INFORMATIC STRATEGIES FOR NONRIBOSOMAL PEPTIDE DISCOVERY

INFORMATIC STRATEGIES AND TECHNOLOGIES FOR THE DIRECTED DISCOVERY OF NONRIBOSOMAL PEPTIDES

By AUBREY BAILEY MORGAN WYATT, HBSc, MRes.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Doctor of Philosophy

McMaster University Doctor of Philosophy (2013) Hamilton, Ontario (Chemical Biology)

TITLE: Informatic Strategies and Technologies for the Directed Discovery of Nonribosomal Peptides. AUTHOR: Aubrey Bailey Morgan Wyatt, Hon. BSc. (University of Toronto), MRes. (Imperial College of Science and Technology). SUPERVISOR: Asst. Professor Nathan A. Magarvey. NUMBER OF PAGES: xxi, 290.

Abstract

Nonribosomal peptides (NRPs) are a major class of natural products known for their biological activities and are employed therapeutically as immunosupressants, anticancer agents, and antibiotics. Nonribosomal peptides are microbial products, biosynthesized by large assembly line-like enzymes, known as nonribosomal peptide synthetases (NRPSs) that can be found in large gene clusters within the genome. With the advent of genome sequencing, the gene clusters for known NRPs are easily identified within producing organisms, but more strikingly, this sequencing reveals that microbes often contain many gene clusters with no known products suggesting traditional methods of isolation are overlooking the majority of NRPs.

Extensive studies of NRPS functions have revealed assembly line logic for the biosynthesis of NRPs and using this knowledge, the NRP products of NRPS gene clusters can be predicted. In this research, products from both a simple dimodular NRPS from *Staphylococcus aureus* and a complex 11 module NRPS from *Delftia acidovorans* were predicted and used to successfully identify and isolate two novel NRPs, aureusimine and delftibactin. Theses compounds fell outside traditional NRP activities, one being a virulence regulator and the other a gold-complexing metallophore. Subsequent biosynthetic studies of the aureusimine gene cluster within the heterologous host, *Escherichia coli*, provide insight into NRPS flexibility for the creation of NRP natural variants and highlighted the utility of *E. coli* for the heterologous production of NRPs.

Realizing single NRP predictions are not always accurate, a strategy was devised to use a genomically predicted NRP fragment barcode databases with the LC-MS/MS dereplication algorithm, iSNAP, to chemoinformatically identify and physically locate genetically predicted NRPs within crude extracts. This final contribution eliminates the need for bioactivity guided approaches to discovery and provides a strategy to systematically discover all predicted NRPs from cryptic gene clusters. This thesis delivers strategies and technologies for the directed discovery of NRPs from microbial sources.

Acknowledgements

First and foremost, I would like to thank my supervisor Nathan Magarvey. Nathan has been an invaluable mentor, contributing to my growth both academically and personally during my tenure at McMaster. His ideas and enthusiasm know no bounds and it has been a privilege to work with such an outstanding scientist. My time spent working with Nathan will continue to benefit me for the remainder of my career.

I would also like to recognize my committee members, Gerry Wright for providing a keen scientific mind and eye for detail during my committee meetings and a collaborative attitude, allowing me access to his lab throughout my tenure at McMaster, and Paul Harrison for his amazing teaching ability and allowing me to contribute to his own scientific research program.

I would also like to thank all the members of the Magarvey Lab who have given advice and helped me throughout my research. In particular, I would like to thank Chad Johnston for being such an intelligent and relentlessly hardworking partner throughout our collaborative projects. The teamwork achieved by us amounted to some truly amazing research and discoveries.

On a personal level, I would like to thank my family, Alexis, for her continuous support and love, my parents, Marion and Jonathan, for supporting me throughout my time as a student, over and above what was necessary, my brother Jackson, for always encouraging me and my successes and my friends, who have made my life fun and exciting through my time as a graduate student.

Table of Contents

Abstract iii

Acknowledgements v

Table of Contents vi

List of Tables xi

Lists of Figures xiii

Abbreviations xx

Declaration of Academic Achievement xxi

Chapter 1. Introduction 1

1.1 Thesis Context 1

1.2 Scope and Nature of this Work. 13

1.3 Genome Mining for Nonribosomal Peptides 14

1.4 New NRP Discovery Methods 16

1.5 Thesis Overview 19

Chapter 2. Genome Mining Nonribosomal Peptide Synthetases 21

2.1 Chapter Preface 21

2.2 Abstract 22

2.3 Introduction 23

2.4 Cryptic nonribosomal peptide assembly in Staphylococcus. 24

2.5 Isolation of tyrosine-valine dipeptides. 27

2.6 Microarray analysis of virulence expression.....28

2.7 Role of aureusimines *in vivo*.....30

2.8 Materials and Methods.....32

2.8.1 Chemical Characterization of Aureusimines.....32
2.8.2 Genetic Manipulations - *ausA* Allelic Replacement.....33
2.8.3 RNA Isolation and Microarray Analysis.....35
2.8.4 Hemolysis.....35
2.8.5 Systemic Mouse Infections.....35
2.9 Supplementary Information......36

2.9.1 Supporting Figures and Legends (Supplementary Fig. 2.1-7).....36
2.9.2 Supplementary Tables.....46
2.9.3 Supplementary Notes (Supplementary References).....58
2.10 Clarification.....58

2.11 References.....61

Chapter 3. NRPS Flexibility.....64

3.1 Chapter Preface.....64

3.2 Abstract.....65

3.3 Introduction.....65

3.4 Results and Discussion.....69

3.4.1 Reconstitution of the aureusimine nonribosomal peptide assembly line.....69
3.4.2 *In vitro* chemoenzymatic synthesis of aureusimine natural and unnatural products.....70
3.4.3 Heterologous *in vivo* production of *S. aureus* cyclic dipeptides.....74
3.4.4. Structural characterisation of the AusA chain-terminating reductase.....75
3.4 Conclusions.....80

3.5 Materials and Methods.....81

3.5.1 Strains, culture media and general methods.....81
3.5.2 PCR amplification and cloning of *ausA/B* and AusA reductase domain.....81
3.5.3 Cloning and protein purification of AusA and T-Re/Re domains.....82
3.5.4 *in vitro* reconstitution of AusA and analogue production.....83
3.5.5. Analysis of *in vitro* analogue production.....83
3.5.6 Small-molecule analysis of *E. coli*+pAusAB.....84
3.5.7 *In vivo* precursor-directed biosynthesis.....85
3.5.8 Crystal structure generation.....85
3.6 Supplemental Information.....86

3.6.1 Supplemental Methods 94
3.6.1.1 Reductase Structure Determination 94
3.6.1.3 Western Blot 97
3.6.1.4 *Escherichia coli* extract analysis 99
3.6.1.5 ESI-MS/MS 100
3.6.1.6 Quantification of Aureusimine A and B 101
3.6.1.7 Structure verification and elucidation: 101
3.6.2 Supplemental References 103
3.7 References 1. D. Schwarzer, R. Finking, M. A. Marahiel, *Nat. Prod. Rep.* 2003, 20, 275.

..... 103

Chapter 4. Heterologous Production of Nonribosomal Peptides 106

4.1 Chapter Preface 106

4.2 Abstract 107

4.3 Introduction 108

4.4 Results and Discussion 111

4.5 Materials and Methods 119

4.5.1 Strains, culture media, and general methods 119
4.5.2 PCR amplification and cloning of aureusimine biosynthetic cluster 120
4.5.3 Detection and analysis of nonribosomal peptide production by SRM analysis 120
4.5.4 Heterologous production optimization 121
4.5.5 Principle component analysis of *E. coli* + pAusAB 122
4.5.6 Structure elucidation of pyrazines 123
4.6 References 125

Chapter 5. Genome Mining Complex Nonribosomal Peptide Synthetases 128

5.1 Chapter Preface 128

5.2 Abstract 130

5.3 Article 130

Figure 5.2. 136

5.4 Materials and Methods 141

5.4.1 General experimental procedures. 141
5.4.2 Bacterial strains. 142

5.4.3 Gold precipitation on agar plates....142
5.4.4 *D. acidovorans* 96-well plate gold bioassay. 143
5.4.5 Identification of delftibactin biosynthetic gene cluster and adenylation domain specificity. 143
5.4.6 Construction of the Δ*delG D. acidovorans* strain. 144
5.4.7 16S alignment and delftibactin production in environmental strains....146
5.4.9 Transmission electron microscopy of delftibactin–Au(III) complexes....147
5.4.10 Gallium-delftibactin–gold interaction. 148
5.4.11 Gold detoxification by delftibactin....148
5.4.12 Gold detoxification in chronic exposure by delftibactin in presence and absence of iron....149
5.4.13 Gold protective comparison of delftibactin A and B. 150
5.4.14 Delftibactin-mediated protection against gold toxicity. 150
5.4.15 Measuring delftibactin production following depletion by gold....151
5.4.16 MRM-LC/MS measurement of delftibactin production in response to iron. 152
5.4.17 Citrate-gold and delftibactin-gold comparison....152

5.5. Supplementary Figures 153

5.6 Supplementary Tables....165

5.7 Supplementary Note – Additional Structural Information 171

5.7.1 Culture and Isolation 171
5.7.2 High Resolution Mass Spectra 171
5.7.3 NMR Methods and Structural Characterization....172
5.7.4 Delftibactin A 172
5.7.5 Delftibactin B 173

5.8 References....174

Chapter 6. New Technologies for the Direct Identification of Nonribosomal Peptides 176

6.1 Chapter Preface 176

6.2 Abstract 178

6.3 Introduction....179

6.4 Results....182

6.4.1 Hypothetical Barcode Libraries Identify Novel Variants of Knowns....182
6.4.2 Identification of Site Specific Modifications within Desired Pharmacophores....186
6.4.3 Discovery of Known Unknowns through Genome-Predicted Barcode Libraries....189

6.5 Discussion 200

6.6 Materials and Methods....202

6.6.1 General Experimental Procedures 202
6.6.2 Microbial Strains 203
6.6.4 Structure Elucidation 206
6.6.5 Genome Sequencing 206
6.6.6 Identification of delftibactin biosynthetic gene cluster and adenylation domain specificity 206
6.6.8 iSNAP dereplication of trichopolyn 1 and efrapeptin F 207
6.6.9 iSNAP trichopolyn variant identification. 207
6.6.10 iSNAP analysis of *A. citrulli* AAC00-1 extract and identification of acidobactin A and B 208
6.6.11 iSNAP analysis of *V. paradoxus* S110 extract and identification of vacidobactin A and B 208
6.6.13 iSNAP identification of WS-9326A and Neurokinin A-like Analogs 209
6.6.14 Molecular barcode identification of thanamycin 210
6.6.15 Validation of iSNAP Structural Library Specificity 210
6.7 Supplementary Information 211

6.7.1 Structural Characterization 232
6.8 References 255

Chapter 7. Significance and Future Prospective 263

7.1 Genome Mining and NRPS Heterologous Expression 263

7.2 Computational Fragment-based NRP Discovery Strategies 270

7.3 Concluding Remarks 275

References. 277

List of Tables

Chapter 2

Supplementary Table 2.1. Primers used for cloning and construction of recombinant *S. aureus* strains bearing lesions in *ausA*…………………………………………………...46

Supplementary Table 2.2. Genes upregulated in early exponential phase by aureusimines.

…………………………………………………………………………………46

Supplementary Table 2.3. Genes downregulated in early exponential phase by aureusimines ……………………………………………………………………….49

Supplementary Table 2.4. Genes upregulated in late exponential phase by aureusimines.

…………………………………………………………………………………51

Supplementary Table 2.5. Genes downregulated in late exponential phase by aureusimines. ……………………………………………………………………...54

Table 2.8.1. Table of genes up-regulated by the aureusimine. ……….……...………60

Table 2.8.2 Table of genes down-regulated by the aureusimines. ……………………61

Chapter 3

Supplementary Table 3.1 ……………………………………………………………98

Supplementary Table 3.2 ……………………………………………………………99

Supplementary Table 3.3 …………………………………………………………..102

Chapter 5

Supplementary Table 5.1 ^{1}H and ^{13}C NMR and NOESY spectral data of delftibactin A in D_2O ……………………………………………………………………..…165

Supplementary Table 5.2 ^{1}H and ^{13}C NMR spectral data of delftibactin B in D_2O ………………………………………………………………………..……**167**

Supplementary Table 5.3. Adenylation domain specificities. …………………..**168**

Supplementary Table 5.4: Delftibactin gene cluster analysis ……………..……**169**

Supplementary Table 5.5 – Primers used in construction of the Δ*delG* strain. …..**169**

Supplementary Table 5.6 – Delftibactin production in response to [Fe^{3+}] …………..**170**

Chapter 6

Supplementary Table 6.1. Adenylation domain specificities for *A. citrulli* AAC00-1 biosynthetic cluster. ……………………………………………………………..…**231**

Supplementary Table 6.2: Acidobactin A/B and vacidobactin A/B gene cluster analysis ………………………………………………………………………..…**232**

Supplementary Table 6.3. NMR spectroscopic Data for acidobactin A (1), B (2), vacidobactin A (3), and B (4) (700 MHz, in D_2O) …………………………………..**234**

Supplementary Table 6.4. NMR spectroscopic Data for variobactin A (5) (700 MHz, in DMSO-d_6). ………………………………………………………………………..**236**

Supplementary Table 6.5: Summary of ^{1}H (600 MHz) and ^{13}C (125 MHz) spectroscopic data for WS-9326A in DMSO-d_6 ………………………………………………..**251**

Lists of Figures

Chapter 1

Figure 1.1 Overview of NRPS, PKS and Hybrid NRPS/PKS assembly lines. **……..6**

Figure 1.2. Thesis context. **…………………………………………………………13**

Figure 1.3 Overview of the informatics strategy to isolate cryptic secondary metabolites.

……………………………………………………………………………19

Chapter 2

Figure 2.1. Identification of a cryptic NRPS biosynthetic gene cluster within S. aureus.

……………………………………………………………………………26

Figure 2.2. Gene regulation by the aureusimines. **……………………………………30**

Figure 2.3. **………………………………………………………………………32**

Figure S2.1. Gene neighbourhood sequence alignment of the Staphylococcus nonribosomal peptide synthetase gene cluster. **……………………………………37**

Figure S2.2. Phylogenetic analysis of aureusimines biosynthetic cluster in staphylococcus. **……………………………………………………………38**

Figure S2.3. Alignment of the adenylation domains and adenylation codes found within staphylococcal NRPS.**…………………………………………………………39**

Figure S2.4. Chemical mining for cryptic *Staphylococcus aureus* nonribosomal peptides.

……………………………………………………………………………40

Figure S2.5. Chemical Mining of the Staphylococcus aureus nonribosomal peptides: Mass spectral analysis of aureusimine A. **……………………………………………41**

Figure S2.6. Two dimensional-NMR experiments of aureusimine A. **……………43**

Figure S2.7. Two dimensional-NMR experiments of aureusimine B/phevalin. **……44**

Figure S2.8. Growth curves of Newman WT and *NewmanΔausA.* **……………………45**

Scheme 3.1 Aureusimine biosynthetic cluster and assembly line. **……………………68**

Chapter 3

Figure 3.1 Reconstituted *in vitro* chemoenzymatic production of aureusimine and analogues with purified AusA. **…………………………………………………73**

Figure 3.2 **………………………………………………………………………75**

Figure 3.3. Stereo images of A) the AusA T-Re didomain and B) the Re domain active site. **…………………………………………………………………………………78**

Figure 3.4. Structural comparison of AusA reductase domain. **……………………79**

Supplemental Figure 3.1. **…………………………………………………………86**

Supplemental Figure 3.2. MS-MS fragmentation of aureusimine analogues produced through *in vitro* chemoenzymatic synthesis. **…………………………………………87**

Supplemental Figure 3.3. LC-MS extracted ion chromatograms of *S. aureus* Newman wild type and Newman Δ*ausA* extracts. **……………………………………………88**

Supplemental Figure 3.4. MS-MS fragmentation of diketopiperazine analogues produced through *in vitro* chemoenzymatic synthesis. **…………………………89**

Supplemental Figure 3.5. MS/MS fragmentation of pyrazinone congeners produced in *E. coli* pAusAB. **……………………………………………………………90**

Supplemental Figure 3.6. Sequence alignment of AusA with NRPS terminal reductases: **…………………………………………………………………………………91**

Supplemental Figure 3.7. **…………………………………………………………92**

Supplemental Figure 3.8. ……………………………………………………………93

Chapter 4

Figure 4.1. Aureusimine biosynthetic cluster and assembly line. …………………..112

Figure 4.2. Aureusimine A and B production in *E. coli* + pAusAB. …………………..113

Figure 4.3. Aurueusimine A and B production optimization. …………………..115

Figure 4.4. .Identification of 3 pyrazine products by principle component analysis (PCA). ……………………………………………………………………………..118

Supplemental Figure 4.1. ……………………………………………………………..123

Supplemental Figure 4.2. Expression of nonribosomal peptide synthetase (NRPS/AusA) and phosphopantetheinyl transferase (PPtase/AusB) after induction with IPTG. …..124

Supplemental Figure 4.3. Production ratios of aureusimine A and B at the induction concentration and temperature indicated. ……………………………………..125

Chapter 5

Figure 5.1. ……………………………………………………………………..132

Figure 5.3. ……………………………………………………………………..140

Supplementary Figure 5.1. *Delftia acidovorans* possesses a PKS/NRPS gene cluster associated with extracellular gold precipitation. ……………………………..153

Supplementary Figure 5.2. Gold precipitation is caused by a metabolite encoded by the *del* gene cluster. ……………………………………………………………..154

Supplementary Figure 5.3. Structural characterization of delftibactin. …………..155

Supplementary Figure 5.4. Environmental isolates of *D. acidovorans* also produce delftibactin. ……………………………………………………………………..156

Supplementary Figure 5.5 Delftibactin and $AuCl_3$ co-precipitate from solution. **.....157**

Supplementary Figure 5.6. Delftibactin detoxifies $AuCl_3$ and enables growth under chronic gold exposure. **...158**

Supplementary Figure 5.7. Delftibactin production is responsive to gold concentrations through reactive homeostasis.**...159**

Supplementary Figure 5.8. Delftibactin has a single metal binding site and occupancy by other metals impedes $AuCl_3$ precipitation. **......................................160**

Supplementary Figure 5.9. Mass spectral analysis of a gold-bound delftibactin species. **...161**

Supplementary Figure 5.10. Delftibactin B (acetylated hydroxy-ornithine delftibactin A) is impaired in gold precipitation and is less protective against $AuCl_3$. **.............162**

Supplementary Figure 5.11. All delftibactin species eventually co-precipitate with $AuCl_3$. **...163**

Supplementary Figure 5.12. Determination of the delftibactin-$AuCl_3$ transient reaction product. **...164**

Supplementary Figure 5.13. Delftibactin rapidly forms complex precipitates of gold nanoparticles. **...165**

Chapter 6

Figure 6.1. **...181**

Figure 6.2. Identification of trichopolyn from extract natural product bioactivity screens using informatic search for natural products (iSNAP). **..............................184**

Figure 6.3. Directed discovery of Neurokinin A mimics from the iSNAP discovery of WS-9326A. **……………………………………………………………………..…188**

Figure 6.4. Identification of *Acidovorax citrulli* AAC00-1 nonribosomal peptide using iSNAP with a predicted acidobactin structure database. **……………………..….193**

Figure 6.5. iSNAP-guided discovery of delftibactin-acidobactin-like compound from unknown extracts. **……………………………………………………………………..196**

Figure 6.6. Detection and de novo structure solution of the cryptic nonribosomal peptide thanamycin. **……………………………………………………………………………..199**

Supplementary Figure 6.1. Trichopolyn 1 iSNAP hit. **…………………………..211**

Supplementary Figure 6.2. Identification of efrapeptin F from extract bioactivity screen using informatic search for natural products (iSNAP). **…………………………..212**

Supplementary Figure 6.3. Trichopolyn variant 59 (trichopolyn 1) iSNAP hit. **…..213**

Supplementary Figure 6.4. Trichopolyn variant 179 iSNAP hit. **…………………..214**

Supplementary Figure 6.5. Trichopolyn variant 187 iSNAP hit. **…………………..215**

Supplementary Figure 6.6. Trichopolyn variant 11 iSNAP hit. **…………………..216**

Supplementary Figure 6.7. Trichopolyn variant 50 (trichopolyn 4) iSNAP hit. **…..217**

Supplementary Figure 6.8. Trichopolyn variant 51 (trichopolyn 2) iSNAP hit. **…..218**

Figure 6.9. WS-Neurokinin Mimic Library variants found in *S. calvus* extract. **…..219**

Supplementary Figure 6.10.……………………………………………………………..220

Supplementary Figure 6.11.……………………………………………………………..221

Supplementary Figure 6.12. MS fragmentation of WS9326F (left) and WS9326A (right). **…………………………………………………………………………………..222**

Supplementary Figure 6.13. ^{1}H NMR spectra comparison of WS9326A (Top) and novel WS compound WSneuro_5 (Bottom). **…………………………………………..223**

Supplementary Figure 6.14.……………………………………………………………..223

Supplementary Figure 6.15. Identification of *Variovorax paradoxus* S110 nonribosomal peptide using iSNAP with a predicted acidobactin structure database. **………..…225**

Supplementary Figure 6.16. Identification of acidobactin variants with iSNAP using an acidobactin variant library. **…………………………………………………………..227**

Supplementary Figure 6.17. Identification of variobactin variants with iSNAP using a variobactin variant library. **…………………………………………………………..229**

Supplementary Figure 6.18. Theoretical database controls. **…………………..230**

Supplementary Figure 6.19. ^{1}H NMR spectrum of acidobactin A (**1**) in D_2O. **…..237**

Supplementary Figure 6.20. ^{1}H-^{1}H COSY spectrum of acidobactin A (**1**) in D_2O…**...238**

Supplementary Figure 6.21. ^{1}H-^{13}C HMBC spectrum of acidobactin A (**1**) in D_2O…**.238**

Supplementary Figure 6.22. ^{1}H-^{13}C HMQC spectrum of acidobactin A (**1**) in D_2O**....239**

Supplementary Figure 6.23. ^{1}H NMR spectrum of acidobactin B (**2**) in D_2O. **…..239**

Supplementary Figure 6.24. DEPTq spectrum of acidobactin B (**2**) in D_2O. **…..240**

Supplementary Figure 6.25. ^{1}H-^{1}H COSY spectrum of acidobactin B (**2**) in D_2O.**…..240**

Supplementary Figure 6.26. ^{1}H-^{13}C HMBC spectrum of acidobactin B (**2**) in D_2O**.....241**

Supplementary Figure 6.27. ^{1}H-^{13}C HMQC spectrum of acidobactin B (**2**) in D_2O.Supplementary Figure 6.28. 1H-1H NOESY spectrum of acidobactin B (2) in D2O. **…………………………………………………………………………..…241**

Supplementary Figure 6.29. Determination of absolute configuration of amino acid of acidobactin A and B (**1**-**2**) by Marfery reactions ………………………………..**243**

Supplementary Figure 6.30. ^{1}H NMR spectrum of variobactin A (**3**) in D_2O. …..**245**

Supplementary Figure 6.31. ^{1}H-^{1}H COSY spectrum of variobactin A (**3**) in D_2O.…..**245**

Supplementary Figure 6.32. ^{1}H-^{13}C HMBC spectrum of variobactin A (**3**) in D_2O….**246**

Supplementary Figure 6.33. ^{1}H-^{13}C HMQC spectrum of variobactin A (**3**) in D_2O….**246**

Supplementary Figure 6.34. ^{1}H NMR spectrum of variobactin B (**4**) in D_2O. …..**247**

Supplementary Figure 6.35. ^{1}H-^{1}H COSY spectrum of variobactin B (**4**) in D_2O.…..**247**

Supplementary Figure 6.36. ^{1}H-^{13}C HMBC spectrum of variobactin B (**4**) in D_2O….**248**

Supplementary Figure 6.37. ^{1}H-^{13}C HMQC spectrum of variobactin B (**4**) in D_2O….**248**

Supplementary Figure 6.38. ^{1}H NMR spectrum of variobactin A (**5**) in D_2O. …..**249**

Supplementary Figure 6.39. DEPTq spectrum of variobactin A (**5**) in D_2O. …..**249**

Supplementary Figure 6.40. ^{1}H-^{1}H COSY spectrum of variobactin A (**5**) in D_2O.…..**250**

Supplementary Figure 6.41. ^{1}H-^{13}C HMBC spectrum of variobactin A (**5**) in D_2O….**250**

Supplementary Figure 6.42. ^{1}H-^{13}C HMQC spectrum of variobactin A (**5**) in D_2O….**251**

Supplementary Figure 6.43. Summary of key 2D-NMR spectroscopic data that supports the structure of WS-9326A. ……………………………………………………..…**254**

Chapter 7

Figure 7.1 Discovery of Novel Natural Products. ………………………………..**275**

Abbreviations

DNA	deoxyribonucleic acid
RNA	ribonucleic acid
NRP	nonribosomal peptide
PK	polyketide
NRPS	nonribosomal peptide synthetase
PKS	polyketide synthase
MS	mass spectrometry
LC-MS	liquid chromatography – mass spectrometry
HRMS	high resolution mass spectrometry
HTS	high throughput screening
NMR	nuclear magnetic resonance
iSNAP	**i**nformatic **S**earch algorithm for **NA**tural **P**roducts
A	adenylation
C	condensation
T	thiolation
AT	acyltransferase
KS	ketosynthase
Re	reductase
TE	thioesterase
IPTG	isopropyl β-D-1-thiogalactopyranoside

Declaration of Academic Achievement

This thesis is formatted for review as a sandwich thesis. Specific details regarding each research contribution are described within each chapter preface.

Chapter 1. Introduction

1.1 Thesis Context

The effect of natural chemicals on human health is well appreciated.[1-3] Civilizations have relied on the botanical knowledge passed down from previous generations to know which plants are safe to eat, poisonous or have healing properties.[4-7] For many thousands of years, medicinal plants have been used within indigenous medical practices as poultices, teas, powders, and other herbal remedies.[4] In more recent Western culture, we isolate with an emphasis on selected agents that impart the perceived/observed activity.[4-7] . Such emphasis enabled by analytical separations, yielded individual bioactive compounds to formulate the first pharmaceutical agents in the 19th century. Early evidence of the utility of this approach was evinced with analgesic plant compounds such as the opiate morphine, from poppies, and aromatic acids, such as salicylate (the precursor to aspirin) from willow bark.[8, 9] These demonstrations led to investigative inquiry from other macro-organismic sources and selective identifications of other important therapeutics such as insulin from mammalian pancreatic tissues.[10, 11] Perhaps the most transformative discoveries in modern medicine came from micro-organisms.[12-18] In 1928, the first widely used antibiotic, penicillin, was serendipitously discovered following an observation of bioactivity on a plate of *Staphylococcus aureus* inadvertantly contaminated with a blue-green mould (*Penicillium chrysogenum(notatum)*).[12, 19] Zones of inhibited S. aureus growth radiated from the blue – green mould and by tracing the bioactive principle, enriched fractions with the chemical penicillin.[19] Systematic interrogation of environmental moulds and ray-fungi (later named

Streptomyces) demonstrated that microbes construct large suites of esoteric chemistries that impart fitness on themselves (but are not necessary for their growth). This specialized metabolism, like the secondary metabolism in plants, produces the most structurally diverse and therapeutically important small molecules known.[1, 13, 18]

Microbes are found across the planet, and their evolution has occured over millions of years. Sculpted microbial chemistries drive their survival, competitiveness, and capability to proliferate within unique and often refractory niches.[20-23] These specialized chemistries diverge from primary ones necessary for growth, but aid to sequester nutrients, provide protection and promote their proliferation and communication within their biological surroundings.[24-29] These fortuitous adaptions have served as a rich base for new human pharmaceuticals deployed as antibacterials, antifungals, anticancer agents, and because of their selective targeting, regulation of numerous cellular/biological processes.[30]

The success of penicillin, initiated a movement for the systematic investigation of microbial chemicals. By 1940 a strategy to identify microbes that produce natural substances with antibiotic properties was established.[31] In particular, Selman Waksman identified the actinobacterial soil microbes as particularly gifted antibiotic producers.[32] Tracking the bioactive principles from culture supernatants led to the identification of antibiotics such as actinomycin, neomycin and later streptomycin, the first antibiotic for TB treatment.[14, 15] In these instances and several others that followed, successive chemical separations and fractionations led to pure compounds for structure and biological activity assessment.[33] The next two decades, described as the golden age of microbial drug discovery, marked success with the isolation of hundreds of antimicrobial agents and demonstrated the utility of bioactivity-based screening.[30] Bioactivity-based screening continues, but unlike the 'golden age' of natural product drug discovery, discovery wanes.[34, 35]

Around the same time the first antibiotics were isolated from actinomycetes, Avery and colleagues demonstrated that DNA carried the genetic information of life using the model bacterium, pneumococci.[36] As natural product chemistry exploited bioactivity to discover some of the most important therapeutics, seminal studies in molecular biology provided incredible advances in the understanding of life including uncovering the central dogma (DNA->RNA->Protein), the cracking of the genetic code, and the ability to predict protein function directly from genetic sequences.[37, 38] Similar advances for understanding how natural products are made were not realized until the 1990s as gene sequencing became more commonplace and the biosynthetic logic leading to bioactive microbial compounds appreciated.

Early natural product research identified two major classes of microbial natural products: polyketides (PKs) and nonribosomal peptides (NRPs).[1, 2, 39] Collectively, these two classes make up the majority of the most important therapeutically used natural products today including immunosupressants (cyclosporine, rapamycin), anticancer agents (bleomycin) and antibacterials (penicillin, erythromycin, tetracycline, vancomycin).[2] Polyketides and nonribosomal peptides are recognized through retro-biosynthetic (looking back for assembly from simple building blocks) analysis whereby established building blocks, acyl-CoAs and amino acids (proteinogenic and non-proteinogenic) respectively, are identified within the structure of a given compound.[2, 24, 40-46] The first studies on the biosynthesis of NRPs and PKs were carried out using isotope-labelled precursor experiments, but with the advancement of sequencing technologies the enzyme architecture responsible for their construction was ascertained. The erythromycin PK biosynthetic gene cluster was the first sequenced.[47] It contains 3 large genes encoding multi-functional megaenzymes, consisting of 7 modules, each containing 3-6 enzymatic domains. These domains were distributed throughout each gene, where each domain's function could be predicted for the biosynthesis of the 6-deoxyerythreonolide product in a collinear or assembly-line like fashion.[47] Nonribosomal peptide gene clusters share a similar assembly line-like architecture and intermixing of these two assembly systems also occurs to produce even greater structural variation. It is now established that both PKs and NRPs are generated by these megaenzymes, known as polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs), which are located in large gene clusters within microbial genomes.[2, 24, 40-46]

At the time the erythromycin gene cluster sequence was published, the most advanced protein structure and function sequence analysis algorithm was released. [48] The basic local alignment search tool (BLAST) algorithm allows for fast sequence comparison, database and motif searches and could analyze multiple regions of similarity along large DNA stretches.[48] Collectively, this aided the rapid annotation of microbial genomes, and identification of NRPS and PKS gene clusters. This algorithm is still used for DNA and amino acid sequence analysis and can identify PKS and NRPS domain motifs directly from gene sequences. Identifying the modular organization of NRPS and

PKS genes with the simultaneous advancement of gene annotation revolutionized how these important small molecules are studied.

The gene and enzymatic logic surrounding PK and NRP biosynthesis is now revealed.[44, 45] Type I PKSs and NRPSs, as well as hybrids thereof, act to elaborate PKs and NRPs through a series of modular protein domains in a collinear or assembly line-like fashion.[49, 50] Each module consists of at least three domains: a domain for monomer selection and activation, a domain that tethers the growing chain to the assembly line and a domain that catalyzes chain elongation to adjacent monomer building blocks.[49, 50] For NRPSs, these are adenylation (A), thiolation (T), and condensation (C) domains, respectively **(Figure 1.1)**. First, the A domain selects for its cognate amino acid (proteinogenic/non-proteinogenic), activates it as an aminoacyl adenylate and transfers it onto the proceeding post-translationally modified T domain.[50] This post-translation modification occurs through an attachment of a phosphopantetheine moiety to a conserved serine found within all T domains by a phosphopantetheinyl transferase (PPtase) and is required for a fully functioning NRPS assembly line.[51] The free thiol of the phosphopantetheine arm tethers the growing chain to the enzyme through nucleophilic attack on the activated amino acid, creating a thioester-linkage between the amino acid monomer and the assembly line.[50] Finally, the C domain acts to catalyze peptide bond formation between adjacent, tethered building blocks. The process of monomer selection and subsequent peptide bond formation continues along the assembly line, further elaborating the product until it is released by a thioesterase (TE), forming a macrocycle or free acid, or a reductase (Re) domain, releasing a terminal aldehyde or alcohol product.[52]

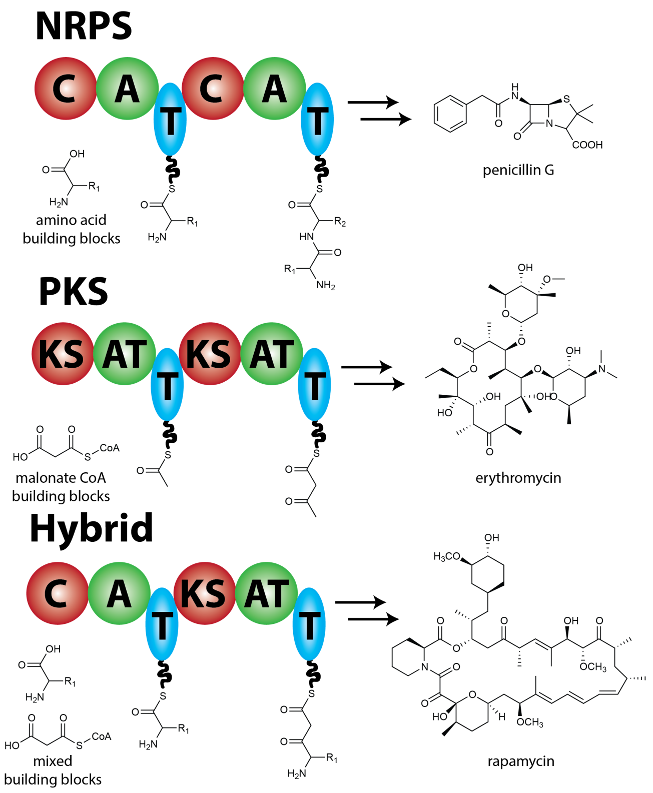


Figure 1.1 Overview of NRPS, PKS and Hybrid NRPS/PKS assembly lines. Building block units and examples of natural products produced by each type are shown.

The module architecture of NRPSs allows increased NRP chemical diversity compared to ribosomal peptides and their coding capability ensures it (20 proteinogenic amino acids, versus over 500 proteinogenic and non-proteinogenic amino acids).[53, 54] Ten conserved amino acids found within the A domain protein sequence make up what is known as the 'NRPS Code' and are predictive of the amino acid monomer that is selected. The NRPS code imparts similar predictive information to the growing NRP as the 3 letter genetic codons do for amino acid sequence prediction.[55] This coding imparts a predictive scheme to reveal monomers activated by the A domains in an NRPS, and the rules of modular collinearity dictate the order of these in the NRP.[56-58]

Further structural complexity is achieved through the presence of auxiliary domains within the assembly line mega-enzyme, including methyltransferases, epimerases, cyclases, etc.[2, 59, 60] The functions of these enzymatic domains are well characterized and the chemical modifications that occur on the growing NRP backbone can also be predicted.[2, 59, 60] In addition to the assembly lines themselves, many tailoring enzymes exist within NRPS and PKS gene clusters that act on released assembly line products to impart even further chemical complexity including glycosyltransferases, oxygenases, reductases, etc.[51, 52] However, it is difficult to define exact locations that modification will occur on a released product, hindering completely error-free prediction of NRPs and PKs. Since the sequencing of the erythromycin assembly line, the ability to predict NRPS products has improved and small molecules can now be included in the central dogma of life (DNA->RNA->Protein->Small Molecule).[61]

Advances in natural product biosynthesis understanding have not translated to increased discovery of new natural products for therapeutic development. Decades of bioactivity-guided investigations without advancements in isolation technology are largely responsible for this decline. Using bioactivity as a sole method for discovery, favoured isolation of abundant and generally cytotoxic compounds, increasingly leading to the rediscovery of known structures. During this time, pharmaceutical companies were under pressure to increase drug leads and as a result, developed methods for high throughput screening (HTS) of desired targets.[62] The low throughput of natural product screening has hindered natural products adoption within the HTS platform causing many companies to reduce or terminate their natural product discovery programs to meet the demands of HTS.[63] Companies realized that coupling HTS with combinatorial synthesis methodologies could create enormous synthetic chemical libraries consisting of millions of compounds that could be screened against a target of interest. The suggested benefits of combinatorial chemistry never truly materialized and between 1981 and 2010, 75% of anticancer and antimicrobial new drug approvals were either natural products or natural product inspired drugs.[30] For the pharmaceutical industry the problems with natural product discovery remain as: low throughput, high rate of rediscovery, and time consuming purification processes. However, advances in DNA sequencing technologies combined with increased biosynthetic understanding are spurring a new era of drug discovery from natural sources using a genome guided approach.[64]

In the late 1990s, the genomics revolution began and the true biosynthetic potential of microbes was revealed. In 2002, the first genome of an antibiotic producing soil bacterium (*Streptomyces coelicolor*) was published.[65] *Streptomyces coelicolor* is a model, soil-dwelling bacterium, and has been extensively studied for decades. At the time of sequencing, it was known to produce several antibiotics, including actinorhodin, prodiginines, and the calcium-dependent antibiotic (CDA).[65] Through genome analysis, the biosynthetic clusters encoding their production could be identified; more surprisingly, the *S. coelicolor* genome revealed an additional 18 ‘cryptic’ biosynthetic clusters with no known products.[65] The realization that researchers had only isolated a fraction of the compounds predicted by genome sequencing, inspired an era of ‘genome mining’ for new natural products from sequenced organisms.

Natural product discoveries have expanded chemical space and identified many classes of chemicals, or pharmacophores, with defined bioactivities.[1, 66] It is well appreciated that molecules that occupy similar areas within this privileged chemical space will have a propensity to act on similar targets.[66, 67] This chemical-biological relationship is important for pharmaceutical development, where lead natural product scaffolds are often modified through synthetic, medicinal chemistry approaches to expand chemical space around an active pharmacophore to optimize potency, toxicity and/or pharmacokinetic profiles.[67] However, in many cases, natural product drug leads are challenged in development programs due to their complexity and inability to be readily chemically synthesized or modified to make important pharmacologic improvements. However, microbes already supply natural variation where a series of analogs arise from NRPS and PKS A/AT domain loading flexibility and/or inconsistent tailoring enzyme action.[29, 68, 69] These variants will usually act on the same target, but with varying potency, providing early access to chemical space surrounding an active pharmacophore. From this, initial structural activity relationships and medicinal chemistry experiments are carried out to identify a lead compound for development.

In addition, available genome sequence data reveals that structurally similar natural products produced by different organisms have similar biosynthetic gene clusters. This connection allows for the identification of potential variants of a desired scaffold of pharmacophore using only genomic information. Therefore, methods to directly identify and target desired chemical pharmacophores within microbial genomes and/or complex extracts are useful for expanding chemical space around desired biological targets where synthetic chemistry is precluded. Ecopia Biosciences Inc. provided one of the first examples of genome mining using genomic information to guide isolation of new enediyne small molecules and represents one of the first hypothesis driven approaches to the directed discovery of a natural product.[70] The enediyne class of natural products are the most potent antitumour compounds ever isolated and have a characteristic chemical 'warhead' responsible for their activity.[71] By identifying the biosynthetic genes responsible for the production of the 'warhead' in known producers, the researchers could extend genomic analysis to identify enediyne biosynthetic genes in other organisms not known to produce enediynes, resulting in the detection of 11 cryptic enediyne clusters.[70, 72, 73] Besides the similar genes encoding for warhead biosynthesis, these clusters often differed from the known producer, suggesting they produced variant molecules while maintaining the desired pharmacophore.[70] This highlights how genome mining can expand theoretical chemical space around known pharmacophores potentially leading to the identification of molecules with more desirable properties for drug development. Although genome mining revealed the presence of enediyne clusters within new microbes, the researchers were still tasked to find conditions in which these molecules could be produced and required bioactivity tests specific to the enediyne mechanism of action to identify and physically locate the new compounds within the extract. In other cases, where gene cluster products are unknown a prediction-based approach can be used to find these unidentified, but genetically encoded small molecules ('known unknowns') to direct isolation.

Coelichelin was the first reported NRP isolated through prediction-guided genome mining.[70] Within the *S. coelicolor* genome, a 3 module, cryptic NRPS was identified.[65, 74] Using the established NRPS code and rules of collinearity, the researchers were able to predict the amino acid monomers and propose a structure for the unknown NRP.[74] By recognizing the prediction contained amino acids associated with metal binding, the researchers induced production with desferric culture conditions, and through comparison to an NRPS knockout mutant, were able to isolate the siderophore, coelichelin.[74] This proposed a new prediction-based approach to identify NRPs from other cryptic gene clusters without a known product.

The amount of data generated from genome sequencing experiments is enormous and researchers quickly understood the need to create automated software to identify and annotate biosynthetic gene clusters. The fact that on average soil bacteria have up to 10 times more clusters within their genome than isolated compounds is astounding.[57, 58, 65, 75] This led to the creation of programs such as ClustScan, CLUSEAN, SBSPKS, SMURF, NRPS Predictor, NPsearcher and others that could annotate domain organization within an NRPS or PKS gene sequence, while simultaneously applying the ‘NRPS code” to reveal A domain monomer selectivity.[76-81] This software has made it relatively easy for manually predicting PKS and NRPS products. Currently software such as AntiSMASH, is available for analyzing entire genomes, including fragmented contig sequence data, for biosynthetic gene clusters.[82, 83] It can also apply the NRPS code and automatically predict products generated by the assembly line. This software is still in development and often does not accurately elaborate the released assembly line product, predict auxiliary domain modification, or other modifications arising from nearby tailoring enzymes. As such, manual prediction is still required.

Identification and annotation of biosynthetic gene clusters from microbes is now commonplace. The current high speed and low cost of genome sequencing is exponentially increasing the available microbial genomes, revealing even more biosynthetic potential outside the traditional soil producers and new strategies will be required for exploring these non-traditional sources. Although this biosynthetic potential exists it is difficult to access with traditional bioactivity based screening. Molecules still need to be potent or abundant enough to be detected in traditional bioactivity assays or time-consuming genetic knockout procedures must be used to identify and isolate the target gene cluster products. The overall problem remains for how to determine what clusters to target, how to induce production in silent clusters, and how to directly identify a predicted metabolite or desired pharmacophore without the aid of bioactivity information. The biosynthetic potential of microbes is now revealed, but suitable technology to target the unknown yet genetically predicted NRPs is lacking. My thesis addresses these problems, and advances technology and methodology for informatically targeting privileged chemical space surrounding NRPs for both therapeutic and industrial purposes (**Figure 1.2**).

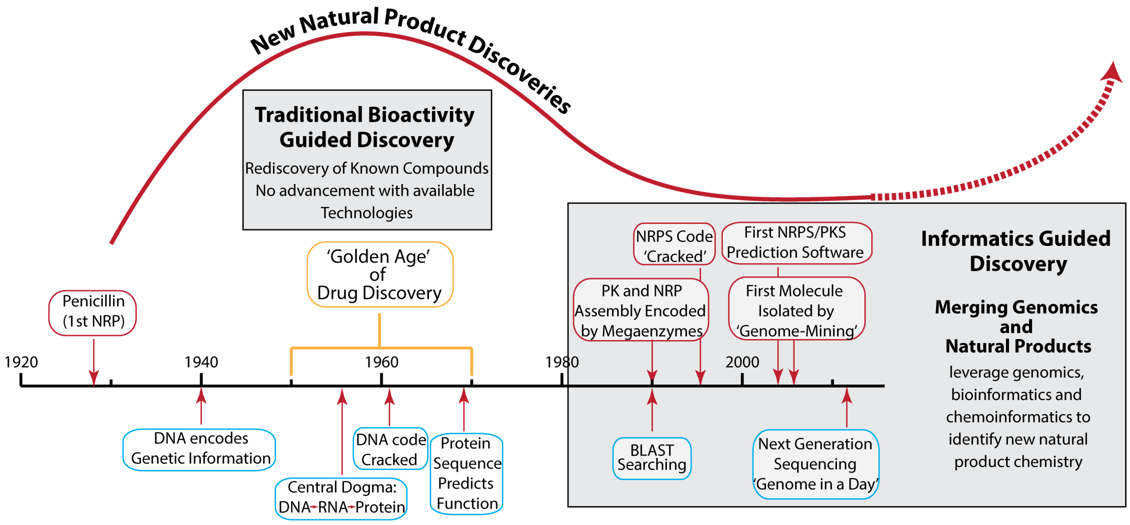

New Natural Product Discoveries
Traditional Bioactivity Guided Discovery
Rediscovery of Known Compounds
No advancement with available Technologies
'Golden Age' of Drug Discovery
Penicillin (1st NRP)
1920
1940
1960
1980
2000
DNA encodes Genetic Information
Central Dogma: DNA-RNA-Protein
DNA code Cracked
Protein Sequence Predicts Function
NRPS Code 'Cracked'
First NRPS/PKS Prediction Software
PK and NRP Assembly Encoded by Megaenzymes
First Molecule Isolated by 'Genome-Mining'
BLAST Searching
Next Generation Sequencing 'Genome in a Day'
Informatics Guided Discovery
Merging Genomics and Natural Products
leverage genomics, bioinformatics and chemoinformatics to identify new natural product chemistry

Figure 1.2. Thesis context.
Timeline of major scientific genetic breakthroughs (blue) and natural product breakthroughs (red). My thesis addresses the merging of genomics and natural products to create an informatics-based strategy and technology for directing natural product discoveries.

1.2 Scope and Nature of this Work.

Natural product drug discovery has consistently relied on bioactivity to find new NRPs from microorganisms. This body of work provides alternative methods and technology that leverages genomic information to selectively identify new NRPs from their genomic predictions to their targeted isolation from complex extracts. The central hypothesis surrounding this thesis is that nonribosomal peptides can be targeted through the application of genomic, metabolomic, and chemoinformatic strategies. Three related research projects, consisting of 5 published scientific articles, expand approaches and technology to identify and isolate novel NRPs.

1.3 Genome Mining for Nonribosomal Peptides

The ability to predict 'cryptic' natural products is increasing and addressing and refining methods using model systems will lead to selective NRP targeting strategies. This is an important step forward for natural product drug discovery as it removes the barriers and randomness associated with natural product research. Distinct hypotheses can be set to query their natural activities, which in turn will spur deeper insights into the secondary metabolism that coordinates complex behaviours, chemical communication and modes by which microbes defend themselves against other organisms and harsh environments. In the first two projects, I describe how genomic information can provide structural predictions to guide researchers to target NRP natural products and direct their isolation.

Genomic sequencing shows other bacterial families from other environmental niches that are also biosynthetically gifted.[84, 85] Small molecules isolated from bacteria that have evolved in unique environments have similarly developed activities that provide the producer with increased fitness. By using genomic information, I demonstrate the directed discovery and characterization of new pharmaceutically and industrially important NRPs that increase bacterial fitness within two disparate, but similarly harsh niches: humans, where pathogens are continually under attack by the immune system, and gold-nuggets, an environment with high heavy metal toxicity.[86-89]

The structural complexity of NRPs often precludes our ability to accurately predict their final structure from genomic information. For this reason, it is important to first apply and perfect NRP prediction rules to small gene cluster systems, so that predictions of increasingly complex gene clusters are more accurate. When an NRP biosynthetic gene cluster is relatively small, predictions are more accurate due to a decreased number of possible building blocks and a reduced number of chemical sites that can be modified by tailoring enzymes. In **Chapter 2**, a dimodular NRPS was identified and its product predicted, within the human pathogen, *S. aureus.*[89] This provided a very accurate prediction of the isolated products, the aureusimines, differing by only an unforeseen ring rearrangement in the isolated NRP. Using this small NRPS as a model system in **Chapter 3** and **Chapter 4**, additional in depth biosynthetic research projects were completed that add to the understanding of NRP assembly, flexibility and heterologous production in *E. coli*, whose findings can be applied to complex NRPS gene clusters to facilitate more accurate NRP predictions.[90, 91]

The successful prediction and examination of a dimodular NRPS aided in the NRP prediction of an 11 module system within *D. acidovorans.* In **Chapter 5***,* I applied several aspects of genome mining to hone in on the NRP of interest.[88] First, a preliminary analysis of the cryptic NRPS cluster revealed several flanking genes associated with heavy metal resistance that might infer its activity, gold toxicity remediation. Second, examination of the assembly line genes, and tailoring genes allowed the prediction of a putative structure that could be leveraged in two ways. First, the prediction showed the incorporation of several modified amino acids whose structures are directly linked to metal binding, suggesting the cryptic metabolite was a metallophore, and second, the structure helped to focus on only molecules that fell within a range of masses surrounding the predicted mass (mass window). Together, this information led to the isolation of the gold complexing small molecule, delftibactin and provides a successful application of several genome mining strategies to large, complex assembly line systems for the identification and isolation of an important NRP natural product.

1.4 New NRP Discovery Methods

Genome mining provides information on the final NRP structure, which aids in its isolation. A major hurdle still exists in actually identifying the peak/compound associated with the NRP of interest within complex extracts. Often, a cryptic NRP is identified using genetic techniques, where a gene cluster of interest is silenced through genetic insertion or deletion. By comparing the extract of this genetic mutant to its wild-type parent, the target NRP can be directly identified. Genetically inactivating gene clusters is a time consuming process and is often difficult within disparate organisms where no or minimal genetic manipulations are known. However, this technique is effective and can directly target the cryptic NRP for isolation of sufficient quantities for full structure elucidation. When using a prediction alone, many molecular possibilities exist that fall within the predicted mass window and are only linked to the target gene cluster after full isolation and structure elucidation. It is necessary to have a more effective and sensitive approach to identify predicted metabolites, distinct from bioactivity and time consuming genetic techniques. By employing more sensitive compound identification techniques, such as mass spectrometry, potential leads within mass windows surrounding the targeted predictions are identified and additional structural information can be acquired using tandem MS techniques, needing only picogram quantities of the target metabolite. By comparing hypothetical prediction information with real structural information from MS/MS experiments, it is possible to use a statistical scoring algorithm to accurately identify a desired compound.[92]

The sequencing and structural characterization of ribosomal peptides by *de novo* tandem MS is common.[93] Ribosomal proteins undergo MS fragmentation generating patterns arising from energy-induced cleavage at amide bonds. Ribosomal peptides are made up of a pool of only 20 amino acid building blocks, each with a characteristic MS fragment. Mass spectrometry fragmentation of the ribosomal peptide leads to characteristic amino acid fragments that are used to determine its peptide sequence. There are several proteomic software programs available for the dereplication of ribosomal proteins that are able to match *de novo* peptide MS sequences to real and hypothetical peptide databases, including PEAKS, Mascot and Sequest.[94-97] Nonribosomal peptide dereplication by MS is more difficult due to the presence of over 500 known amino acid building blocks, additional downstream chemical tailoring events, the presence of hybrid PK-NRP molecules, and the existence of linear, cyclic and branched forms that contribute to the overall NRP fragmentation complexity. The chemical diversity created by NRPS and NRPS-PKS hybrids makes *de novo* peptide sequencing extremely difficult even when the chemical structure is known and is virtually impossible for use in structure elucidation without additional data from other sources such as NMR.

Early endeavours to dereplicate NRPs have tried to identify mass spectral fragments of amino acids from *de novo* MS^n experiments and match them back to a set of known monomers.[98-103] Problems in this approach can happen if fragmentation does not occur for every monomer, or if structural complexity exceeds the pool of known monomers. Another barrier to *de novo* sequencing of NRPs is that experiments are not automated and require manual annotation of MS^n spectra of purified compounds. For genome mining it means they effectively only match structures back to the gene cluster rather than using genomic data to aid in identification.[104, 105] Advancements in computer algorithms to create a more automated approach to dereplication are necessary to access the entire chemical diversity generated by NRPs and other natural products.

Within our lab, we developed a tandem MS approach and algorithm that addresses the issues surrounding dereplicating diverse NRPs from complex mixtures known as the informatic search for natural products (iSNAP).[92] iSNAP uses a curated database of all known NRPs to generate hypothetical fragmentation patterns for each compound and compares these hypothetical fragments to real fragments generated by automated LC-MS/MS experiments of complex extracts. By statistically scoring matched fragment hits, a known NRP is accurately dereplicated from a complex extract. Theoretically if an NRP is predicted with 100% accuracy from an NRPS gene cluster that cryptic molecule could be directly identified within a complex extract irrespective of activity using the iSNAP algorithm. As previously mentioned, predictability is limited due to flexible monomer incorporation and unpredictable tailoring chemistries, however, if a hypothetical compound database incorporates all predicted/possible monomer substitutions and tailoring modifications, an accurate prediction may be generated that facilitates the identification of the cryptic compound using iSNAP. My final project in **Chapter 6** develops the iSNAP platform to target predicted compounds and desired pharmacophores using a user generated hypothetical structural database.[106]

1.5 Thesis Overview

The major projects of my thesis provide a comprehensive methodology of how to target and isolate microbial NRPs in the current genomic era (**Figure 1.3**). Using genome mining, I show that hypotheses for natural product isolations are possible and that metabolite predictions from di-modular NRPSs to more complex multi-modular (11 module) systems are useful for directing isolations of genomically-encoded natural products. Finally, I have advanced technology and methodologies for the directed discovery of NRPs, which can be further applied to other natural product classes. This represents the first instance where an unknown metabolite can be targeted and localized within a complex extract without the need for purification or bioactivity guided fractionation, providing a technological advance for natural product drug discovery.

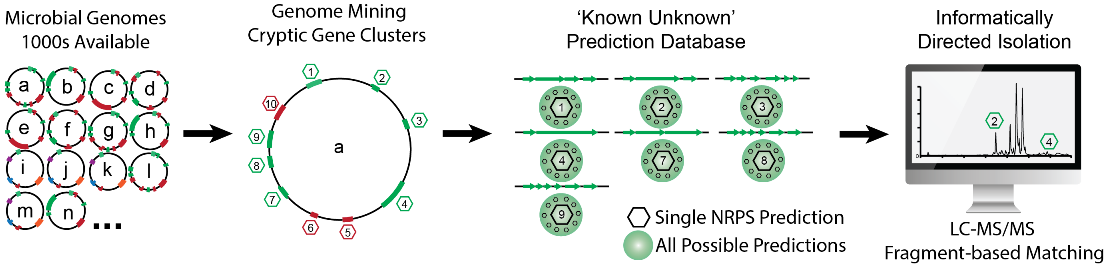


Figure 1.3 Overview of the informatics strategy to isolate cryptic secondary metabolites. Thousands of bacterial genomes are available containing biosynthetic gene clusters with known compounds (red) and cryptic biosynthetic gene clusters with no known product

(known unknown) (green). Hexagons are representative of the known molecules (red) and the predicted molecules (green) genetically encoded within the genome of bacterium **a**. Prediction databases can be generated based on predicted scaffolds of 'known unknowns' from bacterium **a**, representing monomer variation or varied tailoring events. Target cryptic metabolites are directly targeted informatically by iSNAP using the 'known unknown' database to interrogate LC-MS/MS data from crude bacterial extracts of organism **a**.

Chapter 2. Genome Mining Nonribosomal Peptide Synthetases

2.1 Chapter Preface

Using a genome mining approach, a single dimodular NRPS biosynthetic gene cluster was identified within all sequenced *S. aureus* strains. This provided an ideal system for applying NRPS prediction rules to a small NRPS gene cluster to guide isolations through the predicted NRP. *Staphylococcus aureus* has been studied for over a century and produces several small molecules, but was not known to produce NRPs. Genome mining revealed this cryptic metabolite and knowing the importance of secreted factors for *S. aureus* virulence and the propensity of NRPs towards biological targets, it was hypothesized that the NRP may be important for *S. aureus* pathogenesis.

Using the predicted mass of the cryptic dipeptide, I scanned within the LC-MS chromatogram to hone in on compounds within a mass range surrounding the prediction. This led to the isolation of two cyclic dipeptide NRPs, the aureusimines. By creating a gene knockout of the NRPS, we were able to confirm the aureusimines were products of the AusA NRPS and comparison of wild type and knockout strains determined the aureusimines play a role in virulence using both a mouse bacteremia model and RNA expression experiments. These studies indicated the aureusimines were important for virulence and were responsible for initiating the expression of many virulence-related proteins, providing validity for our hypothesis driven approach to NRP discovery. Unfortunately, an inadvertent mutation occurred within the knockout in an important virulence regulatory protein (*saeR)*, which resulted in an over estimation of the aureusimine involvement in virulence within the published article; however, subsequent experiments using a ‘clean’ NRPS knockout in mouse models and microarrays verifies the aureusimines involvement in virulence, as stated in the correction (see Clarification Section 2.8). Most importantly, this research presents the success of a prediction-guided approach within *S. aureus* to reveal important NRPs that were not discovered by traditional methods.

The following chapter is a modified version of a previously published journal article in which I was the lead author. I contributed to experimental design, interpreting results and writing the manuscript. I was a contributor to all experiments; assisting in isolating the compound, generation of the *ausA* deletion mutant, gene expression experiments and mouse virulence models. Wenliang Wang determined the final structure of the aureusimines. Christelle Roux provided technical assistance running the microarray. Federico Beasley conducted the mouse virulence experiments with assistance from me. Nathan Magarvey provided significant assistance and advice during the entirety of the research project. The citation for this publication is as follows:

Wyatt, M. A., Wang, W., Roux, C. M., Beasley, F.C., Heinrichs, D.E., Dunman, P.M., and Magarvey, N.A., ***Staphylococcus aureus*** **Nonribosomal Peptide Secondary Metabolites Regulate Virulence.** *Science* 16 July 2010: **329** (5989), 294-296.Published online 3 June 2010

2.2 Abstract

Staphylococcus aureus is a major human pathogen that is resistant to numerous antibiotics in clinical use. We found two nonribosomal peptide secondary metabolites—the aureusimines, made by *S. aureus*—that are not antibiotics, but function as regulators of virulence factor expression and are necessary for productive infections. *In vivo* mouse

models of bacteremia showed that strains of S. aureus unable to produce aureusimines were attenuated and/or cleared from major organs, including the spleen, liver, and heart. Targeting aureusimine synthesis may offer novel leads for anti-infective drugs.

2.3 Introduction

Staphylococcus aureus is a human pathogen commonly causing hospital and community-acquired infectious diseases (1). It has an array of virulence factors, including surface proteins responsible for adhesion and invasion of host tissues (e.g., fibrinogen and fibronectin-binding proteins), exoproteins responsible for immune evasion (e.g., chemotaxis-inhibitory protein), and numerous hemolytic and pore-forming toxins (e.g., hemolysins, leukocidins, and enterotoxins) (2–4). For successful infection, a coordinated release of virulence factors is necessary, and redundancies exist, such that, if one factor is ablated, a productive infection can still ensue. Early research by Novick and colleagues identified an accessory gene regulator (agr) that controls several virulence factors (5). Expression of the agr locus is positively regulated by the agr pheromone, a ribosomally encoded secondary metabolite (6). Subsequent genomic sequencing has revealed that homologs of the agr pheromones exist in several Gram-positive cocci, many of which are not pathogenic (7–11). Although referred to as the "master" regulator of S. aureus virulence, expression of agr is not always detected *in vivo*, and agr-deficient clinical isolates are known, which raises the possibility that other small molecules factor prominently in the regulation of virulence factor expression (12).

A major class of bacterial secondary metabolites comprises the nonribosomal peptides, which are produced, in microorganisms, by multifunctional enzyme assembly lines known as nonribosomal peptide synthetases (NRPSs) (13). Antibiotics are the best known nonribosomal peptides produced by soil-dwelling microbes, which use them as weapons and for cell-cell communication (14). Penicillin, for example, is not constructed ribosomally but is dependent on an NRPS that uses valine, cysteine, and α-aminoadipic acid precursors (15). Although penicillin was the first nonribosomal peptide used for *S. aureus* infections, *S. aureus* itself has not previously been shown to construct nonribosomal peptides.

2.4 Cryptic nonribosomal peptide assembly in Staphylococcus.

An NRPS uses adenylation (A) domains and adenosine triphosphate (ATP) to activate adenosine monophosphate esters of selected amino acids and delivers them to posttranslationally modified (phosphopantetheine) NRPS thiolation (T) domains (13). Condensation reactions of the T domain–tethered amino acids produce growing peptide chains, which are released by thioesterase or reductase (Re) domains at the C terminus of the NRPS, the latter as peptide aldehydes or alcohols that frequently undergo intramolecular cyclization reactions (13, 16). Details of NRPS-catalyzed reactions, structural assignments of several NRPS domains, and assembly rules for nonribosomal peptide on NRPSs are reasonably well known. Genes encoding for a given nonribosomal peptide are also clustered. Coopting these genetic and/or biochemical parameters with microbial genomic sequencing has created a link between gene and small-molecule prediction, which has assisted in the discovery of unidentified, or “cryptic,” nonribosomal peptides, a process referred to as secondary metabolite genome mining (17, 18).

We used a genome-mining approach to predict nonribosomal peptides that are exclusive and highly conserved within *S. aureus*. Scanning in excess of 50 S. aureus sequenced genomes led to the identification of a universally conserved (average of 97% identical and 97% similar), yet undescribed, NRPS gene cluster (annotated as a gramicidin synthetase or hypothetical protein) (**Fig. 2.1A** and **Fig. S2.1**) (19, 20). This cluster contains an NRPS gene (7.17 kb) that takes up 0.25% of the *S. aureus* genome. An ortholog is present in other staphylococci pathogenic to humans, including *Staphylococcus epidermidis* (53% identical and 71% similar), *Staphylococcus capitis* (53% identical and 70% similar), and *Staphylococcus lugdunensis* (53% identical and 70% similar), but is absent in other staphylococci or Gram-positive cocci (**Figs. S2.1** and **S2.2**) (20, 21). Buoyed by the association of this NRPS with staphylococci pathogenic to humans, we predicted the structure of the encoded nonribosomal peptide.

The *S. aureus* NRPS (2389 amino acids) is a dimodular NRPS with two adenylation (A) domains having strictly conserved NRPS codes for valine and tyrosine for all staphylococci containing the NRPS (**Fig. 2.1B** and **Fig. S2.3**). A Re domain is found at the C terminus of all the *S. aureus* NRPSs and probably releases the valine-tyrosine dipeptide as an aldehyde (22). The proposed linear dipeptide aldehyde is likely to result in the formation of a cyclic imine (predicted mass of 262.17), promoted by nucleophilic attack of the aldehyde by the α-amine of valine (**Fig. 2.1B**) (16, 22).

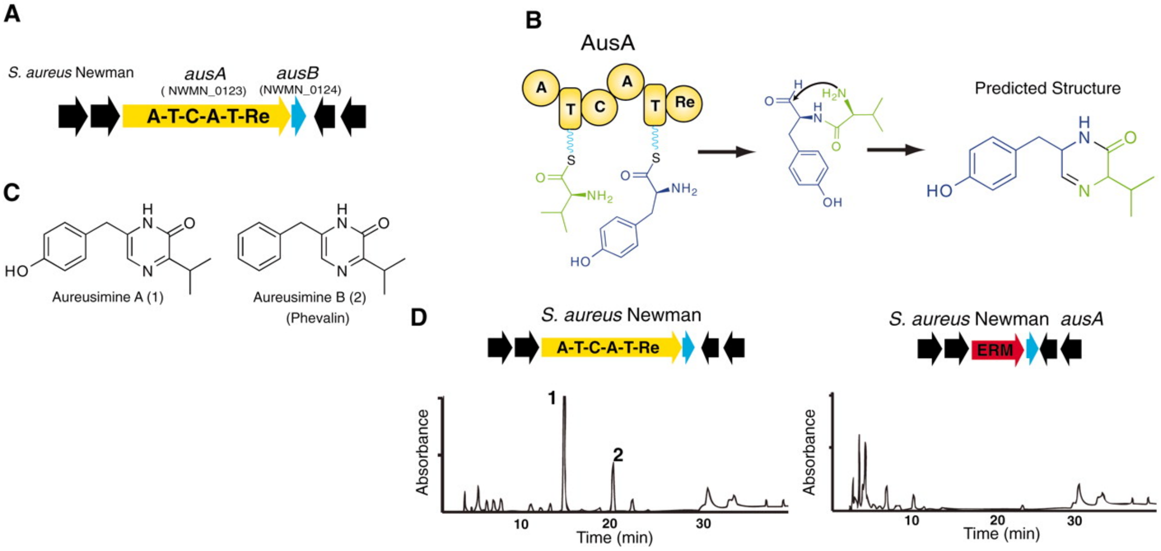


Figure 2.1. Identification of a cryptic NRPS biosynthetic gene cluster within S. aureus. (**A**) Genetic loci of *S. aureus* Newman containing the NRPS gene. The NRPS locus is found in all sequenced *S. aureus* genomes. The NRPS cluster contains two open reading frames: ausA (the NRPS gene) and immediately downstream of it *ausB* (phosphopantetheinyl transferase). *ausB* encodes the enzyme (AusB) predicted to posttranslationally modify AusA with a 4′-phosphopantetheine prosthetic group. (**B**) *S. aureus* NRPS is a dimodular nonribosomal peptide assembly line encoding a putative cyclic dipeptide. Domains A, C, T, and Re within the S. aureus NRPS (AusA) are shown as round spheres shaded in yellow. Curved blue lines originate from the T domain and indicate the phosphopantetheinyl arm that is predicted to be delivered via action of AusB. Amino acid substrates (valine and tyrosine) were predicted according to established NRPS codes (**Fig. S2.2**) (17, 20). Release of a linear valine-tyrosine dipeptide aldehyde and the predicted nonribosomal peptide structure are shown. (**C**) Identification of *S. aureus* nonribosomal peptides. Structures of aureusimine A and aureusimine B (phevalin) were determined by mass spectrometry and NMR experiments (**Figs. S2.4 to S2.7**). (**D**)

Liquid chromatographic separations (HPLC chromatograms) of organic extracts of *S. aureus* Newman and S. *aureus* Newman ΔausA. Aureusimine A (peak 1) and aureusimine B (phevalin) (peak 2) are present within extracts of *S. aureus* Newman but absent in extracts of *S. aureus* Newman ΔausA strain. ERM, erythromycin.

2.5 Isolation of tyrosine-valine dipeptides.

To isolate the cryptic *S. aureus* nonribosomal peptide, we collected organic solvent extracts of *S. aureus* culture broths and subjected them to high-performance liquid chromatography (HPLC) and mass spectrometry analysis with mass-spectral filtering software to identify products within the range of the predicted dipeptide mass (**Fig. S2.4**). Two peaks were obtained, one providing a nearly exact match and the second with a slightly later retention time and differing by 16 mass units (**Fig. S2.4**). Both molecules had a common absorbance spectrum, indicating that the two were congeners (produced in a ~3:1 ratio) or sufficiently similar to suggest that they stem from a common NRPS pathway (fig. S4). One-dimensional and two-dimensional nuclear magnetic resonance (NMR) experiments (**Figs. S2.5** to **S2.7**) provided the structures of both metabolites (**Fig. 2.1C**). The most abundant matched the prediction of the cyclic valine-tyrosine dipeptide bearing a pyrazinone core (**Fig. 2.1C**), and we called it aureusimine A. The second molecule (aureusimine B) had a phenylalanine in place of the tyrosine with a structure that matched a previously identified cyclic dipeptide (phevalin) from *Streptomyces* sp. SC433 (23). Production of two related nonribosomal peptides from a single NRPS is commonplace and is consistent with the second A domain's incorporation of both tyrosine (aureusimine A) and phenylalanine (aureusimine B or phevalin).

To verify that the aureusimines are synthesized by the *S. aureus* NRPS (encoded by the gene we have named *ausA*), an allelic replacement was used to replace ausA with an erythromycin-resistance cassette. Culture broths of the resulting Δ*ausA S. aureus* strain were devoid of aureusimine A and B (**Fig. 2.1D**). We next compared the growth of the *ausA* deletion strain with that of the wild type and found that the aureusimines are not necessary for growth and that the *ausA* deletion strains actually grew better than the wild type (**Fig. S2.8**).

2.6 Microarray analysis of virulence expression.

Our discovery of a nonribosomal peptide unique to *S. aureus* raises the possibility for its role as a regulator of *S. aureus* virulence factor expression. To evaluate the impact of aureusimine on virulence gene expression, we conducted global microarray analysis. Both *S. aureus* Newman and Newman Δ*ausA* overnight cultures were diluted 1:100 in tryptic soy broth and grown until early exponential [absorbance at 600 nm (A_{600}) = 0.3] and late exponential phase (A_{600} = 1.2). mRNA was isolated from each strain and used for microarray experiments. In three separate experiments, primary metabolic genes were largely unchanged in the Δ*ausA* strain, a result consistent with growth studies. However, in comparison with its isogenic parent, the *ausA* mutant displayed significant differences in expression of a large number of virulence genes, including genes encoding immunomodulatory proteins, host cell adhesins, chemotaxis-blocking proteins, and host-targeted lytic proteins and cytotoxins (**Fig. 2.2A**) (**Tables S2.2** to **S2.5**). For example, genes encoding chemotaxis-inhibiting protein and formyl peptide receptor–like

1inhibitory proteins, important for *S. aureus* immune evasion, are massively up-regulated by aureusimine production, >100 times (145.9) and >50 times (73.5), respectively (3, 24). *S. aureus* adhesion molecules, such as fibrinogen-binding protein (Efb) and fibronectin-binding protein A (FnbA), are necessary for endothelial cell invasion and endocarditis and are also up-regulated 187.5- and 75.2-fold, respectively (25–27). Genes encoding for hemolysins were also significantly up-regulated in strains producing aureusimines. As an illustration showing that the transcriptional profiling corresponds to phenotypic alterations, the blood-lysing capacity of both wild-type and mutant strains was compared on blood agar plates (**Fig. 2.2B**). *S. aureus* Newman colonies lyse red blood cells, whereas little to no clearing or lysis was observed by Δ*ausA* colonies. The hemolytic property of S. aureus could be restored to the Δ*ausA* strains by the addition of aureusimines A and B (100 μg/ml) into the blood agar plates (**Fig. 2.2B**).

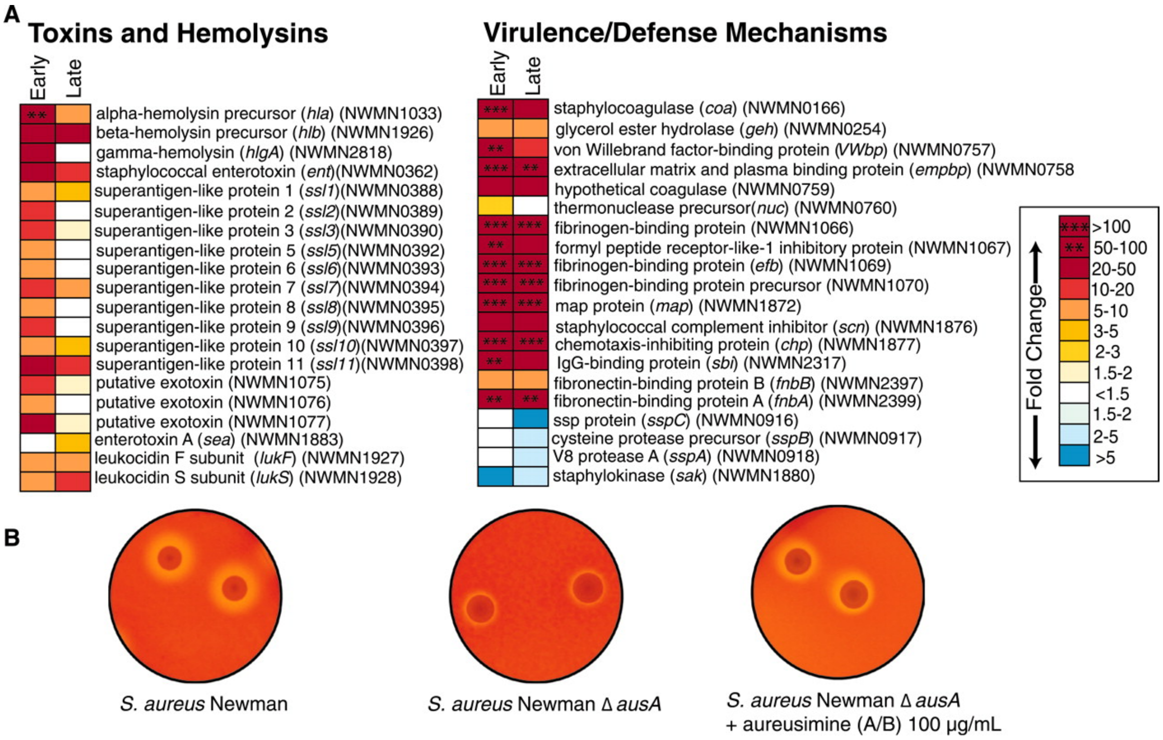


Figure 2.2. Gene regulation by the aureusimines.

(**A**) Differential gene expression caused by the presence of aureusimines A and B in S. aureus in early and late exponential phase growth. Results are presented as mean fold up-regulation (shades of red) and down-regulation (shades of blue) in three separate experiments (see scale bar). The complete microarray results of genes regulated by the aureusimines can be seen in **Tables S2.2** to **S2.5**. (**B**) Aureusimines induce hemolysis. (Left) S. aureus Newman wild-type and (center) *S. aureus* Newman Δ*ausA* were grown on 5% sheep blood agar; (right) *S. aureus* Newman Δ*ausA* was grown on 5% sheep blood agar supplemented with 100 µg/mL aureusimines A and B. Zones of clearance around colonies indicate hemolysis.

2.7 Role of aureusimines *in vivo.*

To gain further insight into the role that aureusimines play in the infectivity and virulence of *S. aureus*, groups of BALB/c mice were injected intravenously with either *S. aureus* strain Newman or the isogenic *ausA* mutant. Over the course of 4 days, mice infected with wild-type *S. aureus* Newman lost, on average, 22% of their original weight, consistent with a productive *S. aureus* infection (**Fig. 2.3A**). In contrast, mice infected with the *S. aureus* Δ*ausA* deletion strain lost, on average, only 7.5% of their original weight (**Fig. 2.3A**). The weight change data for the two groups of mice are significantly different as determined by the Student's t test ($P < 0.001$). Organs from both groups of mice were removed and homogenized, and the resulting suspensions were surveyed for viable *S. aureus* colony-forming units (CFUs). In the group infected with wild-type bacteria, CFUs were high in all organs examined (**Fig. 2.3, B to E**). Although CFUs in samples recovered from kidneys of mice infected with the *ausA* deletion strain were comparable to those from kidneys of mice infected with wild-type *S. aureus*, CFUs obtained from the hearts, spleens, and livers of mice infected with the *ausA* mutant were all significantly less ($P < 0.01$) than those from the respective organs of mice infected with wild-type bacteria, as determined by the Student's t test (**Fig. 2.3, B to E**). In fact, we could not recover detectable CFUs from the hearts of mice infected with the *ausA* mutant (**Fig. 2.3E**).

The aureusimines are previously unidentified nonribosomal peptide secondary metabolites that are integral to the ability of *S. aureus* to act as an infectious agent. Discovery of the aureusimines' control over a wide range of *S. aureus* virulence factors presents opportunities for novel anti-infective strategies. Unlike many other nonribosomal peptide secondary metabolites produced by soil microbes, such as penicillin (28), they do not appear to act as antibiotics. However, the original isolation of phevalin (aureusimines B) from a soil-dwelling actinomycete suggests a possible origin of the aureusimine NRPS (23). For *S. aureus*, acquisition of the aureusimine NRPS biosynthetic machinery was a defining moment.

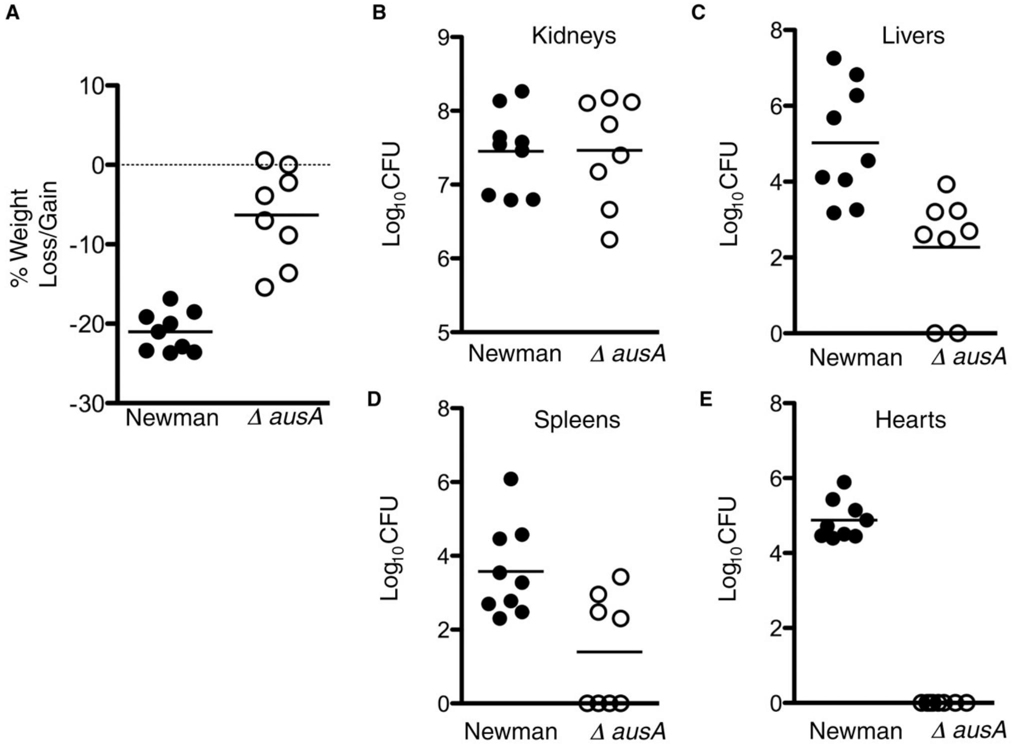


Figure 2.3.

(**A**) Weight change (4 days after infection) for mice infected with *S. aureus* Newman (filled circles) or *S. aureus* Newman Δ*ausA* (open circles). Solid bars represent average weight change. (**B** to **E**) CFUs obtained from kidneys, livers, spleens, and hearts 4 days after infection. Solid bars represent the average $\log_{10}$ CFUs for the group.

2.8 Materials and Methods

2.8.1 Chemical Characterization of Aureusimines

Staphylococcus aureus strains (RN4220, UAMS-1, USA300, and Newman,) and *Staphylococcus epidermidis* strains (ATCC 12228 and RP62A) were grown in Fernbach flasks containing 1.5 L of tryptic soy broth (TSB). The TSB was inoculated with an overnight culture (1:100) and grown for 3 days in an incubated rotary shaker (37°C, 175rpm). The cells and broth were extracted with ethyl acetate, evaporated and dissolved in 1mL of methanol. Concentrated extract was analyzed by HPLC with a Phenomenex Luna 5u C18 column. The mobile phase (was linear from 20 % acetonitrile+ 0.1% trifluoroacetic acid (TFA), 80% water + 0.1% TFA at 5min to 100% acetonitrile+ 0.1% TFA at 45min. Detection of metabolites was carried out using a photodiodearray. Newman ΔausA was fermented and extracted similarly for comparison. Aureusimine A and B eluted at 14.65 min and 20.79 min respectively. Mass spectra were obtained using a Q-Trap (Applied Biosystems) with a mass window set between 220 and 300. The mass for aureusimine A was m+/z = 245.13 and aureusimine B m+/z = 229.11. High resolution mass was obtained on an LTQ OrbiTrap XL (Thermo Scientific) giving m+/z = 245.1292 and m+/z = 229.1343 (Fig. S2.3). 1H spectra of purified S. aureus small molecules were recorded on a Bruker 600MHz NMR spectrometer. Samples were dissolved in deuterated DMSO. δ (integration, multiplicity; J in Hz): Aureusimine A: 1.09 (6-H, doublet; 7.2), 3.22 (1-H, septet; 7.2), 3.61 (2-H, singlet), 6.69 (2-H, doublet; 8.4), 7.05 (1-H, singlet), 7.10 (2-H, doublet; 8.4), 9.30 (1-H, singlet), 12.11 (1-H, singlet). Additional 2D NMR experiments of aureusimine A and B can be seen above (Fig. S2.4 and S2.5). Aureusimine B/phevalin: 1.09 (6-H, doublet; 7.0), 3.23 (1-H, septet; 7.0), 3.74 (2-H, singlet), 7.10 (1-H, singlet), 7.24 (1-H, triplet; 6.7), 7.31 (4-H, multiplet), 12.13 (1-H, singlet). Aureusimine A and phevalin were found in all fermented *S. aureus* and *S. epidermidis* strains.

