UNIVERSITY OF CALIFORNIA, SAN DIEGO

Mapping Molecular Space with Mass Spectrometry

A Thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Chemistry

by

Don Duy Nguyen

Committee in charge:

Professor Pieter C. Dorrestein, Chair
Professor Susan S. Golden
Professor Carlos A. Guerrero

2013

The Thesis of Don Duy Nguyen is approved, and it is acceptable in quality and form for the publication of microfilm and electronically:

_____

_____

_____

Chair

University of California, San Diego

2013

DEDICATION

To Joe Immel, my friend and teacher, for all your wisdom and guidance.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

DNP	2,4-dinitrophenol
DHB	2,5-dihydroxy benzoic acid
DAPI	4',6-diamidino-2-phenylindole
A-domain	Adenylation Domain
Asn	Asparagine
CID	Collision Induced Dissociation
C-domain	Condensation Domain
Da	Dalton
DOE JGI	Department of Energy Joint Genome Institute
DMSO	Dimethyl Sulfoxide
ETD	Electron Transfer Dissocation
ESI	Electrospray Ionization
ESI-MS	Electrospray Ionization Mass Spectrometry
E-domain	Epimerization Domain
FT	Fourier Transform
FTICR-MS	Fourier Transform Ion Cyclotron Resonance Mass Spectrometry
FT-MS	Fourier Transform Mass Spectrometry
GCFs	Gene Cluster Families
GCF	Gene Cluster Family
HMBC	Heteronuclear Multiple Bond Correlation
HSQC	Heteronuclear Single Quantom Coherence
HPLC	High-Performance Liquid Chromatography
IMS	Imaging Mass Spectrometry
IRMPD	Infrared Multiphoton Dissociation
ISP2	International Streptomyces Project 2
ICR	Ion Cyclotron Resonance
Ile	Isoleucine
kV	Kilovolts
Leu	Leucine
LTQ	Linear Trap Quadrupole
LTQ-FT	Linear Trap Quadrupole Fourier Transform
LTQ-FTICR-MS	Linear Trap Quadrupole Fourier Transform Ion Cyclotron Resonance Mass Spectrometry
LC-MS/MS	Liquid Chromatography Tandem Mass Spectrometry
MS	Mass Spectrometry
MS^1	Mass Spectrometry
m/z	Mass-to-Charge
MALDI	Matrix Assisted Laser Desorption Ionization
MALDI-IMS	Matrix Assisted Laser Desorption Ionization Imaging Mass Spectrometry

MALDI-MS	Matrix Assisted Laser Desorption Ionization Mass Spectrometry
MALDI-TOF	Matrix Assisted Laser Desorption Ionization Time of Flight
MALDI-TOF-MS	Matrix Assisted Laser Desorption Ionization Time of Flight Mass Spectrometry
MHz	Megahertz
µL	Microliter
µL	Micrometer
mg/mL	Miligrams per Mililiter
mL/min	Mililiter per Minute
ms	milisecond
MFs	Molecular Families
MF	Molecular Family
nanoDESI	Nanospray Desorption Electrosrapy Ionization
nanoDESI-MS	Nanospray Desorption Electrosrapy Ionization Mass Spectrometry
NCBI	National Center for Biotechnology Information
NRP	Nonribosomal Peptide
NRPS	Nonribosomal Peptide Synthetase
NRPS/PKS	Nonribosomal Peptide Synthetase/Polyketide Synthase
NRPSs	Nonribosomal Peptide Synthetases
MS^n	n^{th} level of Tandem Mass Spectrometry
NMR	Nuclear Magnetic Resonance
blastn	Nucleotide Basic Local Alignment Search Tool
OD	Optical Density
PKS	Polyketide Synthase
PSD	Pst Source Decay
Ser	Serine
MS/MS	Tandem Mass Spectrometry
MS^2	Tandem Mass Spectrometry
T	Tesla
Thr	Threonine
TOF	Time of Flight
Val	Valine
WLIP	White Line-Inducing Principle

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my research advisor, Pieter Dorrestein, for his support as the chair of my committee and for granting me the opportunity to work in his lab. It is a great environment to be in, one that is fast-paced and always enthusiastic about the science. I would also like to thank the other members of my committee, Professors Susan Golden and Professor Carlos Guerrero for taking the time to review this thesis, providing valuable comments, and attending my defense.

Of course, none of this could have been accomplished without the help, guidance, and support of the Dorrestein Lab members (past and present)—Mike Meehan, Jane Yang, Wilna Moree, Xiling Zhao, Jeramie Watrous, Roland Kersten, Wei-Ting Liu, David Gonzalez, Cheng-Chih (Richard) Hsu, Cheng-Hsuan (Bill) Wu, Vanessa Phelan, Chris Rath, Laura Sanchez, Carla Porto, Amina Bouslimani, Cliff Kapono, Yao Peng, Jimmy Zeng, Jennifer Fang, and of course Kathleen and Tatiana Dorrestein.

Chapter 2, in full, has been submitted for publication of the material as it may appear in the Proceedings of the National Academy of Sciences, 2013, Nguyen, Don D.; Wu, Cheng-Hsuan; Moree, Wilna J; Lamsa, Anne; Medema, Marnix H.; Zhao, Xiling; Gavilan, Ronnie G.; Aparicio, Marystella; Atencio, Librada; Jackson, Chanaye; Ballesteros, Javier; Sanchez, Joel; Watrous, Jeramie D.; Phelan, Vanessa V.; van de Wiel, Corine; Kersten, Roland D.; Mehnaz, Samina; De Mot, René; Shank, Elizabeth A.; Charusanti, Pep; Nagarajan, Harish; Duggan, Brendan M.; Moore, Bradley S.; Bandeira, Nuno. Palsson, Bernhard Ø.; Pogliano, Kit; Gutiérrez, Marcelino; Dorrestein, Pieter C. The thesis author as well as Cheng-Hsuan Wu and Wilna J. Moree were the primary investigators and authors of this paper. I would like to thank them, as well as the co-authors for granting me permission to use this work.

ABSTRACT OF THE THESIS

Mapping Molecular Space with Mass Spectrometry

by

Don Duy Nguyen

Master of Science in Chemistry

University of California, San Diego, 2013

Professor Pieter C. Dorrestein, Chair

Mass spectrometry has become an invaluable tool in the discovery and characterization of specialized metabolites and biotechnologically relevant molecules. One aspect of characterizing specialized metabolites involves connecting these molecules to their biosynthetic machineries.

Here we employ several mass spectrometry methods and introduce the idea of connecting molecular families to their gene cluster families on a large scale that takes advantage of publicly accessible genomes. As proof of principle, we used molecular networking to analyze sixty bacteria, 42 bacilli and 18 pseudomonads, simultaneously and matched eight molecular families to their gene cluster families. To illustrate the effectiveness of this technique, we combined it with imaging mass spectrometry to examine two marine *Pseudoalteromonas* that both showed inhibition against *Bacillus subtilis* 3610. The signals showing bioactivity against *B. subtilis* were shown to be a family of molecules for which the biosynthetic gene cluster was discovered using a publiclyavailable genome sequence from *Pseudoalteromonas piscicida* JCM 20779^{T}.

Chapter 1

Specialized Metabolites and Mass Spectrometry

Specialized metabolites are chemical compounds produced by living organisms that include natural products and secondary metabolites (1). All other metabolites, such as those involved in primary metabolism, are common metabolites. The mechanisms that create specialized metabolites, referred to as secondary metabolic pathways, are found in only certain organisms or groups of organisms, and are not produced under all conditions (1, 2). Specialized metabolites include classes of compounds such as alkaloids, terpenoids, polyketides, fatty acids, and nonribosomal peptides and can be produced by bacteria, fungi, lichens, marine invertebrates, plants, and insects. While not essential for growth or reproduction, these compounds can have antibiotic, antifungal, or antitumor properties. Additionally, these compounds can act as signaling molecules for cell-cell communication, metal complexing and transporting agents, and can modulate bacterial communities, although, it is still unclear what the true role of these compounds are in nature (3). Historically, humans have been using specialized metabolites primarily for their medicinal capabilities but also as dyes and fragrances. More often than not, herbs and plant products were the primary resource for these medications, but over time, chemical techniques were developed and advanced to show that there were specific constituents responsible for healing properties. Herein, the focus of discussion will be on several mass spectrometry based methods used to study the parvome, or the world of low molecular weight bioactive compounds from microbial sources (3, 4).

It was not until the discovery of penicillin (Figure 1-1) in late 1920's that specialized metabolites derived from microorganisms could be put to use. Alexander Fleming showed that if a fungus, *Penicillium rubens,* was grown on specific substrates, it would exhibit antibacterial properties against a strain of *Staphylococcus* (5, 6). This began the age of antibiotics. Screening microorganisms became a popular technique to elucidate bioactivity of specialized metabolites. Samples from soil and water were tested against pathogenic organisms leading to the discovery of new and different compounds for use as antibiotic drugs. For example, vancomycin, first isolated

in the early 1950's from the bacteria *Amycolatopsis orientalis*, is a glycosylated nonrribosomal peptide originally used for the treatment against penicillin resistant *Staphylococcus aureus*, and daptomycin, originating from *Streptomyces roseosporus*, is a cyclic lipopeptide used to treat bacterial skin infections (Figure 1-1) (7, 8). Additional benefits from specialized metabolites also began to reveal themselves—compounds contained properties ranging from antiviral, antitumor, antiparasitics, antimalarials, and immunosuppressive, amongst others. Lovastatin, discovered in the 1970's and produced by the fungus *Aspergillus terreus*, was found to reduce a patient's cholesterol, while cyclosporine, discovered from the fungus *Tolypodcladium inflatum* serves as an immunosuppressant to aid in organ transplant operations (Figure 1-1) (9, 10). As such, specialized metabolites have made a huge impact on human health. Even when specialized metabolites do not make it directly through clinical trials to become drugs, many drug candidates are derived from specialized metabolite scaffolds (11, 12). Although there is a relatively high success rate for specialized metabolites to become drugs, pharmaceutical companies began to reduce the effort that goes into this area of research.

This decline in industrial discovery is due to certain challenges that face specialized metabolites research. First and foremost is the inherent difficulty of accessibility to new sources of specialized metabolites. While terrestrial based microbes may be easy to reach, beyond skirting the edge and shallow surfaces of the water, utilizing the entirety of the marine environment can be quite difficult for humans (13). It is also estimated that less than 1% of all species of microorganisms can actually be cultured, which makes performing studies in a controlled lab environment difficult. Seasonal and environmental variations can alter the production and detection of specialized metabolites from specialized metabolite sources. And not all sources of specialized metabolites are renewable, such as taxol and the pacific yew tree. While microbes can be regrown for studies, it is believed that thousands of medicinal plant species are threatened with extinction. Lastly, if the supply of specialized metabolites is not an issue, obtaining a pure sample of a

particular compound is not a trivial task. Extracts of microbial cultures always consist of a complex mixture of compounds and are sometimes produced in very small amounts (12, 14). Therefore, this thesis seeks to examine how recently developed mass spectrometry tools and methodologies applied to studying specialized metabolites have led to genome mining of unsequenced organisms, to the introduction of molecular families and gene cluster families, and how all of these can be tied together to increase the speed, efficiency, and scale of specialized metabolites research.

