, even after evaluating multiple promoters and copy numbers. These results highlight the inherent difficulties in heterologous expression, particularly with a large number of genes. As optimal expression levels for every gene in the pathway are difficult to predict *a priori*, an over-expression strategy is generally favored as a first-pass approach. However, this may have detrimental effects on the host's metabolism that significantly limit flux through the pathway [29]. On a more fundamental level, some proteins simply are not amenable to functional expression in a given heterologous host for reasons beyond our current level of understanding; as the number of genes in a pathway increases, so too does to likelihood that such a protein will be encountered.

Moving forward, the pathway refactoring approach demonstrated here is likely to benefit significantly from advances in DNA synthesis, DNA assembly, and process automation. The decreasing cost of DNA synthesis will facilitate refactoring not just through promoter replacement, but through recoding genes at the codon level. This will help to facilitate soluble expression of the encoded proteins in heterologous hosts, as well as to prepare standardized parts

for generalizable assembly strategies. With advances in DNA assembly protocols, different promoters and genes can be paired combinatorially to better explore the solution space. Finally, automated assembly would allow for many pathways to be evaluated simultaneously, effectively mitigating the risk of an unsuccessful refactoring effort.

4.4 Materials and Methods

4.4.1 Strains, media and reagents

The *Escherichia coli* expression strain BL21(DE3) and the expression vector pET28a were obtained from EMD Biosciences (San Diego, CA). *E. coli* strains BW25141 (for general cloning of *pir*-dependent plasmids) and WM6026 (for conjugation to *S. lividans*), along with *Streptomyces* sp. WM6378 and the corresponding phosphonate-cluster-containing fosmid, were gifts from Prof. William Metcalf (University of Illinois at Urbana-Champaign). *S. lividans* 66 was obtained from the Agricultural Resource Service Culture Collection (Peoria, IL). *E. coli* strains were grown in LB medium supplemented with the appropriate antibiotics and, for strain WM6026, diaminopimelic acid (19 μg/mL). *Streptomyces* strains were grown on modified MYG medium (10 g/L malt extract broth, 4 g/L yeast extract, 4 g/L glucose) with 50 μg/mL apramycin or thiostrepton as needed for plasmid selection. Medium R2 (without sucrose) was used for conjugation [28]. All media components and supplements were purchased from Sigma-Aldrich (St. Louis, MO) with the exception of yeast extract (BD Biosciences, San Jose, CA) and LB broth (Fisher Scientific, Pittsburgh, PA). Antibiotics and isopropyl-β-D-1-thiogalactopyranoside (IPTG) were purchased from Gold Biotechnology (St. Louis, MO). Organic solvents and consumables were purchased from Thermo-Fisher Scientific (Pittsburgh, PA). PCR primers were synthesized by Integrated DNA Technologies (Coralville, IA), and PCR

reactions were performed in FailSafe PCR PreMix G (Epicentre Biotechnologies, Madison, WI) with Q5 DNA polymerase (New England Biolabs, Ipswich, MA). All PCR products were purified using the DNA Clean & Concentrator or Zymoclean Gel DNA Recovery Kit (Zymo Research, Irvine, CA). Plasmids were recovered using the QIAprep Spin Miniprep Kit (Qiagen, Valencia, CA). Restriction enzymes and T4 ligase were purchased from New England Biolabs (Ipswich, MA). Talon Cobalt immobilized metal affinity chromatography (IMAC) resin was purchased from Clontech Laboratories (Mountain View, CA). Total RNA was isolated with the RNeasy Mini Kit (Qiagen) and treated with TURBO DNA-free DNase (Life Technologies, Carlsbad, CA). Complimentary DNA was synthesized with the ProtoScript II First Strand cDNA Synthesis Kit (NEB).

4.4.2 Plasmid construction

Construction of FR-900098 pathway complementation plasmids was performed by restriction digestion and Gibson isothermal assembly. Briefly, unique restriction sites were located upstream and downstream of the target gene. The corresponding replacement gene was PCR amplified from the WM6378 fosmid, along with bridging fragments to fill the gaps between the restriction sites and the termini of the replacement gene. Gibson isothermal assembly was performed according to literature precedent [30].

Plasmids 6378-3p, 6378-6g, and 4GS were assembled via yeast homologous recombination in *S. cerevisiae* HZ848 following previously established protocol [31]. Plasmids recovered from yeast assemblies were electroporated into *E. coli* BW25141 for confirmation by diagnostic restriction

digestion and sequencing. Correct plasmids were then electroporated into *E. coli* WM6026 for conjugation to *S. lividans*.

Single gene expression plasmids for *S. lividans* were constructed using Gibson isothermal assembly Master Mix (NEB) according to the manufacturer's suggested protocol. Single gene expression plasmids for *E. coli* were assembled via ligation of the PCR-amplified gene into the *Nde*I/*Hin*dIII sites (for the flavin reductase) or *Nde*I/*Bam*HI sites (for the NTA-Mo homolog) of pET28a. For the dual expression plasmid, the flavin reductase and monooxygenase were sequentially ligated into the *Nco*I/*Bam*HI and *Nde*I/*Avr*II sites, respectively, of pRSFDuet.

4.4.3 Transformation

E. coli strains BW25141, WM6026, and BL21(DE3), as well as *S. cerevisiae* strain HZ848, were transformed by electroporation. Conjugation of plasmids into *Streptomyces* spores was performed using the modified protocol described elsewhere [17].

4.4.4 FR-900098 complementation assays

E. coli BL21(DE3) FR-900098 complementation strains were grown in 50 mL LB medium supplemented with ampicillin, kanamycin, and chloramphenicol. Cultures were grown at 37 °C, 250 rpm to a density of 0.8, at which time the temperature was reduced to 30 °C and IPTG was added to a final concentration of 0.3 mM. After 40 hr of growth, 1 mL samples were collected, lysed by freeze-thaw at -80 °C, and centrifuged to remove cellular debris. The clarified sample was analyzed by HPLC-MS/MS as described previously [25].

4.4.5 Protein expression and purification

The *E. coli* BL21(DE3) pET28a expression strains were grown in Terrific Broth (TB) media supplemented with kanamycin (50 μg/mL) at 37 °C to an OD_{600} of ~0.8, after which induction was carried out by addition of 0.3 mM IPTG at 20 °C. After 18 – 24 hr, the cells were harvested by centrifugation at 7500 rpm for 15 min and resuspended in 20 mM Tris-HCl (pH 7.65), 0.5 M NaCl, and 15 % glycerol supplemented with 1 mg/mL lysozyme. After a freeze-thaw cycle at -80 °C, the cell suspension was sonicated to ensure sufficient lysis. The lysate was clarified multiple times by centrifugation at 15000 rpm for 15 min, after which the His_6-tagged proteins were purified by affinity chromatography on TALON Superflow Co^{2+} resin coupled to fast-performance liquid chromatography. The eluted proteins were washed three times in 50 mM HEPES (pH 7.25), concentrated, and stored in 15 % glycerol at -80 °C.

4.4.6 SDS-PAGE and Western blotting

S. lividans strains were grown in 5 mL selective MYG cultures for all protein expression studies. Cells were pelleted by centrifugation and resuspended in 500 μL B-PER reagent (Thermo Scientific, Pittsburgh, PA) for lysis following the manufacturer's suggested protocol. Lysates were analyzed on 4-20% Min-PROTEAN TGX precast polyacrylamide gels (Bio-Rad, Hercules, CA) stained with SimplyBlue SafeStain (Life Technologies, Carlsbad, CA). Western blotting was performed with His-probe primary antibody sc-8036 (Santa Cruz Biotechnology, Dallas, TX) and goat anti-mouse–alkaline phosphatase secondary antibody A5153 (Sigma-Aldrich), and visualized with Western Blue stabilized substrate for alkaline phosphatase (Promega, Madison, WI).

4.4.7 *In vitro* enzyme assays

Flavin reductase assays were performed in 800 µL of 30 mM HEPES buffer, pH 7.8, with 300 µM NAD(P)H, 30 µM flavin cofactor (FMN/FAD), and ca. 0.5 µM flavin reductase. Decrease in absorbance at 340 nm was measured for 1 – 2 minutes at room temperature following addition of the enzyme. Monooxygenase assays were performed in 400 µL of 30 mM HEPES buffer, pH 7.8, with 300 µM NAD(P)H, 30 µM flavin cofactor (FMN/FAD), 2 mM $MgCl_2$, ca. 10 µM monooxygenase, and ca. 1 µM flavin reductase. Samples were incubated at 30 °C for 4 – 18 hr (for aerobic reactions) or at room temperature for 18 hr (for anaerobic reactions). For some trials, 10 mM glucose and 5 nM glucose dehydrogenase were added for cofactor regeneration. Samples were desalted with Chelex 100 resin (Bio-Rad) and spiked with 100 µL D_2O for ^{31}P-NMR analysis.