2.8.2 Genetic Manipulations - *ausA* Allelic Replacement

All primers were purchased from IDT laboratories and amplified using Phusion HF (New England Biolabs). The erythromycin resistance cassette was amplified from pMUTIN4 (S2.6)using the primers: A and B. The left and right flanks were independently amplified from mu50 genomic DNA (ATCC#700699D) using the primers: Left Flank: C and D. Right Flank: E and F. A final PCR was performed using the amplified resistance cassette, the left and right flanks using the primers C and F. This product was purified (Qiagen, PCR Purification Kit) and ligated into the NcoI and PstI sites of the temperature sensitive cloning plasmid pCL52.2 (S2.7). The resulting plasmid, pCNKO, was transformed into DH5-α maximum efficiency competent cells (Invitrogen) using manufacturer protocols. The plasmid was purified and sequenced (Mobix Labs, Hamilton). The recombinant plasmid pCNKO was transformed into S. aureus RN4220 using standard electroporation protocols and selected for on erythromycin (10 μg/mL)(S8). The plasmid pCNKO was transduced from RN4220 to Newman via Φ11 and selected for on tryptic soy agar (TSA) containing erythromycin (10μg/mL). Newman containing pCNKO were grown overnight at 45°C (non-permissive temperature for pCL52.2) on TSA containing erythromycin. Large colonies were re-streaked to check the integrity of the colony. A single colony was inoculated into a 250mL flask containing 25mL TSB and grown at 225 rpm, 30°C. for 5 days. Dilutions (1:1000) were made each day into a fresh flask containing 25mL TSB. After 5 days, the culture was streaked onto TSA containing[12] erythromycin. Colonies to have undergone recombination were confirmed through PCR using primers G and H from outside the flanking region to primers I and J inside the erythromycin resistance cassette for both the left and right flank respectively. These products were purified and sequenced to verify the recombination. Subsequent fermentations showed an absence of aureusimine production.

2.8.3 RNA Isolation and Microarray Analysis

Overnight cultures of both *S. aureus* Newman WT and NewmanΔausA were diluted 1:100 in 250 mL Erlenmeyer flasks containing 50 mL TSB (n=3). Cultures were grown in an incubated shaker at 37°C, 175 rpm and samples were taken at O.D.600 = 0.3 (early exponential) and O.D.600 = 1.2 (late exponential). Cells were centrifuged and resuspended in 100 µL of water containing 300 µg/mL lysostaphin. Samples were incubated for 1 h at 37°C. RNA from lysed samples was obtained using an Aurum Total RNA Kit (BioRad). RNA integrity was verified using automated gel electrophoresis on an Experion system (BioRad). RNA was converted to cDNA and hybridized to an Affymetrix GeneChip constituting the *S. aureus* genome as described by Beenken (S9).

2.8.4 Hemolysis

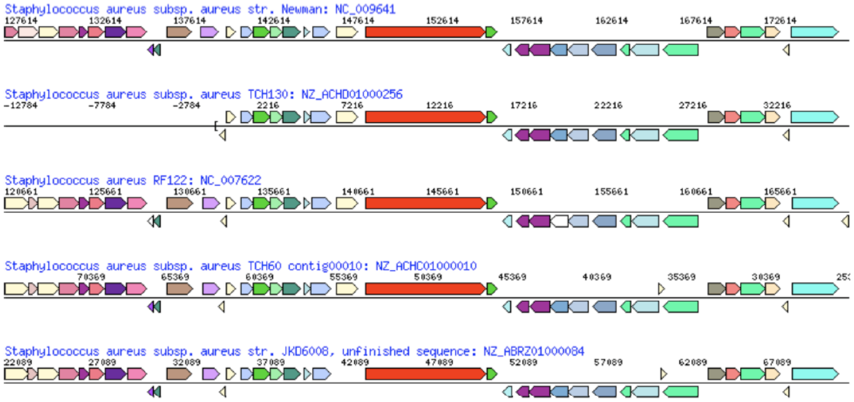
Staphylococcus aureus Newman and S. aureus Newman ΔausA were grown for two days at 37°C on blood agar base No. 2 (BD Biosciences) containing 5% sheeps blood (Biomerieux) followed by an additional 4 days at room temperature.

2.8.5 Systemic Mouse Infections

Seven-week-old female immunocompetent BALB/c mice were purchased from Charles River Laboratories Canada, Inc., and housed in microisolator cages. Overnight S. aureus cultures were diluted 1:40 into fresh Tris succinate (S10) broth and grown to mid-log phase (OD600 approx. 1.0), centrifuged, washed twice and resuspended in saline to yield an OD600 of 0.2 (5 x 10^{7} CFU mL-). Staphylococci were enumerated by colony formation on tryptic soy agar plates to determine the exact infectious dose (Newman, 6.8 x 10^{6} CFU; Newman ΔausA, 7.0 x 10^{6} CFU). Staphylococcal suspension (100 μL) was administered into mice via the tail vein. Ninety hours post-infection, mice were euthanized via intraperitoneal injection of sodium pentobarbital. Organs were aseptically excised and homogenized in PBS containing 0.1 % v/v Triton X-100. Aliquots of homogenized organs were diluted and spotted on tryptic soy agar in quadruplicate for CFU determination. Statistical analysis was performed using the student's unpaired t test using Graphpad Prism software (GraphPad Software, Inc.).

2.9 Supplementary Information.

2.9.1 Supporting Figures and Legends (Supplementary Fig. 2.1-7)


Staphylococcus aureus subsp. aureus str. Newman: NC_009641
127614
132614
137614
142614
147614
152614
157614
162614
167614
172614
Staphylococcus aureus subsp. aureus TCH130: NZ_ACHD01000256
-12784
-7784
-2784
2216
7216
12216
17216
22216
27216
32216
Staphylococcus aureus RF122: NC_007622
120661
125661
130661
135661
140661
145661
150661
155661
160661
165661
Staphylococcus aureus subsp. aureus TCH60 contig00010: NZ_ACHC01000010
70369
65369
60369
55369
50369
45369
40369
35369
30369
25:
Staphylococcus aureus subsp. aureus str. JKD6008, unfinished sequence: NZ_ABRZ01000084
22089
27089
32089
37089
42089
47089
52089
57089
62089
67089

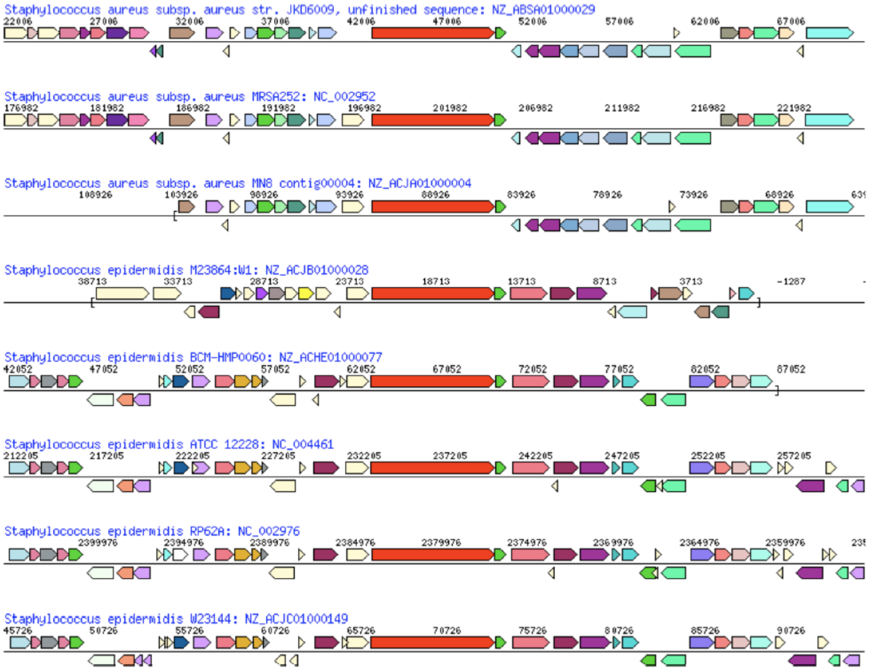

Staphylococcus aureus subsp. aureus str. JKD6009, unfinished sequence: NZ_ABSA01000029
22006
27006
32006
37006
42006
47006
52006
57006
62006
67006
Staphylococcus aureus subsp. aureus MRSA252: NC_002952
176982
181982
186982
191982
196982
201982
206982
211982
216982
221982
Staphylococcus aureus subsp. aureus MN8 contig00004: NZ_ACJA01000004
108926
103926
98926
93926
88926
83926
78926
73926
68926
Staphylococcus epidermidis M23864:W1: NZ_ACJB01000028
38713
33713
28713
23713
18713
13713
8713
3713
-1287
Staphylococcus epidermidis BCM-HMP0060: NZ_ACHE01000077
42052
47052
52052
57052
62052
67052
72052
77052
82052
87052
Staphylococcus epidermidis ATCC 12228: NC_004461
212205
217205
222205
227205
232205
237205
242205
247205
252205
257205
Staphylococcus epidermidis RP62A: NC_002976
2399976
2394976
2389976
2384976
2379976
2374976
2369976
2364976
2359976
Staphylococcus epidermidis W23144: NZ_ACJC01000149
45726
50726
55726
60726
65726
70726
75726
80726
85726
90726

Figure S2.1. Gene neighbourhood sequence alignment of the Staphylococcus nonribosomal peptide synthetase gene cluster.

The red gene is the 7.17 kb NRPS and the gene immediately adjacent in green is a phosphopantetheinyl transferase. Gene neighbourhood alignments were gathered using the IMG-JGI database and the Gene Ortholog Neighborhood Finder program (S1).

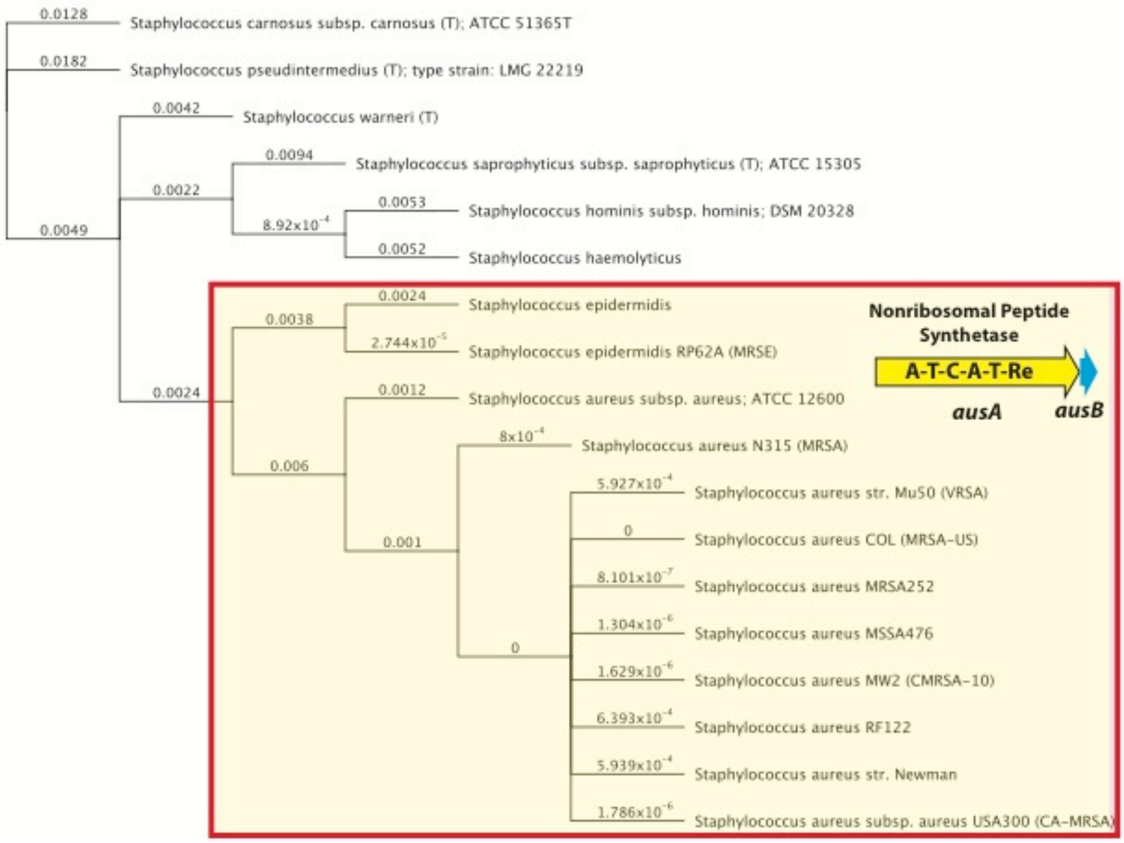

0.0128
Staphylococcus carnosus subsp. carnosus (T); ATCC 51365T
0.0182
Staphylococcus pseudintermedius (T); type strain: LMG 22219
0.0042
Staphylococcus warneri (T)
0.0094
Staphylococcus saprophyticus subsp. saprophyticus (T); ATCC 15305
0.0022
0.0053
Staphylococcus hominis subsp. hominis; DSM 20328
8.92x10^-4
0.0049
0.0052
Staphylococcus haemolyticus
0.0024
Staphylococcus epidermidis
0.0038
2.744x10^-5
Staphylococcus epidermidis RP62A (MRSE)
Nonribosomal Peptide Synthetase
A-T-C-A-T-Re
ausA
ausB
0.0024
0.0012
Staphylococcus aureus subsp. aureus; ATCC 12600
8x10^-4
Staphylococcus aureus N315 (MRSA)
0.006
5.927x10^-4
Staphylococcus aureus str. Mu50 (VRSA)
0.001
0
Staphylococcus aureus COL (MRSA-US)
8.101x10^-7
Staphylococcus aureus MRSA252
1.304x10^-6
Staphylococcus aureus MSSA476
0
1.629x10^-6
Staphylococcus aureus MW2 (CMRSA-10)
6.393x10^-4
Staphylococcus aureus RF122
5.939x10^-4
Staphylococcus aureus str. Newman
1.786x10^-6
Staphylococcus aureus subsp. aureus USA300 (CA-MRSA)

Figure S2.2. Phylogenetic analysis of aureusimines biosynthetic cluster in staphylococcus.

ausA and *ausB* are found in three staphylococcal species: *S. capitis, S. epidermidis* and *S. aureus*.

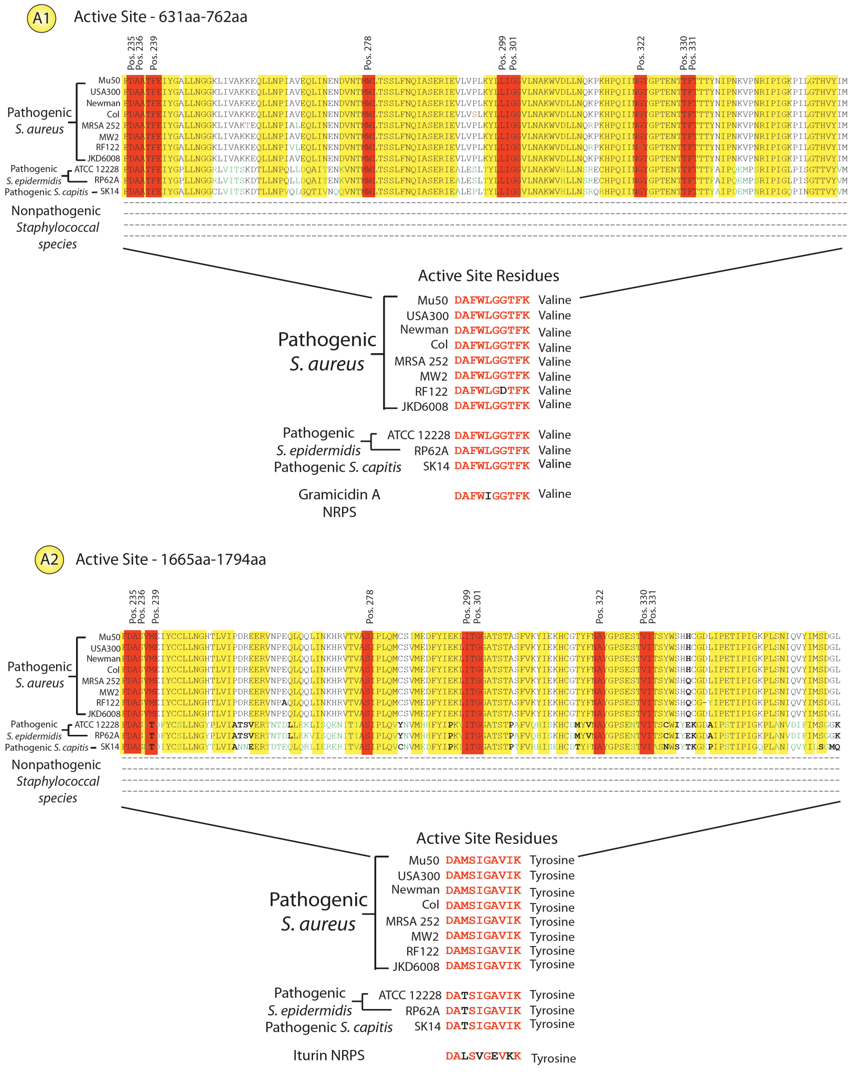


Figure S2.3. Alignment of the adenylation domains and adenylation codes found within staphylococcal NRPS.

Pathogenic strains of *S. aureus* adjacent in green is a phosphopantetheinyl transferase. Gene neighbourhood alignments were gathered using the IMG-JGI database and the Gene Ortholog Neighborhood Finder program. NRPS A domain code for the tyrosine-activating A domain of the iturin NRPS (*S2*) is shown in the lower panel; in the upper panel the gramicidin valine-activating A domain code (*S3*).

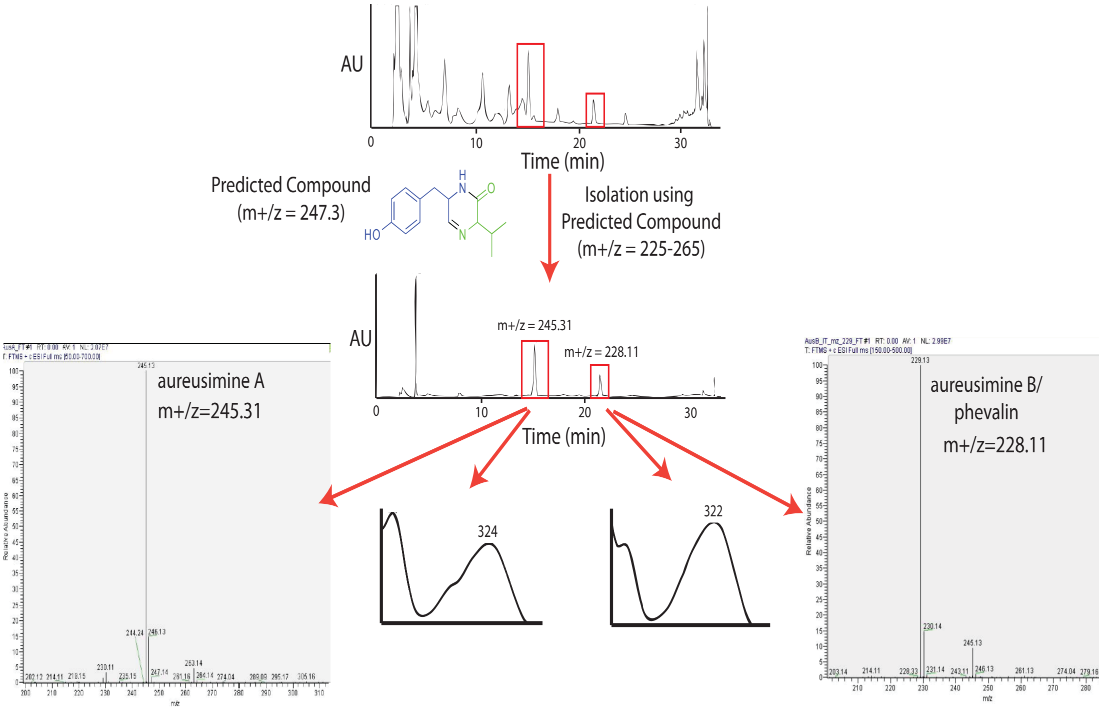


Figure S2.4. Chemical mining for cryptic *Staphylococcus aureus* nonribosomal peptides. At the top is the ethyl acetate crude organic extract of *S. aureus* cultivated in Tryptic Soy Broth and subjected to LC/MS. In the middle part of the figure is the same *S. aureus* extract chromatograph following application of a mass spectral filter (using MassLynx, Waters) and absorbance (322 nm) of select compounds. In the bottom portion is the absorbance spectrum of the two identified chemicals, with the corresponding mass spectrum (FT/MS). Images of mass spectral profiles of the two chemicals (aureusimine A and aureusimine B/phevalin) are also shown in at a larger scale below (Supplementary Figure S4 Con't).

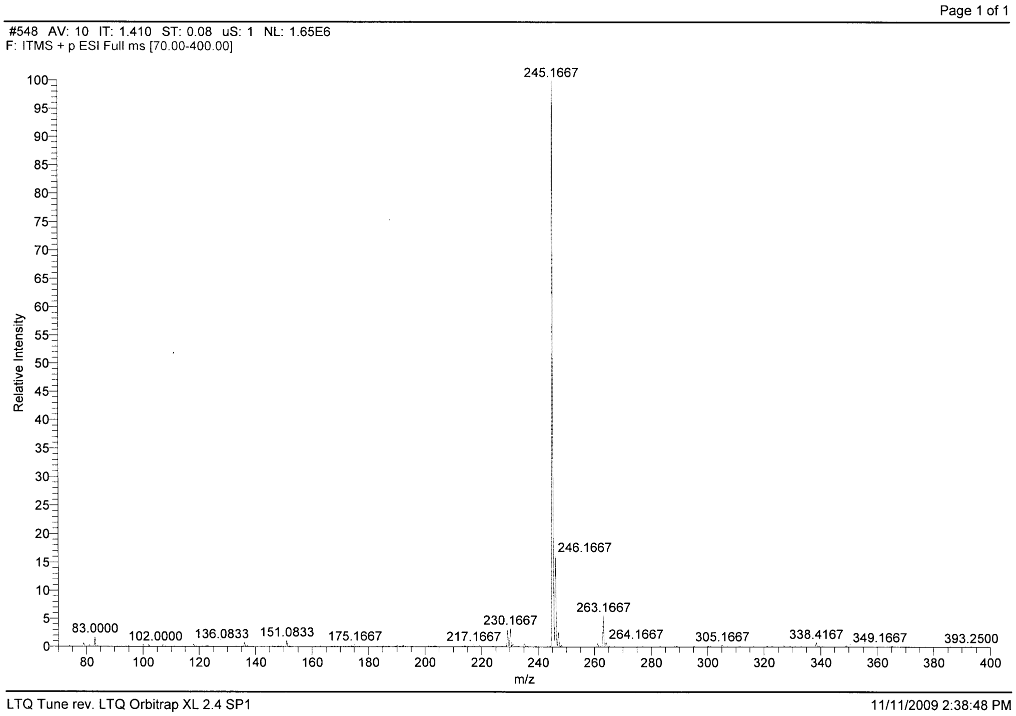


Figure S2.5. Chemical Mining of the Staphylococcus aureus nonribosomal peptides: Mass spectral analysis of aureusimine A.

High resolution mass was obtained on an LTQ OrbiTrap XL (Thermo Scientific).

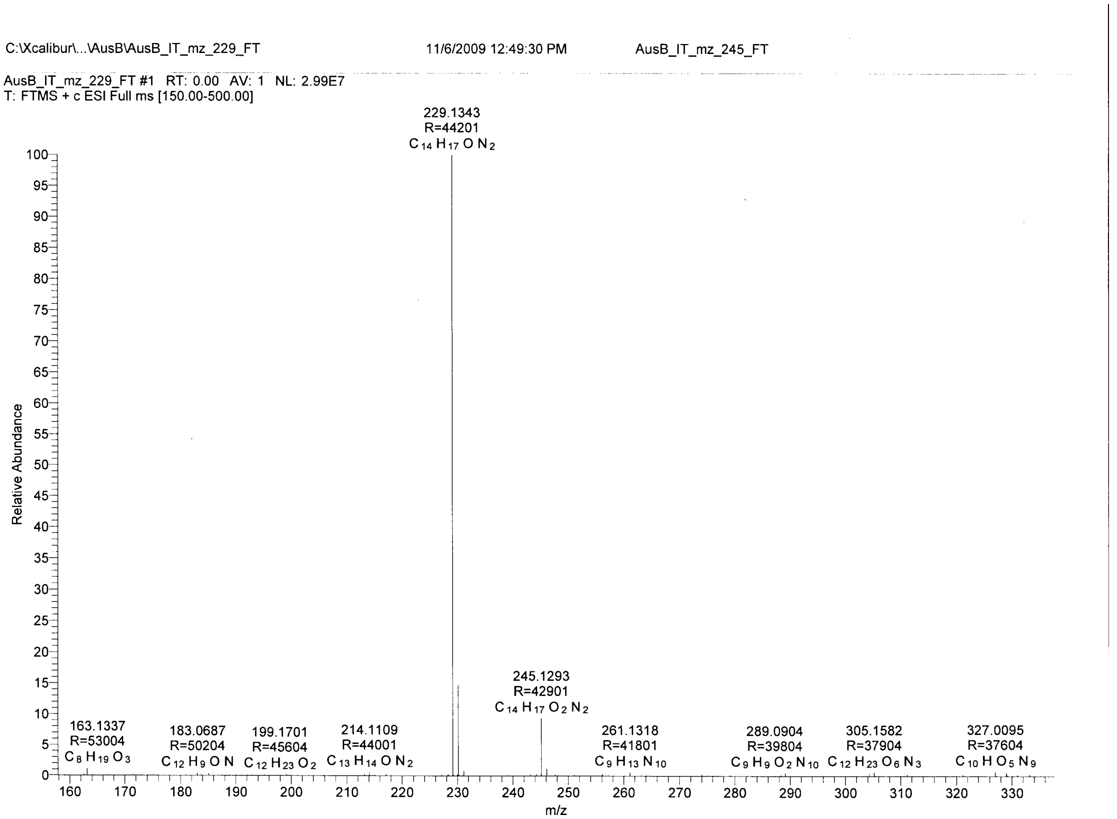


Figure S2.5 (Con't.). Chemical Mining of the Staphylococcus aureus nonribosomal peptides: Mass spectral analysis of aureusimine B. High resolution mass was obtained on an LTQ OrbiTrap XL (Thermo Scientific).

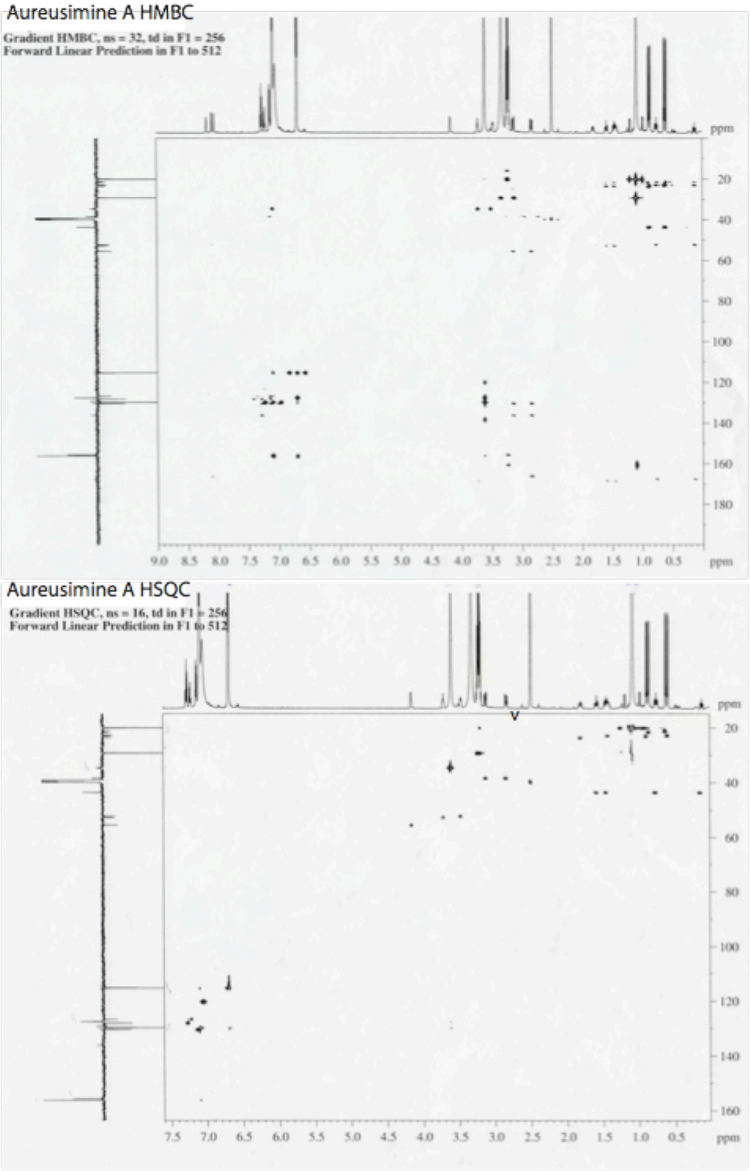


Figure S2.6. Two dimensional-NMR experiments of aureusimine A. In the top panel is the HMBC and in the bottom panel is the HSQC.

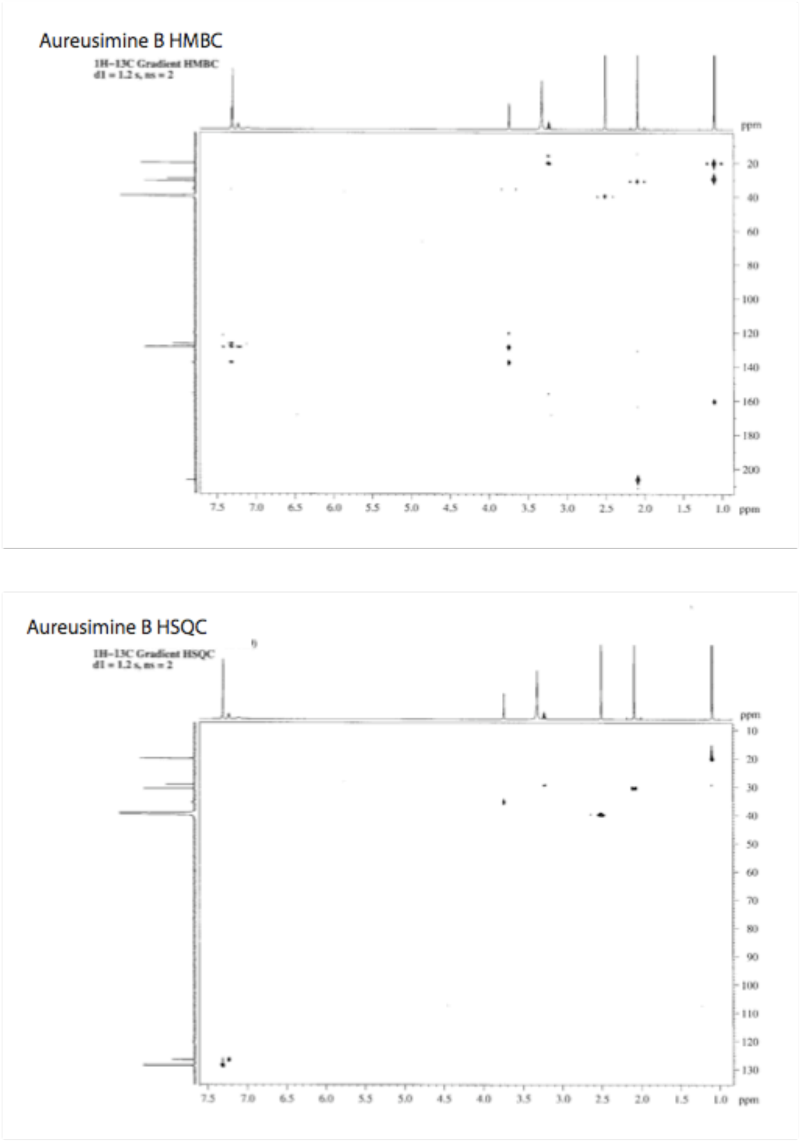


Figure S2.7. Two dimensional-NMR experiments of aureusimine B/phevalin.
In the top panel is the HMBC and in the bottom panel is the HSQC. Spectra is identical to that reported for phevalin (*S4, S5*).

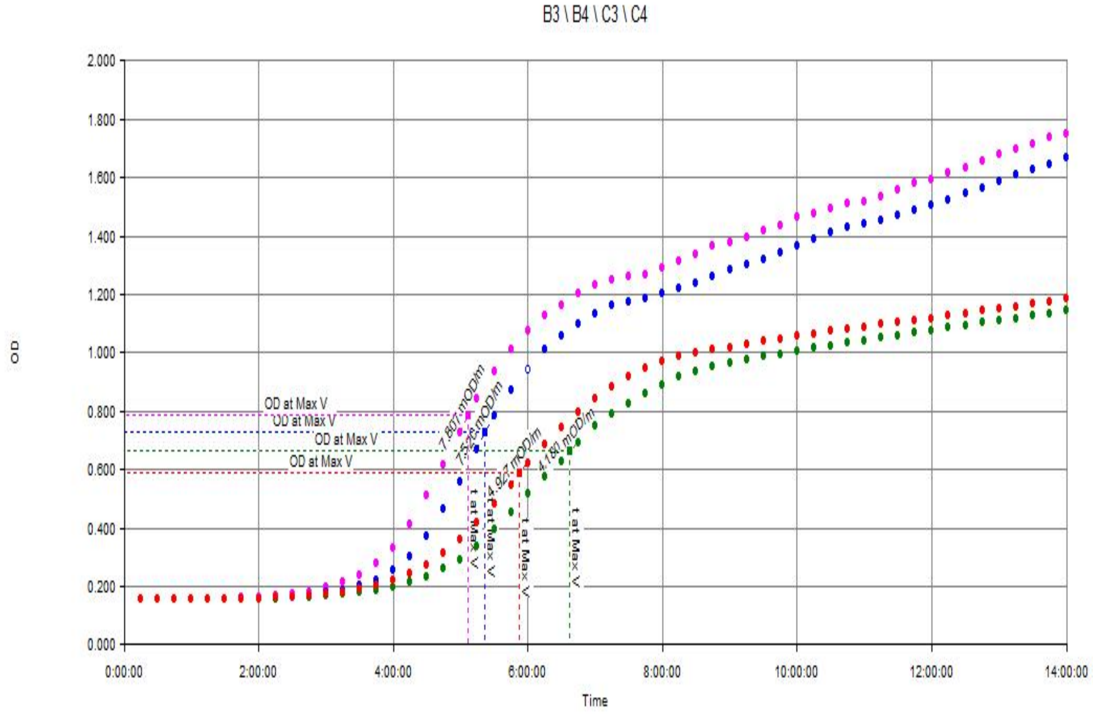


Figure S2.8. Growth curves of Newman WT and *NewmanΔausA*.

Growth curves of *S. aureus* Newman WT and *S. aureus* NewmanΔ*ausA* grown in TSB at 37°C. *Staphylococcus aureus* Newman WT (red and green) growth is delayed, has a lower Max V_{600}, and has a lower max $O.D._{600}$ compared to *S. aureus* NewmanΔ*ausA* (purple and blue).

2.9.2 Supplementary Tables

Microarray Data

Microarray data wass deposited at NCBI (accession: GSE21373).

Supplementary Table 2.1. Primers used for cloning and construction of recombinant *S. aureus* strains bearing lesions in *ausA*.

Primers	Sequence
A	5'-CTT AGA AGC AAA CTT AAG AGT G-3'
B	5'-GGG TCT AGA GTC TAG GGA CC-3'
C	5'-TTT TTC CAT GGA TAT ATT TGG TTA CCT CAA G-3'
D	5'-CAC TCT TAA GTT TGC TTC TAA GTA CAT CTT CGA TAA ATA GCA CA-3'
E	5'-AGA GGT CCC TAG ACT CTA GAC CCA ATA TCA CAA AGT CGT TGG TCA-3'
F	5'-TTT TCT GCA GAC ATT TTG TCT ATC GCT ATA-3'
G	5'-CGA CAG CAT TTA TAC CGT TTA TAG CA-3'
H	5'-GGA CAA GGT GCT GTT TTT GGT ATT-3'
I	5'- TAC TTT GGC GTG TTT CAT TGC TTG-3'
J	5'-GTA TTG TCT ATT TTT TAA TAG TTA TCT ATT-3'

*Restriction sites are underlined

Supplementary Table 2.2. Genes upregulated in early exponential phase by aureusimines. ***Please see correction (2.8.1 and 2.8.2)**

Fold Upregulated			
WT			
Early Exp	Late Exp	Common	Locus
6.7	3.9		BA000017 /// SA0468
6.6	2.5		BA000017 /// SA0468
35.7	16.0		BA000017 /// SA0478
32.3	18.7		BA000017 /// SA0478
28.9	8.1		BA000017 /// SA0479

58.9	8.4		BA000017 /// SA0857
55.0	59.1		BA000017 /// SA0859
50.0	48.3		BA000017 /// SA0859
3.7	22.6	ilvA	BA000017 /// SA1477
39.0	47.9	hlb	BA000017 /// SA2003
9.9	6.4	fnbB	BA000017 /// SA2509
75.2	51.3	fnbA	BA000017 /// SA2511
2.4	1.8		BA000017 /// SA2541
2.6	2.9		BA000018
17.4	1.6		BA000018 /// SA0472
3.6	2.5		SA0007
2.5	1.2		SA0024
2.4	1.1		SA0024
2.1	1.0		SA0024
2.3	3.1		SA0111
17.8	5.2		SA0199
14.9	4.0		SA0208
100.6	23.9		SA0209
3.8	3.3	lytS	SA0245
3.2	1.0	lrgA	SA0247
3.6	1.5	lrgB	SA0248
2.9	2.6	lytM	SA0263
8.1	5.3	geh	SA0309
16.1	1.2	set8	SA0384
17.7	6.7	set11	SA0387
7.3	1.2	set12	SA0388
44.9	15.1		SA0442
6.2	1.6		SA0443
4.2	2.6		SA0468
18.3	1.3		SA0469
13.0	1.5		SA0473
8.8	1.4		SA0474
27.8	12.8		SA0478
27.8	7.8		SA0479
60.7	106.9		SA0480

9.2	8.7	saeS	SA0765
12.8	9.2	saeR	SA0766
82.8	70.5		SA0767
137.5	97.6		SA0768
77.3	12.4		SA0857
107.4	50.2	empBP	SA0858
44.7	40.1		SA0859
2.7	1.7		SA0882
2.0	1.6	pdhC	SA1104
150.9	127.4		SA1164
7.4	9.6	(1:2)	SA1165
7.3	11.2	(2:2)	SA1165
73.5	41.5		SA1166
187.5	172.7	fbp	SA1168
158.7	146.4		SA1169
11.9	2.7		SA1170
4.2	2.2		SA1170
50.0	3.2	hla	SA1173
12.9	2.0		SA1178
22.0	2.1		SA1179
29.9	1.9		SA1180
2.1	0.9	pyrB	SA1212
2.1	1.3	guaC	SA1371
2.6	26.2		SA1476
3.9	19.3	ilvA	SA1477
4.6	24.6	ald	SA1478
33.9	25.8		SA1754
4.4	1.5		SA1870
113.6	100.3	map	SA2002
96.7	89.1	map	SA2002
89.6	71.0	map (1:2)	SA2002
89.6	77.8	map (2:2)	SA2002
8.7	5.6		SA2004
19.1	11.1		SA2006
8.5	5.4		SA2006
2.2	3.8		SA2291
3.0	2.3		SA2293
2.6	1.5		SA2295
11.6	9.0	tcaR	SA2353
10.7	8.9	tcaR	SA2353
72.7	35.2		SA2418
20.0	0.9	hlgA	SA2419

145.9	153.6	chp	SAR2036
7.0	0.9		SAS0388
7.9	1.0		SAS0389
6.1	6.2		SAV1941

Supplementary Table 2.3. Genes downregulated in early exponential phase by aureusimines. ***Please see correction (2.8.1 and 2.8.2)**

Fold Upregulated NRP			
Early Exp	Late Exp	Common	Locus
2.5	1.6	pls	SA0050
2.5	1.6		SA0076
4.0	4.2		SA0085
4.8	2.1		SA0086
5.7	4.7		SA0089
2.1	1.7		SA0092
2.3	4.9	deoD	SA0121
2.6	1.8	tet38	SA0122
3.2	9.5	cap5A	SA0136
2.1	4.7	cap5D	SA0139
2.0	4.3	cap5E	SA0140
2.2	3.8	cap5I	SA0144
2.2	2.6	cap5K	SA0146
2.3	3.4	cap5M	SA0148
2.1	2.8	cap5N	SA0149
3.4	1.4		SA0165
2.2	1.8	rocD	SA0170
2.7	1.6	pflA	SA0205
2.7	3.5		SA0299
2.8	5.3		SA0300
2.3	2.9		SA0507
4.1	3.3		SA0507
2.6	1.5		SA0736
2.6	2.1	clfA	SA0742
2.7	1.7		SA0751
3.1	2.5		SA0764
2.1	1.4		SA0832
2.8	1.7		SA0881
3.4	1.2	clpB	SA0979
3.9	3.3	atl	SA1062
4.9	5.0	atl	SA1062

2.1	1.7	arcC	SA1182
2.9	1.9		SA1252
2.4	1.5		SA1452
2.3	1.8		SA1453
4.0	3.3	epiG	SA1871
4.7	2.9	epiG	SA1871
5.4	3.1	epiG	SA1871
4.8	5.0	epiE	SA1872
4.5	3.6	epiE	SA1872
3.5	5.4	epiE	SA1872
3.5	5.2	epiF	SA1873
3.5	4.5	epiF	SA1873
2.2	5.9		SA1882
2.1	1.3		SA1883
2.6	2.5		SA2007
2.1	0.6	leuD	SA2049
2.1	0.9		SA2132
2.4	4.1		SA2147
3.4	3.2		SA2148
3.5	2.8		SA2149
2.9	0.6	aldC	SA2198
2.4	0.6	budB	SA2199
2.5	1.9	sarV	SA2258
2.0	2.7	ureB	SA2281
2.1	1.9	ureC	SA2282
2.6	2.1		SA2303
2.3	2.3	hutI	SA2323
2.1	2.0		SA2384
4.6	4.3		SA2386
2.5	3.4		SA2389
2.1	3.8		SA2391
4.8	4.5	narJ	SA2393
3.9	4.9	narH	SA2394
6.6	4.1		SA2396
5.0	4.1	nirD	SA2397
4.3	4.1	nirB	SA2398
2.1	1.7		SA2445
3.1	2.1		SA2481
3.3	3.1	fabG	SA2482
2.6	1.9		SA2484
2.8	2.3	sdhB	SA2545
2.8	2.4	crtN	SA2576
2.3	2.5	crtM	SA2577
3.1	3.0		SA2578

2.1	1.4		SA2596
3.6	2.9		SA2621
2.4	3.8		SA2625
2.3	1.6		SA2669
2.5	1.6		SA2670
2.2	2.4		SA2673
2.6	0.5	hisI	SA2696
2.2	1.7		SAS1634

Supplementary Table 2.4. Genes upregulated in late exponential phase by aureusimines.

***Please see correction (2.8.1 and 2.8.2)**

Fold Upregulated WT			
Early Exp	**Late Exp**	Common	Locus
32.3	18.7		BA000017 /// SA0478
35.7	16.0		BA000017 /// SA0478
28.9	8.1		BA000017 /// SA0479
58.9	8.4		BA000017 /// SA0857
55.0	59.1		BA000017 /// SA0859
50.0	48.3		BA000017 /// SA0859
3.7	22.6	ilvA	BA000017 /// SA1477
39.0	47.9	hlb	BA000017 /// SA2003
9.9	6.4	fnbB	BA000017 /// SA2509
75.2	51.3	fnbA	BA000017 /// SA2511
2.3	3.1		SA0111
77.3	255.6		SA0164
13.5	16.0		SA0164
1.4	3.6		SA0191
17.8	5.2		SA0199
14.9	4.0		SA0208
100.6	23.9		SA0209

2.0	6.4	lytR	SA0246
1.2	2.0		SA0261
8.1	5.3	geh	SA0309
17.7	6.7	set11	SA0387
1.5	5.5		SA0417
0.5	2.5	metE	SA0428
44.9	15.1		SA0442
0.9	2.4		SA0450
27.8	12.8		SA0478
27.8	7.8		SA0479
60.7	106.9		SA0480
0.8	2.4		SA0495
1.3	3.0		SA0511
0.9	3.2		SA0512
1.1	2.6	gltB	SA0514
0.7	2.0	mnhF	SA0685
0.6	2.8	rbf	SA0725
9.2	8.7	saeS	SA0765
12.8	9.2	saeR	SA0766
82.8	70.5		SA0767
137.5	97.6		SA0768
0.9	2.6		SA0772
1.5	5.1		SA0781
2.1	3.7	opuBB	SA0783
1.7	2.6	hisC	SA0784
1.0	3.1		SA0788
77.3	12.4		SA0857
107.4	50.2	empBP	SA0858
44.7	40.1		SA0859
2.1	6.6	nuc	SA0860
0.7	2.0		SA0922
0.9	2.8		SA0923
1.0	3.4		SA1115
1.0	2.1		SA1135
150.9	127.4		SA1164
7.3	11.2	(2:2)	SA1165
7.4	9.6	(1:2)	SA1165
73.5	41.5		SA1166
187.5	172.7	fbp	SA1168
158.7	146.4		SA1169
11.9	2.7		SA1170
0.8	2.5		SA1296
0.8	3.4		SA1297
1.4	2.4	hom	SA1362

0.8	2.9	trpE	SA1403
0.6	3.1	trpG	SA1404
0.8	2.7	trpC	SA1406
0.5	3.3	trpF	SA1407
0.8	2.3	trpB	SA1408
1.2	3.7		SA1419
0.9	5.5	lysC	SA1428
0.7	5.2	asd	SA1429
0.7	5.3	dapA	SA1430
0.6	6.0	dapB	SA1431
0.8	5.1	dapD	SA1432
0.8	6.3		SA1433
1.3	4.1		SA1471
1.6	16.6		SA1475
2.6	26.2		SA1476
2.7	17.9		SA1476
3.9	19.3	ilvA	SA1477
4.6	24.6	ald	SA1478
1.0	2.2	hisS	SA1686
1.0	4.1		SA1687
33.9	25.8		SA1754
1.7	2.7		SA1818
1.2	2.1		SA1840
0.8	2.6		SA1841
1.3	3.0		SA1892
0.5	3.3		SA1994
0.3	3.7		SA1995
0.6	3.8		SA1996
0.7	3.2		SA1997
113.6	100.3	map	SA2002
96.7	89.1	map	SA2002
89.6	77.8	map (2:2)	SA2002
89.6	71.0	map (1:2)	SA2002
8.7	5.6		SA2004
19.1	11.1		SA2006
8.5	5.4		SA2006
0.6	2.6	groEL	SA2016
1.0	2.3	ilvD	SA2042
0.7	2.4	leuA	SA2046
0.5	2.5	leuB	SA2047
0.9	2.9	czrA	SA2137
1.1	5.6		SA2164

2.2	3.8		SA2291
11.6	9.0	tcaR	SA2353
10.7	8.9	tcaR	SA2353
72.7	35.2		SA2418
1.4	2.7		SA2533
1.3	2.3	frp	SA2534
1.4	3.4		SA2563
0.8	2.9		SA2585
0.7	2.1		SA2701
0.6	2.3	hisD	SA2702
0.8	2.0	hisG	SA2703
1.6	2.8		SA2713
145.9	153.6	chp	SAR2036
1.1	2.6		SAS1559
6.1	6.2		SAV1941

Supplementary Table 2.5. Genes downregulated in late exponential phase by aureusimines. ***Please see correction (2.8.1 and 2.8.2)**

Fold Upregulated NRP			
Early Exp	**Late Exp**	Common	Locus
1.4	3.4		SA0082
4.0	4.2		SA0085
5.7	4.7		SA0089
1.6	3.2		SA0119
0.9	2.4		SA0119
2.3	4.9	deoD	SA0121
4.0	8.8	cap5B	SA0137
1.7	6.3	cap5C	SA0138
2.1	4.7	cap5D	SA0139
2.0	4.3	cap5E	SA0140
3.2	4.9	cap5F	SA0141
1.9	4.0	cap5G	SA0142
1.6	4.0	cap5H	SA0143
2.2	3.8	cap5I	SA0144
1.7	2.6	cap5J	SA0145
2.2	2.6	cap5K	SA0146
1.5	2.4	cap5L	SA0147
2.3	3.4	cap5M	SA0148
2.1	2.8	cap5N	SA0149
1.4	2.2	aldA	SA0154

1.2	8.0		SA0176
1.3	6.7		SA0177
1.5	4.2		SA0178
1.2	5.6		SA0178
1.5	4.0		SA0179
1.3	5.8		SA0193
1.4	4.0		SA0194
1.8	4.1		SA0195
1.8	2.2		SA0198
2.8	32.1		SA0200
2.4	33.3		SA0200
2.5	2.0		SA0212
1.7	3.9		SA0250
1.2	3.2	bglA	SA0251
1.4	3.2		SA0271
2.1	3.1		SA0272
1.5	2.2		SA0273
1.8	2.9	yukA	SA0276
1.4	3.5		SA0278
1.4	2.5		SA0280
1.3	2.2		SA0289
2.7	3.5		SA0299
2.8	5.3		SA0300
1.2	3.8	nanA	SA0312
1.2	2.9		SA0313
1.7	2.6		SA0399
3.5	10.9		SA0400
3.2	11.9		SA0401
3.1	20.5		SA0402
2.8	15.0		SA0403
1.8	2.1		SA0414
1.9	2.4		SA0486
1.5	3.9		SA0517
1.4	2.4	sdrD	SA0520
1.7	3.8		SA0599
3.1	2.5		SA0671
0.8	3.5		SA0688
2.4	4.5		SA0707
2.0	4.4		SA0708
2.5	3.4		SA0709
2.6	2.2		SA0763
3.1	2.5		SA0764
2.2	4.1		SA0850
2.0	9.0		SA0851

2.8	3.0		SA0854
3.1	3.9		SA0855
1.3	5.8	sspA	SA0901
1.2	2.3	gluD	SA0961
1.2	4.2	sspC	SA1055
1.5	4.9	sspB	SA1056
4.9	5.0	atl	SA1062
3.9	3.3	atl	SA1062
0.9	2.0		SA1111
0.9	2.2	pyrF	SA1216
2.5	2.1		SA1252
1.9	3.3	glpF	SA1319
1.4	2.5	sucB	SA1448
1.2	2.8	sucA	SA1449
1.6	6.6	malA	SA1551
1.3	3.3	ald	SA1758
17.0	4.0	sak	SA1758
1.3	3.3	acs	SA1783
1.1	2.8		SA1785
2.6	6.9	putA	SA1816
1.4	4.4	pckA	SA1838
1.2	3.3		SA1846
1.5	4.2		SA1847
1.5	3.7		SA1850
4.0	3.3	epiG	SA1871
5.4	3.1	epiG	SA1871
4.5	3.6	epiE	SA1872
3.5	5.4	epiE	SA1872
4.8	5.0	epiE	SA1872
3.5	5.2	epiF	SA1873
3.5	4.5	epiF	SA1873
2.6	2.6		SA2007
2.6	2.5		SA2007
2.2	2.5		SA2007
3.5	4.9		SA2013
1.7	2.0	kdpE	SA2071
2.6	2.1	thiE	SA2083
1.7	4.7		SA2146
2.4	4.1		SA2147
3.4	3.2		SA2148
3.5	2.8		SA2149
2.3	13.6		SA2197
1.5	2.1	ureA	SA2280
2.0	2.7	ureB	SA2281

2.6	2.1		SA2303
1.5	4.1		SA2316
2.3	2.3	hutI	SA2323
2.0	3.1	hutU	SA2324
3.5	4.1		SA2374
1.2	3.8		SA2376
1.4	8.2		SA2376
2.1	3.8		SA2391
2.3	5.2	narI	SA2392
4.8	4.5	narJ	SA2393
3.9	4.9	narH	SA2394
5.0	4.1	nirD	SA2397
4.3	4.1	nirB	SA2398
2.5	4.9		SA2441
1.8	4.5		SA2441
1.6	2.9		SA2443
3.6	10.2	lip	SA2463
0.6	3.0	drp35	SA2480
3.1	2.1		SA2481
3.3	3.1	fabG	SA2482
2.1	2.0	(2:2)	SA2505
2.1	2.6		SA2505
1.4	5.4	gntP	SA2514
0.9	7.1	gntK	SA2515
1.4	6.7		SA2521
2.8	2.3	sdhB	SA2545
2.3	2.3		SA2546
1.1	4.3		SA2552
1.3	2.1	cidB	SA2554
1.8	2.7		SA2568
1.8	4.0		SA2569
3.1	3.0		SA2578
1.9	3.7		SA2579
3.6	2.9		SA2621
2.4	3.8		SA2625
1.6	3.1	aur	SA2659
1.9	2.7	isaB	SA2660
1.2	4.2		SA2662
1.5	5.8		SA2663
1.8	4.4	manA	SA2664
2.4	3.3		SA2671
2.4	2.6		SA2717
1.8	4.0		SAV0217
0.9	2.2		SAV0902

1.4	2.1	SAV0907

2.9.3 Supplementary Notes (Supplementary References)

S1 V. M. Markowitz *et al.*, *Nucleic Acids Res* **38**, D382 (2010)

S2. E. H. Duitman *et al.*, *Proc Natl Acad Sci U S A* **96**, 13294 (1999).

S3. G. L. Challis, J. Ravel, C. A. Townsend, *Chem Biol* **7**, 211 (2000).

S4. M. E. Alvarez *et al.*, in *J Antibiot*, **48**, 1165-7, (1995).

S5. Y. Zeng, Q. Li, R. P. Hanzlik, J. Aube, *Bioorg Med Chem Lett* **15**, 3034 (2005).

S6. V. Vagner, E. Dervyn, S. D. Ehrlich, *Microbiol* **144 (Pt 11)**, 3097 (1998).

S7 W. S. Lin, T. Cunneen, C. Y. Lee, *J Bacteriol* **176**, 7005 (1994).

S8 J. C. Lee, *Methods Mol Biol* **47**, 209 (1995).

S9 K. E. Beenken, *J Bacteriol* **186** (14), 4665-84 (2004)

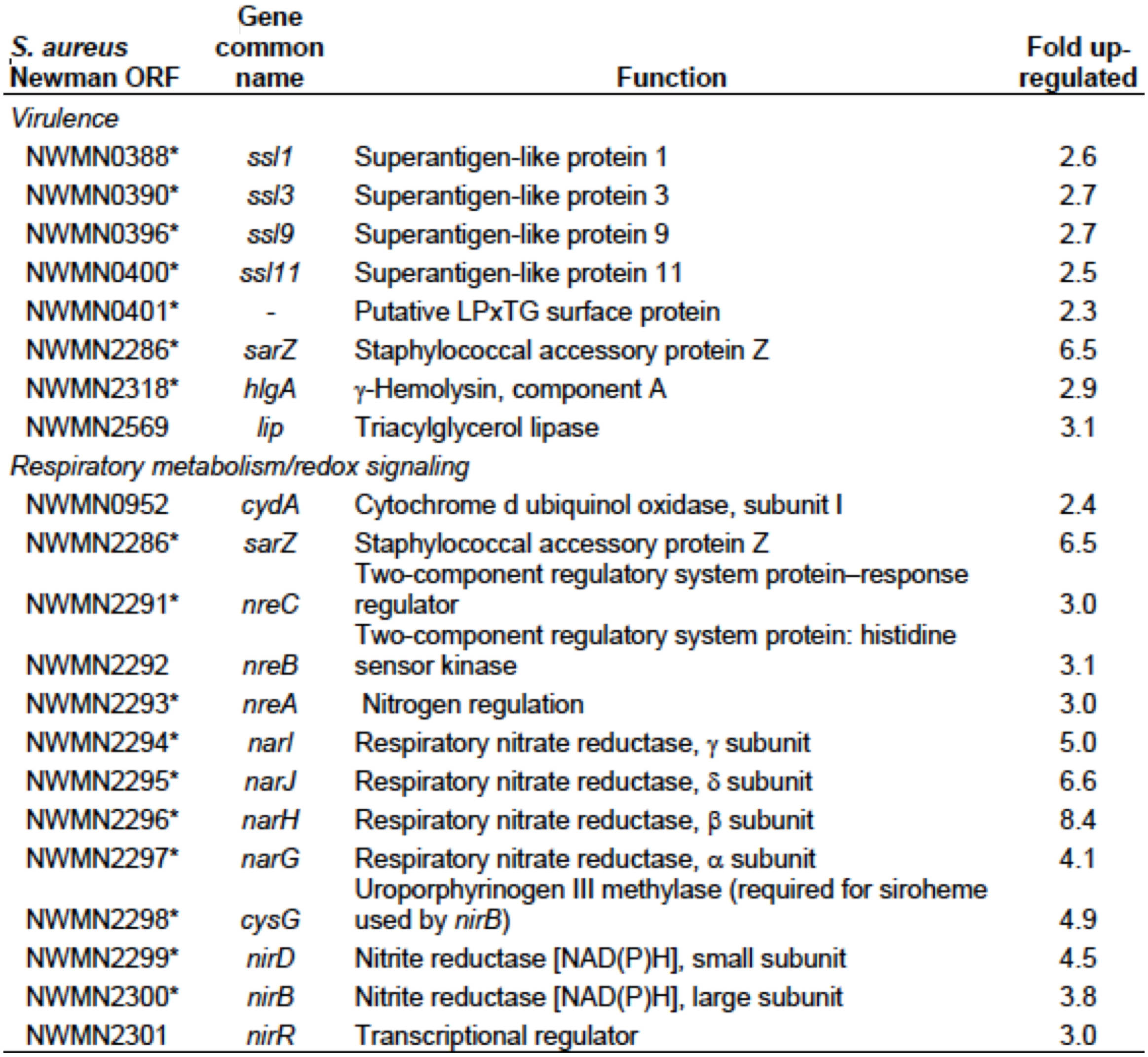
S10 M. T. Sebulsky, B. H. Shilton, C. D. Speziali, D. E. Heinrichs, *J Biol Chem* **278**, 49890 (2003).

2.10 Clarification

Research Article: "***Staphylococcus aureus*** **nonribosomal peptide secondary metabolites regulate virulence**," by Wyatt et al. (16 July 2010). During the construction of the *ausA* deletion strain, an inadvertent secondary site mutation in the *sae* two-component sensor kinase gene *saeS* occurred. The *sae* two-component system is a known regulator of virulence factor expression. Use of this double mutation strain was reported by F. Sun et al. [PLoS One 5, e15703 (2010)], and they concluded there was "no evidence indicating that the dipeptide aureusimines play a role in *sae*-mediated virulence factor production or contribute to staphylococcal virulence."

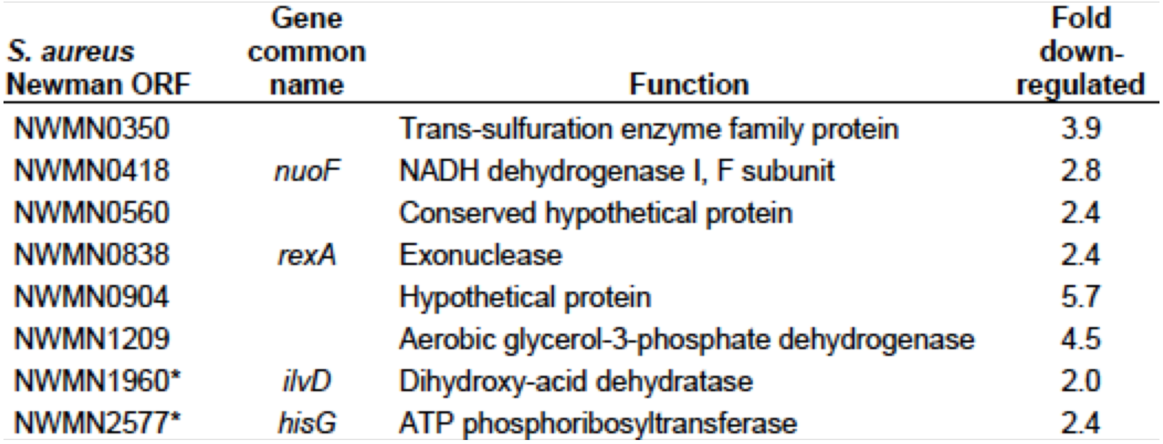
Use of this double-mutant led the authors to make an incorrect causal association between the aureusimine cyclic dipeptides with hemolysis on blood agar plates and *in vivo* virulence data that is attributed to the *sae*-dependent regulon (see microarray data in **Fig. 2.2**). Below are the results of a newly generated global microarray experiment for the *ausA* mutant. These data show the virulence/exotoxins, regulatory and redox-associated genes regulated by the aureusimine cyclic dipeptides. In the new microarray data, it is important to note that, of the hemolysin genes reported in **Fig. 2.2** (genes for α, β, γ hemolysins), only *hlgA*, the gene for γ hemolysin (hlgA) is regulated by the aureusimines. These new data provided below are the collection of virulence genes [including the superantigen-like (*ssl*) genes] and respiratory metabolic genes regulated by the aureusimines. From the microarray data provided below, it appears that aureusimines direct a metabolic switch that requires coordinate expression of genes encoding proteins associated with electron transfer processes (nitrite, nitrate reduction) and redox signaling. The data in **Tables 2.8.1** and **2.8.2** were obtained by the same methods used in the construction of the original microarray presented in the research article.

Table 2.8.1. Table of genes up-regulated by the aureusimines.
First column descriptors are based on the published sequence of *S. aureus* Newman. ORF, open reading frame. Asterisks indicate genes identified in the original microarray. Transcription data were generated similarly to those in the original article at late exponential phase growth.



S. aureus Newman ORF	Gene common name	Function	Fold up-regulated
Virulence			
NWMN0388*	*ssl1*	Superantigen-like protein 1	2.6
NWMN0390*	*ssl3*	Superantigen-like protein 3	2.7
NWMN0396*	*ssl9*	Superantigen-like protein 9	2.7
NWMN0400*	*ssl11*	Superantigen-like protein 11	2.5
NWMN0401*	-	Putative LPxTG surface protein	2.3
NWMN2286*	*sarZ*	Staphylococcal accessory protein Z	6.5
NWMN2318*	*hlgA*	γ-Hemolysin, component A	2.9
NWMN2569	*lip*	Triacylglycerol lipase	3.1
Respiratory metabolism/redox signaling			
NWMN0952	*cydA*	Cytochrome d ubiquinol oxidase, subunit I	2.4
NWMN2286*	*sarZ*	Staphylococcal accessory protein Z	6.5
NWMN2291*	*nreC*	Two-component regulatory system protein–response regulator	3.0
NWMN2292	*nreB*	Two-component regulatory system protein: histidine sensor kinase	3.1
NWMN2293*	*nreA*	Nitrogen regulation	3.0
NWMN2294*	*narI*	Respiratory nitrate reductase, γ subunit	5.0
NWMN2295*	*narJ*	Respiratory nitrate reductase, δ subunit	6.6
NWMN2296*	*narH*	Respiratory nitrate reductase, β subunit	8.4
NWMN2297*	*narG*	Respiratory nitrate reductase, α subunit	4.1
NWMN2298*	*cysG*	Uroporphyrinogen III methylase (required for siroheme used by *nirB*)	4.9
NWMN2299*	*nirD*	Nitrite reductase [NAD(P)H], small subunit	4.5
NWMN2300*	*nirB*	Nitrite reductase [NAD(P)H], large subunit	3.8
NWMN2301	*nirR*	Transcriptional regulator	3.0

Table 2.8.2 Table of genes down-regulated by the aureusimines.
First column descriptors are based on the published sequence of *S. aureus* Newman. Asterisks indicate genes identified in the original microarray. Transcription data were generated similarly to those in the original article at late exponential phase growth.



S. aureus Newman ORF	Gene common name	Function	Fold down-regulated
NWMN0350		Trans-sulfuration enzyme family protein	3.9
NWMN0418	*nuoF*	NADH dehydrogenase I, F subunit	2.8
NWMN0560		Conserved hypothetical protein	2.4
NWMN0838	*rexA*	Exonuclease	2.4
NWMN0904		Hypothetical protein	5.7
NWMN1209		Aerobic glycerol-3-phosphate dehydrogenase	4.5
NWMN1960*	*ilvD*	Dihydroxy-acid dehydratase	2.0
NWMN2577*	*hisG*	ATP phosphoribosyltransferase	2.4

2.11 References

1. F. D. Lowy, N. Engl. J. Med. 339, 520 (1998).

2. R. J. Gordon, F. D. Lowy, Clin. Infect. Dis. 46, (Suppl 5), S350 (2008).

3. T. J. Foster, Nat. Rev. Microbiol. 3, 948 (2005).

4. A. L. Cheung, A. S. Bayer, G. Zhang, H. Gresham, Y. Q. Xiong, FEMS Immunol. Med. Microbiol. 40, 1 (2004).

5. P. Recsei et al., Mol. Gen. Genet. 202, 58 (1986).

6. R. P. Novick, E. Geisinger, Annu. Rev. Genet. 42, 541 (2008).

7. N. Autret, C. Raynaud, I. Dubail, P. Berche, A. Charbit, Infect. Immun. 71, 4463 (2003).

8. J. Nakayama et al., Mol. Microbiol. 41, 145 (2001).

9. L. E. Hancock, M. Perego, J. Bacteriol. 186, 5629 (2004).

10. T. Fujii et al., J. Bacteriol. 190, 7655 (2008).

11. K. P. Scott, J. C. Martin, G. Campbell, C. D. Mayer, H. J. Flint, J. Bacteriol. 188, 4340 (2006).

12. C. Goerke et al., Infect. Immun. 68, 1304 (2000).

13. M. A. Fischbach, C. T. Walsh, Chem. Rev. 106, 3468 (2006).

14. G. Yim, H. H. Wang, J. Davies, Philos. Trans. R. Soc. London B Biol. Sci. 362, 1195 (2007).

15. M. F. Byford, J. E. Baldwin, C. Y. Shiau, C. J. Schofield, Chem. Rev. 97, 2631 (1997).

16. J. E. Becker, R. E. Moore, B. S. Moore, Gene 325, 35 (2004).

17. S. Lautru, R. J. Deeth, L. M. Bailey, G. L. Challis, Nat. Chem. Biol. 1, 265 (2005).

18. G. L. Challis, J. Med. Chem. 51, 2618 (2008).

19. Materials and methods are available as supporting material on Science Online.

20. K. Liolios, N. Tavernarakis, P. Hugenholtz, N. C. Kyrpides, Nucleic Acids Res. 34, (Database issue), D332 (2006).

21. T. Baba, T. Bae, O. Schneewind, F. Takeuchi, K. Hiramatsu, J. Bacteriol. 190, 300 (2008).

22. Y. Q. Zhang et al., Mol. Microbiol. 49, 1577 (2003).

23. F. Kopp, C. Mahlert, J. Grünewald, M. A. Marahiel, J. Am. Chem. Soc. 128, 16478 (2006).

24. M. E. Alvarez et al., J Antibiot 48, 1165 (1995).

25. C. J. C. de Haas et al., J. Exp. Med. 199, 687 (2004).

26. L. Piroth et al., Infect. Immun. 76, 3824 (2008).

27. R. Heying, J. van de Gevel, Y. A. Que, P. Moreillon, H. Beekhuizen, Thromb. Haemost. 97, 617 (2007).

28. P. Moreillon et al., Infect. Immun. 63, 4738 (1995).

29. A. Fleming, Br. J. Exp. Pathol. 10, 226 (1929).

Chapter 3. NRPS Flexibility

3.1 Chapter Preface

The simple architecture and relatively small size of the aureusimine biosynthetic gene cluster lends itself as an ideal model system to examine the structure, and flexibility of a simple NRP assembly-line system. Inherent promiscuity of NRPS assembly line A domains allow for many analogs of an NRP to be produced from a single assembly line and this has ramifications for the accurate identification of predicted NRPs in downstream methodologies and technologies. By understanding how this flexibility is controlled, the ability to predict and isolate cryptic compounds is increased. To provide insight into the flexibility of simple NRPS assembly lines, the entire aureusimine biosynthetic gene cluster was cloned into an *E. coli* heterologous host where the assembly-line was overexpressed allowing production of the aureusimines both *in vitro* using purified recombinant AusA and *in vivo* in growing *E. coli* cultures. This first publication of aureusimine biosynthesis used purified AusA to demonstrate the flexibility of both the AusA A domains and the terminal Re domain through *in vitro* feeding experiments, giving insight into how these assembly-line systems can create chemical diversity without any genetic modifications. In addition, I was able to overexpress and structurally characterize the Re domain of AusA, providing the first structure of an NRPS ring closing Re domain.

The following chapter is a modified version of a previously published article. I was the lead author on this work and contributed significantly to the conception, design, and interpretation of all experimental results. Other than the generation of the AusA Re protein crystal and subsequent structural assembly, performed by Mac Mok and Murray Junop, I generated and purified all AusA proteins and conducted all other experiments and analysis. The citation for this publication is as follows:

Wyatt, M. A., Mok, M. C. Y., Junop, M. and Magarvey, N. A. (2012), **Heterologous Expression and Structural Characterisation of a Pyrazinone Natural Product Assembly Line**. *ChemBioChem*, **13**: 2408–2415. doi: 10.1002/cbic.201200340

3.2 Abstract

Through a number of strategies nonribosomal peptide assembly lines give rise to a metabolic diversity not possible by ribosomal synthesis. One distinction within nonribosomal assembly is that products are elaborated on an enzyme-tethered substrate, and their release is enzyme catalysed. Reductive release by NAD(P)H-dependent catalysts is one observed nonribosomal termination and release strategy. Here we probed the selectivity of a terminal reductase domain by using a full-length heterologously expressed nonribosomal peptide synthetase for the dipeptide aureusimine and were able to generate 17 new analogues. Further, we generated an X-ray structure of aureusimine terminal reductase to gain insight into the structural details associated with this enzymatic domain.

3.3 Introduction

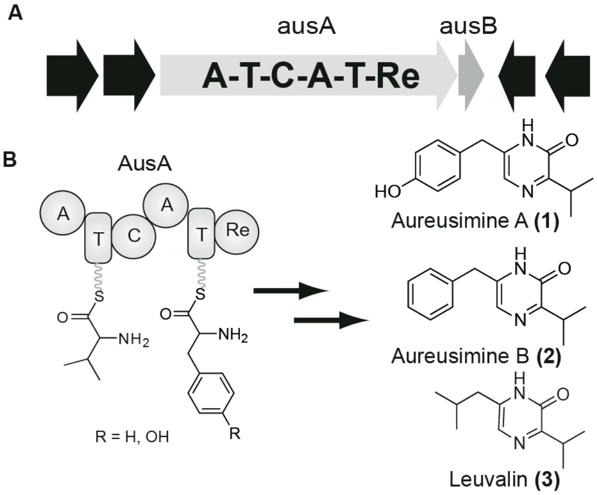
Nonribosomal peptide synthetases (NRPSs) are responsible for the biosynthesis of many pharmacologically significant natural products, including antibiotics (e.g., vancomycin, penicillin), anticancer compounds (e.g., bleomycin, epothilone), and immunosuppressants (e.g., cyclosporin). In contrast to ribosomal peptides, nonribosomal peptides are elaborated from an enzyme template through the action of catalytic domains or free-standing enzyme catalysts. Contained within the NRPS modules are the condensation (C), adenylation (A) and thiolation (T) domains followed by either a thioesterase (TE) or a reductase (Re).[1–3] The final TE and Re domains are responsible for terminating chain elongation of the majority of nonribosomal products arising from NRPS biosynthesis and are a critical step for final product formation.[3, 4]

The TE chain termination selectivity within the NRPSs of gramicidin (Grs), tyrocidine (Tyc) and surfactin (Srf) synthetases is known. In these cases, the SrfTE and TycTE were probed for selectivity determinants *in vitro* by using a heterologously expressed, excised TE.[5, 6] In these contexts, each amino acid position was substituted with alanine on a synthesised linear precursor, and it was found that the most important residues for ring closure and release were those at the proximal end and site of ring closure.[5] In these cases, it was possible to alter the macrocycle size by using TycTE to give six-, eight-, ten-, 12- and 14-membered rings.[6] These TE studies focused on the release and cyclisation of synthesised *N*-acetylcysteamine thioester (SNAC) peptides or peptides chemically linked to solid supports as a potential method to generate libraries by merging biosynthetic enzymes and combinatorial chemistry.[5, 7] In these cases, the in vitro capabilities were not matched with actual *in vivo* processing.

The Re domain is another termination strategy that can result in macrocyclisaton. Several outcomes for the peptide chain are known to result from Re reductive action, they include alcohols (myocbacterial glycopeptidolipids, myxochelin A, lyngbyatoxin), linear aldehydes, and cyclic imine-containing compounds (nostocyclopeptide, aureusimine A and B).[8–12] Studies of the nostocyclopeptide Re demonstrate that Re catalyses linear aldehyde formation and directs the self-assembly of imine macrocycle formation.[11] Similar to the TEs, investigations into the specificity of the C- and N-terminal amino acids show some flexibility within the Re domains, which therefore have utility as catalysts in creating diversity. For total biosynthesis, however, efforts at coupling *in vitro* and *in vivo* analysis would be of value in order to fully understand their biological flexibility.

Recently, we characterised the nonribosomal peptide aureusimine from *Staphylococcus aureus* str. Newman, and identified its corresponding nonribosomal peptide synthetase cluster.[12] This biosynthetic locus is present in pathogenic strains of *S. aureus, S. epidermidis, S. capitis*, and *S. lugdenensis*. The biosynthetic gene cluster consists of a phosphopantetheinyl transferase (PPtase) gene (*ausB*) and a dimodular NRPS (*ausA*) that comprises the following enzymatic domains: A-T-C-A-T-Re (Scheme 3.1). Each of the known products has a pyrazinone core, which is probably generated through nucleophilic attack by the terminal amine onto an aldehyde intermediate generated by the terminal Re. This would lead to imine bond formation and then a putative spontaneous air oxidation, probably driven by the force of aromatisation, to form the final valine–tyrosine (aureusimine A), valine–phenylalanine (aureusimine B), and valine–leucine (leuvalin) pyrazinone products.[12, 13]

The dimodular AusA NRPS alone encodes for the end products, and the rather straightforward architecture presents an opportunity to probe a full-length NRPS. Other end-product-encoding dimodular NRPS's have been interrogated, including those for the production of enniatin, PF1022, and beauvericin; however, they represent an iterative NRPS elongation mechanism without a termination catalyst.[14–16] AusA, offers a unique opportunity to reveal the promiscuity of an intact enzyme with a natively embedded Re domain, which itself has not been studied in depth. In this way, we can more easily probe the flexibility of this system, in particular the terminal Re domain. In this paper, we report investigations into the biosynthesis of the aureusimines that were conducted by heterologously expressing the entire aureusimine gene cluster behind the inducible T7 promoter in *Escherichia coli* to characterise its ability, both *in vitro* and *in vivo,* to produce new analogues through Re-mediated ring closure within the AusA dimodular system. In addition, we show the crystal structure of the AusA terminal Re domain.



Scheme 3.1 Aureusimine biosynthetic cluster and assembly line.
A) Aureusimine bimodular NRPS gene (*ausA*) and post-translational modifying enzyme, phosphopantetheneiyl transferase (*ausB*), in *S. aureus*. Black arrows indicate adjacent genes not required for aureusimine biosynthesis. B) AusA biosynthetic assembly line is a

bimodular NRPS. AusA contains NRPS adenylation (A), thiolation (T), condensation (C), and reductase (Re) domains that are indicated by shaded spheres. AusA has been shown to produce three natural products in *S. aureus*: aureusimine A (**1**), aureusimine B (**2**), and leuvalin (**3**).

3.4 Results and Discussion

3.4.1 Reconstitution of the aureusimine nonribosomal peptide assembly line

To probe the flexibility of an Re domain fused in its cognate modular association within an NRPS assembly line, we overexpressed full-length AusA and used precursor-directed biosynthesis.[17, 18] We heterologously expressed the entire aureusimine gene cluster, *ausA* and *ausB*, within *E. coli* and used that as a host to probe selectivity. Co-expression of AusB, the presumed cognate AusA PPTase, was part of the reconstitution, as NRP production demands NRPS phosphopantetheinylation. Because the two genes are immediately adjacent, and we sought to track whether both gene products were overexpressed, a vector construct was designed such that both the resulting AusA and AusB would bear N- and C-terminal histidine tags. An *ausA*/*ausB* amplicon was obtained by PCR and cloned within the pET28b expression vector at BamHI and NotI sites. The resulting construct, pAusAB, has both genes under isopropyl β-d-1-thiogalactopyranoside (IPTG) control and leads to an N-terminal His-tagged AusA and a C-terminally tagged AusB (Figure S3.1 A in the Supporting Information). Since the AusA and AusB fusions are sufficiently distinct in size (273 and 25 kDa, respectively), these products can be simultaneously traced by anti-histidine antibodies and SDS-PAGE (Figure S1 B). The histidine-tagged *holo*-AusA fusion protein was purified by using Ni-NTA (nickel-nitriloacetic acid) affinity chromatography from induced *E. coli* cultures. Purified AusA was then probed for its promiscuity and capability to elaborate dipeptide variants not naturally observed.

3.4.2 *In vitro* chemoenzymatic synthesis of aureusimine natural and unnatural products

With the established A domain specificity-conferring codes, the substrates for the two AusA adenylation domains are predicted to be valine and tyrosine.[12, 19] Based on the naturally occurring products, it can be inferred that the first A domain activates valine and the second activates tyrosine, phenylalanine or leucine. These variations correspond to the natural products aureusimine A, aureusimine B and leuvalin.[12, 13] Using purified AusA, we reconstituted aureusimine assembly *in vitro* and assessed the flexibility of pyrazinone formation through a chemoenzymatic approach.

The activity of AusA was first assessed by using its natural amino acid substrates, Val/Tyr, Val/Phe or Val/Leu, as described in the Experimental Section. Extraction of the *in vitro* AusA reaction products and LC-MS analysis confirmed the production of aureusimines A and B, as well as leuvalin, respectively, when pairs of amino acids were used (Figure 3.1 B). Comparison with authentic standards and a diagnostic neutral loss of 70 Da in the MS-MS data confirmed this production (Figure S3.2). This verified that that *holo*-AusA is all that is required for biosynthesis of the final products and that the oxidation that leads to the pyrazinone ring is not catalysed by adjacent enzymes surrounding the aureusimine cluster. Assays to reveal the flexibility of AusA included ones in which the amino acid pairs were scrambled with structural amino acid variants (Table S3.2). In these instances, the LC-MS assay was used to detect the projected pyrazinone analogues (Figure 3.1). Using pair-wise amino acid testing and substituting the valine with alanine, leucine or isoleucine, while keeping phenylalanine (second amino acid) constant, did not yield the corresponding pyrazinone product. This observed selectivity is in keeping with the high-homology of the valine-activating A domain of GrsB.[12] The second adenylation domain, however, had a lower degree of similarity to the specificity-conferring amino acid residues for tyrosine (70 %). In addition, aromatic amino acid activating A domains are known to be flexible.[19] Using aromatic amino acid analogues, we tested the flexibility of the Re domain to catalyse pyrazinone formation.