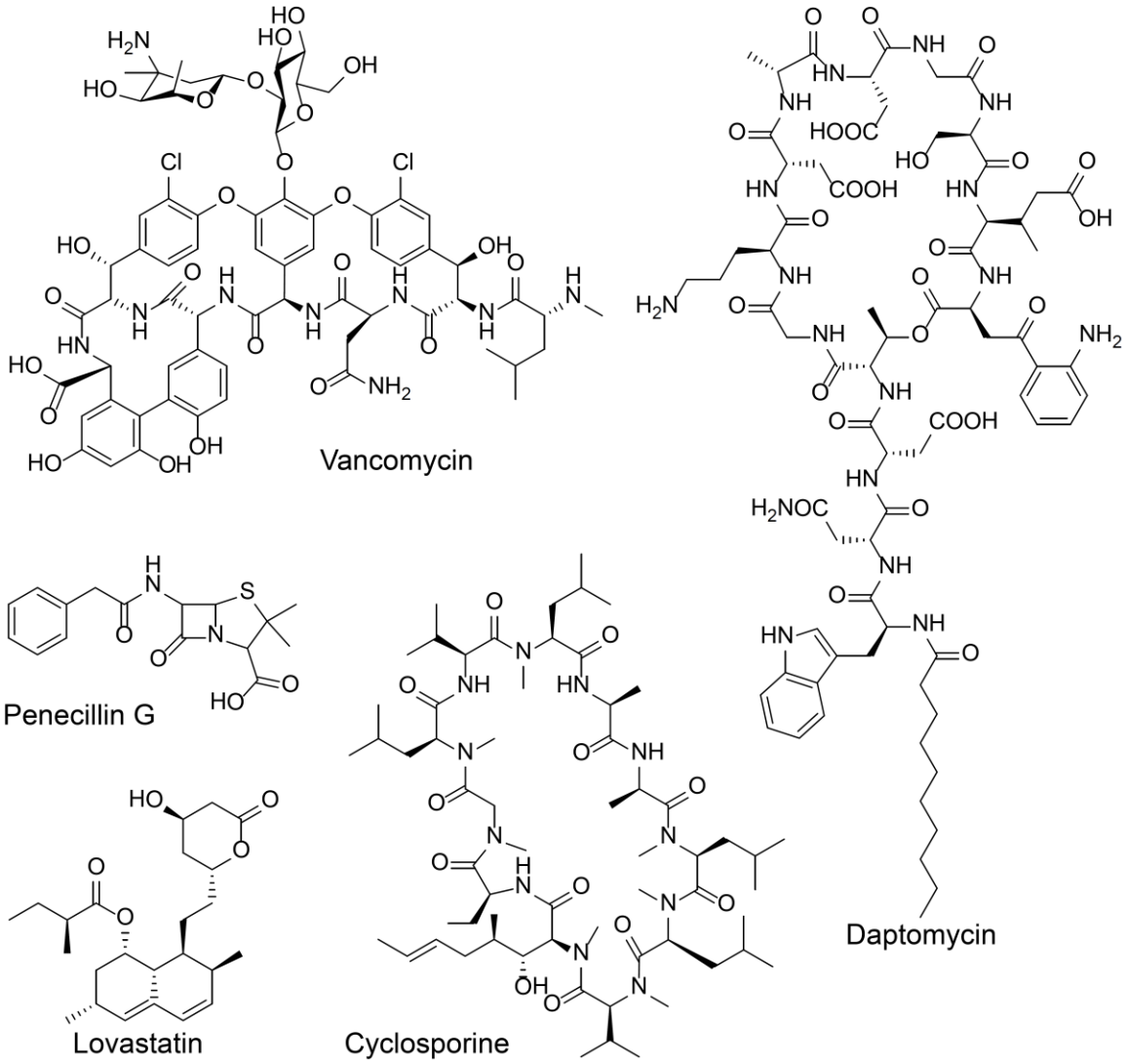


Figure 1-1. Examples of specialized metabolites used in aiding human health and disease. Antibiotics vancomycin, penicillin g, and daptomycin. Pharmaceuticals lovastatin and cyclosporine.

In order to aid in the discovery and characterization of specialized metabolites from microbial sources, we make use of mass spectrometry and introduce the concepts of molecular families and gene cluster families (explained in detail in Chapter 2). For these investigations, we used two types of mass spectrometers: matrix assisted laser desorption ionization imaging mass spectrometry (MALDI-IMS) and electrospray ionization mass spectrometry (ESI-MS), specifically by nanospray desorption electrospray ionization (nanoDESI). In general, all mass spectrometers

consist of three basic components: the ion source, analyzer, and detector (Figure 1-2). Mass spectrometry is a technique that generates ions from a sample of organic or inorganic compounds and then sorts and detects these ions based on their mass-to-charge ratios (*m/z*). Analytes can be ionized in a variety of fashions such as firing energetic electrons at a sample, by bombardment of energetic neutral atoms, or by dissolving sample in a solvent and spraying it out of a capillary or small nozzle. Upon entering the mass spectrometer, ions are separated and sorted based on the behavior of the ion's *m/z* within the presence of an electric or magnetic field, but can also be used in field free regions. Eventually these ions pass by or hit a surface that causes an induced charge to be detected whereupon this information is digitized and converted into a mass spectrum. With this information, molecular weight and chemical composition can be determined, as well as discerning hints of structural features from a molecule.

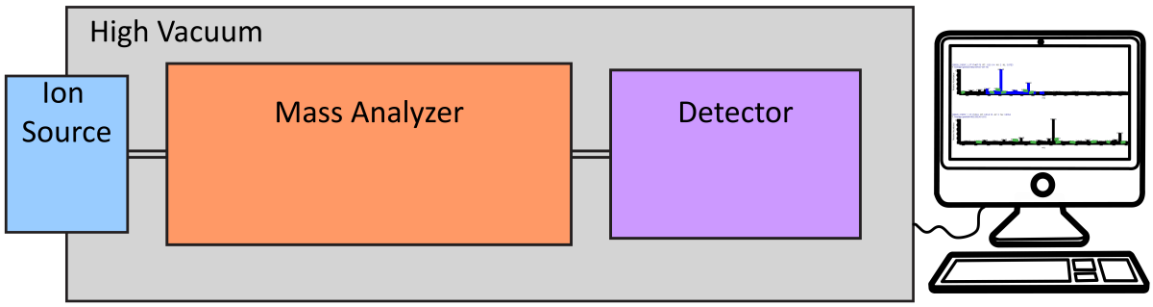


Figure 1-2. Schematic of a mass spectrometer. The ion source is responsible for generating molecular ions from a sample and can be done in a variety of fashions. The mass analyzer sorts the ions based on their *m/z*, while the detector registers the presence of an ion. This information is converted by software into a mass spectrum.

Here, MALDI-IMS was used to examine the specialized metabolites involved in the interaction of two different marine bacteria *Pseudoalteormonas* OT59 and 04M1A, which both showed inhibition against *Bacillus subtilis* 3610, respectively. Matrix Assisted Laser Desorption Ionization (MALDI) is an ionization method most widely used for the analysis of peptides, proteins, and many other biomolecules (15-17). To carry out this form of analysis, a sample is mixed and crystalized with an excess of matrix and desorbed by a laser. There are a variety of matrices, but

most are generally low molecular weight molecules that are weak organic acids with an absorbing chromophore in the infrared or ultraviolet regions. This mixture of sample and matrix is spotted onto a MALDI target plate and placed into the mass spectrometer where a laser is fired at the sample causing the matrix to vaporize and carry analyte with it (Figure 1-3). The purpose of the matrix is absorb the laser light and transfer this energy into the sample, and to act as a proton donor or acceptor, indirectly causing the analyte to vaporize and ionize.

The mass analyzers for MALDIs are typically time-of-flight (TOF) analyzers and come in two varieties, linear as well as reflectron analyzers. Detection in these instruments occurs when ions collide with a resistive material, which results in the release of secondary particles that intensify the original signal, a process known as secondary emission, and is recorded and digitized as a mass spectrum. The principles of both linear and reflectron analyzers are the same; a set of ions with different *m/z* are accelerated towards a detector all with the same amount of energy. The ions of differing *m/z* are dispersed in time as they fly down the TOF tube along a path of known length (Figure 1-3) (15, 16, 18). Because all the ions start with the same amount of kinetic energy, smaller ions that are less massive will travel along the TOF tube at a greater velocity, thus reaching the detector first, while larger ions that are more massive will travel along the TOF tube at a slower velocity reaching the detector last (Figure 1-3). In a linear analyzer, the ions are accelerated straight towards the detector, however, a reflectron detector increases the time ions require to reach a secondary detector by reflecting the ions back in the opposite direction they originally came from. This is done by placing a reflecting electric field at the end of the TOF tube. As ions enter this electric field, their kinetic energy is reduced to zero, at which point they are ejected in the opposite direction towards the reflectron detector (Figure 1-3). In essence, a reflectron analyzer lengthens the TOF tube by changing the direction of the ions, thus increasing the resolution of the instrument, albeit at the expense of sensitivity. A linear detector, on the other hand, has a decreased path length

for ions to fly down, which decreases the chances for collisions with residual gas, thus offering better sensitivity at the expense of resolution.

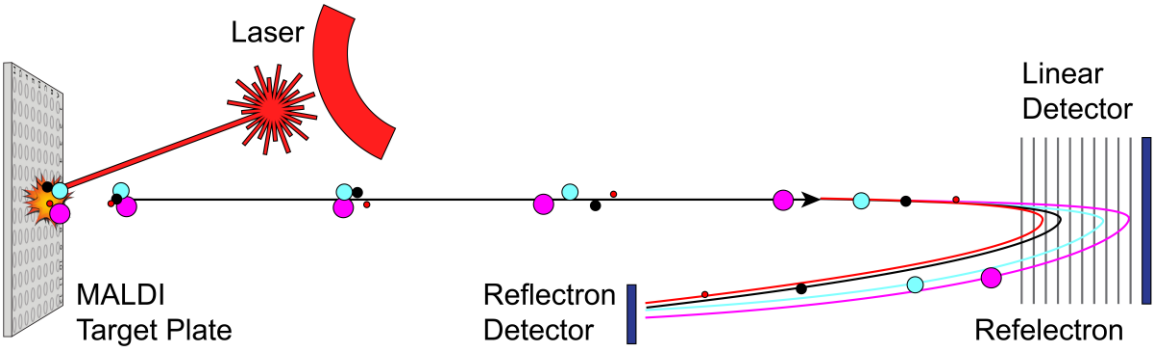


Figure 1-3. Schematic of MALDI-TOF-MS. After desorption by the laser, the ions are separated in time along their flight path and can be detected by either the linear or reflectron detector.

MALDI-MS has traditionally been used for the analysis of peptides, proteins, and large biomolecules (17). A technique known as imaging mass spectrometry (IMS) has been combined with MALDI to study proteins and peptides distributed along tissue sections (19, 20). While the molecular weight of a given ion does not inherently lead to its identification, further information is gained while examining the distribution of this ion in two dimensions, such as its localization within a sample (Figure 1-4). In our laboratory, we apply MALDI-IMS to study metabolic exchange or the use of specialized metabolites in an organism's interaction with its environment or with neighboring organisms, in single bacterial cultures or in microbial interactions (21, 22). When performing MALDI-IMS, the sample on the MALDI target plate is moved in the x-y plane while the laser is fired at discrete locations, or pixels, in a predefined raster area (Figure 1-4). This allows for ionization of analytes at every pixel. Mass spectra are recorded at each pixel and contain hundreds to thousands of individual ions, ultimately resulting in a two dimensional mass spectral profile. After data has been recorded throughout the entire raster area, a single ion can be selected for through specialized software, given a false color, and then overlaid onto an optical image of the sample. This shows the location as well as the intensity of the selected ion in relation to the sample (Figure 1-4). Our laboratory has applied this technology to a variety of microbes including

terrestrial, freshwater, and marine bacteria, as well as amoeba, fungi, and yeast for the purposes of studying metabolic exchange (21). In the present study, two *Pseudoalteromonas* (OT59 and 04M1A) both showed inhibition against *B. subtilis* 3610 when cocultured on an agar petri dish. Subjecting these interactions to MALDI-IMS allowed us to prioritize which molecules produced by the *Pseudoalteromonas* should be characterized, by revealing a set of ions with similar distributions residing at the zone of inhibition. This suggests that the ions specifically at the zone of inhibition are responsible for bioactivity against *B. subtilis*.

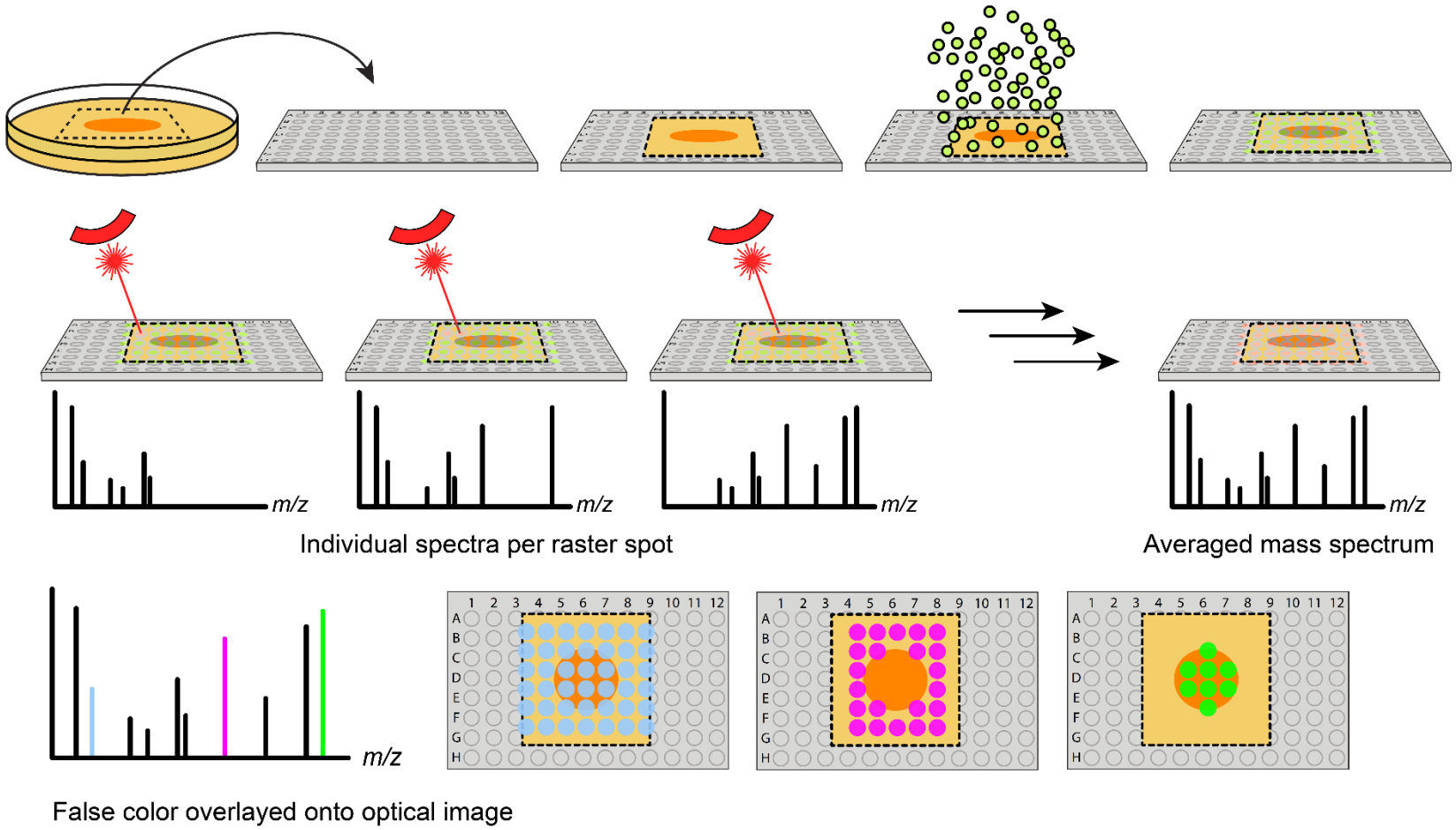


Figure 1-4. MALDI-IMS workflow. A sample grown on agar is transferred to a MALDI target plate, coated with matrix, dehydrated, and measured over a predefined raster region. Individual mass spectra are collected at each pixel and are then averaged together. Ions can are given a false color which is then overlaid on top of an optical image of the sample and target plate, showing the distribution of that particular ion.