4.4.8 *Streptomyces lividans* cultivation

Streptomyces lividans liquid seed cultures (2 mL MYG with the appropriate antibiotic) were inoculated from a plate or frozen stock with a sterile 200 µL pipet tip, which was left in the 14 mL culture tube. Seed cultures were incubated at 30 °C with 250 rpm shaking until achieving turbidity or high particle density (typically 2 to 3 days). A 500 µL inoculum was then added to 50 mL of fresh MYG medium with the appropriate antibiotic in a baffled 250 mL Erlenmeyer flask with ca. fifty 4 mm glass beads. The 50 mL cultures were incubated at 30 °C with 250 rpm shaking for 5 to 7 days. For large-scale cultivation, 600 mL cultures in baffled 2 L Erlenmeyer flasks (with beads) were inoculated with 6 mL of a 2 – 3 day old culture.

4.4.9 NMR analysis

S. lividans cultures were transferred to 50 mL conical tubes via serological pipet and pelleted at 4000 rpm in an Eppendorf 5810R centrifuge for 10 min. The supernatants were then split into two 50 mL conical tubes (25 mL of supernatant in each), while the cell pellet was stored at -80 °C. The supernatant samples were then flash frozen in liquid nitrogen and lyophilized to dryness. To each dried supernatant tube was added 25 mL of methanol, while the thawed cell pellet was resuspended in 10 mL methanol. Methanol suspensions were briefly agitated by hand, then vortexed for 2 min, then shaken by hand for 5 to 10 min, and finally incubated at 4 °C in a rolling incubator for 0 to 2 hr. Samples were then clarified via centrifugation twice at 4000 rpm for 10 to 15 min each time, and filtered to remove residual particulate matter (as needed). Methanol extracts from supernatant and pellet were then pooled, evaporated to dryness by rotavap, and resuspended in 700 µL deuterium oxide (added in two 350 µL aliquots). A small spatula-full of Chelex 100 resin (Bio-Rad) was then added to each sample in a 1.7 mL centrifuge tube, which was incubated for 10 to 30 min at room temperature with 650 rpm agitation in a microplate shaker. The samples were then centrifuged for 2 min at maximum speed in a benchtop centrifuge, and the supernatant was added to a 10 kDa or 5 µm filter to remove residual resin and particulates. The filtrate was then transferred to a 5mm NMR tube for ^{31}P-NMR analysis. ^{31}P-NMR was conducted in an Agilent 600 MHz NMR machine with OneNMR probe at the Institute for Genomic Biology. Samples were analyzed with an acquisition time of 0.3 seconds for a total of 2560 transients. Data was analyzed using MNova software (MestreLab Research, Escondido, CA).

4.4.10 Isolation of compounds 1 – 7

To isolate compounds **1** – **6,** 50 L of culture supernatant from strain 3p2.1 was dried completely by lyophilization. Compound **7** was isolated from 5 L of culture supernatant from strain 3pf. After treatment with 90 % methanol, the supernatant was dried and applied to Sephadex G-20 columns of different sizes to separate phosphonic acids **1** – **6**; phosphonic acid **7** was found in the precipitate following an analogous methanol cut. Compounds **1**, **3**, **6**, and **7** were isolated individually, while compounds **2**, **4**, and **5** were isolated as a mixture. Structures were confirmed by one- and two-dimensional NMR analyses, as well as MS analysis performed at the School of Chemical Sciences Mass Spectrometry Facility, University of Illinois, Urbana-Champaign.

4.5 References

1. Newman, D.J. and Cragg, G.M. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod*, **75**, 311-335.
2. Li, J.W. and Vederas, J.C. (2009) Drug discovery and natural products: end of an era or an endless frontier? *Science*, **325**, 161-165.
3. Weiss, U., Yoshihira, K., Highet, R.J., White, R.J. and Wei, T.T. (1982) The chemistry of the antibiotics chrysomycin A and B. Antitumor activity of chrysomycin A. *J Antibiot (Tokyo)*, **35**, 1194-1201.
4. Huigens, R.W., 3rd, Morrison, K.C., Hicklin, R.W., Flood, T.A., Jr., Richter, M.F. and Hergenrother, P.J. (2013) A ring-distortion strategy to construct stereochemically complex and structurally diverse compounds from natural products. *Nat Chem*, **5**, 195-202.

5. Bachmann, B.O., Van Lanen, S.G. and Baltz, R.H. (2014) Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *J Ind Microbiol Biotechnol*, **41**, 175-184.

6. Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H. and Fedorova, N.D. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet Biol*, **47**, 736-741.

7. Blin, K., Medema, M.H., Kazempour, D., Fischbach, M.A., Breitling, R., Takano, E. and Weber, T. (2013) antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res*, **41**, W204-212.

8. Scherlach, K. and Hertweck, C. (2009) Triggering cryptic natural product biosynthesis in microorganisms. *Org Biomol Chem*, **7**, 1753-1760.

9. Luo, Y., Huang, H., Liang, J., Wang, M., Lu, L., Shao, Z., Cobb, R.E. and Zhao, H. (2013) Activation and characterization of a cryptic polycyclic tetramate macrolactam biosynthetic gene cluster. *Nat Commun*, **4**, 2894.

10. Metcalf, W.W. and van der Donk, W.A. (2009) Biosynthesis of phosphonic and phosphinic acid natural products. *Annu Rev Biochem*, **78**, 65-94.

11. Yu, X., Doroghazi, J.R., Janga, S.C., Zhang, J.K., Circello, B., Griffin, B.M., Labeda, D.P. and Metcalf, W.W. (2013) Diversity and abundance of phosphonate biosynthetic genes in nature. *Proc Natl Acad Sci U S A*, **110**, 20759-20764.

12. Peck, S.C. and van der Donk, W.A. (2013) Phosphonate biosynthesis and catabolism: a treasure trove of unusual enzymology. *Curr Opin Chem Biol*, **17**, 580-588.

13. Murakami, T., Anzai, H., Imai, S., Satoh, A., Nagaoka, K. and Thompson, C.J. (1986) The bialaphos biosynthetic genes of *Streptomyces hygroscopicus*: molecular cloning and characterization of the gene cluster. *Mol Gen Genet*, **205**, 42-50.

14. Blodgett, J.A., Thomas, P.M., Li, G., Velasquez, J.E., van der Donk, W.A., Kelleher, N.L. and Metcalf, W.W. (2007) Unusual transformations in the biosynthesis of the antibiotic phosphinothricin tripeptide. *Nat Chem Biol*, **3**, 480-485.

15. Hendlin, D., Stapley, E.O., Jackson, M., Wallick, H., Miller, A.K., Wolf, F.J., Miller, T.W., Chaiet, L., Kahan, F.M., Foltz, E.L. *et al.* (1969) Phosphonomycin, a new antibiotic produced by strains of *Streptomyces*. *Science*, **166**, 122-123.

16. Shoji, J., Kato, T., Hinoo, H., Hattori, T., Hirooka, K., Matsumoto, K., Tanimoto, T. and Kondo, E. (1986) Production of fosfomycin (phosphonomycin) by *Pseudomonas syringae*. *J Antibiot (Tokyo)*, **39**, 1011-1012.

17. Woodyer, R.D., Shao, Z., Thomas, P.M., Kelleher, N.L., Blodgett, J.A., Metcalf, W.W., van der Donk, W.A. and Zhao, H. (2006) Heterologous production of fosfomycin and identification of the minimal biosynthetic gene cluster. *Chem Biol*, **13**, 1171-1182.

18. Circello, B.T., Eliot, A.C., Lee, J.H., van der Donk, W.A. and Metcalf, W.W. (2010) Molecular cloning and heterologous expression of the dehydrophos biosynthetic gene cluster. *Chem Biol*, **17**, 402-411.

19. Borisova, S.A., Circello, B.T., Zhang, J.K., van der Donk, W.A. and Metcalf, W.W. (2010) Biosynthesis of rhizocticins, antifungal phosphonate oligopeptides produced by *Bacillus subtilis* ATCC6633. *Chem Biol*, **17**, 28-37.