A series of aromatic amino acid analogues was chosen for *in vitro* studies; these included *p*-chloro-phenylalanine, *p*-fluoro-phenylalanine, *p*-bromo-phenylalanine, 2-chloro-tyrosine, 3-amino-cyclohexane propionic acid, *p*-(4-hydroxybenzoyl)phenylalanine, 4-methyl-phenylalanine, and l-4-methylamino-phenylalanine. *In vitro* biosynthesis afforded the aureusimine analogues **4**–**7**, and LC-MS-extracted ion chromatograms of those compounds are shown (Figure 3.1). AusA-dependent, Re-catalysed pyrazinone formation was confirmed through comparison with negative controls that were missing Re co-factors (no NAD(P)H added), predicted mass, and a diagnostic fragmentation of the pyrazinone core (Figure S3.2). The inability of AusA to use *p*-(4-hydroxybenzoyl)phenylalanine, 4-methyl-phenylalanine, and l-4-methylamino-phenylalanine indicated that the kinds of substitutions that can be made on the benzene ring of phenylalanine are restricted. This might, in part, be due to size restrictions for amino acids such as *p*-(4-hydroxybenzoyl)phenylalanine, and is supported by the fact that *p*-bromo-phenylalanine could not be incorporated, whereas *p*-chloro-phenylalanine and *p*-fluoro-phenylalanine could generate the pyrazinone product, probably due to smaller van der Waals radii. In addition to the pyrazinone analogues of the aureusimines, we also produced the cognate diketopiperazine analogues (**8**–**13**) of all the pyrazinone compounds that were produced, except for aureusimine A.

The diketopiperazine series of analogues is probably produced by an intramolecular attack of the free amine onto the thioester-activated carboxyl of the first amino acid. This type of nonenzymatic release from NRPSs has been observed previously in heterologously expressed dimodular NRPSs, such as TycA/TycB1 for the *in vitro* production of cyclo(d-Phe-Pro) diketopiperazines, and can be observed in the premature release of the dipeptide beauvericin from BbBEAS.[16, 20, 21] The absence of the Val-Tyr diketopiperazine within the *in vitro* studies suggests that the Re acts to release the Val-Tyr product before it can undergo an intramolecular attack to form the diketopiperazine side product. This indicates that there is some substrate preference observed by the AusA Re domain. Comparison of *S. aureus* str. Newman wild-type and its Δ*ausA* mutant show that cyclo(Val-Phe) (**9**), and cyclo(Val-Tyr) (not observed *in in vitro* reactions) are observed in the natural host and stem from AusA, rather than other cyclic dipeptide biosynthesis systems such as cyclodipeptide synthetases in albonoursin biosynthesis (Figure S3.3).[22] AusA-dependent production of the diketopiperazines was verified by comparison to a negative control by using predicted mass, diagnostic fragmentation and the presence of each precursor's immonium ion after MS fragmentation (Figure S3.4). In all cases in which no pyrazinone congener was formed, the corresponding diketopiperazine was also absent; this suggests that the flexibility of the system was limited by the second adenylation in these experiments. In addition, within each *in vitro* reaction, a series of pyrazine analogues was produced (compounds **14**–**20**). These are likely to be the result of a reductive release of aromatic amino acids from the second A domain to yield aldehyde amino acids that react with each other to form a pyrazine product according to a mechanism analogous to that by which the pyrazinone congeners are formed. Unlike the diketopiperazines, the pyrazines have not been identified in *S. aureus* culture. The activity of AusA *in vitro* led us to investigate the production of aureusimine and its analogues in growing *E. coli* cultures so as to determine the *in vivo* flexibility of the NRPS system.

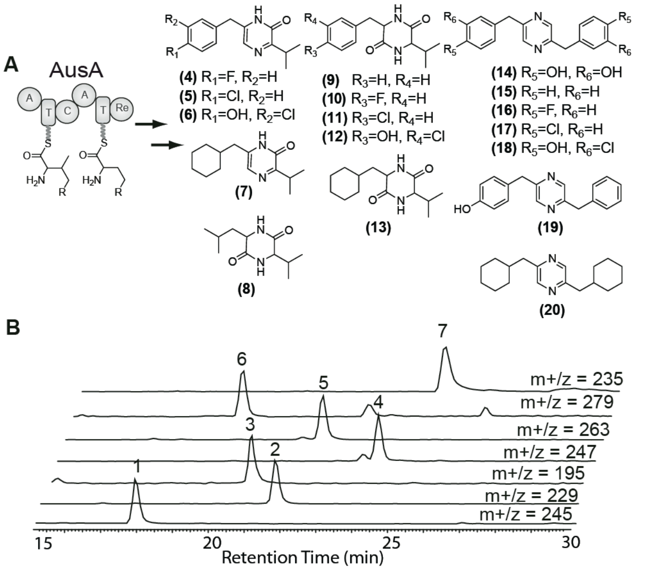


Figure 3.1 Reconstituted *in vitro* chemoenzymatic production of aureusimine and analogues with purified AusA.

A) Compounds identified in LC-MS analysis as products from AusA. B) LC-MS-extracted ion chromatograms of pyrazinone products biosynthesised by recombinant AusA.

3.4.3 Heterologous *in vivo* production of *S. aureus* cyclic dipeptides

E. coli bearing the pAusAB plasmid and an empty pET28b plasmid were induced with IPTG and grown overnight in lysogeny broth. Cultures were extracted with ethyl acetate and subjected to LC-MS analysis. Comparison with the control culture showed the production of aureusimines A (**1**) and B (**2**) as well as leuvalin (**3**) and the Tyr-Tyr (**16**), Phe-Phe (**17**) and Phe-Tyr (**21**) pyrazine series, in accordance with the *in vitro* study (Figure 3.2). Production of aureusimines A and B was 2.65 and 3.49 mg L^{-1}, respectively, which is comparable to *S. aureus* UAMS-1, whose titre in tryptic soy broth is 1.3 and 0.5 mg L^{-1} for aureusimines A and B, respectively.

After heterologously producing the aureusimines in a non-pathogenic Gram-negative organism, we probed the flexibility of AusA to make analogues *in vivo* through precursor-directed biosynthesis, similar to the *in vitro* studies. Growth of *E. coli*+pAusAB in LB supplemented with *p*-chloro-phenylalanine, *p*-fluoro-phenylalanine, 2-chloro-tyrosine and 3-amino-cyclohexane-propionic acid yielded compounds **4**–**7** (Figure 3.2). These molecules were isolated in trace amounts by utilising the characteristic UV absorption of aureusimine (320 nm). Products were confirmed by high-resolution mass spectrometry and MS-MS fragmentation (Figure S3.5). An in-depth analysis of all components was difficult due to the high metabolic background produced by *E. coli* and the LB medium. The diketopiperazine and pyrazine analogues were probably also produced; however, due to the high concentration of natural substrates available to AusA, they could not be identified in our experiments. The presence of the pyrazinone products, however, does verifies the flexibility of AusA *in vivo* to accept substituted aromatic rings.

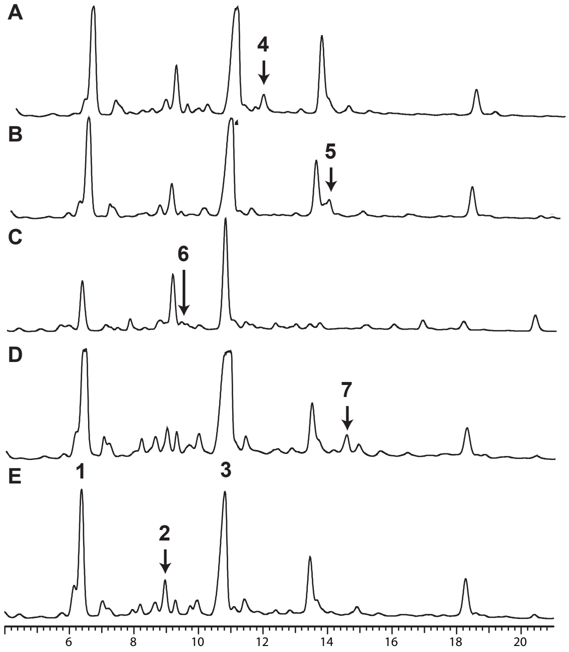


Figure 3.2

HPLC UV chromatograms (UV=320 nm) of *E. coli*+pAusAB extracts supplemented with A) 4-fluoro-phenylalanine, B) 4-chloro-phenylalanine, C) 3-chloro-tyrosine, D) (*S*)-(+)-α-amino-cyclohexane propionic acid, and E) LB alone. The pyrazinone analogues identified are indicated.

3.4.4. Structural characterisation of the AusA chain-terminating reductase

Although the structure of a terminal NRPS Re domain from *Mycobacteria tuberculosis* was recently published,[8] this represents an Re that performs two iterative $[2+2]e^{-}$ reductions to generate a terminal alcohol. The AusA Re domain catalyses only a single reduction to release a linear aldehyde, and its characterisation will give additional structural details for reductive release macrocyclisation catalysts.[8]

The Re of the AusA dimodular NRPS is predicted to be a short-chain dehydrogenase/reductase (SDR) based on sequence homology in BLAST searches. Alignment of the C terminus of AusA shows homology to other known C-terminal Re domains from NRPS biosynthetic clusters, such as those from glycopeptidolipid, saframycin, myxochelin A, peptaibol, gramicidin and nostocyclopeptide NRPS (Figure S3.6).[8, 11, 23–26] Within AusA, a conserved NAD(P)H binding domain is identified, and the conserved catalytic triad of Thr(/Ser)-Tyr-Lys common to SDR is present (Figure S3.6).[27, 28] The aureusimine C-terminal Re domain is believed to take part in a similar mechanism to nostocyclopeptide biosynthesis, in that the Re catalyses a one-step reduction to form a released aldehyde product that subsequently undergoes nucleophilic attack by a terminal amine to generate a cyclic imine product.[11] In the case of aureusimine biosynthesis, the α-amine of valine would attack the released electrophilic aldehyde to spontaneously form a cyclic imine, in a similar manner to nostocyclopeptide biosynthesis, followed by a putative spontaneous air oxidation.[11]

As several analogues are produced by AusA, we hypothesised that not only did we require A domain flexibility, but we also required a flexible terminal Re mechanism to accommodate the various substrates. To determine whether AusA Re was dependent on NADH or NADPH, we carried out an *in vitro* assay on purified AusA with either NADH or NADPH as coenzyme. Analysis of the *in vitro* products showed that AusA was functional and could produce aureusimine B by using either NADH or NADPH (Figure

S3.7).

The structure of AusA Re was determined by X-ray crystallography to 2.8 Å (Figure 3.3). AusA Re contains the highly conserved Rossman fold, which contains a central β-sheet surrounded by α-helices. Within the structure, the NAD(P)H binding site can be seen at position 2050–2057 of AusA, with the common extended SDR Gly-motif of TGATGFLG.[29] Although no electron density can be seen for NADPH, structural alignment of the T-Re didomain with both type I (*eryA1*; PDB ID: 2FR0) and type II (*actIII*) ketoreductase domains (PDB ID: 1W4Z) shows a similar Rossman-fold structure and a coenzyme binding site containing NADPH within the N terminus of the AusA Re structure (Figures 3.4 and S3.8).[29, 30] Structural comparison also shows that the catalytic triad of the AusA Re domain comprised Thr-Tyr-Lys at positions 2172 , 2203, and 2207, respectively, with threonine replacing the chemically equivalent amino acid, serine; this is in agreement with the structural alignment of the EryA1 and ActIII ketoreductase domains (Figure 3.4 A and B). This is also consistent with the recently published crystal structure of the active site of the *Mycobacterium tuberculosis* NRPS terminal Re domain (Mps2 R_{nrp}).[8] Sequence and structural alignment of the AusA Re domain with Mps2 R_{nrp} show considerable homology between the two termination catalysts, except for the presence of a loop (residues 2270–2275) in R_{nrp} that occludes the co-factor binding site, but is disordered and therefore absent in our structure (Figure 3.4 C and D, Figure S3.8).[8] In addition, comparison of AusA Re domain and R_{nrp} reveals a common C terminus that is proposed to be responsible for substrate recognition and binding; this is absent in the non-terminal reductases ActIII and EryA1 (Figure 3.4).[8]

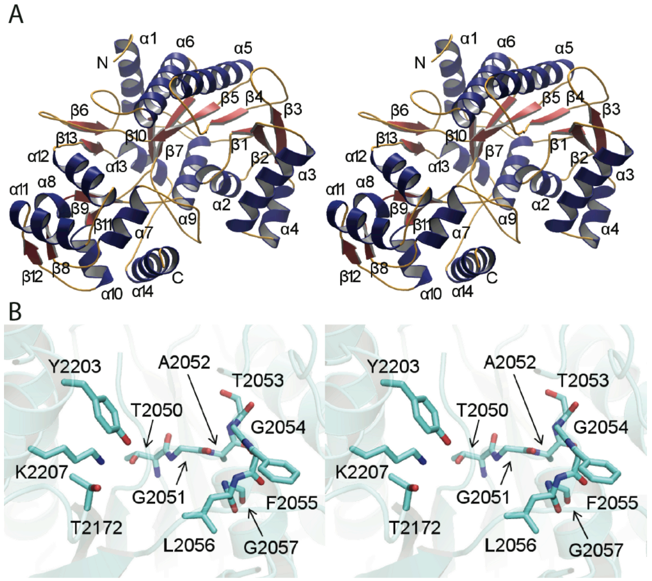


Figure 3.3. Stereo images of A) the AusA T-Re didomain and B) the Re domain active site.

Catalytic residues Y2203, K2207 and T2172, as well as potential NADPH binding residues T2050, G2051, A2052, T2053, G2054, F2055, L2056 and G2057 are labelled.

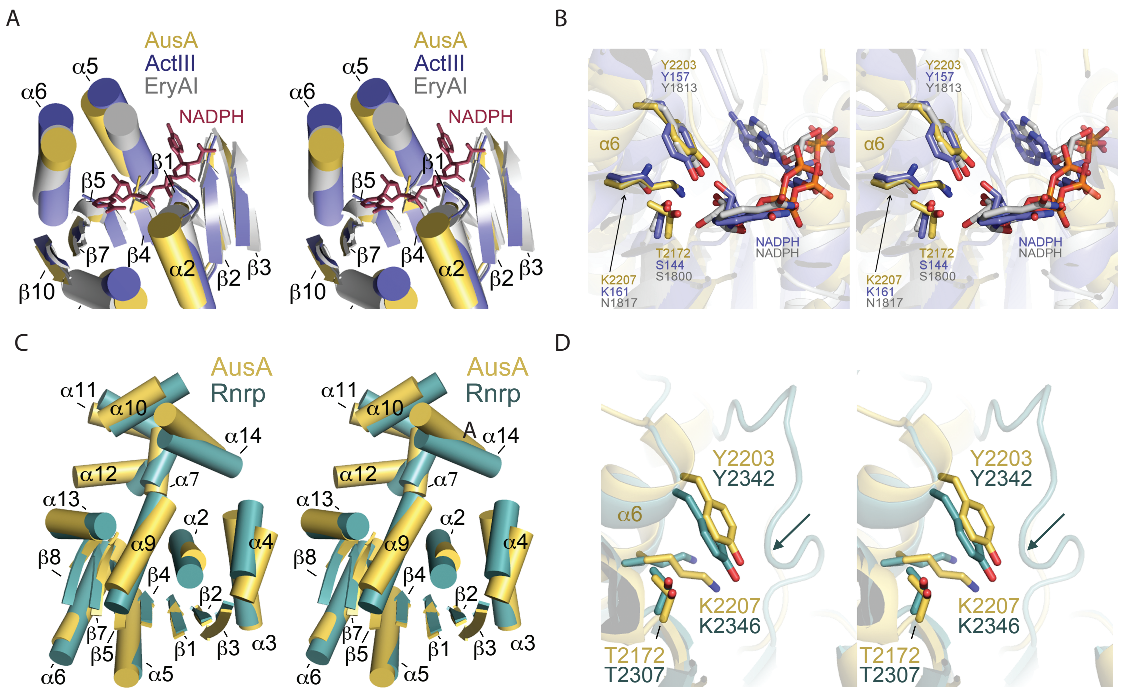


Figure 3.4. Structural comparison of AusA reductase domain.

A) and B) Comparison with ketoreductase ActIII (*Streptomyces coelicolor*) and EryAI (*Saccharopolyspora erythraea*) active sites. Structural comparison with C) entire R_{nrp} (*Mycobacteria tuberculosis*) domain and D) active site. Regions of AusA (yellow), ActIII (blue; PDB ID: 1W4Z), EryAI (grey; PDB ID: 2FR0) and R_{nrp} (green: PDB ID: 4DQV) are superimposed. Secondary structure elements of AusA are labelled. NADPH from the co-crystal structure with ActIII is shown in red. Arrow indicates loop of R_{nrp} (residues 2270–2275) that occlude the cofactor binding site, are disordered in the AusA Re domain and not observed in the structure.

3.4 Conclusions

We have demonstrated the successful heterologous expression of the aureusimine biosynthetic cluster in *E. coli*, and used precursor-directed biosynthesis, both *in vitro* and *in vivo*, to probe the promiscuity and flexibility of the AusA terminal Re domain in its native context within full-length NRPS. We have demonstrated the flexibility of the A domains and the C-terminal Re domain by generating a series of unnatural pyrazinone analogues and have observed the formation of an additional diketopiperazine and pyrazine series of compounds as side products of the AusA assembly line. However, since neither the *in vitro* or *in vivo* experiments yielded additional diketopiperazines when a pyrazinone congener could not be produced, most of the analogue restriction appears to come from the second A domain and the Re domain seems to be flexible in reducing substrates loaded onto the assembly line.

From these studies, a subset of novel pyrazinone variants was generated firstly *in vitro* and subsequently *in vivo*. These studies demonstrated the utility of coupling *in vitro* probing with the ability to produce the corresponding products *in vivo*, which is often a challenge in biosynthetic manipulations. Importantly, the flexibility of Re shows that AusA is capable of processing a wide variety of substrates, and we have generated a 2.8 Å crystal structure of this C-terminal macrocyclising Re domain; the first such Re to be structurally characterised. These studies provide structural insight into these flexible nonribosomal peptide termination catalysts, and these data could be critical to revealing how they may be successfully integrated into synthetically created natural-product assembly lines either by combinatorial biosynthesis or de novo assembly-line construction using synthetic biology.

3.5 Materials and Methods

3.5.1 Strains, culture media and general methods

Escherichia coli strains BL21 (DE3), and DH5α were used in this study. Strains bearing the plasmid pAusAB were routinely grown in lysogeny broth (LB) supplemented with kanamycin (25 µg mL^{-1}). Growth conditions and supplements varied as indicated. Fermentation was carried out in 250 mL flasks containing LB (50 mL, Bioshop Canada Inc, Ontario) in an incubated shaker at 175 rpm. Standard procedures were used for DNA manipulations.[31] Oligonucleotides were purchased from Integrated DNA Technologies (Coralville, Iowa), and DNA sequencing was carried out at the Mobix Lab (Hamilton, Ontario).

3.5.2 PCR amplification and cloning of *ausA/B* and AusA reductase domain

SAV0179 and *SAV0180* from *S. aureus* Mu50 (ATCC#: 700699) were amplified by using the following primers: GGAAGG ATCCTA AAGAAG GACTTT TTATGA TTATGG (BamHI site underlined), antisense: GGAAGC GGCCGC ACTACT CAATAA CTGAAA TACAG (NotI site underlined). The Re domain was amplified by using the following primers: sense GGAAGG ATCCAC AATCAT TAGTTG CA (BamHI site underlined); antisense GGAAGC GGCCGC CTTATT GAATAT (NotI site underlined). The T-Re domain was amplified by using the following primers: sense: GGAAGG ATCCGT CTAATA AAGTGT AT (BamHI site underlined); antisense, GGAAGC

GGCCGC CTTATT GAATAT (NotI site underlined). Reactions were subjected to PCR and purified by using a Qiaquick PCR purification kit (Qiagen). The amplicon was digested with BamHI and NotI (New England Biolabs), and ligated with T4 DNA ligase (New England Biolabs) into a similarly digested pET28b vector. According to the design, the expressed SAV0179 and SAV0180 would both bear His$_6$-tags expressed from *E. coli* following induction with IPTG. The resulting constructs, pAusAB, pAusARe and pAusATRE, were then transformed into chemically competent *E. coli* DH5α cells (Invitrogen) by standard heat-shock methods.

3.5.3 Cloning and protein purification of AusA and T-Re/Re domains

LB was inoculated (1:100) from a fresh overnight culture of *E. coli* BL21(DE3) bearing the pAusAB, pAusATRe or pAusARe plasmid. Cells were grown for 2 h at 37 °C. The cultures were cooled on ice and induced with IPTG (50 μM). Cultures were grown for an additional 5 h at 28 °C (pAusAB) or overnight at 15 °C (pAusATRe or pAusARe). Cells were harvested by centrifugation and resuspended in lysis buffer (50 mL, 20 mm Tris, 500 mm NaCl, 20 mm imidazole, pH 8.0) per l.5 L of culture. Cells were lysed by using a continuous-flow cell disruptor at 172 000 kPa, and the lysate was clarified by centrifugation (20 000 *g*). Ni-NTA agarose slurry (50 %, 1 mL, Qiagen) was added to the supernatant, and the mixture was incubated at 4 °C for 1 h. The resin was subjected to step-gradient elution with imidazole. The fraction containing AusA was concentrated by using 30 MWCO protein concentrators (Millipore) and applied to a size-exclusion column (HiLoad 26/60 Superdex 200 column (GE Life Sciences, USA)). The fraction containing AusA was concentrated once again and further purified on an anion-exchange column (Mono Q 5/50 GL (GE Life Sciences). The final fractions containing purified AusA were used for further *in vitro* studies. The sample containing T-Re and Re was concentrated and frozen for crystallisation. Re labelled with selenomethionine (Re-SeMet) was expressed in the methionine auxotrophic strain *E. coli* B834 by using SeMet M9 medium from Shanghai Medicilon Inc. (http://www.mediciloninc.com). Re-SeMet was expressed and purified in the same manner as wild-type Re, except that cells were induced with IPTG when the OD_{600} reached ~1.2. Pure Re-SeMet was buffer exchanged into Tris (20 mm, pH 8) supplemented with NaCl (200 mm).

3.5.4 *in vitro* reconstitution of AusA and analogue production

In vitro production of aureusimine and unnatural analogues was performed in 1 mL volumes under the standard reaction conditions of Tris (25 mM, pH 8.0), NaCl (50 mm), $MgCl_2$ (1.5 mM), ATP (3.0 mM), NADH/NADPH (1.5 mM), valine (1.5 mM), amino acid **2** (1.5 mM, varied depending on reaction), and purified AusA (20 μg). The reaction was initiated by the addition of NADH/NADPH and kept at 30 °C for 18 h. The entire reaction mixture was extracted with ethyl acetate (4×1 mL). The ethyl acetate was evaporated, and the dried products were dissolved in methanol (200 mL).

3.5.5. Analysis of *in vitro* analogue production

Products were analysed by HPLC-MS using a Dionex Ultimate 3000 HPLC system and a Bruker AmazonX ion-trap mass spectrometer. Separation was achieved on a

Phenomenex Luna 5u C_{18} column. The mobile phase was linear from 15 % acetonitrile+0.1 % formic acid (FA), 85 % water+0.1 % FA at 5 min to 80 % acetonitrile+0.1 % FA at 25.5 min. Metabolites were detected by using a Bruker AmazonX ion-trap mass spectrometer with a dry flow rate of 8 L min^{-1}, capillary voltage of −4500 V and an offset potential of −500 V with scanning from 70–1600 *m*/*z* in UltraScan mode. Auto-MS-MS was used for fragmentation. The number of precursor ions was set to 3. Smart fragment parameters were used with a CID of 1 V and ramping from 25.0 % to a final amplitude of 180.0 %. Aureusimine analogues were identified by predicted molecular weight and verified by a common neutral loss of 70 Da (see the Supporting Information). Diketopiperazine analogues were verified by molecular weight and by MS-MS fragmentation. Pyrazine derivatives were identified by MS and NMR . A complete list of retention times, and identified masses can be seen in Table S3.4.

3.5.6 Small-molecule analysis of *E. coli*+pAusAB

Lysogeny broth was inoculated (1:100) from a fresh overnight culture of *E. coli* BL21(DE3) bearing the pAusAB plasmid. Cells were grown for 2 h at 37 °C. The cultures were cooled on ice and induced with IPTG (50 μm). Cultures were grown for an additional 16 h at 28 °C. Ethyl acetate (1.0 equiv) was added to the cultures, and the organic layer was collected and dried by using a rotary evaporator. Each culture was extracted twice. The crude extract was dissolved in methanol (1 mL) and an aliquot (100 μL) was used for analysis by LC-MS.

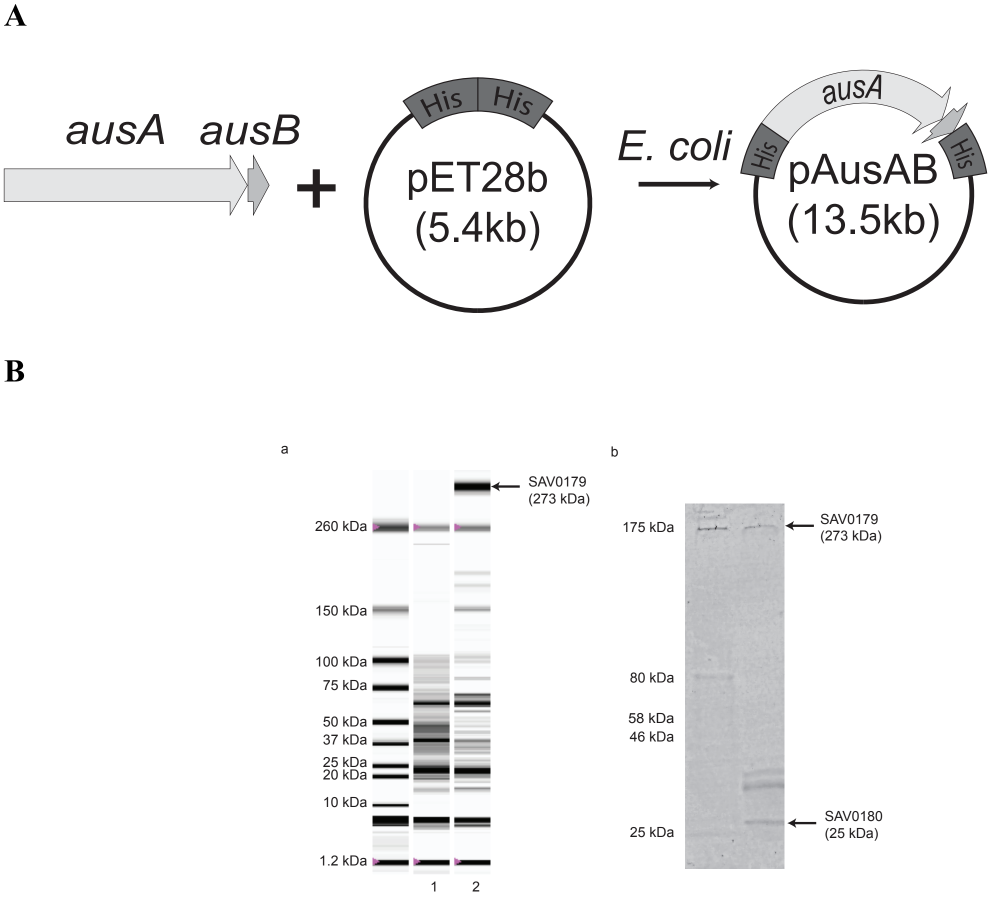
3.5.7 *In vivo* precursor-directed biosynthesis

Lysogeny broth supplemented with amino acid derivatives [4-fluoro-l-phenylalanine (Bachem, USA), 4-chloro-l-phenylalanine (Bachem) 4-bromo-l-phenylalanine (Bachem), 3-chloro-l-tyrosine (Sigma–Aldrich) and (*S*)-(+)-α-aminocyclohexanepropionic acid (Sigma–Aldrich) 1 mm] was inoculated (1:100) from a fresh overnight culture of *E. coli* BL21(DE3) bearing the pAusAB plasmid. Cells were grown for 2 h at 37 °C. The cultures were cooled on ice, induced with IPTG (50 µm) and grown for an additional 16 h at 28 °C. Ethyl acetate (1.0 equiv) was added to the cultures, and the organic layer was collected and dried by using a rotary evaporator. Each culture was extracted twice. Crude extracts were dissolved in methanol (1 mL) and an aliquot (20 µL) was analysed by using a Sunfire C_{18} 5 µm column (4.6×50 mm). The gradient method is described in the Supporting Information. Metabolites were detected by using a Waters 2998 photodiode array and a Micromass ZQ in ES+ mode.

3.5.8 Crystal structure generation

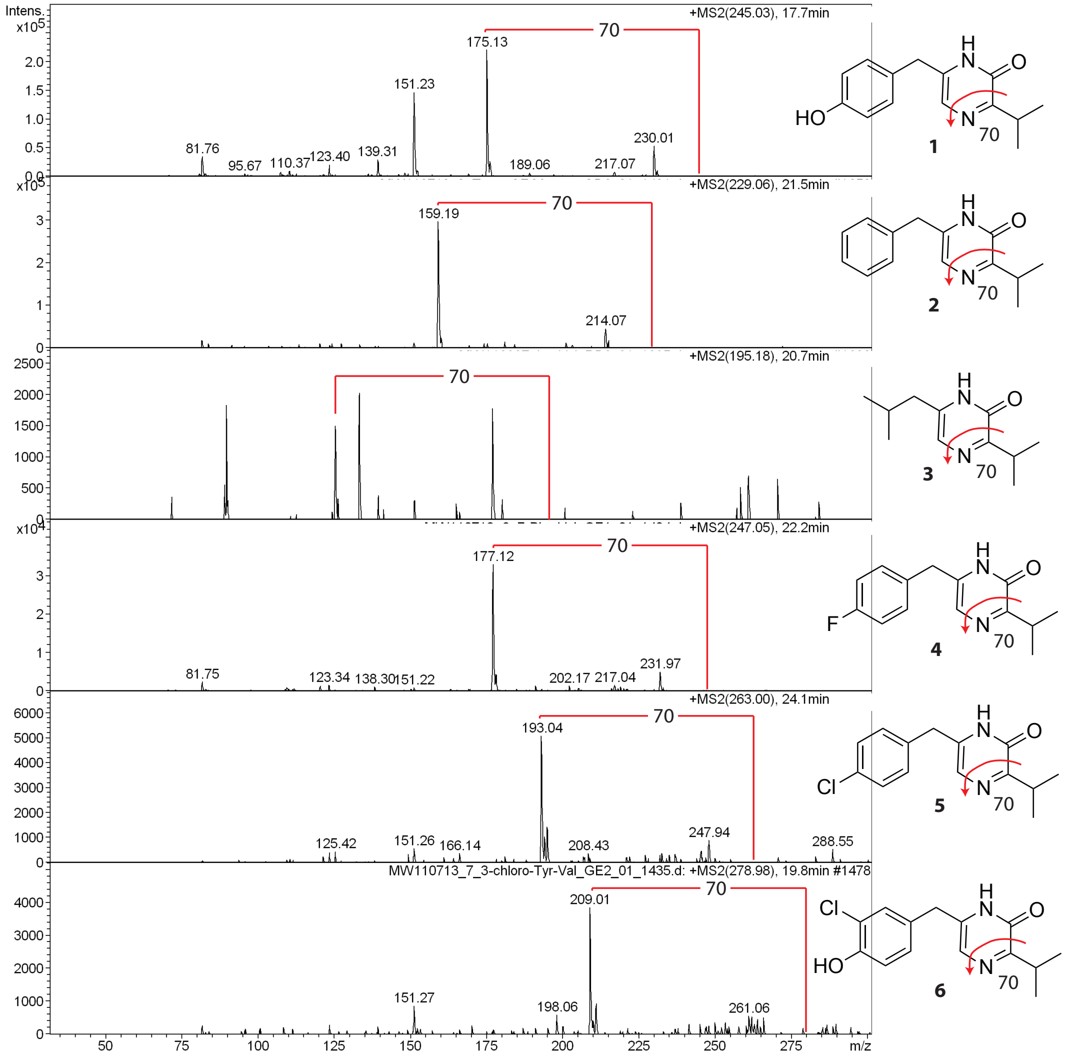
See the Supporting Information. The PDB accession numbers for the Re and T-Re structures are 4F6C and 4F6L, respectively.

3.6 Supplemental Information



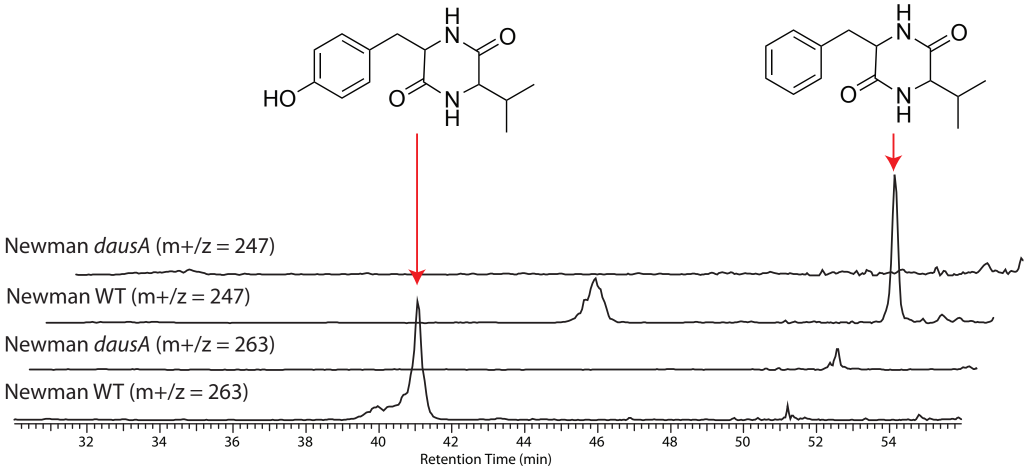
Supplemental Figure 3.1.

A Aureusimine biosynthetic cluster (*ausA* and *ausB***)** was cloned into an inducible pET28b vector allowing for histidine tags of the N-terminal of AusA and C-terminal of AusB. **B** (a) Automated gel electrophoresis of *E. coli* + pAusAB. Lane 1: 0µM IPTG, grown at 37°C overnight. Lane 2: 50µM IPTG induction, grown at 28°C overnight (optimal conditions). Bands are representative of absorption peaks (λ=280 nm) eluting off an automated electrophoresis capillary and compared to an internal standard. (b) Western blot of purified protein from *E. coli* + pAusAB using anti-histidine antibodies. Protein SAV0179 and SAV0180 correspond to Aus A and B respectively.



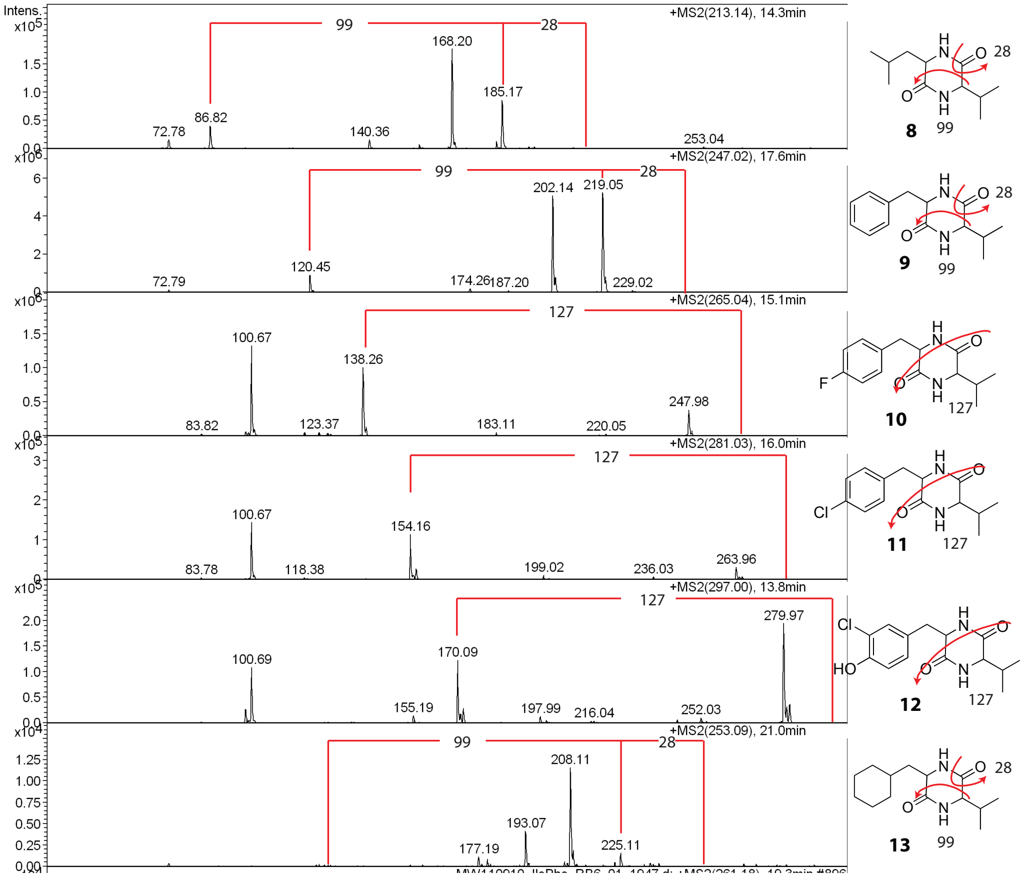
Supplemental Figure 3.2. MS-MS fragmentation of aureusimine analogues produced through *in vitro* chemoenzymatic synthesis.

Diagnostic fragments are indicated for each aureusimine analogue **1-6.**



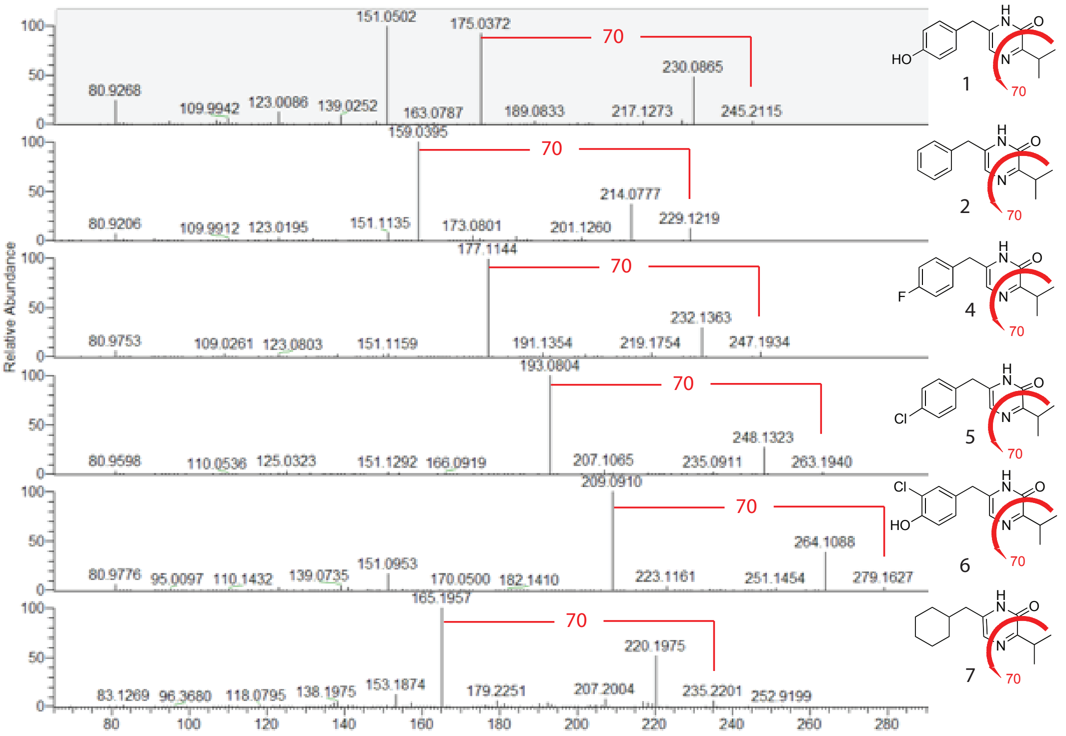
Supplemental Figure 3.3. LC-MS extracted ion chromatograms of *S. aureus* Newman wild type and Newman Δ*ausA* extracts.

Cyclo (Tyr-Val) and cyclo(Phe-Val) diketopiperazines are only produced in Newman wildtype, indicating they are products of AusA *in vivo*.



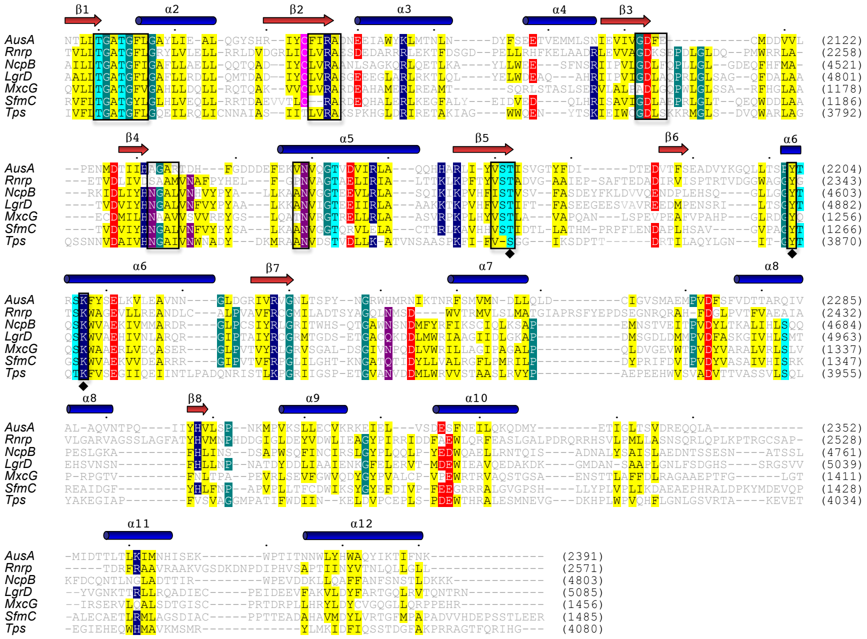
Supplemental Figure 3.4. MS-MS fragmentation of diketopiperazine analogues produced through *in vitro* chemoenzymatic synthesis.

Diagnostic fragments are indicated for each diketopiperazine **8-13.**



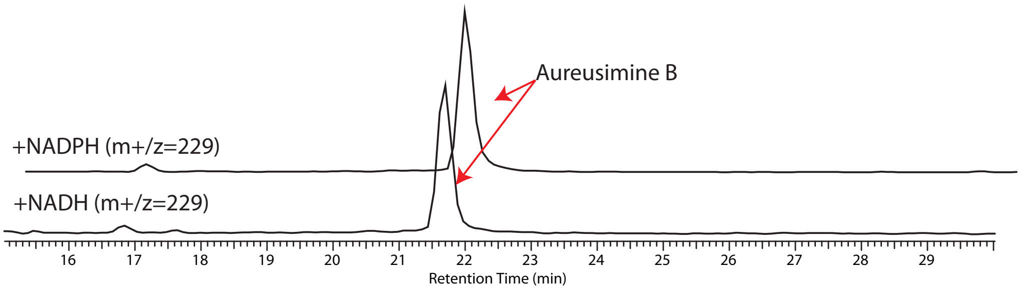
Supplemental Figure 3.5. MS/MS fragmentation of pyrazinone congeners produced in *E. coli* pAusAB.

Samples were analysed by ESI in positive ion mode: FT and IT modes. MS/MS experiments were performed on the most abundant ions observed in the full MS spectrum. The fragmentation behavior of each compound in in MS + mode is shown for the orbitrap.



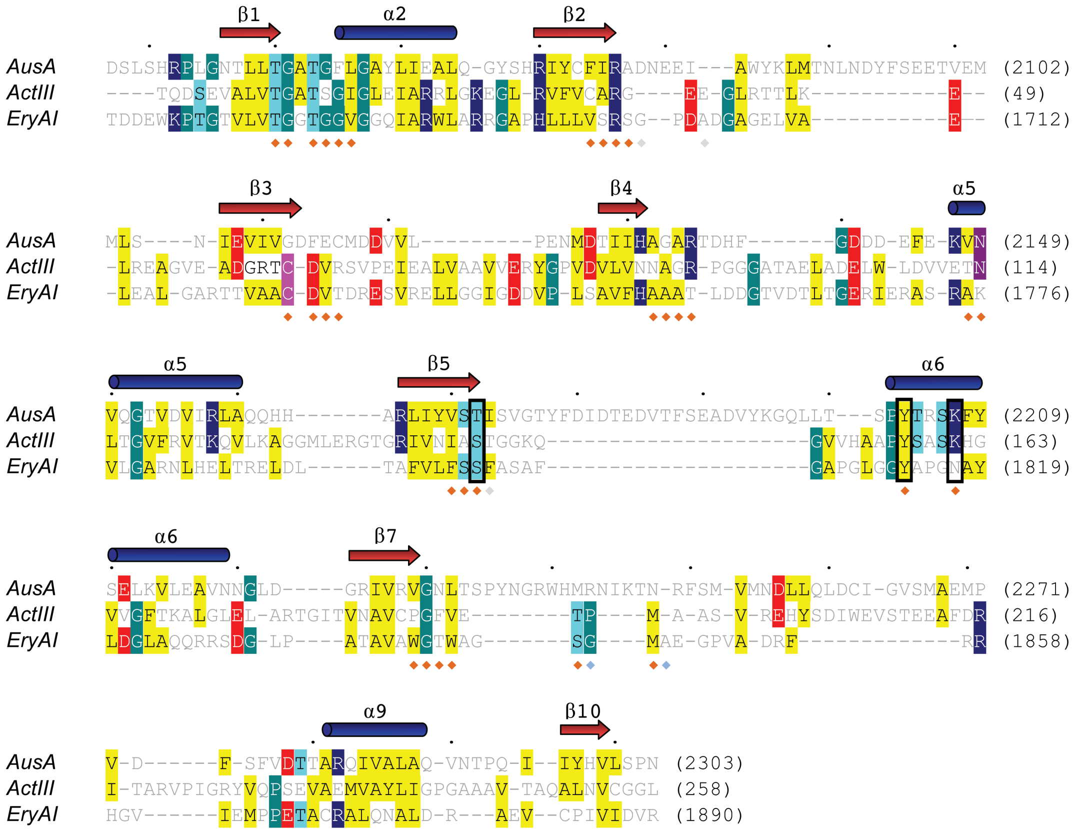
Supplemental Figure 3.6. Sequence alignment of AusA with NRPS terminal reductases:

$\mathrm{R}_{\mathrm{nrp}}$ (*M. tuberculosis* UT205), ncpB (*Nostoc sp.* ATCC 53789), IgrD (*Brevibacillus parabrevi*), mxcG (*Stigmatella aurantiaca* Sg a15), sfmC (*Streptomyces lavendulae*), and TPS (*Hypocrea virens*). Secondary structure elements of AusA (red arrows, beta strands; blue cylinders, alpha helices) are indicated above the corresponding sequence. Conserved residues are coloured as follows: hydrophobic in yellow, positively charged in blue, negatively charged in red, serine and threonine in cyan, proline and glycine in green, cysteine in light purple and asparagine in dark purple. Active site residues are indicated by a black diamond. Residues that interact with NADPH are enclosed in black boxes. The accession numbers for the synthetases indicated are: nostocyclopeptide (AY167420), gramicidin (Q70LM4), myxochelin (AAG31130), saframycin (AAC44129), and peptaibol (AAM78457).



Supplemental Figure 3.7.

LC-MS extracted ion chromatogram of *in vitro* production of aureusimine B using both NADH and NADPH as coenzymes.


β1
α2
β2
AusA DSLSHRPLGNTLLTGATGFLGAYLIEALQ-GYSHRIYCFIRADNEEI---AWYKLMTNLNDYFSEETVEM (2102)
ActIII ----TQDSEVALVTGATSGIGLEIARRLGKEGL-RVFVCARG----EE-GLRTTLK-----------E-- (49)
EryAI TDDEWKPTGTVLVTGGTGGVGGQIARWLARRGAPHLLLVSRSG--PDADGAGELVA-----------E-- (1712)
β3
β4
α5
AusA MLS----N-IEVIVGDFECMDDVVL---------PENMDTIIHAGARTDHF-------GDDD-EFE-KVN (2149)
ActIII -LREAGVE-ADGRTC-DVRSVPEIEALVAAVVERYGPVDVLVNNAGR-PGGGATAELADELW-LDVVETN (114)
EryAI -LEAL-GARTTVAAC-DVTDRESVRELLGGIGDDVP-LSAVFHAAAT-LDDGTVDTLTGERIERAS-RAK (1776)
α5
β5
α6
AusA VQGTVDVIRLAQQHH-------ARLIYVSTISVGTYFDIDTEDVTFSEADVYKGQLLT---SPYTRSKFY (2209)
ActIII LTGVFRVTKQVLKAGGMLERGTGRIVNIASTGGKQ--------------------GVVHAAPYSASKHG (163)
EryAI VLGARNLHELTRELDL------TAFVLFSSFASAF--------------------GAPGLGGYAPGNAY (1819)
α6
β7
AusA SELKVLEAVNNGLD-----GRIVRVGNLTSPYNGRWHMRNIKTN-RFSM-VMNDLLQLDCI-GVSMAEMP (2271)
ActIII VVGFTKALGLEL-ARTGITVNAVCPGFVE--------TP----M-A-AS-V-REHYSDIWEVSTEEAFDR (216)
EryAI LDGLAQQRRSDG-LP----ATAVAWGTWAG-------SG----MAE-GPVA-DRF------------RR (1858)
α9
β10
AusA V-D------F-SFVDTTARQIVALAQ-VNTPQ-I--IYHVLSPN (2303)
ActIII I-TARVPIGRYVQPSEVAEMVAYLIGPGAAAV-TAQALNVCGGL (258)
EryAI HGV------IEMPPETACRALQNALD-R---AEV--CPIVIDVR (1890)

Supplemental Figure 3.8.

Aligned catalytic residues of reductase domains from AusA, ActIII and EryAI . Elements are coloured as follows: oxygen in red, nitrogen in blue and carbon coloured in yellow, blue and gray for AusA, ActIII and EryAI respectively. NADPH bound to ActIII and EryAI are labelled in blue and gray respectively. C) Structure-based sequence alignment of AusA, ActIII and EryAI. Conserved secondary elements (red arrows, beta strands; blue cylinders, alpha helices) between the three proteins are indicated over respective sequence alignment and are labelled corresponding to the AusA structure. Conserved residues are coloured as follows: hydrophobic in yellow, positively charged in blue, negatively charged in red, serine and threonine in cyan, proline and glycine in green, cysteine in light purple and asparagine in dark purple. Active site residues are enclosed by a black box. Orange diamonds indicate common NADPH binding residues shared between ActIII and EryAI; blue diamonds are ActIII-specific NADPH binding residues, gray diamonds are EryAI-specific NADPH binding residues.

3.6.1 Supplemental Methods

3.6.1.1 Reductase Structure Determination

All crystals were grown at 20°C using the hanging-drop vapor diffusion method. Purified Re-SeMet protein was concentrated to 25 mg/mL and mixed with 1 μL of the precipitant (0.2 M tri-sodium citrate, 0.1 M HEPES pH 7.5, 10 % isopropanol) and equilibrated against 800 μL of 1.4 M ammonium sulphate at 20°C. Crystals were grown for 1 week before moving to 4°C. Crystals remained at 4°C for two days before being dehydrated over 800 μL of 4 M ammonium sulphate for an additional two days. Crystals of native TRe were grown at 4°C by mixing 2 μL of 21 mg/mL protein (20 mM Tris pH 8, 200 mM NaCl) with 1 μL of precipitant containing 10 % (w/v) PEG8000, 0.1 M Tris pH 7, 0.2 M $MgCl_2$. TRe crystals were processed similar to Re-SeMet prior to data collection. All crystals were flash frozen with liquid nitrogen prior to data collection.

A single-wavelength anomalous diffraction (SAD) data set for Re-SeMet and a native data set for TRe were collected at wavelengths of 0.9802 and 1.1 Å, respectively. Re-SeMet and TRe data sets were collected under cryogenic conditions (100 K) at beam lines X25 and X29 of the National Synchrotron Light Source at Brookhaven National Labs, respectively. Data was processed using *HKL2000*[1]. For Re-SeMet, *Phenix-*

AutoSol was used to locate 35 of the 36 predicted Se sites and also for phasing and density modification[2]. The structure of TRe was determined by molecular replacement using *Phenix-AutoMR*, and the structure of Re-SeMet as a search model[2]. Model building and refinement of both Re and TRe structures were carried out utilizing multiple iterations of *Coot* and *Phenix-Refine* until R and Rfree values converged and geometry statistics reached suitable ranges[2,3]. In both structures, no density corresponding to NADPH was observed. Excluding the His tag fusions and residues added during cloning (see preparation of expression constructs), the final Re model, contained several residues that were not modeled due to disorder, including: residues 2010-2012, 2115-2117, 2137-2143 in chain A; as well as 2010-2011, 2115-2116, 2137-2141, 2344-2347 in chain B. Similarly, a number of residues were not able to be modeled in the TRe structure, including: 1929-2012 for both chain A and B; as well as 2136-2143 and 2391 in chain B. Structural images were generated using PyMol[4].

Supplemental Table 3.1. Crystallographic data and refinement statistics.

Data Collection		
Crystal	Re (SeMet)	TRe (Native)
Wavelength (Å)	0.9802	1.100
Space group	$P2_12_12_1$	$P2_12_12_1$
Unit-cell parameters		
(a,b,c Å)	105.7 ,106.4, 124.9 $\alpha = \gamma = \beta = 90.0$	104.8, 107.8, 127.3 $\alpha = \gamma = \beta = 90.0$
Molecules in A.U.	2	2
Resolution range (Å)[a]	50.0 – 2.8 (2.9 – 2.8)	50.0 – 3.9 (3.97 – 3.9)
Data Redundancy[a]	7.6 (7.6)	7.1 (7.4)
Completeness (%)[a]	99.9 (100)	100 (100)
I/σ(I)[a]	27.1 (3.4)	12.5 (5.7)
R_{merge} (%)[a]	6.5 (61.8)	13.2 (45.9)
Wilson *B* factor	75.59	97.5
Model and refinement		
Resolution range (Å)[a]	50.0 – 2.8	50.0 – 3.9
R_{work} (%)	20.6	20.0
R_{free} (%)	24.0	24.6
Refl. observed	33,530	13,894
Refl. test set	1,935	1,393
Protein atoms	5,545	5,714
No. of waters	4	0
rmsd bond lengths (Å)	0.009	0.004
rmsd bond angles (°)	1.12	0.84
Average *B* factor (Å^2)	94.64	140.1

[a] Data for the highest resolution shell are shown in parentheses.

3.6.1.2 *S. aureus* Fermentation

S. aureus Newman and *S. aureus* Newman Δ*ausA* were grown in Fernbach flasks for 3 days at 37 °C, 175 rpm in 1 L of chemically defined media consisting of: 3 g KCl, 9.5 g NaCl, 1.3 g $MgSO_4$, 4 g $(NH_4)_2SO_4$, 12.1 g Tris, 125 mg arginine, 200 mg proline, 250 mg glutamic acid, 150 mg valine, 150 mg threonine, 150 mg phenylalanine, 150 mg leucine, 50 mg cystine, 22 mg $CaCl_2$, 140 mg KH_2PO_4, 0.1 mg biotin, 2 mg thiamine, 2 mg nicotinic acid, 2 mg calcium pantothenate, 6 mg $FeSO_4$, 10 mg $MnSO_4$, 6 mg citric acid, 5 g glucose, per 1 L distilled water**.** Cultures were extracted twice with equal volumes of ethyl acetate and evaporated. Extracts were resuspended in 1 mL methanol. Extracts were analysed by HPLC-MS using a Dionex Ultimate 3000 HPLC system and a Bruker AmazonX ion trap mass spectrometer. Separation was achieved with a Phenomenex Luna 5u C_{18} column. The mobile phase was Curved (8) from 5 % acetonitrile + 0.1 % FA, 95 % water + 0.1 % FA at 2 min to 95% acetonitrile + 0.1 % FA at 54 min. Detection of metabolites was carried out using a Bruker AmazonX ion trap mass spectrometer with a dry flow rate of 8 L/min, capillary voltage of -4500 V and an offset potential of -500 V scanning from 70-1600 m/z in UltraScan mode. Auto-MS-MS was used for fragmentation. Number of precursor ions was set to 3. Smart fragment parameters were used with a CID of 1 V using ramping from 25.0 % to a final amplitude of 180.0 %. Aureusimine analogues were identified by predicted molecular weight and verified by a common neutral loss of 70 molecular weight units (see supplemental information). Diketopiperazine AusA products were verified by molecular weight and by MS-MS fragmentation.

3.6.1.3 Western Blot

A Western blot was performed on protein from *E. coli* + pAusAB to determine the presence of heterologously expressed proteins bearing a histidine tag. The protein sample was run on an 8-16 % gradient SDS-PAGE gel for 1 h at 110 V. The protein was then transferred onto Immobilon-P polyvinyl difluoride (PVDF) membrane (Millipore) for 1.5 h at 95 V. The membrane was shaken at room temperature in PBS, 5 % milk for 1

h. Mouse Anti-His antibody (1:2000 dilution) (Bioshop Canada Inc.) was subsequently added and shaken for 1 h. The membrane was then washed with PBS + 0.05 % Tween20 (3 x 5 minutes). HRP conjugated, rabbit anti-mouse secondary antibody (25 µg) (Bioshop Canada Inc.) was added in 5 % milk and shaken for 1 h. The membrane was rinsed with TBST (10 mM Tris, pH 8.0, 0.05 % Tween20) (3 X 5 min). The blot was developed using 3,3'-diaminobenzidine (DAB) and hydrogen peroxide according to Western Blot Kit Protocol (BioBasic, Inc., Can.).

Supplementary Table 3.1

Amino Acid 1	**Amino Acid 2**	**Production (Y/N)**
valine (Bioshop Inc, Canada)	tyrosine (Bioshop Inc, Canada)	N
valine	phenylalanine (Bioshop Inc, Canada)	Y
valine	leucine (Bioshop Inc, Canada)	Y
valine	Amino cyclohexane propionic acid (Sigma Aldrich, Canada)	Y
valine	*p*-Chloro-phenylalanine (Sigma Aldrich, Canada)	Y
valine	*p*-Fluoro-phenylalanine	Y

	(Sigma Aldrich, Canada)	
Valine	*p*-bromo-phenylalanine (Sigma Aldrich, Canada)	N
valine	3-chloro-4-hydroxy-phenyalanine	Y
Valine	*p*-(4-Hydroxybenzoyl)phenylalanine (Bachem, USA)	N
valine	4-methyl-phenyalanine (Peptech, USA)	N
valine	L-4-methylamino-phenylalanine (Peptech, USA)	N
alanine (Sigma Aldrich, Canada)	phenylalanine	N
leucine	phenylalanine	N
Isoleucine (Bioshop Inc, Canada)	phenylalanine	Y
Photoleucine	Phenyalanine	N

3.6.1.4 *Escherichia coli* extract analysis

Crude extracts were dissolved in 1 mL methanol and 20 µL was analysed using a Sunfire C18 5 µm column (4.6x250 mm) using the method below. The flow rate was constant at 1 mL/min.

Supplementary Table 3.2

Time	% Water	% ACN
Initial	80	20
2 min	80	20
20 min	39	61
21min	5	95
24 min	5	95
25 min	80	20
27 min	80	20

3.6.1.5 ESI-MS/MS

Electrospray ionization (ESI) experiments were performed on *E. coli* + pAusAB purified compounds using a ThermoFisher LTQ -XL - Orbitrap Hybrid Mass Spectrometer (ThermoFisher, Bremen, Germany), equipped with an electrospray interface operated in positive or negative ion mode. A 0.1-mg/mL solution of each compound was directly infused into the mass spectrometer, the flow of sample solution was 5 μL/min. The ESI source and MS parameters were automatically optimized and saved in a tune file for the base peak in the mass spectrum. The temperature of the heated capillary was set at 274C. A voltage of 3.9 kV applied to the ESI needle resulted in a distinct signal. Nitrogen was used as auxiliary and sheath gas. The flow rate of the sheath and auxiliary gas was set at 5 (arbitrary units). Helium was used as the damping gas and as the collision gas. The LTQ-XL- Orbitrap mass spectrometer experiment was set to perform a FT full scan from 100-2000 m/z with resolution set at 60,000 (at 500 m/z), followed by linear ion trap MS/MS scans on the top three ions. Dynamic exclusion was set to 2 and selected ions are placed on an exclusion list for 30 seconds. The lock-mass option was enabled for the FT full scans using the ambient air polydimethylcyclosiloxane (PCM) ion of m/z =

445.120024 or a common phthalate ion m/z = 391.284286 for real time internal calibration.

3.6.1.6 Quantification of Aureusimine A and B

Titres of aureusimine A and B were determined by HPLC analysis by integration of aureusimine A and B peaks at a wavelength of 320 nm. For *S.* aureus Newman, 3 X 100 mL of tryptic soy broth (TSB) were inoculated 1:1000 from an overnight culture and fermented at 37°C for 72 h. Each culture was extracted 3 times with ethyl acetate, dried, and resuspended in 1 mL of methanol. 50 µL of extract was analysed using the method from supplementary Table 3. Similarily 3 X 100 mL of LB was inoculated with 1:100 of an overnight culture of *E. coli* + pAusAB and grown for 2h at 37°C. Cultures were cooled on ice and induced with 50µM IPTG and grown at 28°C for 16 h. Cultures were extracted and analysed similar to above. Exact titres were determined by generating a calibration curve using known concentrations of Aureusimine A and B dissolved in 100 mL TSB, which were then subjected to similar extraction and analysis as above.

3.6.1.7 Structure verification and elucidation:

Aureusimine A and B (**1** and **2**) High resolution mass was obtained on an LTQ OrbiTrap XL (Thermo Scientific) giving m+/z = 245.1292 and m+/z = 229.1343. The mass and retention times of the *in vitro* biosynthesized products can be seen in supplementary table 4. Pyrazine structures were initially confirmed by NMR for compound **14** and **15**. All other pyrazine analogues were verified by LC-MS and comparison to control reactions. ^{1}H spectra were recorded on a Bruker 600 MHz NMR spectrometer. Samples were dissolved in deuterated DMSO. δ (integration, multiplicity): Compound **14**: 3.944 (H-4, s), 6.663 (H-4, d), 7.047 (H-4, d), 8.426 (H-2, S), 9.215 (H-2, s). Compound **15:** 3.951 (H-2, s), 4.076 (H-2, s), 6.660 (H-2, d), 7.052 (H-2, d), 7.193 (H-1, s), 7.266 (H-4, m), 8.447 (H-1, s), 8.492 (H-1, s), 9.217 (H-1, s).

Supplementary Table 3.3

Mass and Retention Time of *in vitro* generated AusA products

Compound	**Retention Time (min)**	**Mass (m+/z)**	**Calculated Mass (m+/z)**
1	**17.7 min**	**244.98**	**245.12**
2	**21.6**	**229.06**	**229.13**
3	**20.7**	**195.16**	**195.14**
4	**22.2**	**247.05**	**247.12**
5	**23.9**	**263.01**	**263.09**
6	**19.7**	**278.99**	**279.08**
7	**25**	**235.10**	**235.17**
8	**14.3**	**213.14**	**213.15**
9	**17.6**	**247.02**	**247.14**
10	**15.1**	**265.03**	**265.13**
11	**16.1**	**281.02**	**281.10**
12	**13.8**	**297.00**	**297.09**
13	**21.0**	**253.09**	**253.18**

14	**19.4**	**292.96**	**293.12**
15	**19.3**	**261.18**	**261.15**
16	**28.5**	**296.98**	**297.11**
17	**31.5**	**328.96**	**329.05**
18	**23.0**	**360.94**	**361.04**
19	***only *in vivo***	**277.22**	**277.13**
20	**34.0**	**273.12**	**273.23**

3.6.2 Supplemental References

1. Z. Otwinowski, W. Minor, *Method. Enzymol.* **1997.** 276, 307-326.
2. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. Chen, I. W. Davis, N. Echoo Is, J. J.Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, *Acta. Cryst.* **2010.** D66, 213-221.
3. P. Emsley, K. Cowtan, *Acta. Cryst.* **2004.** D66, 486-501.
4. The PyMOL Molecular Graphics System, Version 1.2, Schrödinger, LLC.

3.7 References

1. D. Schwarzer, R. Finking, M. A. Marahiel, *Nat. Prod. Rep.* 2003, 20, 275.

2. H. D. Mootz, M. A. Marahiel, *Curr. Opin. Chem. Biol.* 1997, **1**, 543–551.

3. C. T. Walsh, M. A. Fischbach, *J. Am. Chem. Soc.* 2010, **132**, 2469–2493.

4. L. Du, L. Lou, *Nat. Prod. Rep.* 2010, **27**, 255.

5. J. W. Trauger, R. M. Kohli, H. D. Mootz, M. A. Marahiel, C. T. Walsh, *Nature* 2000, **407**, 215–218.

6. R. M. Kohli, J. W. Trauger, D. Schwarzer, M. A. Marahiel, C. T. Walsh, *Biochemistry* 2001, **40**, 7099–7108.

7. R. M. Kohli, C. T. Walsh, M. D. Burkart, *Nature* 2002, **418**, 658–661.

8. A. Chhabra, A. S. Haque, R. K. Pal, A. Goyal, R. Rai, S. Joshi, S. Panjikar, S. Pasha, R. Sankaranarayanan, R. S. Gokhale, *Proc. Natl. Acad. Sci. USA* 2012, **109**, 5681–5686.

9. N. Gaitatzis, B. Kunze, R. Müller, *Proc. Natl. Acad. Sci. USA* 2001, **98**, 11136–11141.

10. J. A. Read, C. T. Walsh, *J. Am. Chem. Soc.* 2007, **129**, 15762–15763.

11. F. Kopp, C. Mahlert, J. Grünewald, M. A. Marahiel, *J. Am. Chem. Soc.* 2006, **128**, 16478–16479.

12. M. A. Wyatt, W. Wang, C. M. Roux, F. C. Beasley, D. E. Heinrichs, P. M. Dunman, N. A. Magarvey, *Science* 2010, **329**, 294–296.

13. M. Zimmermann, M. A. Fischbach, *Chem. Biol.* 2010, **17**, 925–930.

14. S. C. Feifel, T. Schmiederer, T. Hornbogen, H. Berg, R. D. Süssmuth, R. Zocher, *ChemBioChem* 2007, **8**, 1767–1770. Direct Link:

15. J. Müller, S. C. Feifel, T. Schmiederer, R. Zocher, R. D. Süssmuth, *ChemBioChem* 2009, **10**, 323–328. Direct Link:

16. D. Matthes, L. Richter, J. Müller, A. Denisiuk, S. C. Feifel, Y. Xu, P. Espinosa-Artiles, R. D. Süssmuth, I. Molnár, *Chem. Commun.* 2012, **48**, 5674–5676.

17. R. Thiericke, J. Rohr, *Nat. Prod. Rep.* 1993, **10**, 265–289.

18. A. Kirschning, F. Taft, T. Knobloch, *Org. Biomol. Chem.* 2007, **5**, 3245–3259.

19. T. Stachelhaus, H. D. Mootz, M. A. Marahiel, *Chem. Biol.* 1999, **6**, 493–505.

20. T. Stachelhaus, H. D. Mootz, V. Bergendahl, M. A. Marahiel, *J. Biol. Chem.* 1998,

273, 22773–22781.

21. S. Gruenewald, H. D. Mootz, P. Stehmeier, T. Stachelhaus, *Appl. Environ. Microbiol.* 2004, **70**, 3282–3291.

22. M. Gondry, L. Sauguet, P. Belin, R. Thai, R. Amouroux, C. Tellier, K. Tuphile, M. Jacquet, S. Braud, M. Courçon, C. Masson, S. Dubois, S. Lautru, A. Lecoq, S. Hashimoto, R. Genet, J. Pernodet, *Nat. Chem. Biol.* 2009, **5**, 414–420.

23. L. Li, W. Deng, J. Song, W. Ding, Q. Zhao, C. Peng, W. Song, G. Tang, W. Liu, *J. Bacteriol.* 2008, **190**, 251–263.

24. Y. Li, K. J. Weissman, R. Müller, *J. Am. Chem. Soc.* 2008, **130**, 7554–7555.

25. B. Manavalan, S. K. Murugapiran, G. Lee, S. Choi, *BMC Struct. Biol.* 2010, **10**, 1.

26. N. Schracke, U. Linne, C. Mahlert, M. A. Marahiel, *Biochemistry* 2005, **44**, 8507–8513.

27. H. Joernvall, B. Persson, M. Krook, S. Atrian, *Biochemistry* 1995, **34**, 6003–6013.

28. U. Oppermann, C. Filling, M. Hult, N. Shafqat, X. Wu, M. Lindh, J. Shafqat, E. Nordling, Y. Kallberg, B. Persson, H. Jornvall, *Chem.-Biol. Interact.* 2003, **143**, 247–253.

29. A. T. Keatinge-Clay, R. M. Stroud, *Structure* 2006, **14**, 737–748.

30. A. T. Hadfield, C. Limpkin, W. Teartasin, T. J. Simpson, J. Crosby, M. P. Crump, *Structure* 2004, **12**, 1865–1875.

31. J. Sambrook, T. M. Fritsch, T. Maniatis, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, New York, 1989.

Chapter 4. Heterologous Production of Nonribosomal Peptides

4.1 Chapter Preface

Often NRPS clusters are identified within unculturable or difficult to culture environmental organisms where NRP production is silenced or extremely low. One technique to target these inaccessible or silent clusters is to genetically express them within amenable heterologous hosts, such as *E. coli*, allowing access to additional chemical diversity not achievable from normal culturing conditions. At the time of this experiment, many NRPS systems had been heterologously expressed, however, there was only limited data involving the optimization of these systems for NRP production. In this third project using the aureusimine biosynthetic locus, the *E. coli* strain bearing an active AusA recombinant assembly line is used as a model system for NRP heterologous production in *E. coli.* This research project optimizes NRP (aureusimine) production within a pET28 expression system in *E. coli* using IPTG concentration, expression temperature, and precursor feeding experiments to facilitate future discoveries using *E. coli* as a heterologous host for NRP production. The simple heterologous system also reveals additional NRPS products of the AusA assembly line and provides examples of how metabolomics techniques, such as principal component analysis, can identify low yield NRP products from complex extracts that are left unidentified using standard analysis techniques. Together, chapters 2-4 contribute to the overall understanding of how chemical diversity is created by NRPS assembly-lines and how we can target these important small molecules within complex extracts using both genomic and metabolomic analysis techniques in a hypothesis-driven approach to natural product discovery.

The following chapter is a modified version of a previously published article. I was the lead author and took a major role in conception, experimental design, execution and analysis. The citation for the original publication is as follows:

Wyatt, M.A., Magarvey, N.A. (2013) **Optimizing dimodular nonribosomal peptide synthetases and natural dipeptides in an *Escherichia coli* heterologous host.** *Biochem. Cell. Biol.*, Published Online 2 January 2013, 10.1139/bcb-2012-0097.

4.2 Abstract

Nonribosomal peptides are an important class of natural products that have a broad range of biological activities. Their structural complexity often prevents simple chemical synthesis, and production from the natural producer is often low, which deters pharmaceutical development. Expression of biosynthetic machinery in heterologous host organisms like *Escherichia coli* is one way to access these structures, and subsequent optimization of these systems is critical for future development. We utilized the aureusimine biosynthetic gene cluster as a model system to identify the optimal conditions to produce nonribosomal peptides in the isopropyl β-d-1-thiogalactopyranoside (IPTG)-inducible T7 promoter system of pET28. Single reaction monitoring of nonribosomal products was used to find the optimal concentration of IPTG, postinduction temperature, and the effect of amino acid precursor supplementation. In addition, principle component analysis of these extracts identified 3 previously undiscovered pyrazine products of the aureusimine biosynthetic locus, highlighting the utility of heterologously expressing nonribosomal peptide synthetases to find new products.

4.3 Introduction

Nonribosomal peptides are an important class of natural products constructed via modular assembly-line enzymes known as nonribosomal peptide synthetases (NRPSs). Intrinsic bioactivity and a range of therapeutic uses make them valuable as human therapeutic agents, agricultural bioproducts (e.g., pesticides), and as agents for other diverse medical and industrial applications (Walsh and Fischbach 2010). Often their study and application are blunted by the lack of significant production from native producers, or by an inability to rapidly alter their structure either chemically or biosynthetically from genetically intractable strains (Walsh and Fischbach 2010). Heterologous expression of nonribosomal peptides in ectopic hosts can afford a situation to readily tune expression and is an approach increasingly taken in an effort to obtain larger amounts, interrogate biosynthetic machineries, and generate chemical variants for bioactivity testing (Gruenewald et al. 2004; Pfeifer et al. 2003; Watanabe et al. 2006; Watanabe and Oikawa 2007). Unfortunately, there are few model studies that reveal how to best control and direct heterologously expressed modular nonribosomal peptide pathways to biosynthesize their products within model hosts such as *Escherichia coli*.

Nonribosomal peptide biosynthetic machinery is encoded within clusters and the encoded NRPSs function like an assembly line. Nonribosomal peptides synthetases recognize, activate, and incorporate amino acid building blocks into growing peptide products. Each elongation module consists of a condensation (C), adenylation (A), and thiolation (T) domain. Adenylation domains recognize and activate specific amino acid building blocks onto a posttranslationally modified phosphopantetheinyl arm of the T domain. Condensation domains then catalyze peptide bond formation between adjacent, T domain–loaded, amino acids. This cycle of condensation reactions continues along the assembly line, increasing the length of the growing peptide product by one amino acid at every module. Ultimately, the fully matured modified peptide chain is released from the NRPS by a thioesterase or reductase domain (Mootz and Marahiel 1997; Schwarzer et al. 2003).

In many instances, studies have used *E. coli* as a host to produce individual enzyme domains, full-length modules, or tailoring enzymes to reveal biosynthetic logic imparted within nonribosomal peptide assembly (Stachelhaus et al. 1995, 1998; Walsh et al. 2001). Demands for understanding how best to direct nonribosomal peptide assembly in heterologous hosts have been spurred by accumulating evidence that many biosynthetic gene clusters are not expressed, or their products are not detected, from native hosts (Lautru et al. 2005). Gaining access to more of these evolved nonribosomal peptides may require decoupling their biosynthetic machineries from their native context and expressing them within heterologous host systems. Transferring such genetic clusters to heterologous hosts may act to decouple these biosynthetic clusters from repressive transcriptional controls that preclude their natural production. Moreover, placing these clusters within microbial cell factory platforms that provide strong and inducible promoters is one approach to achieve tuneable expression of the enzymatic machinery and thus greater yields of the nonribosomal products. In this way, reliable heterologous hosts require defined experiments that establish how best to proceed with heterologous production of these natural products. Also, having well-established heterologous production conditions within genetically amenable hosts such as *E. coli* results in ready access to downstream genetic manipulations of these assembly systems using combinatorial biosynthesis approaches to prepare unnatural natural products via module swapping, deletion, or fusion strategies (Mootz et al. 2000; Stachelhaus et al. 1995). Previous heterologously produced nonribosomal peptide studies within *E. coli* have reported little on the fermentation process, but focused more closely on the genetic organization and protein expression levels in different expression systems to increase yield (Gruenewald et al. 2004; Pfeifer et al. 2003). However, once an expression system has been chosen, it is necessary to understand how these artificial hosts can be turned into cellular factories that maximize the desired product. In some cases, it is beneficial to feed amino acid analogues to cultures expressing NRPSs as a directed biosynthesis strategy resulting in altered end products. Due to the known flexibility of A domains, selective production of congeners can be achieved through supplementing the growth media with amino acids or their derivatives rather than through genetic manipulation (Grüschow et al. 2009; Moran et al. 2009). In this study, we used a simple NRPS system from *Staphylococcus aureus* (AusA and AusB) to examine the effect general fermentation conditions have on nonribosomal peptide formation. The *S. aureus* NRPS gene cluster is a good model system because it represents an entire NRPS cluster and can completely synthesize the nonribosomal products without the addition of other enzymes or genetic elements (Wyatt et al. 2010). We used AusA cloned within an inducible expression vector and monitored aureusimine production using single reaction monitoring (SRM) to determine the effect induction level and expression temperature had on nonribosomal peptide titre and our ability to drive congener production through supplementation with amino acid precursors.

4.4 Results and Discussion

The aureusimine biosynthetic gene cluster from *S. aureus* was used as a model system to examine the effects of common induction parameters within *E. coli* for the production of nonribosomal natural products. This cluster contains a single dimodular NRPS (AusA) followed by a phosphopantetheinyl transferase (PPTase) (AusB) (Fig. 4.1a). This cluster is responsible for the production of the *S. aureus* nonribosomal peptides, aureusimine A and B (1 and 2), and includes the enzymatic domains A-T-C-A-T-Re, where Re is a reductase domain (Fig. 4.1b) (Wyatt et al. 2010; Zimmermann and Fischbach 2010). Based on the isolated compounds and amino acid specificity conferring code, the first A domain is predicted to activate a valine onto the assembly line, whereas the second A domain is flexible, incorporating either a tyrosine or phenylalanine in compound 1 and 2, respectively (Wyatt et al. 2010; Stachelhaus et al. 1999). Previous studies have shown aromatic A domains to be flexible and this system allows for a unique opportunity to explore the ability to drive production of one congener through amino acid supplementation in an *E. coli* heterologous host (Magarvey et al. 2006, Moran et al. 2009; Weist et al. 2002).

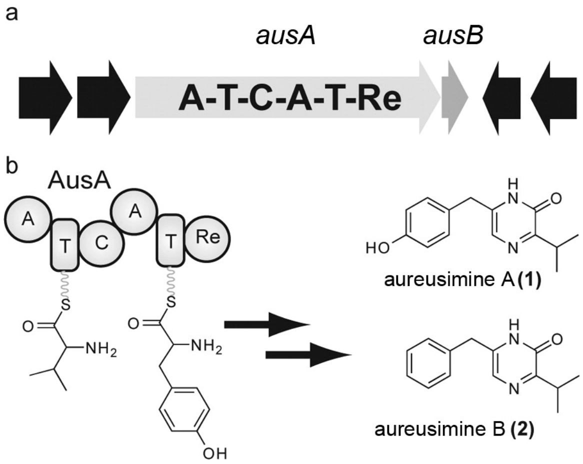


Figure 4.1. Aureusimine biosynthetic cluster and assembly line.
a) Aureusimine dimodular NRPS gene (*ausA*) and posttranslational modifying enzyme, phosphopantetheinyl transferase, (*ausB*) in *S. aureus* Mu50. *b*) AusA biosynthetic assembly line is a dimodular NRPS containing the following domains indicated by shaded spheres: adenylation (A), thiolation (T), condensation (C), and reductase (Re). AusA produces the natural products aureusimine A (1) and aureusimine B (2) in *S. aureus.*

All NRPSs are produced in an inactive state until posttranslationally modified by a PPTase (Lambalot et al. 1996; Schlumbohm et al. 1991). Phosphopantetheinyl transferases act to attach a phosphopantetheine arm onto a conserved serine within T domains of the NRPS, which later are used as a chemical handle for the activated amino acids and growing peptide chain (Schlumbohm et al. 1991). Most systems to date have used a promiscuous PPTase, Sfp, from *Bacillus subtilis* through the addition of either a second expression plasmid containing the *sfp* gene or by incorporation of *sfp* into the *E. coli* genome (Balibar and Walsh 2006; Fortin et al. 2007; Gruenewald et al. 2004). In this instance, the small size of the *S. aureus* gene cluster enabled the direct amplification of both the NRPS (*ausA*) and the native PPTase (*ausB*) in a single contiguous piece (7.9 kb) by PCR and subsequent ligation into a pET28 expression vector generating the plasmid, pAusAB. This allowed equal induction of both enzymes after the addition of IPTG, which is most similar to the native *S. aureus* NRPS in which all components are under the control of a single promoter (Sun et al. 2010). After induction with IPTG, there was an increase in a protein band at >175 kDa (calculated AusA ~275 kDa) in the cell-free extract, compared to an uninduced culture, and this band was confirmed to be AusA through additional Western blot analysis using antihistidine antibodies (Supplementary data Fig. 4.1)[1]. AusA was shown to be active in growing *E. coli* cultures through the identification of aureusimine A and B by LC-MS analysis of induced *E. coli* + pAusAB extracts and comparison with authentic aureusimine standards (Fig. 4.2).