However, due to the low resolution and mass accuracy of most MALDI-TOF instruments, orthogonal methods are required to further probe the molecules that inhibit the growth of *B. subtilis*. To do this, we must take advantage of another form of mass spectrometry. Electrospray ionization (ESI) is another technique for generating intact molecular ions where a sample is dissolved in the electrospray solvent (a mixture of water with volatile organic solvent and a small percentage of acid) and sprayed out of a fine capillary or needle that is held at high voltage (15, 23). The emerging liquid is charged and forced into a fine spray of droplets directed towards the inlet of the mass spectrometer. Generally, a heated inert gas is circulated near the inlet of the mass spectrometer to help desolvate the droplets. As the droplets desolvates and drifts towards the inlet of the mass spectrometer, the diameter of each droplet decreases, thus increasing the charge density on the droplet's surface until Coulombic repulsion causes the droplet to explode, creating daughter droplets that also continue to desolvate and undergo Coulombic explosions. The charge is then localized onto analytes within the sample, and with the assistance of a small percentage of acid, a molecular ion is created, and taken into the mass spectrometer. One advantage of ESI is the ability to analyze very large biomolecules, molecules and even entire viruses ranging from several to tens of megadaltons, due to the ability to observe ions with multiple charges (24, 25). Thus, larger molecules with multiple charges can be detected at lower *m/z*, ultimately extending the range with which molecules can be observed.

Here, we make use of nanoDESI. NanoDESI uses a syringe pump to deliver solvent through a fused silica capillary (the primary capillary) to which high voltage is applied (Figure 1-5) (26, 27). The primary capillary terminates just above the sample surface and creates a solvent bubble that is placed into contact with the sample. A second fused silica capillary (the secondary capillary) is positioned near the primary capillary such that a liquid bridge is formed between the primary and secondary capillary using the solvent bubble. The terminating end of the secondary capillary is positioned at the MS inlet (Figure 1-5). Analyte is desorbed into the solvent bubble and

pulled up the secondary capillary by a combinatorial effect of capillary action and the vacuum of the MS inlet, and is sprayed due to the high voltage applied at the primary capillary. Using nanoDESI, one can analyze molecules directly from surfaces without any prior sample preparation and has been used to analyze organic aerosols, microbial colonies directly from petri dishes, as well as tissue sections for imaging mass spectrometry (27-30).

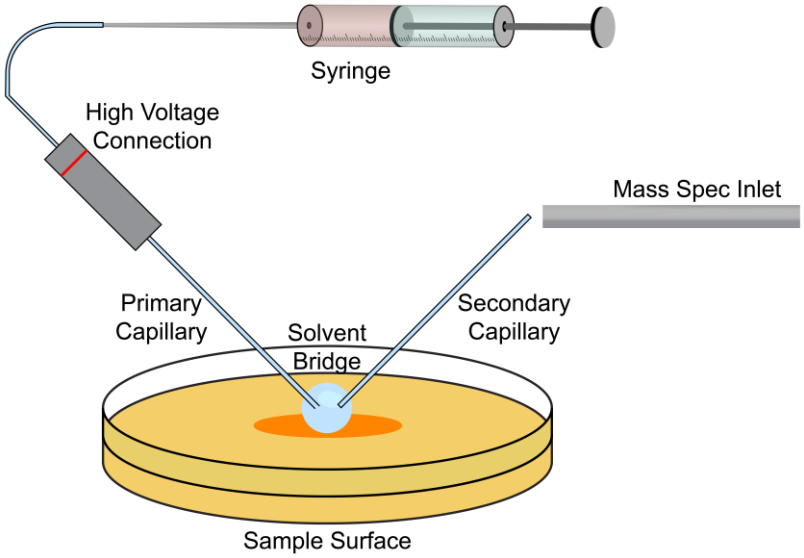


Figure 1-5. NanoDESI setup. ESI solvent is delivered via s syringe pump to form a liquid bridge between the primary and secondary capillary. Analyte is dissolved in the liquid bridge and delivered through the secondary capillary into the mass spectrometer.

The nanoDESI source utilized for experiments within this thesis are coupled with a Thermo Linear Trap Quadrupole Fourier Transform Ion Cyclotron Resonance mass spectrometer (LTQ-FTICR-MS) (15, 31, 32). This is a hybrid instrument consisting of a linear ion trap on the front end combined with an FTICR-MS. This setup allows for the use of both low resolution/low mass accuracy that is very quick and sensitive using the ion trap, with the slower and less sensitive FTICR that is capable of high resolution/high mass accuracy. For high resolution/high mass accuracy measurements, ions generated by the ESI source are introduced into the ICR cell that is held within a strong magnetic field. In the presence of a strong magnetic field, ions begin to orbit in a cyclotron motion, where the angular frequency of the ion is dependent only on its mass, charge, and the

strength of the magnetic field (Figure 1-6, A and B). The orbiting ions are detected by two detector plates and amplified to give a time domain free induction decay (31) (Figure 1-6, C). A Fourier transform is applied to the time domain signal giving a frequency. For less massive ions, the frequency of the orbit is greater, while more massive ions have a smaller frequency. The frequency is converted to a mass spectrum, and results in very high resolution, or the ability to distinguish the peaks in the mass spectrum from one another, allowing for accurate mass measurements (Figure 1-6, D). With sufficiently high mass accuracy, one can calculate a chemical formula for a given *m/z* or match an accurate mass to a compound reported in the literature, both of which aid in identifying a compound (15). Using nanoDESI in conjunction with FTICR-MS allowed us to very quickly to sample the surface of *Pseudoalteromonas* OT59 and 04M1A directly from agar petri dishes. The ions observed by MALDI-IMS were now measured with high mass accuracy and could be cross referenced with known molecules in the AntiMarin database. One of the ions was putatively identified and matched to the previously reported bromoalterochromide A/A' at a mass accuracy of -0.1 ppm error. While this is an extremely accurate match, it is possible that multiple compounds can have the exact same chemical formula (and therefore the same molecular weight and *m/z*) and still be structurally unique.

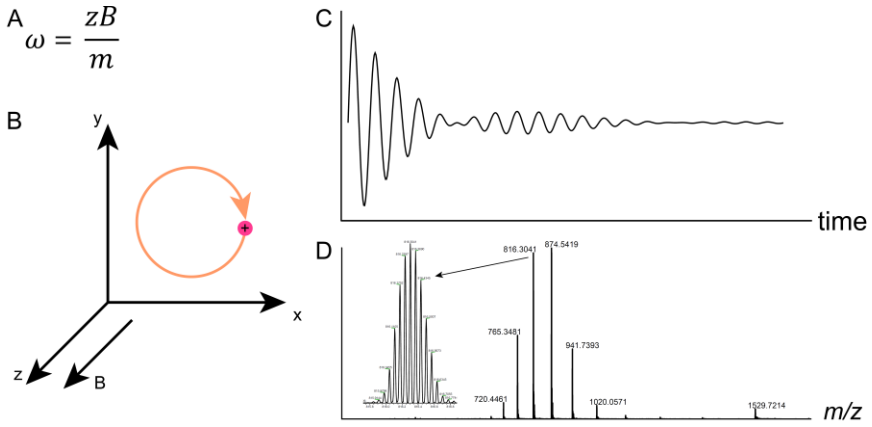


Figure 1-6. Principles of FTICR-MS. (A) Equation relating angular frequency to mass, charge, and strength of a magnetic field. (**B)** An ion in cyclotron motion. (**C)** Free induction decay. (**D)** Fourier transformed mass spectrum and isotopic distribution of a 12 kDa protein.

Because an *m/z* and molecular weight alone does not inherently identify a compound, we have made use of tandem mass spectrometry to help determine the structure of specialized metabolites. In tandem mass spectrometry (MS/MS, MS^2, MS^n), a mass selected ion (MS^1), also known as the precursor mass or precursor ion, is subject to a second level of mass spectrometric analysis (MS^2) (Figure 1-7, A). The second mass analysis is generally a measure of the mass-to-charge of the dissociated ions from the precursor mass; these are known as product ions. In essence, one is measuring the *m/z* of the smaller components and building blocks of the original molecule and this information can be used to partially and sometimes fully elucidate the structure of a molecule. While there are many methods to generate fragment ions, such as post source decay (PSD), electron transfer dissociation (ETD), infrared multiphoton dissociation (IRMPD), we primarily use collision induced dissociation (CID) (33-36).

Our LTQ-FTICR is capable of CID, where an ion beam is passed through a collision cell where a neutral gas such as helium, nitrogen, or argon, is introduced. The selected ion is accelerated by an electric potential into the collision cell and collides with molecules of neutral gas. The collision with neutral gas transfers kinetic energy to and subsequently increases the internal energy of the molecule. The increased internal energy dissociates through the bonds of the molecule causing bond breakage along any available fragmentation pathways (15, 33-36). In general, each individual molecule results in a unique MS/MS spectrum giving us a molecular fingerprint. This fingerprint can be used to identify molecules with the aid of metabolomics databases such as METLIN, MassBank, LIPID MAPS, as well as a variety of other databases (37-39). However, there are few databases specifically for specialized metabolites offering tandem MS data that are accessible to the public or are of sufficient size to encompass the molecules found within the range of organisms that produce them. Therefore, this analysis must be done manually. To assist in this analysis of specialized metabolites, a technique called MS/MS networking (or molecular MS/MS networking) has been developed in the Dorrestein Laboratory (Figure 1-7, C) (27).

Manual analysis of a single MS/MS spectrum is generally time intensive and can vary depending on the nature and complexity of the molecule. In a single microbial extract, there may be thousands to millions of MS/MS spectra, which makes manual analysis of a single data set impractical. Harnessing the unique molecular fingerprints from molecular fragmentation, a network-based method has been developed to visualize molecules and any of their related analogues (27). Molecular networking visualizes the similarity and differences between fragmentation spectra of all ions that are fragmented in a mass spectral data set. The benefit of visualizing and organizing data based on tandem MS similarities and differences is the ability to identify analogues and related compounds much more easily.

There are several steps to utilizing molecular networking to analyze MS/MS data. Pairwise comparisons are made between all MS^2 spectra to identify, at minimum, six matching fragment peaks, taking into account relative intensities of each peak as well as the *m/z* difference between two precursor ions; this results in a cosine score for each pair of spectra compared. Each spectrum is then connected to its top ten scoring matches (based on the cosine score). Based on a user defined cutoff, connections above the minimum cosine score are retained while connections below the minimum cosine score are discarded. The network is then processed to remove MS/MS spectra belonging to controls such as media or solvent blanks. Finally, this information is imported into the free software, Cytoscape, which allows for visualization of the entire molecular network. At the smallest level, Cytoscape visualizes this data as nodes; small circles that represent the parent mass that was fragmented as well as its underlying MS/MS spectra (Figure 1-7, A). These nodes are then connected to their related ions of differing precursor mass (but still similar fragmentation spectra) by edges (Figure 1-7, B). For the two strains of *Pseudoalteromonas*, nanoDESI in conjunction with molecular networking revealed that the multiple molecules observed at the zone of inhibition by MALDI-IMS all cluster together. This suggests the multiple molecules involved with the inhibition of *B. subtilis* are a family of molecules related to bromoalterochromide A/A'. The family of

bromoalterochromides was then connected to their biosynthetic gene clusters, not with genomic information from OT59 or 04M1A, but with previously published genomic information from *Pseudoalteromonas piscicida* JCM 20779^{T} (40). Upon further investigation, it was found that the same biosynthetic gene cluster that produces the family of molecules responsible for the inhibition of *B. subtilis* is also present in OT59 and 04M1A. We therefore proposed the idea of connecting molecular families to their gene cluster families using molecular networking and publicly available genome sequences to increase the speed, scale, and efficiency of how specialized metabolites are studied. In this way, molecular networking can be used for visualization of large tandem MS data sets, dereplication, as well as prioritizing the analysis of specific specialized metabolites to greatly increase the speed and efficiency of the analysis of the parvome (41).