20. Jomaa, H., Wiesner, J., Sanderbrand, S., Altincicek, B., Weidemeyer, C., Hintz, M., Turbachova, I., Eberl, M., Zeidler, J., Lichtenthaler, H.K. *et al.* (1999) Inhibitors of the

nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science*, **285**, 1573-1576.

21. Park, B.K., Hirota, A. and Sakai, H. (1977) Structure of plumbemycin A and B, antagonists of L-threonine from *Streptomyces plumbeus*. *Agric Biol Chem*, **41**, 573-579.

22. Gao, J., Ju, K.-S., Yu, X., Velásquez, J.E., Mukherjee, S., Lee, J., Zhao, C., Evans, B.S., Doroghazi, J.R., Metcalf, W.W. *et al.* (2014) Use of a phosphonate methyltransferase in the identification of the fosfazinomycin biosynthetic gene cluster. *Angewandte Chemie International Edition*, **53**, 1334-1337.

23. Circello, B.T., Miller, C.G., Lee, J.H., van der Donk, W.A. and Metcalf, W.W. (2011) The antibiotic dehydrophos is converted to a toxic pyruvate analog by peptide bond cleavage in *Salmonella enterica*. *Antimicrob Agents Chemother*, **55**, 3357-3362.

24. Eliot, A.C., Griffin, B.M., Thomas, P.M., Johannes, T.W., Kelleher, N.L., Zhao, H. and Metcalf, W.W. (2008) Cloning, expression, and biochemical characterization of *Streptomyces rubellomurinus* genes required for biosynthesis of antimalarial compound FR900098. *Chem Biol*, **15**, 765-770.

25. Johannes, T.W., DeSieno, M.A., Griffin, B.M., Thomas, P.M., Kelleher, N.L., Metcalf, W.W. and Zhao, H. (2010) Deciphering the late biosynthetic steps of antimalarial compound FR-900098. *Chem Biol*, **17**, 57-64.

26. Shao, Z., Rao, G., Li, C., Abil, Z., Luo, Y. and Zhao, H. (2013) Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold. *ACS Synth Biol*, **2**, 662-669.

27. Uetz, T., Schneider, R., Snozzi, M. and Egli, T. (1992) Purification and characterization of a two-component monooxygenase that hydroxylates nitrilotriacetate from "*Chelatobacter*" strain ATCC 29600. *J Bacteriol*, **174**, 1179-1188.

28. Kieser, T., Bibb, M.J., Buttner, M.J., Chater, K.F. and Hopwood, D.A. (2000) *Practical Streptomyces Genetics*. John Innes Foundation, Norwich, UK.

29. Ajikumar, P.K., Xiao, W.H., Tyo, K.E., Wang, Y., Simeon, F., Leonard, E., Mucha, O., Phon, T.H., Pfeifer, B. and Stephanopoulos, G. (2010) Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. *Science*, **330**, 70-74.

30. Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd and Smith, H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343-345.

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

CHAPTER 5. Efficient multiplex genome editing of *Streptomyces* species via an engineered CRISPR/Cas9 system

5.1 Introduction

Actinobacteria of the genus *Streptomyces* are among the most prolific and well-studied producers of diverse secondary metabolites [1,2]. Over the past several decades, *Streptomyces* strains have been found to produce a number of important bioactive natural products, such as the anticancer compound daunorubicin from *Streptomyces peucetius* [3], the herbicide phosphinothricin from *Streptomyces hygroscopicus* [4] and *Streptomyces viridochromogenes* [5], and the antibacterial daptomycin from *Streptomyces roseosporus* [6]. While these decades of study might suggest that the supply of *Streptomyces* natural products is nearing exhaustion, in fact genome sequencing efforts have revealed that the well is far from dry, even in the most comprehensively studied strains [7].

Access to this "silent" majority of uncharacterized natural product gene clusters would benefit greatly from the development of new genetic manipulation tools that leverage genomic information to aid natural product discovery, characterization, engineering and production. In the context of a *Streptomyces* strain of interest, for example, facile genome manipulation would aid both discovery and validation of new natural products from uncharacterized gene clusters, as well as ensuing biochemical and mechanistic studies. In heterologous production strains, such techniques would further enable genomic remodeling to direct metabolic flux toward a pathway of interest and eliminate competing pathways, as well as pathway engineering for product

diversification. Nevertheless, the current *Streptomyces* genetic toolkit, though well-developed and widely employed, often mandates a significant investment of time and effort.

Typically, for gene disruption in *Streptomyces*, single cross-over integration of a suicide plasmid can be employed, resulting in disruption of the gene of interest with a selectable marker [8]. However, the limited number of selectable markers limits the reusability of this approach. Further, disruption via single-crossover can revert in the absence of selective pressure, resulting in undesired restoration of the wild type allele. Inclusion of flanking recombinase target sites and expression of the corresponding recombinase can enable recycling of markers and improve mutation stability, but this mandates additional steps and leaves a scar sequence at the target site.

Alternatively, clean genomic deletions can be made via double-crossover integration [8]. However, this multi-step process is often labor and time intensive. First, integration of the disruption vector at the target locus (the first single-crossover event) is identified with a selectable marker. Next, loss of the disruption vector and its selectable marker (the second single-crossover event) is identified via non-selective culture. Finally, colonies exhibiting loss of the selectable marker (upon replica plating) must be further screened to separate those that have lost the vector via the desired second crossover event from those that have reverted the first crossover event to restore the original genotype. Counterselectable markers such as *rpsL* and *glkA* can facilitate identification of the second crossover event, but are limited to use only in particular mutant hosts [8]. To facilitate identification of double-crossover integrants in one step, a double-strand break (DSB) can be introduced at the genomic locus of interest, as has been demonstrated using a homing endonuclease [9]. However, this method is dependent upon prior

integration of the homing endonuclease recognition site at the target locus, as homing endonucleases are minimally amenable to specificity alteration [10].

Recently, DSB-mediated genome editing has been achieved via the type II Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/CRISPR associated proteins (Cas) system of *Streptococcus pyogenes* [11]. Functioning like a bacterial immune system, CRISPR/Cas requires three components to realize targeted cleavage of foreign DNA: Cas9, the nuclease and scaffold for the recognition elements; CRISPR RNA (crRNA), short RNAs conferring target site specificity, encoded as spacers in a CRISPR array; and trans-activating crRNA (tracrRNA), a short RNA that facilitates crRNA processing and recruitment to Cas9 [12]. The Cas9/crRNA/tracrRNA complex can target any DNA sequence, known as a protospacer, provided that the requisite protospacer-adjacent motif (PAM) is present at the 3' end (NGG in the case of *S. pyogenes*, where N represents any nucleotide) [13].

To repurpose this system for genome engineering, spacer sequences matching genomic loci of interest can be directly programmed into a heterologously expressed CRISPR array. To further simplify the system, fusion of the crRNA and tracrRNA into a single synthetic guide RNA (sgRNA) transcript has been demonstrated, obviating the need for processing of the transcribed CRISPR array (pre-crRNA) into individual crRNA components [14]. The *S. pyogenes* CRISPR/Cas system has been successfully reconstituted in a variety of hosts across all domains of life, including (but not limited to) *Escherichia coli* [15], *Saccharomyces cerevisiae* [16], and human cell lines [17]. Compared with other tools for site-specific genome engineering, such as zinc-finger nucleases and transcription activator-like effector (TALE) nucleases, the engineered

CRISPR/Cas system offers unprecedented modularity. Targeting any site of interest requires only the insertion of a short spacer into a CRISPR array/sgRNA construct, which can be achieved rapidly and with high throughput using modern DNA assembly techniques. A suitable target site needs only to have an adjacent NGG sequence, which are notably abundant in GC-rich *Streptomyces* genomes.

The goal of the work presented here is to design, construct, and evaluate an engineered CRISPR/Cas9 system in actinobacteria. The system will be reconstituted in multiple *Streptomyces* species for both single-locus and multiplex targeting to introduce precise deletions of various lengths. To facilitate rapid customization to any task of interest, the full CRISPR/Cas9 system will be contained on a single plasmid that is amenable to insertion of custom spacer sequences and editing templates.