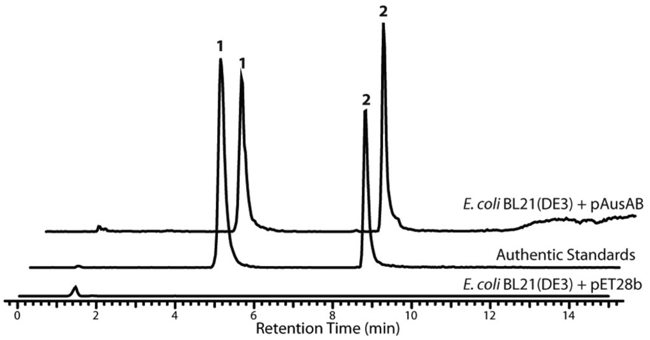


Figure 4.2. Aureusimine A and B production in *E. coli* + pAusAB. Supernatant of *E. coli* + pAusAB and *E. coli* + pET28b empty vector was analyzed by single reaction monitoring (see methods) and compared with aureusimine A and B authentic standards, confirming production in the heterologous host.

After verifying the production of both aureusimine A and B, we sought to identify the culture conditions that would maximize aureusimine A and B titres. To determine the production titres effectively, a HPLC–MS method was developed that used SRM to determine aureusimine production directly from culture supernatants (see Materials and Methods). This method enabled us to quantify nanogram quantities of aureusimine A and B within complex supernatants without any further extraction. As a result, we were able to determine the optimal IPTG concentration and postinduction growth temperature for maximum yield of aureusimine A and B in *E. coli* cultures and further increase titres through amino acid supplementation.

Induction of nonribosomal peptides (aureusimine A and B) by IPTG was examined first. In all conditions, aureusimine B was produced in greater quantities than aureusimine A, which is contrary to the natural producer, *S. aureus,* in which the most abundant compound is aureusimine A (Wyatt et al. 2010). This may be a result of varied media components, LB vs. TSB. By varying the IPTG concentration, we identified an optimal induction concentration of 10 µmol/L for aureusimine A and B (titres of 3.08 and 7.47 mg/L, respectively) (Fig. 4.3a). At higher concentrations, there is a decrease in production likely due to the stress AusA translation and aureusimine production has on the *E. coli* host. Nonribosomal peptide synthetases are large enzymes, and even the relatively small dimodular NRPS used in this study is 273 kDa and likely has a significant metabolic cost to the host. Overexpression of these enzymes likely takes resources away from general *E. coli* metabolism and replication, resulting in an overall lowered AusA expression of soluble protein at higher concentrations of IPTG (Supplementary data Fig.

4.2)[1]. Cultures not induced with IPTG also had detectable levels of aureusimines A and B (0.61 and 1.31 mg/L), which is likely due to "leaky" gene transcription, a common occurrence in T7 promoter systems (Mertens et al. 1995). Monitoring of the final products has an advantage in these instances as one ismonitoring active NRPS only and can eliminate the need for optimization and analysis of protein that is found in inclusion bodies or is inactive resulting in more direct optimization of nonribosomal peptide production.

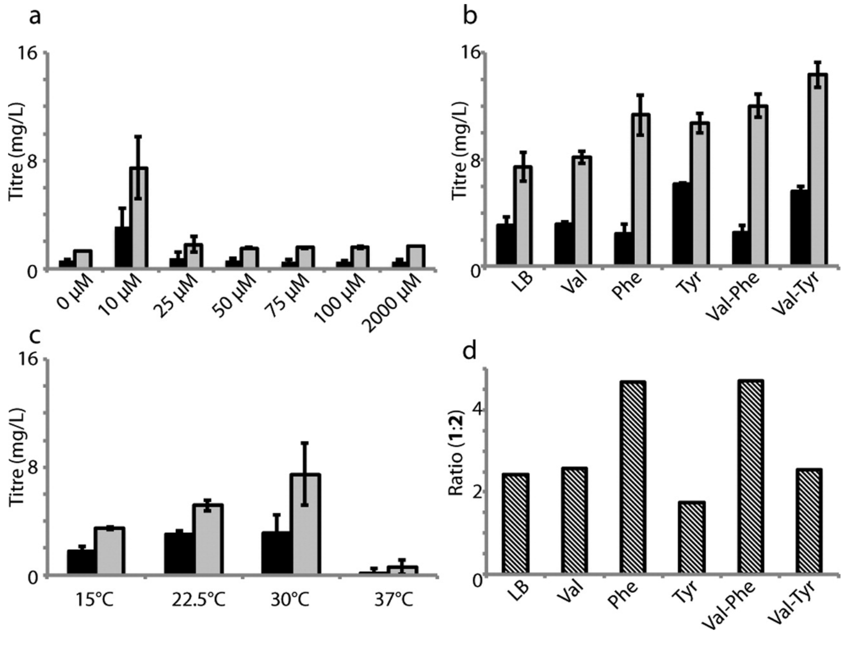


Figure 4.3. Aurueusimine A and B production optimization.

The influence of (a) IPTG concentration at 30 °C, (b) precursor supplementation (2 mmol/L),.and (c) postinduction temperature at 10 µmol/L IPTG on aureusimine A and B production using single reaction monitoring (d) Ratio of aureusimine B to aureusimine A with amino acid precursor supplementation. Aureusimine A is represented by solid black bars and aureusimine B is represented by solid grey bars. Error bars represent the standard deviation of the mean (*n* = 4 for all experiments).

Postinduction temperature also affects aureusimine production. *E. coli* is generally maintained at 37 °C, optimal for metabolism and replication. After induction in heterologous systems, however, maintaining the same metabolic rate concurrently with high heterologous protein expression can quickly exhaust *E. coli* translation machinery and also result in higher amounts of AusA deposited within inclusion bodies that is inactive. *E. coli* grown at high temperatures (37 °C) produced the lowest concentrations of aureusimines, whereas optimal production was found at 30 °C (Fig. 4.3a). It is also possible to alter congener ratios using temperature (Supplementary data Fig. 4.3). One hypothesis is that the flexibility of the A domain may be, in some part, temperature dependent where the active site of the A domain may change conformation at different temperatures. This temperature shifting can be used to drive nonribosomal peptide heterologous systems where one congener is preferred over another.

Another method to enhance the production of one congener over another or to increase titres is to supplement the media with additional amino acid building blocks required for final product formation. Media supplementation with 2 mmol/L tyrosine, phenylalanine, valine, or a combination of both phenyalanine–valine and tyrosine–valine at optimal induction conditions was carried out (Fig. 4.3c). The results suggest that in the *E. coli* system, phenylalanine and tyrosine are limiting factors for aureusimine production, whereas the addition of the common precursor valine has minimal effects.

The production ratio of aureusimine A and B could be altered by increasing the relative production of aureusimine B with the addition of phenylalanine, and the production of aureusimine A with the addition of tyrosine (Fig. 4.3d). Final optimization conditions using 10 μmol/L IPTG, an induction temperature of 30 °C and precursor supplementation with 2 mmol/L tyrosine and 2 mmol/L valine afforded maximum titres of aureusimine A and B of 6.16 mg/mL and 14.34 mg/mL, respectively. These titres can be compared to aureusimine A and B production in *S. aureus* UAMS-1 after 3 days of growth, being 1.3 and 0.5 mg/L, respectively. This significant increase in production highlights the utility of heterologous host's ability to decrease fermentation time and increase titres of compounds derived from difficult to culture bacteria.

Heterologous expression of a biosynthetic assembly system is not only a tool to increase titres not available from the natural producer, but it can also generate structural analogues or "unnatural" natural products due to the increased expression in the inducible host. Extracts of *E. coli* BL21(DE3) + pAusAB and *E. coli* BL21(DE3) bearing an empty pET28b plasmid were subjected to LC–MS analysis and subsequently analyzed using principle component analysis (PCA) (ProfileAnalysis, Bruker Daltonik, Germany). In addition to identifying aureusimine A and B in the pAusAB extract, PCA identified 3 additional compounds (3–5) corresponding to retention times of 48.4, 53.6, and 57.6 min with *m/z* values of 260.03, 276.01, and 292.03, respectively (Fig. 4.4). Purification and elucidation of these peaks revealed 3 additional pyrazines produced by AusA, each incorporating either 2 tyrosines (3), a phenylalanine and tyrosine (4), or 2 phenylalanines (5). These additional products are not identified in *S. aureus* extracts and may be due to reductive release of a single tyrosine or phenylalanine aldehyde from the second A domain that can spontaneously react with another to form compounds 3, 4, and 5. The identification of these new products shows how overexpression in *E. coli* can enable identification of new nonribosomal peptide products that are not detected in the natural host.

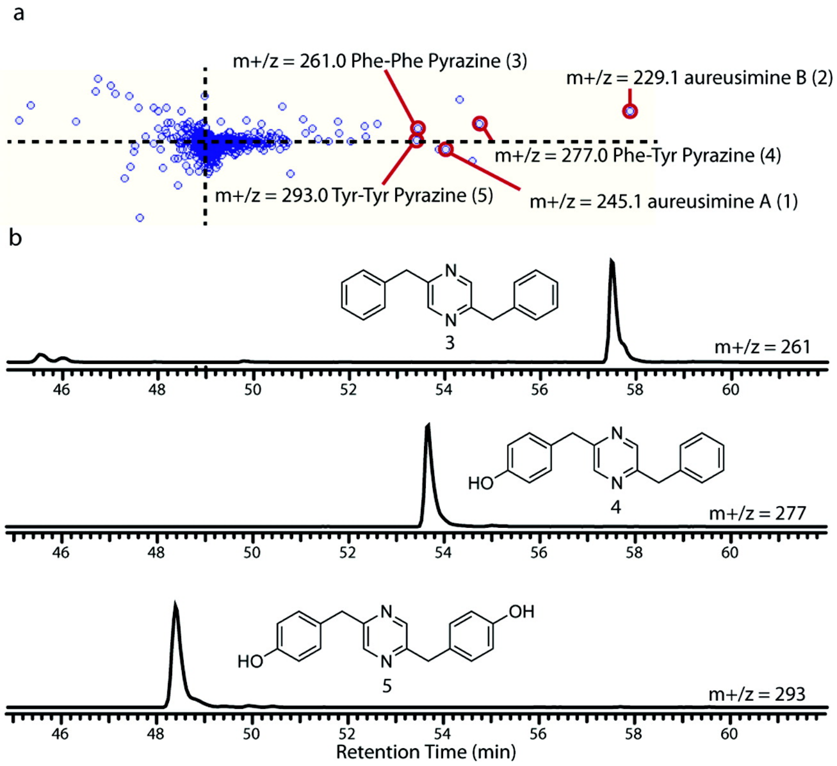


Figure 4.4. .Identification of 3 pyrazine products by principle component analysis (PCA). *a*) PCA scatter plot of *E. coli* + pAusAB and *E. coli* + pET28b empty vector. Compounds contributing to the *E. coli* + pAusAB extract are circled and their masses indicated. These include aureusimine A and B and 3 new pyrazine products. *b*) Extracted ion chromatogram and structures of AusA produced pyrazines 3–5.

Using the AusA dimodular NRPS as a model system, we have identified optimal conditions for nonribosomal peptide production, providing insight into fermentation conditions for secondary metabolite production in inducible *E. coli* heterologous systems. Furthermore, expression of AusA within *E. coli* has successfully identified 3 new pyrazine products from AusA that are either not produced or are undetectable in the native organism, *S. aureus.* Flexibility in heterologously expressed NRPSs to generate new metabolites is an advantage when directing biosynthesis within the natural host is more difficult. By placing the biosynthetic machinery behind inducible promoters within *E. coli*, we can achieve increased production of complex natural products through simple changes in induction conditions and generate analogues not seen in the natural producer facilitating the development of nonribosomal products for industrial and pharmaceutical applications.

4.5 Materials and Methods

4.5.1 Strains, culture media, and general methods

The *E. coli* strains BL21(DE3) and DH5α were used in this study. Cells were grown in Luria-Bertani broth (LB) supplemented with 25 μg/mL kanamycin. Growth conditions and supplements were as indicated. Fermentations were carried out in round-bottom 96-well plates containing 150 μL of LB (Bioshop Canada Inc., ON) in an incubated shaker at 200 rpm. Amino acids were purchased from Bioshop Canada Inc. (Mississauga, ON). Standard procedures were used for DNA manipulations (Sambrook et al. 1989). Oligonucleotides were purchased from Integrated DNA Technologies

(Coralville, IA) and DNA sequencing was carried out at Mobix Lab (Hamilton, ON).

4.5.2 PCR amplification and cloning of aureusimine biosynthetic cluster

The genes *ausA* (*SAV0179*) and *ausB* (*SAV0180*) were amplified in a contiguous piece from *S. aureus* Mu50 genomic DNA (ATCC 700699) using the following primers: GGAA*GGATCC*TAAAGAAGGACTTTTATGATTATGG (*Bam*HI site italicized), antisense: GGAA*GCGGCCGC*ACTACTCAATAACTGAAATACAG (*Not*I site underlined). Reactions were subjected to the following conditions and purified using Qiaquick PCR purification kit (Qiagen):1× (2 min at 98 °C), 25× (10 s at 98 °C, 30 s at 49 °C, 4:15 min at 72 °C), and 1× (10 min at 72 °C). The amplicon was digested with *Bam*HI and *Not*I (New England Biolabs), ligated with T4 DNA Ligase (New England Biolabs) into a similarly digested pET28b vector. According to the design, the expressed SAV0179 and SAV0180 would bear hexahistidine tags on the N-terminus and C-terminus, respectively, following induction with isopropyl β-d-1-thiogalactopyranoside (IPTG). The resulting construct, pAusAB, was then transferred into BL21(DE3) *E.* ***coli*** by electroporation, or into chemically competent cells by standard heat shock methods using DH5α *E. coli* maximum efficiency competent cells (Invitrogen).

4.5.3 Detection and analysis of nonribosomal peptide production by SRM analysis

Supernatants were analyzed by high performance liquid chromatography – mass spectrometry HPLC–MS using a Dionex Ultimate 3000 HPLC system and a Bruker AmazonX ion trap mass spectrometer. Separation was achieved with a Phenomenex Luna

5 µm C18 column. The mobile phase was isocratic from 0 to 2 min at 30% acetonitrile + 0.1% formic acid (FA) increasing linearly to 85% acetonitrile + 0.1% FA at 8.5 min and continuing to 100% acetonitrile at 10 min returning to 30% acetonitrile + 0.1% FA at 11 min at a flow rate of 1.1 mL/min. Detection of metabolites was carried out using SRM with a Bruker AmazonX ion trap mass spectrometer with a dry flow rate of 5 L/min, capillary voltage of −4500 V and an offset potential of −500 V scanning from 100–500 *m/z* in UltraScan mode. Reaction monitoring for aureusimine A: isolated precursor ion = 244.9 with ion monitoring at 174.8 from 0 to 7 min. Reaction monitoring for aureusimine B: isolated precursor ion = 228.9 with ion monitoring at 194.9 from 9.5 to 15 min. Smart fragment parameters were used with a charge injection device (CID) of 1 V using ramping from 50% to 200%. All samples were compared to standards curves generated from authentic standards of aureusimine A and B.

4.5.4 Heterologous production optimization

The LB broth was inoculated (1:100) from a fresh overnight culture of *E. coli* BL21(DE3) bearing the pAusAB plasmid. Cells were grown for 2.5 h at 37 °C, 200 rpm in round-bottom 96-well plates. The cultures were cooled on ice for 10 min and induced with IPTG at varying concentrations (0 µmol/L, 10 µmol/L, 25 µmol/L, 50 µmol/L, 75 µmol/L, and 100 µmol/L). Cultures continued to grow at either 15 °C, 22.5 °C, 30 °C, or 37 °C. After 24 h, the cultures were pelleted by centrifugation and the supernatant was transferred to a separate 96-well plate for analysis by SRM (see previous).

For precursor-driven production, LB broth was supplemented with 2 mmol/L valine, 2 mmol/L phenylalanine, 2 mmol/L tyrosine or a combination of both phenylalanine (2 mmol/L) and valine (2 mmol/L) or tyrosine (2 mmol/L) and valine (2 mmol/L) (L- amino acids were used in all experiments). Cells were grown for 2.5 h at 37 °C, 200 rpm in round-bottom 96-well plates. The cultures were cooled on ice for 10 min and induced with 10 µmol/L IPTG and continued to grow at 30 °C for 24 h. Supernatants were analyzed by SRM as previously described.

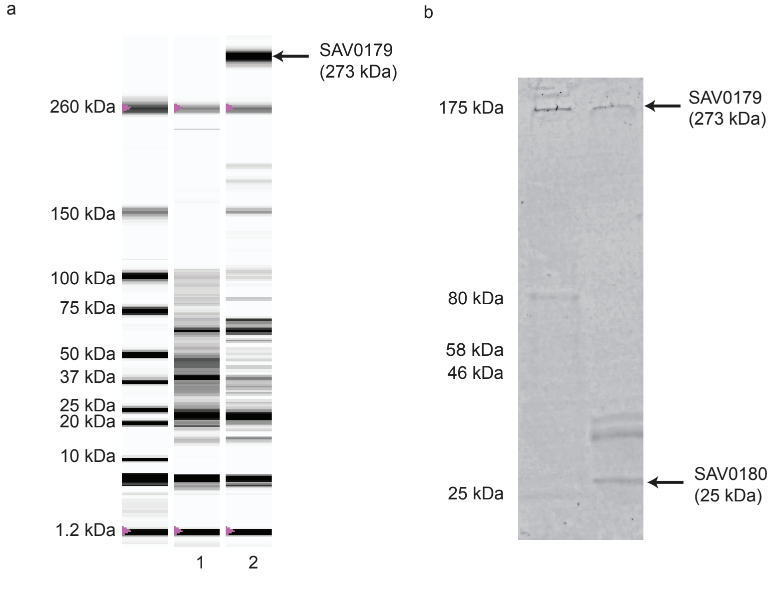
4.5.5 Principle component analysis of *E. coli* + pAusAB

Cultures of *E.* coli + pAusAB and *E. coli* + pET28b were grown in 100 mL LB (triplicate) and induced with 10 µmol/L IPTG and grown at 30 °C as previously described. After 24 h, cultures were extracted 3 times with ethyl acetate and dried. Dried extracts were dissolved in 1 mL methanol and analyzed by HPLC–MS using a Dionex Ultimate 3000 HPLC system and a Bruker AmazonX ion trap mass spectrometer. Separation was achieved with a Phenomenex Luna 5 µm C18 column. The mobile phase was curved (8) from 5% acetonitrile + 0.1% FA, 95% water + 0.1% FA at 2 min to 95% acetonitrile + 0.1% FA at 54 min. Principle component analysis was carried out on the extracts using ProfileAnalysis (Bruker Daltonik, Bremen). The LC–MS data was prepared for PCA using a bucketing approach of the raw data. The LC–MS data was integrated into time and *m/z* buckets of 0.1 m and 1 *m/z* from 2 min to 70 min in a mass range from 120 to 600. Each dataset was normalized to the largest bucket in the analysis.

4.5.6 Structure elucidation of pyrazines

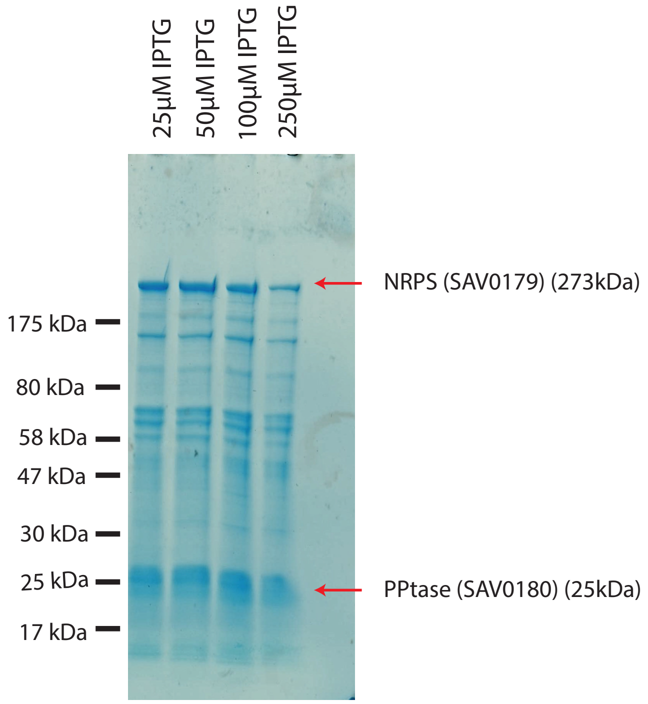
Compounds 3–5 were determined by 1D ^{1}H NMR and mass spectrometry. Compounds 3–5 had masses of *m/z* = 260.24, *m/z* = 276.22, and *m/z* = 292.21, respectively. Samples were dissolved in deuterated DMSO. δ (integration, multiplicity): Compound 3: 3.94 (H-2, s), 6.663 (H-4, d), 7.05 (H-4, 2), 8.43 (H-2, S), 9.22 (H-2, s). Compound 4: 3.951 (H-2, s), 4.08 (H-2, s), 6.66 (H-2, d), 7.05 (H-2, d), 7.19 (H-1, s), 7.266 (H-4, m), 8.45 (H-1, s), 8.49 (H-1, s), 9.22 (H-1, s). Compound 5: 3.94 (H-4, s), 6.660 (H4, d), 7.048 (H-4, d), 8.426 (H-2, s), 9.216 (H-2, s).

4.6 Supplementary Information

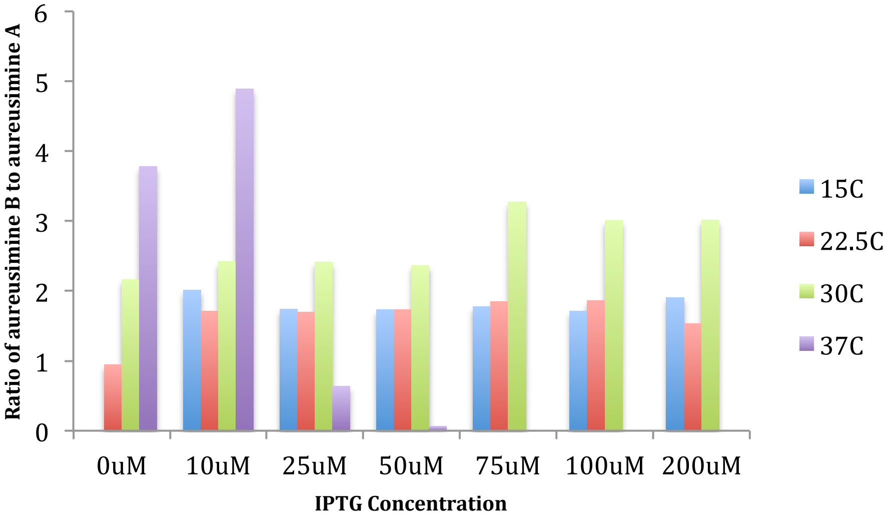


Supplemental Figure 4.1.

(a) Automated gel electrophoresis of *E. coli* + pAusAB. Lane 1: 0μM IPTG, grown at 37°C overnight. Lane 2: 50μM IPTG induction, grown at 28°C overnight. Bands are representative of absorption peaks (λ=280 nm) eluting off an automated electrophoresis capillary and compared to an internal standard. (b) Western blot of purified protein from *E. coli* + pAusAB using anti-histidine antibodies.



Supplemental Figure 4.2. Expression of nonribosomal peptide synthetase (NRPS/AusA) and phosphopantetheinyl transferase (PPtase/AusB) after induction with IPTG. *Escherichia coli* bearing the plasmid pAusAB was induced with IPTG at the concentrations indicated and protein was purified as described. Protein bands corresponding to SAV0179 (NRPS/AusA) and SAV0180 (PPtase/AusB) are indicated with arrows.



Supplemental Figure 4.3. Production ratios of aureusimine A and B at the induction concentration and temperature indicated.

4.6 References

Balibar CJ, Walsh CT. 2006. GliP, a multimodular nonribosomal peptide synthetase in *Aspergillus fumigatus*, makes the diketopiperazine scaffold of gliotoxin. Biochemistry **45**: 15029-15038

Fortin PD, Walsh CT, Magarvey NA. 2007. A transglutaminase homologue as a condensation catalyst in antibiotic assembly lines. Nature **448**: 824-827

Gruenewald S, Mootz HD, Stehmeier P, Stachelhaus T. 2004. In vivo production of artificial nonribosomal peptide products in the heterologous host Escherichia coli. Appl. Environ. Microbiol. **70**: 3282-3291

Grüschow S, Rackham EJ, Elkins B, Newill PLA, Hill LM, Goss RJM. 2009. New Pacidamycin Antibiotics Through Precursor-Directed Biosynthesis. ChemBioChem **10**: 55-360

Lambalot RH, Gehring AM, Flugel RS, Zuber P, LaCelle M, Marahiel MA, et al. 1996. A new enzyme superfamily - the phosphopantetheinyl transferases. Chem. Biol. **3**: 923-936

Lautru S, Deeth RJ, Bailey LM, Challis GL. 2005. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. Nat. Chem. Biol. **1**: 265-269

Magarvey NA, Beck ZQ, Golakoti T, Ding Y, Huber U, Hemscheidt TK, et al. 2006. Biosynthetic characterization and chemoenzymatic assembly of the cryptophycins. Potent anticancer agents from cyanobionts. ACS Chem. Biol. **1**: 766-779

Mertens N, Remaut E, Fiers W. 1995. Tight transcriptional control mechanism ensures stable high-level expression from T7 promoter-based expression plasmids. Biotechnology **13**: 175-179

Mootz HD, Marahiel MA. 1997. Biosynthetic systems for nonribosomal peptide antibiotic assembly. Curr. Opin. Chem. Biol. **1**: 543-551

Mootz HD, Schwarzer D, Marahiel MA. 2000. Construction of hybrid peptide synthetases by module and domain fusions. Proc. Natl. Acad. Sci. U.S.A. **97**: 5848-5853

Moran S, Rai DK, Clark BR, Murphy CD. 2009. Precursor-directed biosynthesis of fluorinated iturin A in *Bacillus* spp. Org. Biomol. Chem. **7**: 644

Pfeifer BA, Wang CC, Walsh CT, Khosla C. 2003. Biosynthesis of yersiniabactin, a complex polyketide-nonribosomal peptide, using *Escherichia coli* as a heterologous host. Appl. Environ. Microbiol. **69**: 6698-6702

Sambrook, J., Fritsch, T.M., and Maniatis, T. 1989. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, New York.

Schlumbohm W, Stein R, Ullrich C, Vater J, Krause M, Marahiel MA, et al. 1991. An active serine is involved in covalent substrate amino acid binding at each reaction center of gramicidin S synthetase. J. Biochem. **266**: 23135-23141

Schwarzer D, Finking R, Marahiel MA. 2003. Nonribosomal peptides: from genes to products. Nat. Prod. Rep. **20**: 275

Stachelhaus T, Mootz HD, Bergendahl V, Marahiel MA. 1998. Peptide Bond Formation in Nonribosomal Peptide Biosynthesis. Catalytic Role of the Condensation Domain. J. Biol. Chem. **273**: 22773-22781

Stachelhaus T, Mootz HD, Marahiel MA. 1999. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chem. Biol. **6**: 493-505

Stachelhaus T, Schneider A, Marahiel MA. 1995. Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. Science **269**: 69-72

Sun F, Li C, Sohn C, He C, Bae T. 2010. In the Staphylococcus aureus two component system sae, the response regulator SaeR binds to a direct repeat sequence and the DNA binding requires phosphorylation by the sensor kinase SaeS. J. Bacteriol. **192**: 2111-2122

Walsh CT, Chen H, Keating TA, Hubbard BK, Losey HC, Luo L, et al. 2001. Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines. Curr. Opin. Chem. Biol. **5**: 525-534

Walsh CT, Fischbach MA. 2010. Natural products version 2.0: connecting genes to molecules. J. Am. Chem. Soc. **132**: 2469-2493

Watanabe K, Hotta K, Praseuth AP, Koketsu K, Migita A, Boddy CN, et al. 2006. Total biosynthesis of antitumor nonribosomal peptides in *Escherichia coli*. Nat. Chem. Biol. **2**: 423-428

Watanabe K, Oikawa H. 2007. Robust platform for de novo production of heterologous polyketides and nonribosomal peptides in Escherichia coli. Org. Biomol. Chem. **5**: 593-602

Weist S, Bister B, Puk O, Bischoff D, Pelzer S, Nicholson GJ, et al. 2002. Fluorobalhimycin–a new chapter in glycopeptide antibiotic research. Angew. Chem. Int. Ed. Engl. **41**: 3383-3385

Wyatt MA, Wang W, Roux CM, Beasley FC, Heinrichs DE, Dunman PM, Magarvey NA. 2010. *Staphylococcus aureus* Nonribosomal Peptide Secondary Metabolites Regulate Virulence. Science **329**: 294-296

Zimmermann M, Fischbach MA. 2010. A Family of Pyrazinone Natural Products from a Conserved Nonribosomal Peptide Synthetase in Staphylococcus aureus. Chem. Biol. **17**: 925-930

Chapter 5. Genome Mining Complex Nonribosomal Peptide Synthetases

5.1 Chapter Preface

Using genomic predictions to identify novel NRPs is relatively facile for small NRPS assembly lines, however, the additional structural complexity arising from larger NRP gene clusters often creates molecules with even more interesting activities. The ability to apply the same genome mining principles to large clusters is necessary for the full illumination of all NRPs nature has to offer. In this project, the NRPS prediction rules are applied to a complex NRP assembly system within a gold associated bacterium to aid in the identification of a gold interacting NRP.

Many bacteria are known to excrete chemicals that bind metals for cellular processes and it was envisioned that the gold dwelling microbe, *Delftia acidovorans*, excretes similar substances that interact with gold that allow it to survive in such toxic environments. Again, using genome mining in conjunction with this hypothesis, we sought to identify directly whether a small molecule NRP was involved in gold biomineralizing and/or heavy metal resistance using a genome guided approach for NRP discovery.

Analysis of the *D. acidovorans* genome, identified a single NRPS-PKS hybrid gene cluster with several regulatory and tailoring genes that suggested the ‘cryptic’ NRP-PK was a metallophore and may be involved in gold biomineralization. Sequence analysis of the *D. acidovorans* gene cluster revealed 3 NRPS and 1 PKS gene comprising of 11 assembly-line modules, a significantly more complicated biosynthetic system compared with the dimodular aureusimine assembly line. The same principles used to identify aureusimine were applied to predict this much larger NRP and using a mass window based on this prediction, in conjunction with an in-house developed gold interaction assay, the cryptic NRP, delftibactin, was revealed. Delftibactin is secreted from *D. acidovorans* and interacts with soluble ionic gold to form solid gold nanoparticles. An NRPS knockout strain was constructed in *D. acidovorans* showing decreased fitness in the presence of toxic gold ions. The prediction of even large NRPS gene clusters provides us with the ability to use hypotheses to direct the discovery of interesting and potentially industrial important bacterial small molecules removing the randomness associated with traditional natural product discovery.

The following chapter is a modified version of a published article. I was a co-first author on this work with Chad Johnston. I contributed significantly to the conception, experimental design, execution and analysis of the research. I took major roles in the characterization of the delftibactin-gold complex including: compound purification, experimental design, characterization of the delftibactin-gold complex, survival assays, and HPLC-MS analysis of the complex. Chad's major roles included compound identification (delftibactin A and B), purification and genetic manipulation of *D. acidovorans*. Xiang Li carried out NMR structure experiments. Ashraf Ibrahim provided HR-MS for delftibactin A. Jeremiah Shuster and Gordon Southam provided initial insight into the delftibactin-gold complex and generated the electron microscope images. Nathan Magarvey provided valuable insight and advice throughout the project. The citation for this work is as follows:

Johnston, C.W.*, **Wyatt, M.A.***, Li, X., Ibrahim, A., Shuster, J., Southam, G., Magarvey, N.A. (2013) **Gold biomineralization by a metallophore from a gold-associated microbe.** *Nat. Chem. Biol.* **9**, 241–243 (2013) doi:10.1038/nchembio.1179
* contributed equally to this work.

5.2 Abstract

Microorganisms produce and secrete secondary metabolites to assist in their survival. We report that the gold resident bacterium *Delftia acidovorans* produces a secondary metabolite that protects from soluble gold through the generation of solid gold forms. This finding is the first demonstration that a secreted metabolite can protect against toxic gold and cause gold biomineralization.

5.3 Article

Microorganisms inhabit nearly all surfaces on the planet, an achievement typically attributed to their metabolic versatility. Frequently, secondary metabolic pathways and secreted products of these specialized branches of metabolism are complicit in an organisms' ability to capture niches, enhance fitness and overcome environmental stress and often have considerable industrial importance[1]. Metals represent a notable environmental condition for microbes, as some are required for growth (for example, Fe^{3+}), whereas others inhibit it (for example, Au^{3+}, Ag^{+} and Hg^{2+})[2]. Bacterial biofilms exist on the surface of gold nuggets[3, 4]; though soluble gold is inherently toxic2, these bacteria are implicated in its accumulation and deposition[5, 6]. The existence of bacterial biofilms coating gold nuggets and the discovery of bacterioform gold suggest that bacteria and specialized bacterial metabolic processes are involved in gold biomineralization[3, 4, 5, 6]. Sequencing gold nugget microbiota has revealed that *Cupriavidus metallidurans* and *Delftia acidovorans* are dominant organisms within such communities and comprise over 90% of these populations[4]. Investigations into *C. metallidurans* have revealed that it bioaccumulates inert gold nanoparticles within its cytoplasm as a mechanism to protect itself from soluble gold[5].

We sought to test whether *D. acidovorans* has any appreciable differences with *C. metallidurans* with respect to such gold biomineralization. An assay was developed to define whether the mechanism of biomineralization was extracellular or intracellular and to reveal mechanisms of how *D. acidovorans* protects itself from toxic soluble Au^{3+} and how these interactions may relate to gold deposition. We reasoned that if a cell-associated mechanism was predominant or exclusive, the bacteria would accumulate insoluble gold particles[7]; in contrast, if an extracellular gold reduction occurs at the cell surface or within the area surrounding microbial colonies, blackening would result because of gold reduction and the creation of solid gold particles. *D. acidovorans* and *C. metallidurans* were grown on agar plates and then flooded with solutions of Au(III), the dominant form of soluble gold found in terrestrial conditions[5, 8]. Following gold exposure, darkened zones developed around colonies of *D. acidovorans* but not *C. metallidurans* (Fig. 5.1a). These blackened zones suggested that *D. acidovorans* generated a diffusible metabolite that acts to generate reduced solid gold forms.

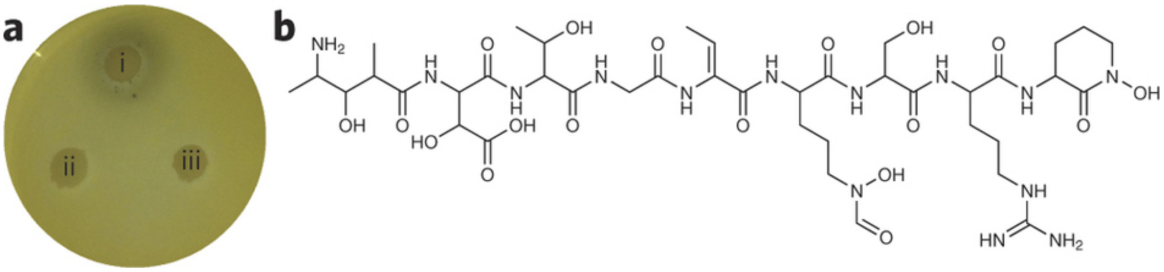


Figure 5.1.

(**a**) Gold resident bacteria *D. acidovorans* (wild type, i), the NRPS-null *D. acidovorans* mutant strain (Δ*delG*, ii) and *C. metallidurans* (iii) were grown for 3 d and overlaid with soft agarose containing 10 mM $AuCl_3$ for 2 h. Black halos are formations of gold nanoparticles. (**b**) Final structure of the gold-interacting nonribosomal peptide delftibactin.

We investigated the *D. acidovorans* genome for genes that may be associated with a unique small-molecule biosynthesis pathway that is absent from *C. metallidurans*. For example, polyketides and nonribosomal peptides are classes of secondary metabolites that bacteria use to promote environmental fitness[1] and include members that function to bind metals (for example, iron and copper)[9, 10, 11]. Indeed, our analysis identified a candidate nonribosomal peptide synthetase/polyketide synthase (NRPS/PKS) gene cluster (*Daci_4753*, *Daci_4754*, *Daci_4755*, *Daci_4756*, *Daci_4757*, *Daci_4758*, *Daci_4759* ; henceforth referred to as the Daci_4753–4759 or del cluster for an unknown secondary metabolite that, according to bioinformatic analysis[12] and *in silico* predictions, was expected to be a polar peptidic small molecule (Supplementary Results, Supplementary Fig. 1a)). Upstream, flanking these biosynthetic genes is a tripartite heavy metal efflux pump[13, 14] (*Daci_4763*, *Daci_4764*, *Daci_4765*; 68% identical and 83% similar to the CzcA-like HmyA heavy metal efflux pump from *C. metallidurans* CH34 (*Rmet_4123*)), perhaps supporting a role of this cluster being associated with gold detoxification. Downstream genes were associated with metallophores that bind iron (siderophores) and, specifically, genes for their reception and regulation[9]. To reveal whether the Daci_4753–4759 (del) cluster (Supplementary Fig. 5.1a) was associated with the observed gold precipitation, we constructed an insertional inactivation of the nonribosomal peptide synthetase gene (*Daci_4754*; referred to here as *delG*), and the resulting mutant strain was compared to the wild-type in the soluble gold exposure agar plate assay (Fig. 1a and Supplementary Fig. 5.1b). Unlike the wild-type colonies, colonies of the Δ*delG* strain were deficient in producing a blackening zone. To further reveal whether end products from this biosynthetic locus were solely responsible for the gold precipitation, we generated broth extracts of the entire *D. acidovorans* secreted metabolome, subjected the mixtures to chromatographic separations with LC/MS and eluted the separated contents into a 96-well plate. Within the water-soluble fractions, a select number of wells recapitulated the gold activity (Supplementary Fig. 5.2a). Well fractions capable of gold precipitation were analyzed further and were found to share a peptidic compound that closely matched the molecular weight of the predicted del nonribosomal peptide, which was absent in extracts from the Δ*delG* strain (Supplementary Fig. 5.2b) that lack gold-precipitating metabolites. This peptidic compound could be identified in supernatants in concentrations in excess of 200 μM (Supplementary Table 5.6), enabling its isolation and structure determination by high-resolution MS (Supplementary Fig. 5.3a) and NMR

(Supplementary Fig. 5.3b), which revealed a linear polyketide-nonribosomal peptide consistent with the structural prediction, which was named delftibactin (**1**; Fig. 5.1b). *D. acidovorans* environmental isolates were also screened for their ability to produce delftibactin, resulting in its identification in all tested isolates (Supplementary Fig. 5.4a,b).

Purified delftibactin was observed to coprecipitate with gold from solution, recapitulating the original findings in end-point assays (Supplementary Fig. 5.5a–c). We sought to address whether the gold precipitation caused by delftibactin confers a protective advantage to *D. acidovorans* and assists in ameliorating gold toxicity. In initial assays, this question was addressed with an acute toxic exposure of the wild-type and the Δ*delG* strains, whereby broth cultures of each were exposed for 30 min and the colony-forming units subsequently determined. The results of this assay showed that a 102.8-fold increase in sensitivity to gold toxicity could be observed in the Δ*delG* strain and that this increase could be rescued with exogenous addition of purified delftibactin (Fig. 5.2a). A detoxifying effect was observed in dose escalations of delftibactin to overtly toxic concentrations of $AuCl_3$ (Supplementary Fig. 5.6a), and subsequent examination revealed that although soluble gold is toxic at 10 μM, the blackened precipitate did not show any obvious toxicity when supplied in excess of 10 mM (data not shown). Metals found within secondary gold deposits have been outlined previously[4], specifically revealing that the concentration of iron is low relative to gold. However, we probed the fate of delftibactin when presented with equimolar concentrations of soluble gold and iron. This simultaneous exposure revealed that gold precipitation would proceed in the presence of high concentrations of iron (Fig. 5.2b). To assess what impacts this precipitation would have on *D. acidovorans* viability, we set up cultures of the Δ*delG* strain containing the resultant gold reactions. Growth curves demonstrate that although iron-free conditions are optimal for detoxification owing to metal competition, sufficient detoxification occurs in the presence of iron to support the growth of *D. acidovorans* (Fig. 5.2c). Chronic exposures were also tested, demonstrating that exogenous delftibactin addition to cultures of the Δ*delG* strain was sufficient to overcome chronic gold toxicity (Supplementary Fig. 5.6b). The results of these experiments, though not conducted in a natural context, may inform on the protective nature of delftibactin for *D. acidovorans*. We next performed an experiment aimed at revealing whether *D. acidovorans* may maintain protective extracellular concentrations of delftibactin by monitoring the loss of delftibactin by gold coprecipitation, leading to an increase in delftibactin production through a positive feedback mechanism[15]. *D. acidovorans* supernatants treated with 10 μM and 30 μM $AuCl_3$ caused delftibactin depletion, resulting in a compensatory increase in delftibactin concentrations compared to an untreated control, representing a form of reactive homeostasis (Supplementary Fig. 5.7). Delftibactin concentrations were also responsive to iron concentrations (Supplementary Table 5.6), indicating that delftibactin is most likely a siderophore that serves at least two purposes for this organism.

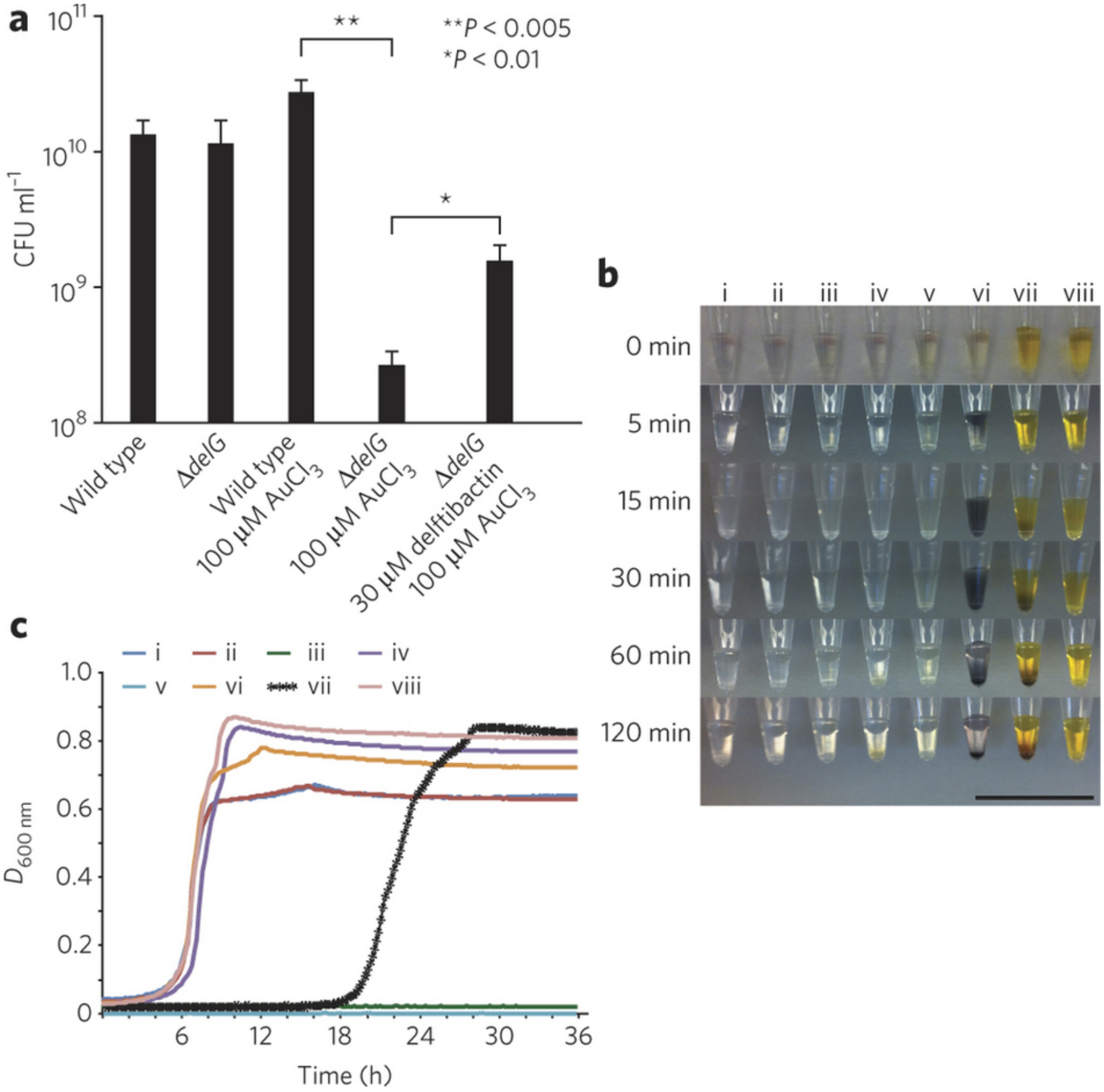


Figure 5.2.

(**a**) Delftibactin-null *D. acidovorans* shows increased sensitivity to gold toxicity and can be rescued by the addition of delftibactin. *D. acidovorans* wild-type and Δ*delG* cultures were grown for 48 h at 30 °C in deferrated Acidovorax complex medium (ACM). Cultures were incubated in the presence and absence of 100 μM $AuCl_3$ for 30 min, revealing increased sensitivity in the Δ*delG* strain. Addition of delftibactin (30 μM) to

Δ*delG D. acidovorans* ameliorated gold toxicity. Results are shown as mean ± s.d.; *n* = 4; Two-tailed student's *t*-test. CFU, colony-forming units. (**b**) Delftibactin is capable of precipitating gold in the presence of iron. Time course progression of 5 mM $AuCl_3$ reacted with: (i) water only, (ii) 5 mM delftibactin A, (iii) 5 mM $AuCl_3$, (iv) 5 mM $FeCl_3$, (v) 5 mM $AuCl_3$ + 5 mM $FeCl_3$, (vi) 5 mM delftibactin A + 5 mM $AuCl_3$, (vii) delftibactin A + 5 mM $AuCl_3$ + 5 mM $FeCl_3$ and (viii) delftibactin A + 5 mM $FeCl_3$. Scale bar, 20 mm. (**c**) Growth curves of *D. acidovorans* Δ*delG* in the presence of each reaction mixture shown in **b** at a final concentration of 30 µM in ACM. Results are a mean of three growth curves for each condition from a single representative experiment.

Metallophores are recognized to create complexes with metals, and we wished to reveal whether such complexation was part of the gold-delftibactin interaction. As the gold-delftibactin association leads to coprecipitation and formation of an insoluble material, we examined how delftibactin may bind metals using gallium. NMR analysis of the delftibactin–gallium complex showed the coordinating activity of delftibactin. These results indicated that N^{δ}-hydroxy-N^{δ}-formylornithine, the polyketide-extended portion of the *N*-terminal alanine, and cyclic N^{δ}-hydroxyornithine form ligands for metal binding (Fig. 5.3a). This complexation is relevant to gold, as purified gallium–delftibactin exposed to gold showed considerably decreased precipitation (Supplementary Fig. 5.8). Although the gallium-gold competition may inform on how gold interacts with delftibactin, we next sought direct evidence of gold binding by delftibactin. We identified an initial gold–delftibactin complex through MS and confirmed its identity through diagnostic MS/MS fragmentation (Supplementary Fig. 5.9). To reveal in more depth the mechanisms that lead to gold precipitation and which sites within delftibactin are associated with gold complexation, we made use of natural delftibactin variants. Several compounds were observed to elute at a similar time to delftibactin and had comparable fragmentation patterns and masses, indicating that they may be structural analogs that would be useful if they had modifications within the proposed chelation core. One promising candidate, bearing a hydroxylated and acetylated ornithine (delftibactin B (**2**) *m+/z* = 1,047; Supplementary Fig. 5.10a), was identified and subsequently characterized. Subsequent examination of the complexing properties and protective nature of delftibactin B indicated that it was less efficient in gold reduction than delftibactin (renamed delftibactin A; Supplementary Fig. 5.10b), resulting in decreased detoxification (Supplementary Fig. 5.10c). Exposing delftibactin A and delftibactin B to $AuCl_3$ leads to their depletion; new peaks, however, emerge following the exposure, with the predominant one at *m+/z* = 989 (Fig. 5.3b); the molecules at this molecular weight continue to react with gold and are also lost from solution (Fig. 5.3b and Supplementary Fig. 5.11). Structural characterization indicates that this reaction product does not bear the ornithine modifications observed in delftibactin A and B (Supplementary Fig. 5.12), indicating that delftibactin can chelate gold and also react with it. This observation most likely explains why delftibactin B is less protective, as it has a ketone moiety that is less easily oxidized than the aldehyde found on delftibactin A, which may be partially responsible for gold reduction. Transmission electron microscopy was used to better assess the nature of the gold precipitate and revealed an abundance of colloidal gold nanoparticles and octahedral gold platelets (Fig. 5.3c). Such solid gold forms are authentic morphologies found in gold nuggets and bacterioform gold[6, 16]. These data show that pure delftibactin A is capable of creating naturally occurring complex gold structures from Au^{3+} on short timescales (seconds) at room temperature and neutral pH and at rates that far exceed those observed for traditional gold nanoparticle–producing agents such as citrate[17] (Supplementary Fig. 5.13), providing a potential mechanism for bacterial gold biomineralization. We propose that delftibactin facilitates this biomineralization and protects *D. acidovorans* by chelating soluble Au^{3+} and directly precipitating it as a complex or by binding and reducing gold through oxidative decarboxylation before chelating a second Au^{3+} ion and precipitating as a complex.

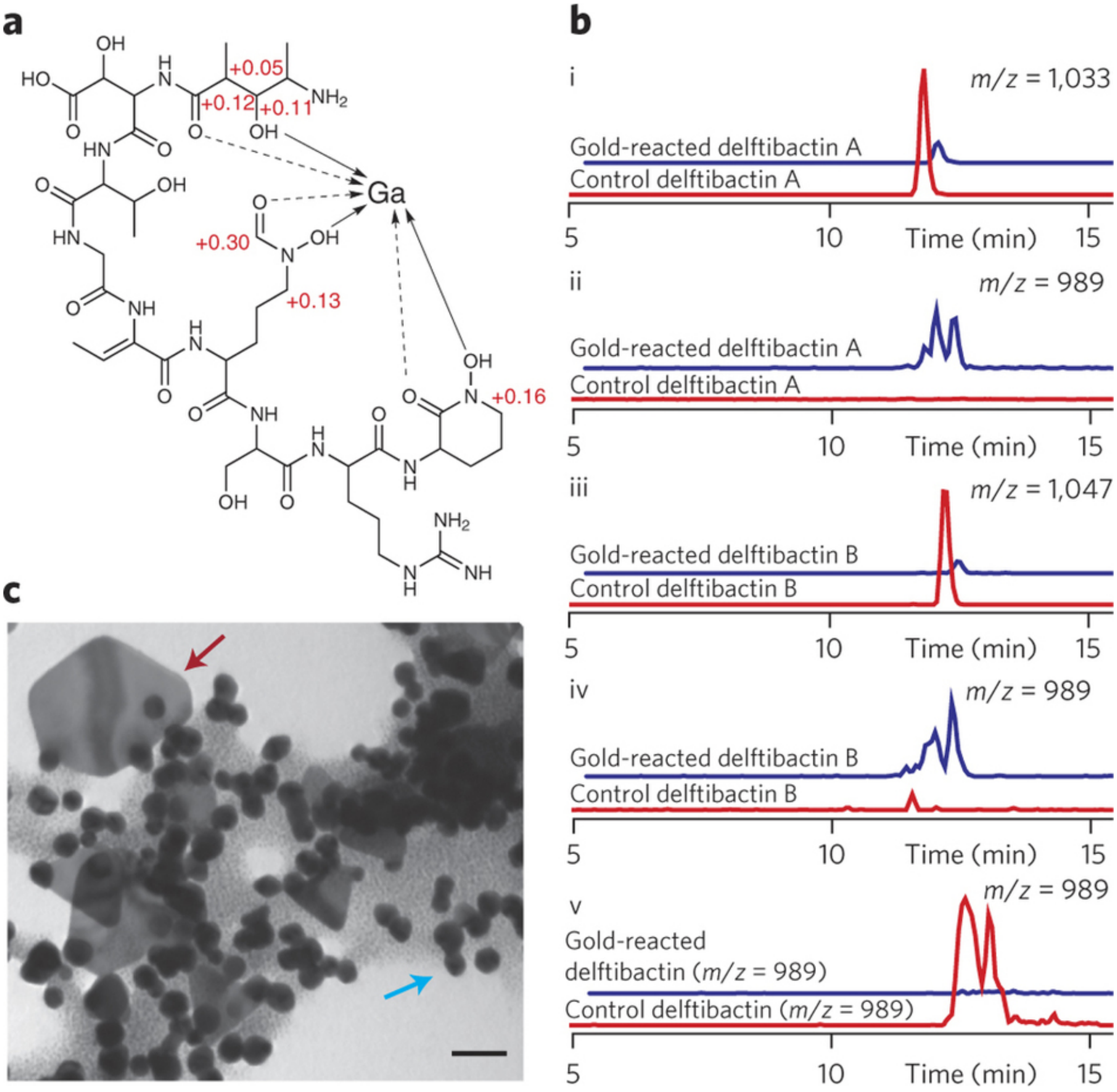


Figure 5.3.

(**a**) Gallium NMR confirms that delftibactin has a single metal-binding site. (**b**) Chelation core modification affects the rate of gold precipitation but does not affect the generation of a common transient intermediate. Delftibactin A and B were reacted with equimolar (5 mM) $AuCl_3$ for 2 h before reactions (blue) were analyzed by LC/MS and compared to an unreacted control (red). Extracted ion chromatograms show a depletion of delftibactin A (i) and B (iii) following incubation with $AuCl_3$, accompanied by the emergence of a new delftibactin species (*m*/*z* = 989; ii, iv) that could be further reacted with $AuCl_3$ and depleted from solution (v). (**c**) TEM of delftibactin–gold (2:1) complex after 10 min reveals the presence of colloidal and octahedral gold nanoparticles, reminiscent of those seen in natural deposits. Blue arrow, colloidal gold. Red arrow, octahedral gold. Scale bar, 50 nm.

Collectively, these results indicate that although delftibactin is dispensable in culture, consistent with other secondary metabolites, it has an important role in protecting its gold-resident producer from toxic soluble gold. Further, we have shown that gold biomineralization can take place through the secretion of this metallophore from a gold resident bacterium. This phenomenon echoes situations observed previously including boron chelation by vibroferrin[18] and copper chelation by yersiniabactin[11] and methanobactin[10], wherein bacterial siderophores have dual physiological roles that are important in their environments. Delftibactin seems to be what is to our knowledge the first example of a co-opted metallophore that protects its producer from toxic soluble gold and provides a mechanism for bacterial gold biomineralization.

5.4 Materials and Methods

5.4.1 General experimental procedures.

One-dimensional (^1H and ^{13}C) and two-dimensional (^1H-^{13}C and ^1H-^{15}N HMBC, HSQC, NOESY and COSY) NMR spectra were recorded on a Bruker AVIII 700 MHz NMR spectrometer in D_2O (D_2O; Cambridge Isotope Laboratories). High-resolution MS spectra were collected on a Thermo LTQ OrbiTrap XL mass spectrometer (ThermoFisher Scientific, USA) with an electrospray ionization source (ESI) and using CID with helium for fragmentation. LC/MS data was collected using a Bruker AmazonX ion trap mass spectrometer coupled with a Dionex UltiMate 3000 HPLC system, using a Luna C18 column (250 mm × 4.6 mm, Phenomenex) for analytical separations, running acetonitrile and ddH2O as the mobile phase.

5.4.2 Bacterial strains.

Delftia acidovorans was ordered from the German Resource Centre for Biological Material (DSMZ, DSM no. 39). *Delftia acidovorans* was cultured on Acidovorax complex medium[19] (ACM) plates at 30 °C. The Δ*delG* strain was initially grown in the presence of 30 µg/mL tetracycline. Environmentally isolated *D. acidovorans* strains D27L and D126L were found in soil samples collected around McMaster University from June–August 2010. Environmental isolates were identified as *D. acidovorans* strains based on 16S sequence alignment, using 16S sequences that were amplified from single colonies using the universal 16S primers[20]: 27f (AGAGTTTGATCMTGGCTCAG) and 1525r (AAGGAGGTGATCCAGCC).

5.4.3 Gold precipitation on agar plates.

Wild-type and Δ*delG Delftia acidovorans* were streaked onto a Chelex-treated (deferrated) ACM agar plate and grown for 3 d at 30 °C. The plate was then overlaid with 10 mL of 0.5% agarose containing 10 mM $AuCl_3$. Gold complexing comparison to other bacteria was carried out as follows: 10 μL of an overnight culture of *D. acidovorans*, *D. acidovorans* Δ*delG* or *C. metallidurans* were placed onto deferrated ACM plates and grown for 3 d at 30 °C. The plates were overlaid with $AuCl_3$ as described above. Agar plate overlay images were taken after 2 h of incubation at room temperature.

5.4.4 *D. acidovorans* 96-well plate gold bioassay.

After brief centrifugation to remove particulates, 100 μL of the *D. acidovorans* HP20 extract was loaded onto a Waters Alliance 2695 separations module HPLC equipped with a photodiode array and fractionated into a 96-deep well plate, collecting 96 fractions starting at 2 min and finishing at 56 min. Fractions were obtained approximately every 30 s. The mobile phase was curved (curve 8) from 5% acetonitrile, 95% water at 2 min to 80% acetonitrile at 45 min at a flow rate of 3 mL/min. Plates were dried overnight in a GeneVac HT4 series 2 and resuspended in 60 μL of ddH_2O, and 25 μL were placed in fresh plates along with 25 μL 10 mM $AuCl_3$ and left to react at room temperature for 30 min.

5.4.5 Identification of delftibactin biosynthetic gene cluster and adenylation domain specificity.

Delftibactin genes encoding NRPS and PKS were identified using the BLAST function of IMG (http://img.jgi.doe.gov/), using the sequence of *pksJ* as a query. Adenylation domain specificities were assessed using NRPS Predictor[21] or NRPS-PKS[22], and the ten residue codes[12] of each entry and its top scoring hit were recorded. For the alignment of the adenylation domains specific for hydroxylated ornithine, the delftibactin adenylation code and the vicibactin adenylation code were determined with NRPS Predictor[21] and aligned manually as neither database contained domains with homologous sequences.

5.4.6 Construction of the Δ*delG D. acidovorans* strain.

All primers and plasmids used in this process are described in Supplementary Table 5.5. If not stated explicitly, genetic manipulations and molecular biology techniques followed those from Cold Spring Harbor Protocols, available at http://www.molecularcloning.com/.

A knockout plasmid for *D. acidovorans* was constructed by inserting a 2-kb PCR product of *delG* (primers 2kbNRPS2Xba2 and 2kbNRPS2Sac2) into pUC19 using Xba1 and Sac1 digest sites, ligating with T4 ligase, transforming into chemically competent DH5α (Invitrogen) and plating on LB medium with 100 μg/mL ampicillin. Positive clones were identified by colony PCR with 2kbNRPS2Xba2 and 2kbNRPS2Sac2 and verified through digestion following an overnight growth and plasmid miniprep using a QIAprep Spin Miniprep Kit (Qiagen). A clone containing a 2-kb insert was digested with Not1 to cut in the middle of the 2-kb insert, treated with calf intestinal phosphatase (CIP) and gel extracted to remove remaining CIP. A tetracycline resistance cassette was amplified from pLLX13 (primers TetNotF and TetNotR), purified, digested with Not1 and ligated with the digested vector. This ligation was transformed into chemically competent DH5α (Invitrogen) and plated on LB medium with 100 μg/mL ampicillin and 10 μg/mL tetracycline. Positive colonies were confirmed by PCR with TetNotF and TetNotR and verified with digestion with both Xba1 and Sac1 or with Not1, following an overnight growth and plasmid miniprep. The resulting plasmid was modified further by digesting with HindIII, treating with CIP and gel extracting to remove remaining CIP. An oriT was amplified from pLLX13 with primers OriTF and OriTR, purified, digested with HindIII and ligated with the cut vector before transforming into chemically competent DH5α and plating on LB medium with 100 μg/mL ampicillin and 10 μg/mL tetracycline. Positive colonies were identified with PCR using OriTF and OriTR and verified by digestion following an overnight growth and plasmid miniprep. The final plasmid (pDEL19) was transformed into chemically competent *E. coli* ET12567 carrying the helper plasmid pUZ8002, plating on LB medium with 10 μg/mL tetracycline, 30 μg/mL chloramphenicol and 25 μg/mL kanamycin. Positive transformants were grown in 3 mL LB medium with antibiotics overnight at 37 °C, alongside a 3-mL ACM growth of *D. acidovorans* at 30 °C. One milliliter of the donor *E. coli* and 1 mL of the recipient *D. acidovorans* were centrifuged separately and washed twice with fresh LB medium. Cells were resuspended in 1 mL LB medium and mixed 1:1, and 300 μL was dispensed on nutrient agar plates and left to grow overnight at 30 °C. Cells were scraped and resuspended in 3 mL LB medium, plating 50 μL on LB plates containing 30 μg/mL tetracycline and 100 μg/mL apramycin to remove *E. coli*. Colonies were observed and tested for growth in LB medium with antibiotics at 30 °C and 200 r.p.m., with viable cultures streaked on LB plates with 30 μg/mL tetracycline. Colony PCR with TetNotF and TetNotR was used to confirm the presence of the tetracycline cassette. Chromosomal integration of the tetracycline cassette was confirmed by PCR with TetNotR and NRPS2Seq2 primers.

5.4.7 16S alignment and delftibactin production in environmental strains.

Environmental isolates from around the McMaster University campus were identified as *D. acidovorans* strains on the basis of 16S sequence alignment, using 16S sequences that were PCR amplified from single colonies (see above). Using these sequences and the 16S sequence for the *D. acidovorans* genome strain SPH1 from GenBank, along with the sequence for a gold biofilm isolate of *D. acidovorans* (accession number: GU013673)4 were aligned with Geneious software version 4.8.5 (http://www.geneious.com/), using a Tamura-Nei genetic distance model, a neighbor-joining tree building method featuring a global alignment with free end gaps. Isolated strains were grown for 3 d at 30 °C and 190 r.p.m. in 1 L of ACM that had been treated with Chelex100 resin to limit the iron concentration. Cultures were centrifuged at 7,000 r.p.m. to remove the cell mass, and supernatants were treated with 20 g/L washed HP20 resin (Dialon). After 1 h of shaking with the supernatant, HP20 was collected by Buchner funnel vacuum filtration and eluted with 400 mL of methanol. The resin eluent was evaporated to dryness, resuspended in 50% methanol and water and injected into a Waters AutoPure LC/MS using a similar method as above. MassLynx software was used to generate the 1,033-*m/z* extracted ion chromatograms for each extract. Fragmentation of these compounds was carried out on a Bruker AmazonX ion trap mass spectrometer.

5.4.8 Delftibactin–Au(III) precipitation measurements.

The gold-delftibactin interaction was determined through two separate experiments. First,

$AuCl_3$ was held constant at 2.5 mM, and the interaction with delftibactin was monitored by measuring the absorption of $AuCl_3$ remaining in solution after precipitation by delftibactin through comparison with a standard curve. Briefly, 2.5 mM $AuCl_3$ was incubated with 5 mM, 2.5 mM, 1.25 mM, 0.6125 mM, 0.3063 mM and 0.1531 mM delftibactin for 1 h. Solutions were filtered with a 0.22-µM Acrodisc (Pall) to remove insoluble delftibactin-gold precipitate. one hundred microliters of the filtered reaction were placed in a 96-well plate, and the absorbance was read at 300 nm using a SpectraMax 384 Plus (Molecular Devices). Absorption was compared to a $AuCl_3$ standard curve to determine the concentration of $AuCl_3$ remaining in solution. To monitor the amount of delftibactin remaining in solution after reaction with gold, a similar experiment was conducted. Delftibactin (2.5 mM) was incubated with 5 mM, 2.5 mM, 1.25 mM, 0.6125 mM, 0.3063 mM and 0.1531 mM $AuCl_3$ for 1 h. Solutions were filtered similar to above. Reaction mixtures were analyzed using a Waters Alliance 2695 RP-HPLC separations module, equipped with a Waters 2998 photodiode array and a Luna 5u C18 column (250 × 4.60 mm, Phenomenex). The mobile phase was linear from 2% acetonitrile, 98% water + 5 mM $(NH_4)_2CO_3$ at 2 min to 14% acetonitrile at 18 min at a flow rate of 1 mL/min. The UV peak associated with delftibactin (*T*r = 12.29 min) was integrated and compared to a standard curve.

5.4.9 Transmission electron microscopy of delftibactin–Au(III) complexes.

Delftibactin was reacted with $AuCl_3$ with a molar ratio equal to 2:1 for 10 min. Each separate reaction of delftibactin with $AuCl_3$ was examined using a Phillips CM-10 transmission electron microscope operating at 80 kV. The whole-mount sample was absorbed and dried on a formvar-carbon–coated 100-square mesh copper grid and rinsed with filter-sterilized, de-ionized water to remove any salt precipitates.

5.4.10 Gallium-delftibactin–gold interaction.

Gallium-bound delftibactin was adjusted to 10 mM and mixed 1:1 with an equimolar solution of $AuCl_3$, alongside purified delftibactin and water (control). The reaction mixture was monitored at room temperature for 30 min.

5.4.11 Gold detoxification by delftibactin.

The assay was set up as follows: 50 µL of ddH2O containing 3.2 mM, 1.6 mM, 1.4 mM, 1.2 mM, 1.0 mM, 0.8 mM, 0.6 mM, 0.4 mM and 0.2 mM delftibactin was added in quadruplicate to wells within a 96-well plate. A 1.6-mM stock solution of AuCl3 was made, and 50 µL was added to each well containing delftibactin. No-delftibactin and no-$AuCl_3$ controls containing only water were also added to the 96-well plate in quadruplicate. These were incubated for 30 min at room temperature, during which time 10 mL of an overnight culture of *D. acidovorans* grown in ACM was centrifuged and resuspended in 5 mL sterilized ddH2O. After 30 min incubation of AuCl3 with delftibactin, 100 µL of concentrated culture was added to each well. Final concentration of $AuCl_3$ was 400 µM, and the final concentrations of delftibactin were 800 µM, 400 µM, 350 µM, 300 µM, 250 µM, 200 µM, 150 µM, 100 µM and 50 µM. After 30 min of incubation at room temperature, mixtures were serially diluted and plated onto nutrient agar plates and incubated at 30 °C. Colonies were counted after 24 h of growth. Results are shown as mean ± s.d.; $n = 4$. To assess whether *D. acidovorans* could grow in the presence of the gold precipitate, several milligrams of delftibactin and $AuCl_3$ were reacted 1:1 overnight, centrifuged and washed once with water to concentrate the precipitate. The precipitate was resuspended in ddH2O at a final concentration of 100 mM, calculated using a molecular weight of 1,227 g/mol, corresponding to a gold-delftibactin species. *D. acidovorans* was grown overnight in a 96-well plate in 100 µL of ACM containing 20 µM–10 mM gold precipitate or $AuCl_3$. No growth was observed in any well containing AuCl3, whereas full growth was observed in every well containing the corresponding amount of precipitate. We have determined the MIC of $AuCl_3$ to be roughly 10 µM.

5.4.12 Gold detoxification in chronic exposure by delftibactin in presence and absence of iron.

Twenty-microliter reactions were set up as follows: (i) water only, (ii) 5 mM delftibactin, (iii) 5 mM $AuCl_3$, (iv) 5 mM $FeCl_3$, (v) 5 mM $AuCl_3$ + 5 mM $FeCl_3$, (vi) 5 mM delftibactin B + 5 mM $AuCl_3$, (vii) delftibactin B + 5 mM AuCl + 5 mM $FeCl_3$, (viii) delftibactin B + 5 mM $FeCl_3$. Reactions were initiated by the addition of delftibactin, and images were taken at the time points as indicated in Figure 5.2b and Supplementary Figure 5.10. After 2 h, reactions were serially diluted to a 30-µM final concentration in ACM containing *D. acidovorans* Δ*delG* diluted 1:1,000 from an overnight culture. Optical density was monitored using a TECAN Sunrise microplate reader at 600 nm for

36 h. Results are a mean of three growth curves for each condition from a single representative experiment. As a second test of delftibactin protective capacity, *D. acidovorans* Δd*elG* cells from an overnight culture were inoculated 1:1,000 into 100 µL ACM in a 96-well plate containing 0 µM or 10 µM $AuCl_3$ and then provided 0 µM or 100 µM delftibactin. Cultures were grown for 84 h at 250 r.p.m. at 30° in a TECAN Sunrise microplate reader and measured at 600 nm to assess growth. Results are a mean of three growth curves for each condition from a single representative experiment.

5.4.13 Gold protective comparison of delftibactin A and B.

Twenty-microliter reactions were set up as follows: water only, 5 mM $AuCl_3$, 5 mM $AuCl_3$ + delftibactin A, and 5 mM $AuCl_3$ + delftibactin B. Reactions were initiated with the addition of delftibactin A or B. After 2 h, reactions were serially diluted to 125 µM in ACM containing *D. acidovorans* Δ*delG* diluted 1:1,000 from an overnight culture. Optical density was monitored using a TECAN Sunrise microplate reader at 600 nm for 36 h. Results are a mean of three growth curves for each condition from a single representative experiment.

5.4.14 Delftibactin-mediated protection against gold toxicity.

Cultures of *D. acidovorans* wild type and *D. acidovorans* Δ*delG* were grown in deferrated ACM for 2 d at 30 °C. In 96-well plates, 200 µL of wild-type or mutant grown culture were incubated in the presence and absence of 100 µM $AuCl_3$. At the same time, the mutant culture was also complemented with 30 µM delftibactin (biological concentration) in the presence of 100 μM $AuCl_3$. After 30 min, cultures were serially diluted in water and plated on LB agar to determine the colony-forming units of *D. acidovorans* after gold exposure. Results are shown as mean ± s.d.; *n* = 4; Two-tailed student's *t*-test.

5.4.15 Measuring delftibactin production following depletion by gold.

Cultures of *D. acidovorans* were grown in 10 mL ACM in 50-mL Falcon tubes for 48 h, at which point both growth and delftibactin production had ceased. Cultures were then pelleted by centrifugation at 4,500 r.p.m. for 30 min at 4 °C. Supernatants were kept separate and moved into labeled, sterile 50-mL Falcon tubes while cell pellets were kept on ice. Initial delftibactin concentrations were assessed by filter sterilizing and then placing 400 μL of each supernatant into a HPLC sample vial, storing at 4 °C. Supernatants were then adjusted to 0 μM, 10 μM or 30 μM $AuCl_3$ with the appropriate volume of 10 mM $AuCl_3$ and left to react at room temperature. After 12 h, supernatants were returned the appropriate cell pellets and resuspended by vortexing before taking a second 400-μL sample to assess delftibactin depletion by gold treatment. Cultures were returned to the incubator and left shaking for another 48 h before a final sample was taken. Delftibactin concentrations were assessed by MRM-LC/MS for delftibactin, with washes between each sample. Values of integrated delftibactin peaks were normalized to the untreated control and represent the percent increase in delftibactin concentration (± propagated error) observed 48 h following gold precipitation; *n* = 6.

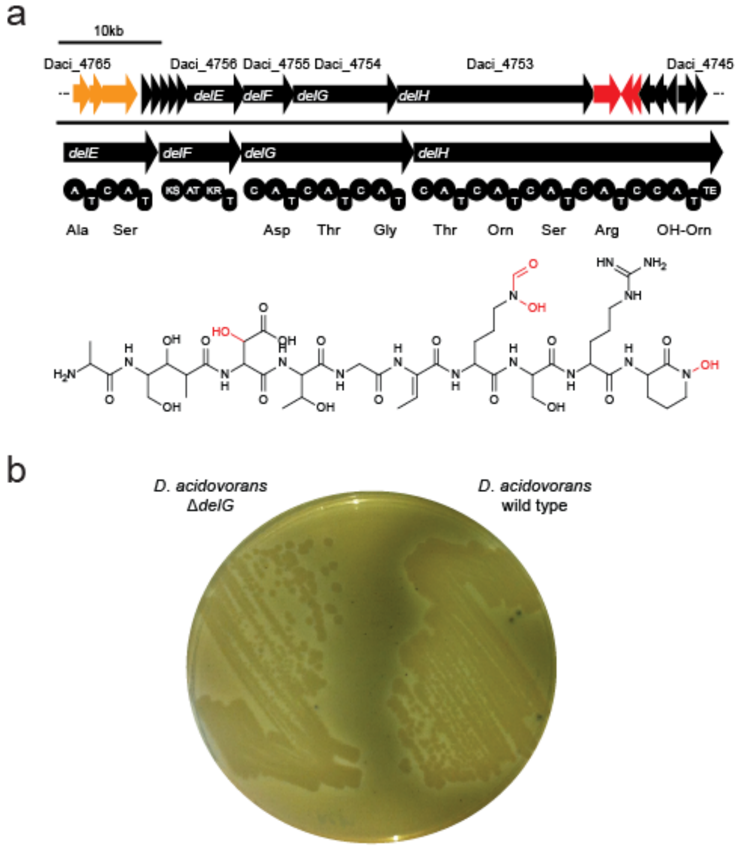
5.4.16 MRM-LC/MS measurement of delftibactin production in response to iron.

Cultures of *D. acidovorans* were grown for 48 h in 10-mL cultures of deferrated ACM resupplied with $FeCl_3$ in varying concentrations. Iron concentrations and corresponding delftibactin concentrations in the filter-sterilized supernatants are listed in Supplementary Table 5.6. Delftibactin concentrations were established using MRM-LC/MS. Results are shown as mean ± s.d.; $n = 3$.

5.4.17 Citrate-gold and delftibactin-gold comparison.

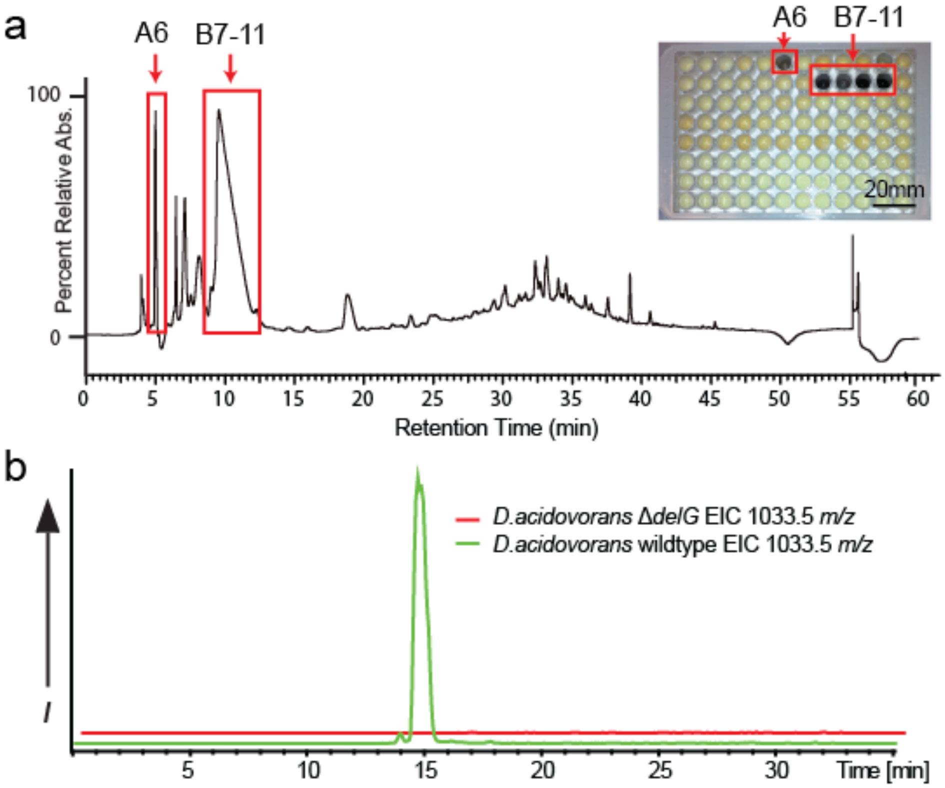
Stock 10-mM solutions of sodium citrate and purified delftibactin were mixed with an equimolar solution of $AuCl_3$ and allowed to react at room temperature in 1.5-mL Eppendorf tubes. Photographs were taken from the initial addition of gold to 1-h exposure. Similarly, TEM experiments were performed by mixing 10-mM stock solutions of $AuCl_3$ and either delftibactin or sodium citrate 1:1 on a formvar-carbon–coated 100-square mesh copper grid, imaging after 10 min of $AuCl_3$ exposure.

5.5. Supplementary Figures



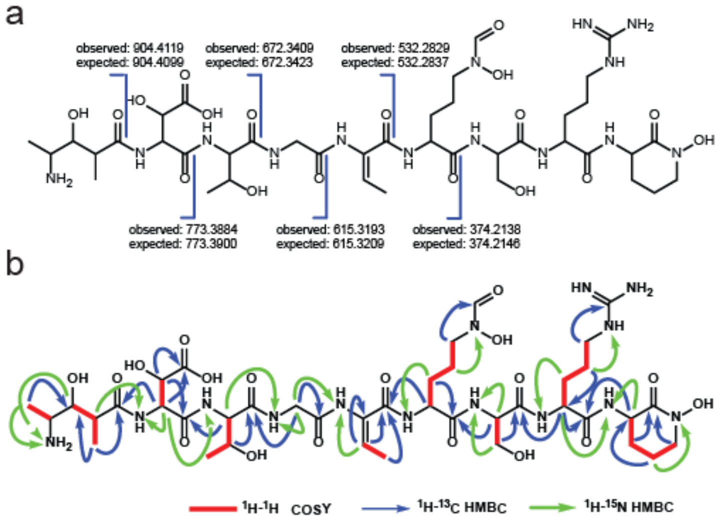
Supplementary Figure 5.1. *Delftia acidovorans* possesses a PKS/NRPS gene cluster associated with extracellular gold precipitation.