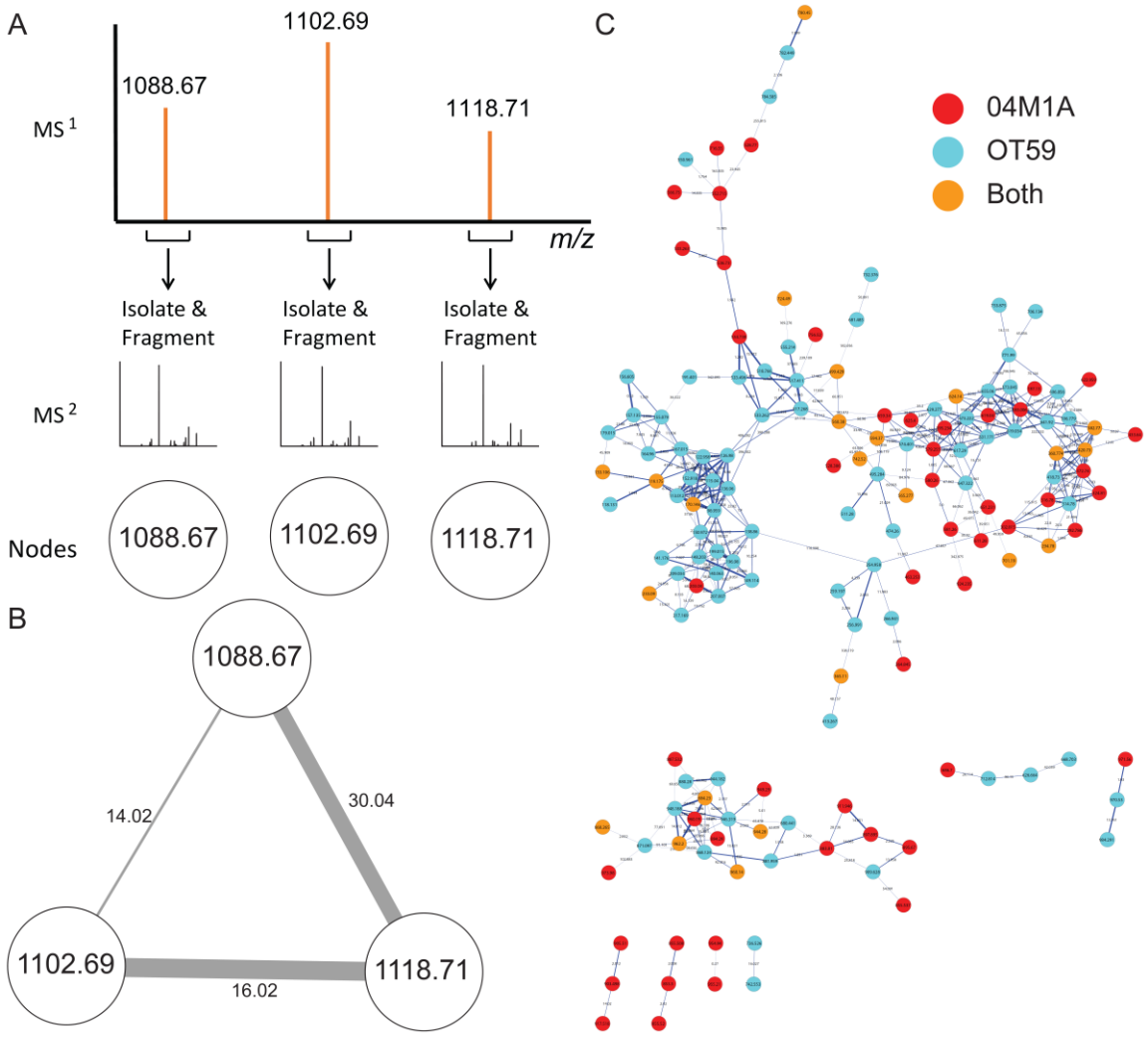


Figure 1-7. Concepts of molecular networking. (A) All ions above an intensity threshold are isolated and fragmented in the mass spectrometer. Upon molecular networking, the precursor mass and unique MS^2 fragmentation pattern are represented as nodes. (**B)** Nodes are connected to one another by edges, which signify similar unique, but related, MS^2 fragmentation patterns. (**C)** Example of a typical molecular network. In this case, of the metabolites produced by two strains of *Pseudoalteromonas*, OT59 and 04M1A.

To summarize, the goals of this thesis are to illustrate the applications of the mass spectrometry methodologies—MALDI-IMS, ESI-MS, and molecular networking—in the study of microbial specialized metabolites, but also to systems outside of specialized metabolites research. Chapter 2 describes the concept of molecular families and gene cluster families and how the two are connected to one another using molecular networking. This technique is used on a large scale

to map the nonribosomal peptide molecular families of more than sixty organisms as well as to discover the biosynthetic machinery that produces the family of molecules known as the bromoalterochromides from the genus of marine bacteria *Pseudoalteromonas*. Lastly, Chapter 3 briefly describes the future applications of these mass spectrometry methodologies towards two different projects. The first is an expansion on the idea of molecular families paired to gene cluster families, but specifically towards hundreds of *Pseudoalteromonas*. The goals of this study are to examine the metabolic similarities and differences of *Pseudoalteromonas* collected from oceanic waters around the world but also from dairy products. The second project is the application of MALDI-IMS to the study of grazing activity of predators on photosynthetic organisms. Here we examine the molecular signatures involved between the amoeba, HGG1, grazing on the cyanobacteria *Synechococcus elongatus* PCC 7942 in the attempt to identify and characterize molecules that are specific to amoeba grazing. Identification of such signatures can then be used as molecular markers to signify the contamination of large scale algae biofuel ponds, allowing for proper counter measures to be taken that would prevent the pond from crashing and losing thousands of dollars in fuel crops.

1.1 REFERENCES

1. Weng JK & Noel JP (2012) The Remarkable Pliability and Promiscuity of Specialized Metabolism. *Cold Spring Harbor symposia on quantitative biology*.

2. Dewick PM (2009) *Medicinal Natural Products: A Biosynthetic Approach* (Wiley).

3. Davies J & Ryan KS (2012) Introducing the parvome: bioactive compounds in the microbial world. *ACS Chem Biol* 7(2):252-259.

4. Davies J (2009) Darwin and microbiomes. *EMBO reports* 10(8):805.

5. Fleming A (1929) On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. *British Journal of Experimental Pathology* 10:226-236.

6. Houbraken J, Frisvad JC, & Samson RA (2011) Fleming's penicillin producing strain is not Penicillium chrysogenum but P. rubens. *IMA fungus* 2(1):87-95.

7. Levine DP (2006) Vancomycin: a history. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 42 Suppl 1:S5-12.

8. Eliopoulos GM, Willey S, Reiszner E, Spitzer PG, Caputo G, & Moellering RC, Jr. (1986) In vitro and in vivo activity of LY 146032, a new cyclic lipopeptide antibiotic. *Antimicrobial agents and chemotherapy* 30(4):532-535.

9. Endo A, Kuroda M, & Tanzawa K (1976) Competitive inhibition of 3-hydroxy-3-methylglutaryl coenzyme a reductase by ML-236A and ML-236B fungal metabolites, having hypocholesterolemic activity. *FEBS Letters* 72(2):323-326.

10. Borel JF (2002) History of the discovery of cyclosporin and of its early pharmacological development. *Wiener klinische Wochenschrift* 114(12):433-437.

11. Butler MS (2008) Natural products to drugs: natural product-derived compounds in clinical trials. *Natural Product Reports* 25(3):475-516.

12. Cragg GM & Newman DJ (2013) Natural products: A continuing source of novel drug leads. *Biochim Biophys Acta* 1830(6):3670-3695.

13. Gerwick WH & Moore BS (2012) Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. *Chem Biol* 19(1):85-98.

14. Li JW & Vederas JC (2009) Drug discovery and natural products: end of an era or an endless frontier? *Science* 325(5937):161-165.

15. Gross JH (2004) *Mass Spectrometry: A Textbook* (Springer, Germany) 1 Ed.

16. Lewis JK, Wei J, & Siuzdak G (2006) Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry in Peptide and Protein Analysis. *Encyclopedia of Analytical Chemistry*, (John Wiley & Sons, Ltd).

17. Karas M & Hillenkamp F (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* 60(20):2299-2301.

18. Stephens WE (1946) Proceedings of the American Physical Society. *Physical Review* 69(11-12):674-674.

19. Cornett DS, Reyzer ML, Chaurand P, & Caprioli RM (2007) MALDI imaging mass spectrometry: molecular snapshots of biochemical systems. *Nat Methods* 4(10):828-833.

20. Seeley EH & Caprioli RM (2008) Molecular imaging of proteins in tissues by mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* 105(47):18126-18131.

21. Yang JY, Phelan VV, Simkovsky R, Watrous JD, Trial RM, Fleming TC, Wenter R, Moore BS, Golden SS, Pogliano K, & Dorrestein PC (2012) Primer on agar-based microbial imaging mass spectrometry. *J Bacteriol* 194(22):6023-6028.

22. Yang YL, Xu Y, Straight P, & Dorrestein PC (2009) Translating metabolic exchange with imaging mass spectrometry. *Nat Chem Biol* 5(12):885-887.

23. Fenn JB, Mann M, Meng CK, Wong SF, & Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246(4926):64-71.

24. Fuerstenau SD & Benner WH (1995) Molecular weight determination of megadalton DNA electrospray ions using charge detection time-of-flight mass spectrometry. *Rapid communications in mass spectrometry : RCM* 9(15):1528-1538.

25. Fuerstenau SD, Benner WH, Thomas JJ, Brugidou C, Bothner B, & Siuzdak G (2001) Mass Spectrometry of an Intact Virus The authors gratefully acknowledge Jennifer Boydston for her helpful comments and suggestions. G.S. is grateful for support from the NIH (GM55775). The work at LBL was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program, U.S. Department of Energy under contract number DE-AC03-76SF00098. *Angew Chem Int Ed Engl* 40(6):9822.

26. Roach PJ, Laskin J, & Laskin A (2010) Nanospray desorption electrospray ionization: an ambient method for liquid-extraction surface sampling in mass spectrometry. *The Analyst* 135(9):2233-2236.

27. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, van der Voort M, Pogliano K, Gross H, Raaijmakers JM, Moore BS, Laskin J, Bandeira N, & Dorrestein PC (2012) Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences of the United States of America* 109(26):E1743-1752.

28. Roach PJ, Laskin J, & Laskin A (2010) Molecular characterization of organic aerosols using nanospray-desorption/electrospray ionization-mass spectrometry. *Analytical chemistry* 82(19):7979-7986.

29. Laskin J, Heath BS, Roach PJ, Cazares L, & Semmes OJ (2012) Tissue imaging using nanospray desorption electrospray ionization mass spectrometry. *Analytical chemistry* 84(1):141-148.

30. Lanekoff I, Thomas M, Carson JP, Smith JN, Timchalk C, & Laskin J (2013) Imaging nicotine in rat brain tissue by use of nanospray desorption electrospray ionization mass spectrometry. *Analytical chemistry* 85(2):882-889.

31. Marshall AG & Grosshans PB (1991) Fourier transform ion cyclotron resonance mass spectrometry: the teenage years. *Anal Chem* 63(4):215A-229A.

32. Marshall AG, Hendrickson CL, & Jackson GS (1998) Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrometry Reviews* 17(1):1-35.

33. Sleno L & Volmer DA (2004) Ion activation methods for tandem mass spectrometry. *Journal of Mass Spectrometry* 39(10):1091-1112.

34. Little DP, Speir JP, Senko MW, O'Connor PB, & McLafferty FW (1994) Infrared multiphoton dissociation of large multiply charged ions for biomolecule sequencing. *Anal Chem* 66(18):2809-2815.

35. McLafferty FW, Kelleher NL, Begley TP, Fridriksson EK, Zubarev RA, & Horn DM (1998) Two-dimensional mass spectrometry of biomolecules at the subfemtomole level. *Curr Opin Chem Biol* 2(5):571-578.