5.2 Results and Discussion

5.2.1 Plasmid design

To harness the CRISPR/Cas system for genome editing in *Streptomyces* species, the pCRISPomyces expression system was designed (Figure 5.1). The initial design included both tracrRNA and CRISPR array expression cassettes along with *cas9*. Previously characterized strong promoters [18] were selected to drive expression of the CRISPR/Cas elements, along with widely-used terminators from phages fd, lambda, and T7. Analysis of the *Streptococcus pyogenes cas9* gene revealed the presence of several rare codons. Among them were many *bldA* codons, translation-level regulators of secondary metabolism in *Streptomyces* species [19]. As a result, for the initial pCRISPomyces design (pCRISPomyces-0), a *cas9* gene codon-optimized

for human cell lines (*hSpcas9*) was chosen, as it was conveniently found to be devoid of *bldA* codons.

To facilitate seamless, one-step Golden Gate assembly of custom-designed spacers into the CRISPR array, a *lacZ* cassette flanked by unique *Bbs*I restriction sites was incorporated between two direct repeat sequences in the empty CRISPR array. A unique *Xba*I restriction site was included to enable linearization of the backbone for insertion of additional elements, such as editing template sequences for recombination-driven repair, via Gibson assembly or traditional ligation. Inclusion of *aac(3)-IV* allows selection in both *Escherichia coli* and *Streptomyces* hosts, the colE1 origin enables replication in *E. coli*, and the RP4 origin of transfer *oriT* enables conjugative transfer of pCRISPomyces plasmids from *E. coli* to *Streptomyces* hosts. Finally, the temperature-sensitive *rep* region from pSG5 allows rapid clearance of the pCRISPomyces plasmid following the desired genome editing.

5.2.2 Assembly of pCRISPomyces-0

To enable Golden Gate assembly of spacers into the pCRISPomyces backbone, all other recognition sequences for the Type IIS restriction enzyme of choice had to be removed. Because it had the fewest sites present in the designed construct (5 total: 2 in *hSpcas9*, 2 in the rpsLp(XC) promoter, and 1 in the pSG5 *rep*), the Type IIS restriction enzyme *Bbs*I was chosen for the pCRISPomyces system. To perform all desired *Bbs*I-removing mutations simultaneously, an assembly scheme was designed in which mutations in rpsL(XC) were introduced in a gBlock synthetic DNA construct and the junctions between the fragments were positioned at the

remaining mutation sites. Given the number of parts this assembly scheme mandated (9 total; Figure 5.1), yeast homologous recombination was chosen to maximize assembly efficiency.

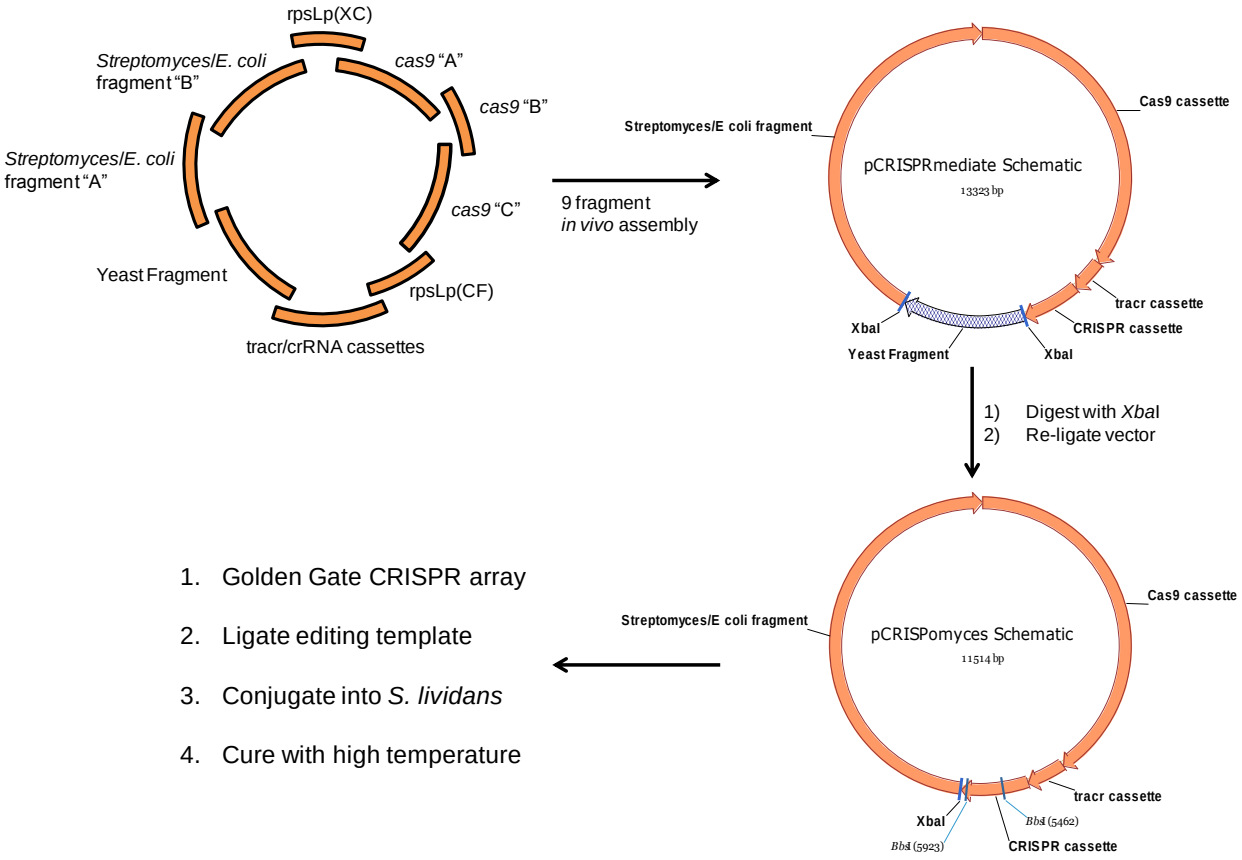


Figure 5.1: Assembly scheme for pCRISPomyces-0 plasmid.

Implementation of yeast *in vivo* assembly required addition of an extra fragment for plasmid maintenance and selection in yeast. As a result, a piece of the pRS416 *E. coli*-yeast shuttle vector was also included in the assembly, bringing the total number of fragments to 9. *Xba*I recognition sites were added to the end of this yeast helper fragment to allow its excision following assembly, obviating the need to remove *Bbs*I recognition sites from this fragment and reducing the overall size of the final pCRISPomyces construct. Further, the resulting unique

*Xba*I site could then be repurposed for linearization of the pCRISPomyces backbone. Assembly was carried out successfully, and confirmed via diagnostic digestion and sequencing of the *cas9*, tracrRNA cassette, and CRISPR array cassette.

5.2.3 Evaluation of pCRISPomyces-0

To assess the functionality of the pCRISPomyces system, initial experiments were carried out in the well-studied strain *Streptomyces lividans* 66 [20]. Plasmids were designed and constructed to target either one gene (*redN*) or two genes (*redD* and *redF*) from the undecylprodigiosin gene cluster [21]. In both cases, a 20 bp sequence with the requisite NGG PAM sequence was chosen, with preference given to sites on the non-coding strand and those with multiple purine bases at the 3' end [22]. To minimize off-target effects, sites were chosen in which the last 12 bp of the protospacer plus the PAM (15 bp in total) were unique [17], as confirmed by BLAST analysis against the published genome sequence. Both constructs were initially assembled without an editing template to assess the potential for NHEJ-mediated repair.

Delivery of the pCRISPomyces-0 plasmids to *S. lividans* was attempted via conjugation. However, repeated attempts failed to yield any exconjugants. As a troubleshooting measure, a series of plasmids was assembled to test the different elements of the pCRISPomyces-0 plasmid. Conjugation of the empty pCRISPomyces-0 plasmid again failed to yield exconjugants, ruling out the possibility that cell death was mediated only by RNA-guided digestion at the targeted loci. Replacement of the pSG5 *rep* origin with the pIJ101 origin also did not yield any exconjugants, suggesting that the toxicity of the pCRISPomyces-0 plasmid stems from the CRISPR/Cas machinery and not the general plasmid maintenance elements. Removal of the full

hSpcas9 gene, in contrast, restored normal conjugation efficiency, suggesting a general toxicity associated with over-expression of the Cas9 protein. This was further confirmed by constructing a plasmid containing only the *hSpcas9* gene without the tracr/crRNA cassettes, which again did not yield any exconjugants. Interestingly, integration of the *hSpcas9* cassette to the chromosome via the ΦC31 integrase was well tolerated by the cell, suggesting that Cas9 toxicity was only apparent when the gene was present at high copy number. Notably, however, no genome editing activity was observed when all elements of the CRISPR/Cas system were co-integrated to the chromosome, as all exconjugants with integrated *redN* or *redD*/*redF* targeting cassettes maintained the ability to produce undecylprodigiosin (evidenced by their red pigmentation).