A) *del* gene cluster and domain architecture of NRPS-PKS hybrid assembly-line is shown; consisting of adenylation (A), thiolation (T), condensation, (C), ketosynthase (KS), acyltransferase (AT), ketoreductase (KR), and thioesterase domains. Flanking genes for heavy metal resistance (orange) and iron metabolism (red) are shown. Predicted activated amino acids are indicated below their respective A domains (see supplementary table 3). The final predicted structure of the unknown *del* metabolite is shown. B) *D. acidovorans* Δ*delG* does not produce the gold precipitate halo**.** *D.* acidovorans wildtype and Δ*delG* grown for 3 d at 30°C on ACM agar, overlaid with 0.5% agarose containing 10 mM $AuCl_3$ for 2 h.

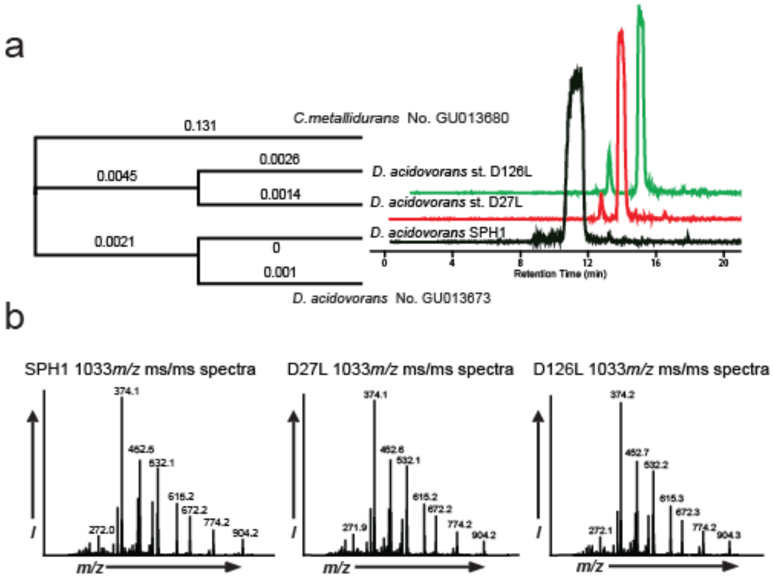


Supplementary Figure 5.2. Gold precipitation is caused by a metabolite encoded by the *del* gene cluster.

a) Metabolites from *D. acidovorans* cultures were extracted, separated by HPLC into a 96-well plate, and reacted with 5 mM $AuCl_3$. Blackening indicates gold nanoparticle formation. Active wells with corresponding UV peaks are highlighted and were found to contain a common peptidic metabolite. b) Extracts of wildtype and Δ*delG D.acidovorans* analyzed by LCMS. The extracted ion chromatogram of the wildtype specific compound associated with gold precipitation is shown.

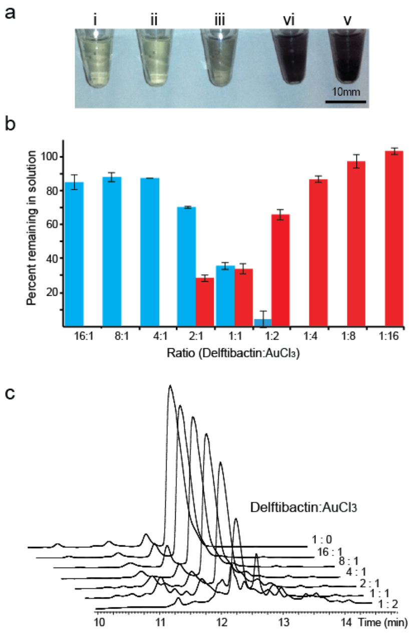


Supplementary Figure 5.3. Structural characterization of delftibactin.
a) High resolution mass fragmentation and b) 2D NMR spin systems for ^{1}H-^{13}C HMBC, ^{1}H-^{1}H COSY and ^{1}H-^{15}N HMBC for delftibactin in D_2O.



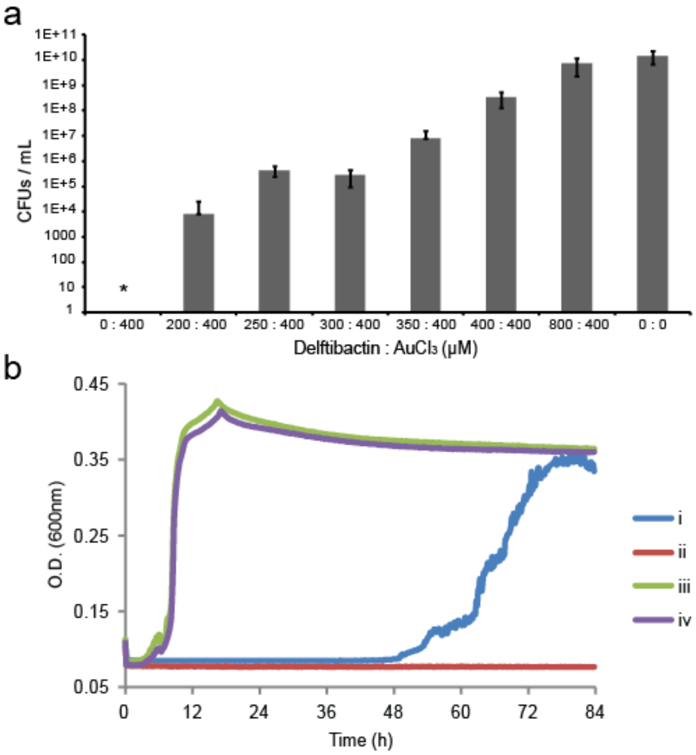
Supplementary Figure 5.4. Environmental isolates of *D. acidovorans* also produce delftibactin.

a) A 16S rDNA phylogenetic tree of *D. acidovorans* environmental isolates (D126L and D27L), genome strain (SPH-1), and sequences from samples associated with gold nugget biofilms, with respective extracted ion chromatogram (m^+/z = 1033.5) from HP-20 extracts (see supplementary methods). b) Fragmentation pattern of m^+/z = 1033.5 (delftibactin) from SPH-1, D27L, and D126L.



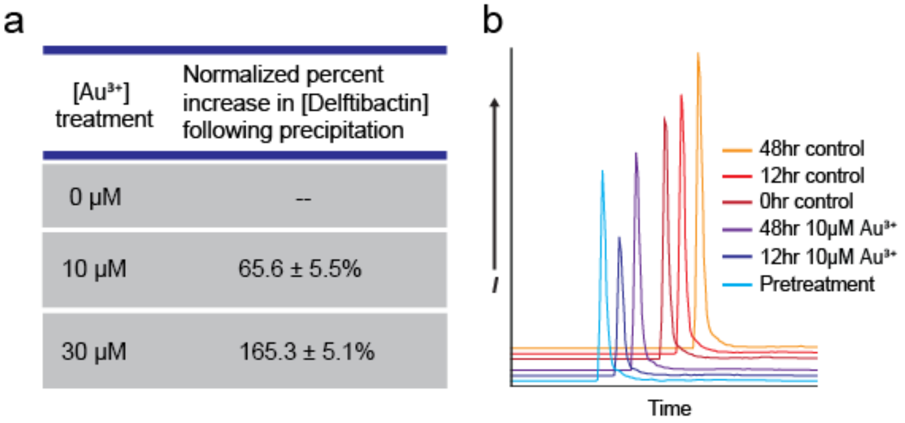
Supplementary Figure 5.5 Delftibactin and $AuCl_3$ co-precipitate from solution. a) Increasing concentrations of delftibactin in the presence of 2.5 mM $AuCl_3$ cause an increase in gold nanoparticle formation. Images were taken 30 min after the addition of delftibactin in the following concentrations: i) 0.3125 mM ii) 0.625 mM iii) 1.25 mM iv) 2.5 mM and v) 5 mM. b) Analysis of delftibactin-$AuCl_3$ reaction supernatants. Solutions of 2.5 mM $AuCl_3$ were reacted with 1:8, 1:4, 1:2, 1:1, and 2:1 equivalents of delftibactin for 30 minutes. Delftibactin remaining in solution was determined by integration at 220 nm (blue) by HPLC. The amount of gold remaining in solution was measured at 300 nm (red) by UV absorbance spectrometry. Results are shown as mean ± s.d; n = 3. c)

Delftibactin depletion in representative HPLC chromatograms of delftibactin-$AuCl_3$ supernatants.



Supplementary Figure 5.6. Delftibactin detoxifies $AuCl_3$ and enables growth under chronic gold exposure.

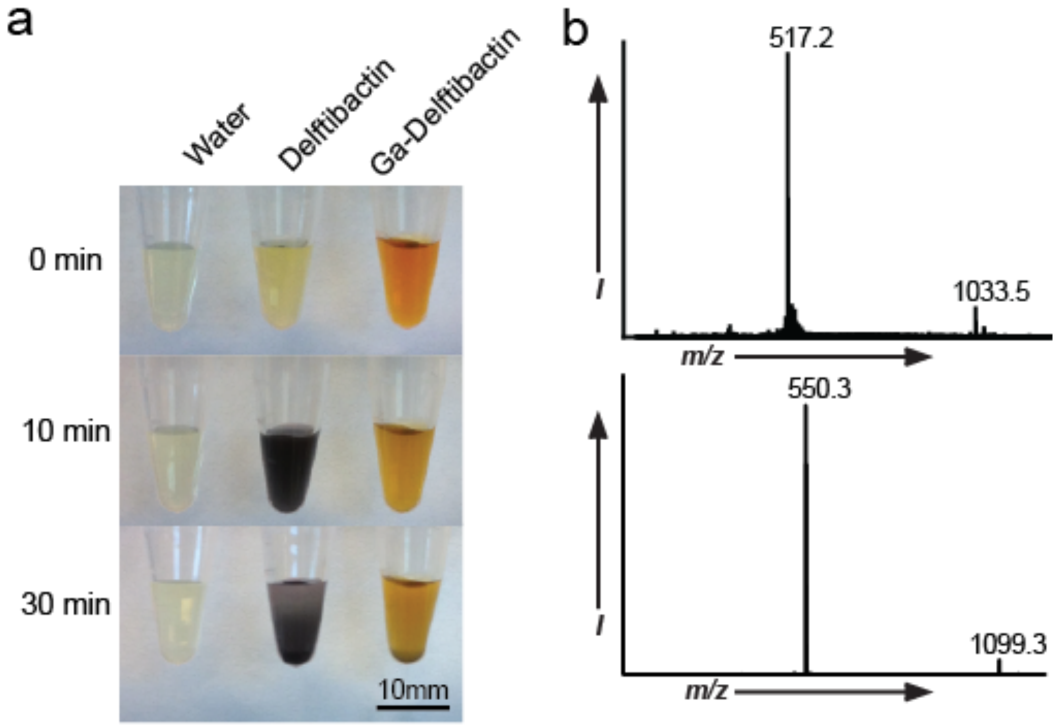
a) *Delftia acidovorans* colony forming units (CFUs) after exposure to 400 µM $AuCl_3$ for 30 minutes. Delftibactin was added to $AuCl_3$ solutions as indicated before exposing to *D. acidovorans.* *No survival was observed without the addition of delftibactin. Results are shown as mean ± s.d.; n = 4. b) Growth curves of *D. acidovorans* Δ*delG* cultures inoculated 1:1000 into ACM followed by the addition of delftibactin and/or gold as follows: i) 100 μM Delftibactin + 10 μM $AuCl_3$, ii) 10 μM $AuCl_3$ only, iii) 10 μM delftibactin only, and iv) water only. Results are a mean of three growth curves for each condition from a single representative experiment.



$[Au^{3+}]$ treatment	Normalized percent increase in [Delftibactin] following precipitation
0 μM	--
10 μM	65.6 ± 5.5%
30 μM	165.3 ± 5.1%

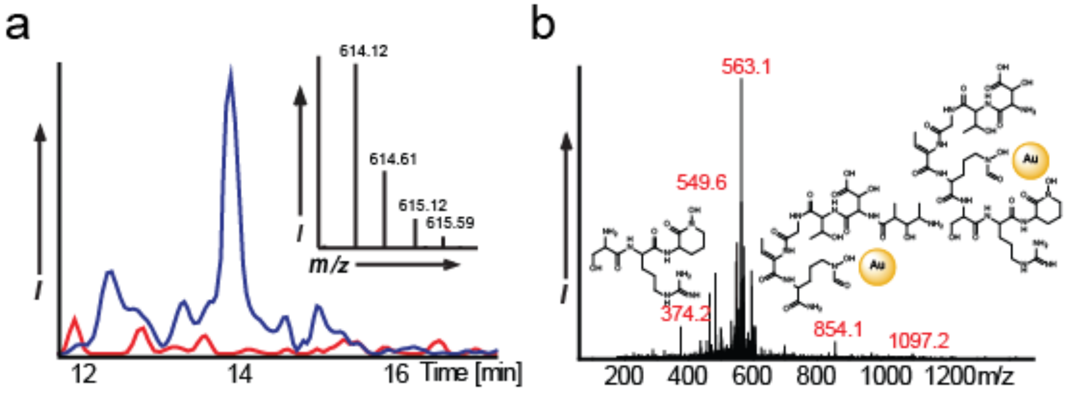
Supplementary Figure 5.7. Delftibactin production is responsive to gold concentrations through reactive homeostasis.

a) MRM-LCMS quantification of delftibactin concentrations. Values represent the percent increase in delftibactin concentrations following precipitation by $AuCl_3$, normalized to an untreated control, ± propagated error; n = 6. b) Sample MRM-LCMS chromatogram of (a).

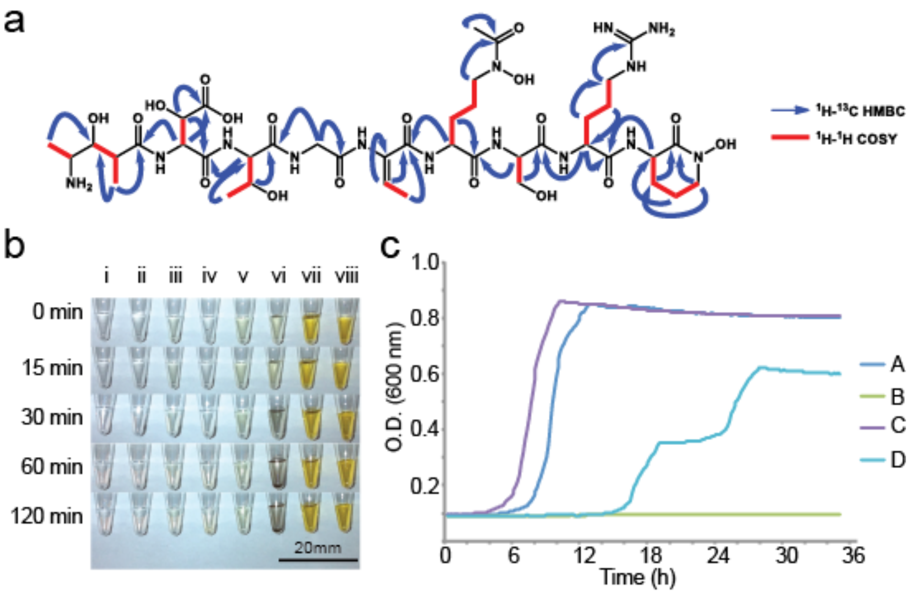


Supplementary Figure 5.8. Delftibactin has a single metal binding site and occupancy by other metals impedes $AuCl_3$ precipitation.

a) Delftibactin (5 mM) and Ga-delftibactin (5 mM) reacted with equimolar concentrations of $AuCl_3$. b) Mass spectra of free (top) and gallium-bound delftibactin (bottom).

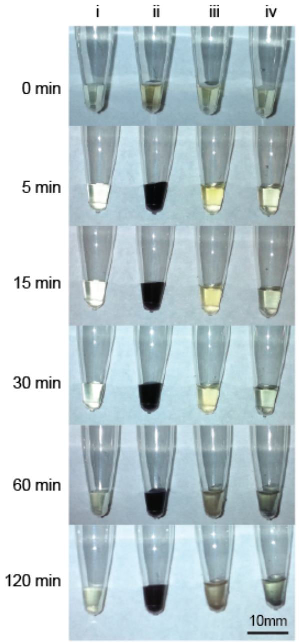


Supplementary Figure 5.9. Mass spectral analysis of a gold-bound delftibactin species. a) A mixture of 50 µM $AuCl_3$ and 100 µM delftibactin shows a double charged ion (blue) corresponding to a delftibactin-gold 1:1 complex not observed when gold is absent (red). b) The fragmentation pattern of the delftibactin-gold ion, showing the characteristic tripeptide loss and remaining delftibactin-gold complex (+374.2, +854.1 m/z), the double- and single-charged ions of the delftibactin-gold complex following the first amide cleavage (+549.6, +1097.2 m/z), and a double-charged delftibactin-gold ion following loss of one side chain (+563.1 m/z). Spectra averaged from a total of 96 scans.



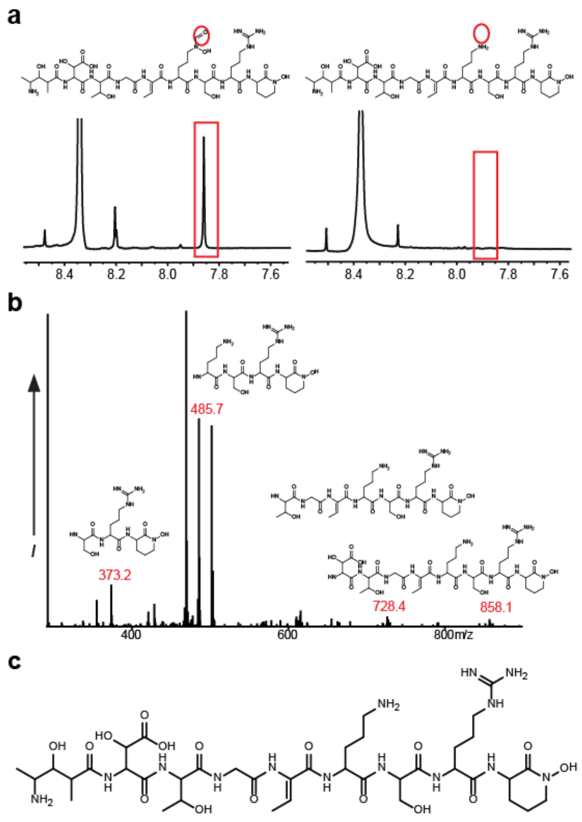
Supplementary Figure 5.10. Delftibactin B (acetylated hydroxy-ornithine delftibactin A) is impaired in gold precipitation and is less protective against $AuCl_3$.

a) Structure of the acetylated analog and spin systems for ^1H-^{13}C HMBC, ^1H-^1H COSY and ^1H-^{15}N HMBC in D_2O for delftibactin B. b) $AuCl_3$ (5mM) precipitation by delftibactin B over time in the absence and presence of $FeCl_3$ as follows: i) water only ii) 5 mM delftibactin B iii) 5 mM $AuCl_3$ iv) 5 mM $FeCl_3$ v) 5 mM $AuCl_3$ + 5 mM $FeCl_3$ vi) 5 mM delftibactin B + 5 mM $AuCl_3$ vii) delftibactin B + 5 mM $AuCl_3$ + 5 mM $FeCl_3$ viii) delftibactin B + 5 mM $FeCl_3$. c) Delftibactin A (formylated) is more protective against $AuCl_3$ toxicity than delftibactin B (acetylated). Growth curves of *D. acidovorans* Δ*delG* in the presence of reaction mixture of water (A), 125 μM $AuCl_3$ (B), 125 μM $AuCl_3$ + 125 μM delftibactin A (C), and 125 μM $AuCl_3$ + 125 μM delftibactin B (D). Results are a mean of three growth curves for each condition from a single representative experiment.



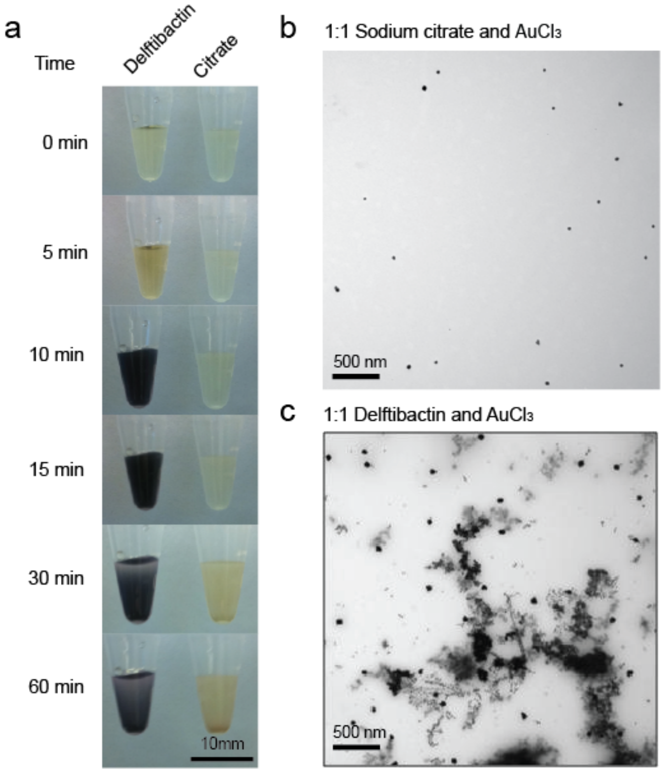
Supplementary Figure 5.11. All delftibactin species eventually co-precipitate with $AuCl_3$.

Timecourse of 5 mM $AuCl_3$ reacted with i) water ii) 5 mM delftibactin A iii) 5 mM delftibactin-$AuCl_3$ reaction product and iv) 5 mM delftibactin B were monitored over 2 h for $AuCl_3$ precipitation.



Supplementary Figure 5.12. Determination of the delftibactin-AuCl$_3$ transient reaction product.

a) Aromatic region of 1D proton NMR of delftibactin A (left) and the reacted delftibactin intermediate (m+/z = 989) (right). The formyl hydrogen signal is absent in the m+/z = 989 delftibactin intermediate (red). b) Fragmentation of the double charged ion of the reacted delftibactin species (m+/z = 989) with diagnostic fragments shown. Mass loss of 44 is localized to ornithine functional group. c) Final structure of transient delftibactin-AuCl$_3$ reaction intermediate is shown. Structure and mass deviation is consistent with the loss of the formyl and hydroxyl group from ornithine.



Supplementary Figure 5.13. Delftibactin rapidly forms complex precipitates of gold nanoparticles.

a) $AuCl_3$ (5 mM) was reacted with equimolar concentrations of delftibactin or sodium citrate. TEM images of sodium citrate-gold (b) and delftibactin-gold (c) reaction after 10 minutes.

5.6 Supplementary Tables

Supplementary Table 5.1 ^{1}H and ^{13}C NMR and NOESY spectral data of delftibactin A in $D_2O^{a,b}$

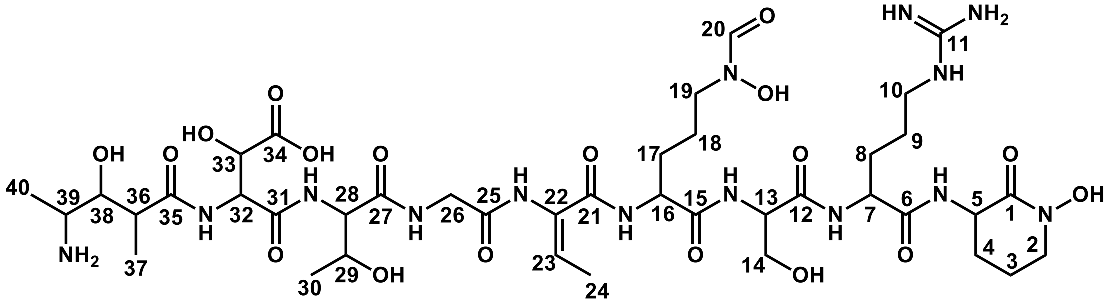
C/H	^{1}H NMR	^{13}C NMR	NOE	Ga-binding	C/H	^{1}H NMR	^{13}C NMR	NOE	Ga-binding
1	–	165.9	–		19	3.49	49.4	H-18	3.62

		(C)					(CH)		
2a	3.54	51.5 (CH$_2$)	H-3a	3.67	20	7.86	159.1 (CH)	–	8.14
2b	3.59		H-3b	3.70	21	–	166.4 (C)	–	
3a	1.87	19.4 (CH$_2$)	H-2a	1.93	22	–	127.2 (CH)	–	
3b	1.92		H-2b		23	6.63	134.1 (CH)	H-24	6.62
4a	1.69	26.0 (CH$_2$)	H-3b	1.80	24	1.65	12.1 (CH$_3$)	H-23	1.66
4b	1.94		H-5	2.01	25	–	172.4 (C)	–	
5	4.36	49.8 (CH)	H-4b	4.39	26	3.99	42.6 (CH$_2$)	–	4.06
6	–	172.5 (C)	–		27	–	171.1 (C)	–	
7	4.25	53.0 (CH)	H-8	4.22	28	4.32	58.5 (CH)	H-29	4.36
8	1.80	27.5 (CH$_2$)	H-7, H-9	1.76	29	4.32	66.1 (CH)	H-28, H-30	4.27
9	1.54	23.8 (CH$_2$)	H-8, H-10	1.54	30	1.13	18.3 (CH$_3$)	H-29	1.13
10	3.10	40.0 (CH$_2$)	H-9	3.12	31	–	171.0 (C)	–	
11	–	156.2 (C)	–		32	4.68	55.8 (CH)	H-33	4.68
12	–	171.3 (C)	–		33	4.21	71.8 (CH)	H-32	4.24
13	4.32	55.3 (C)	H-14a	4.37	34	–	176.4 (C)	–	
14a	3.76	60.3 (CH$_2$)	H-13	3.80	35	–	175.6 (C)	–	
14b	3.78		–		36	2.54	42.5 (CH)	H-37	2.67
15	–	173.5 (C)	–		37	1.18	10.3 (CH$_3$)	H-36, H-37	1.19
16	4.29	53.5 (CH)	H-17a	4.24	38	3.79	71.5 (CH)	H-37, H-40	3.89
17a	1.67	26.8 (CH$_2$)	H-16	1.78	39	3.33	48.6 (CH)	–	3.44
17b	1.79		–		40	1.17	14.1 (CH$_3$)	H-38	1.18
18a	1.63	22.3	H-17a	1.67					

18b	1.65	(CH_2)	–

[a] Chemical shift δ and (multiplicity, J in Hz).

[b] Proton chemical shift changes after gallium binding are highlighted in yellow.



Supplementary Table 5.2 ^1H and ^{13}C NMR spectral data of delftibactin B in D_2O[a]

C/H	^1H NMR	^{13}C NMR	C/H	^1H NMR	^{13}C NMR
1	–	164.9 (C)	19	3.49	49.4 (CH)
2a	3.54	51.7 (CH_2)	20	7.86	160.0 (CH)
2b	3.59		21	–	166.4 (C)
3a	1.87	19.6 (CH_2)	22	–	127.2 (CH)
3b	1.92		23	6.62	134.7 (CH)
4a	1.69	25.8 (CH_2)	24	1.65	12.1 (CH_3)
4b	1.94		25	–	172.4 (C)
5	4.35	49.7 (CH)	26	3.99	42.6 (CH_2)
6	–	172.1 (C)	27	–	171.1 (C)
7	4.27	53.3 (CH)	28	4.35	58.5 (CH)
8	1.80	27.1 (CH_2)	29	4.35	66.1 (CH)
9	1.54	24.1 (CH_2)	30	1.12	18.2 (CH_3)
10	3.12	40.1 (CH_2)	31	–	171.3 (C)
11	–	156.7 (C)	32	4.68	55.8 (CH)
12	–	171.1 (C)	33	4.21	71.8 (CH)
13	4.31	55.1 (C)	34	–	176.1 (C)

14a	3.76	60.0 (CH_2)	35	–	175.6 (C)
14b	3.77		36	2.56	42.1 (CH)
15	–	173.5 (C)	37	1.16	10.3 (CH_3)
16	4.27	53.7 (CH)	38	3.76	71.6 (CH)
17a	1.59	26.4 (CH_2)	39	3.34	48.6 (CH)
17b	1.73		40	1.17	14.1 (CH_3)
18a	1.66	22.1 (CH_2)			
18b	1.63				

[a] Chemical shift *δ* and (multiplicity, *J* in Hz).

Supplementary Table 5.3. Adenylation domain specificities**.**

Prediction of incorporated amino acids was performed using NRPSPredictor and NRPS PKS, providing a probability sequence for the NRPS/PKS product.

Adenylation Domain	**Active Site Residues**	**Substrate**	**Product**
DelE A1	DMGGYGCLFK DAGGCAMVAK	Alanine Alanine	HC Toxin
DelE A2	DIWHISLIEK DVWHISLIDK	Inactive Serine	Nostopeptolide
DelG A1	DLTKVGHVGK DLTKVGHIGK	Aspartic acid Aspartic acid	Surfactin
DelG A2	DFWNIGMVHK DFWNIGMVHK	Threonine Threonine	Syringopeptin
DelG A3	DILQLGLIWK DILQLGLIWK	Glycine Glycine	Nostopeptolide
DelH A1	DFWNIGMVHK DFWNIGMVHK	Threonine Threonine	Syringopeptin
DelH A2	DVWNIGLIHK DVGEIGSIDK	Ornithine Ornithine	Fengycin
DelH A3	DVWHLSLIDK DVWHLSLIDK	Serine Serine	Syringopeptin
DelH A4	DGEDHGAVTK DAEDIGAITK	Arginine Arginine	Pederin
DelH A5	DGEAVGGVTK DGESSGGMTK	N^{δ}-hydroxyornithine N^{δ}-	Vicibactin

		hydroxyornithine	

Supplementary Table 5.4: Delftibactin gene cluster analysis

Locus	Gene	Predicted Function	Strand	Amino Acids
Daci_4765	-	Heavy metal efflux outer membrane component	-	485
Daci_4764	-	Heavy metal efflux periplasmic component	-	405
Daci_4763	-	Heavy metal efflux inner membrane component	-	1029
Daci_4762	-	Nitrogen regulatory protein P-II	-	111
Daci_4761	-	RNA polymerase, sigma subunit	-	174
Daci_4760	*delA*	MbtH domain protein	-	121
Daci_4759	*delB*	Thioesterase	-	247
Daci_4758	*delC*	Phosphopantetheinyl transferase	-	229
Daci_4757	*delD*	Aspartic acid dioxygenase	-	329
Daci_4756	*delE*	Nonribosomal peptide synthetase	-	1789
Daci_4755	*delF*	Polyketide synthase	-	1560
Daci_4754	*delG*	Nonribosomal peptide synthetase	-	3331
Daci_4753	*delH*	Nonribosomal peptide synthetase	-	6176
Daci_4752	*delI*	Siderophore receptor	-	799
Daci_4751	*delJ*	anti-FecI sigma factor, FecR	+	344
Daci_4750	*delK*	RNA polymerase, sigma subunit	+	199
Daci_4749	*delL*	Lysine/Ornithine N-monooxygenase	+	432
Daci_4748	*delM*	Acetyltransferase	+	400
Daci_4747	*delN*	Esterase/Lipase	+	321
Daci_4746	*delO*	Siderophore Export Pump	-	571
Daci_4745	*delP*	N5-hydroxyornithine formyltransferase	-	284

Supplementary Table 5.5 – Primers used in construction of the Δ*delG* strain.

PCR Primer	Sequence (5' – 3')	Purpose
TetNotF	TTT TGC GGC CGC TGC TGA ACC	Amplification of

	CCC AA	tetracycline resistance cassette from pLLX13.
TetNotR	TTT TGC GGC CGC TAT CGT TTC CAC GA	Amplifcation of the tetracycline resistance cassette from pLLX13.
2kbNRPS2Xba2	TTT TTC TAG ACG CAT TGC TGA ACT ACC	Amplification of the 2kb *delG* fragment from *D.acidovorans*.
2kbNRPS2Sac2	TTT TGA GCT CAG CAG TTG CAC CAC CT	Amplification of the 2kb *delG* fragment from *D.acidovorans*.
OriTF	TTT TAA GCT TTT CCT CAA TCG CTC TTC	Amplification of an oriT from pLLX13.
OriTR	TTT TAA GCT TTT TTC GCA CGA TAT ACA	Amplification of an oriT from pLLX13.
NRPS2Seq2	GGG GTG CGG AAA ATG TCC TG	Confirmation of genomic integration of the tetracycline resistance cassette.

.

Supplementary Table 5.6 – Delftibactin production in response to [Fe^{3+}]

Media [Iron]	**[Delftibactin]**
0	205.8 ± 28.0 µM
100 nM	196.1 ± 30.7 µM
1 µM	149.3 ± 21.0 µM
10 µM	21 ± 9.4 µM
100 µM	2.2 ± 1.4 µM
1 mM	Growth Inhibitory

5.7 Supplementary Note – Additional Structural Information

5.7.1 Culture and Isolation

A colony from a fresh plate of *D. acidovorans* was inoculated into a 2.8 L glass Fernbach flask containing 1 L of Acidovorax Complete Media[19] (ACM) that had been treated with 4 grams of Chelex-100 resin (*Sigma*). Cultures were grown at 30°C, shaking at 190 rpm for three days, after which cells were pelleted by centrifugation at 7000 rpm for 15 min. HP20 resin (*Dialon*) was added to the supernatant at 20 g/L and shaken for ~2 h at 220 rpm. The resin was harvested by Buchner funnel filtration and washed with 400 mL of distilled water. The resin was washed with 400 mL of methanol. The methanol eluent was evaporated to dryness under rotary vacuum and resuspended in 2 mL of 50:50 ddH_2O:MeOH. Delftibactin A and B were purified using a Waters Alliance 2695 RP-HPLC separations module, equipped with a Waters 2998 photodiode array and a Luna 5 μm C_{18} column (250 x 10.0 mm, *Phenomenex*). The mobile phase was linear from 2 % acetonitrile, 98 % water + 5 mM $(NH_4)_2CO_3$ at 2 minutes to 14 % acetonitrile at 18 min at a flow rate of 3 mL/min. Delftibactin A eluted at 14.20 min and delftibactin B eluted at 17.5 min.

5.7.2 High Resolution Mass Spectra

A stock solution of 20 mg/ml of delftibactin was diluted to a final concentration of 10 μg/ml in water with 0.1% formic acid. This solution was directly infused at a rate of ~3 μL per min into a Thermo Finnigan LTQ OrbiTrap XL mass spectrometer running Xcaliber 2.07 and TunePlus 2.4 SP1. High resolution MS was acquired using an electrospray ionization source and fragmentation was obtained through collision induced dissociation (CID). The instrument was operated in the positive mode using a maximum resolution of 100, 000. Data was acquired for approximately 1 min for a total of 32 scans. Bradykinin was used as an internal standard, and was premixed with delftibactin to a final concentration of 5 μg/ml. The lock mass feature was applied using the bradykinin standard at [M+H] = 1060.56922 *m/z*.

Compound	***Calculated m/z***	***Observed m/z***	***Δppm***
Delftibactin [M+H]	1033.49143	1033.49154	0.106 ppm

5.7.3 NMR Methods and Structural Characterization

NMR spectra were measured on a Bruker Avance 700 spectrometer equipped with a 5 mm inverse detection probe and using TMS as an internal standard. Lyophilized samples were dissolved in D_2O and spectra were recorded at 297 K. NMR experiments were processed and analyzed with Bruker TOPSPIN 2.1. Chemical shifts (*δ*) expressed in parts per million (ppm) and coupling constants (*J*) are reported in Hertz (Hz). Assembly of individual amino acids to form the final linear structure was accomplished by considering long-range ^{1}H-^{13}C and ^{1}H-^{15}N HMBC correlations from both protons adjacent carbonyl carbons and nitrogens, as well as by assignments of 2D ^{1}H-^{1}H COSY and 2D ^{1}H-^{13}C HSQC correlations.

5.7.4 Delftibactin A

Comprehensive analysis of 2D NMR data, including the results of ^{1}H-^{1}H COSY, HSQC, and HMBC experiments have been used to elucidate the planar structure of delftibactin A, and the chemical shifts from these experiments are provided in Supplementary Table 1. The molecular formula of delftibactin was established as $C_{40}H_{68}$ $N_{14}O_{18}$ based on positive HR-ESI-MS *m/z*: 1033.4915 $[M+H]^+$ (calculated 1033.4914) indicating 15 degrees of unsaturation. ^{13}C NMR (DEPT) and gHMBC spectra revealed ten amide carbons at 156.2, 165.9, 166.4, 171.0, 171.1, 171.3, 172.4, 172.4, 173.5 and 175.6 ppm along with a formyl singlet at 7.86 ppm (^{13}C NMR, 159.1 ppm). Analysis of COSY cross peaks gave eight spin systems. On the basis of long range ^1H-^{13}C, ^1H-^{15}N and COSY interactions and high resolution MS-MS fragmentation, we link the novel structure shown in Fig S3. Ga-delftibactin A was obtained by reacting 1 mg/mL of delftibactin (20 mg in total) in a 100 mL flask with ~5-fold excess of solid $GaBr_3$, added slowly over ~5 min and stirred gently overnight at room temperature. Ga-delftiabactin A was purified using a C18 column (1.6 g, 20 x 0.5 cm), and eluted with 0 to 5% MeOH aqueous solution over 30 min. This procedure provided ~13 mg Ga-delftibactin.

5.7.5 Delftibactin B

Comprehensive analysis of 2D NMR data, including the results of ^1H-^1H COSY, HSQC, and HMBC experiments have been used to elucidate the planar structure of delftibactin B. The chemical shifts of the protons and carbons (Supplementary Table 2) of delftibactin B (m/z = 1047.4) were similar to those of delftibactin A, the main differences between the two metabolites concerned a newly appeared acetyl group [δ_C 17.2, δ_H 2.06 (s)] in compound 2. The location of the methoxyl group was established taking into account the correlation observed between acetyl group δ_H 2.06 (s) and C-20 (δ160.0) in the HMBC experiment of 2.

5.8 References

1. Vining, L. C. *Annu. Rev. Microbiol.* **44**, 395-427 (1990)

2. Nies, D. H. *Appl. Microbiol. Biotechnol.* **51**, 730-750 (1999)

3. Reith, F., Rogers, S. L., McPhail, D. C., & Webb, D. *Science* **313**, 233-236 (2006)

4. Reith, F. *et al*. *Geology* **38**, 843-846 (2010)

5. Reith, F. *et al*. *Proc. Natl Acad. Sci. USA* **106**, 17757-17762 (2009)

6. Reith, F., Lengke, M. F., Falconer, D., Craw, D., & Southam, G. *The ISME Journal* **1**, 567-584 (2007)

7. Kashefi, K., Tor, J. M., Nevin, K. P., & Lovely, D. R. *Appl. Env. Microbiology* **67**, 3275-3279 (2001)

8. Usher, A., McPhail, D. C., & Brugger, J. *Geochim. Cosmochim. Acta* **73**, 3359–3380 (2009)

9. Stachelhaus, T., Mootz, H. D., & Marahiel, M. A. *Chemistry & Biology* **6**, 493–505 (1999)

10. Diels, L., Dong, Q., van der Lelie, D., Baeyens, W., & Mergeay, M. *J. Ind. Microbiol. Biot.* **14**, 142-153 (1995)

11. Salem, I. B., *et al*. *Ann. Microbiol.*, 1-12 (2012)

12. Hider, R. C., & Kong, X. *Nat. Prod. Rep.* **27**, 637-57 (2010)

13. Kim, H. J., *et al*. *Science* **305**, 1612-1615 (2004)

14. Chaturvedi, K. S., *et al*. *Nat. Chem. Bio.* **8**, 731–736 (2012)

15. Miller, M. C., *et al. Microbiology* **156**, 2226-38 (2010)

16. Hough, R. M., *et al*. *Geology* **36**, 571-574 (2008)

17. Ojea-Jiménez, I., Romero, F. M., Bastús, N. G., & Puntes, V. *J. Phys. Chem.* **114**, 1800–1804 (2010)

18. Amin, S. A., *et al*. *J. Am. Chem. Soc.* **129**, 478–479 (2007)

19. Pinel, N., Davidson, S. K., & Stahl, D. A. *IJSEM* **58**, 2147–2157 (2008)

20. Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. *J. Bacteriol.* **173**, 697-703 (1991)

21. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W., & Huson, D. H. *Nucleic Acids Res.* **33**, 5799-5808 (2005)

22. Ansari, M. Z., Yadav, G., Gokhale, R. S., & Mohanty, D. *Nucleic Acids Res.* **32**,W405-413 (2004)

Chapter 6. New Technologies for the Direct Identification of Nonribosomal Peptides

6.1 Chapter Preface

A major hurdle in natural product discovery is the rediscovery of known compounds. The development of iSNAP within our lab is the first example of an algorithm that can dereplicate known NRPs from complex mixtures in an automated and statistically significant manner from LC-MS data. I recognized that similar principles could be applied to identify unknown/cryptic NRPs and that, in theory, if an NRP prediction is 100% accurate, iSNAP can identify the cryptic NRP. However, predictions are not always accurate, but we can often limit the possibilities of building block selection to just a few amino acids for each module and tailoring enzyme functions can be limited to only the available functional groups found on the predicted scaffold. Using the gene cluster as a guide, a database of hypothetical compounds with all iterations and permutations created through building block variation and tailoring enzyme functionalization is created and will have a high likelihood of containing a prediction that matches the final cryptic NRP product or at least containing structural fragments found in the real structure. In such a way, we can use the iSNAP algorithm to successfully identify unknown NRPs prior to bioactivity or purification. For simple NRPSs, such as the aureusimine NRPS, a small database is generated, whereas more complex NRPS will have larger prediction databases due to the increased number of possible variations. Theoretically in both cases, this prediction database approach can be successfully employed using iSNAP to identify cryptic NRPs, irrespective of cluster size and complexity. In this final chapter, I develop the iSNAP algorithm to identify cryptic NRPs from complex extracts using genomic predictions in conjunction with automated LC-MS data acquisition. This is the first technology that directly identifies predicted metabolites from complex extracts and builds onto the genome mining methodologies used in **Chapter 2** and **Chapter 4**.

The following chapter is a modified version of a publication in review at *Nature Biotechnology*. I am a co-first author of this publication. I contributed greatly to the conception and development of iSNAP as a tool to identify cryptic NRPs. In particular, I generated the prediction database methodology for using iSNAP as a tool for identifying unknown NRPs, performed all cluster and iSNAP analysis, generated the figures, identified and isolated variobactin, and contributed to the writing of the final manuscript. Chad Johnston contributed to identifying and isolating the acidobactin, vacidobactins, and the WS series of compounds, and writing the manuscript. He also identified and isolated thanamycin and performed initial trial experiments showing iSNAP is capable of identifying unknown metabolites from small, predicted NRP libraries. Xiang Li performed all NMR structure elucidations. Lian Yang helped develop the original iSNAP algorithm and generated modified iSNAP outputs use in this project. Ashraf Ibrahim helped develop the original iSNAP algorithm and generated the HRMS for acidobactin and vacidobactin. Alyssa Grunwald and Russell Kerr provided the extracts containing the peptaibols. Stephanie Vanner and David Zechel provided the original extracts containing WS-9326A. Nathan Magarvey contributed greatly to project conception and provided extensive feedback throughout. The following is a proposed citation for the article in review at Nature Biotechnology:

Wyatt, M.A.*, Johnston, C.W.*, Li, X. Yang, L. Grunwald, A., Ibrahim, A., Vanner, S.A., Zechel, D.L., Kerr, R.G., Ma, B., Magarvey, N.A. (**2013**). **Directed discovery of unknown natural products using fragment-based molecular barcodes.** *Submitted for review in Nature Biotechnology.*
*These authors contributed equally to the work.

6.2 Abstract

Natural products possess unique activities and are profoundly important as therapeutic and industrial agents. Microbial natural products created by modular assembly-line like enzymes encoded by clusters of polyketide synthase and nonribosomal peptide synthetase genes are particularly notable for their diverse bioactivities and chemical structures. Genome sequencing suggests that many novel polyketides and nonribosomal peptides remain (so called known unknown natural products), but there is no direct method for their targeted isolation or for the generalized exploration of their pharmacophores and the chemical space they occupy. Here, we describe an informatic strategy for automated, targeted detection of predicted or hypothetical natural products based on chemical barcodes and MS/MS fragmentation. Using this approach we have identified over 25 new genetically-envisioned compounds and molecules with desired pharmacophores from complex extracts, leading to the selective isolation of over 10 novel compounds. These examples demonstrate an ability to identify and elucidate ‘known unknown’ nonribosomal peptides, including unknown variants and relatives of knowns, as well as ones predicted from genomic data.

6.3 Introduction

Natural products serve as central pillars in human therapeutic development and are major drivers in the innovation and inspiration of drugs used in modern medicine[1-4]. Evolved chemical patterns and pharmacophores direct specific binding and lead to selective modulation of cellular processes (**Fig. 6.1**)[5-7]. Strategic exploration and expansion of privileged natural product chemical space is recognized component of drug discovery, and natural products and their derivatives comprise a diverse array of clinically used antimicrobial/anticancer agents, immunomodulatory entities, and cholesterol-lowering therapies (**Fig. 6.1**)[1, 5, 7]. Commonly, the initial natural product hit is not optimal as a drug, and new variants must be isolated or created to realize human therapeutics with optimal efficacy, stability, and/or safety [8, 9]. Methodologies from synthetic chemistry, such as diversity oriented synthesis or medicinal chemistry techniques are important contributors to drug creation from natural product leads, but they are often hindered by costly, time-consuming syntheses due to the complexity of the natural product scaffolds [10, 9]. Microbes, however, are prolific in their combinatorialization around bioactive scaffolds, taking advantage of the diversity-oriented biosynthesis achieved by modular assembly lines (i.e. polyketide synthases [PKSs] and nonribosomal peptide synthetases [NRPSs]) that are chemically promiscuous, and seemingly genetically recombinogenic[5, 11, 12]. These natural diversity oriented biosyntheses lead to the production of series of bioactive metabolites present as dominant products or minor constituents, in concentrations that may be below the limits of bioactivity detection [13-16] (**Fig. 6.1**). Sole use of bioactivity based navigation of naturally evolved drug space acts to pre-select for abundant compounds and is often confronted with isolation of knowns[17], and is low-throughput, cumbersome, and ambiguous with respect to the chemical nature of the lead[18,19]. Recent metabolomics works have sought to reveal known natural products through informatics strategies both from microbial genomic-level information and predictions of natural products from PKS and NRPS gene clusters, exposing the wider chemical space genetically encoded molecules may occupy[20-22]. Accessing the full collection of natural products and explicitly these '*known unknown*' molecules is suggested as a key challenge in tapping into undiscovered drug leads visible within microbial genomes[23-26]. The significant numbers of these unknowns has lead to proposals that these are cryptic metabolites, meaning either that they are not expressed or perhaps they have not yet been identified within the complex natural product extracts. Several lines of evidence suggest specific manipulations either varying culturing conditions or specific inductions (e.g. temperature and chemical epigenetic modulators) tune their production. Rapidly connecting these cryptic gene clusters to their products and selectively targeting these molecules within crude and complex extracts, is important for focusing efforts on new leads for activity testing. Certainly strategies that can specifically leverage the predictability of these known unknowns and selectively target these molecules for drug discovery purposes will spur the advancement of these genetically evolved natural products for biotechnological and pharmaceutical uses. Integrating and connecting these biosynthetic gene clusters to their metabolic products will, however, demand new tools and technologies. Here we presents a strategy that provides a mechanism to strategically expand drug space of noted nonribosomal peptide pharmacophores directly from natural product extracts and specifically detect and physically locate the predicted known unknown metabolites from NRPS biosynthetic clusters.

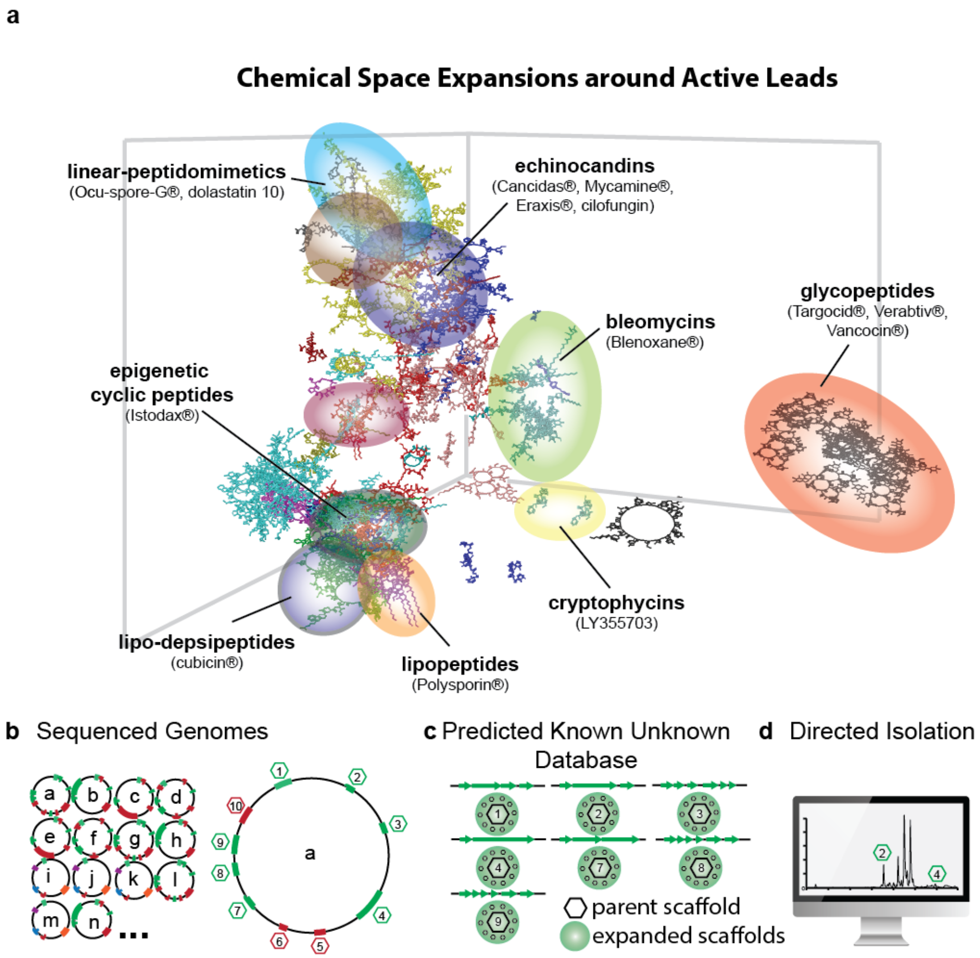


Figure 6.1.

a) Three-dimensional chemical space of select nonribosomal peptide compounds. A series of 225 compounds taken from the iSNAP database and plotted for the following features: molecular weight, OpenBabel linear fragments (FP2), and CDK Klekota-Roth biological activity features using Chemical Space-Mapper (CheS-

Mapper v1.0.27)[56]. Natural product families with defined pharmacophores and activities are indicated by circles, Examples of family members that have entered clinical trials or become FDA approved drugs are indicated[57-62]. b) Representative bacterial genomes containing biosynthetic gene clusters with known compounds (red) and cryptic biosynthetic gene clusters with no known product (known unknown) (green). Hexagons are representative of the known molecules (red) and the predicted molecules (green) genetically encoded within the genome of bacterium **a**. b) Generation of prediction database of based on predicted scaffolds of the 'known unknowns' from bacterium **a**, representing possible monomer variation or tailoring events. c) Representation image of iSNAP analysis of bacterial extract by LC-MS using the generated 'known unknown' database of organism **a**. Green hexagons.

6.4 Results

6.4.1 Hypothetical Barcode Libraries Identify Novel Variants of Knowns

Within the course of a cell-based bioactivity screening campaign of a natural product extract library using the model eukaryote *Saccharomyces cerevisiae*, we identified a number of cell death inducing extracts. Microbial extracts with activity were profiled using LC-MS/MS with automated data dependent acquisition to obtain spectra and fragmentation patterns of analytes within the extract (**Fig. 2a**). To detect whether this bioactive extract contained known natural products, we chose to employ the recently developed informatic search algorithm, iSNAP[29]. As amino acid sequences found within peptides are specific to their identity, the MS/MS fragments of natural peptides can be used as chemical barcodes to automatically match acquired MS/MS data to barcode libraries of hypothetical MS/MS fragments, facilitating the automated identification of natural peptides within chromatograms of complex extracts without isolation. MS/MS scans of bioactive extracts were analyzed using the iSNAP program, which identified a metabolite within the first extract at a retention time of 27.8 min whose MS/MS barcode provided a P1/P2 match score of 24/15 to that of the known peptaibol trichopolyn 1 (**Fig. 6.2b**). This structure was subsequently verified by high resolution MS and MS^n fragmentation (**Supplementary Information; Supplementary Fig. 6.1**). Other extracts within this library also yielded barcode matches to known peptaibols, including efrapeptin (**Supplementary Fig. 6.2**). Peptaibols, like other natural products, have distinct chemical features (i.e. pharmacophores) that define their actions. For instance, repetitive aminoisobutryic acid (Aib) monomers frequently installed within their peptide backbones often leads to an alpha helical structure and directs membrane channel formation[30]. Interest in membrane selective peptaibols has prompted further analysis, and has led to the discovery of family members that are cancer cell specific[31]. Peptaibols are known to have a nonribosomal origin, and like other nonribosomal peptides variation occurs through tailoring modifications and nonspecific amino acid incorporation during their assembly, creating within-family diversity[32]. Currently, there are no automated strategies for unveiling variants within chemical families, or revealing their locations within extracts. The utility of a database driven approach such as iSNAP is that new barcodes can be added to the existing database for new hypothetical variants that may then be identified by iSNAP within extracts with statistical validity.

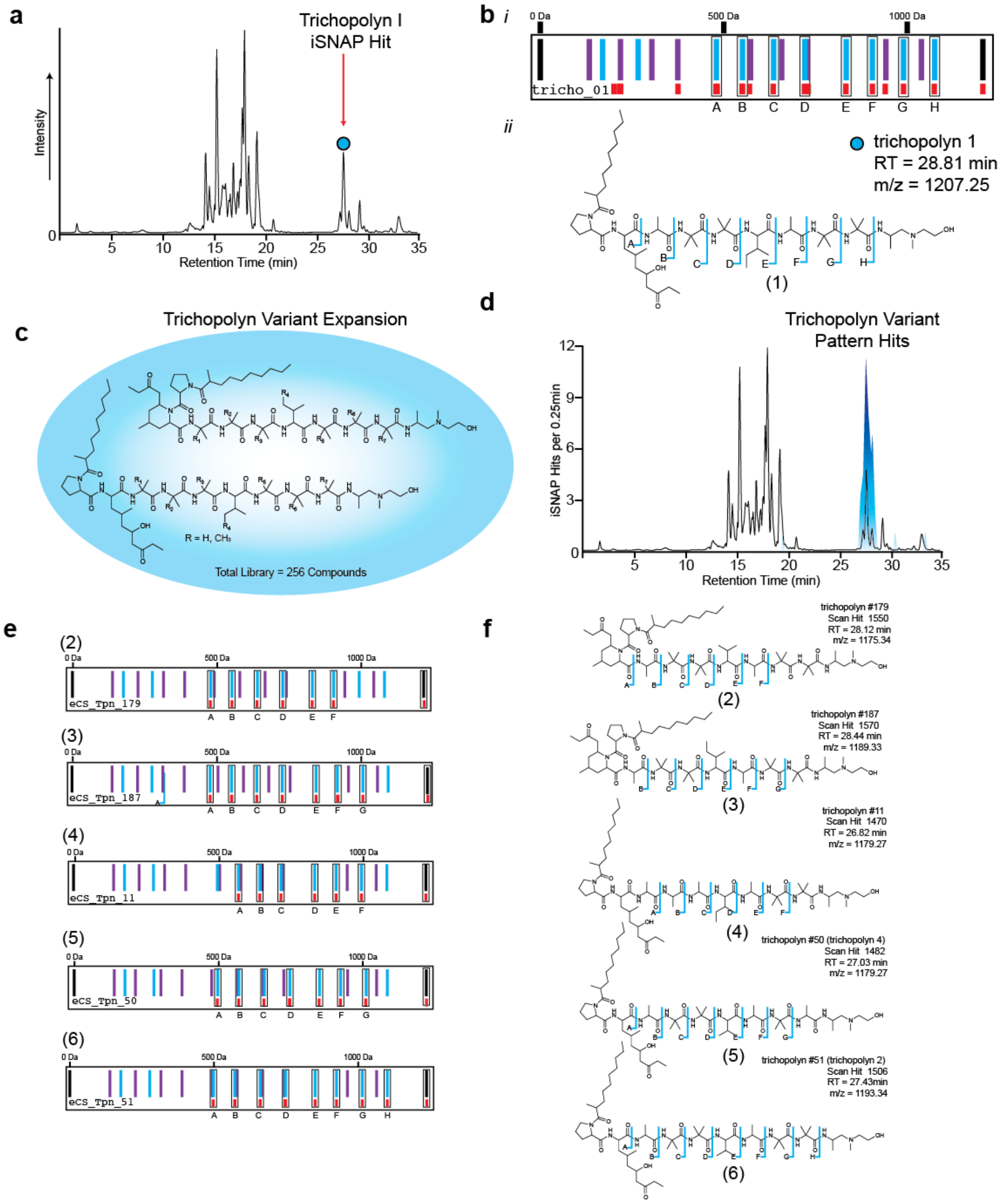


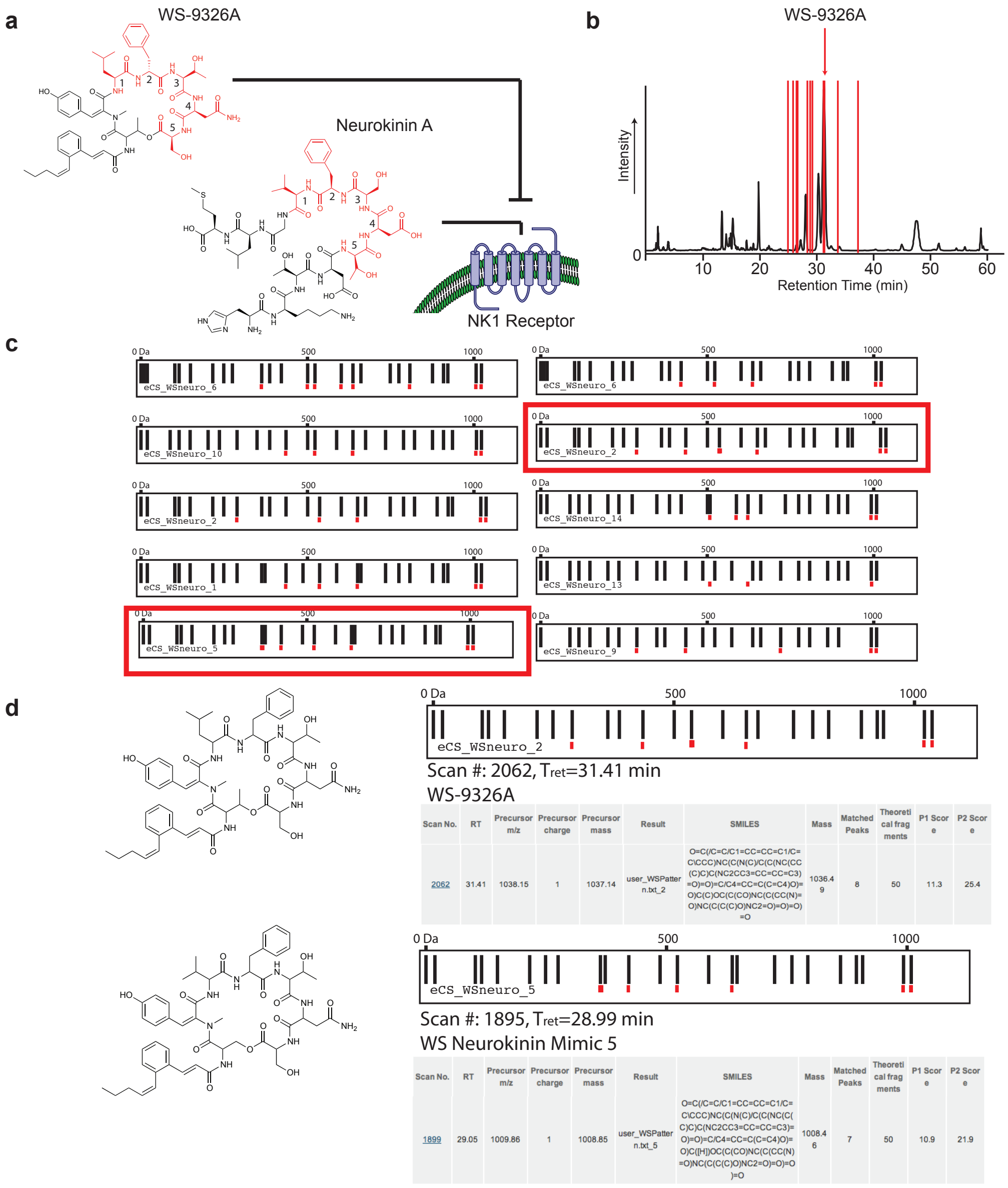
Figure 6.2. Identification of trichopolyn from extract natural product bioactivity screens using informatic search for natural products (iSNAP).

a) LC-MS base peak chromatogram of environmental extract, RKDO-M33, is shown with trichopolyn 1 iSNAP hit. b) Fragmentation pattern of trichopolyn 1. *i)* MS2 spectra of trichopolyn 1 hit, scan 1530. *ii*) Predicted b (blue) and y (purple) ion fragmentation of trichopolyn 1 compared to iSNAP matched fragment hits (red). *iii*) B ion fragment matches are indicated below on the structure of trichopolyn 1. c) Computer generated trichopolyn variant library consisting of all combinations of either alanine of aminoisobutyric acid (AIB) and valine or isoleucine on trichopolyn scaffolds. d) An iSNAP direct mass hit (mass window = 1 Da) frequency plot of iSNAP trichopolyn variant hits per 0.25 min in LC retention, overlaid on the 344 M3 LC-MS base peak chromatogram. e) Trichopolyn Barcode Hits using predicted b (blue) and y (purple) ion fragmentation of trichopolyn variant hits compared to iSNAP fragment hits (red). F) B ion fragment hits are indicated on identified trichopolyn variant structure found using the trichopolyn barcodes (for structural details see Supplementary Figures 2-8).

In an attempt to identify undescribed members of the peptaibol family, we generated a barcode library of hypothetical trichopolyn variants, using the chemoinformatic program SmiLib v.2.0 to incorporate seven sites of modification within the peptide core[33]. This approach simulated the combinations and permutations of natural product diversity that could plausibly arise from the trichopolyn assembly-line, generating 256 hypothetical variant barcodes (**Fig. 6.2c**). These hypothetical variants were added to the defined library of NRPs currently populating iSNAP to form a new library including hypothetical unknown trichopolyn structures. Reanalysis of LC-MS/MS data of the trichopolyn extract with this extended barcode library enhanced the number of hits to include 3 novel and 3 known structures (**Fig. 6.2d, e**) including trichopolyn 1 (**Supplementary Fig. 6.3**). The putative variants detected by iSNAP at retention times 28.81 (1), 28.12 (2), 28.44 (3), 26.82 (4), 27.03 (5), and 27.43 min (6) had MS/MS barcodes that matched the iSNAP-generated fragmentation patterns of the hypothetical variant structures # 59 (1; trichopolyn 1), #179 (2), #187 (3), #11 (4), #50 (5; trichopolyn 4), and #51 (6; trichopolyn 2), corresponding to their iSNAP hits (**Fig. 6.2e; Supplementary Fig. 6.4-8**). High-resolution mass spectrometry and manual MS/MS annotation confirmed the identity of the iSNAP variant barcode hits demonstrating that this approach is useful and accurate in expanding and exploring natural chemical space around known natural product structures (**Fig. 6.2f; Supplementary Fig. 6.4-8**). The validity of the hypothetical variant barcode approach is further illustrated by the identification of the hypothetical trichopolyn variant #59, which iSNAP correctly matched to the known structure of trichopolyn 1 from the large structure library of 256 trichopolyn variants (**Supplementary Fig. 6.1 and 6.3**).

6.4.2 Identification of Site Specific Modifications within Desired Pharmacophores

While screening natural product extracts for both known and predicted molecules using iSNAP, we identified a strain of *Streptomyces calvus* as a novel producer of the lipodepsipeptide WS-9326A, and confirmed this with 1D and 2D NMR experiments (see **Supplementary Information**). Though this strain had previously been studied for production of anti-bacterial and anti-trypanosomal compounds[38], this new molecule functions as a potent antagonist of the G-protein coupled receptor NK-1 [39, 40], whose natural ligands are tachykinin peptide hormones such as neurokinin A [41] (**Fig. 6.3a**). Neuropeptide mimics, accessed through synthesis or targeted isolation, have been well characterized as potent modulators of GPCR activity [42]. Analysis of the WS-9326A structure revealed that a large portion bore a significant resemblance to neurokinin A (**Fig. 6.3a**) which could seemingly explain its potent antagonistic activity. To identify analogs of WS-9326A with increased similarity to neurokinin A, we generated a targeted barcode library with combinatorialized substitutions at macrocycle ring positions 1, 3, 5, and 7, where modifications may lead to increased homology (**Supplementary Fig. 6.9**). This targeted library was used with a mass window of one to exclusively identify specified analogs from a culture extract of *S. calvus*. A series of barcodes matching WS-9326A variants were identified, including recently described congeners [43], and novel analogs with increased homology to neurokinin A through detected amino acid substitutions (**Fig. 6.3b**). One minor analog, which co-elutes with several other structural analogs and was produced at 0.4% the titer of WS-9326A, was shown to possess serine and valine substitutions at peptide macrocycle positions 1 and 3, and was targeted for time-dependent fractionation using the iSNAP directed retention time of 28.9 min (**Fig. 6.3c; Supplementary Fig. 6.10-11**). Subsequent MS/MS and NMR studies confirmed the match to the hypothetical chemical barcode (**Supplementary Fig. 6.12-13**), thus outlining how a target pharmacophore-associated chemical space can be enlarged and mapped informatically by using barcode libraries to identify site-specific modifications of bioactive natural product leads, even when such compounds are present in vanishing quantities.



Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	Matched Peaks	Theoretical fragments	P1 Score	P2 Score
2062	31.41	1038.15	1	1037.14	user_WSPattern.txt_2	O=C(/C=C/C1=CC=CC=C1/C=C\CCC)NC(C(N(C)/C(C(NC(CC(C)C)C(NC2CC3=CC=CC=C3)=O)=O)=C/C4=CC=C(C=C4)O)=O)C(C)OC(C(CO)NC(C(CC(N)=O)NC(C(C(C)O)NC2=O)=O)=O)=O	1036.49	8	50	11.3	25.4

Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	Matched Peaks	Theoretical fragments	P1 Score	P2 Score
1899	29.05	1009.86	1	1008.85	user_WSPattern.txt_5	O=C(/C=C/C1=CC=CC=C1/C=C\CCC)NC(C(N(C)/C(C(NC(C(C)C)C(NC2CC3=CC=CC=C3)=O)=O)=C/C4=CC=C(C=C4)O)=O)C([H])OC(C(CO)NC(C(CC(N)=O)NC(C(C(C)O)NC2=O)=O)=O)=O	1008.46	7	50	10.9	21.9

Figure 6.3. Directed discovery of Neurokinin A mimics from the iSNAP discovery of WS-9326A.

a) The natural substrate (neurokinin A) and the iSNAP detected depsipeptide inhibitor (WS-9326A) for the NK-1 GPCR are shown. Structural similarities between neurokinin A and WS-9326A are shown in red and labeled 1-5. b) A structural library of WS variants based on the structure of neurokinin A (WS-Neurokinin Mimic Library) was used to screen the LC-MS chromatogram using iSNAP. iSNAP detected WS-Neurokinin mimic variants are indicated in red on the chromatogram. c) The theoretical WS-Neurokinin A Mimic Variant hits found within the extract are shown with their barcodes (black) and iSNAP fragment hits (red). d) iSNAP identified WS-Neurokinin A mimic library hits including WSneuro_2 (WS-9326A) and WSneuro_5 are shown with their structure, barcode (black) with iSNAP hits (red), retention time, and the associated iSNAP screenshot. WSneuro_5 and WS-9326A were isolated and assayed for their antagonistic activity against NK-1 receptors.

6.4.3 Discovery of Known Unknowns through Genome-Predicted Barcode Libraries

Directly identifying specific nonribosomal peptide biosynthetic gene cluster products or previously unidentified, known unknowns, will demand forward strategies that leverage established principles that infer their biosynthetic predictions. Such a strategy also needs to accommodate the realty that biosynthetic predictions are often incomplete, and lack sufficient accuracy, but nonetheless provide insight into components of these genomically-envisioned known unknowns. We sought to apply our molecular barcoding strategy with genomic predictions of nonribosomal peptides and seek to use this strategy to identify the physical location of such genomically-envisioned known unknowns. As a first test case we applied this to genetic clusters that were identified within previously unstudied bacteria but had biosynthetic clusters with some sequence homology to the cluster for the molecule, delftibactin, a novel metallophore from a microbe found on gold deposits.

A series of undescribed NRPSs from organisms including *Acidovorax* (*A. citrulli* AAC00-1) and *Variovorax* (*V. paradoxus* S110 and *V. paradoxus* EPS) (**Supplementary Fig. 6.14**). These uncharacterized NRPS clusters have high overall similarity and a pattern of homology that exists between the modules, indicative of common natural product architectures. Inspection of the *A. citrulli* AAC00-1 NRPS assembly-line exposed the specificity codes of the adenylation domains (alanine, ornithine, serine, threonine, and aspartic acid) with additional gene cluster analysis suggesting plausible modifications to yield aspartic acid β-hydroxylation, and ornithine N^{δ}-hydroxylation, formylation, or acetylation. a peptide scaffold for a barcode library of hypothetical structures (**Fig. 6.4a**). Further combinatoralizations were also included to include polyketide assembly line variance including malonate and methyl malonate.   The importance of this combinatorialization is to increase the likelihood that one of the structures is a more accurate match of the real structure and so it can be readily detected. Further it is known in NRP predictions from genomic information that variance frequently occurs between the core prediction and the final structure, as post assembly-line tailoring of the peptide core leads to further diversification[14, 15, 25].   In total a molecular library of 576 and their associated chemical barcodes could be realized when considering these molecules could be found in two forms, linear and cyclic structures.

We reasoned that one of the predicted *A. citrulli* hypothetical NRP structures would sufficiently reflect the real structure (within the extract) that when their fragmentation barcodes were matched they would overlap and the iSNAP algorithm would discern the physical location of the 'known unknown' natural product by utilizing the library of hypothetical chemical barcodes. Loading these molecular barcodes into the software and running the extract *A. citrulli* AAC00-1 strain, fermented it, extracted with resin, and interrogated the extract for the predicted unknown. The *A. citrulli* AAC00-1 NRP barcode library was loaded into iSNAP and a precursor ion mass window of 50 *Da* was used to analyze the LC-MS/MS chromatogram and account for minor differences between predicted and matched structures. Taking the resulting ranking of the iSNAP hits for each scan at increasing retention time, and plotting this distribution of hits into a frequency plot over the LC chromatogram provided a ranking of the most closely matched hypothetical variants with a specific retention time (**Fig. 6.4b**). The two predominant hits are shown in a representative fractal tree which tracks sequential permutations of the chemical structures of hypothetical *A. citrulli* NRPs (**Fig. 6.4c**). These two hits occupy a common branch of this tree, indicating a high degree of similarity between the hypothetical cyclic compounds #200 (140 iSNAP hits) and #202 (47 iSNAP hits) at retention times of 12.13 and 11.80 min, respectively. The molecules were obtained from broth extractions and their structures were determined by 1D and 2D NMR experiments (**see Supplementary Information**), and named acidobactin A and B **(Fig. 6.4e**). Not only were these structures analogs of each other, they exhibited significant structural similarity to hypothetical compounds #200 and #202. Since the gene clusters of *A. citrulli* AAC00-1 and *V. paradoxus* S110 have extensive homology to each other, with identical adenylation domain specificity and assembly-line architecture, we sought to examine whether the acidobactin prediction library would succeed in identifying the unknown *Variovorax* NRP from a different metabolic background. Similar fermentation and iSNAP analysis of the *V. paradoxus* S110 extract revealed matched barcodes within the same retention time region as acidobactin A and B, with a top hit of hypothetical structure #200 (21 iSNAP hits; **Supplementary Fig. 6.15**). Subsequent isolation of iSNAP hit peaks revealed vacidobactin A and B, whose structures and MS2 fragmentation patterns were analogous to the acidobactins (**Supplementary Fig. 6.15**), with an extra methyl group derived from the PKS mediated incorporation of methyl malonate. Upon realization of the isobutyric acid substitution for the predicted alanine within the structures of acidobactin A and B (as well as the vacidobactins), a new barcoded hypothetical structure library was created, including 72 hypothetical variants of these macrocyclic compounds, incorporating more varied ornithine modifications, along with malonate and methyl malonate units. Using this refined structural library to reanalyze the original *A. citrulli* extract exposed a new hit similar to hypothetical variant #56 which eluted at 14.0 min, and upon further MS/MS analysis, is proposed to be an additional acidobactin analog, acidobactin C (**Supplementary Fig. 6.16**).

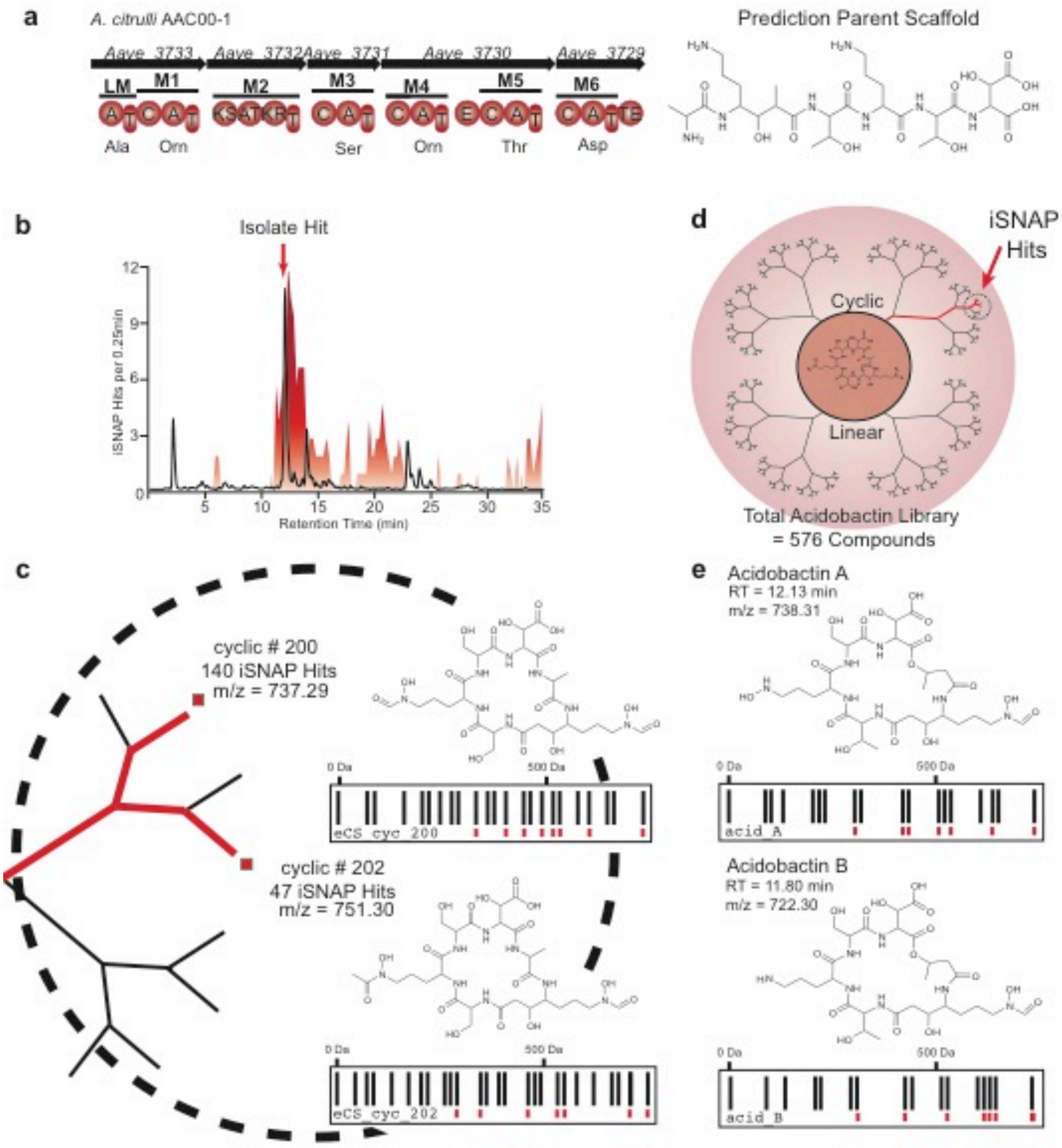


Figure 6.4. Identification of *Acidovorax citrulli* AAC00-1 nonribosomal peptide using iSNAP with a predicted acidobactin structure database.

a) *Acidovorax citrulli* AAC00-1 NRPS-PKS gene cluster and protein domain organization including adenylation (A), thiolation (T), condensation (C), ketosynthase (KS), acyltransferase (AT), ketoreductase (KR), and thiolation domains. The predicted amino acids loaded onto the assembly line are indicated below their respective A domain and the predicted parent scaffold of the cryptic Acidovorax NRP-PK is shown. b) *A. citrulli* AAC00-1 extract LC-MS base peak chromatogram overlaid with a frequency plot of predicted acidobactin library iSNAP hits per 0.25 min in LC retention (mass window =50). c) Fractal tree representation of expanded chemical space using acidobactin prediction compound library with the two major iSNAP hits indicated. d) Close-up of predicted acidobactin fractal tree showing the structure of the two major iSNAP hits, their chemical barcode (black) with iSNAP fragment hits (red), theoretical molecular weight, and retention time of hits are shown. e) Final structures, retention time, and molecular weight of isolated compounds corresponding to iSNAP hits, acidobactin A and B.

We sought to extend our barcode-based targeted isolation strategy for delftibactin-like molecules beyond organisms with sequenced genomes, into a natural product extract screening campaign comprising extracts from uncharacterized environmental microbes. We first profiled organisms on the basis of whether they produced agents that exhibit likeness to a series of 14, 592 hypothetical variants of delftibactin-like structures, including delftibactin, acidobactin, vacidobactin, and the predicted NRP from *V. paradoxus* EPS (**Fig. 6.3, 6.5a**). One of the extracts within the natural product library produced two significant hits including hypothetical barcodes #3278 and #2953 (**Fig. 6.5b, c**). The retention time of the associated hits were 23.1 min and included precursor masses of 1149.58 and 1134.64 (**see Supplementary Information for statistical values and P1/P2 scores**). MS-guided isolation of this peak was conducted, and subsequent scaling of broth extracts yielded sufficient quantities for structural characterization by NMR. The complete structure was assigned and defined as an acylated cyclic depsipeptide with components in common with delftibactin, acidobactin, and vacidobactin, including common modified ornithine units, β-hydroxy aspartic acids, and serine, which indicated that it may also arise from an analogous gene cluster and was given the name variobactin A (**Fig. 6.5d**). Following the description of this novel molecule, a new library of chemical barcodes was created to incorporate new sites of modification on the identified scaffold, including variations in ornithine decoration and fatty acid chain length (C_9-C_{14}). Re-testing the environmental extract LC-MS/MS file with this extended barcode library revealed a series of related natural products which were named variobactin B-E. The structures of these compounds were inferred from the molecular barcodes corresponding to their iSNAP hits, specifically chemical barcodes #209 (variobactin B), #210 (variobactin C), #209 (variobactin D), and #178 (variobactin E) (**Supplementary Fig. 6.17**). Although variobactin B and E shared the same barcode match (#209), they differed by 2 *Da* and indicated a divergent desaturation in the fatty acid tail which was not included in the variant library. This example demonstrates the utility of an informatic search strategy by selectively identifying and locating novel nonribosomal peptides from select regions of chemical space directly from natural product extracts. In total, we have defined five new cyclic lipopeptides from this isolate which appear to share a common biosynthetic origin, arising from a PKS-NRPS gene cluster that is suspected to share sequence homology with the delftibactin / acidobactin / vacidobactin gene clusters (**Fig. 6.3**). To determine if this was true, the genome of the environmental isolate P4B was sequenced and scanned for PKS-NRPS gene clusters using the acidobactin NRPS genes as a query sequence. Results from these searches revealed that it indeed harbors a PKS-NRPS assembly-line that bears significant similarity to the *Delftia*, *Acidovorax*, and *Variovorax* biosynthetic gene clusters (**Fig. 6.5e**), validating our hypothesis that hypothetical barcodes could be used to populate chemical space and identify related unknown molecules.