36. Haselmann KF, Budnik BA, Olsen JV, Nielsen ML, Reis CA, Clausen H, Johnsen AH, & Zubarev RA (2001) Advantages of external accumulation for electron capture dissociation in Fourier transform mass spectrometry. *Anal Chem* 73(13):2998-3005.

37. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, & Siuzdak G (2005) METLIN: a metabolite mass spectral database. *Ther Drug Monit* 27(6):747-751.

38. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, & Nishioka T (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* 45(7):703-714.

39. Fahy E, Subramaniam S, Murphy RC, Nishijima M, Raetz CR, Shimizu T, Spener F, van Meer G, Wakelam MJ, & Dennis EA (2009) Update of the LIPID MAPS comprehensive classification system for lipids. *J Lipid Res* 50 Suppl:S9-14.

40. Xie BB, Shu YL, Qin QL, Rong JC, Zhang XY, Chen XL, Shi M, He HL, Zhou BC, & Zhang YZ (2012) Genome sequences of type strains of seven species of the marine bacterium Pseudoalteromonas. *Journal of bacteriology* 194(10):2746-2747.

41. Corley DG & Durley RC (1994) Strategies for Database Dereplication of Natural Products. *J Nat Prod* 57(11):1484-1490.

Chapter 2

MS/MS networking guided analysis of molecule and gene cluster families

2.1 Abstract

The ability to correlate the production of specialized metabolites to the genetic capacity of the organism that produces such molecules has become an invaluable tool in aiding the discovery of biotechnologically applicable molecules. Here we accomplish this task by matching molecular families with gene cluster families, making these correlations to 60 microbes at once, instead of connecting one molecule-one organism at a time as it is traditionally done. We can correlate these families through the use of nanoDESI MS/MS, an ambient pressure mass spectrometry technique, in conjunction with MS/MS networking and peptidogenomics. We matched the molecular families of peptide natural products produced by 42 bacilli and 18 pseudomonads through the generation of amino acid sequence tags from MS/MS data of specific clusters found in the MS/MS network. These sequence tags were then linked to biosynthetic gene clusters in publicly accessible genomes, providing us with the ability to link particular molecules with the genes that produced them. As an example of its utility, this approach was applied to two unsequenced *Pseudoalteromonas* species leading to the discovery of the gene cluster for a molecular family, the bromoalterochromides, in the previously sequenced strain *Pseudoalteromonas piscicida* JCM 20779^{T}. The approach itself is not limited to 60 related strains as spectral networking can be readily adopted to look at molecular family-gene cluster families of hundreds or more diverse organisms in one single MS/MS network.

2.2 Keywords

MS/MS molecular networking, molecular families, gene cluster families, nanoDESI

2.3 Significance

The paper introduces the concepts of molecular families (MFs) and gene cluster families (GCFs). We define MFs as structurally related molecules based on their mass spectral fragmentation patterns, while GCFs are biosynthetic gene clusters that show similar gene cluster

organization with a high degree of sequence similarity. We use MS/MS networking as a tool to map the molecular network of more than sixty organisms, most of which are unsequenced, and locate their nonribosomal peptide MFs. These MFs from unsequenced organisms are then connected to GCFs to publicly available genome sequences of closely related organisms.

2.4 Introduction

Tens of thousands of sequenced microbial genomes or rough drafts of genomes are available at this time, and this number is predicted to grow into the millions over the next decades. This wealth of sequence data has the potential to be utilized for the discovery of small bioactive molecules through genome mining (1-6). Genome mining is a process in which small molecules are discovered by predicting what compound will be genetically encoded based on the sequences of biosynthetic gene clusters. However, the process of mining genetically encoded small molecules is not keeping pace with the rate by which genome sequences are being obtained. In general, genome mining is still done one gene cluster at a time and requires many person-years of effort to annotate a single molecule. The time and significant expertise that current genome mining requires also makes genome mining very expensive. In light of this, alternative approaches towards genome mining and annotating specialized metabolites must be developed that not only take advantage of the sequenced resources available and make it efficient to perform genome mining on a more global scale, but also enable the molecular analysis of unsequenced organisms. Such methods will then significantly reduce the cost of genome mining, by increasing the speed with which molecules are connected to candidate genes and by using resources already available. Here we put forward such a mass spectrometry-based strategy that enables the genome mining of small molecule families from unsequenced organisms. This strategy uses partial *de novo* structures inferred from nanoDESI-based MS/MS networking to connect to structures predicted from genomic resources available in sequence repositories (2, 7). The MS/MS network-based genome mining approach

presented in this paper takes a more global approach than is currently the norm. This paper builds on many advances that have happened over the past decade. First, an enormous amount of microbial sequencing data has been deposited in public databases and is waiting to be mined (8-10). Second, our understanding of biosynthetic pathways and the function of specific enzymes found in gene clusters—especially for complex peptides made by nonribosomal peptide synthetases (NRPSs)—has dramatically increased (11-19). Finally, the last decade has seen very significant advances in mass spectrometry with respect to ion sources and the sensitivity of the instruments themselves (20-27). Ambient ionization methods, in combination with significant improvements in sensitivity and mass accuracy of mass spectrometry instrumentation, now enable the detection of intact molecules directly from surfaces (7, 28-40). Using the ambient method nanoDESI, the molecular characterization of microbial colonies directly from agar surfaces without any prior sample preparation has become possible (7).

In this study, nanoDESI is used to observe detectable metabolites, where we focused on nonribosomally synthesized peptides, from unsequenced bacterial strains, as well as from representative sequenced *Pseudomonas* and *Bacillus* strains (Table 2-1). These metabolites were subsequently subjected to MS/MS networking to first, generate a molecular network representing the detectable metabolites that are then related to one another based on similarity of their fragmentation spectra which is dictated by their molecular structure (7). MS/MS networking was then second, used to generate *de novo* peptide sequences from nonribosomally synthesized peptides as well as their respective molecular families (MFs). MFs are defined in this paper as a series of related molecules based on their fragmentation behavior that translates to structural similarity. Mass spectrometry-based genome mining utilizing genomes in sequences repositories from related organisms were then used to connect these MFs to their gene cluster families (GCFs) (2, 7). GCFs are defined as gene clusters that exhibit similar gene cluster organization with a high degree of sequence similarity and where the A-domain specificity is minimally altered. We targeted the well-

studied family of molecules, the nonribosomal peptide (NRP) systems, with our MS/MS network-based genome mining strategy to demonstrate that mass spectrometric signatures can be used to group families of molecules from multiple organisms. Grouping these molecular families can in turn can be used to find candidate biosynthetic gene clusters found in sequence repositories that could be responsible for the biosynthesis of such specialized metabolites at a more global scale. For a detailed description on how biology creates peptides without a ribosome, one should consult several detailed reviews from the literature (18, 19). In short, NRPS-derived peptides are produced by protein machineries that build the peptides from a collection of more than 500 different amino acid building blocks. Genome based predictions of peptide cores created by NRPS assemblies have now been automated and integrated into informatic tools where a sequence is uploaded and predictions generated (41-45). In our opinion, NRPS-derived molecules are the most readily achievable goals with respect to genome mining due to the availability of extensive biosynthetic studies in the last decades. It is, however, expected that through creative adaptation of the approach or related approaches, additional small molecule classes such as isoprenoids, polyketides, oligosaccharides, glycolipids, lipids, and other natural products can be mined as well.

To accomplish genome mining of unsequenced organisms, we utilized sequencing information from publicly available databases as well as the predictive power of NRPS A-domains, encoded by NRPS genes in the sequenced genomes of related taxa, to link the MS/MS peptide signatures of compounds produced to candidate biosynthetic gene clusters. This allowed us to correlate MFs (e.g. surfactin/lichenysin, and viscosin/WLIP/Massetolide) from 60 strains of Bacilli and Pseudomonads to their respective GCFs (46-49). We then applied this same methodology to assign the gene cluster of the membrane-disrupting antimicrobial agent, the bromoalterochromides (50-52), from a Panamanian octocoral-associated *Pseudoalteromonas* species.

Table 2-1. *Pseudomonas* and *Bacilli* strains subjected to nanoDESI based MS/MS networking.

Sequenced	Name	Source
Pseudomonads		
No	*Pseudomonas aurantiaca*	plant/soil
Yes	*Pseudomonas putida* RW10S2	plant/soil
No	*Pseudomonas tolaasii* CH36	plant/soil
Partial	*Pseudomonas putida* RW10S1	plant/soil
No	*Pseudomonas fluorescens* BW11P2	plant/soil
No	*Pseudomonas putida* BW11M1	plant/soil
No	*Pseudomonas fluorescens* RW9S1	plant/soil
Yes	*Pseudomonas aeruginosa* PAO1	human
Yes	*Pseudomonas aeruginosa* PA14	human
No	*Pseudomonas chlororaphis* 200B1	plant/soil
No	*Pseudomonas moraviensis* ES97	plant/soil
No	*Pseudomonas moraviensis* ES16	plant/soil
No	*Pseudomonas* spp ES11	plant/soil
No	*Pseudomonas para* ES60	plant/soil
Partial	*Pseudomonas fluorescens* C52	plant/soil
No	*Pseudomonas fluorescens* Tn5	plant/soil
No	*Pseudomonas putida*	mosquito
No	*Pseudomonas* HCL17	mosquito
Bacilli		
No	*B. megaterium* QMB1551	soil
No	*B. megaterium* ES-190	soil
Yes	*B. amyloliquefaciens* FZB42	soil
No	*B. thuringiensis* ES-20	soil
No	*B. firmus* ES-118	soil
No	*B. firmus* ES-115	soil
No	*B. subtilis* ES-73	soil
Yes	*B. subtilis* 3610	soil
No	*B. subtilis* KP1302	soil
Yes	*B. subtilis* PY79	soil
No	*B. licheniformis* ES-114	soil
No	*B. licheniformis* ES-44	soil
No	*B. marisflavi* ES-120	soil
No	*B. pumilus* ES-76	soil
No	*B. vallismortis* ES113	soil
No	*B. licheniformis* ES221	soil

Table 2-1, Continued		
Sequenced	**Name**	**Source**
No	*B. sp. (fusiformis)* ES222	soil
Yes	*B. amyloliquefaciens* FZB42 ES223	soil
No	*B. weihenstephanensis* ES332	soil
No	*B. megaterium* ES333	soil
No	*Lysinibacillus sphaericus* ES335	soil
No	*B. clausii* ES336	soil
No	*B. circulans* ES337	soil
No	*B. firmus* ES341	soil
No	*B. lentus* ES342	soil
No	*B. coagulans* ES343	soil
No	*Aeribacillus pallidus* ES345	soil
No	*B. subtilis subsp. subtilis* ES382	soil
No	*B. subtilis subsp. subtilis* ES386	soil
No	*B. subtilis subsp. subtilis* ES387	soil
No	*B. subtilis subsp. spizizenii* ES392	soil
No	*B. amyloliquefaciens* ZK4633	soil
No	*B. pumilus* ZK4636	soil
No	*B. vallismortis* ZK4678	soil
No	*B. cereus* GP24	soil
No	*B. thuringiensis* GP25	soil
No	*B. mycoides* LCT22	soil
No	*B. cereus* LCT5	soil
No	*B. cereus* LS18	soil
No	*B. megaterium* LS27B	soil
No	*B. megaterium* LS28	soil
No	*B. thuringiensis* LS29	soil

2.5 Results and Discussion

2.5.1 MS/MS network-guided genome mining

Matching of molecular families with gene cluster families of unsequenced microbes through association with sequenced genomes was accomplished via a four-step process (outlined in Figure 2-1). In the first step, fragmentation data for the molecules produced by these microbes was obtained for analysis by molecular MS/MS networking, effectively creating a searchable molecular network for these organisms. For this purpose we chose nanoDESI mass spectrometry as the ionization method. NanoDESI, through a real-time liquid extraction, enables ionization of molecules directly from colonies grown on agar surfaces in Petri-dishes without any sample preparation, but other mass spectrometry techniques such as LC-MS/MS or direct infusion MS/MS could also have been utilized (7). Because our nanoDESI is interfaced with an ion trap, it was possible to directly fragment all the ions that were detected to obtain MS/MS spectra. We subjected 60 different strains of bacteria to nanoDESI analysis: there were 42 bacilli and 18 pseudomonads, and their resulting MS/MS spectra were networked and visualized with Cytoscape (Figure 2-1A, step 2, and Figure 2-1B) (7, 53-55). Such organization into networks enables the relationships between spectrally identical and related molecules to be mapped based on the spectral similarity of their MS/MS signatures. An MS/MS cluster where many nodes are connected by edges indicates that many related molecules were observed, while an MS/MS cluster with few nodes may be a unique set of molecules with few alternative forms, which result in unique spectra. Furthermore, MS/MS networking enables the visualization of groups possessing unique spectral signatures that indicate the molecules are distinct from the other molecules in a given data set.