Since the *cas9*-integrated *S. lividans* exhibited normal growth characteristics, *hSpcas9* expression at the transcriptional and translational levels was evaluated in this strain. Quantitative PCR (qPCR) of the *hSpcas9* transcript revealed high expression (120-fold) relative to an *hrdB* internal standard at 24 hr, followed by a significant decrease in expression at 48 hr and 72 hr (~5-fold). At the translational level, expression of the Cas9 protein was not seen via Coomassie staining (Figure 5.2). After addition of a C-terminal 3xFLAG tag to the integrated *cas9*, Western blotting revealed only the expression of a much smaller protein than the expected 160 kDa size. Thus, it was determined that the human codon-optimized *hSpcas9* was not properly expressed in *S. lividans*.

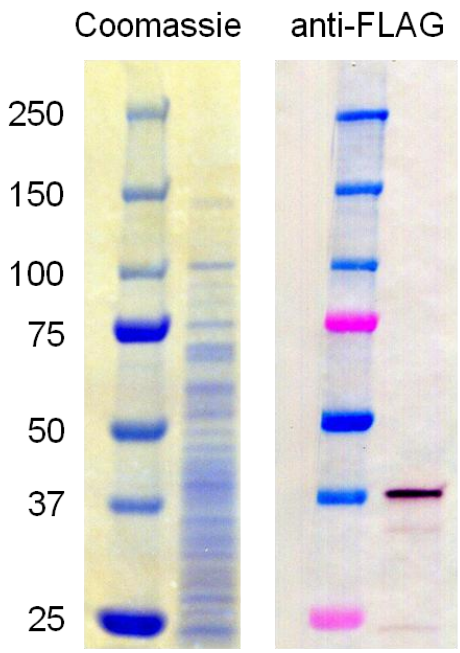


Figure 5.2: Expression analysis of human codon-optimized Cas9 in *S. lividans* by SDS-PAGE and Western blot.

5.2.4 Assembly and evaluation of pCRISPomyces-1

For the next iteration of pCRISPomyces design, the sequenced genomes of *Streptomyces* and related genera were queried to identify a more favorable *cas9* for *S. lividans* expression. While many strains were identified with putative CRISPR arrays and related Cas proteins, no *cas9* homologs were found. Rather, CRISPR/Cas systems in *Streptomyces* appear to be exclusively of the Type I or Type III varieties, utilizing multi-enzyme complexes to perform functions analogous to those of the single-polypeptide Cas9. To avoid the complications associated with coordinated expression of several Cas genes, a codon-optimized *cas9* for *Streptomyces* expression (*sSpcas9*) was instead designed and synthesized *de novo*.

Replacement of *hSpcas9* with *sSpcas9* in pCRISPomyces-0 resulted in construct pCRISPomyces-1. At the same time, the *sSpcas9* cassette was integrated to the *S. lividans*

chromosome for comparison with the previous *hSpcas9* expression results. For this strain, SDS-PAGE and Western blotting revealed high-level expression of the Cas9 protein at the expected 160 kDa size (Figure 5.3).

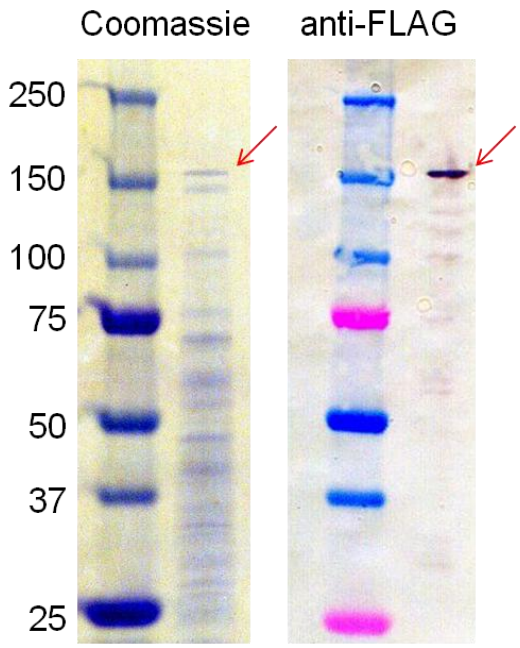


Figure 5.3: Expression analysis of *Streptomyces* codon-optimized Cas9 in *S. lividans* by SDS-PAGE and Western blot. Red arrows indicate the expected band (160 kDa).

To evaluate targeted genome editing with the sSpCas9 protein, subsequent experiments were carried out with the episomal pCRISPomyces-1 plasmid. Two genomic protospacer sequences were targeted using pCRISPomyces-1: one in *redN*, as described previously; and one in *actVA-ORF5*, from the actinorhodin gene cluster [23]. Protospacer sequences were identified as described above. To enable defined editing via homologous recombination, a 2 kb editing template was supplied on the pCRISPomyces-1 plasmid. The editing template consisted of two 1 kb arms homologous to the corresponding sequences upstream and downstream of the

protospacer, designed to introduce a short deletion (20 – 34 bp) to partially or fully eliminate the protospacer sequence and create a frame shift.

The pCRISPomyces-1 plasmids were transferred to *S. lividans* via conjugation. Although notably reduced conjugation efficiency was observed for both of the targeting plasmids bearing the *sSpcas9* gene, enough exconjugants were obtained for downstream analysis. To screen for the desired editing event, genotyping of multiple exconjugants was carried out by first isolating the genomic DNA and then PCR amplifying the target locus. To ensure that the PCR product was amplified from the chromosome rather than the pCRISPomyces plasmid, primers were designed to anneal slightly upstream and downstream of the editing template sequence. Each PCR product was sequenced with internal primers to determine if the intended deletion had been introduced.

Using the pCRISPomyces-1 system, genome editing events were observed. However, editing efficiency was found to be low (Table 5.1). For the *redN* target, three out of fourteen exconjugants possessed the desired deletion, while the remaining eleven strains were unedited at the *redN* locus. As a negative control, a pCRISPomyces-1 derivative plasmid bearing all CRISPR elements (tracrRNA, *redN* spacer, and *redN* editing template) except for *sSpcas9* was constructed and conjugated into *S. lividans*. In this case, twelve exconjugants were screened, and each possessed the unedited genotype. For the *actVA-ORF5* target, a similar efficiency was observed as the *redN* target, with two out of eight exconjugants displaying the edited genotype.

Table 5.1: pCRISPomyces-1 editing results in *S. lividans* 66

Plasmid	Target	Deletion Size	Result
pCRISPomyces-1 w/o Cas9	*redN*	20 bp	0/12
pCRISPomyces-1	*redN*	20 bp	3/14
pCRISPomyces-1	*actVA-ORF5*	34 bp	2/8

To demonstrate clearance of the pCRISPomyces-1 plasmid, one of the identified *S. lividans* strains carrying the desired *redN* deletion was cultured non-selectively at high temperature (37-39 °C). After growth to stationary phase, a fraction of the culture was plated to isolate individual colonies. Multiple colonies were obtained that had regained apramycin sensitivity and did not produce the red undecylprodigiosin pigment, indicating successful clearance of the temperature-sensitive plasmid. This sequence allows the apramycin selection marker to be reused in this strain for future applications, such as further genome editing or introduction of heterologous genes.

5.2.5 Evaluation of pCRISPomyces-2

5.2.5.1 Single-locus targeting

Since the three-component CRISPR/Cas9 system of the pCRISPomyces-1 plasmid did not exhibit high editing efficiency, it was speculated that employing the simpler two-component system could yield improvement. To test this hypothesis, the same protospacers in the *redN* and *actVA-ORF5* target genes were evaluated with the sgRNA-expressing pCRISPomyces-2 plasmid. Editing efficiency was again evaluated by genotyping multiple isolated exconjugants. With the pCRISPomyces-2 *redN*-targeting construct, significantly higher editing efficiency was observed, as six out of six randomly selected exconjugants were revealed to possess the edited phenotype

(Table 5.2). Similar results were observed for the *actVA-ORF5* target, where eight out of eight randomly selected exconjugants were found to be correctly edited. Taken together, these results show that sgRNA targeting affords higher efficiency than tracr/crRNA targeting. A possible explanation could be that pre-crRNA processing by native RNase enzymes is inefficient when applying CRISPR/Cas in *Streptomyces*, in contrast to results in other hosts where native RNase enzymes appear sufficient [17].