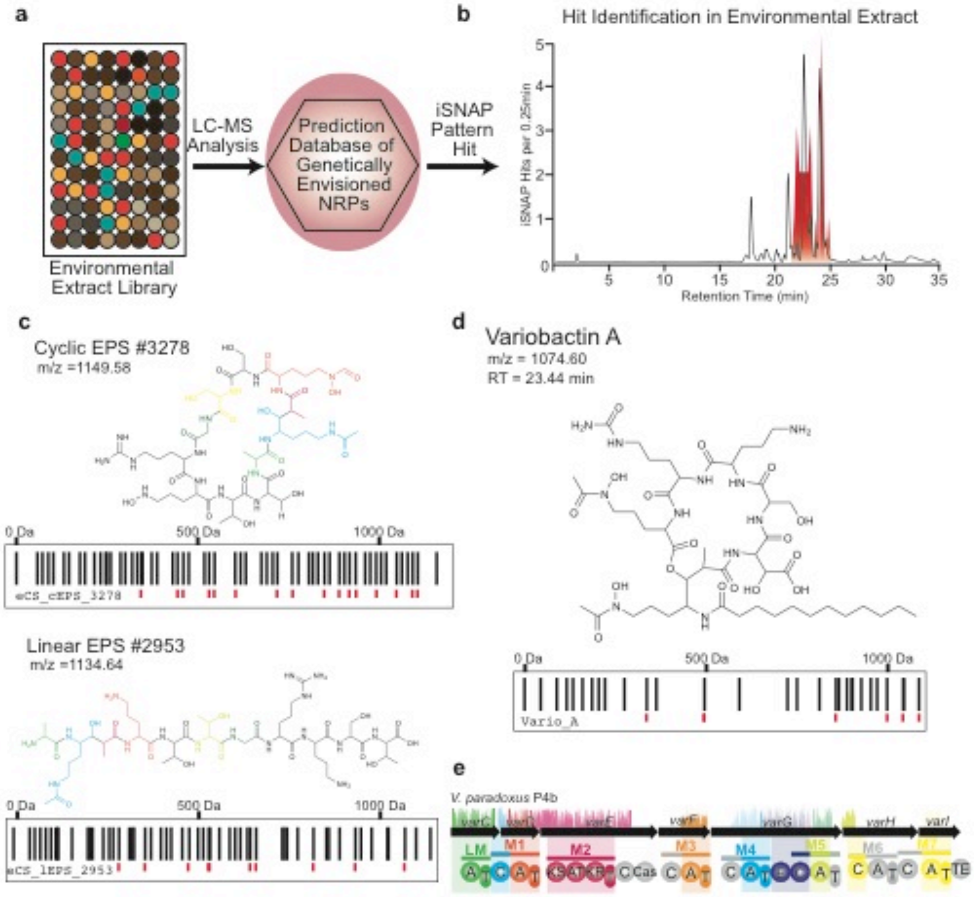


Figure 6.5. iSNAP-guided discovery of delftibactin-acidobactin-like compound from unknown extracts.

a) LC-MS screen of environmental extract library identifies prediction library iSNAP hits in unknown strain P4b. b) LC-MS base peak chromatogram of environmental strain P4b overlaid with a frequency plot of the predicted library iSNAP hits per 0.25 min in LC retention time (mass window =50). c) Top iSNAP prediction library hits for environmental strain P4b showing their theoretical barcodes (black) with iSNAP detected hits (red), structure, and mass. d) Final structure of isolated iSNAP hit and novel natural product, variobactin A, including the structural barcode (black) and fragmentation hits found by iSNAP (red). e) Variobactin gene cluster from sequenced genome, *V. paradoxus* str. P4b. Gene similarity to related acidobactin and delftibactin gene clusters is shown using Mauve analysis (see Figure 3).

To further demonstrate the strength of this approach we chose to investigate genetically predicted metabolites with noteworthy activities, but whose complete structures have remained elusive. With the advent of next generation sequencing the discovery of such 'cryptic' gene clusters is accelerating as exemplified by recent examples such as colibactin (a PK/NRP from *E. coli* found within the human microbiome implicated in colon cancer) and thanamycin. The thanamycin case is an intriguing one, as the molecule was not identified but the corresponding gene cluster was found to be responsible for disease suppressive soils. We recently sequenced the genome of a *Pseudomonas fluorescens* strain that possesses a cluster that is identical to that of the complete thanamycin gene cluster (89% identity/93% similarity) found within *Pseudomonas sp*. SHC52 (**Fig. 6.6a**). Using this gene cluster, we were able to predict the general peptide skeleton of thanamycin, but with four sites of low prediction confidence, specifically within the lipid tail, as well as monomer positions 2, 3, and 4 which are only suggestive of arginine, aspartic acid, and ornithine respectively (**Fig. 6.6b**). This structure prediction was used to query an extract of *P. fluorescens* DSM 11579, using a mass window of 50 Da and facilitate detection of structures similar to the prediction. This search yielded a single hit at 21.16 min, corresponding to a doubly charged peptide with a molecular weight of 1273 Da. With this in mind, sites of low prediction confidence were combinatorialized with similar lipids and amino acids also observed in monochlorinated cyclic lipopeptides from Pseudomonads, generating a second library of 96 hypothetical structures. This hypothetical structure library was used to query an enriched fraction of thanamycin produced with size exclusion chromatography, using a stricter mass window of 20 Da to further narrow in on the true structure. The predominant structure returned from this search included modifications at each of the four positions, and differed from the putative thanamycin by a mass consistent with an additional hydroxyl group. Generation of another 96-membered structure library with an additional hydroxyl group was able to consistently identify the putative thanamycin peak within a window of 1 Da, suggesting *de novo* structure solution of this previously cryptic metabolite. To verify this, thanamycin was isolated and solved by NMR, confirming the structure identified through our molecular barcode analysis (**Fig. 6.6b**).

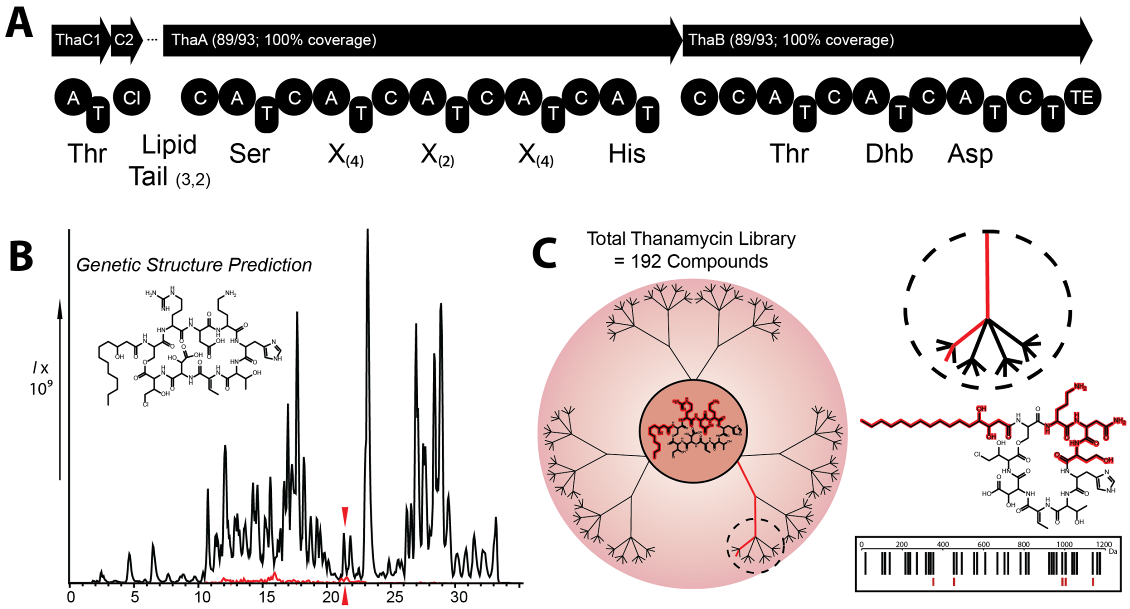

A
ThaC1
C2
ThaA (89/93; 100% coverage)
ThaB (89/93; 100% coverage)
A T Cl C A T C A T C A T C A T C A T C C A T C A T C A T C T TE
Thr
Lipid Tail (3,2)
Ser
X(4)
X(2)
X(4)
His
Thr
Dhb
Asp
B
Genetic Structure Prediction
I x 10⁹
0
5
10
15
20
25
30
C
Total Thanamycin Library = 192 Compounds
0
200
400
600
800
1000
1200
Da

Figure 6.6. Detection and de novo structure solution of the cryptic nonribosomal peptide thanamycin.

A) The recently sequenced genome of Pseudomonas fluorescens DSM 11579 reveals that it is closely related to Pseudmonas sp. SHC52, and contains a matching gene cluster for the cryptic nonribosomal peptide thanamycin. Standard genetic prediction yields three potential points of variation, as well as a potentially variable lipid tail. B) The genetic prediction of thanamycin was used to screen P.fluorescens DSM 1179 extracts for potential ledas, revealing one candidate compound present in vanishing quantities (red). C) Combinatorialization of four sites of low prediction confidence (red) enables the generation of hypothetical chemical barcodes. Analysis of thanamycin enriched fractions with this hypothetical library enabled the detection and de novo structure solution of the cryptic NRP thanamycin, later confirmed by NMR.

6.5 Discussion

Identifying molecular patterns and pharmacophores is a critical process during the discovery and development of natural product leads[2, 4, 44]. Such chemical features define points of entry either for synthetic efforts[4, 45], or for expanded natural product searches to probe specific regions of chemical space and realize the potential of the lead molecule as a therapeutic agent[13-15]. For instance, during the development of glycopeptide antibiotics, natural product libraries were extensively screened to identify novel variants with improved therapeutic potential [46]. Expansion of the chemical space of promising anticancer natural products, including those active against taxol-resistant tumors (e.g. hemiasterlins), has been achieved through screening and synthetic efforts to produce improved chemical entities [47-49]. Several promising known natural products with unique pharmacophores, such as HUN-7293, bengamide, and abyssomicin, still occupy sparsely populated areas of chemical space, indicating that systematic discovery efforts are likely to yield new related compounds with improved therapeutic values. Although explorations of chemical space have previously been conducted through bioactivity-guided approaches, such methodologies are time-consuming and relatively untargeted[17, 18]. While directed bioactivity screens of natural product extracts can provide data on specific queries, such as inhibition of targets of interest, these methods neglect and discard valuable information regarding the chemical structures and pharmacophores of hits, the presence of novel structural variants, and the location and relative abundance of lead compounds[19, 23]. Informatic search methods like the one presented here can effectively detect patterns and pharmacophores of interest from within complex extracts using MS/MS information collected during LC-MS analysis. Our approach expedites the discovery of targeted bioactive pharmacophores and structural variations, dereplicates known compounds that would otherwise complicate subsequent analysis, and facilitates preliminary structure elucidation of novel compounds through statistically validated matching of observed and hypothetical chemical barcodes.

In several instances in this work, we have detected novel genetically-predicted compounds from sequenced genomes based on similarity of biosynthetic gene clusters and presumed sharing of chemical space, validating the utility of our approach for using genomic data to profile metabolomes. Though most of these compounds were subsequently characterized by NMR, our barcoding strategy has provided partial or total structural identification in many instances, without the explicit need for isolation and further characterization. We have also used expansive barcode libraries to populate hypothetical areas of chemical space occupied by a specific family of natural products and identify novel related compounds, and confirmed this relatedness by genome sequencing. Finally, the value of this program for detecting desired pharmacophores has been demonstrated through the selective identification of minor structural variants bearing pre-selected site specific modifications, providing unobscured access to nature's combinatorial prowess. Within the provided examples of large libraries of nonribosomal peptide chemical barcodes, we have observed the robust and accurate dereplication of knowns and detection of unknowns corresponding to their respective libraries (**Supplementary Fig. 17**). These experiments have revealed the efficacy of this approach for detecting desired or unique agents with defined molecular patterns and pharmacophores, which serve to impart natural products with their sought after activities.

Peptide natural products exhibit a broad range of bioactivities and are widely used within clinical settings as antibiotic-, anticancer-, and immunomodulatory-agents[1, 5, 7]. Screening large collections of natural product extract libraries for bioactivity is the only strategy that has been systematically applied to expand and explore natural product chemical space[3, 7]. We propose that targeted discovery of peptide drug leads may also be realized using informatic search strategies to automatically detect desired pharmacophores, genetically envisioned families of 'cryptic' natural products, or site specific modifications that may improve pharmacological properties.

6.6 Materials and Methods

6.6.1 General Experimental Procedures

1D (^{1}H and ^{13}C) and 2D (^{1}H-^{13}C HMBC, HSQC, NOESY, and COSY) NMR spectra were recorded on a Bruker AVIII 700 MHz NMR spectrometer in D_2O (D_2O; Cambridge Isotope Laboratories). High resolution MS spectra were collected on a Thermo LTQ OrbiTrap XL mass spectrometer (*ThermoFisher Scientific, USA*) with an electrospray ionization source (ESI) and using CID with helium for fragmentation. LCMS data was collected using a Bruker AmazonX ion trap mass spectrometer coupled with a Dionex UltiMate 3000 HPLC system, using an Ascentis Express C_{18} column (150 mm x 4.6 mm, *Supelco*) or Luna C18 column (150 mm or 250 mm x 4.6 mm, *Phenomenex*) for analytical separations, running acetonitrile with 0.1% formic acid and ddH_2O with 0.1% formic acid as the mobile phase.

6.6.2 Microbial Strains

Acidovorax citrulli AAC00-1 and *Variovorax paradoxus* S110 were ordered from the German Resource Centre for Biological Material (DSMZ, DSM No. **17060 and 30034**) and cultured on Acidovorax Complex Media[50] (ACM) plates at 30°C. Environmental isolates including strain P4B were found in soil samples collected around McMaster University from June to August 2010 and maintained on casitone yeast extract (CYE) or tryptic soy broth (TSB) media. Environmental isolate *Elaphocordyceps* sp. RKGE-151 was isolated from brown algae collected from Prince Edward Island, Canada. Isolate *Hypocrea minutispora* RKDO-344 was isolated from Great Slave Lake, Northwest Territories, Canada. *Streptomyces sp.* used for screening were obtained from other laboratories and strain repositories including DSMZ and ATCC. *Streptomyces calvus* was obtained from DSMZ (DSM No. 40010) and was cultured on mannitol soya agar. The thanamycin-producing strain of *Pseudomonas fluorescens* was obtained from DSM (DSM No. 11579) and was regularly maintained on LB agar at 30°C.

6.6.3 Fermentation and Small Molecule Isolation

RKDO-344 and RKGE-151 was inoculated from a 5 day shaking culture in SMYA media (10 g/L peptone, 40 g/L maltose, 10 g/L yeast extract) at 22°C into MMK2 media (40 g/L mannitol, 5 g/L yeast extract, 4.3 g/L murashuge and Skoog salts) and grown standing at 22°C at a 20 degree angle. Cultures were extracted with 5% XAD7 and 5% HP20 activated resins. Extracts were subjected to LC-MS/MS analysis. The mobile phase was 2% acetonitrile until 5min and increased nonlinearly (curve 7) to 100% acetonitrile at 25 min and was held for an additional 5 min. Trichopolyn 1 eluted at 28.81 min and efrapeptin F eluted at 25.33 min. Dissolved RKDO-344 extract in 8:2 H_2O:MeOH was fractionated over a C18 SEP-PAK. Elution was stepwise with: 8:2 H_2O :MeOH 2) 1:1 H_2O :MeOH 3) EtOH 4) 1:1 DCM: MeOH. Trichopolyn 1 eluted in fraction 3.

A colony from a fresh plate of *A. citrulli* AAC00-1*,* and *V. paradoxus* S110 was inoculated into a 2.8 L glass Fernbach flask containing 1 L of Acidovorax Complete Media (ACM)[50]. Environmental strain *V. paradoxus* P4b was inoculated from a fresh plate into a 2.8 L glass Fernbach flask containing 1 L water, 10 g casitone, 1 g $MgSO_4$ x 7 H_2O, 1 g $CaCl_2$ x 2 H_2O, 50 mM Hepes buffer, and 20 g/L HP20 resin (Dialon) with pH adjusted to 7.0[51]. All cultures were grown at 30°C, shaking at 190 rpm for three days, after which *A. citrulli* AAC00-1 and *V. paradoxus* S110 cells were pelleted by centrifugation at 7000 rpm for 15 min. HP20 resin (*Dialon*) was added to the *A. citrulli* AAC00-1 and *V. paradoxus* S110 supernatant at 20 g/L and shaken for ~2 h at 220 rpm. The resin for all was harvested by Buchner funnel filtration and washed with 400 mL of distilled water. The resin was eluted three times with 400 mL of methanol. The methanol eluent was evaporated to dryness under rotary vacuum. Acidobactin A, B, C, and D were purified using a Luna 5 μm C_{18} column (250 x 10.0 mm, *Phenomenex*). The mobile phase was 2 % acetonitrile with 0.1% formic acid, and 98 % water with 0.1% formic acid at 2 minutes, increasing along curve 7 to 9 % acetonitrile at 23 min at a flow rate of 6 mL/min. Acidobactin A eluted at 15.5 min, acidobactin B eluted at 15.9 min, vacidobactin A eluted at 15.7 min, and vacidobactin B eluted at 16.2 min. Variobactin was purified using a Luna 5 μm C_{18} column (250 x 15.0 mm, *Phenomenex*). The mobile phase was 5 % acetonitrile with 0.1% formic acid, and 95 % water with 0.1% formic acid at 0 min with a flow rate of 2.5 mL/min increasing to 8 mL/min at 1.5min for an additional 3.5 min. The gradient increased linear from 5 to 10 min to 10% acetonitrile then from 10-52 min the gradient was linear to 50% acetonitrile. Variobactin A eluted at 38.03 min.

Single colonies of *S. calvus* were used to initiate 50mL cultures of TSB, and grown for several days at 28°C and 200 rpm. For production of WS-9326A, 10mL of starter culture was inoculated into 1 L of production media (10 g potato dextrin, 10 g peptone, 2 g NaCl, 2 g ammonium phosphate dibasic, 1.5 g potassium phosphate monobasic, 0.5 g potassium phosphate dibasic, 0.25 $MgSO_4$ x 7 H_2O, and 5mL of trace element solution [2 g/L $MgSO_4$, 2 g/L $ZnSO_4$ x 7 H_2O, 2 g/L $FeSO_4$ x 7 H_2O, 2 g/L

$MnCl_2$ x 4 H_2O, 2 g/L $CaCl_2$ x 2 H_2O, 2 g/L NaCl, 0.4 g/L $CuCl_2$ x 2 H_2O, 0.4 g/L boric acid, 0.2 g/L sodium molybdenate hydrate, 0.2 g/L $CoCl_2$, and 2.2 g/L sodium citrate.], and grown for three days at 225 rpm and 28°C. Cultures were harvested by extracting twice with 2:1 ethyl acetate and evaporating until dry. Culture extracts were resuspended in methanol and applied to an open column of LH20 resin in methanol. Fractions containing WS-9326A were pooled and dried, resuspended in methanol, and analyzed by LCMS, using a Luna 5 µm C_{18} column (250 x 10.0 mm, *Phenomenex*) and mobile phases of acetonitrile with 0.1% formic acid, and water with 0.1% formic acid. To optimize detection of WS-9326 analogs, a method was devised with a flow rate of 1.4 mL/min, starting at 5% acetonitrile for the first 4 min, ramping with curve 7 to 42% acetonitrile by 14 min, slowly ramping with curve 7 to 53% acetonitrile by 50 min, and finally ramping with curve 7 to 100% acetonitrile by 60 min. WS-9326A eluted at 32.8 min, and the 1009 *m*/*z* analogue eluted at 28.7 min.

For production of thanamycin, *P. fluorescens* was grown in 400 mL of YM media per 2.8 L Fernbach flask, or in 40 mL YM media per 250 mL Erlenmeyer flask. Cultures were repeatedly grown at 22° and 250 rpm for 48 hr until 40 L of material was accumulated. Cultures were harvested by centrifugation at 7000 rpm, followed by methanol extraction of the cell pellet, and HP20 extraction of the supernatant. Methanol eluent of the HP20 resin was pooled with the methanol extract of the cell pellet and dried under rotary vacuum, and resuspended in methanol. Extracts were analyzed by LCMS with a 150 mm Luna C18 column with the mobile phase flowing at 1 ml/min, ramping from 2% acetonitrile at 5 minutes to 100% acetonitrile at 25 minutes (curve 7). LH20 column chromatography with methanol as the mobile phase was used to generate thanamycin enriched fractions. Fractions were analyzed again using LCMS with a 150 mm Luna C18 column with the mobile phase flowing at 1 ml/min, ramping from 5% acetonitrile at 4 minutes to 100% at 22 minutes (curve 5). Preparative HPLC was used to isolate thanamycin, using a 250 mm x 10 mm Luna C_{18} column with the mobile phase flowing at 8 ml/min, starting with 5% acetonitrile for the first 5 minutes, ramping to 35%

by 10 minutes and maintaining that until 35 minutes, followed by a wash of 100% acetonitrile.

6.6.4 Structure Elucidation

See supplementary information for details.

6.6.5 Genome Sequencing

A single colony of environmental isolate P4B was grown in 3mL TSB overnight at 30°C, 250 rpm. Genomic DNA was harvested using a GenElute Bacterial Genomic DNA Kit (Sigma). Genomic DNA was sent for library preparation and Illumina sequencing at the Farncombe Metagenomics Facility at McMaster University, using an Illumina MiSeq DNA sequencer. Contigs were assembled using the ABySS genome assembly program and with Geneious bioinformatic software.

6.6.6 Identification of delftibactin biosynthetic gene cluster and adenylation domain specificity

Homologous delftibactin NRPS gene clusters were found in *A. citrulli* AAC00-1, *V. paradoxus* S110 and *V. paradoxus* EPS using the BLAST function of IMG (http://img.jgi.doe.gov), using the *delG* sequence as the query. Adenylation domain specificities were assessed using NRPS Predictor or NRPS-PKS, and the 10 residue codes of each entry and its top scoring hit were recorded and compared to the delftibactin adenylation code (see supplementary information for more details)[52-54].

6.6.7 MAUVE alignment of biosynthetic gene clusters

Gene cluster alignments of *D. acidovorax* SPH-1 (Daci_4756-4753), *V. paradoxus* EPS (Varpa_4327-4324), *V. paradoxus* S110 (Vapar_3746-3742), *A. citrulli* AAC00-1 (3733-3729), and environmental isolate *V. paradoxus* P4b (*varC-I)* were carried out in Geneious software (v5.6.4) using a progressive Mauve algorithm plugin with a seed weight of 20 and a local collinear block setting of 3000[55]**.**

6.6.8 iSNAP dereplication of trichopolyn 1 and efrapeptin F

Trichopolyn 1 and Efrapeptin F were identified from environmental extract 344-M3 and GE-151 respectively through the iSNAP program for dereplication and is summarized briefly below[29]. The iSNAP nonribosomal peptide SMILES database was assembled from NORINE, Pubchem, and J of Antibiotics databases, among others. This conglomerate database has been periodically updated since its assembly by Ibrahim and Yang and now includes >1100 chemical structures in SMILES code. Each of these structures are fragmented at amide bonds and neutral loss functional groups to generate a library of hypothetical structural fragments (hSFs) that would be diagnostic of the real fragmented NRP. This approach was validated for a diverse array of peptide architectures, including cyclic, branched, and linear structures containing proteinogenic and nonproteinogenic amino acids.

6.6.9 iSNAP trichopolyn variant identification.

A structural database of all aminoisobutyric acid (aib) and alanine combinations was created for the trichopolyn scaffolds based on trichopolyn 1 and trichopolyn B. In addition, all structural combinations of valine and isoleucine were also included to afford a final structural database consisting of 254 compounds. This combinatorial database was facilitated through the use of SmiLib v2.0 online software[33]. The 254 trichopolyn variant database was uploaded onto iSNAP and analysis was performed on the LC-MS/MS mzxmL file for the 344-M3 extract with the mass window set to one, affording only direct mass hits from the extract to the database. Structural confirmation was carried out through manual MS2 annotation, iSNAP fragment hit analysis, and high resolution mass spectrometry.

6.6.10 iSNAP analysis of *A. citrulli* AAC00-1 extract and identification of acidobactin A and B

The acidobactin prediction database was constructed similar to above using the *A. citrulli* AAC00-1 gene cluster prediction as the scaffold (**Fig. 3, 4**). Variants included both cyclic and linear structures with variations of the ornithine groups (hydroxylation, formylation, and acetylation) and the polyketide portion (malonate or methyl malonate), which afforded a library of 576 compounds. The 576 compound acidobactin prediction database was uploaded onto iSNAP and analysis was performed on the LC-MS/MS mzxmL file for the *A. citrulli* AAC00-1 extract with the mass window set to fifty without P1/P2 score cutoffs. All acidobactin prediction library iSNAP scan hits were summed for each 0.25 min in retention time and plotted against retention time. These iSNAP hit frequency plots were overlaid with LC-MS/MS chromatograms for compound peak identification using Adobe Illustrator CS6. Variants of acidobactin were identified using the final structure of acidobactin A as the scaffold for variant library generation, resulting in a library of 72 compounds. This identified three variant analogs (54, 48, 56), which corresponded to acidobactin A, B and putative acidobactin C. Subsequent MS2 fragment analysis confirmed these as the true structures (**Supplementary Fig. 16**).

6.6.11 iSNAP analysis of *V. paradoxus* S110 extract and identification of vacidobactin A and B

The *V. paradoxus* S110 extract was analyzed similar to *A. citrulli* AAC00-1 using the acidobactin prediction library and an iSNAP mass window of 50 without P1/P2 score cutoffs (**Supplementary Fig. 9**).

6.6.12 iSNAP identification of the delftibactin-acidobactin-vacidobactin-like compound, variobactin A

The combined prediction database was compiled using the structure of delftibactin A, the predicted structure of the *V. paradoxus* EPS, and *A. citrulli* AAC00-1 gene cluster.

Variants included both cyclic and linear structures for *V. paradoxus* EPS and *A. citrulli* AAC00-1 and linear structures for delftibactin A with variations on the ornithine groups (hydroxylation, formylation, and acetylation) and the polyketide portion (malonate or methyl malonate) afforded a combined prediction library of 14, 592 compounds (**Fig. 5**).

Extracts generated from a bacterial environmental library consisting of 80 unknown organisms were analyzed analytically by LC-MS/MS similar to the trichopolyn and efrapeptin producer extracts. Base peak ion chromatograms were converted to mzxml format using CompassXport and uploaded onto iSNAP where they were analyzed using the combined prediction database with a window of 50. All combined prediction library iSNAP scan hits were summed for each 0.25 min in retention time and plotted against retention time. These iSNAP hit frequency plots were overlaid with LC-MS/MS chromatograms for each environmental extract using Adobe Illustrator CS6. Strain P4b was identified from the library based on the high frequency of iSNAP hits around an unknown metabolite peak. This peak was revealed to be the novel compound, variobactin A (**Fig. 5**). Variobactin A variants were identified similarly to acidobactin A variants, with the exception that the final structure of variobactin A was used as the scaffold for library generation, resulting in a total of 216 variant structures (**Supplementary Fig. 11**).

6.6.13 iSNAP identification of WS-9326A and Neurokinin A-like Analogs

During the course of screening *Streptomyces* extracts we identified WS-9326A from an extract of *S. calvus*, which was not previously known to produce the WS-9326 series of compounds. This was done using the standard iSNAP nonribosomal peptide SMILES database as outlined above. To detect analogs with increased homology to neurokinin A, a targeted library of 16 hypothetical structures was constructed to include: a serine or threonine at position 1, a valine or leucine at position 3, a serine or threonine at position 5, and a serine or threonine at position 7. This tailored 16 compound library was uploaded onto iSNAP and analysis was performed on the LC-MS/MS mzxmL file for the *S. calvus* extract with the mass window set to one without P1/P2 score cutoffs. WS-

9326A variants were identified and the retention time is indicated by red lines overlaid on the LC-MS/MS chromatogram using Adobe Illustrator CS6.

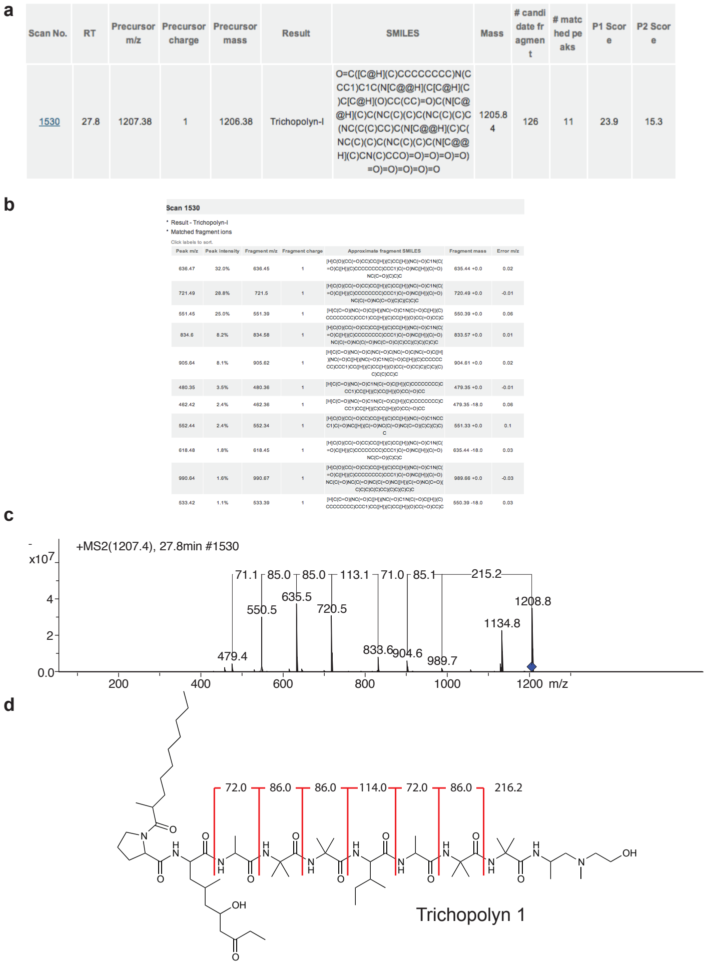
6.6.14 Molecular barcode identification of thanamycin

A genetic prediction of the thanamycin structure was generated using previously described techniques, including the use of NRPSPredictor to arrive at plausible amino acid monomers. A C_{12} 3-hydroxy fatty acid was selected for the prediction based on its presence in related chlorinated lipopeptides from *Pseudomonas sp*. The predicted structure was used to query a *Pseudomonas* extract previously generated for production of the unrelated nonribosomal peptide SB-253514, using a mass window of 50 Da without P1/P2 score cutoffs, and yielded a hit at 21.16 min to a doubly charged peptide with a molecular weight of 1273 Da. Four monomers with low prediction confidence in this hit were selected for combinatorialization, generating 96 hypothetical chemical structures. This hypothetical library was used to query a new thanamycin enriched chromatogram in hopes of improving the structure, using a stricter mass window of 20 Da without P1/P2 score cutoffs. The predominant result (4 of 11 library hits) was a hypothetical structure with modifications to the lipid tail and each of the three combinatorialized amino acids. This lead structure had a mass difference from the putative thanamycin hit corresponding to an errant hydroxyl group. Addition of the hydroxyl group to the lipid tail of the hypothetical structure library generated another 96 hypothetical structures and facilitated the efficient detection of thanamycin with a mass window of 1 Da with standard P1/P2 score cutoffs, suggesting the proposed hypothetical was indeed the final structure. This suggestion was subsequently validated by NMR.

6.6.15 Validation of iSNAP Structural Library Specificity

The specificity of each library was verified by comparison to each extract analyzed in this study (**Supplementary Fig. 6.18**). Each iSNAP analysis indicated above was reexamined using the other iSNAP structure databases developed in this work. Comparison of the iSNAP hit frequency plots indicates mutual exclusivity of the databases for only the extracts that contain similar compounds.

6.7 Supplementary Information



Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	# candidate fragment	# matched peaks	P1 Score	P2 Score
1530	27.8	1207.38	1	1206.38	Trichopolyn-I	O=C([C@H](C)CCCCCCCC)N(CCC1)C1C(N[C@@H](C[C@H](C)C[C@H](O)CC(CC)=O)C(N[C@@H](C)C(NC(C)(C)C(NC(C)(C)C(NC(C(C)CC)C(N[C@@H](C)C(NC(C)(C)C(NC(C)(C)C(N[C@@H](C)CN(C)CCO)=O)=O)=O)=O)=O)=O)=O)=O)=O	1205.84	126	11	23.9	15.3

Scan 1530

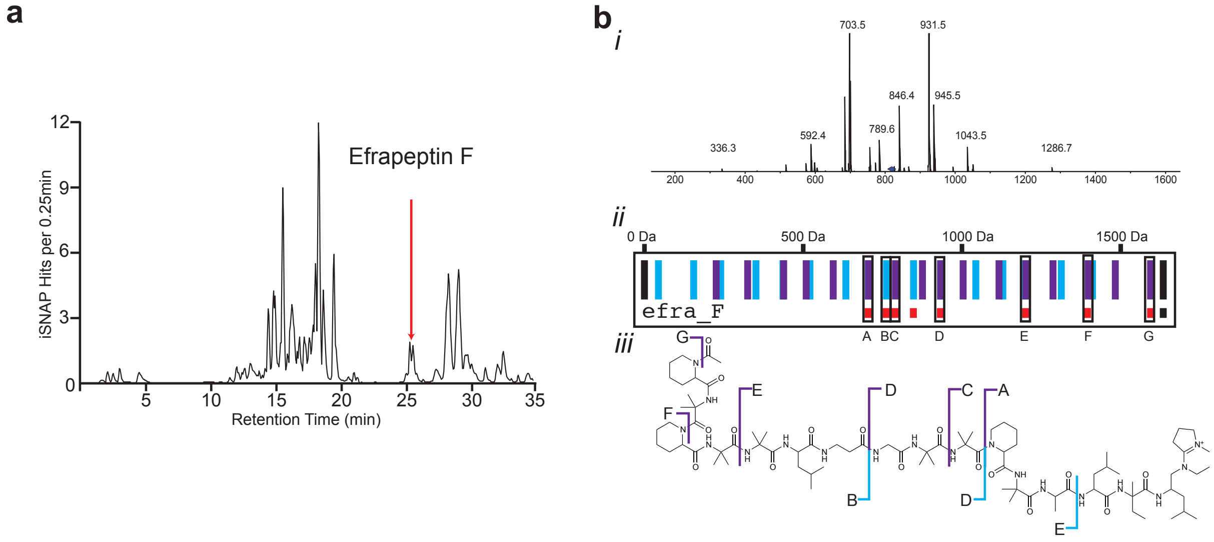
- Result - Trichopolyn-I
- Matched fragment ions

Click labels to sort.

Peak m/z	Peak intensity	Fragment m/z	Fragment charge	Approximate fragment SMILES	Fragment mass	Error m/z
636.47	32.0%	636.45	1	[H]C(O)(CC(=O)CC)CC([H])(C)CC([H])(NC(=O)C1N(C(=O)C([H])(C)CCCCCCCC)CCC1)C(=O)NC([H])(C(=O)NC(C=O)(C)C)C	635.44 +0.0	0.02
721.49	28.8%	721.5	1	[H]C(O)(CC(=O)CC)CC([H])(C)CC([H])(NC(=O)C1N(C(=O)C([H])(C)CCCCCCCC)CCC1)C(=O)NC([H])(C(=O)NC(C(=O)NC(C=O)(C)C)(C)C)C	720.49 +0.0	-0.01
551.45	25.0%	551.39	1	[H]C(C=O)(NC(=O)C([H])(NC(=O)C1N(C(=O)C([H])(C)CCCCCCCC)CCC1)CC([H])(C)CC([H])(O)CC(=O)CC)C	550.39 +0.0	0.06
834.6	8.2%	834.58	1	[H]C(O)(CC(=O)CC)CC([H])(C)CC([H])(NC(=O)C1N(C(=O)C([H])(C)CCCCCCCC)CCC1)C(=O)NC([H])(C(=O)NC(C(=O)NC(C(=O)NC(C=O)C(C)CC)(C)C)(C)C)C	833.57 +0.0	0.01
905.64	8.1%	905.62	1	[H]C(C=O)(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C([H])(NC(=O)C([H])(NC(=O)C1N(C(=O)C([H])(C)CCCCCCCC)CCC1)CC([H])(C)CC([H])(O)CC(=O)CC)C)(C)C)(C)C)C(C)CC)C	904.61 +0.0	0.02
480.35	3.5%	480.36	1	[H]C(C=O)(NC(=O)C1N(C(=O)C([H])(C)CCCCCCCC)CCC1)CC([H])(C)CC([H])(O)CC(=O)CC	479.35 +0.0	-0.01
462.42	2.4%	462.36	1	[H]C(C=O)(NC(=O)C1N(C(=O)C([H])(C)CCCCCCCC)CCC1)CC([H])(C)CC([H])(O)CC(=O)CC	479.35 -18.0	0.06
552.44	2.4%	552.34	1	[H]C(O)(CC(=O)CC)CC([H])(C)CC([H])(NC(=O)C1NCCC1)C(=O)NC([H])(C(=O)NC(C(=O)NC(C=O)(C)C)(C)C)C	551.33 +0.0	0.1
618.48	1.8%	618.45	1	[H]C(O)(CC(=O)CC)CC([H])(C)CC([H])(NC(=O)C1N(C(=O)C([H])(C)CCCCCCCC)CCC1)C(=O)NC([H])(C(=O)NC(C=O)(C)C)C	635.44 -18.0	0.03
990.64	1.6%	990.67	1	[H]C(O)(CC(=O)CC)CC([H])(C)CC([H])(NC(=O)C1N(C(=O)C([H])(C)CCCCCCCC)CCC1)C(=O)NC([H])(C(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC([H])(C(=O)NC(C=O)(C)C)C)(C)C)C(C)CC)(C)C)(C)C)C	989.66 +0.0	-0.03
533.42	1.1%	533.39	1	[H]C(C=O)(NC(=O)C([H])(NC(=O)C1N(C(=O)C([H])(C)CCCCCCCC)CCC1)CC([H])(C)CC([H])(O)CC(=O)CC)C	550.39 -18.0	0.03

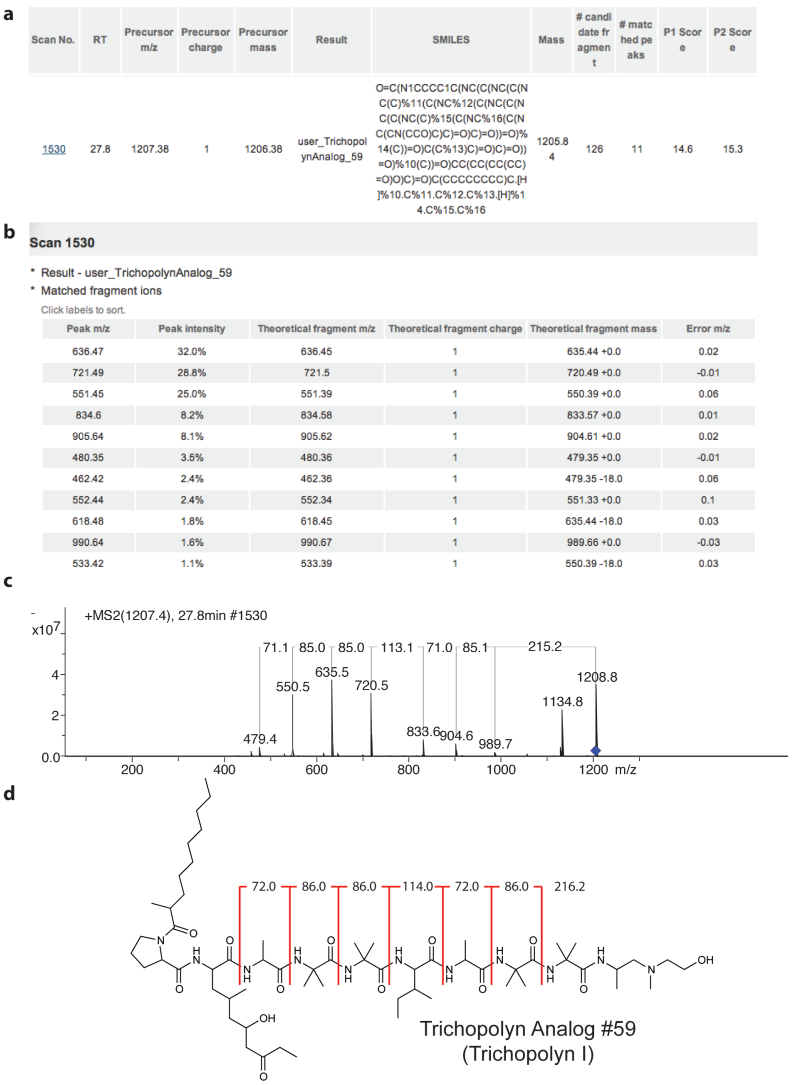
Supplementary Figure 6.1. Trichopolyn 1 iSNAP hit.

A) Screen shot of trichopolyn 1 iSNAP hit in 344 M3 extract. B) iSNAP fragmentation match screenshot of scan 1530 to trichopolyn 1 theoretical fragmentations. C) MS2 fragmentation of scan 1530 (trichopolyn 1), neutral losses are indicated. D) Structure and fragmentation pattern of trichopolyn 1.



Supplementary Figure 6.2. Identification of efrapeptin F from extract bioactivity screen using informatic search for natural products (iSNAP).

A) LC-MS base peak chromatogram of environmental extract GE-151 is shown with efrapeptin F iSNAP scan hit. B) Fragmentation pattern of efrapeptin f. *i)* MS2 spectra of efrapeptin F hit, scan 1243. *ii*) Predicted b (blue) and y (purple) ion fragmentation of efrapeptin F compared to iSNAP matched fragment hits (red). *iii*) B and Y ion fragment matches are indicated below on the structure of efrapeptin F.



Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	# candidate fragment	# matched peaks	P1 Score	P2 Score
1530	27.8	1207.38	1	1206.38	user_TrichopolynAnalog_59	O=C(N1CCCC1C(NC(C(NC(C(NC(C)%11(C(NC%12(C(NC(C(NC(C(NC(C)%15(C(NC%16(C(NC(CN(CCO)C)C)=O)C)=O))=O)%14(C))=O)C(C%13)C)=O)C)=O))=O)%10(C))=O)CC(CC(CC(CC)=O)O)C)=O)C(CCCCCCCC)C.[H]%10.C%11.C%12.C%13.[H]%14.C%15.C%16	1205.84	126	11	14.6	15.3

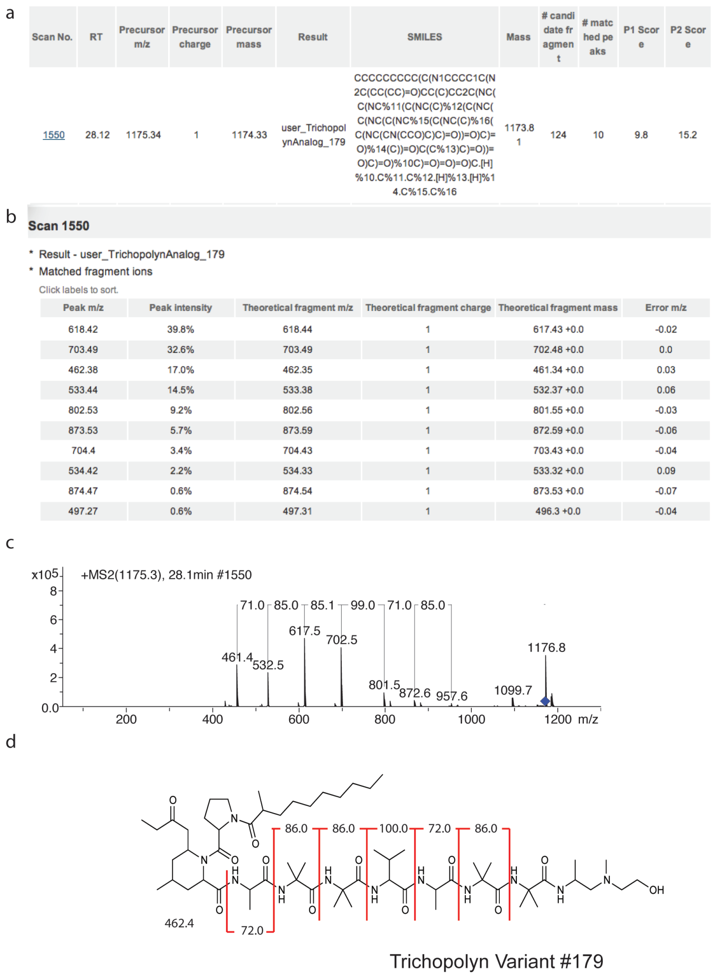
Scan 1530

- Result - user_TrichopolynAnalog_59
- Matched fragment ions

Click labels to sort.

Peak m/z	Peak intensity	Theoretical fragment m/z	Theoretical fragment charge	Theoretical fragment mass	Error m/z
636.47	32.0%	636.45	1	635.44 +0.0	0.02
721.49	28.8%	721.5	1	720.49 +0.0	-0.01
551.45	25.0%	551.39	1	550.39 +0.0	0.06
834.6	8.2%	834.58	1	833.57 +0.0	0.01
905.64	8.1%	905.62	1	904.61 +0.0	0.02
480.35	3.5%	480.36	1	479.35 +0.0	-0.01
462.42	2.4%	462.36	1	479.35 -18.0	0.06
552.44	2.4%	552.34	1	551.33 +0.0	0.1
618.48	1.8%	618.45	1	635.44 -18.0	0.03
990.64	1.6%	990.67	1	989.66 +0.0	-0.03
533.42	1.1%	533.39	1	550.39 -18.0	0.03

Supplementary Figure 6.3. Trichopolyn variant 59 (trichopolyn 1) iSNAP hit. A) Screen shot of trichopolyn variant 59 iSNAP hit in 344 M3 extract. B) iSNAP fragmentation match screenshot of scan 1530 to trichopolyn variant 59 theoretical fragmentations. C) MS2 fragmentation of scan 1530(trichopolyn variant 59/trichopolyn 1), neutral losses are indicated. D) Structure and fragmentation pattern of trichopolyn variant 59 (trichopolyn 1).



Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	# candidate fragment	# matched peaks	P1 Score	P2 Score
1550	28.12	1175.34	1	1174.33	user_TrichopolynAnalog_179	CCCCCCCCC(C(N1CCCC1C(N2C(CC(CC)=O)CC(C)CC2C(NC(C(NC%11(C(NC(C)%12(C(NC(C(NC(C(NC%15(C(NC(C)%16(C(NC(CN(CCO)C)C)=O))=O)C)=O)%14(C))=O)C(C%13)C)=O))=O)C)=O)%10C)=O)=O)=O)C.[H]%10.C%11.C%12.[H]%13.[H]%14.C%15.C%16	1173.81	124	10	9.8	15.2

Scan 1550

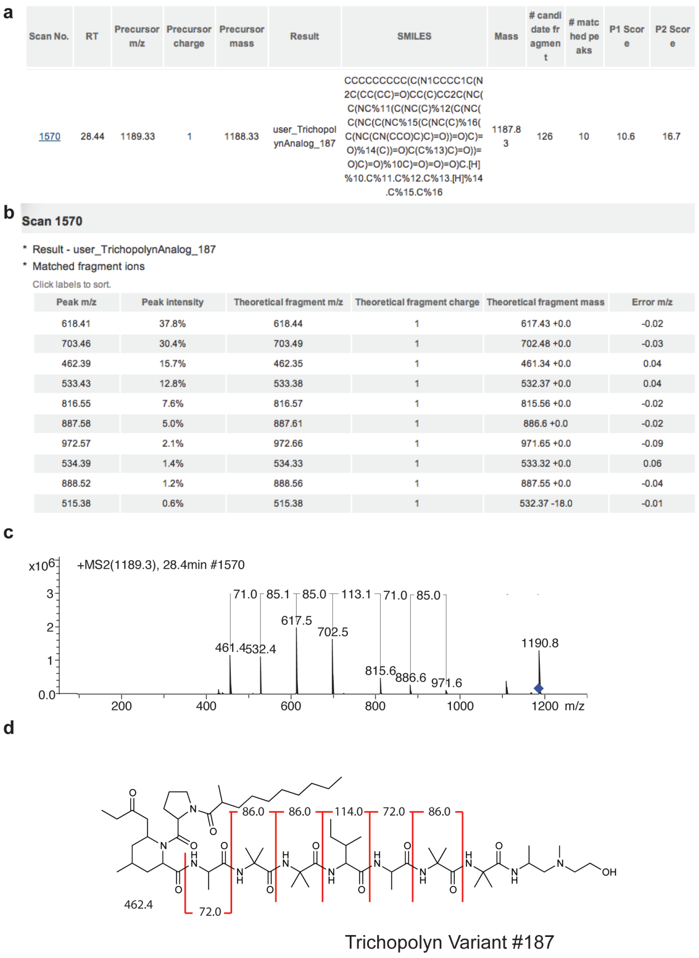
- Result - user_TrichopolynAnalog_179
- Matched fragment ions

Click labels to sort.

Peak m/z	Peak intensity	Theoretical fragment m/z	Theoretical fragment charge	Theoretical fragment mass	Error m/z
618.42	39.8%	618.44	1	617.43 +0.0	-0.02
703.49	32.6%	703.49	1	702.48 +0.0	0.0
462.38	17.0%	462.35	1	461.34 +0.0	0.03
533.44	14.5%	533.38	1	532.37 +0.0	0.06
802.53	9.2%	802.56	1	801.55 +0.0	-0.03
873.53	5.7%	873.59	1	872.59 +0.0	-0.06
704.4	3.4%	704.43	1	703.43 +0.0	-0.04
534.42	2.2%	534.33	1	533.32 +0.0	0.09
874.47	0.6%	874.54	1	873.53 +0.0	-0.07
497.27	0.6%	497.31	1	496.3 +0.0	-0.04

Supplementary Figure 6.4. Trichopolyn variant 179 iSNAP hit.

A) Screen shot of trichopolyn variant 179 iSNAP hit in *Trichoderma* ***X*** extract. B) iSNAP fragmentation match screenshot of scan 1550 to trichopolyn variant 179 theoretical fragmentations. C) MS2 fragmentation of scan 1550 (trichopolyn variant 179), neutral losses are indicated. D) Structure and fragmentation pattern of trichopolyn variant 179.



Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	# candidate fragment	# matched peaks	P1 Score	P2 Score
1570	28.44	1189.33	1	1188.33	user_TrichopolynAnalog_187	CCCCCCCCC(C(N1CCCC1C(N2C(CC(CC)=O)CC(C)CC2C(NC(C(NC%11(C(NC(C)%12(C(NC(C(NC(C(NC%15(C(NC(C)%16(C(NC(CN(CCO)C)C)=O))=O)C)=O)%14(C))=O)C(C%13)C)=O))=O)C)=O)%10C)=O)=O)=O)C.[H]%10.C%11.C%12.C%13.[H]%14.C%15.C%16	1187.83	126	10	10.6	16.7

Scan 1570

- Result - user_TrichopolynAnalog_187
- Matched fragment ions

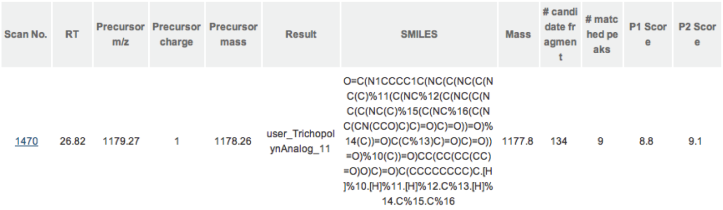
Click labels to sort.

Peak m/z	Peak intensity	Theoretical fragment m/z	Theoretical fragment charge	Theoretical fragment mass	Error m/z
618.41	37.8%	618.44	1	617.43 +0.0	-0.02
703.46	30.4%	703.49	1	702.48 +0.0	-0.03
462.39	15.7%	462.35	1	461.34 +0.0	0.04
533.43	12.8%	533.38	1	532.37 +0.0	0.04
816.55	7.6%	816.57	1	815.56 +0.0	-0.02
887.58	5.0%	887.61	1	886.6 +0.0	-0.02
972.57	2.1%	972.66	1	971.65 +0.0	-0.09
534.39	1.4%	534.33	1	533.32 +0.0	0.06
888.52	1.2%	888.56	1	887.55 +0.0	-0.04
515.38	0.6%	515.38	1	532.37 -18.0	-0.01

Supplementary Figure 6.5. Trichopolyn variant 187 iSNAP hit.

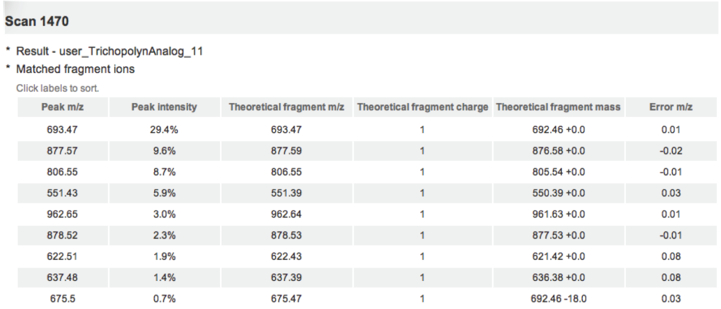
A) Screen shot of trichopolyn variant 187 iSNAP hit in *Trichoderma* ***X*** extract. B) iSNAP fragmentation match screenshot of scan 1570 to trichopolyn variant 187 theoretical fragmentations. C) MS2 fragmentation of scan 1570 (trichopolyn variant 187), neutral losses are indicated. D) Structure and fragmentation pattern of trichopolyn variant 187.

a



Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	# candidate fragment	# matched peaks	P1 Score	P2 Score
1470	26.82	1179.27	1	1178.26	user_TrichopolynAnalog_11	O=C(N1CCCC1C(NC(C(NC(C(NC(C)%11(C(NC%12(C(NC(C(NC(C(NC(C)%15(C(NC%16(C(NC(CN(CCO)C)C)=O)C)=O))=O)%14(C))=O)C(C%13)C)=O)C)=O))=O)%10(C))=O)CC(CC(CC(CC)=O)O)C)=O)C(CCCCCCCC)C.[H]%10.[H]%11.[H]%12.C%13.[H]%14.C%15.C%16	1177.8	134	9	8.8	9.1

b



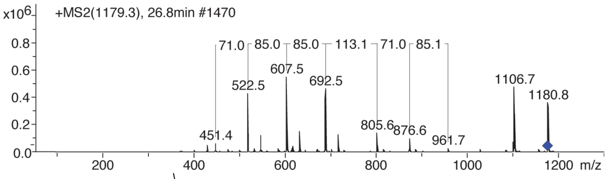
Scan 1470

- Result - user_TrichopolynAnalog_11
- Matched fragment ions

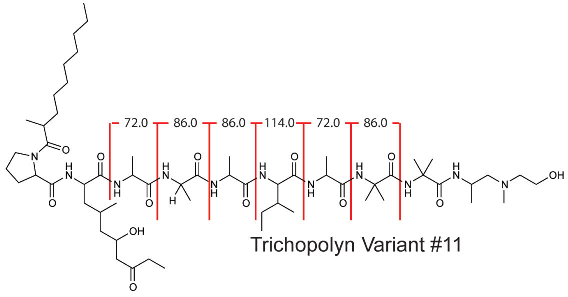
Click labels to sort.

Peak m/z	Peak intensity	Theoretical fragment m/z	Theoretical fragment charge	Theoretical fragment mass	Error m/z
693.47	29.4%	693.47	1	692.46 +0.0	0.01
877.57	9.6%	877.59	1	876.58 +0.0	-0.02
806.55	8.7%	806.55	1	805.54 +0.0	-0.01
551.43	5.9%	551.39	1	550.39 +0.0	0.03
962.65	3.0%	962.64	1	961.63 +0.0	0.01
878.52	2.3%	878.53	1	877.53 +0.0	-0.01
622.51	1.9%	622.43	1	621.42 +0.0	0.08
637.48	1.4%	637.39	1	636.38 +0.0	0.08
675.5	0.7%	675.47	1	692.46 -18.0	0.03

c

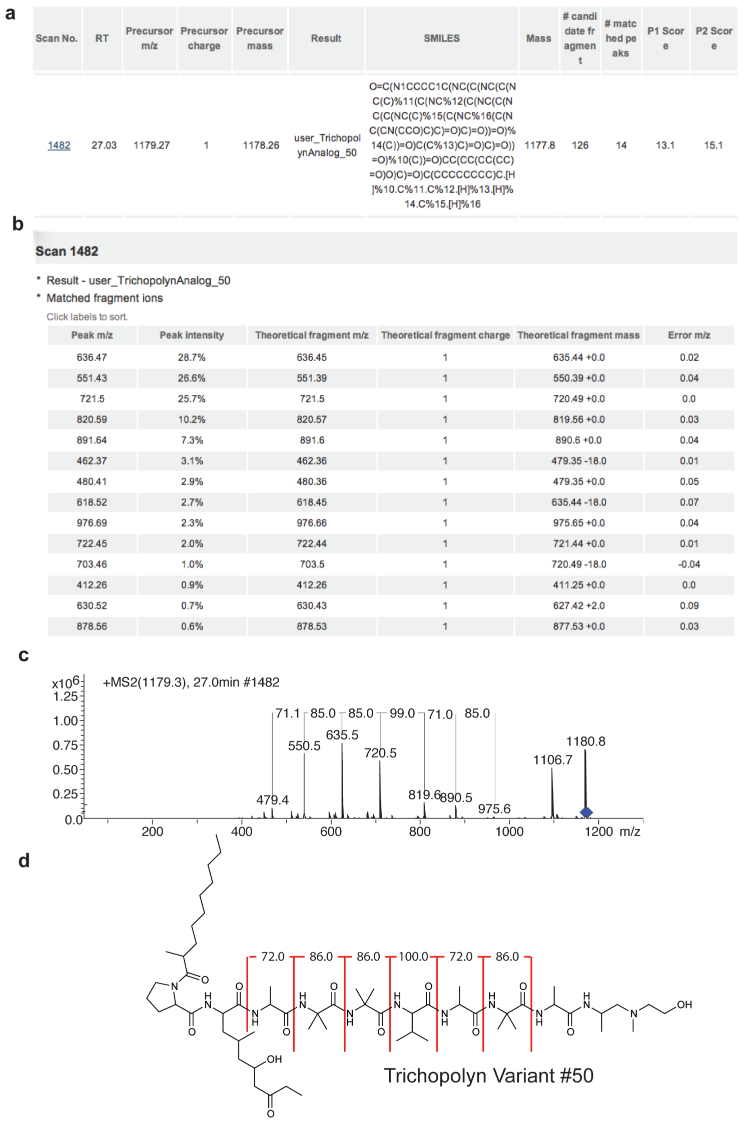


d



Supplementary Figure 6.6. Trichopolyn variant 11 iSNAP hit.

A) Screen shot of trichopolyn variant 11 iSNAP hit in *Trichoderma* ***X*** extract. B) iSNAP fragmentation match screenshot of scan 1470 to trichopolyn variant 11 theoretical fragmentations. C) MS2 fragmentation of scan 1470 (trichopolyn variant 11), neutral losses are indicated. D) Structure and fragmentation pattern of trichopolyn variant 11.



Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	# candidate fragment	# matched peaks	P1 Score	P2 Score
1482	27.03	1179.27	1	1178.26	user_TrichopolynAnalog_50	O=C(N1CCCC1C(NC(C(NC(C(NC(C)%11(C(NC%12(C(NC(C(NC(C(NC(C)%15(C(NC%16(C(NC(CN(CCO)C)C)=O)C)=O))=O)%14(C))=O)C(C%13)C)=O)C)=O))=O)%10(C))=O)CC(CC(CC(CC)=O)O)C)=O)C(CCCCCCCC)C.[H]%10.C%11.C%12.[H]%13.[H]%14.C%15.[H]%16	1177.8	126	14	13.1	15.1

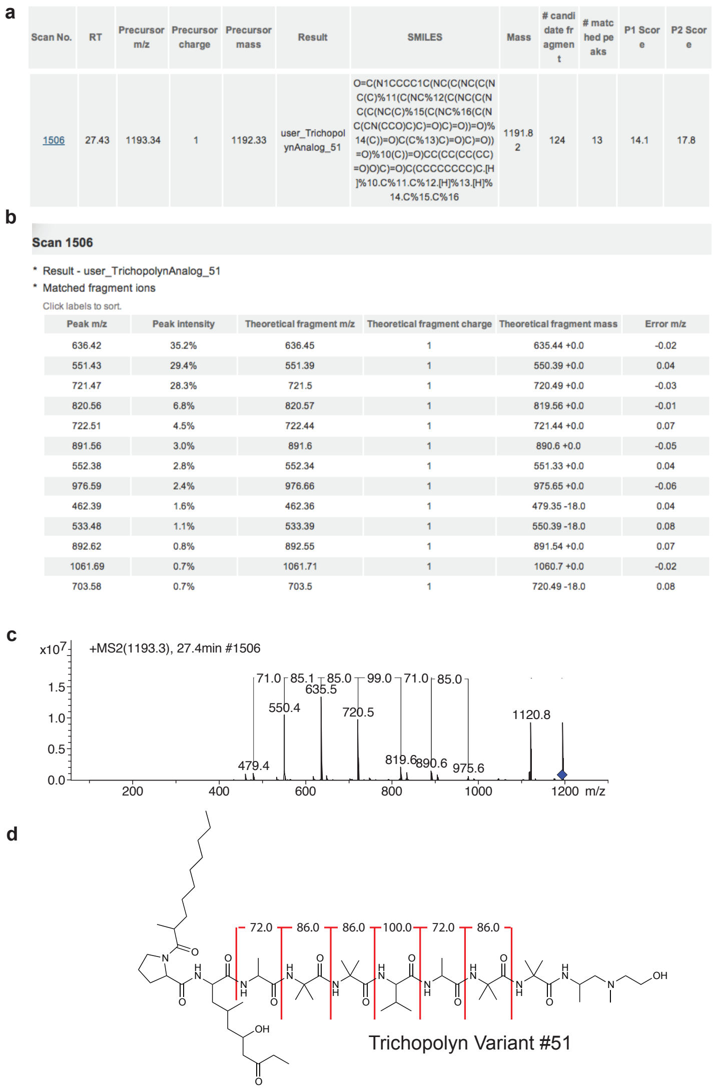
Scan 1482

- Result - user_TrichopolynAnalog_50
- Matched fragment ions

Click labels to sort.

Peak m/z	Peak intensity	Theoretical fragment m/z	Theoretical fragment charge	Theoretical fragment mass	Error m/z
636.47	28.7%	636.45	1	635.44 +0.0	0.02
551.43	26.6%	551.39	1	550.39 +0.0	0.04
721.5	25.7%	721.5	1	720.49 +0.0	0.0
820.59	10.2%	820.57	1	819.56 +0.0	0.03
891.64	7.3%	891.6	1	890.6 +0.0	0.04
462.37	3.1%	462.36	1	479.35 -18.0	0.01
480.41	2.9%	480.36	1	479.35 +0.0	0.05
618.52	2.7%	618.45	1	635.44 -18.0	0.07
976.69	2.3%	976.66	1	975.65 +0.0	0.04
722.45	2.0%	722.44	1	721.44 +0.0	0.01
703.46	1.0%	703.5	1	720.49 -18.0	-0.04
412.26	0.9%	412.26	1	411.25 +0.0	0.0
630.52	0.7%	630.43	1	627.42 +2.0	0.09
878.56	0.6%	878.53	1	877.53 +0.0	0.03

Supplementary Figure 6.7. Trichopolyn variant 50 (trichopolyn 4) iSNAP hit. A) Screen shot of trichopolyn variant 50 iSNAP hit in *Trichoderma* ***X*** extract. B) iSNAP fragmentation match screenshot of scan 1482 to trichopolyn variant 50 (trichopolyn 4) theoretical fragmentations. C) MS2 fragmentation of scan 1482 (trichopolyn variant 179/ trichopolyn 4), neutral losses are indicated. D) Structure and fragmentation pattern of trichopolyn variant 50 (trichopolyn 4).



Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	# candidate fragment	# matched peaks	P1 Score	P2 Score
1506	27.43	1193.34	1	1192.33	user_TrichopolynAnalog_51	O=C(N1CCCC1C(NC(C(NC(C(NC(C)%11(C(NC%12(C(NC(C(NC(C(NC(C)%15(C(NC%16(C(NC(CN(CCO)C)C)=O)C)=O))=O)%14(C))=O)C(C%13)C)=O)C)=O))=O)%10(C))=O)CC(CC(CC(CC)=O)O)C)=O)C(CCCCCCCC)C.[H]%10.C%11.C%12.[H]%13.[H]%14.C%15.C%16	1191.82	124	13	14.1	17.8

Scan 1506

* Result - user_TrichopolynAnalog_51
* Matched fragment ions

Click labels to sort.

Peak m/z	Peak intensity	Theoretical fragment m/z	Theoretical fragment charge	Theoretical fragment mass	Error m/z
636.42	35.2%	636.45	1	635.44 +0.0	-0.02
551.43	29.4%	551.39	1	550.39 +0.0	0.04
721.47	28.3%	721.5	1	720.49 +0.0	-0.03
820.56	6.8%	820.57	1	819.56 +0.0	-0.01
722.51	4.5%	722.44	1	721.44 +0.0	0.07
891.56	3.0%	891.6	1	890.6 +0.0	-0.05
552.38	2.8%	552.34	1	551.33 +0.0	0.04
976.59	2.4%	976.66	1	975.65 +0.0	-0.06
462.39	1.6%	462.36	1	479.35 -18.0	0.04
533.48	1.1%	533.39	1	550.39 -18.0	0.08
892.62	0.8%	892.55	1	891.54 +0.0	0.07
1061.69	0.7%	1061.71	1	1060.7 +0.0	-0.02
703.58	0.7%	703.5	1	720.49 -18.0	0.08

Supplementary Figure 6.8. Trichopolyn variant 51 (trichopolyn 2) iSNAP hit. A) Screen shot of trichopolyn variant 51 iSNAP hit in *Trichoderma* ***X*** extract. B) iSNAP fragmentation match screenshot of scan 1506 to trichopolyn variant 51 (trichopolyn 2) theoretical fragmentations. C) MS2 fragmentation of scan 1506 (trichopolyn variant 51/ trichopolyn 2), neutral losses are indicated. D) Structure and fragmentation pattern of trichopolyn variant 51 (trichopolyn 2).

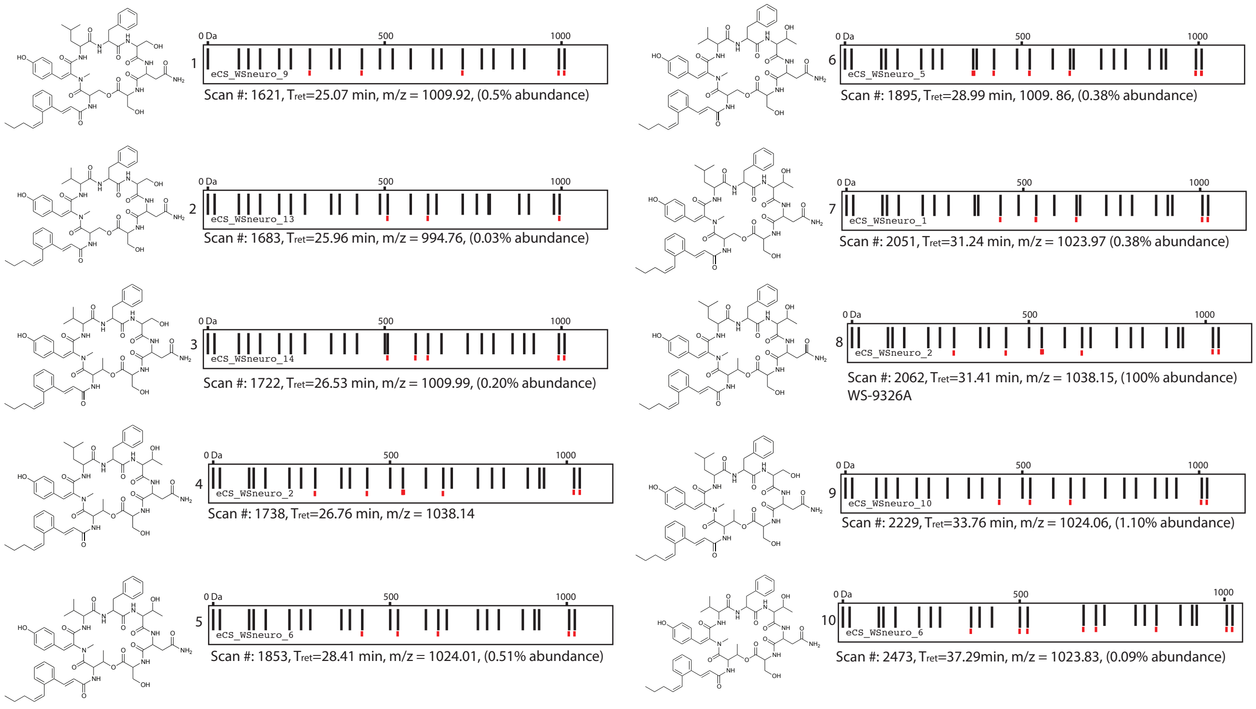
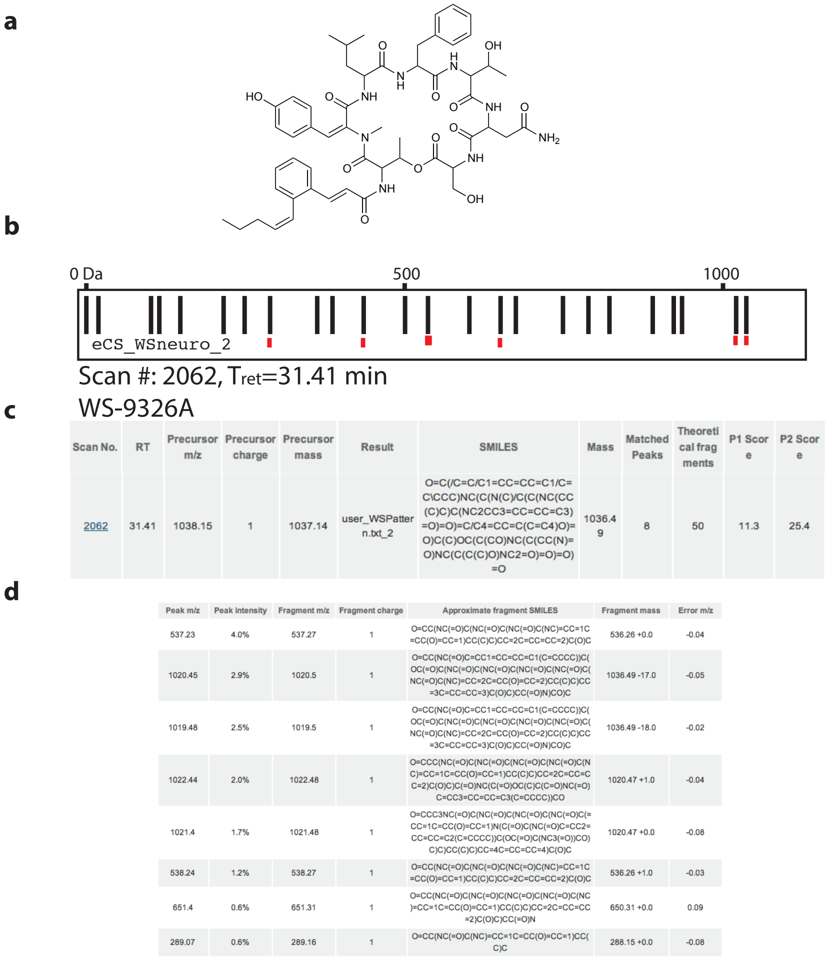

1
0 Da
500
1000
eCS_WSneuro_9
Scan #: 1621, Tret=25.07 min, m/z = 1009.92, (0.5% abundance)
2
eCS_WSneuro_13
Scan #: 1683, Tret=25.96 min, m/z = 994.76, (0.03% abundance)
3
eCS_WSneuro_14
Scan #: 1722, Tret=26.53 min, m/z = 1009.99, (0.20% abundance)
4
eCS_WSneuro_2
Scan #: 1738, Tret=26.76 min, m/z = 1038.14
5
eCS_WSneuro_6
Scan #: 1853, Tret=28.41 min, m/z = 1024.01, (0.51% abundance)
6
eCS_WSneuro_5
Scan #: 1895, Tret=28.99 min, 1009. 86, (0.38% abundance)
7
eCS_WSneuro_1
Scan #: 2051, Tret=31.24 min, m/z = 1023.97 (0.38% abundance)
8
eCS_WSneuro_2
Scan #: 2062, Tret=31.41 min, m/z = 1038.15, (100% abundance)
WS-9326A
9
eCS_WSneuro_10
Scan #: 2229, Tret=33.76 min, m/z = 1024.06, (1.10% abundance)
10
eCS_WSneuro_6
Scan #: 2473, Tret=37.29min, m/z = 1023.83, (0.09% abundance)
HO
NH
N
O
OH
NH₂

Figure 6.9. WS-Neurokinin Mimic Library variants found in *S. calvus* extract. WS-Neurokinin variants identified by iSNAP are shown with their theoretical fragment barcode (black) and the matched fragments identified by iSNAP underneath (red).