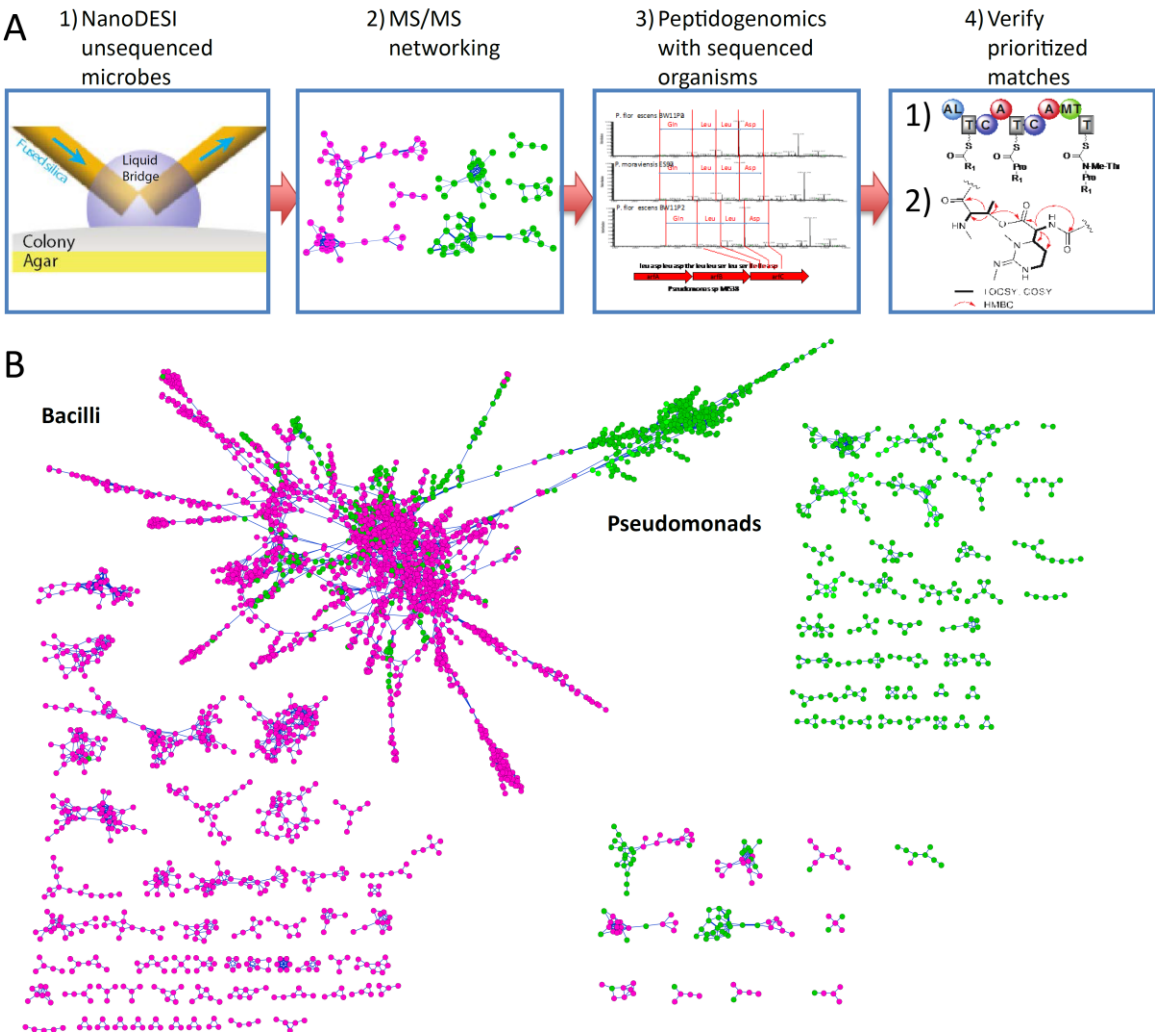


Figure 2-1. Process of MS/MS networking-guided genome mining of nonribosomal peptides produced by unsequenced organisms and the molecular network generated from 42 bacillus and 18 pseudomonad strains. (**A**) Step 1: NanoDESI MS on live microbial colonies to determine molecular weight and obtain MS/MS fragmentation data. Step 2: Generation of molecular networks and visualization using Cytoscape. Step 3: Peptide sequence tag generation from raw spectra of MS/MS clusters and prediction of NRPS gene clusters (antiSMASH) from genomic data available in public databases. Step 4-1: Biosynthetic gene cluster analysis to verify candidate molecules. Step 4-2: If the putative matches are of high priority, proceed with full structure elucidation from MS^n and NMR data. (**B**) Molecular network of 42 bacilli (pink) and 18 pseudomonad (green) strains.

Both the pseudomonad and bacilli nanoDESI MS/MS data were combined into one MS/MS network in order to create a searchable molecular network. It was anticipated that very few of the MS/MS signals that come from nonribosomal peptides would overlap between the two organisms;

there are no nonribosomal peptides that have been described in the literature that are found in both bacilli and pseudomonads and no NRPS gene clusters between these genera are related to one another (46-49). The merging of the data also enables the removal of overlapping signals that are not of interest, including any signals derived from the growth medium, although individual networks for the bacilli and pseudomonads, respectively, could have been created. Combining the data from all organisms assists in the peptidogenomics-based genome mining, since only one MS/MS node needs to be matched to its corresponding genome. This genome can then be related to the surrounding nodes in the MS/MS cluster so that not every MS/MS spectrum has to be individually correlated to candidate gene clusters, even when they originate from different organisms. This effort required to correctly correlate a GCF-MF pair is additive and only required once. Any newer molecules that are added to this network and that cluster within a particular MF, can then have the previously linked GCF-MF pair related to it. Finally, the vast majority of nonribosomal peptides isolated from these genera contain proteinogenic amino acids, thus simplifying the peptidogenomic analysis. Although there are non-peptidic molecules that are observed in the MS/MS network, such as the rhamnolipids and quinolones (56) (Figure 2-2), the goal of the analysis of these 60 strains was to provide a proof-of-principle to correlate nonribosomal peptides to their candidate gene clusters. While we only used 42 bacili and 18 pseudomonads, this technique can be scaled to hundreds or even thousands or even tens of thousands of organisms, sequenced or unsequenced, and still require only one MS/MS node to be matched to its corresponding genome. Combining the MS/MS data from the bacilli and the pseudomonads resulted in about 22% (972/4311 nodes) overlap in signals. Likely sources for these common molecules are primary metabolism, the nanoDESI solvent, the growth medium, and molecules that fragment poorly. The majority of spectra (78%) are unique to either the bacilli or the pseudomonads (Figure 2-1B). There are 121 MS/MS clusters that contain three or more nodes of unique fragmentation patterns; these MS/MS clusters visualize individual MFs.

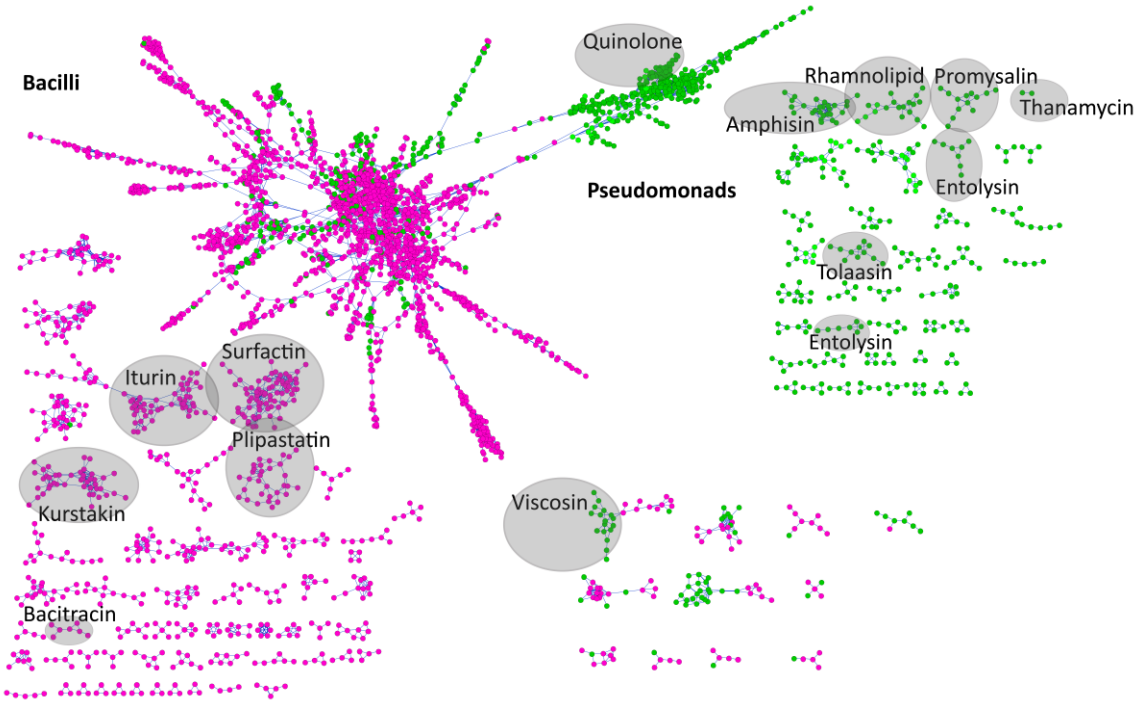


Figure 2-2. Identification of 8 GCF-MF pairs through the combination of peptidogenomic analysis with publicly available bacilli genomes and identification of 4 MFs. Identification of 4 GCF-MF pairs from the Bacilli: iturin, surfactin, kurstakin, bacitracin, and 4 pairs from the Pseudomonads: viscosin, thanamycin, entolysin, amphisin. Identification of 1 MF from the Bacilli: plipastatin and 4 MFs from Pseudomonads: promysalin, rhamnolipid, quinolone, and tolaasin.

Once the MS/MS network had been generated in step 2, we examined the raw MS/MS data looking for mass shifts between adjacent ions corresponding to the mass of an amino acid thereby creating a sequence tag that would indicate that a particular MS/MS cluster is peptidic in nature (Table 2-2). The peptidic clusters were then subjected to peptidogenomic analysis (Figure 2-1, step 3). We limited this search space to only proteinogenic amino acids (with and without an N-methyl group) because manual annotation of spectra with the more than 500 possible unique amino acids that can be incorporated into nonribosomally produced peptides, is a nearly impossible task, especially with low resolution MS/MS data. In the future, specific algorithms can be developed that will overcome this limitation, especially in conjunction with high-resolution MS/MS data. Instead of carrying out peptidogenomics analysis on a single organism as previously described, the peptide

backbones of all of the NRPS gene clusters from all available genome sequences of bacilli and pseudomonads in the public databases was predicted using a batch form of antiSMASH as well as curation using other A-domain prediction tools (4, 16, 43, 44, 57) (Table 2-2 and 2-3). By combining amino acid MS/MS signatures with the predicted amino acid specificity of NRPS A-domains, we obtained candidate matches of MS/MS signatures to particular GCFs. At this stage, the iterative process of examining and matching MS data to the gene clusters, as is done in peptidogenomics analysis, was carried out because this information is needed to correlate MFs and GCFs and can then be related to surrounding nodes within the MS/MS cluster (2). To improve our confidence in the peptidogenomics analysis, we looked for several correlations: 1) Can we find additional amino acids that correlate to the A-domain specificity predictions that were missed when generating the initial sequence tag from the MS/MS data? 2) Does the biosynthetic gene cluster contain tailoring domains in the NRPS or biosynthetic enzymes to make non-proteinogenic amino acids, and can those amino acids be found in the MS/MS data? 3) Did we observe patterns of mass shifts in the parent ions such as +/- 14 Da indicative of different set of amino acid substitutions (e.g. Gly vs Ala), methylations, or different fatty acid chain lengths that are common to nonribosomal peptides? 4) Does the size of the molecule match up to the size of the gene cluster (e.g. an NRPS with 20 A-domains is unlikely to encode for a molecule that is ~1000 Da)? 5) Does the biosynthetic pathway match the MF structural prediction? If these correlations all agree then it is possible to state that a potential GCF and MF match has been found. When the molecule or gene cluster is very important, based on biological prioritization, the GCF-MF correlation will need to be confirmed via other means as described below. Below are more detailed examples of how such GCF-MF correlations were obtained for known compounds from bacilli and pseudomonads.

Table 2-2. Manual curation of amino acid loading predicted by antiSMASH and their amino acid mass shifts. Non-standard amino acids predicted are: ornithine (orn), D-2-hydroxyisovalerate (hyv-d), 2,4-diaminobutyric acid (dab), 4-Hydroxyphenylpyruvic acid (4-hppa).