Table 5.2: pCRISPomyces-2 editing results in *S. lividans* 66

Plasmid	Target(s)	Deletion Size	Result
pCRISPomyces-2	*redN*	20 bp	6/6
pCRISPomyces-2	*actVA-ORF5*	34 bp	8/8
pCRISPomyces-2	*redN*/*actVA-ORF5*	20 bp and 34 bp	4/4
pCRISPomyces-2	*redD*/*redF*	31,415 bp	4/4

5.2.5.2 Multiplex targeting

One of the key advantages of the CRISPR/Cas system compared with other targeted nucleases is the relative ease with which multiple sequences can be simultaneously targeted using the same Cas9 endonuclease. Given the observed superiority of sgRNA over the tracrRNA/CRISPR array configuration for single-target genome editing, additional constructs were designed with two sgRNA cassettes to evaluate the potential for multiplex targeting. In these constructs, tandem sgRNA cassettes were driven by two copies of the same strong promoter. Insertion of both cassettes into the same pCRISPomyces-2 backbone was carried out in one step by Golden Gate assembly with a synthetic DNA fragment containing the first sgRNA sequence, a terminator, a promoter, and the second sgRNA sequence, all flanked by *Bbs*I sites.

The first dual-targeting construct was designed to introduce simultaneous short deletions in both the *redN* and *actVA-ORF5* loci, utilizing the same protospacer sequences and 2 kb editing templates previously evaluated. Following conjugation and isolation of genomic DNA from multiple exconjugants, both loci of interest were sequenced. In four out of four strains evaluated, the desired deletion was observed at both loci, demonstrating the potential for multiplex genome editing. The ability to cut two loci in the genome simultaneously opens up the possibility for excision of larger chromosomal segments by introducing a DSB at both ends and bridging the gap with a plasmid-borne editing template (Figure 5.4a). To test this method, a second dual-targeting construct was designed to delete the entire 31 kb *red* cluster. Protospacer sequences were selected within the genes at the start and end of the cluster (*redD* and *redF*, respectively), and a 2 kb editing template was constructed from the 1 kb sequences immediately flanking the cluster. In total, four exconjugants were picked and evaluated by PCR from genomic DNA with primers that anneal outside of the 1 kb homology arms. In all four cases (Table 5.2), the 2.1 kb band indicative of cluster deletion was observed, while no band was observed when wild type genomic DNA was used as template (Figure 5.4b). Similarly, PCR amplification with a primer annealing within the deleted region only produced a 1.1 kb band from the wild type genomic DNA, but not from the strains in which the *red* cluster had been deleted. Phenotypic screening (Figure 5.4C) confirmed loss of red pigmentation in all four edited strains.

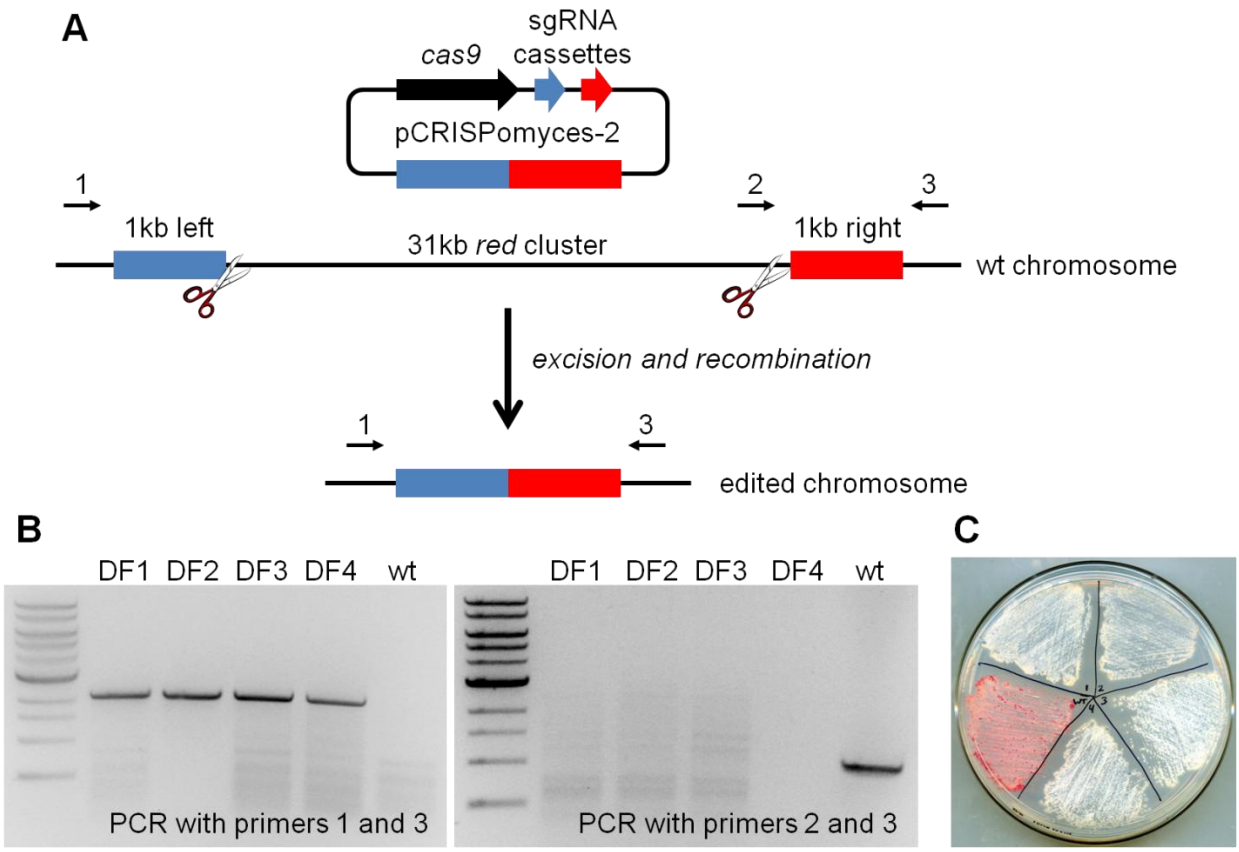


Figure 5.4: Deletion of the 31 kb *red* cluster in *S. lividans*. (a) Two sgRNA transcripts guide Cas9 to introduce DSBs at both ends of the cluster, while a co-delivered editing template bridges the gap via homologous recombination. (b) PCR evaluation of *red* cluster deletion from four exconjugants (DF1 – DF4) with wild type control (wt). (c) Phenotypic screening confirms loss of red pigmentation (wild type shown for comparison).

5.2.6 Evaluation in multiple *Streptomyces* species

Outside of *S. lividans*, it would be desirable to utilize the pCRISPomyces plasmids in other *Streptomyces* strains to realize targeted editing of genes of interest, such as natural product gene clusters, in the native producers. To evaluate the possibility for broader applicability, two additional *Streptomyces* species were selected: *S. viridochromogenes* DSM 40736 and *Streptomyces albus* J1074. In *S. viridochromogenes*, two genes within the phosphinothricin tripeptide (PTT) gene cluster [24], *phpD* and *phpM*, were chosen as individual targets. For both

genes, the corresponding single-targeting pCRISPomyces-2 plasmid was constructed with an appropriate spacer and 2 kb editing template to introduce a short frame-shift deletion early in the coding sequence. Following an analogous conjugation and genotyping protocol as described for *S. lividans*, seven out of seven exconjugants isolated with the *phpD*-targeting plasmid were confirmed to possess the intended deletion. For *phpM*, a total of six exconjugants were screened, and four were found to exhibit the desired mutant genotype (Table 5.3). The other two exconjugants possessed the unedited wild type sequence.