Scan #: 2062, T_{ret}=31.41 min

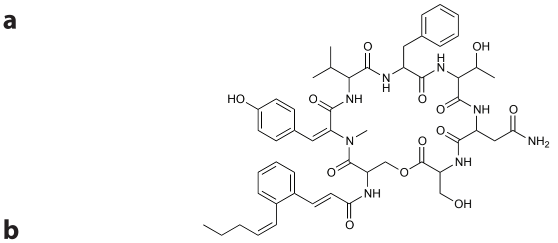
WS-9326A

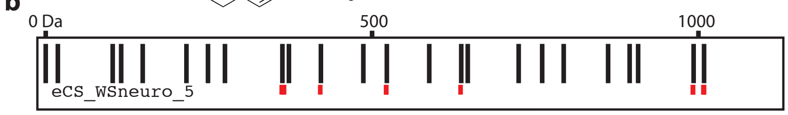
Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	Matched Peaks	Theoretical fragments	P1 Score	P2 Score
2062	31.41	1038.15	1	1037.14	user_WSPattern.txt_2	O=C(/C=C/C1=CC=CC=C1/C=C\CCC)NC(C(N(C)/C(C(NC(CC(C)C)C(NC2CC3=CC=CC=C3)=O)=O)=C/C4=CC=C(C=C4)O)=O)C(C)OC(C(CO)NC(C(CC(N)=O)NC(C(C(C)O)NC2=O)=O)=O)=O	1036.49	8	50	11.3	25.4

Peak m/z	Peak intensity	Fragment m/z	Fragment charge	Approximate fragment SMILES	Fragment mass	Error m/z
537.23	4.0%	537.27	1	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)CC(C)C)CC=2C=CC=CC=2)C(O)C	536.26 +0.0	-0.04
1020.45	2.9%	1020.5	1	O=CC(NC(=O)C=CC1=CC=CC=C1(C=CCCC))C(OC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=2C=CC(O)=CC=2)CC(C)C)CC=3C=CC=CC=3)C(O)C)CC(=O)N)CO)C	1036.49 -17.0	-0.05
1019.48	2.5%	1019.5	1	O=CC(NC(=O)C=CC1=CC=CC=C1(C=CCCC))C(OC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=2C=CC(O)=CC=2)CC(C)C)CC=3C=CC=CC=3)C(O)C)CC(=O)N)CO)C	1036.49 -18.0	-0.02
1022.44	2.0%	1022.48	1	O=CCC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)CC(C)C)CC=2C=CC=2)C(O)C)C(=O)NC(C(=O)OC(C)C(C=O)NC(=O)C=CC3=CC=CC=C3(C=CCCC))CO	1020.47 +1.0	-0.04
1021.4	1.7%	1021.48	1	O=CCC3NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(=CC=1C=CC(O)=CC=1)N(C(=O)C(NC(=O)C=CC2=CC=CC=C2(C=CCCC))C(OC(=O)C(NC3(=O))CO)C)C)CC(C)C)CC=4C=CC=CC=4)C(O)C	1020.47 +0.0	-0.08
538.24	1.2%	538.27	1	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)CC(C)C)CC=2C=CC=CC=2)C(O)C	536.26 +1.0	-0.03
651.4	0.6%	651.31	1	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)CC(C)C)CC=2C=CC=CC=2)C(O)C)CC(=O)N	650.31 +0.0	0.09
289.07	0.6%	289.16	1	O=CC(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)CC(C)C	288.15 +0.0	-0.08

Supplementary Figure 6.10.

iSNAP hit information for WS-9326A positive identification from WS-Neurokinin A Mimic Library Variants, including structure (a), barcode (black) with fragment hits (red)(b), iSNAP hit screenshot(c), and iSNAP fragmentation data screenshot (d)

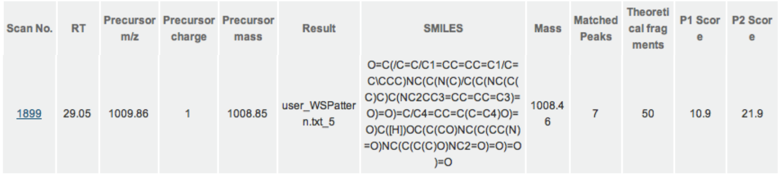




Scan #: 1895, T$_{ret}$=28.99 min

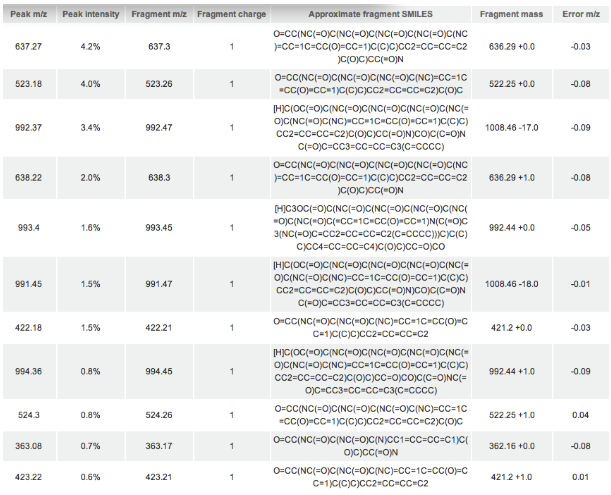
WS Neurokinin Mimic 5

c



Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	Matched Peaks	Theoretical fragments	P1 Score	P2 Score
1899	29.05	1009.86	1	1008.85	user_WSPattern.txt_5	O=C(/C=C/C1=CC=CC=C1/C=C\CCC)NC(C(N(C)/C(C(NC(C(C)C)C(NC2CC3=CC=CC=C3)=O)=O)=C/C4=CC=C(C=C4)O)=O)C([H])OC(C(CO)NC(C(CC(N)=O)NC(C(C(C)O)NC2=O)=O)=O)=O	1008.46	7	50	10.9	21.9

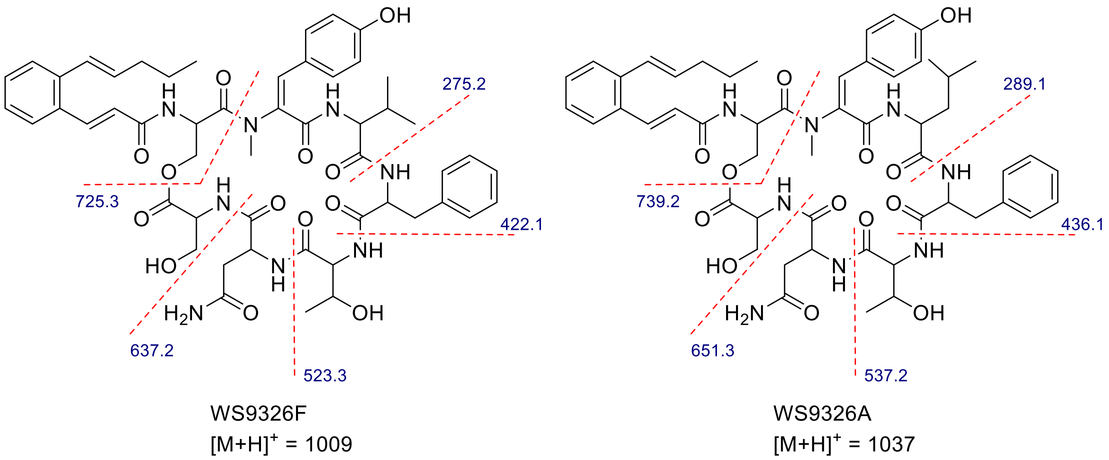
d



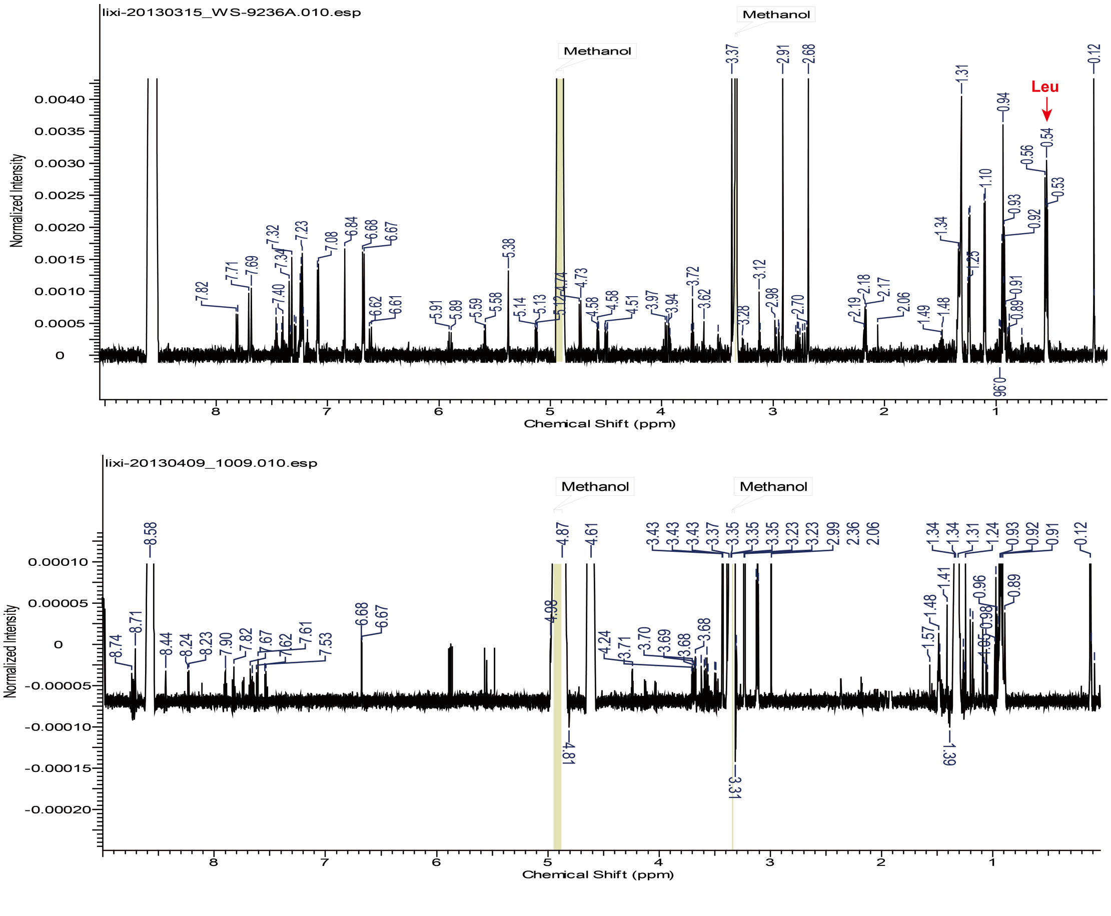
Peak m/z	Peak intensity	Fragment m/z	Fragment charge	Approximate fragment SMILES	Fragment mass	Error m/z
637.27	4.2%	637.3	1	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)C(C)C)CC2=CC=CC=C2)C(O)C)CC(=O)N	636.29 +0.0	-0.03
523.18	4.0%	523.26	1	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)C(C)C)CC2=CC=CC=C2)C(O)C	522.25 +0.0	-0.08
992.37	3.4%	992.47	1	[H]C(OC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)C(C)C)CC2=CC=CC=C2)C(O)C)CC(=O)N)CO)C(C=O)NC(=O)C=CC3=CC=CC=C3(C=CCCC)	1008.46 -17.0	-0.09
638.22	2.0%	638.3	1	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)C(C)C)CC2=CC=CC=C2)C(O)C)CC(=O)N	636.29 +1.0	-0.08
993.4	1.6%	993.45	1	[H]C3OC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(=CC=1C=CC(O)=CC=1)N(C(=O)C3(NC(=O)C=CC2=CC=CC=C2(C=CCCC)))C)C(C)C)CC4=CC=CC=C4)C(O)C)CC=O)CO	992.44 +0.0	-0.05
991.45	1.5%	991.47	1	[H]C(OC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)C(C)C)CC2=CC=CC=C2)C(O)C)CC(=O)N)CO)C(C=O)NC(=O)C=CC3=CC=CC=C3(C=CCCC)	1008.46 -18.0	-0.01
422.18	1.5%	422.21	1	O=CC(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)C(C)C)CC2=CC=CC=C2	421.2 +0.0	-0.03
994.36	0.8%	994.45	1	[H]C(OC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)C(C)C)CC2=CC=CC=C2)C(O)C)CC=O)CO)C(C=O)NC(=O)C=CC3=CC=CC=C3(C=CCCC)	992.44 +1.0	-0.09
524.3	0.8%	524.26	1	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)C(C)C)CC2=CC=CC=C2)C(O)C	522.25 +1.0	0.04
363.08	0.7%	363.17	1	O=CC(NC(=O)C(NC(=O)C(N)CC1=CC=CC=C1)C(O)C)CC(=O)N	362.16 +0.0	-0.08
423.22	0.6%	423.21	1	O=CC(NC(=O)C(NC(=O)C(NC)=CC=1C=CC(O)=CC=1)C(C)C)CC2=CC=CC=C2	421.2 +1.0	0.01

Supplementary Figure 6.11.

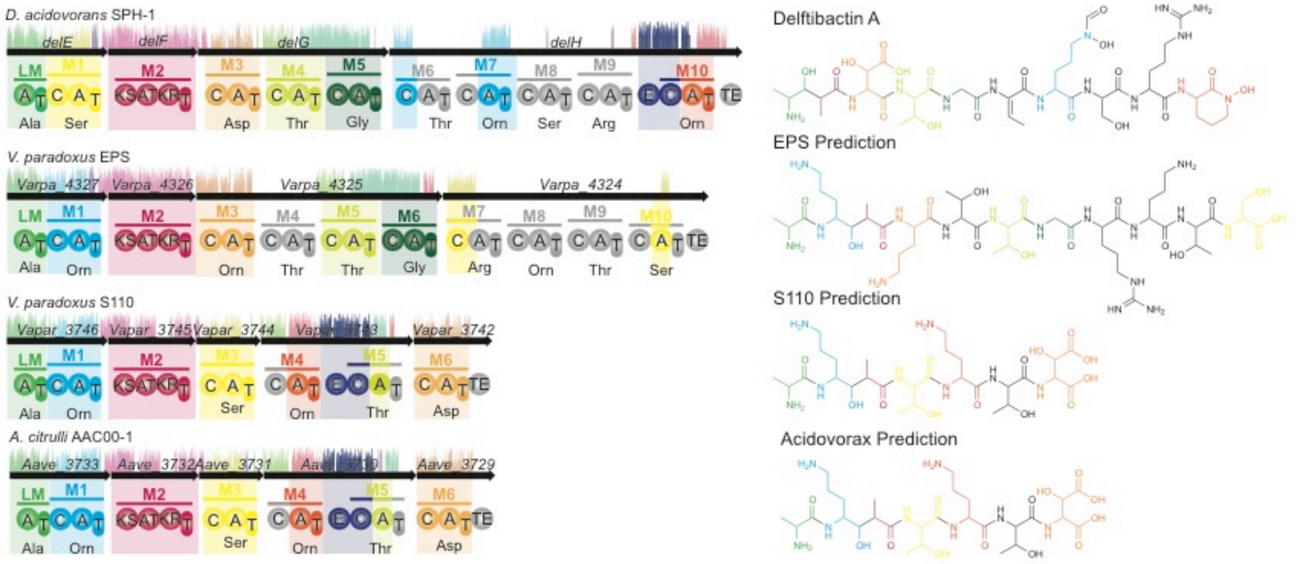
iSNAP hit information for WS-Neurokinin Mimic 5 positive identification from WS-Neurokinin A Mimic Library Variants, including structure (a), barcode (black) with fragment hits (red)(b), iSNAP hit screenshot(c), and iSNAP fragmentation data screenshot (d)



Supplementary Figure 6.12. MS fragmentation of WS9326F (left) and WS9326A (right).

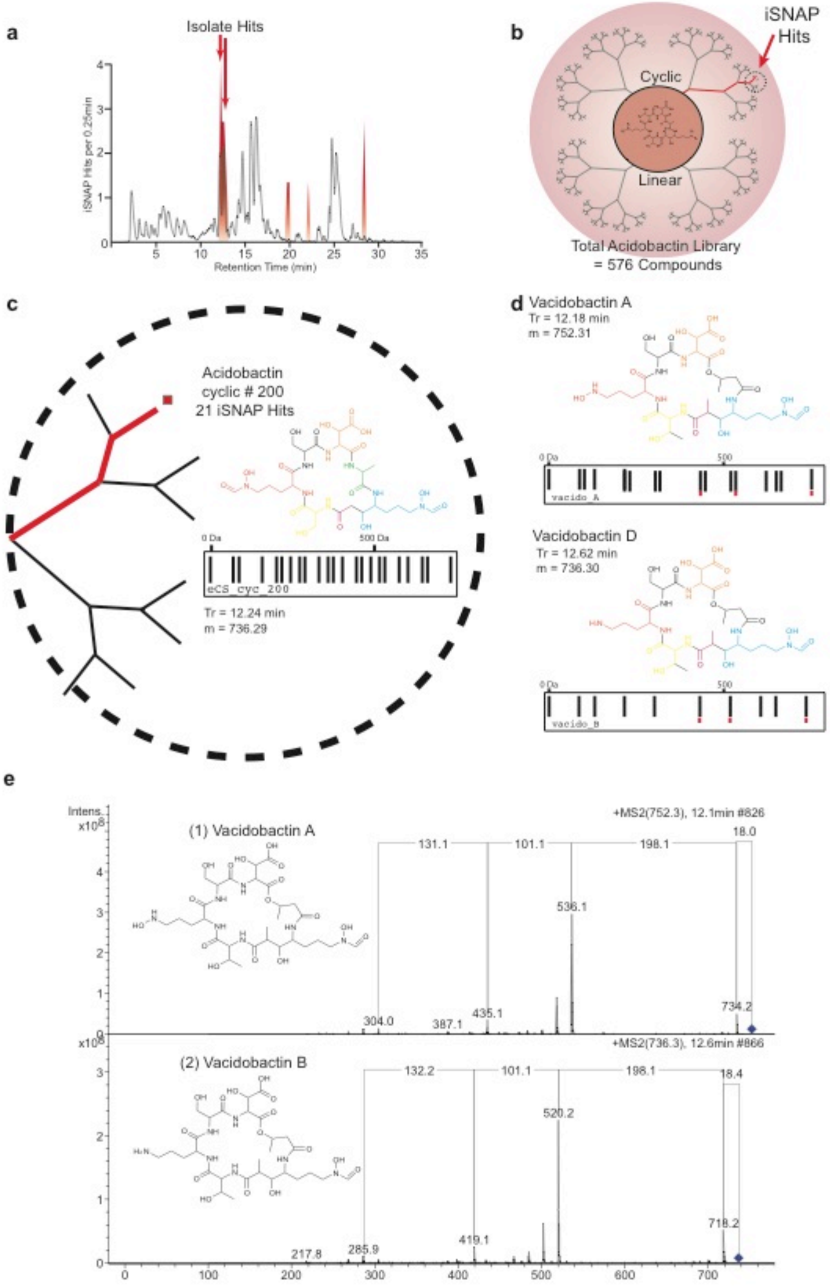


Supplementary Figure 6.13. ^{1}H NMR spectra comparison of WS9326A (Top) and novel WS compound WSneuro_5 (Bottom).



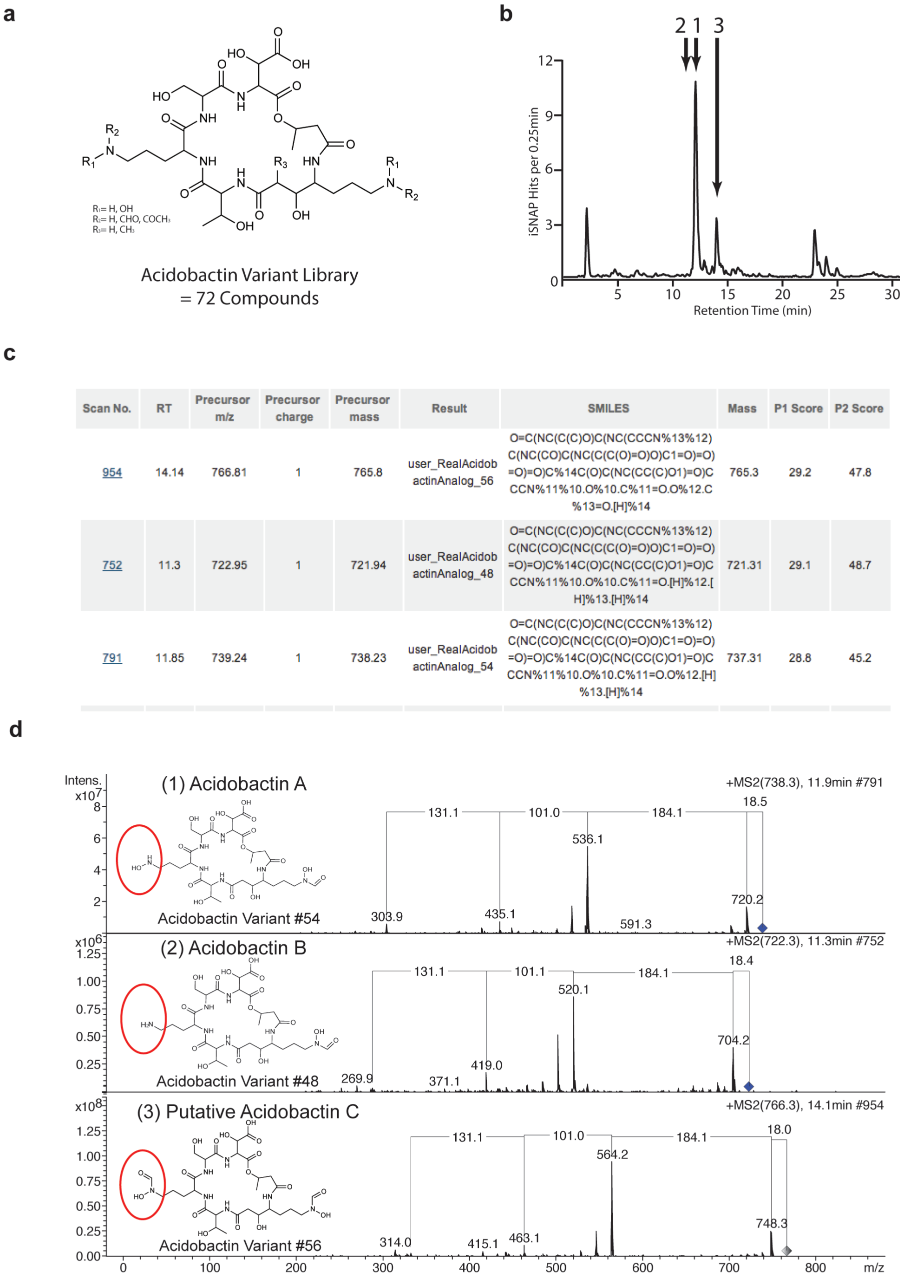
Supplementary Figure 6.14.

Mauve alignment of delftibactin gene cluster and related gene clusters found in *Variovorax paradoxus* EPS (NC_014931), *V. paradoxus* S110 (NC_012791), and *Acidovorax citrulli* AAC00-1 (NC_008752). Nonribosomal peptide and polyetide genes are shown in black and the protein domain architecture is shown, where A, T, C, KS, AT, KR, E and TE are adenylation, thiolation, condensation, ketosynthase, acyltransferase, ketoreductase, epimerase and thioesterase domains respectively. Genetic similarity is indicated above the gene by mauve analysis and highlighted on the protein domains in accordance with the encoded protein domains. The amino acid specificity for each A domain is indicated as a three letter amino acid code (for more details see Supplementary Information). The structure of delftibactin A is shown and the assembly-line predicted structures for the *V. paradoxus* EPS, *V. paradoxus* S110 and *A. citrulli* AAC00-1 gene clusters are shown.



Supplementary Figure 6.15. Identification of *Variovorax paradoxus* S110 nonribosomal peptide using iSNAP with a predicted acidobactin structure database.

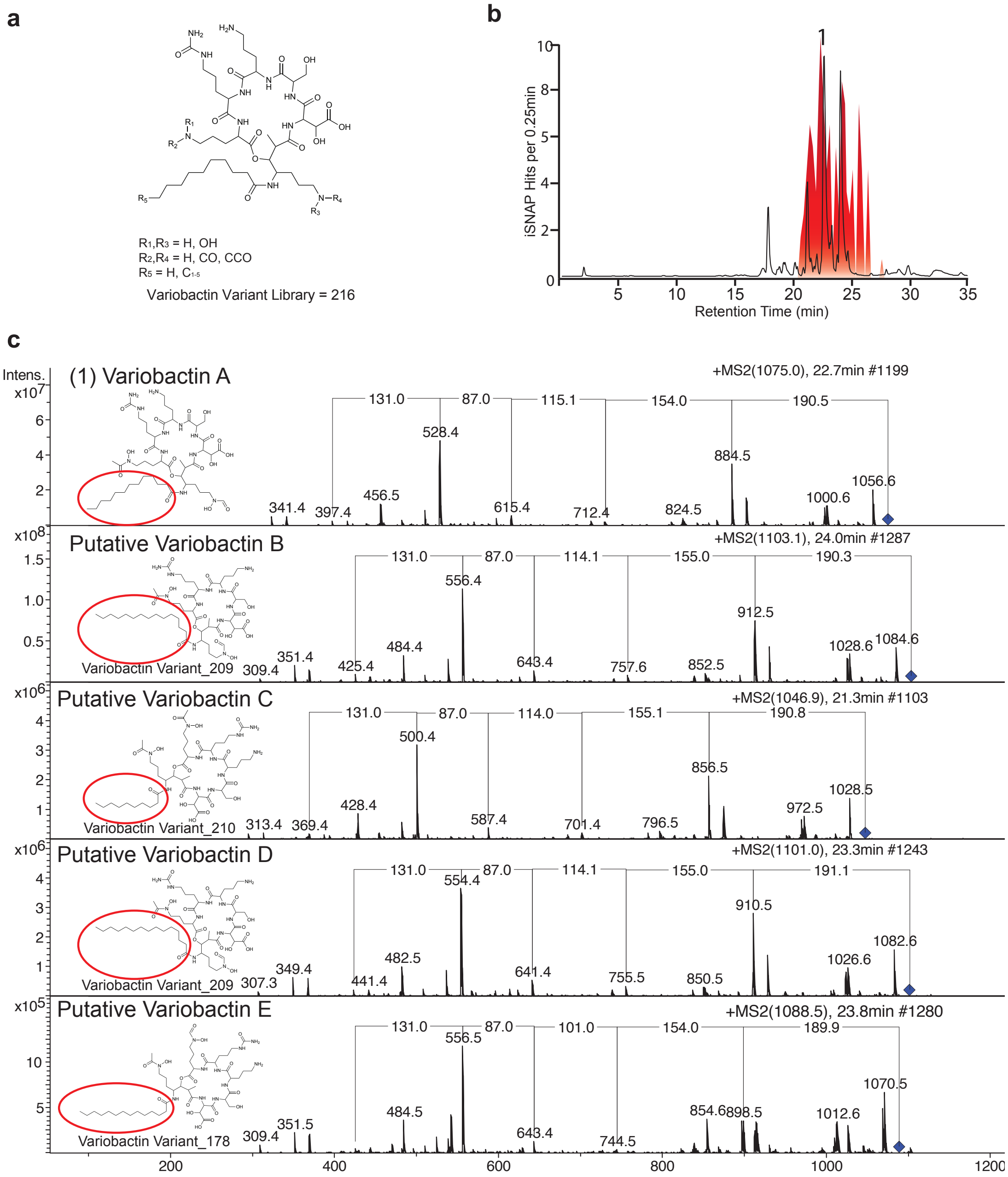
A) *V. paradoxus* S110 extract LC-MS base peak chromatogram overlaid with a frequency plot of predicted acidobactin library iSNAP hits per 0.25 min in LC retention time (mass window = 50). B) Fractal tree representation of acidobactin prediction compound library with the major iSNAP hit indicated on the tree. C) Close-up of predicted acidobactin fractal tree showing the structure of the major iSNAP hit, molecular weight, and retention time. D) Final structures, retention time, and molecular weight of isolated compounds, vacidobactin A and B. E) MS2 spectrum of vacidobactin A and B.



Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	P1 Score	P2 Score
954	14.14	766.81	1	765.8	user_RealAcidobactinAnalog_56	O=C(NC(C(C)O)C(NC(CCCN%13%12)C(NC(CO)C(NC(C(C(O)=O)O)C1=O)=O)=O)=O)C%14C(O)C(NC(CC(C)O1)=O)CCCN%11%10.O%10.C%11=O.O%12.C%13=O.[H]%14	765.3	29.2	47.8
752	11.3	722.95	1	721.94	user_RealAcidobactinAnalog_48	O=C(NC(C(C)O)C(NC(CCCN%13%12)C(NC(CO)C(NC(C(C(O)=O)O)C1=O)=O)=O)=O)C%14C(O)C(NC(CC(C)O1)=O)CCCN%11%10.O%10.C%11=O.[H]%12.[H]%13.[H]%14	721.31	29.1	48.7
791	11.85	739.24	1	738.23	user_RealAcidobactinAnalog_54	O=C(NC(C(C)O)C(NC(CCCN%13%12)C(NC(CO)C(NC(C(C(O)=O)O)C1=O)=O)=O)=O)C%14C(O)C(NC(CC(C)O1)=O)CCCN%11%10.O%10.C%11=O.O%12.[H]%13.[H]%14	737.31	28.8	45.2

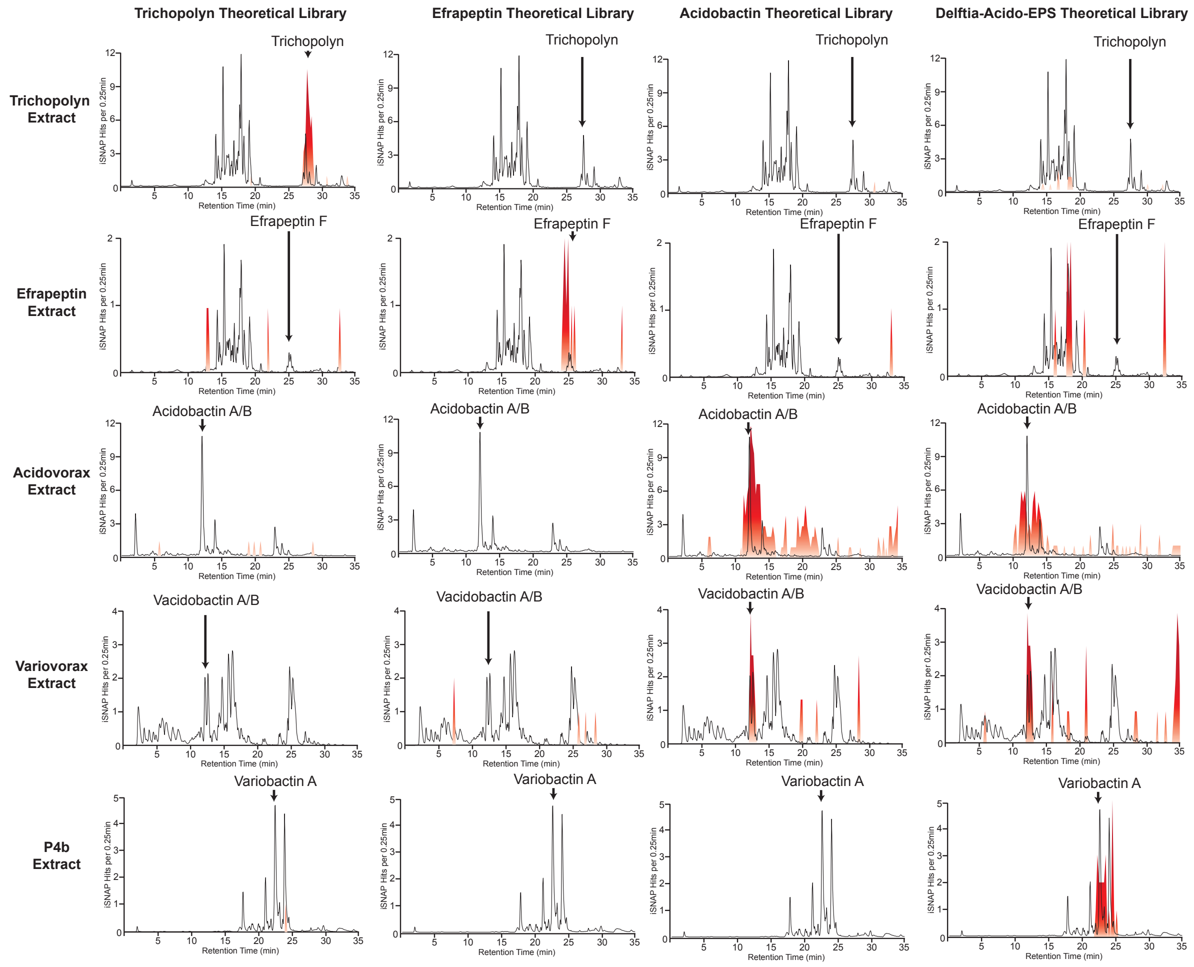
Supplementary Figure 6.16. Identification of acidobactin variants with iSNAP using an acidobactin variant library.

A) Acidobactin variant library was generated using acidobactin A as the scaffold as indicated. B) *A. citrulli* AAC00-1 extract LC-MS base peak chromatogram with iSNAP acidobactin variant hits shown. C) Screenshot of iSNAP containing the acidobactin variant hits. Acidobactin variant(analog) #54 and #48 have the same structures as acidobactin A and B. D) MS2 spectra of acidobactin A and B, including MS2 spectrum of a putative acidobactin C. The top acidobactin variant library iSNAP hit is shown for the putative acidobactin variant (acidobactin variant 56). Preliminary MS2 analysis suggest this is the correct structure for acidobactin C. Mass differences between the acidobactin variants are localized to the ornithine (red circle).



Supplementary Figure 6.17. Identification of variobactin variants with iSNAP using a variobactin variant library.

A) Variobactin variant library based on final variobactin A structure as indicated. B) *V. paradoxus* str. P4b extract chromatogram overlaid with frequency plot of variobactin variant library iSNAP hits per 0.25 min in LC retention. C) MS2 spectrum of variobactin A, including MS2 spectrum of putative variobactin variants B-E. Variobactin variant library iSNAP hits are shown for each putative compound's associated MS2 scan, variation is localized to fatty acid tail and chain length variation is shown in red.



Supplementary Figure 6.18. Theoretical database controls.

Selected extracts screened by iSNAP using the structure library indicated. LC-MS base peak chromatograms for indicated extracts are displayed in rows and are overlaid with a frequency plot of indicated library (columns) iSNAP hits per 0.25 min in LC retention.

Supplementary Table 6.1. Adenylation domain specificities for *A. citrulli* AAC00-1 biosynthetic cluster**.**

Amino acid adenylation codes were obtained from NRPSPredictor and matched with adenylation codes from the delftibactin biosynthetic locus (below)

Acidovorax citrulli

Aave_3733 A1	DMGGYGCLYK	3-HBA
	DMGGYGCLFK	Ala
Aave_3733 A2	DVWNIGLIHK	Orn
	DVWNIGLIHK	Orn
Aave_3731 A1	DVWHVSLIDK	Ser
	DVWHLSLIDK	Ser
Aave_3730 A1	DGEGSGGVTK	hOrn
	DGEGSGGVTK	hOrn
Aave_3730 A2	DFWNIGMVHK	Thr
	DFWNIGMVHK	Thr
Aave_3729 A1	DLTKVGHVGK	Asp
	DLTKVGHVGK	Asp

Variovorax paradoxus S110

Vapar_3746 A1	DMGGYGCLF-	3-HBA
	DMGGYGCLFK	Ala
Vapar_3746 A2	DVWNIGLIHK	Orn
	DVWNIGLIHK	Orn
Vapar_3744 A1	DVWHVSLIDK	Ser
	DVWHLSLIDK	Ser
Vapar_3743 A1	DGEGSGGVTK	hOrn
	DGEGSGGVTK	hOrn
Vapar_3743 A2	DFWNIGMVHK	Thr
	DFWNIGMVHK	Thr
Vapar_3742 A1	DLTKVGHVGK	Asp
	DLTKVGHVGK	Asp

Supplementary Table 6.2: Acidobactin A/B and vacidobactin A/B gene cluster analysis

Locus	Predicted Function	Strand	Amino Acids
Aave_3737	RNA polymerase, sigma-24 subunit, ECF subfamily	-	177
Aave_3736	MbtH domain protein	-	95
Aave_3735	Thioesterase	-	258
Aave_3734	TauD dioxygenase	-	330
Aave_3733	NRPS	-	1779
Aave_3732	PKS	-	1535
Aave_3731	NRPS	-	1130
Aave_3730	NRPS	-	2651
Aave_3729	NRPS	-	1368
Aave_3728	TonB siderophore receptor	+	733
Aave_3727	N-acetyltransferase	+	366
Aave_3726	N5-hydroxyornithine formyltransferase	+	313
Aave_3725	Ferric iron reductase	+	276
Aave_3724	Cyclic peptide transporter	-	564
Aave_3723	L-lysine 6-monooxygenase	-	452
Aave_3722	Phosphopantetheinyl transferase	-	257

Locus	Predicted Function	Strand	Amino Acids
Vapar_3752	RNA polymerase, sigma-24 subunit, ECF subfamily	-	181
Vapar_3751	anti-FecI sigma factor	-	67
Vapar_3750	MbtH domain protein	-	85
Vapar_3749	Thioesterase	-	246
Vapar_3748	Phosphopantetheinyl transferase	-	229
Vapar_3747	TauD dioxygenase	-	330
Vapar_3746	NRPS	-	1776
Vapar_3745	PKS	-	1520
Vapar_3744	NRPS	-	1110
Vapar_3743	NRPS	-	2626
Vapar_3742	NRPS	-	1358
Vapar_3741	TonB siderophore receptor	+	723
Vapar_3740	L-lysine 6-monooxygenase	+	439
Vapar_3739	N-acetyltransferase	+	344
Vapar_3738	N5-hydroxyornithine formyltransferase	+	281
Vapar_3737	Ferric iron reductase	+	281
Vapar_3736	Cyclic peptide transporter	-	563

6.7.1 Structural Characterization

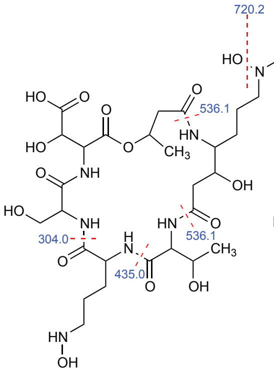
High Resolution Mass Spectra

A stock solution of 20 mg/ml of each compound was diluted to a final concentration of 10 µg/ml in water with 0.1% formic acid. This solution was directly infused at a rate of ~3 µL per min into a Thermo Finnigan LTQ OrbiTrap XL mass spectrometer running

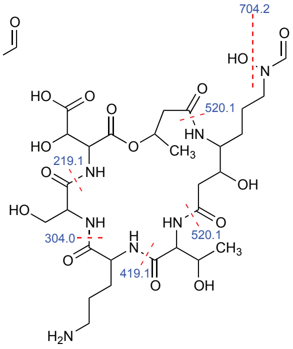
Xcaliber 2.07 and TunePlus 2.4 SP1. High resolution MS was acquired using an electrospray ionization source and fragmentation was obtained through collision induced dissociation (CID). The instrument was operated in the positive mode using a maximum resolution of 100, 000. Data was acquired for approximately 1 min for a total of 32 scans. Bradykinin was used as an internal standard, and was premixed with compounds to a final concentration of 5 µg/ml. The lock mass feature was applied using the bradykinin standard at [M+H] = 1060.56922 *m/z*.

Mass Spectra Fragmentation

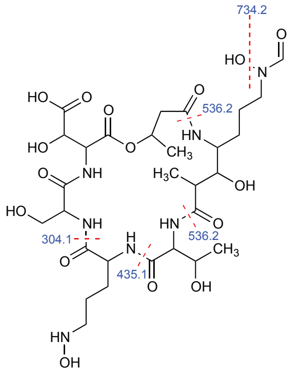
Mass spectral fragmentation patterns are shown for each compound below. Actual MS2 fragmentation can be seen in Supplementary Figures 9, 10 and 12.



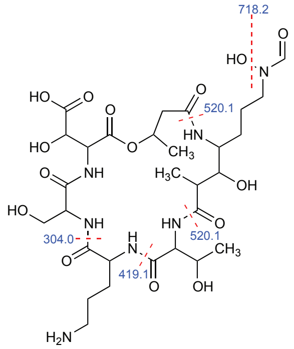
Acidobactin A (1)



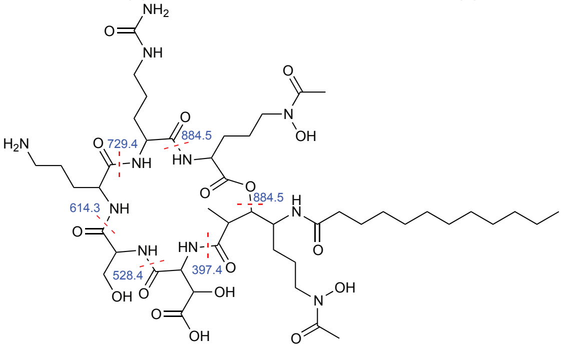
Acidobactin B (2)



Vacidobactin A (3)



Vacidobactin B (4)



Variobactin A (5)

NMR Methods and Structural Characterization

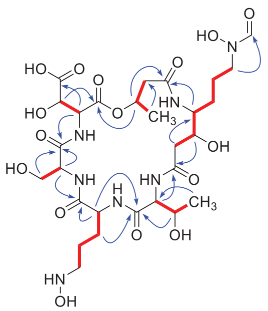
NMR spectra were measured on a Bruker Avance 700 spectrometer equipped with a 5 mm inverse detection probe and using TMS as an internal standard. Lyophilized samples were dissolved in D_2O and spectra were recorded at 297 K. NMR experiments were processed and analyzed with Bruker TOPSPIN 2.1. Chemical shifts (*δ*) expressed in parts per million (ppm) and coupling constants (*J*) are reported in Hertz (Hz). Assembly of individual amino acids to form the final linear structure was accomplished by considering long-range ^1H-^{13}C HMBC correlations from protons adjacent carbonyl carbons, as well as by assignments of 2D ^1H-^1H COSY and 2D ^1H-^{13}C HSQC correlations. Absolute configuration of acidobactin A and B were determined using Marfey's reaction and comparison to L and D amino acid standards

Supplementary Table 6.3. NMR spectroscopic Data for acidobactin A (1), B (2), vacidobactin A (3), and B (4) (700 MHz, in D_2O)

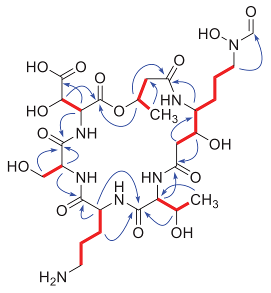
position	**δH mult.**				**δC**			
	1	**2**	**3**	**4**	**1**	**2 a**	**3 a**	**4 a**
1	1.25, d, 4.0	1.20, d, 4.2	1.19, d, 3.6	1.19, d, 3.6	18.6, CH_3	19.3, CH_3	19.0, CH_3	19.0, CH_3
2	5.34, m	5.48, m	5.38, m	5.37, m	68.9, CH	70.7, CH	70.6, CH	70.7, CH
3a	2.51, m	2.73, m	2.58, m	2.58, m	34.9, CH_2	39.7, CH_2	38.4, CH_2	38.3, CH_2
3b	2.80, m	2.81, m	2.77, m	2.71, m	-	-	-	-
4	-	-	-	-	164.2, C	171.8, C	164.6, C	171.4, C
5	3.36, m	3.33, m	3.31, m	3.36, m	54.4, CH	54.9, CH	52.8, CH	52.9, CH
6a	1.61, m	1.50, m	1.63, m	1.63, m	23.1, CH_2	23.7, CH_2	22.4, CH_2	22.6, CH_2
6b	1.59, m	1.68, m	-	-	-	-	-	-
7a	1.71, m	1.63, m	1.83, m	1.87, m	25.4, CH_2	26.4, CH_2	28.6, CH_2	28.4, CH_2
7b	1.80, m	1.85, m	-	-	-	-	-	-
8a	3.55, m	3.49, m	3.38, m	3.33, m	49.9, CH_2	49.8, CH_2	52.1, CH_2	52.7, CH_2
8b	-	3.90, m	-	3.48, m	-	-	-	-
9	7.90, s	7.88, s	7.74, s	7.74, s	153.6, CH	160.0, CH	155.4, CH	155.5, CH
10	4.22, m	4.35, m	4.30, m	4.17, m	68.0, CH	71.9, CH	72.3, CH	66.9, CH
11a	2.79, m	2.59, m	2.71, m	2.69, m	39.0,	38.9,	43.2, CH	43.1, CH

					CH$_2$	CH$_2$		
11b	-	2.69, m	-	-	-	-	-	-
11-CH3	-	-	1.03, d	1.02, d	-	-	12.8, CH$_3$	12.7, CH$_3$
12	-	-	-	-	173.4, C	172.3, C	177.0, C	177.1, C
13	4.23, m	4.40, d, 1.8	4.41, m	4.44, m	58.7, CH	60.8, CH	56.2, CH	55.6, CH
14	4.22, m	4.31, m	4.26, m	4.19, m	66.3, CH	66.5, CH	66.8, CH	66.4, CH
15	1.10, d, 3.8	1.07, d, 4.0	1.10, m	1.08, m	18.8, CH3	19.1, CH3	19.2, CH3	19.1, CH3
16	-	-	-	-	174.1, C	174.1, C	172.3, C	173.9, C
17	4.36, m	4.60, t, 3.6	4.40, m	4.41, m	53.6, CH	54.1, CH	55.3, CH	55.1, CH
18a	1.62, m	1.71, m	1.60, m	1.62, m	27.7, CH$_2$	27.9, CH$_2$	28.4, CH$_2$	28.3, CH$_2$
18b	1.84, m	2.61, m	-	-	-	-	-	-
19a	1.71, m	1.81, m	1.85, m	1.81, m	23.7, CH$_2$	22.1, CH$_2$	22.6, CH$_2$	22.4, CH$_2$
19b	1.87, m	2.12, m			-	-	-	-
20a	3.55, m	3.32, m	3.41	3.37	50.3, CH$_2$	49.3, CH$_2$	51.4, CH$_2$	49.8, CH$_2$
20b	-	3.96, m	-	3.65	-	-	-	-
21	-	-	-	-	172.9, C	174.2, C	172.7, C	174.9, C
22	4.41, m	4.44, m	4.30, m	4.27, m	56.3, CH	55.7, CH	54.7, CH	54.4, CH
23a	3.83, d, 8.5	3.84, d, 9.0	3.90, m	3.85, m	61.3, CH	61.1, CH	61.3, CH	61.5, CH
23b	-	3.95, d, 9.0	-	-	-	-	-	-
24	-	-	-	-	171.6, C	172.5, C	169.3, C	169.4, C
25	4.82, d	4.50, m	4.80, m	4.59, m	56.8, CH	56.2, CH	56.4, CH	51.9, CH
26	4.31, d	4.25, m	3.90, m	3.87, m	74.7, CH	72.4, CH	73.9, CH	73.6, CH
27	-	-	-	-	180.0, C	175.6, C	179.6, C	181.4, C
28	-	-	-	-	169.6, C	169.1, C	177.9, C	177.7, C

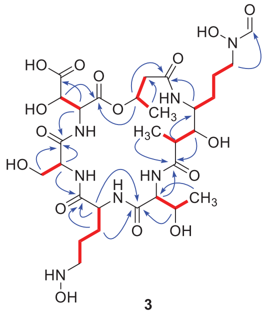
[a] ^{13}C NMR were extracted from HMBC spectra.



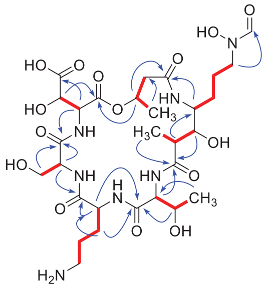
1



2



3

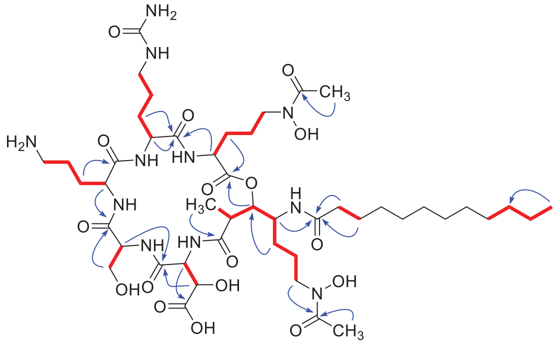


4

Supplementary Table 6.4. NMR spectroscopic Data for variobactin A (5) (700 MHz, in DMSO-d_6)[a].

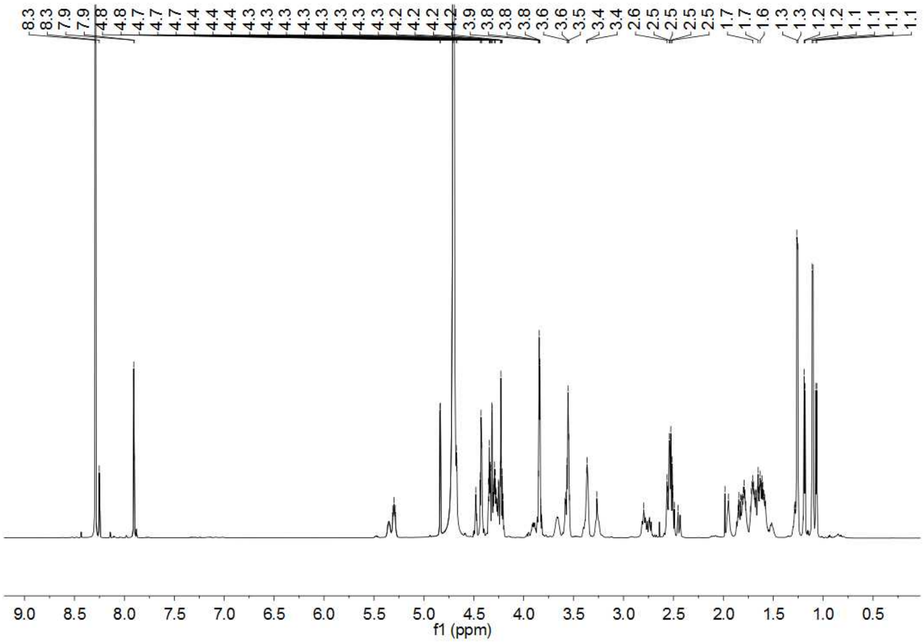
Position	δH mult.	δC	Position	δH mult.	δC
1	1.26 (ov)	16.8, CH_3	**22**	-	157.3, C
2	2.14 (m)	44.3, CH	**23**	-	172.1, C
3	3.40 (m)	74.5, CH	**24**	3.59 (ov)	50.4, CH
4	4.36 (m)	49.9, CH	**25**	1.21 (ov)	29.3, CH_2
5	1.56 (m)	29.0, CH_2	**26**	1.45 (m)	23.7, CH_2
6	1.87 (m)	23.7, CH_2	**27a**	3.32 (m)	49.2, CH_2
7a	3.70 (m)	47.2, CH_2	**27b**	3.40 (ov)	171.1, C
7b	3.90 (m)		**28**	-	
8	-	159.8, C	**29**	4.52 (m)	53.8, CH
9	2.03 (s)	16.6, CH_3	**30a**	3.18 (m)	62.9, CH_2
10	-	175.3, C	**30b**	3.70 (d, 5.7)	
11	4.19 (d, 3.6)	61.3, CH	**31**	-	170.3, C
12	1.24 (ov)	31.8, CH_2	**32**	4.19 (d, 9.0)	58.8, CH
13	1.82 (m)	23.6, CH_2	**33**	4.29 (d, 1.8)	74.1, CH
14a	3.66 (m)	49.4, CH_2	**34**	-	179.6, C
14b	3.84 (m)		**35**	-	174.4, C
15	-	161.7, C	**36**	-	172.2, C
16	1.99 (s)	16.9, CH_3	**37**	2.04 (m)	36.0, CH_2
17	-	170.6, C	**38**	1.45 (m)	25.9, CH_2
18	4.64 (m)	50.5, CH	**39-44**	1.23 (ov)	28.5-29.5, CH_2
19a			**45**	1.23 (ov)	31.7, CH_2
19b			**46**	1.23 (ov)	22.5, CH_2
20a	1.22 (m)	25.5, CH_2	**47**	0.89 (t, 3.6)	14.4, CH_3
20b	1.46 (m)				
21a	2.94 (m)	40.9, CH_2			
21b	3.15 (t, 4.2)				

[a] Chemical shift *δ* and (multiplicity, *J* in Hz).

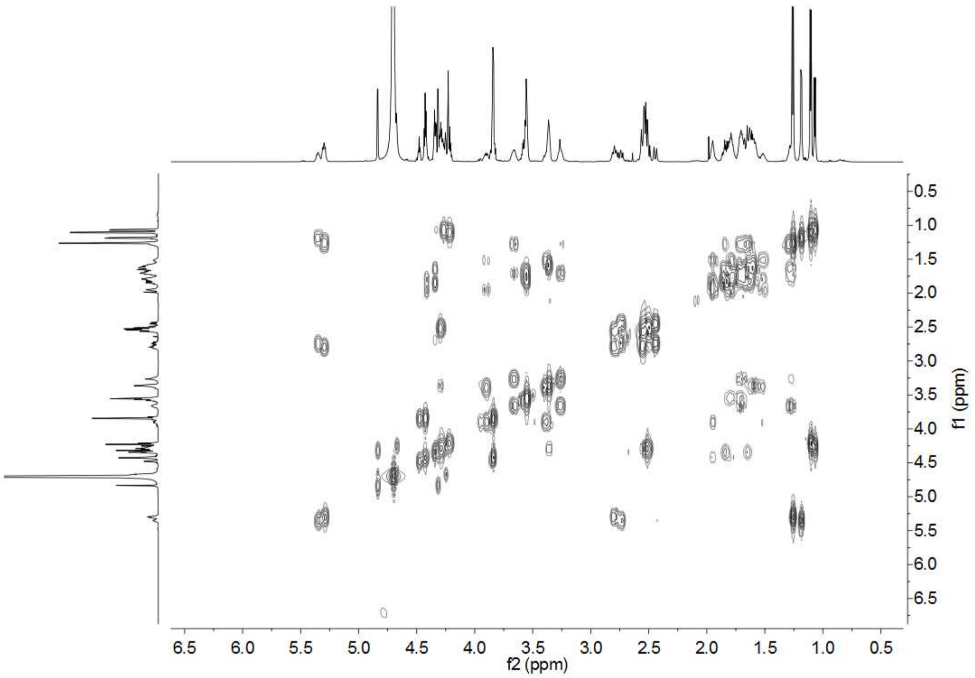


5

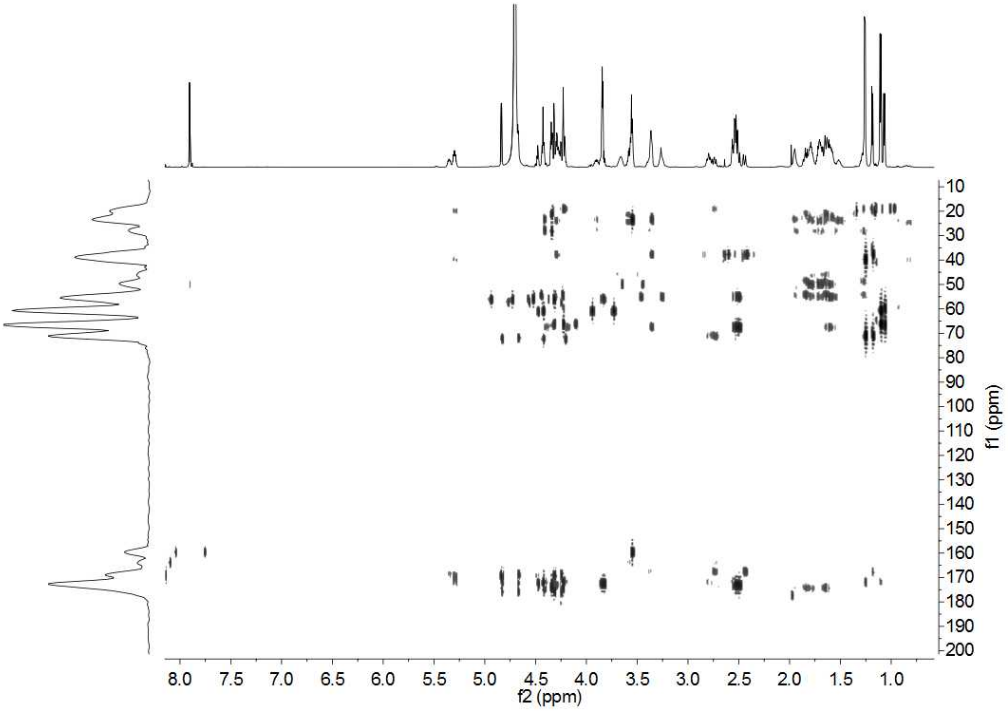
NMR Spectra



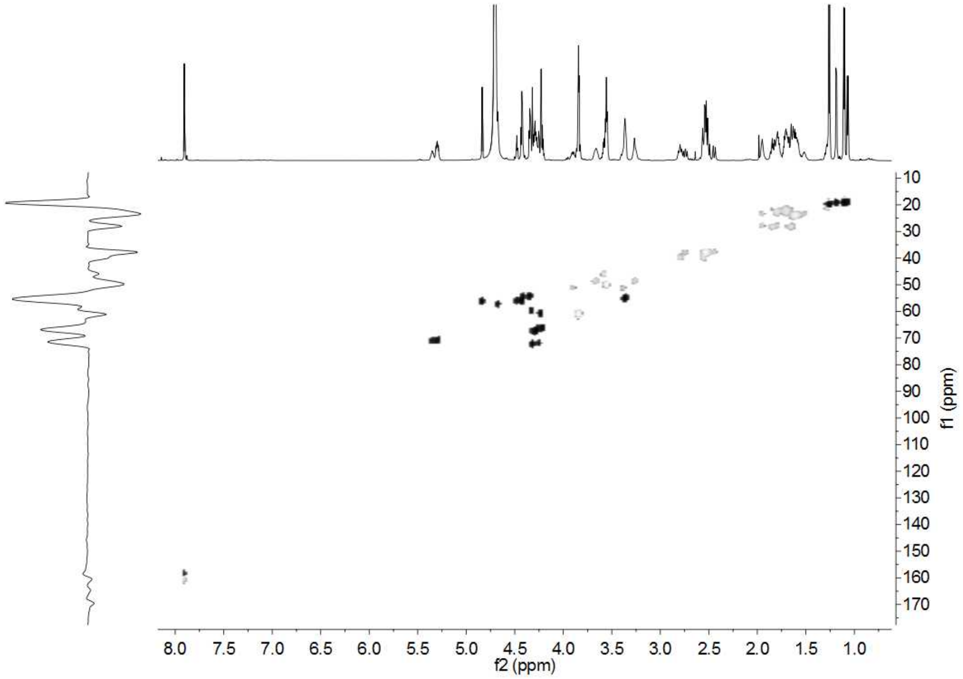
Supplementary Figure 6.19. ^{1}H NMR spectrum of acidobactin A (**1**) in D_2O.



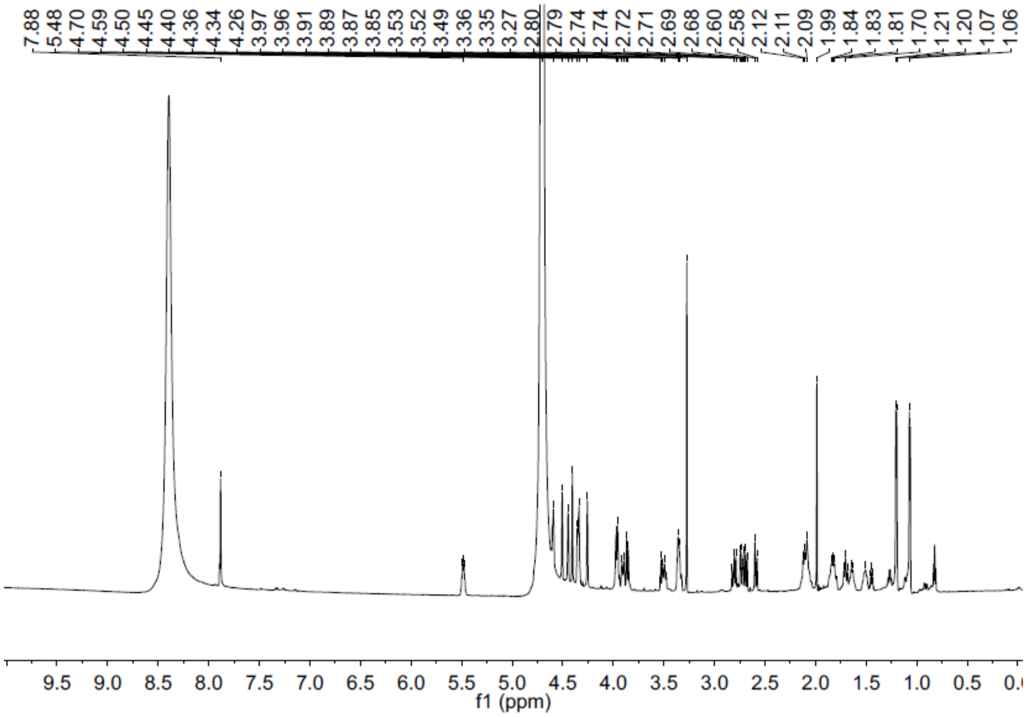
Supplementary Figure 6.20. ^{1}H-^{1}H COSY spectrum of acidobactin A (**1**) in D_2O.



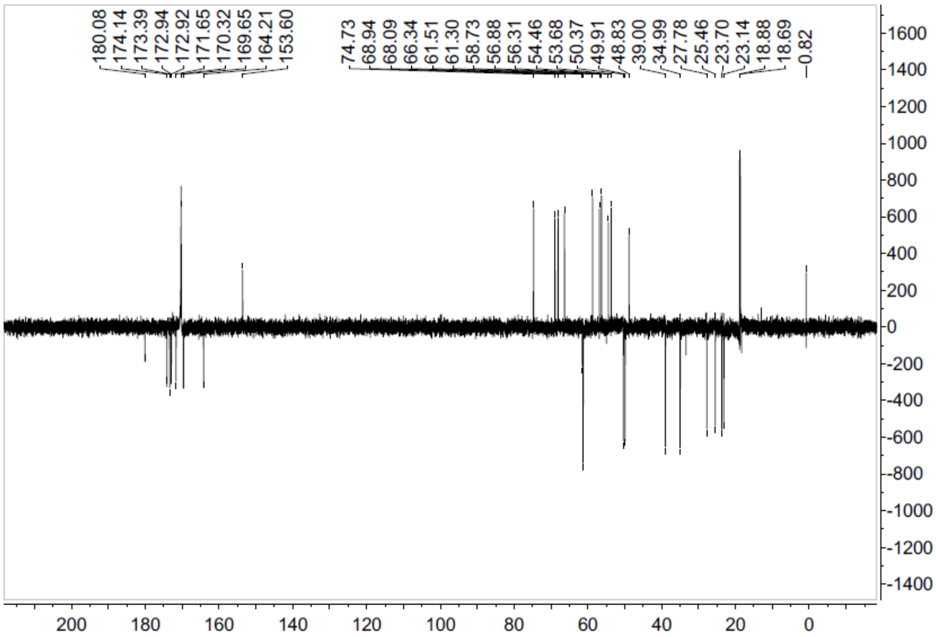
Supplementary Figure 6.21. ^{1}H-^{13}C HMBC spectrum of acidobactin A (**1**) in D_2O.



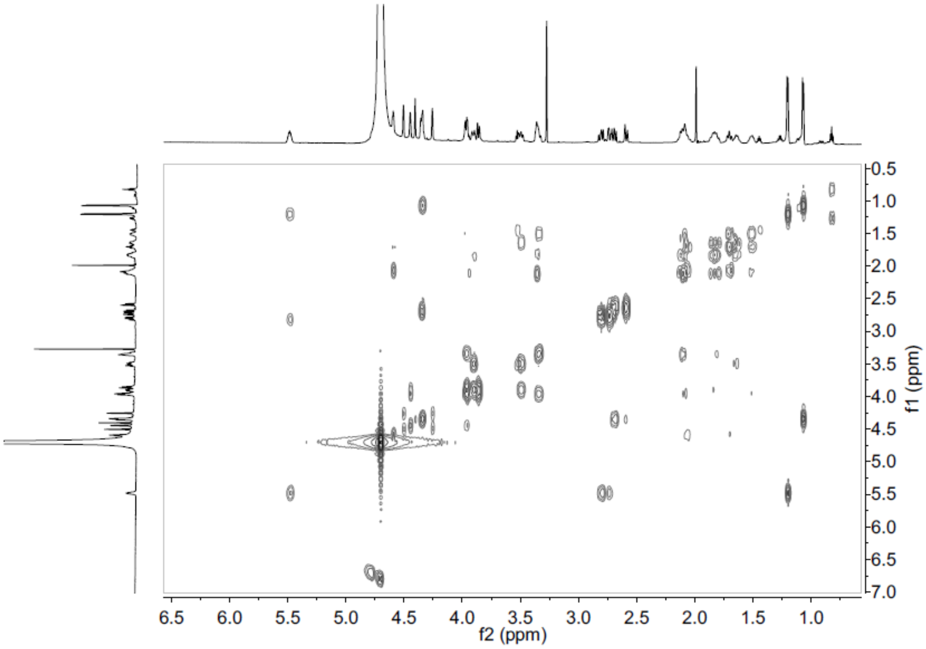
Supplementary Figure 6.22. ^{1}H-^{13}C HMQC spectrum of acidobactin A (**1**) in D_2O.



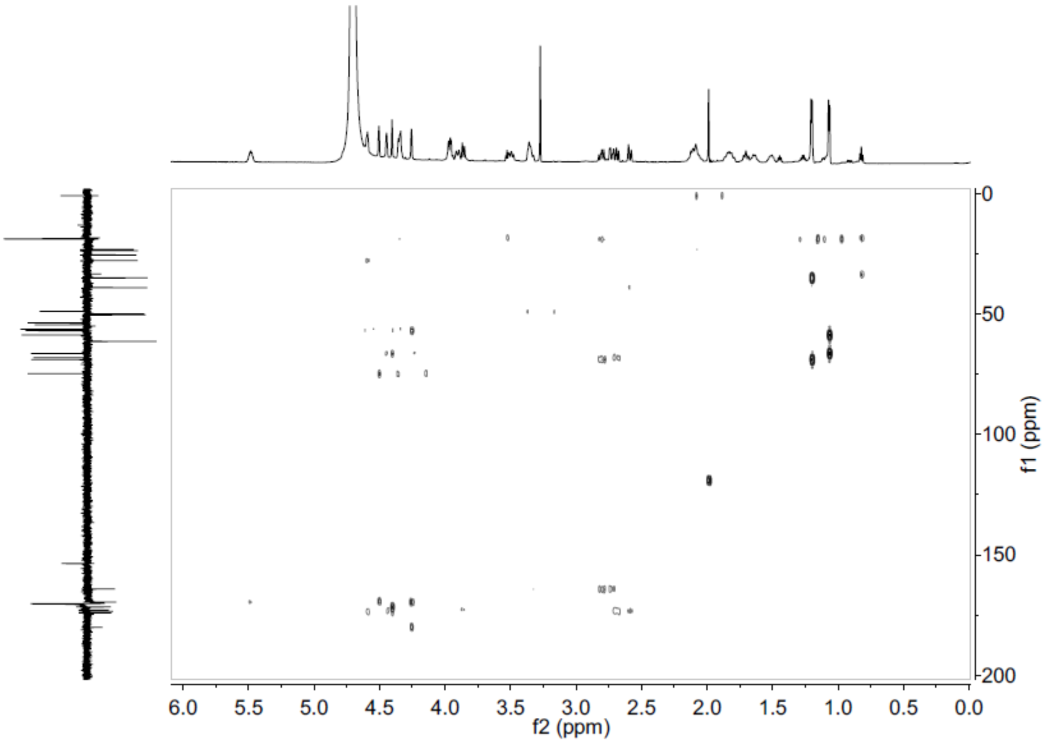
Supplementary Figure 6.23. ^{1}H NMR spectrum of acidobactin B (**2**) in D_2O.



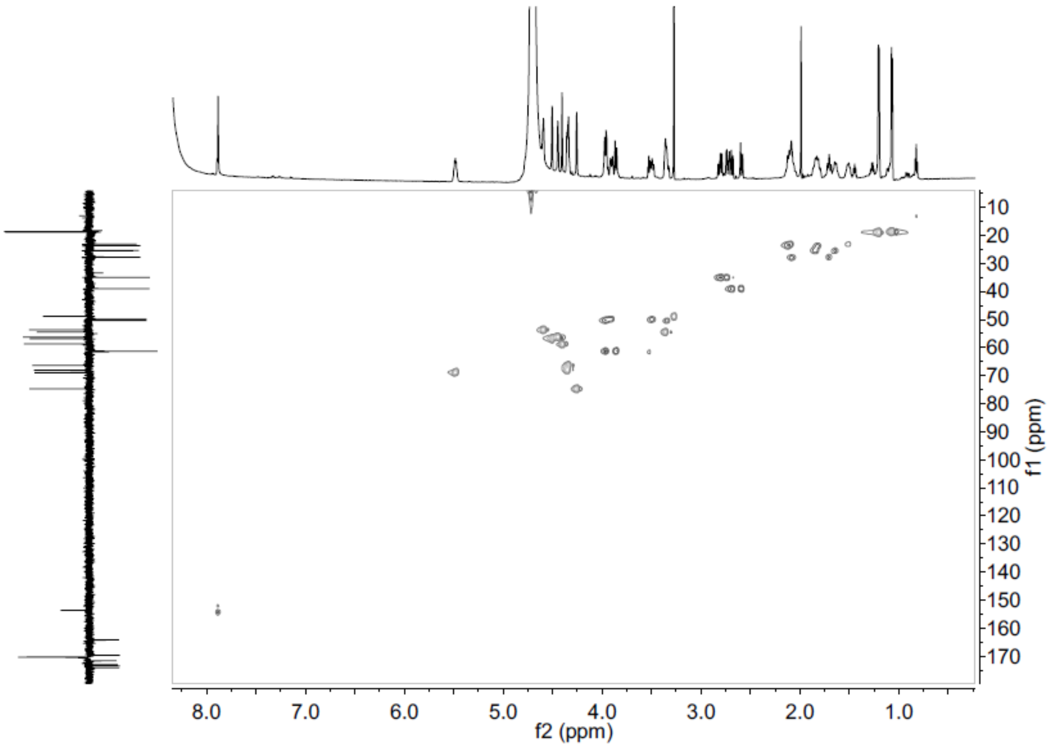
Supplementary Figure 6.24. DEPTq spectrum of acidobactin B (**2**) in D_2O.



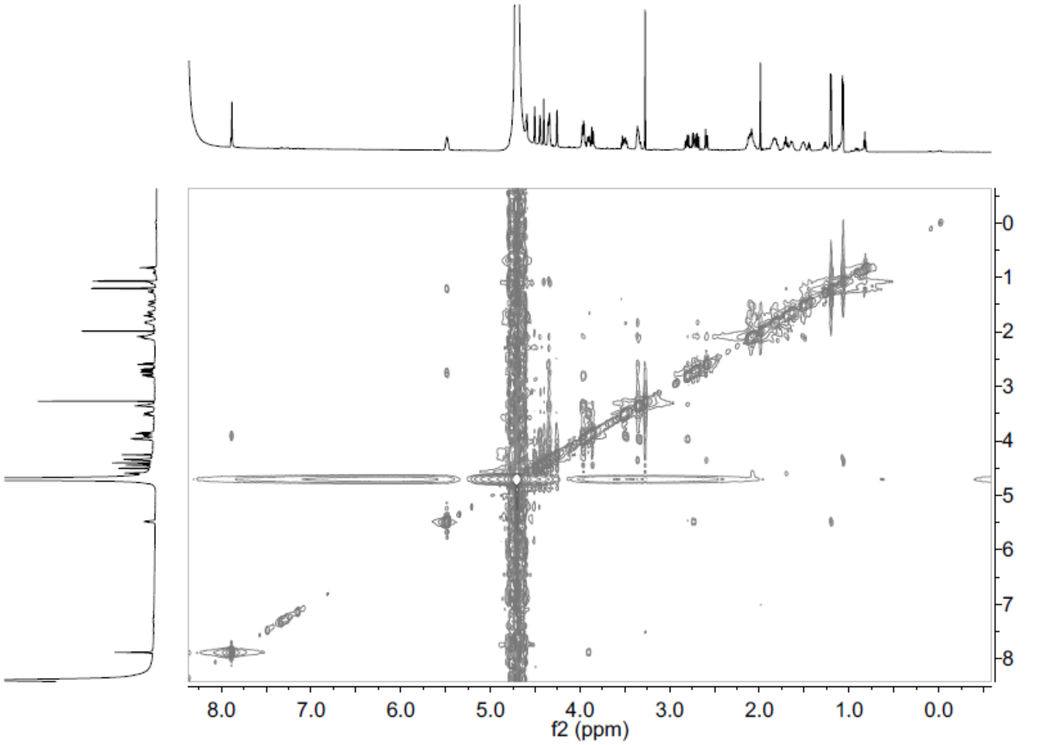
Supplementary Figure 6.25. ^1H-^1H COSY spectrum of acidobactin B (**2**) in D_2O.



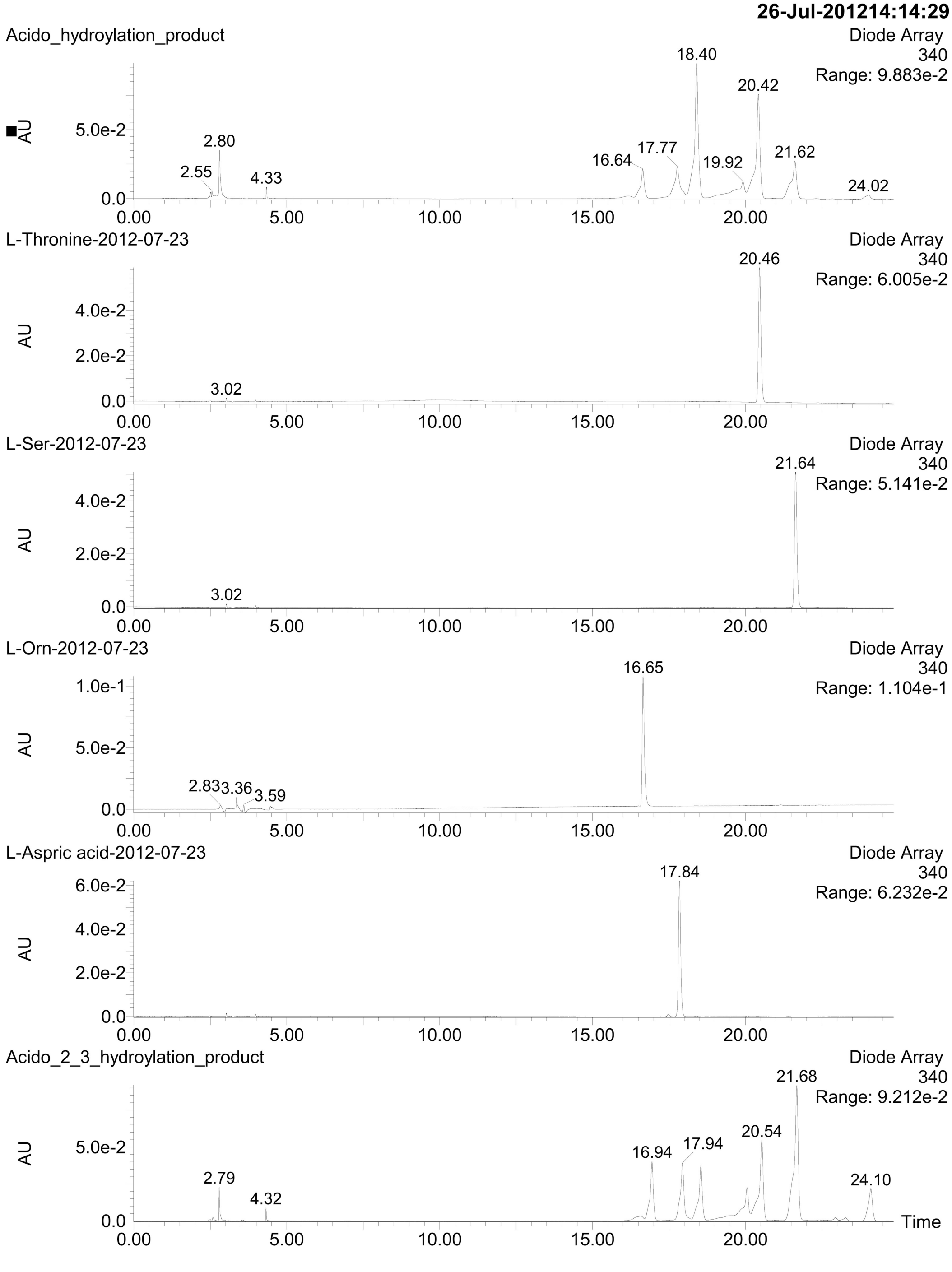
Supplementary Figure 6.26. ^{1}H-^{13}C HMBC spectrum of acidobactin B (**2**) in D_2O.



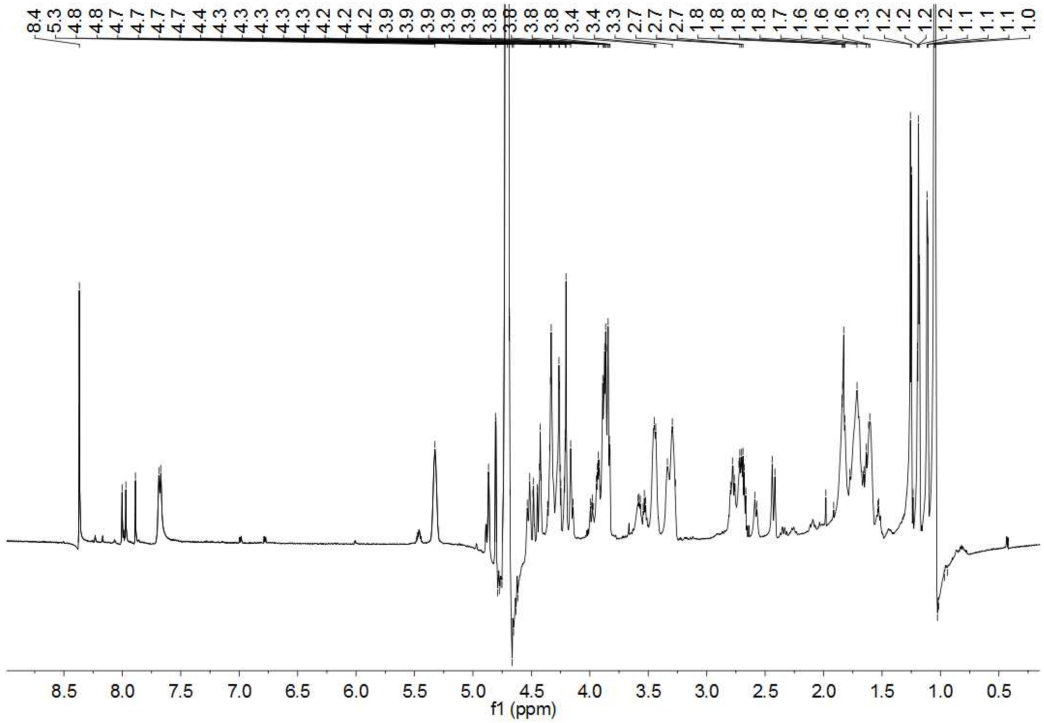
Supplementary Figure 6.27. ^{1}H-^{13}C HMQC spectrum of acidobactin B (**2**) in D$_2$O.



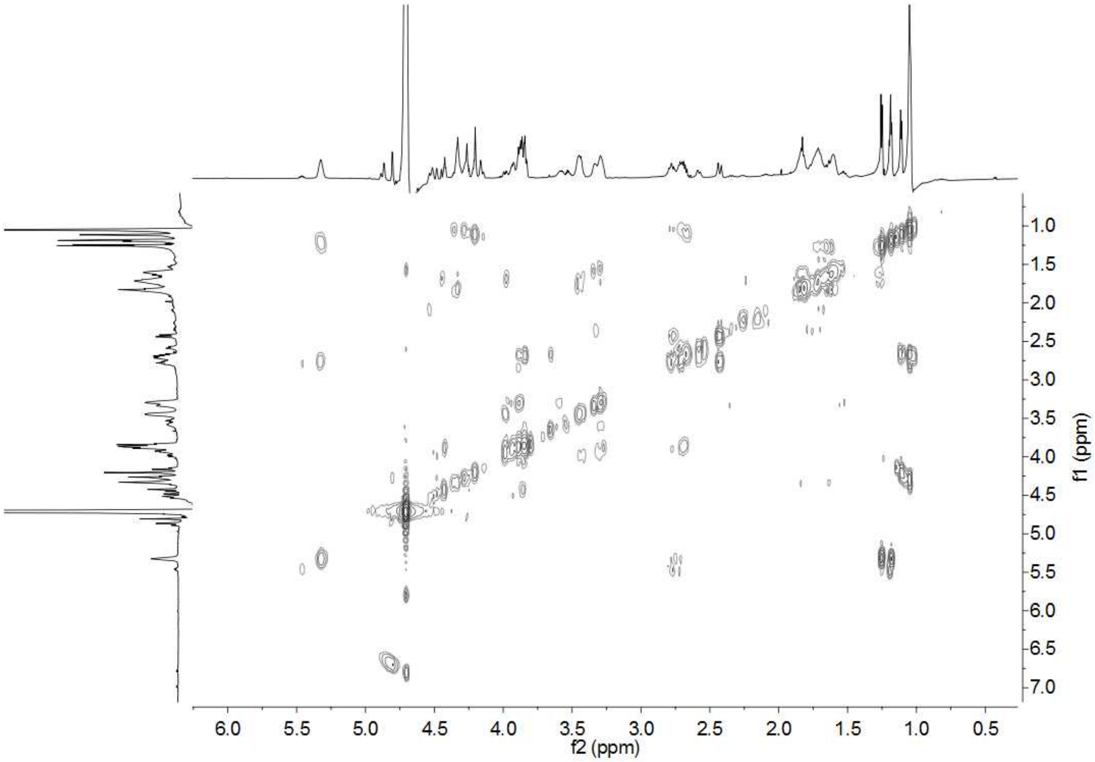
Supplementary Figure 6.28. 1H-1H NOESY spectrum of acidobactin B (2) in D2O.



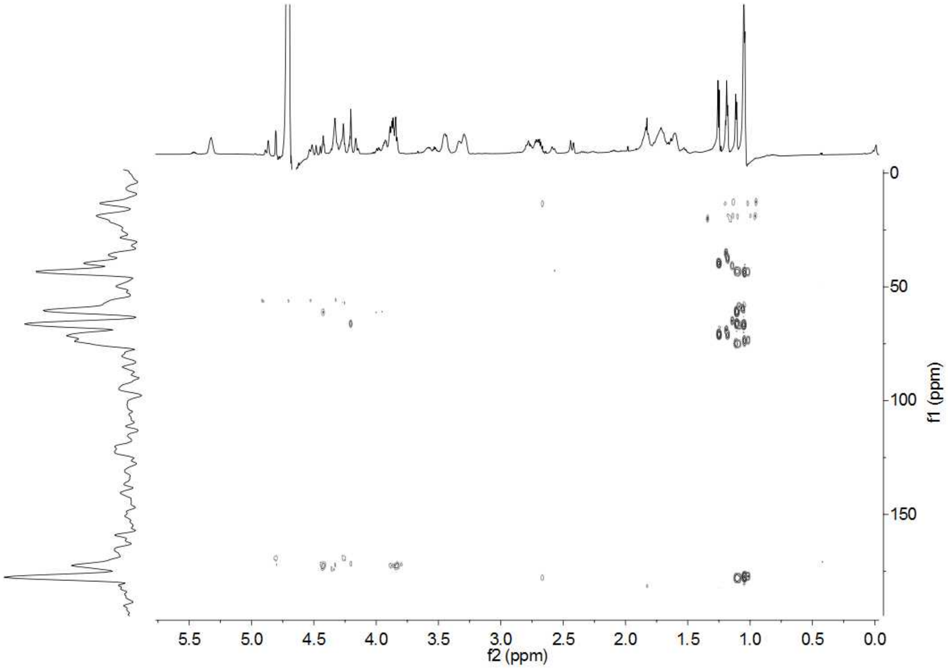
Supplementary Figure 6.29. Determination of absolute configuration of amino acid of acidobactin A and B (**1-2**) by Marfery reactions and comparison to L amino acid standardss



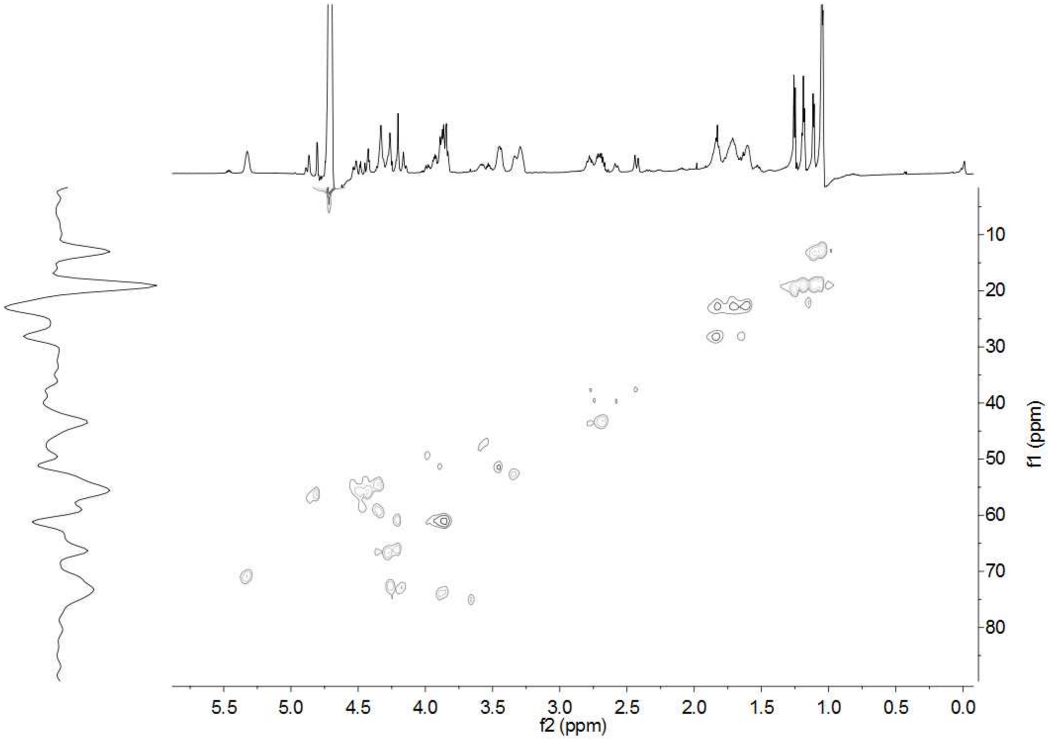
Supplementary Figure 6.30. ^{1}H NMR spectrum of variobactin A (**3**) in D_2O.