Mass Shift	**Sequence Tag**	**Organism**	**Protein Homolog Accession Number**	**Predicted Amino Acid Sequence**
Pseudomonads				
87-113-87-113	ser_leu/ile_ser_leu/ile	*Pseudomonas fluorescens* strain SS101 clone 2	EU199081.2	thr_ile_leu_ser_leu_ser_ile
		Pseudomonas synxantha BG33R ctg1124295418239	NZ_AHPP01000001.1	thr_ile_leu_ser_leu_ser_ile
		Pseudomonas fluorescens SBW25	NC_012660.1	lys_orn_ser_thr_val_leu_ser_leu_ser_ile_ser_lys_gly_orn
		Pseudomonas sp. MIS38	AB107223.1	leu_asp_thr_leu_leu_ser_leu_ser_ile_ile_asp
		Pseudomonas fluorescens Pf-5	NC_004129.6	abu_thr_ile_leu_ser_leu_asp_leu_leu_ser_hyv-d
		Pseudomonas sp. CMR12a	JQ309921.1	leu_leu_ser_hyv-d_thr_hyv-d_leu_ser_leu_asp
		Pseudomonas putida strain RW10S2	JN982333.1	thr_val_leu_ser_leu_ser_ile
		Pseudomonas putida strain PCL1445	DQ151887.2	leu_asp_leu_leu_gln_ser_val_leu_ser_leu_val_ser
		Pseudomonas fluorescens Pf0-1	NC_007492.2	leu_leu_ser_ile_gln_ile_leu_gln_ser_leu_asp
		Pseudomonas sp. CMR12a	JQ309920.1	thr_thr_ile_lys_dab_trp_leu_val_val_gln_leu_val_thr_pro_ser_leu_val_gln
113-99	val_leu/ile	*Pseudomonas fluorescens* SBW25	NC_012660.1	leu_ser_ile_lys_orn_ser_thr_val_leu_ser_ser_lys_gly_orn
		Pseudomonas syringae pv. *syringae* B728a	NC_007005.1	ser_ser_dab_dab_arg_phe_thr_asp_thr_thr_pro_val_leu_ala_ala_ala_val_thr_ala_val_dab_dab_tyr_ala_ala_thr_thr_ser_ala_thr_ala
		Pseudomonas putida strain PCL1445	DQ151887.2	leu_asp_leu_val_ser_leu_leu_gln_ser_val_leu_ser

Table 2-2, Continued				
Mass Shift	**Sequence Tag**	**Organism**	**Protein Homolog Accession Number**	**Predicted Amino Acid Sequence**
		Pseudomonas entomophila L48	NC_008027.1	val_leu_ser_ile_gln_val_leu_gln_val_leu_gln_ser
		Pseudomonas putida strain RW10S2	JN982333.1	leu_ser_ile_thr_val_leu_ser
		Pseudomonas syringae pv. tomato str. DC3000	NC_004578.1	leu_leu_gln_leu_thr_val_leu_leu
		Pseudomonas syringae pv. tomato NCPPB 1108	NZ_ADGA01000001.1	leu_thr_val_leu_leu
		Pseudomonas sp. CMR12a	JQ309921.1	thr_hyv-d_leu_ser_leu_asp_leu_leu_ser_hyv-d
		Pseudomonas syringae pv. syringae 642	NZ_ADGB01000036.1	ile_leu_leu
		Pseudomonas fluorescens Pf-5	NC_004129.6	abu_leu_leu_ser_hyv-d_leu_asp_thr_ile_leu_ser
128-113-113-115	asp_ leu/ile _ leu/ile _gln	*Pseudomonas fluorescens* Pf0-1	NC_007492.2	gln_ile_leu_gln_ser_leu_asp_leu_leu_ser_ile
		Pseudomonas fluorescens Pf-5	NC_004129.6	abu_thr_ile_leu_ser_leu_asp_leu_leu_ser_hyv-d
		Pseudomonas putida strain PCL1445	DQ151887.2	leu_asp_leu_leu_gln_ser_val_leu_ser_leu_val_ser
		Pseudomonas sp. CMR12a	JQ309921.1	thr_hyv-d_leu_ser_leu_asp_leu_leu_ser_hyv-d
		Pseudomonas sp. MIS38	AB107223.1	leu_ser_ile_ile_asp_leu_asp_thr_leu_leu_ser
		Pseudomonas syringae pv. tomato str. DC3000	NC_004578.1	leu_thr_val_leu_leu_leu_leu_gln
		Pseudomonas syringae pv. syringae B728a	NC_007005.1	leu_leu_gln_leu_thr_ile_leu_leu
		Pseudomonas syringae pv. tomato NCPPB 1108	NZ_ADGA01000001.1	leu_thr_val_leu_leu

Table 2-2, Continued				
Mass Shift	**Sequence Tag**	**Organism**	**Protein Homolog Accession Number**	**Predicted Amino Acid Sequence**
		Pseudomonas syringae pv. syringae 642	NZ_ADGB01000121.1	leu_leu_gln_leu
		Pseudomonas syringae pv. tabaci ATCC 11528	AEAP01000030.1.1	leu_thr_ile_leu_leu
137-101	his_thr	*Pseudomonas* sp. SHC52	HQ888764.1	ser_orn_asp_lys_his_thr_thr_asp_thr
		Pseudomonas fluorescens Pf-5	NC_004129.6	leu_asp_thr_ile_leu_ser_leu_leu_ser_hyv-d_abu
		Pseudomonas syringae pv. glycinea str. race 4	AEGH01000062.1	glu_arg_ser_lys_asn_thr_thr_ser_asn_ser
		Pseudomonas fluorescens SBW25	NC_012660.1	leu_ser_ile_lys_orn_ser_thr_val_leu_ser_ser_lys_gly_orn
		Pseudomonas syringae pv. tomato T1	NZ_ABSM01000024.1	asp_ser_thr_ser_asn_thr_lys_glu_arg_ser
		Pseudomonas syringae pv. phaseolicola 1448A	NC_005773.3	glu_arg_ser_lys_asn_thr_thr_ser_asn_ser
		Pseudomonas aeruginosa 152504	AEVW01000121.1	
		Pseudomonas syringae pv. tomato str. DC3000	NC_004578.1	leu_leu_gln_leu_thr_val_leu_leu
		Pseudomonas fluorescens strain SS101 clone 2	EU199081.2	leu_ser_ile_thr_ile_leu_ser
		Pseudomonas aeruginosa PA7	NC_009656.1	gly_thr_ser_orn_ser_orn_lys
Bacilli				
113-113	leu/ile _ leu/ile	*Bacillus amyloliquefaciens* CAU-B946	NC_016784.1	glu_leu_leu_val_asp_leu_leu
		Bacillus subtilis	D13262.1	val_asp_leu_glu_leu_leu
		Bacillus subtilis subsp. *subtilis* RO-NN-1	NC_017195.1	val_asp_leu_leu
		Bacillus atrophaeus 1942	NC_014639.1	val_asp_leu_ile_glu_leu_leu

Table 2-2, Continued

Mass Shift	Sequence Tag	Organism	Protein Homolog Accession Number	Predicted Amino Acid Sequence
		Bacillus licheniformis BNP29	AJ005061.1	ile_gln_leu_leu_val_asp_leu
		Bacillus subtilis	X70356.1	glu_leu_leu_val_asp_leu_leu
		Bacillus licheniformis ATCC 14580	NC_006270.3	ile_val_asp_leu_gln_leu_leu
		Bacillus subtilis BSn5	NC_014976.1	glu_leu_leu_leu
		Bacillus licheniformis	U95370.1	gln_leu_leu_ile_val_asp_leu
		Bacillus subtilis subsp. *subtilis*	JQ073775.1	leu_val_asp_leu_glu_leu_leu
		Bacillus thuringiensis BMB171	NC_014171.1	Thr Gln Ala Ser His Gln Gln
128-71-101	gln_ala_thr	*Bacillus thuringiensis serovar huazhongensis* BGSC 4BD1	NZ_ACNI01000052.1	gln_ala_thr_ser_pro_glu_glu
		Bacillus cereus ATCC 10876	NZ_ACLT01000053.1	gln_ala_thr_ser_pro_glu_glu
		Bacillus cereus Rock1-3	NZ_ACMG01000150.1	gln_ala_thr_ser_leu_glu_glu
		Bacillus cereus BDRD-ST196	NZ_ACMD01000179.1	gln_ala_thr_ser_pro_thr_glu
		Bacillus cereus 172560W	NZ_ACLV01000048.1	gln_ala_thr_ser_pro_glu_glu
		Bacillus cereus AH603	NZ_ACMP01000212.1	gln_ala_thr_ser_pro_thr_glu
		Bacillus cereus AH1134	NZ_ABDA02000007.1	gln_ala_thr_ser_pro_glu_glu
		Bacillus sp. 7_6_55CFAA_CT2	NZ_ACWE01000040.1	ser_pro_glu_gln_ala_thr_glu
		Bacillus cereus Rock1-15	NZ_ACMH01000131.1	gln_ala_thr_ser_pro_glu_glu
		Bacillus cereus m1550	NZ_ACMA01000045.1	gln_ala_thr_ser_pro_glu_glu
114-97-129	asn_pro_glu	*Bacillus amyloliquefaciens* Y2	NC_017912.1	ser_thr_tyr_asn_pro_glu_asn_4-hppa_ile_tyr_val_tyr_thr

Table 2-2, Continued				
Mass Shift	**Sequence Tag**	**Organism**	**Protein Homolog Accession Number**	**Predicted Amino Acid Sequence**
		Bacillus amyloliquefaciens FZB42	NC_009725.1	ser_thr_tyr_asn_pro_glu_asn_4-hppa_ile_pro_glu_tyr_glu_val_tyr_thr_glu_orn
		Bacillus subtilis	AY137375.1	ser_thr_tyr_asn_pro_glu_asn
		Bacillus amyloliquefaciens subsp. *plantarum* YAU B9601-Y2	NC_017061.1	ser_thr_tyr_asn_pro_glu_asn_4-hppa_ile_pro_glu_tyr_val_glu_tyr_thr_glu_orn
		Bacillus amyloliquefaciens	JQ271536.1	ser_thr_tyr_asn_pro_glu_asn_ile_pro_glu_tyr_glu_val_tyr_thr_glu_orn
		Bacillus sp. 5B6 5B6	NZ_AJST01000001.1	ser_thr_tyr_asn_pro_glu_asn_4-hppa_ile_pro_glu_tyr_glu_val_tyr_thr_glu_orn
		Bacillus amyloliquefaciens FZB42.	AJ576102.1	ser_thr_tyr_asn_pro_glu_asn_4-hppa_ile_pro_glu_tyr_glu_val_tyr_thr_glu_orn
		Bacillus subtilis subsp. *spizizenii* TU-B-10	NC_016047.1	asn_tyr_asn_gln_pro_4-hppa_ser_asn
		Bacillus amyloliquefaciens XH7	NC_017191.1	4-hppa_asn_ser_tyr_asn_gln_pro_asn
		Bacillus subtilis mycosubtilin	AF184956.1	4-hppa_ser_asn_tyr_asn_gln_pro_asn
113-129-113	leu/ile_glu_leu/ile	*Bacillus licheniformis*	AB016965	ile_cys_leu_glu_ile_lys_orn_ile_phe_his_asp_asn

Table 2-3. Molecule production by organism based on the literature.

Organism in literature	**Molecules Produced**
B. amyloliquefaciens	Surfactin, Iturin, Plipastatin
B. subtilis 3610	Surfactin, Plipastatin, Iturin
P. putida RW10S2	WLIP
P. fluorescens SH-C52	Thanamycin
P. aeruginosa PA01	Rhamnolipid, Quinolone
P. aeruginosa PA14	Rhamnolipid, Quinolone

2.5.2 Nonribosomally produced peptides from bacilli

Mapping the searchable molecules through the creation of a molecular network from the 60 organisms revealed a large cluster of 78 nodes representing molecules with masses ranging from 1002 to 1116 Da found only within the bacilli data (Figure 2-3A). The data incorporated in the nodes came from 23 different data sets (Table 2-4). Generation of sequence tags using only proteinogenic amino acids revealed a 113 Da and a 113-113 Da pair of signatures characteristic of peptides (Figure 2-3B). For the purposes of this study, the longest consecutive sequence tag was used to carry out peptidogenomics because longer tags are more likely to lead to correct identifications. Future algorithms with high-resolution data will enable one to take in account all the tags that are generated, including those with non-proteinogenic amino acids. For this MS/MS cluster, a search tag of 113-113 Da, corresponding to Leu-Leu, Leu-Ile, Ile-Leu or Ile-Ile, was used to search all of the predicted NRPS sequences, obtained from the A-domain specificity predictions, of the publicly available sequences of bacilli and pseudomonads. This 113-113 Da sequence tag matched to sequence tags from *B. subtilis, B. amyloliquefaciens, B. atrophaeus,* and *B. licheniformis,* but not to predicted sequence tags from the pseudomonads. This gene cluster family included the *B. subtilis* surfactin and *B. lichenformis* lichenysin synthetases. Comparing the gene cluster matches from the bacilli revealed that all of the gene clusters had related A-domain specificities and similar gene cluster organization, with over 80% protein sequence similarity (Figure 2-3D). This is in agreement with the known structures of surfactin and lichenysin. At least 17 lichenysins and 53 surfactins are described in the literature with different fatty acid lengths and

geometries, as well as different amino acids in the backbone of the molecule, since the promiscuity of many A-domains leads to the production of MFs. Thus, the surfactin/lichenysin family-gene cluster family was identified. We had included some sequenced strains, such as *B. subtilis* 3610, making it possible to verify from the fragmentation data alone whether the MF contained surfactin. Indeed the surfactin fragmentation data from *B. subtilis* 3610 is found in this cluster (Figure 2-4). Using this approach, four candidate GCF-MF pairs were identified from bacilli (Figure 2-2).