Table 5.3: pCRISPomyces-2 editing results in other *Streptomyces* species

Strain	Target(s)	Deletion Size	Result
S. viridochromogenes DSM 40736	*phpD*	23 bp	7/7
S. viridochromogenes DSM 40736	*phpM*	20 bp	4/6
S. albus J1074	*sshg_05713*	67 bp	6/6
S. albus J1074	*sshg_00040*/*sshg_00050*	13,214 bp	2/3
S. albus J1074	*sshg_05699*/*sshg_05729*	50,626 bp	2/2

In *S. albus*, the PKS-NRPS hybrid gene *sshg_05713* from a cryptic polycyclic tetramic acid macrolactam gene cluster [25] was chosen to evaluate single locus targeting, again to introduce a short frame-shift deletion using a 2 kb editing template. Six out of six exconjugants evaluated by PCR amplification from genomic DNA and sequencing were confirmed to harbor the intended 67 bp deletion (Table 5.3), confirming Cas9 functionality in this strain. As a second trial, a dual-targeting plasmid was constructed to delete a full 13 kb lanthipeptide cluster encoded by genes *sshg_00040* to *sshg_00050*. After conjugation with the dual-targeting plasmid containing a 2 kb editing template, three *S. albus* exconjugants were isolated and evaluated by PCR. All three

were found to yield the correct PCR product indicative of full-cluster deletion (amplified with primers 1 and 3 in Figure 2a), and the identity of the PCR product was confirmed by sequencing. However, one of the three also yielded the PCR product with a primer annealing within the deleted region (primers 2 and 3 in Figure 2a), indicating a mix of edited and wild type cells in this population.

As a third trial, an analogous dual-targeting plasmid was constructed to target both ends of the polycylic tetramic acid macrolactam gene cluster spanning genes *sshg_05699* to *sshg_05729* for deletion of the full 50 kb region. After conjugation with the dual-targeting plasmid containing a 2 kb editing template, two *S. albus* exconjugants were picked from the conjugation plate and restreaked to isolate individual colonies. Three individual colonies were screened from each exconjugant were screened by PCR, and all were found to yield the correct PCR product indicative of full-cluster deletion. The identity of the PCR product was confirmed by sequencing. PCR amplification with a primer internal to the gene cluster did not yield the target band, as expected for deletion of the cluster. Taken together, the observed efficient genome editing of *S. viridochromogenes* and *S. albus* without strain-specific modification of the pCRISPomyces backbone suggests broader applicability of this system in various *Streptomyces* species.

5.3 Conclusions and Outlook

In summary, we have demonstrated that the type II CRISPR/Cas system of *S. pyogenes* can be reconstituted successfully in *Streptomyces* hosts to realize targeted genome editing. While the native tracr/crRNA configuration yielded only modest genome editing efficiency when evaluated in *S. lividans*, utilization of guide RNA boosted efficiency significantly, enabling multiplexing

for the introduction of short and long deletions. Further, high-efficiency genome editing was realized in two additional *Streptomyces* strains, demonstrating broad host applicability. To facilitate implementation of the CRISPR/Cas9 genome editing system for any target(s) of interest, we have developed pCRISPomyces plasmids amenable to insertion of custom spacers and editing templates via modern DNA assembly techniques. Thus, application of the pCRISPomyces system could significantly reduce the time and labor needed to perform precise chromosomal manipulations compared to previous techniques.

Moving forward, the pCRISPomyces system should facilitate a wide variety of future studies in *Streptomyces* species. In known natural-product-producing strains, for example, pCRISPomyces can be used for gene deletion experiments in the elucidation of biosynthetic pathways. Further, precise mutations can be introduced to facilitate *in vivo* mechanistic studies. In non-producing strains with cryptic natural product gene clusters, promoters can rapidly be knocked-in in front of one or more key genes for gene cluster activation experiments. Alternatively, putative regulators can be up- or down-regulated to de-repress silenced pathways.

In popular heterologous hosts such as *S. lividans* and *S. albus*, implementation of CRISPR/Cas9-mediated genome editing could be used for extensive genomic remodeling to design optimal hosts for heterologous expression. This includes knocking out competing pathways, introducing or up-regulating pathways for precursor biosynthesis, or introducing seamless tags to evaluate temporal expression patterns of heterologous proteins. The CRISPR/Cas9 system further enables rapid alteration of previously integrated genes, opening the door for pathway optimization studies to maximize production of valuable products.

5.4 Materials and Methods

5.4.1 Strains, media and reagents

E. coli strain NEB5-alpha (New England Biolabs, Ipswich, MA) was used for plasmid cloning and maintenance. Yeast *in vivo* plasmid assembly was performed in *S. cerevisiae* HZ848 [26]. *S. lividans* 66 was obtained from the Agricultural Resource Service Culture Collection (Peoria, IL), *S. albus* J1074 was a gift from Prof. Wenjun Zhang (University of California, Berkeley), and *S. viridochromogenes* DSM40736 and the *E. coli* conjugation strain WM6026 [24] were gifts from Prof. William Metcalf (University of Illinois at Urbana-Champaign). *E. coli* strains were grown in LB medium supplemented with apramycin (50 µg/mL) and, for strain WM6026, diaminopimelic acid (19 µg/mL). *Streptomyces* strains were grown on modified MYG medium (10 g/L malt extract broth, 4 g/L yeast extract, 4 g/L glucose) with 50 µg/mL apramycin as needed for plasmid selection. Medium R2 (without sucrose) was used for conjugation [8]. All media components and supplements were purchased from Sigma-Aldrich (St. Louis, MO) with the exception of yeast extract (BD Biosciences, San Jose, CA) and LB broth (Fisher Scientific, Pittsburgh, PA). PCR primers were synthesized by Integrated DNA Technologies (Coralville, IA), and PCR reactions were performed in FailSafe PCR PreMix G (Epicentre Biotechnologies, Madison, WI) with Q5 DNA polymerase (New England Biolabs, Ipswich, MA). All PCR products were purified using the DNA Clean & Concentrator or Zymoclean Gel DNA Recovery Kit (Zymo Research, Irvine, CA). Plasmids were recovered using the QIAprep Spin Miniprep Kit (Qiagen, Valencia, CA). Restriction enzymes and T4 ligase were purchased from New England Biolabs (Ipswich, MA).

5.4.2 Plasmid construction

Plasmid pCRISPomyces-0 was constructed via yeast homologous recombination [26] from the following fragments: promoter rpsLp(XC) [18], synthesized as a gBlock (Integrated DNA Technologies, Coralville, IA) to remove *Bbs*I recognition sites; human codon-optimized *Spcas9* and fd terminator, split into three pieces to remove *BbsI* recognition sites; promoter rpsLp(CF), PCR amplified from a previous construct [18]; tracrRNA, oop terminator, promoter gapdhp(EL) [18], a *lacZ* expression cassette flanked by *Bbs*I recognition sites and direct repeat sequences, and a T7 terminator, synthesized as a gBlock (IDT); yeast helper fragment containing *URA3* and CEN6/ARS4 flanked by *Xba*I recognition sites, PCR amplified from pRS416 (Stratagene, La Jolla, CA); and an *E. coli*/*Streptomyces* helper fragment containing origin colE1, selection marker *aac*(3)IV, pSG5 *rep* origin, and origin of transfer *oriT*, PCR amplified in two pieces from plasmid pJVD52.1 [24] to remove a *Bbs*I recognition site in pSG5 *rep*. The resulting intermediate plasmid was then digested with *Xba*I to liberate the yeast helper fragment, and the backbone was re-ligated to yield pCRISPomyces-0. Replacement of the pSG5 *rep* region with pIJ101 *rep* or *attP*/ΦC31 *int* was performed via digestion of pCRISPomyces-0 with *Cla*I and *Sph*I and ligation of corresponding similarly digested insert. Addition of a C-terminal 3xFLAG tag to *hSpcas9* was performed via digestion with *Ssp*I and *Xba*I and Gibson assembly with fragments comprising the C-terminal end of *hSpcas9*, a 3xFLAG tag, and the remaining downstream elements (tracrRNA and crRNA cassettes).

Plasmid pCRISPomyces-1 was assembled from pCRISPomyces-0 via digestion with *Fse*I and *Eco*RI and ligation of a similarly-digested *Streptomyces* codon-optimized *Spcas9* gene, synthesized by GenScript (Piscataway, NJ). Plasmid pCRISPomyces-2 was constructed via

isothermal assembly of the *Eco*RI/*Xba*I-digested pCRISPomyces-1 backbone with two synthetic gBlocks (IDT) comprising a guide RNA expression cassette (with a *Bbs*I-flanked *lacZ* cassette in place of the spacer sequence). All targeting constructs were assembled by a combination of Golden Gate assembly [27] (for insertion of spacers) and traditional digestion/ligation or isothermal assembly [28] (for insertion of editing templates). Single spacer inserts were generated by annealing two 24 nt oligonucleotides (offset by 4 nt to generate sticky ends), while double spacer inserts were synthesized as gBlocks (IDT). The 1 kb left and right arms of each editing template were amplified from purified genomic DNA, spliced by overlap-extension PCR [29], and ligated into the *Xba*I site of the desired plasmid. Correct plasmid assembly was confirmed by diagnostic digestion and sequencing (GeneWiz, South Plainfield, NJ). Plasmid maps were generated with Vector NTI (Invitrogen, Carlsbad, CA).