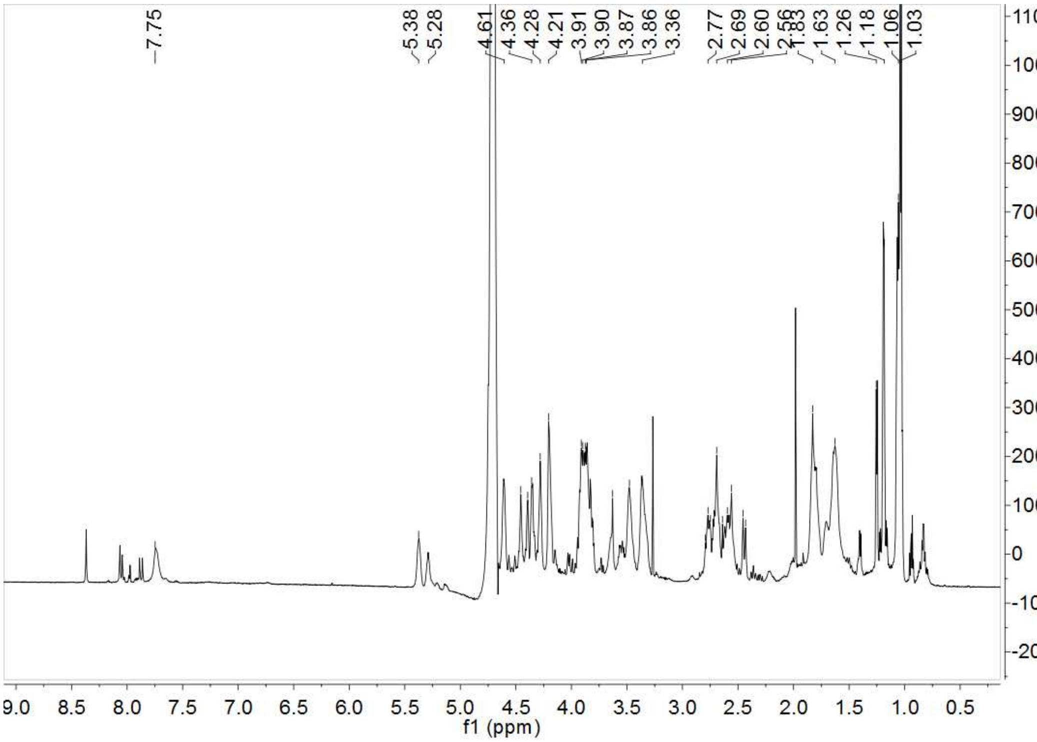
Supplementary Figure 6.31. ^{1}H-^{1}H COSY spectrum of variobactin A (**3**) in D_2O.



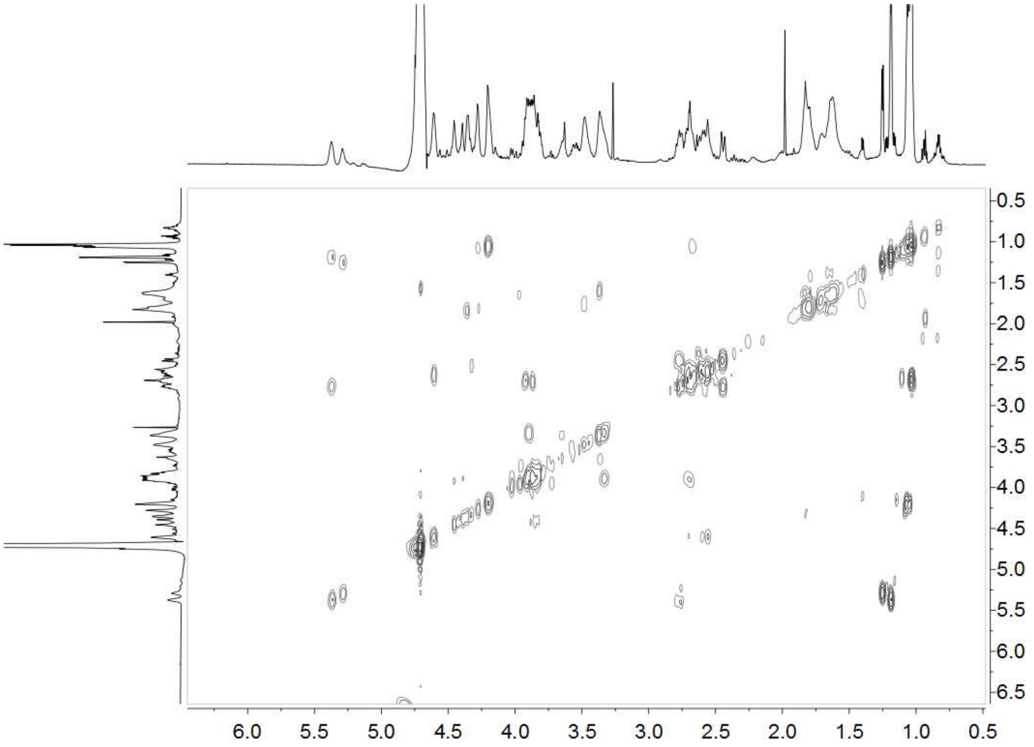
Supplementary Figure 6.32. ^{1}H-^{13}C HMBC spectrum of variobactin A (**3**) in D_2O.



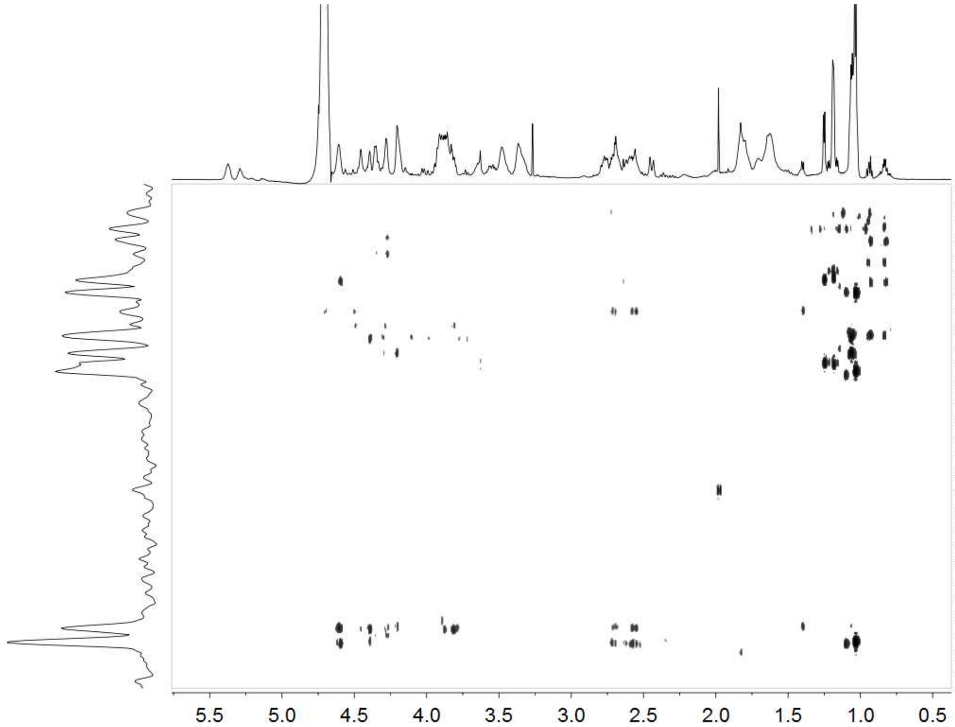
Supplementary Figure 6.33. ^{1}H-^{13}C HMQC spectrum of variobactin A (**3**) in D_2O.



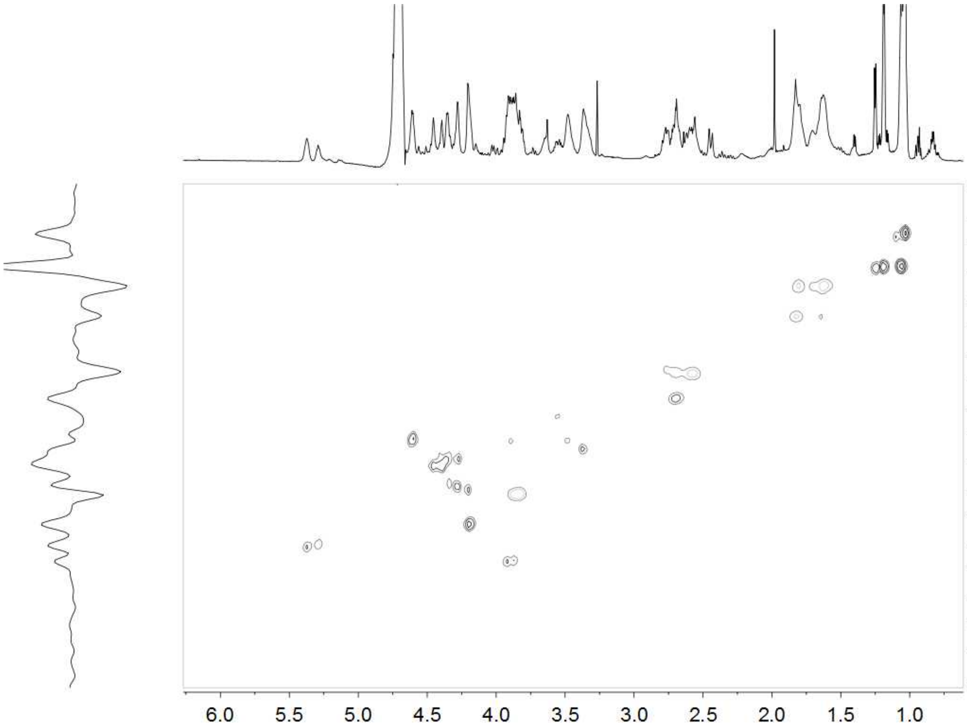
Supplementary Figure 6.34. ^{1}H NMR spectrum of variobactin B (**4**) in D_2O.



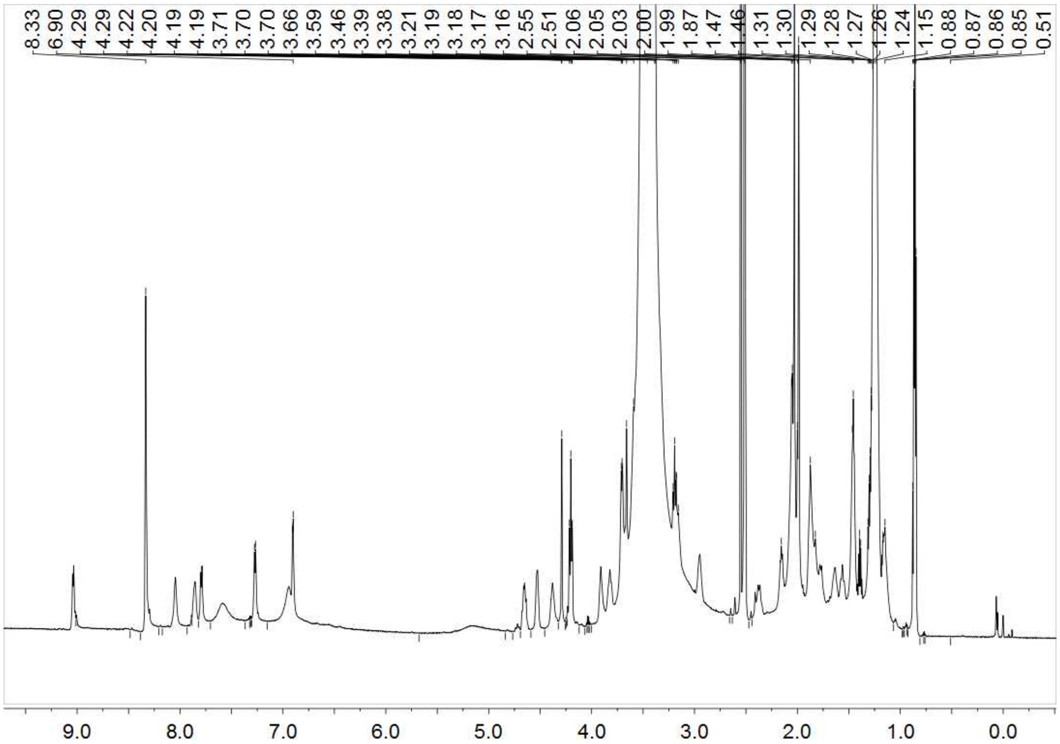
Supplementary Figure 6.35. ^{1}H-^{1}H COSY spectrum of variobactin B (**4**) in D_2O.



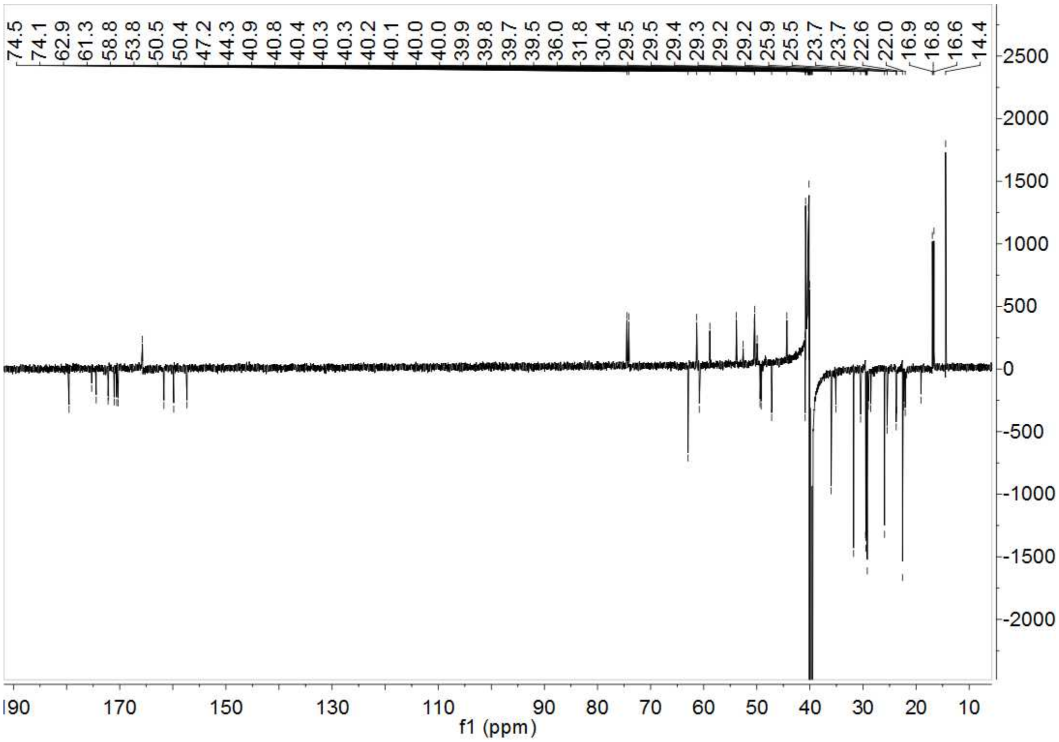
Supplementary Figure 6.36. ^{1}H-^{13}C HMBC spectrum of variobactin B (**4**) in D$_2$O.



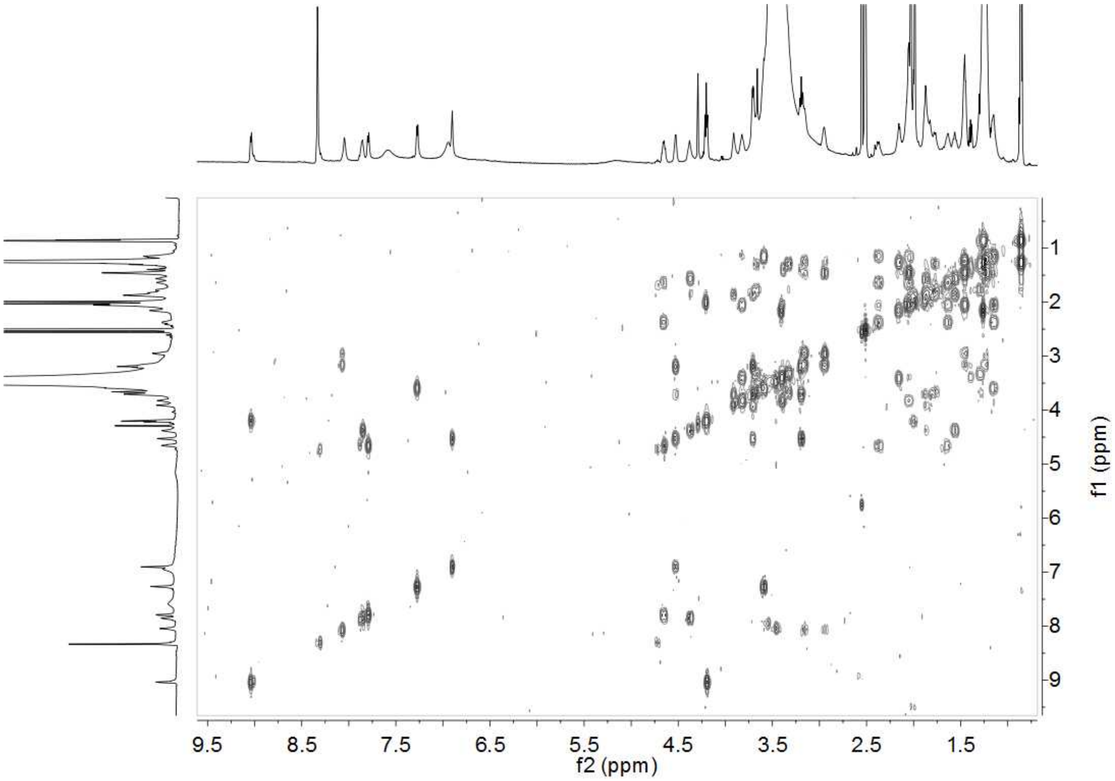
Supplementary Figure 6.37. ^{1}H-^{13}C HMQC spectrum of variobactin B (**4**) in D$_2$O.



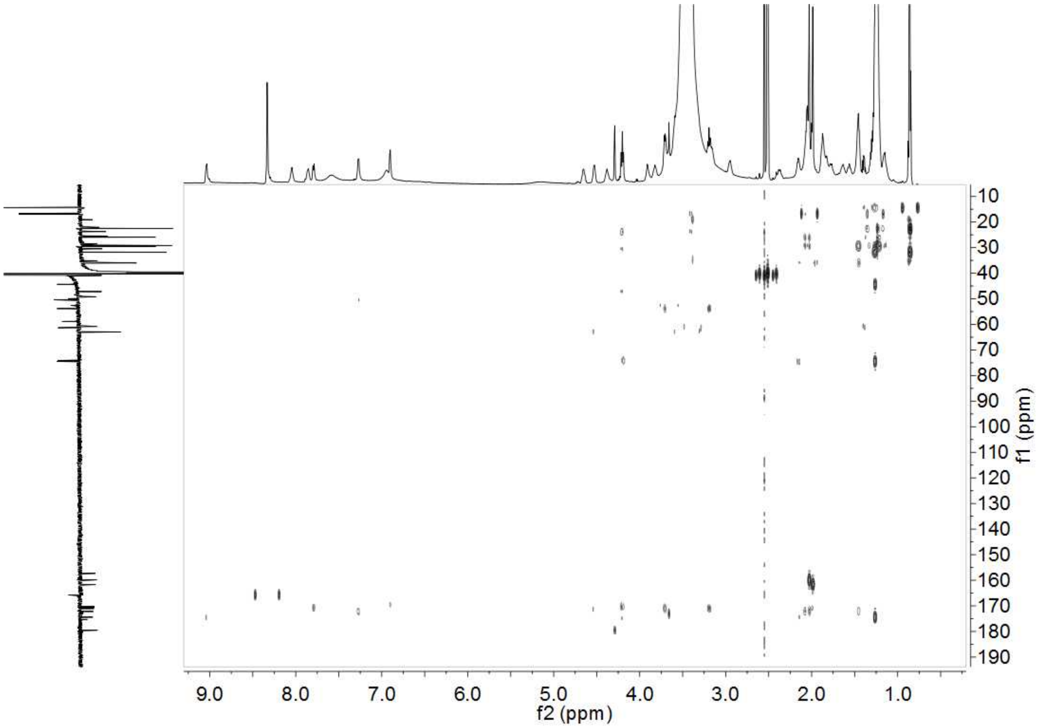
Supplementary Figure 6.38. ^{1}H NMR spectrum of variobactin A (**5**) in D_2O.



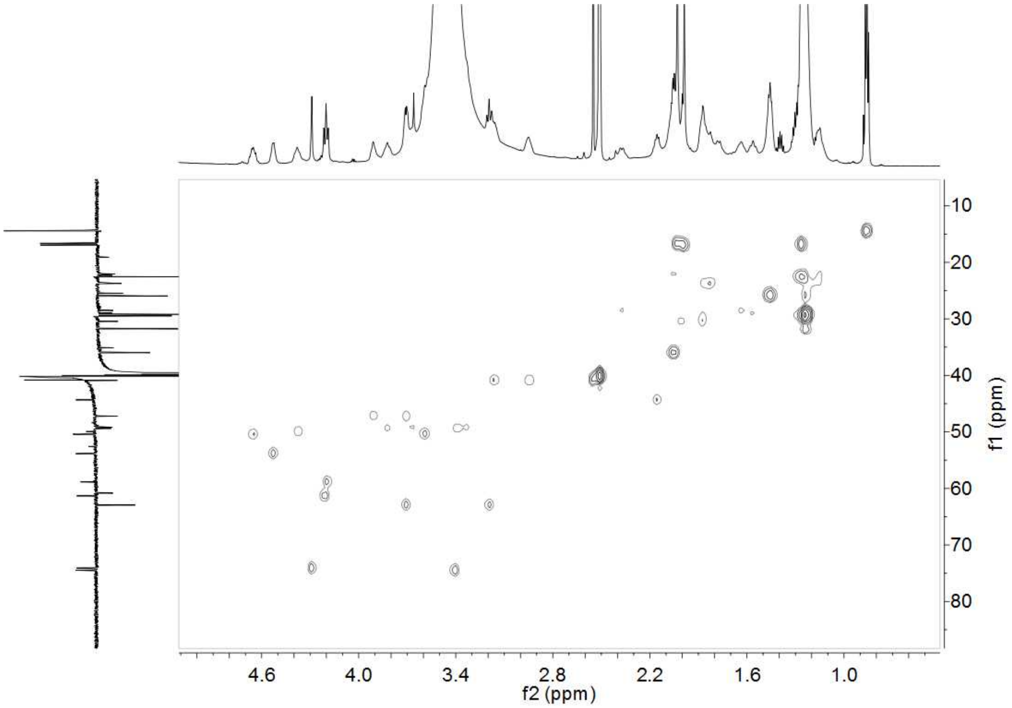
Supplementary Figure 6.39. DEPTq spectrum of variobactin A (**5**) in D_2O.



Supplementary Figure 6.40. ^{1}H-^{1}H COSY spectrum of variobactin A (**5**) in D$_2$O.



Supplementary Figure 6.41. ^{1}H-^{13}C HMBC spectrum of variobactin A (**5**) in D$_2$O.



Supplementary Figure 6.42. ^{1}H-^{13}C HMQC spectrum of variobactin A (**5**) in D_2O.

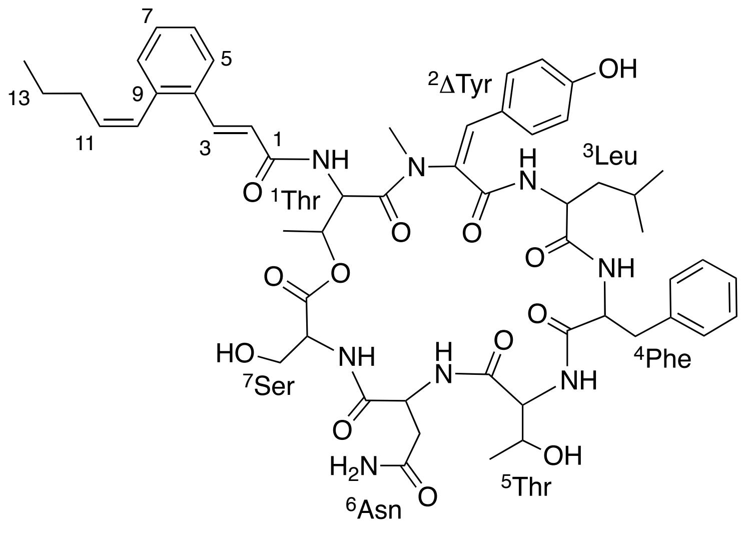
Supplementary Table 6.5: Summary of ^{1}H (600 MHz) and ^{13}C (125 MHz) spectroscopic data for WS-9326A in DMSO-d_6[a]

position		dH (J in Hz)	dC
Acyl	1 (C=O)		165.2, s
	2	6.68, 1H, d (15.6)	122.7, d
	3	7.42, 1H, d (15.6)	127.3, d
	4		133.1, s
	5		126.0, d
	6	n/d	n/d
	7	n/d	n/d
	8		129.6, d
	9		137.0, s
	10	6.50, 1H, d (11.2)	126.8, d
	11	5.83, 1H, dt (11.7, 7.4)	134.0, d
	12	1.99, 2H, m	29.9, t

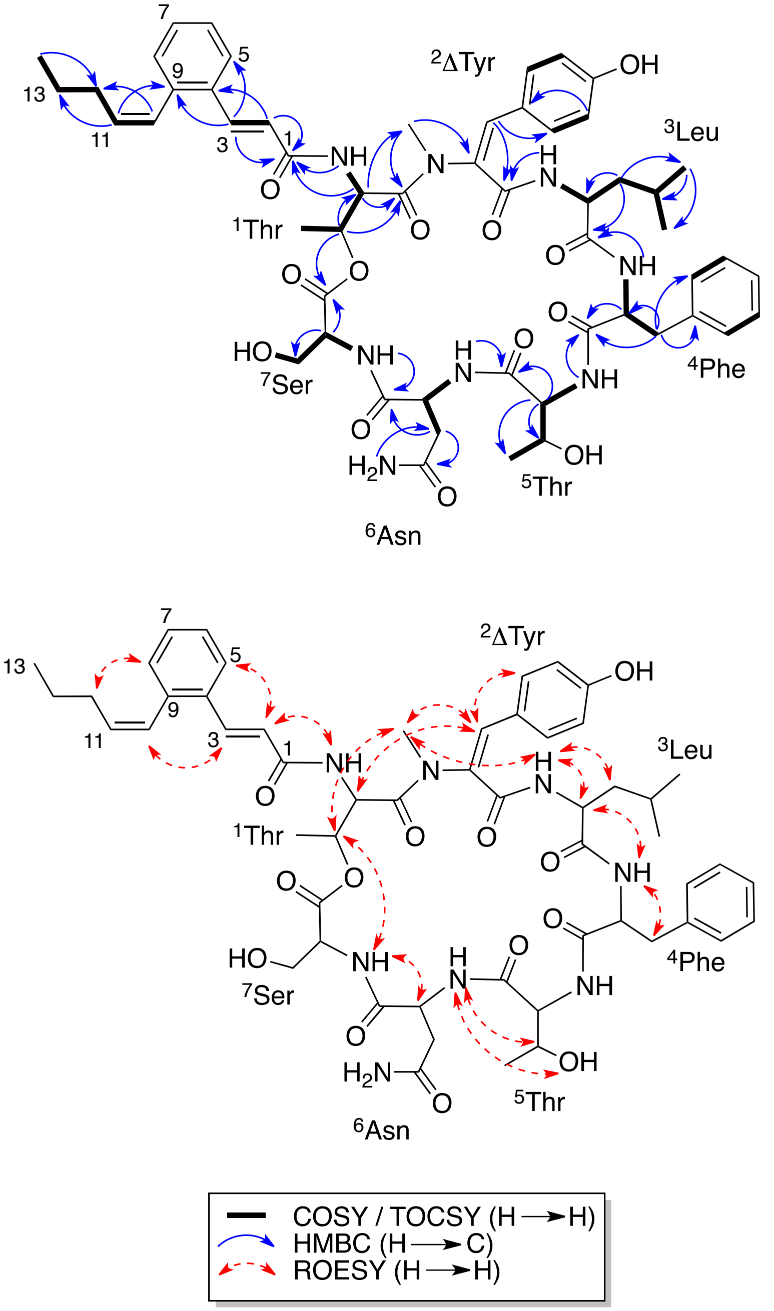
	13	1.36, 2H, m (7.8)	21.9, t
	14	0.79, 3H, t (7.4)	13.53, q
1Thr	**NH**	8.70, 1H, d (9.3)	
	a	5.33, 1H, t (9.6)	53.2, d
	b	5.03, 1H, dq (9.8, 6.2)	73.24, d
	g	1.15, 3H, d, 6.0	16.56, q
	C=O		169.0, s
2DTyr	**NMe**	2.98, 3H, s	34.2, q
	a		128.5, s
	b	6.13, 1H, s	131.6, d
	1		122.9, s
	2,6	7.39, 2H, d (8.6)	131.5, d
	3,5	6.59, 2H, d (8.6)	114.8, d
	4		158.1, s
	C=O		165.6, s
3Leu	**NH**	9.23, 1H, br. s	
	a	4.07, 1H, m	53.7, d
	b	1.26, 2H, m	38.8, t
	g	0.89, 1H, m	23.4, d
	d	0.63, 3H, o. d	22.8, q
		0.75, 3H, d (6.3)	22.0, q
	C=O		172.1, s
4Phe	**NH**	9.16, 1H, d (7.7)	
	a	4.33 m	55.7, d
	b	3.28, 1H, m	36.3, t
		2.73, 1H, o. t (12.9)	
	1		138.7, s
	2,6	7.32, 2H, o. d (8.4)	128.9, d
	3,5	7.27, 2H, t, (7.6)	127.9, d
	4	n/d	n/d
	C=O		170.1, s
5Thr	**NH**	7.59, 1H, d (9.5)	
	a	4.36, 1H, o. t (9.8)	57.2, d
	b	4.26, 1H, m	68.1, d
	g	0.64, 3H, o. d (6.4)	22.0, q
	OH	5.18, 1H, d (2.9)	
	C=O		169.9, s
6Asn	**NH**	8.33, 1H, d (7.3)	
	a	4.44, 1H, m	50.8, d
	b	2.46, 1H, m	36.7, t

		2.41, 1H, dd (15.3, 9.9)	
	g C=O		171.2, s
	g NH2	6.93, 1H, s	
		7.30, 1H, o	
	C=O		171.6, s
7Ser	**NH**	8.48, 1H, d (9.6)	
	a	4.33, 1H, o	56.0, d
	b	3.26, 1H, o	60.8, d
		3.16, 1H, br. t (~9)	
	OH	4.78, 1H, br	
	C=O		168.8, s

a Assignments based on HSQC, COSY, TOCSY, HMBC, and ROESY experiments. o = overlapping signal, br = broad signal, n/d = no data



WS9326a



Supplementary Figure 6.43. Summary of key 2D-NMR spectroscopic data that supports the structure of WS-9326A.

6.8 References

1. Newman, D.J. & Cragg, G.M. Natural Products As Sources of New Drugs over the 30 Years from 1981 to 2010. J. Nat. Prod. 75, 311-335 (2012).

2. Carter, G.T. Natural products and Pharma 2011: strategic changes spur new opportunities. Nat. Prod. Rep. 28, 1783-1789 (2011).

3. Koehn, F.E. & Carter, G.T. Rediscovering natural products as a source of new drugs. Discov. Med. 5, 159-164 (2005).

4. Koehn, F.E. & Carter, G.T. The evolving role of natural products in drug discovery. Nat. Rev. Drug Discov. 4, 206-220 (2005).

5. Clardy, J. & Walsh, C. Lessons from natural molecules. Nature 432, 829-837 (2004).

6. Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. Nature 432, 855-861 (2004).

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

8. Tobert, J. A. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. Nat. Rev. Drug Discov. 2, 517-526 (2003).

9. Cragg, G. M., & Newmann, D. J. Nature: a vital source of leads for anticancer drug development. Phytochem. Rev. 8, 313-331 (2009).

10. Gregory, M. A. et al. Mutasynthesis of rapamycin analogues through the manipulation of a gene governing started unit biosynthesis. Angew. Chem. Int. Ed .117, 4835-4838 (2005).

11. Gu, L. et al. Metamorphic enzyme assembly in polyketide diversification. Nature 459, 731-735 (2009).

12. Sherman, D. H. The Lego-ization of polyketide biosynthesis. Nat. Biotechnol.. 23, 1083-1084 (2005)

13. Starks, C. M., Zhou, Y., Liu, F., & Licari, P. J. Isolation and characterization of new epothilone analogues from recombinant Myxococcus xanthus fermentations. J. Nat. Prod. 66, 1313-1317 (2003).

14. Magarvey, N. A. et al. Biosynthetic characterization and chemoenzymatic assembly of the cryptophycins: potent anticancer agents from Nostoc cyanobionts. ACS Chem. Bio. 1, 766–779 (2006)

15. Chai, Y. et al. Discovery of 23 Natural Tubulysins from Angiococcus disciformisAn d48 and Cystobacter SBCb004. Chem. Biol. 17, 296–309 (2010)

16. Herrmann, J. et al. Pretubulysin: From Hypothetical Biosynthetic Intermediate to Potential Lead in Tumor Therapy. PLoS One 7, e37416 (2012)

17. Johnson, T. A. et al. Natural Product Libraries to Accelerate the High-Throughput Discovery of Therapeutic Leads. J. Nat. Prod. 74, 2545-2555 (2011)

18. Butler. M. S. The Role of Natural Product Chemistry in Drug Discovery. J. Nat. Prod., 67, 2141-2153 (2004)

19. Singh, S. B., Young, K., & Miesel, L. Screening Strategies for Discovery of Antibacterial Natural Products. Exp. Rev. Anti-Infect. Ther. 9, 589–613 (2011)

20. Cane, D.E., Walsh, C.T. & Khosla, C. Harnessing the biosynthetic code: combinations, permutations, and mutations. Science 282, 63-68 (1998).

21. Nett, M., Ikeda, H. & Moore, B.S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. Nat. Prod. Rep. 26, 1362-1384 (2009).

22. Zazopoulos, E. et al. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. Nat. Biotechnol. 21, 187-190 (2003).

23. Johnston, C., Ibrahim, A. & Magarvey, N. Informatic strategies for the discovery of polyketides and nonribosomal peptides. Medchemcomm 3, 932-937 (2012).

24. Challis, G.L. Genome Mining for Novel Natural Product Discovery. J. Med. Chem. 51, 2618-2628 (2008).

25. Walsh, C. T., & Fischbach, M. A. Natural Products Version 2.0: Connecting Genes to Molecules. J. Am. Chem. Soc. 132, 2469–2493 (2010)

26. Little, J. L., Williams, A. J., Pshenichnov, A., & Tkachenko, V. Identification of "Known Unknowns" Utilizing Accurate Mass Data and ChemSpider. J. Am. Soc. Mass Spectrom. 23, 179-185 (2012)

27. Ng, J. et al. Dereplication and de novo sequencing of nonribosomal peptides. Nat. Methods 6, 596-599 (2009).

28. Kersten, R.D. et al. A mass spectrometry–guided genome mining approach for natural product peptidogenomics. Nat. Chem. Biol. 11, 974-802 (2011).

29. Ibrahim, A. et al. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. Proc. Natl. Acad. Sci. USA 109, 19196-19201 (2012).

30. Degenkolb, T., Kirschbaum, J. & Brückner, H. New sequences, constituents, and producers of peptaibiotics: an updated review. Chem. Biodivers. 4, 1052-1067 (2007).

31. He, H. et al. Culicinin D, an antitumor peptaibol produced by the fungus Culicinomyces clavisporus, strain LL-12I252. J. Nat. Prod. 69, 736-741 (2006).

32. Caboche, S., Leclere, V., Pupin, M., Kucherov, G. & Jacques, P. Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. J. Bacteriol. 192, 5143-5150 (2010).

33. Schüller, A., Hähnke, V. & Schneider, G. SmiLib v2.0: A Java-Based Tool for Rapid Combinatorial Library Enumeration. QSAR Combin. Sci. 26, 407-410 (2007).

34. Banik, J.J. & Brady, S.F. Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary. Proc. Natl. Acad. Sci. USA 105, 17273-17277 (2008).

35. Seyedsayamdost, M.R. et al. Mixing and matching siderophore clusters: structure and biosynthesis of serratiochelins from Serratia sp. V4. J. Am. Chem. Soc. 134, 13550-13553 (2012).

36. Piel, J. et al. Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge Theonella swinhoei. Proc. Natl. Acad. Sci. U S A 101, 16222-16227 (2004).

37. Johnston, C.W. et al. Gold biomineralization by a novel metallophore from a gold-associated microbe. Nat. Chem. Biol. 9, 241-243 (2012).

38. Fukuda, K., Tamura, T., Segawa, Y., Mutaguchi, Y., & Inagaki, K. Enhanced Production of the Fluorinated Nucleoside Antibiotic Nucleocidin by a rifR-Resistant Mutant of Streptomyces calvus IFO13200. Actinomycetologica 23, 51-55 (2009)

39. Hayashi, K. et al. WS9326A, A novel tachykinin antagonist isolated from Streptomyces violaceusniger No. 9326; Taxonomy, fermentation, isolation, physico-chemical properties and biological activities. J. Antibiot. (Tokyo) 45, 1055-1063 (1992)

40. Hashimoto, M. et al. WS9326A, A novel tachykinin antagonist isolated from Streptomyces violaceusniger No. 9326; Biological and pharmacological properties of WS9326A and tetrahydro-WS9326A (FK224). J. Antibiot. (Tokyo) 45, 1064-1070 (1992)

41. Pennefather, J. N. et al. Tachykinins and tachykinin receptors: a growing family. Life Sci. 74, 1445–1463 (2004)

42. Hoyer, D., & Bartfai, T. Neuropeptides and neuropeptide receptors: drug targets, and peptide and non-peptide ligands. Chem. Biodivers. 9, 2367-2387 (2012)

43. Yu, Z., Vodanovic-Jankovic, S., Kron, M., & Shen, B. New WS9326 congeners from Streptomyces sp. 9078 inhibiting Brugia malayi asparaginyl-tRNA synthetase. Org. Lett. 14, 4946-4949 (2012)

44. Eberhardt, L., Kumar, K., & Waldmann, H. Exploring and exploiting biologically relevant chemical space. Curr. Drug Targets 11, 1531-46 (2011).

45. Over, B. et al. Natural-product-derived fragments for fragment-based ligand discovery. Nature Chemistry 5, 21–28 (2013)

46. Rake, J.B. et al. Glycopeptide antibiotics: a mechanism-based screen employing a bacterial cell wall receptor mimetic. J. Antibiot. (Tokyo) 39, 58-67 (1986).

47. Loganzo, F. et al. HTI-286, a Synthetic Analogue of the Tripeptide Hemiasterlin, Is a Potent Antimicrotubule Agent that Circumvents P-Glycoprotein-mediated Resistance in Vitro and in Vivo. Cancer Res. 63, 1838-1845 (2003).

48. Gamble, W.R. et al. Cytotoxic and tubulin-interactive hemiasterlins from Auletta sp. and Siphonochalina spp. sponges. Bioorg. Med. Chem. 7, 1611-1615 (1999).

49. Coleman, J.E., Dilip de Silva, E., Kong, F., Andersen, R.J. & Allen, T.M. Cytotoxic peptides from the marine sponge Cymbastela sp. Tetrahedron 51, 10653-10662 (1995).

50. Pinel, N., Davidson, S.K. & Stahl, D.A. Verminephrobacter eiseniae gen. nov., sp. nov., a nephridial symbiont of the earthworm Eisenia foetida (Savigny). Int. J. Syst. Evol. Microbiol. 58, 2147-2157 (2008).

51. Steinmetz, H. et al. Elansolid A, a Unique Macrolide Antibiotic from Chitinophaga sancti Isolated as Two Stable Atropisomers. Angew. Chem. Int. Ed. 50, 532-536 (2010).

52. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D.H. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS)

using transductive support vector machines (TSVMs). Nuc. Acid. Res. 33, 5799-5808 (2005).

53. Ansari, M.Z., Yadav, G., Gokhale, R.S. & Mohanty, D. NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nuc. Acid. Res. 32, W405-W413 (2004).

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

55. Darling, A.E., Mau, B. & Perna, N.T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5, e11147 (2010).

56. Gutlein, M., Karwath, A. & Kramer, S. CheS-Mapper - Chemical Space Mapping and Visualization in 3D. J. Cheminform. 4, 7 (2012).

57. Harrison, S.J., et al. A focus on the preclinical development and clinical statuse of the histone deacetylase inhibitor, romidepsin (depsipeptide, Istodax). Epigen. 5, 571-589 (2012).

58. Mogi, T., Kita, K., Gramicidin S and polymyxins: the revival of cationic cyclic peptide antibiotics. Cell. Mol, Life. Sci. 66, 3821-3826 (2009).

59. Van Bambeke, F. Glycopeptides and glycodepsipeptides in clinical development a comparative review of their antibacterial spectrum, pharmacokinetics and clinical efficacy. Curr. Opin. Investig. Drugs. 8, 740-749 (2006).

60. Steenbergen, J.N., Alder, J., Thorne, G.M., Tally, F.P., Daptomycin: a lipopeptide antibiotic for the treatment of serious Gram-positive infections. J. Antibmicrob. Chemother. 55, 283-288 (2005).

61. Newman D.J. Cragg, G.M., Meeting the Supply Needs of Marine Natural Products. Handbook of Marine Natural Products 26, 1295-1296 (2012).

62. Eggen, M., Georg, G.I., The cryptophycins: their synthesis and anticancer activity. Med. Res. Rev. 22, 85-101 (2002).

Chapter 7. Significance and Future Prospective

Nonribosomal peptides are known for their structural complexity and diverse bioactivities. Although many NRPs are currently used as pharmaceutical agents, there has been a steady decline in their discovery and subsequent introduction into the clinic. Lack of new technologies to find and isolate novel NRPs has caused natural product discoveries to lag behind other scientific disciplines where advancements in technology have led to greater understanding of biology and disease. This thesis identifies new strategies for the direct identification of NRPs with strategies and technologies that combine genomic, bioinformatic, chemoinformatic and analytical techniques, that can be further applied to other natural product classes.

7.1 Genome Mining and NRPS Heterologous Expression

Bioactive small molecules are traditionally found in a random fashion through the systematic fermentation and bioactivity guided fractionation of microbial extracts. Rapid advances in genome sequencing and annotation now identify the genes responsible for the production of these molecules, while also revealing the presence of gene clusters encoding the production of yet to be discovered products. At the beginning of my research, the process of genome mining microbes for new NRPs was relatively nascent, with the majority of genetic analysis of a biosynthetic cluster occurring after the isolation of an NRP. Genome mining for new products was almost exclusively reserved for organisms already known to produce bioactive small molecules. For most researchers, the cost of genome sequencing was prohibitively expensive and the number of genomes publicly available was low and limited to microbes of intense study, such as human pathogens or model organisms, with little focus given for the gene clusters involved in secondary metabolism.

As genome sequencing became more commonplace, it was quickly realized that although the secondary metabolites produced by many of these microbes were extensively studied, the majority of biosynthetic gene clusters had no associated natural product. It was clear that new strategies to find the 'known unknowns' were needed and that a prediction based approach was the most desirable to leverage the newly available genomic information to aid in discovery. *Staphylococcus aureus* was a model organism that had been studied extensively for secreted effector molecules, revealing several important small molecules for virulence, including a reactive oxygen scavenger, staphyloxanthin, and a virulence regulator, the autoinducing peptide.[107-109] However, these studies never revealed the production of a nonribosomal peptide and *S. aureus* was never considered to be among the NRP producers until genome sequencing uncovered the presence of a cryptic dimodular NRPS. Since prediction-guided isolations were relatively new, most of the software currently available for predicting NRPs was not available, requiring manual annotation of the cryptic gene cluster. In this instance, a dimodular NRPS was ideal for proving that a prediction could guide the isolation of a cryptic NRP in an organism studied for over a century.

The study contained within **Chapter 2** was one of the first hypothesis driven approaches to natural product discovery guided by genetically predicting a small molecule from a genome, rather than its bioactivity. The small size of the NRPS generated an accurate prediction of the cyclic dipeptide aureusimine to guide the search for metabolites within a mass range surrounding the predicted mass. The isolation of the aureusimines was by prediction alone, and a genetic knockout of *ausA* was completed subsequent to the isolation. This experiment confirmed the source of the aureusimines and allowed us to examine their role in *S. aureus* pathogenesis. This was the first instance that an NRP without obvious activity was isolated in a directed fashion. Furthermore, the notion that secondary metabolites are important for the producer's fitness suggested there may be a role in pathogenesis. The discovery that the aureusimines control expression of several virulence genes is striking, however, an inadvertent mutation in our original study overestimated the aureusimine's role in virulence. Subsequent analysis shows that they are still relevant in *S. aureus* virulence regulation and may be important in other biological processes.

As a result of genome mining, we reveal the aureusimine NRPS is conserved across all human associated staphylococci including *S. aureus, S. epidermitis,* and *S. capitis*. Furthermore, production of aureusimine A, B and C is present in all Staphylococci strains bearing the *ausA* island. The conservation of the aureusimines is quite interesting considering that *S. aureus* has extremely high strain-to-strain virulence gene variability suggesting the aureusimines are an important aspect of staphylococci biology. Although several studies now show the aureusimines are active in a variety of capacities (gene regulation, protease inhibition), no absolute function is assigned. Emerging evidence suggests that aureusimine may be influencing the host, through gene expression and cytokine regulation (ongoing work in the Magarvey Lab).[110] The developments that *S. aureus* may be modulating host responses to ensure survival are compelling revelations. This may be akin to the discovery that certain *E. coli* lineages contain gene clusters encoding an NRP/PK hybrid molecule (colibactin) that causes DNA damage, increasing the instance of cancer in infected hosts.[111-113] It will be interesting when the full function of the aureusimines is revealed, as it will likely be a marked event for *S. aureus* research and the role microbiome-produced small molecules play in human health and disease. If not for genome mining and the prediction-based approach to NRP discovery, the aureusimines and their activities may be still left undiscovered. This research is a major contribution to not only the *S. aureus* community, but also highlights the new role natural products research can play for revealing biological phenomena in the face of increasing genomic data.

The isolation of 3 aureusimine variants highlights that monomer selection is flexible. This has ramifications for prediction-guided approaches to NRP discovery. Although A domain selectivity is predictable, in many cases this selectivity is directed towards structurally similar amino acids (e.g. valine/isoleucine, serine/threonine, phenylalanine/tyrosine, etc.), often with one amino acid being slightly favoured compared to another. For prediction-guided discovery, this means we need to appreciate that a predicted structure may not be the major NRP product, but represents just one of several possibilities generated by an NRPS assembly line. Understanding the flexibility within an NRPS helps improve the accuracy of predictions and gives an appreciation of other variants that may be present within the extract.

In **Chapter 3**, A domain flexibility and the Re domain gate-keeping properties of AusA were probed. This study shows the flexibility of monomer incorporation by NRPS assembly lines. For example, in AusA, the first A domain encodes the activation of a valine and *in vitro* analysis shows that loading of variant amino acids does not occur, while the second A domain encoding an aromatic amino acid is able to accommodate several proteinogenic and non-proteinogenic amino acids. The degree of flexibility for individual A domain predictions should be accounted for while predicting cryptic metabolites from the genome. The biochemical and structural analysis of the AusA Re domain also reveals Re domains can process a variety of substrates, common for other NRPS terminating domains such as TEs. This study also provides the first structure of a macrocycle forming NRPS Re domain, which adds structural insight into how NRPs are released from assembly lines.

The availability of biosynthetic gene cluster sequences provides an opportunity to clone and express clusters within heterologous hosts when the producer is not easily laboratory culturable or fails to produce the target metabolite. A fortunate consequence of expressing AusA in *E. coli* to probe its flexibility is that it could also be used as a model system for NRPS/NRP heterologous production within *E. coli* (**Chapter 4)**. Most overexpression experiments involving NRPSs involve the cloning and expression of only small portions of an NRPS gene and rarely are used for clusters due to the large size, preventing traditional cloning methods. The small size of the aureusimine biosynthetic gene cluster allowed for PCR-based cloning into an overexpression vector, creating an IPTG inducible production system for an entire NRP gene cluster. Using *E. coli* bearing an inducible plasmid containing aureusimine gene cluster, all of the aureusimines were produced in a non-pathogenic host including several analogs not produced by *S. aureus*. Many studies focus on the cloning and expression of NRPS systems in heterologous hosts due to the limitations of the producer (e.g. slow growing, unculturable, or pathogenic) or the need to manipulate NRP production in some way (e.g. gene swapping or deletion), this is one of the only studies within *E. coli* aimed at optimizing NRP production alone rather than as a function of protein expression. With less than 0.1% of bacteria being laboratory culturable, uncultured bacteria represent one of the largest sources of untapped genetic potential for the expression of NRPs.[114] By overexpressing these systems in genetically amenable laboratory hosts such as *E. coli*, we can further elaborate natural product space, which is exemplified by the identification of 3 additional pyrazine assembly line products only produced by the *E. coli* host.

At the time of the aureusimine discovery, software and online tools for fully predicting and elaborating NRP and PK molecules were in their infancy. Gene cluster domain architecture was determined by BLAST alignment to conserved domain motifs and few online tools were available to predict A domain substrate selectivity. However, many algorithms and online software tools have advanced to more easily predict NRPS gene clusters from raw genomic data, allowing expansion of this prediction-based approach to include complex gene clusters. These algorithms can now identify genes involved in specialized amino acid biosynthesis and more accurately predict A domain selectivities for non-proteinogenic amino acids. These types of advancements were important for the prediction-guided discovery of delftibactin.

In **Chapter 5**, a cryptic gene cluster was identified within a gold-associated microbe, *D. acidovorans*. Applying genome mining techniques, a prediction of a large NRP was generated for the cryptic cluster to guide isolation. Compared to the 2 gene cluster for aureusimine biosynthesis, the *D. acidovorans* cluster is significantly more complex containing over 20 genes, including 4 large NRPS/PKS genes comprising 11 biosynthetic modules and several tailoring/modifying enzymes. Fortunately, the complexity provided clues to the NRPs function, with several genes being previously studied for their roles in the biosynthesis of metal binding moieties and several flanking genes responsible for heavy metal tolerance. As more tailoring functions and NRPS modification rules are revealed, our predictions become more accurate, which can also help reveal the unknown NRPs function by the presence of pharmacophores or other functional chemical moieties.

For *D. acidovorans*, a combined prediction and bioactivity based approach was used to hone in on the large NRP. A positive bioactivity hit for gold complexation from a compound with a mass close to the predicted structure, directed our isolation of the novel metallophore, delftibactin. Delftibactin represents one of the largest NRPs isolated using genomic prediction and reveals a new form of metallophore that provides heavy metal toxicity protection and a mechanism for gold biomineralization. Although internalization of gold ions to form solid gold particles by microbes is known, this is the first defined mechanism of a small molecule-gold complexing agent, which has several industrial implications: such as the biomining of gold from water reservoirs and for use as a gold biosensor.[115]

7.2 Computational Fragment-based NRP Discovery Strategies

The iSNAP algorithm identifies NRPs from extracts by matching *in silico* fragmentation of known NRPs and comparing them to real MS/MS fragmentation patterns obtained from LC-MS/MS experiments. In the previous two examples, the prediction never matched the final structure exactly, precluding the use of iSNAP for identifying genomically envisioned NRPs by a single prediction alone. However, in the previous two examples, the NRPS assembly lines produced several variants rather than a single compound, suggesting a prediction-guided strategy using a single structure is inherently flawed. By predicting every possibility of an NRPS assembly line, an accurate match is likely generated and this database of predicted hypothetical structures can be used by iSNAP to physically localize the NRP of interest within LC-MS/MS chromatrograms of crude extracts alone.

In **Chapter 6**, iSNAP is first shown to identify two peptaibols, efrapeptin and trichopolyn. Like other NRPS derived molecules, many variants are produced by the assembly line and in fact several are known. By generating a hypothetical library based on putative assembly line products arising from the trichopolyn NRPS, we were able to identify 6 additional trichopolyn variants. More impressively, iSNAP was able to identify the exact variant match that corresponded to the real trichopolyn structure revealing 3 completely novel structures. This extends the utility of iSNAP to find unknown but envisioned compounds (known unknowns), and also shows how iSNAP can provide structural information without the need for purification and full structure elucidation.

A similar approach was extended to identify NRPS products from cryptic gene clusters, leading to the generation of larger hypothetical prediction databases due to inherent NRPS flexibility and numerous prediction possibilities. By generating prediction databases for cryptic clusters found within *Acidovorax citrulli* AAC00-1 and *Variovorax paradoxus* S110, 5 'known unknown' NRPs were detected (acidobactin A, B and C and vacidobactin A and B). These libraries contained 576 predicted structures based on genomic prediction and iSNAP physically localized and identified the LC-MS/MS peak associated with the acidobactins and vacidobactins. The physical location of the NRP of interest within the LC-MS/MS chromatogram allowed for quick isolation and structure verification without the need for gene knockouts or bioactivity.

The predicted library approach was used successfully again to identify a cryptic NRP, thanamycin. Thanamycin was previously identified several times by researchers; however, the production was low and isolation strategies were not successful in providing sufficient quantities for full structure elucidation. In this instance, the power of our prediction database for aiding in structure elucidation is exemplified by the accurate selection of the real thanamycin structure from a pool of predicted variants. This is the first instance of real structural information being ascertained from genomic predictions of an unknown compound prior to isolation.

Another extremely powerful aspect of the hypothetical library based approach to NRP identification is that we can identify organisms that produce desired NRP structural classes. This is achieved by examining the microbial extract directly by LC-MS/MS without the need for genomic information. As mentioned earlier, Ecopia Biosciences sought to identify new producers of enediyne variants by sequencing organisms and bioinformatically identifying those with gene clusters associated with the enediyne warhead.[70] However, the time and cost associated with isolating genomic DNA, sequencing, and assembling a microbial genome, makes this approach prohibitive for large-scale research programs attempting to identify target pharmacophores from large microbial collections. In addition, this approach does not overcome the problems associated with natural product discovery: determining NRP production conditions and physically locating the desired molecule within an extract. The pattern-based hypothetical database approach allows for input of all predicted structures surrounding a desired scaffold or pharmacophore into iSNAP and by scanning LC-MS/MS chromatograms of microbial crude extract libraries, the desired structural homolog can be localized within unknown producing organisms. This is a huge advantage and technology for pharmaceutical companies that maintain millions of extracts in libraries with no knowledge of the molecular components within them. Our identification of variobactin from an environmental library using a hypothetical structural database of a desired scaffold, demonstrates the power of this approach for the directed discovery of desired metabolites from crude extracts.

The extension of the iSNAP dereplication software to accept hypothetical compound libraries and evidence that the algorithm locates predicted or sought after peptides within complex extracts represents a major bound forward for directing natural product discovery. This work highlights how NRPs or peptidic compounds can be detected and isolated from crude extracts in a directed fashion using LC-MS/MS data and genomic predictions or structural data without the aid of crude bioactivity assays. This aspect alone will increase the number of NRPs discovered as bioactivity guided experiments enrich in and target generally cytotoxic agents and can miss low abundant molecules that could be pharmaceutical leads. Mass spectrometry is a much more sensitive analytical technique and can identify molecules in the picogram quantities which could only be identified through large-scale fermentation and concentration steps in bioactivity guided discovery programs. Using hypothetical molecular barcodes and matching their fragments to real spectra, we can conduct the same analysis directly from small-scale fermentations. More importantly, the iSNAP platform can determine the structures of unknown, but related compounds found within the extract without the need for purification and extensive structure elucidation experiments (e.g. NMR). This is a major advancement for natural product discovery programs as the majority of the costs are associated with purifying enough quantity of the unknown metabolite for NMR studies and the time associated with full structure elucidation.

Although iSNAP was originally developed for peptidic compounds, a similar *in silico* fragmentation strategy can be applied to generate hypothetical fragmentation libraries. These would identify desired/predicted/unknown natural products and extend its utility for identifying molecules from other secondary metabolite gene clusters. The majority of products from secondary metabolite gene clusters have not been found using traditional bioactivity screens suggesting other means to identify them are necessary. The current reductionist approach to systematically create gene cluster knockouts and isolate these compounds directly is outpaced by genomic sequencing. The creation of prediction databases for all cryptic gene clusters is now possible and these compounds can be systematically identified and isolated using iSNAP without the need for time-consuming knockout experiments or bioactivity (**Figure 7.1**). In addition, by focusing on cryptic clusters, the potential for rediscovery of knowns is mitigated and every compound isolated is a potential lead. It is no longer adequate to conduct natural product research based on chance alone and the use of bioactivity-guided fractionation does not leverage genomic information. Although traditional natural product discovery methods are proven successful, advancements in other sciences can now be applied in the search for natural products. Implementation of informatic strategies such as iSNAP using fragment-based molecular barcodes will pave the way for directed discovery of desired pharmacophores, novel genetically predicted compounds, and analogs surrounding an active compound that may lead to new pharmaceutical drug leads.

.

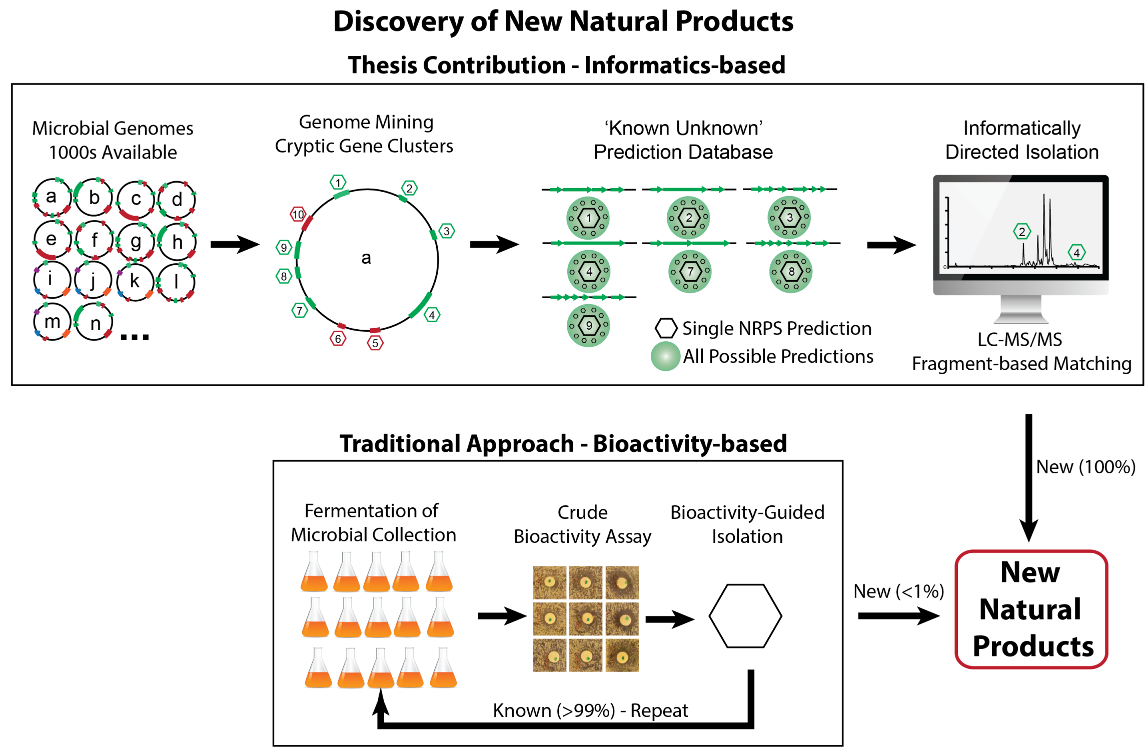


Figure 7.1 Discovery of Novel Natural Products.

Overview of my thesis contribution to informatics based strategies and traditional natural products discovery approaches.

7.3 Concluding Remarks

The success of natural products as drugs is evident with the multitude of treatments available for human diseases. The lack of technologies to identify new potent bioactive small molecules in a timely and cost efficient manner and years of rediscovery of known compounds has ultimately led to their decline within pharmaceutical discovery programs. The wealth of genomic information now available for microbes suggests the majority of natural products remain to be discovered. Directly targeting these predicted compounds is important for revitalizing natural product drug discovery. Advances in other fields, such as molecular biology, genomics, analytical chemistry and bioinformatics, offer new technology platforms to launch natural products research into a new era of drug discovery. With the ability to generate genomic and metabolomic data rapidly, it is necessary to advance natural product informatics, technologies and strategies concurrently to discover new industrial and pharmaceutically important small molecules. Collectively, this thesis sets out new discovery methodology and technology for the discovery of NRPs, advancing significantly from traditional natural product discovery strategies.

References.

1. Clardy, J. & Walsh, C. Lessons from natural molecules. *Nature* **432**, 829-837 (2004).
2. Walsh, C.T. & Fischbach, M.A. Natural products version 2.0: connecting genes to molecules. *J. Am. Chem. Soc.* **132**, 2469-2493 (2010).
3. Cragg, G.M. & Newman, D.J. Biodiversity: A continuing source of novel drug leads. *Pur. App. Chem.* **77**, 7-24 (2005).
4. Dias, D.A., Urban, S. & Roessner, U. A Historical Overview of Natural Products in Drug Discovery. *Metabol.* **2**, 303-336 (2012).
5. Heinrich, M. Ethnobotany and its role in drug development. *Phytother. Res.* **14**, 479-488 (2000).
6. Heinrich, M. & Gibbons, S. Ethnopharmacology in drug discovery: an analysis of its role and potential contribution. *J. Pharm. Pharmacol.* **53**, 425-432 (2001).
7. Yun, U.W. et al. Plant natural products: history, limitations and the potential of cambial meristematic cells. *Biotechnol. Gen. Eng. Rev.* **28**, 47-59 (2012).
8. Schmitz, R. Friedrich Wilhelm Sertürner and the Discovery of Morphine. *Pharm. Hist.* **27**, 61-74 (1985).
9. Ugurlucan, M. et al. Aspirin: from a historical perspective. *Rec. Pat. Cardio. Drug Disc.* **7**, 71-76 (2012).

10. Tata, J.R. One hundred years of hormones - A new name sparked multidisciplinary research in endocrinology, which shed light on chemical communication in multicellular organisms. *Embo Rep.* **6**, 490-496 (2005).

11. Roth, J. et al. Insulin's discovery: New insights on its ninetieth birthday. *Diabetes-Metab. Res.* **28**, 293-304 (2012).

12. Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to Their Use in the Isolation of B. influenzae. *Brit. J. Exp. Pathol.* **10**, 226-236 (1928).

13. Vicente, M.F., Basilio, A., Cabello, A. & Pelaez, F. Microbial natural products as a source of antifungals. *Clin. Microbiol. Infec.* **9**, 15-32 (2003).

14. Schatz, A. & Waksman, S.A. Effect of Streptomycin and Other Antibiotic Substances upon Mycobacterium tuberculosis and Related Organisms. *Exper. Biol. Med.* **57**, 244-248 (1944).

15. Waksman, S.A. & Tishler, M. THE CHEMICAL NATURE OF ACTINOMYCIN, AN ANTI-MICROBIAL SUBSTANCE PRODUCED BY ACTINOMYCES ANTIBIOTICUS. *J. Biol. Chem.* **142**, 519-528 (1942).

16. Griffith, R.S. Introduction to Vancomycin. *Rev. Infect. Dis.* **3**, S200-S204 (1981).

17. Pelaez, F. The historical delivery of antibiotics from microbial natural products - Can history repeat? *Biochem. Pharmacol.* **71**, 981-990 (2006).

18. Newman, D.J. & Cragg, G.M. Microbial antitumor drugs: Natural products of microbial origin as anticancer agents. *Curr. Opin. Invest. Dr.* **10**, 1280-1296 (2009).

19. Fleming, A. The Discovery of Penicillin. *Brit. Med. Bulletin* **2**, 4-5 (1944).

20. Feng, Y. et al. Evolution and pathogenesis of Staphylococcus aureus: lessons learned from genotyping and comparative genomics. *FEMS Microbiol. Rev.* **32**, 23-37 (2008).

21. Wernegreen, J.J. Genome evolution in bacterial endosymbionts of insects. *Nat. Rev. Gen.* **3**, 850-861 (2002).

22. Zaneveld, J. et al. Host-bacterial coevolution and the search for new drug targets. *Curr. Opin. Chem. Biol.* **12**, 109-114 (2008).

23. Madigan, M.T. Extremophilic Bacteria and Microbial Diversity. *Annals Missouri Bot. Gar.***87**, 3-12 (2000).

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

25. Keller, L. & Surette, M.G. Communication in bacteria: an ecological and evolutionary perspective. *Nat. Rev. Microbiol.* **4**, 249-258 (2006).

26. Peláez, F. The historical delivery of antibiotics from microbial natural products--can history repeat? *Biochem. Pharmacol.* **71**, 981-990 (2006).

27. Bérdy, J. Bioactive microbial metabolites. *J. Antibiot.* **58**, 1-26 (2005).

28. Varma, A. & Chincholkar, S.B. Microbial siderophores. (Springer Verlag, 2007).

29. Kleinkauf, H. & von Dohren, H. The nonribosomal peptide biosynthetic system--on the origins of structural diversity of peptides, cyclopeptides and related compounds. *Anton. Leeuwen.* **67**, 229-242 (1995).

30. Newman, D.J. & Cragg, G.M. Natural Products As Sources of New Drugs over the 30 Years from 1981 to 2010. *J. Nat. Prod* **75**, 311-335 (2012).

31. Kresge, N., Simoni, R.D. & Hill, R.L. Selman Waksman: the Father of Antibiotics. *J. Biol. Chem.* **279**, e7 (2004).

32. Waksman, S.A., Schatz, A. & Reynolds, D.M. Production of Antibiotic Substances by Actinomycetes. *Annals NY Acad. Sci.* **1213**, 112-124 (2010).

33. Weller, M.G. A Unifying Review of Bioassay-Guided Fractionation, Effect-Directed Analysis and Related Techniques. *Sensors-Basel.* **12**, 9181-9209 (2012).

34. Clardy, J., Fischbach, M.A. & Walsh, C.T. New antibiotics from bacterial natural products. *Nat, Biotechnol.* **24**, 1541-1550 (2006).

35. Walsh, C. Where will new antibiotics come from? *Nat. Rev. Microbiol.* **1**, 65-70 (2003).

36. Avery, O.T., Macleod, C.M. & Mccarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types - Induction of Transformation by a Deoxyribonucleic-Acid Fraction Isolated from

Pneumococcus Type-Iii (Reprinted from Journal of Experimental Medicine, Vol 79, Pg 137-158, 1944). *Mol. Med.* **1**, 344-365 (1995).

37. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).

38. Matthaei, J.H., Jones, O.W., Martin, R.G. & Nirenberg, M.W. Characteristics and composition of RNA coding units. *Proc. Natl. Acad. Sci. U S A* **48**, 666-677 (1962).

39. Walsh, C.T. Polyketide and nonribosomal peptide antibiotics: modularity and versatility. *Science* **303**, 1805-1810 (2004).

40. Marahiel, M., Stachelhaus, T. & Mootz, H. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem. Rev.* **97**, 2651-2673 (1997).

41. Mootz, H.D. & Marahiel, M.A. Biosynthetic systems for nonribosomal peptide antibiotic assembly. *Curr. Opin. Chem. Biol.* **1**, 543-551 (1997).

42. Stachelhaus, T. & Marahiel, M.A. Modular structure of genes encoding multifunctional peptide synthetases required for non-ribosomal peptide synthesis. *FEMS Microbiol Lett.* **125**, 3-14 (1995).

43. Chan, Y.A., Podevels, A.M., Kevany, B.M. & Thomas, M.G. Biosynthesis of polyketide synthase extender units. *Nat. Prod. Rep.***26**, 90 (2008).

44. Donadio, S., Monciardini, P. & Sosio, M. Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. *Nat. Prod. Rep.* **24**, 1073-1109 (2007).

45. Keating, T.A. & Walsh, C.T. Initiation, elongation, and termination strategies in polyketide and polypeptide antibiotic biosynthesis. *Curr. Opin. Chem. Biol.* **3**, 598-606 (1999).

46. Staunton, J. & Weissman, K.J. Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **18**, 380-416 (2001).

47. Cortes, J., Haydock, S.F., Roberts, G.A., Bevitt, D.J. & Leadlay, P.F. An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of Saccharopolyspora erythraea. *Nature* **348**, 176-178 (1990).

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

49. Fischbach, M.A. & Walsh, C.T. Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.* **106**, 3468-3496 (2006).

50. Mootz, H.D., Schwarzer, D. & Marahiel, M.A. Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *ChemBioChem* **3**, 490-504 (2002).

51. Lambalot, R.H. et al. A new enzyme superfamily - the phosphopantetheinyl transferases. *Chem. Biol.* **3**, 923-936 (1996).

52. Du, L. & Lou, L. PKS and NRPS release mechanisms. *Nat. Prod. Rep.* **27**, 255 (2010).

53. Finking, R. & Marahiel, M.A. Biosynthesis of nonribosomal peptides1. *Ann. Rev Microbiol.* **58**, 453-488 (2004).

54. Walsh, C.T., O'Brien, R.V. & Khosla, C. Nonproteinogenic Amino Acid Building Blocks for Nonribosomal Peptide and Hybrid Polyketide Scaffolds. *Angew. Chem. Int. Ed.* **52**, 7098-7124 (2013).

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

56. Yadav, G., Gokhale, R.S. & Mohanty, D. Towards prediction of metabolic products of polyketide synthases: an in silico analysis. *PLoS Comput. Biol.* **5**, e1000351 (2009).

57. Challis, G.L. Mining microbial genomes for new natural products and biosynthetic pathways. *Microbiol.* **154**, 1555-1569 (2008).

58. Challis, G.L. Genome Mining for Novel Natural Product Discovery. *J. Med. Chem.* **51**, 2618-2628 (2008).

59. Samel, S.A., Marahiel, M.A. & Essen, L.-O. How to tailor non-ribosomal peptide products—new clues about the structures and mechanisms of modifying enzymes. *Mol. Biosyst.* **4**, 387 (2008).

60. Walsh, C.T. et al. Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines. *Curr. Opin Chem. Biol.* **5**, 525-534 (2001).

61. Schreiber, S.L. Small molecules: the missing link in the central dogma. *Nat. Chem. Biol.* **1**, 64-66 (2005).

62. Persidis, A. High-throughput screening. Advances in robotics and miniturization continue to accelerate drug lead identification. *Nat. Biotechnol.* **16**, 488-489 (1998).

63. Koehn, F.E. & Carter, G.T. The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discov.* **4**, 206-220 (2005).

64. Timmis, K.N. Golden age of drug discovery or dark age of missed chances? *Environ. Microbiol.* **7**, 1861-1863 (2005).

65. Bentley, S.D. et al. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141-147 (2002).

66. Koch, M.A. et al. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc. Nat. Acad. Sci. USA.* **102**, 17272-17277 (2005).

67. Renner, S. et al. Bioactivity-guided mapping and navigation of chemical space. *Nat. Chem. Biol.* **5**, 585-592 (2009).

68. Magarvey, N.A. et al. Biosynthetic characterization and chemoenzymatic assembly of the cryptophycins. Potent anticancer agents from cyanobionts. *ACS Chem. Biol.* **1**, 766-779 (2006).

69. Walsh, C.T. The chemical versatility of natural-product assembly lines. *Acc Chem. Res.* **41**, 4-10 (2008).

70. Zazopoulos, E. et al. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. *Nat. Biotechnol.* **21**, 187-190 (2003).

71. Smith, A.L. & Nicolaou, K.C. The enediyne antibiotics. *J. Med. Chem.* **39**, 2103-2117 (1996).

72. Liu, W., Christenson, S.D., Standage, S. & Shen, B. Biosynthesis of the enediyne antitumor antibiotic C-1027. *Science* **297**, 1170-1173 (2002).

73. Ahlert, J. et al. The calicheamicin gene cluster and its iterative type I enediyne PKS. *Science* **297**, 1173-1176 (2002).

74. Lautru, S., Deeth, R.J., Bailey, L.M. & Challis, G.L. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. *Nat. Chem. Biol.* **1**, 265-269 (2005).

75. Tanaka, Y. et al. Activation and Products of the Cryptic Secondary Metabolite Biosynthetic Gene Clusters by Rifampin Resistance (rpoB) Mutations in Actinomycetes. *J. Bacteriol.* **195**, 2959-2970 (2013).

76. Starcevic, A. et al. ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nuc. Acid. Res.* 11 (2008).

77. Weber, T. et al. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J. Biotechnol.* **140**, 13-17 (2009).

78. Anand, S. et al. SBSPKS: structure based sequence analysis of polyketide synthases. *Nuc. Acid. Res.* **38**, W487-496 (2010).

79. Khaldi, N. et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fung. Gen. Biol.* **47**, 736-741 (2010).

80. Rottig, M. et al. NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nuc. Acid. Res.* **39**, W362-367 (2011).

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

82. Blin, K. et al. antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nuc. Acid. Res.* **41**, W204-W212 (2013).

83. Medema, M.H. et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nuc. Acid. Res.*, 1-8 (2011).

84. Weissman, K.J. & Muller, R. Myxobacterial secondary metabolites: bioactivities and modes-of-action. *Nat. Prod. Rep.* **27**, 1276-1295 (2010).

85. Wyatt, M.A., Lee, J., Ahilan, Y. & Magarvey, N.A. Bioinformatic evaluation of the secondary metabolism of antistaphylococcal environmental bacterial isolates. *Can. J. Microbiol.* **59**, 465-471 (2013).

86. Rooijakkers, S.H.M., van Kessel, K.P.M. & van Strijp, J.A.G. Staphylococcal innate immune evasion. *Trend. Microbiol.* **13**, 596-601 (2005).

87. Reith, F. Biomineralization of Gold: Biofilms on Bacterioform Gold. *Science* **313**, 233-236 (2006).

88. Johnston, C.W. et al. Gold biomineralization by a metallophore from a gold-associated microbe. *Nat. Chem. Biol.* **9**, 241-243 (2013).

89. Wyatt, M.A. et al. Staphylococcus aureus Nonribosomal Peptide Secondary Metabolites Regulate Virulence. *Science*, 1-4 (2010).

90. Wyatt, M.A., Mok, M.C., Junop, M. & Magarvey, N.A. Heterologous expression and structural characterisation of a pyrazinone natural product assembly line. *Chembiochem* **13**, 2408-2415 (2012).

91. Wyatt, M.A. & Magarvey, N.A. Optimizing dimodular nonribosomal peptide synthetases and natural dipeptides in an Escherichia coli heterologous host. *Biochem. Cell Biol.* 1-6 (2013).

92. Ibrahim, A. et al. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proc. Natl. Acad. Sci. USA* **109**, 19196-19201 (2012).

93. Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E. & Pevzner, P.A. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **6**, 327-342 (1999).

94. Taylor, J.A. & Johnson, R.S. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid. Commun. Mass Spectrom.* **11**, 1067-1075 (1997).

95. Ma, B. et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337-2342 (2003).

96. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophor.* **20**, 3551-3567 (1999).

97. Eng, J.K., Mccormack, A.L. & Yates, J.R. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **5**, 976-989 (1994).

98. Mohimani, H. et al. Cycloquest: identification of cyclopeptides via database search of their mass spectra against genome databases. *J. Proteome. Res.* **10**, 4505-4512 (2011).

99. Mohimani, H. et al. Sequencing cyclic peptides by multistage mass spectrometry. *Proteom.* **11**, 3642-3650 (2011).

100. Mohimani, H. et al. Multiplex de novo sequencing of peptide antibiotics. *J. Comput. Biol.* **18**, 1371-1381 (2011).

101. Ng, J. et al. Dereplication and de novo sequencing of nonribosomal peptides. *Nat. Meth.* **6**, 596-599 (2009).

102. Liu, W.-T. et al. Interpretation of Tandem Mass Spectra Obtained from Cyclic Nonribosomal Peptides. *Anal. Chem.* **81**, 4200-4209 (2009).

103. Kavan, D., Kuzma, M., Lemr, K., Schug, K. & Havlicek, V. CYCLONE—A Utility for De Novo Sequencing of Microbial Cyclic Peptides. *J. Am. Soc. Mass Spectrom.*, 1-8 (2013).

104. Nguyen, D.D. et al. MS/MS networking guided analysis of molecule and gene cluster families. *Proc. Natl. Acad. Sci. USA* **110**, E2611-2620 (2013).

105. Kersten, R.D. et al. A mass spectrometry–guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* (2011).

106. Wyatt, M.A.J., C.W.; Li, X.; Yang, L.; Grunwald, A.; Vanner, S.A.; Ibrahim, A.; Zechel, D.L; Kerr, R.G.; Ma, B.; Magarvey, N.A. Directed discovery of unknown natural products using fragment-based molecular barcodes. *Nat. Biotechnol.* ***In Review*** (2013).

107. Clauditz, A., Resch, A., Wieland, K.-P., Peschel, A. & Götz, F. Staphyloxanthin plays a role in the fitness of Staphylococcus aureus and its ability to cope with oxidative stress. *Infect. Immun.* **74**, 4950-4953 (2006).

108. Pelz, A. Structure and Biosynthesis of Staphyloxanthin from Staphylococcus aureus. *J. Biol. Chem.* **280**, 32493-32498 (2005).

109. Novick, R.P. & Geisinger, E. Quorum sensing in staphylococci. *Ann. Rev. Gen.* **42**, 541-564 (2008).

110. Secor, P.R. et al. Phevalin (aureusimine B) production by Staphylococcus aureus biofilm and impacts on human keratinocyte gene expression. *PLoS One* **7**, e40973 (2012).

111. Cuevas-Ramos, G. et al. Escherichia coli induces DNA damage *in vivo* and triggers genomic instability in mammalian cells. *Proc. Nat. Acad. Sci. USA* **107**, 11537-11542 (2010).

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

113. Arthur, J.C. et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120-123 (2012).

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

115. Reith, F. et al. Mechanisms of gold biomineralization in the bacterium Cupriavidus metallidurans. *Proc. Nat. Acad. Sci. USA* **106**, 17757-17762 (2009).