5.4.3 Transformation

E. coli NEB5-alpha was transformed by heat shock following the manufacturer's suggested protocol. Yeast transformation was performed as described elsewhere [26]. *E. coli* WM6026 was transformed by electroporation. Conjugation of plasmids into *Streptomyces* spores was performed using the modified protocol described elsewhere [30].

5.4.4 SDS-PAGE and Western blotting

S. lividans strains were grown in 50 mL selective MYG cultures for all protein expression studies. Samples of 1 mL were collected and concentrated 5-fold in B-PER reagent (Thermo Scientific, Pittsburgh, PA) for lysis following the manufacturer's suggested protocol. Lysates were analyzed on 4-20% Min-PROTEAN TGX precast polyacrylamide gels (Bio-Rad, Hercules,

CA) stained with SimplyBlue SafeStain (Life Technologies, Carlsbad, CA). Western blotting was performed with anti-FLAG primary antibody F3165 (Sigma-Aldrich, St. Louis, MO) and goat anti-mouse–alkaline phosphatase secondary antibody A5153 (Sigma-Aldrich), and visualized with Western Blue stabilized substrate for alkaline phosphatase (Promega, Madison, WI).

5.4.5 Screening of *S. lividans* strains

Following conjugation, individual exconjugants were randomly picked and restreaked on MYG agar plates supplemented with 50 µg/mL apramycin and grown at 30°C for 2-3 days. Single colonies were then picked to liquid MYG medium for genomic DNA isolation using the Wizard Genomic DNA Purification Kit (Promega, Madison, WI). The locus of interest was PCR amplified and sequenced using primers annealing ~300 bp upstream and downstream of the deletion site (GeneWiz, South Plainfield, NJ). Clearance of the plasmid was accomplished with high-temperature cultivation (37-39 °C) for 2-3 days, followed by replica plating on selective and non-selective plates to confirm restoration of apramycin sensitivity. For phenotypic screening of undecylprodigiosin-producing strains, saturated liquid cultures were streaked on non-selective R2 plates without sucrose and grown at 30°C for 2-3 days.

5.5 References

1. Bibb, M.J. (2005) Regulation of secondary metabolism in streptomycetes. *Curr Opin Microbiol*, **8**, 208-215.
2. Medema, M.H., Breitling, R. and Takano, E. (2011) Synthetic biology in *Streptomyces* bacteria. *Methods Enzymol*, **497**, 485-502.

3. Otten, S.L., Stutzman-Engwall, K.J. and Hutchinson, C.R. (1990) Cloning and expression of daunorubicin biosynthesis genes from *Streptomyces peucetius* and *S. peucetius subsp. caesius*. *J Bacteriol*, **172**, 3427-3434.

4. Seto, H., Imai, S., Tsuruoka, T., Satoh, A., Kojima, M., Inouye, S., Sasaki, T. and Otake, N. (1982) Studies on the biosynthesis of bialaphos (SF-1293). 1. Incorporation of ^{13}C- and ^{2}H-labeled precursors into bialaphos. *J Antibiot (Tokyo)*, **35**, 1719-1721.

5. Bayer, E., Gugel, K.H., Hagele, K., Hagenmaier, H., Jessipow, S., Konig, W.A. and Zahner, H. (1972) [Metabolic products of microorganisms. 98. Phosphinothricin and phosphinothricyl-alanyl-analine]. *Helv Chim Acta*, **55**, 224-239.

6. Baltz, R.H. (1997) Lipopeptide antibiotics produced by *Streptomyces roseosporus* and *Streptomyces fradiae*. In Strohl, W. R. (ed.), *Biotechnology of Antibiotics*. Marcel Dekker, New York, pp. 415–435.

7. Bachmann, B., Van Lanen, S. and Baltz, R. (2014) Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *J Ind Microbiol Biotechnol*, **41**, 175-184.

8. Kieser, T., Bibb, M.J., Buttner, M.J., Chater, K.F. and Hopwood, D.A. (2000) *Practical Streptomyces Genetics*. John Innes Foundation, Norwich, UK.

9. Siegl, T., Petzke, L., Welle, E. and Luzhetskyy, A. (2010) I-SceI endonuclease: a new tool for DNA repair studies and genetic manipulations in streptomycetes. *Appl Microbiol Biotechnol*, **87**, 1525-1532.

10. Sun, N., Abil, Z. and Zhao, H. (2012) Recent advances in targeted genome engineering in mammalian systems. *Biotechnol J*, **7**, 1074-1087.

11. Mali, P., Esvelt, K.M. and Church, G.M. (2013) Cas9 as a versatile tool for engineering biology. *Nat Methods*, **10**, 957-963.

12. Wiedenheft, B., Sternberg, S.H. and Doudna, J.A. (2012) RNA-guided genetic silencing systems in bacteria and archaea. *Nature*, **482**, 331-338.

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

14. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. and Church, G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823-826.

15. Jiang, W., Bikard, D., Cox, D., Zhang, F. and Marraffini, L.A. (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol*, **31**, 233-239.

16. Bao, Z., Xiao, H., Liang, J., Zhang, L., Xiong, X., Sun, N., Si, T. and Zhao, H. (2014) Homology-integrated CRISPR-Cas (HI-CRISPR) system for one-step multigene disruption in *Saccharomyces cerevisiae*. *ACS Synth. Biol.*, doi: 10.1021/sb500255k.

17. Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819-823.

18. Shao, Z., Rao, G., Li, C., Abil, Z., Luo, Y. and Zhao, H. (2013) Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold. *ACS Synth Biol*, **2**, 662-669.

19. Leskiw, B.K., Bibb, M.J. and Chater, K.F. (1991) The use of a rare codon specifically during development? *Mol Microbiol*, **5**, 2861-2867.

20. Cruz-Morales, P., Vijgenboom, E., Iruegas-Bocardo, F., Girard, G., Yanez-Guerra, L.A., Ramos-Aboites, H.E., Pernodet, J.L., Anne, J., van Wezel, G.P. and Barona-Gomez, F. (2013) The genome sequence of *Streptomyces lividans* 66 reveals a novel tRNA-dependent peptide biosynthetic system within a metal-related genomic island. *Genome Biol Evol*, **5**, 1165-1175.

21. Cerdeno, A.M., Bibb, M.J. and Challis, G.L. (2001) Analysis of the prodiginine biosynthesis gene cluster of *Streptomyces coelicolor* A3(2): new mechanisms for chain initiation and termination in modular multienzymes. *Chem Biol*, **8**, 817-829.

22. Wang, T., Wei, J.J., Sabatini, D.M. and Lander, E.S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**, 80-84.

23. Okamoto, S., Taguchi, T., Ochi, K. and Ichinose, K. (2009) Biosynthesis of actinorhodin and related antibiotics: discovery of alternative routes for quinone formation encoded in the act gene cluster. *Chem Biol*, **16**, 226-236.

24. Blodgett, J.A., Thomas, P.M., Li, G., Velasquez, J.E., van der Donk, W.A., Kelleher, N.L. and Metcalf, W.W. (2007) Unusual transformations in the biosynthesis of the antibiotic phosphinothricin tripeptide. *Nat Chem Biol*, **3**, 480-485.

25. Olano, C., Garcia, I., Gonzalez, A., Rodriguez, M., Rozas, D., Rubio, J., Sanchez-Hidalgo, M., Brana, A.F., Mendez, C. and Salas, J.A. (2014) Activation and identification of five clusters for secondary metabolites in *Streptomyces albus* J1074. *Microb Biotechnol*, **7**, 242-256.

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

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

28. Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd and Smith, H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*, **6**, 343-345.

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

30. Woodyer, R.D., Shao, Z., Thomas, P.M., Kelleher, N.L., Blodgett, J.A., Metcalf, W.W., van der Donk, W.A. and Zhao, H. (2006) Heterologous production of fosfomycin and identification of the minimal biosynthetic gene cluster. *Chem Biol*, **13**, 1171-1182.