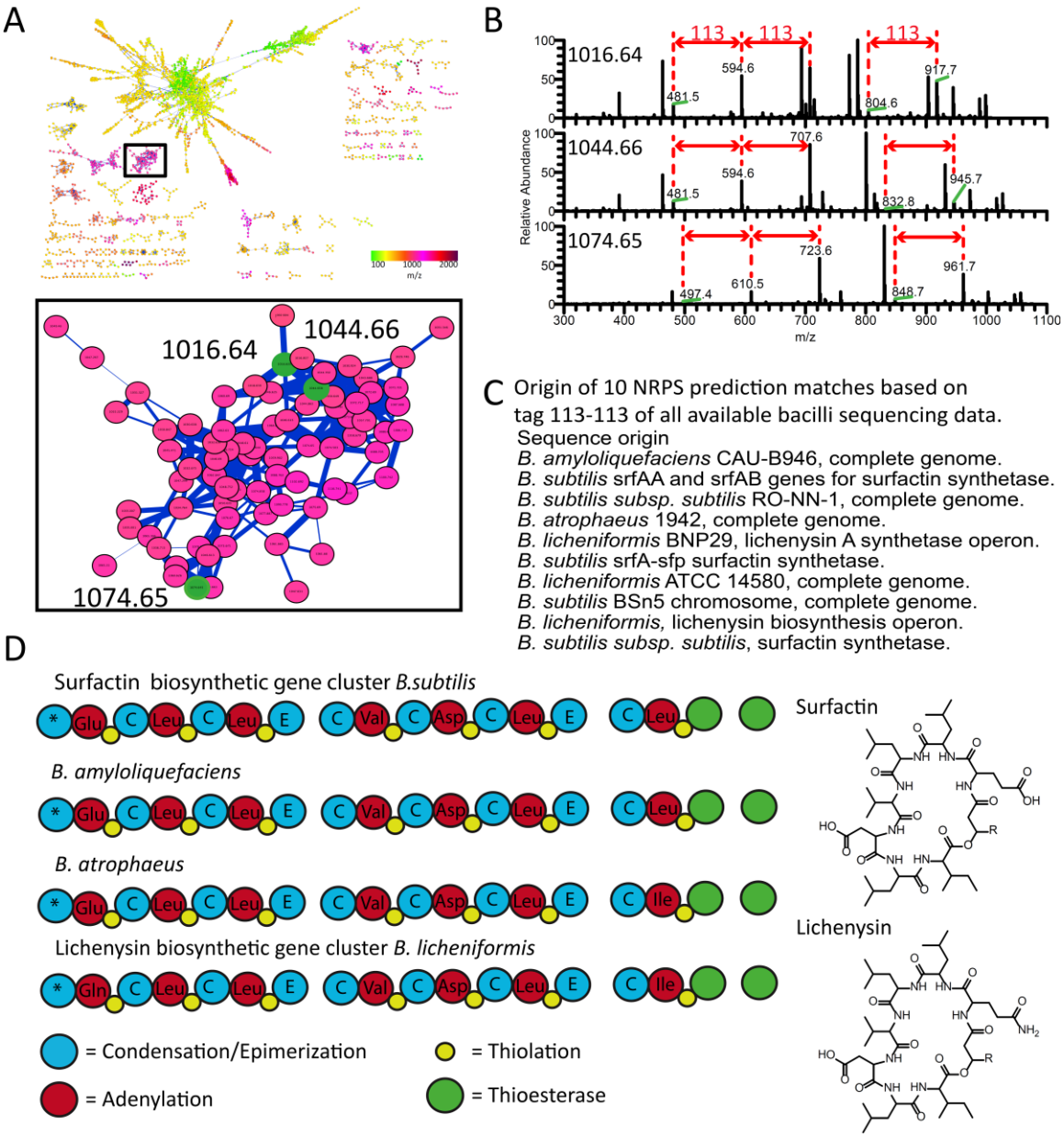


Figure 2-3. Molecular network from bacilli and pseudonomads with the identification of the surfactin molecular family. (**A**) Surfactin molecular family (boxed and enlarged) originating from the bacilli MS/MS clusters. (**B**) Random selection of nodes for raw MS/MS spectra analysis and sequence tag generation. (**C**) Genome mining by antiSMASH of all publicly available sequenced bacilli genomes. (**D**) Matching the generated sequence tags to all of the predicted NRPS gene clusters from bacilli genomes. Starred domains are starter condensation domains.

Table 2-4. Number of organisms contributing to each MF cluster.

Cluster name	***m/z* Range (Da)**	**Organisms found in MS/MS Cluster**	**Number of organisms in cluster**
Iturin	1025-1137	*B. vallismortis* ES113	11
		B. vallismortis ZK4678	
		B. amyloliquefaciens FZB42	
		B. amyloliquefaciens ES223	
		B. subtilis ES73	
		B. amyloliquefaciens ZK4633	
		B. firmus ES118	
		Aeribacillus pallidus ES345	
		B. subtilis ES382	
		B. subtilis 3610	
		B. subtilis KP1032	
Surfactin	1002-1116	*B. firmus* ES118	23
		B. subtilis 3610	
		B. subtilis PY79	
		B. amyloliquefaciens Zk4633	
		B. amyloliquefaciens FZB42	
		B. subtilis KP1032	
		B. firmus ES115	
		B. licheniformis ES44	
		B. marisflavi ES120	
		B. subtilis ES382	
		B. subtilis ES386	
		B. subtilis ES387	
		B. firmus ES118	
		B. subtilis ES73	
		B. vallismortis ES113	
		B. licheniformis ES221	
		B. amyloliquefaciens ES223	
		B. vallismortis ZK4678	
		B. coagulans ES343	
		Aeribacillus pallidus ES345	
		B. pumilus ES76	
		B. sp. (*fusiformis*) ES222	
		B. megaterium LS28	

Table 2-4, Continued			
Cluster name	***m/z* Range (Da)**	**Organisms found in MS/MS Cluster**	**Number of organisms in cluster**
Kurstakin	835-960	*B. cereus* GP24	7
		B. cereus LCT5	
		B. cereus LS18	
		B. firmus ES118	
		B. megaterium ES333	
		B. clausii ES336	
		B. coagulans ES343	
Bacitracin	704-722	*B. firmus* ES115	3
		B. licheniformis ES221	
		B. marisflavi ES120	
Plipastatin	1441-1530	*B. subtilis* 3610	9
		B. subtilis ES73	
		B. licheniformis ES221	
		B. vallismortis ES113	
		B. firmus ES118	
		B. clausii ES336	
		B. subtilis ES73	
		B. subtilis KP1032	
		B. firmus ES115	
Viscosin	1133-1369	*P. tolaasii* CH36	2
Amphisin	1008-1148	*P. moraviensis* ES97	1
Rhamnolipid	499-867	*P. aeruginosa* PA14	4
		P. fluorescens SH-C52	
		P. aeruginosa PAO1	
		P. sp. HCL17 (mosquitos)	
Promysalin	445-489	*P. putida* RW10S1	1
Thanamycin	646	*P. fluorescens* SH-C52	1
Tolaasin	980-1008	*P. tolaasii* CH36	1

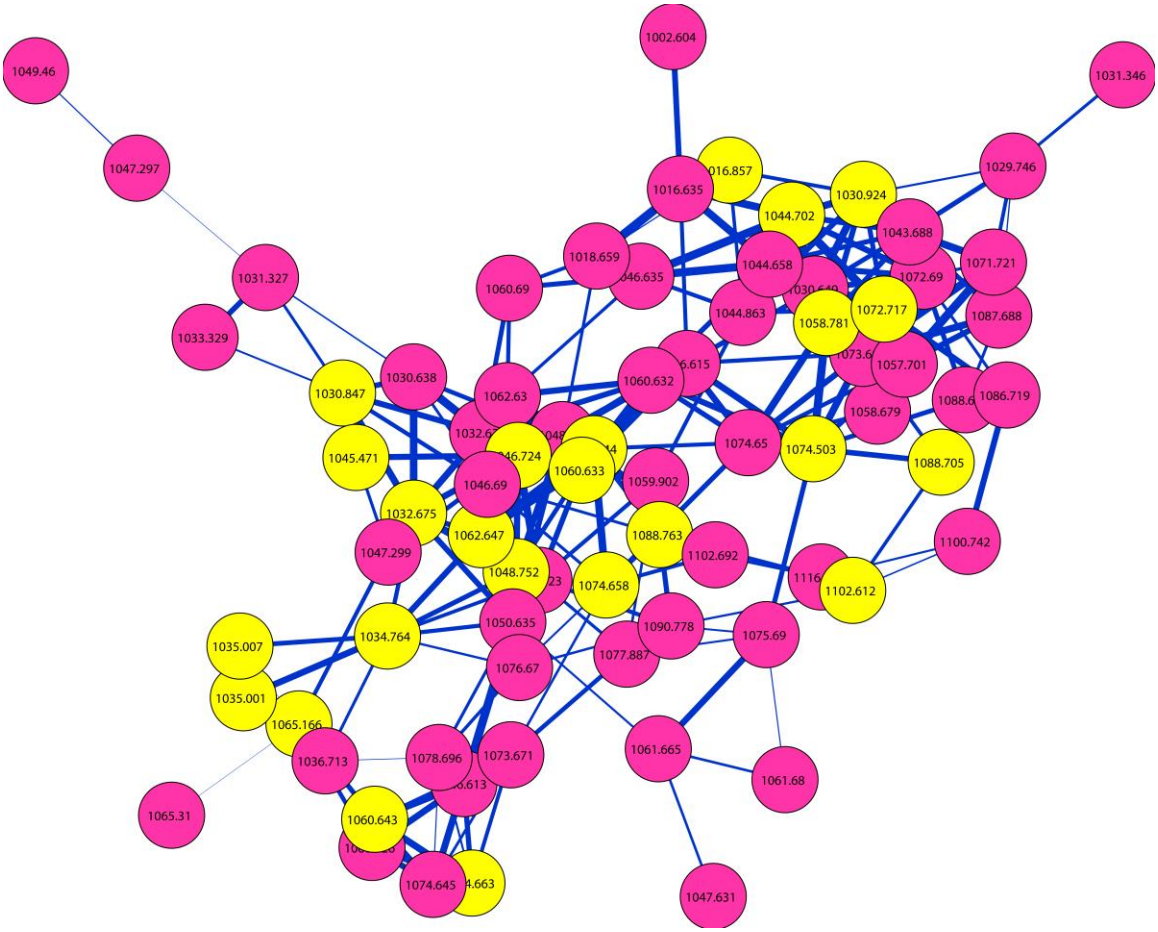


Figure 2-4. The Surfactin MS/MS cluster. Nodes highlighted in yellow are data originating from *Bacillus subtilis* 3610, while nodes in pink are the various *bacilli* found in **Table 2-4.**

2.5.3 Nonribosomally produced peptides from pseudomonads

Similar results were obtained for the pseudomonads. For example, a 17-node MS/MS cluster ranging from 1133 to 1193 Da contained a tag of 87-113-87-113 Da corresponding to Ser-Leu-Ser-Leu, Ser-Ile-Ser-Leu, Ser-Leu-Ser-Ile, Ser-Ile-Ser-Ile, and reverse sequences (Figure 2-5). This sequence tag, from MS/MS data obtained from unsequenced *P. tolaasii*, *P. putida,* and *P. aurantiaca,* matched to the predicted GCFs only from the pseudomonads and not the bacilli. The matches included the viscosin, WLIP, and massetolide gene clusters. Therefore both the gene clusters and molecules that were identified from this GCF-MF pair belong to the viscosin/WLIP and massetolide family of molecules. We confirmed the viscosin cluster by adding MS/MS data

from *P.putida* RW10S2, a known WLIP producer (58). Four candidate GCF-MF pairs were identified from the pseudomonads (Figure 2-2). To date despite the importance of the strain in agriculture, *P.tolaasii* strains are not described to make a molecule belonging to this MF. Upon closer inspection however, a GCF, with the correct gene and domain organization to the viscosin MF is found in the draft genomes of *P.tolaasii* PMS117, 2192 and 6264, consistent with our observations. Based on these two proof-of-principle results, we applied this strategy of genome mining using molecular families to all MS/MS clusters with peptidic signatures. This revealed that 8 out of 121 MS/MS signatures could be correlated to GCFs (Figure 2-2): the GCFs of iturin, surfactin/lichenysin, kurstakin, bacitracin, viscosin, thanamycin, entolysin, and amphisin were successfully paired with their respective MFs.

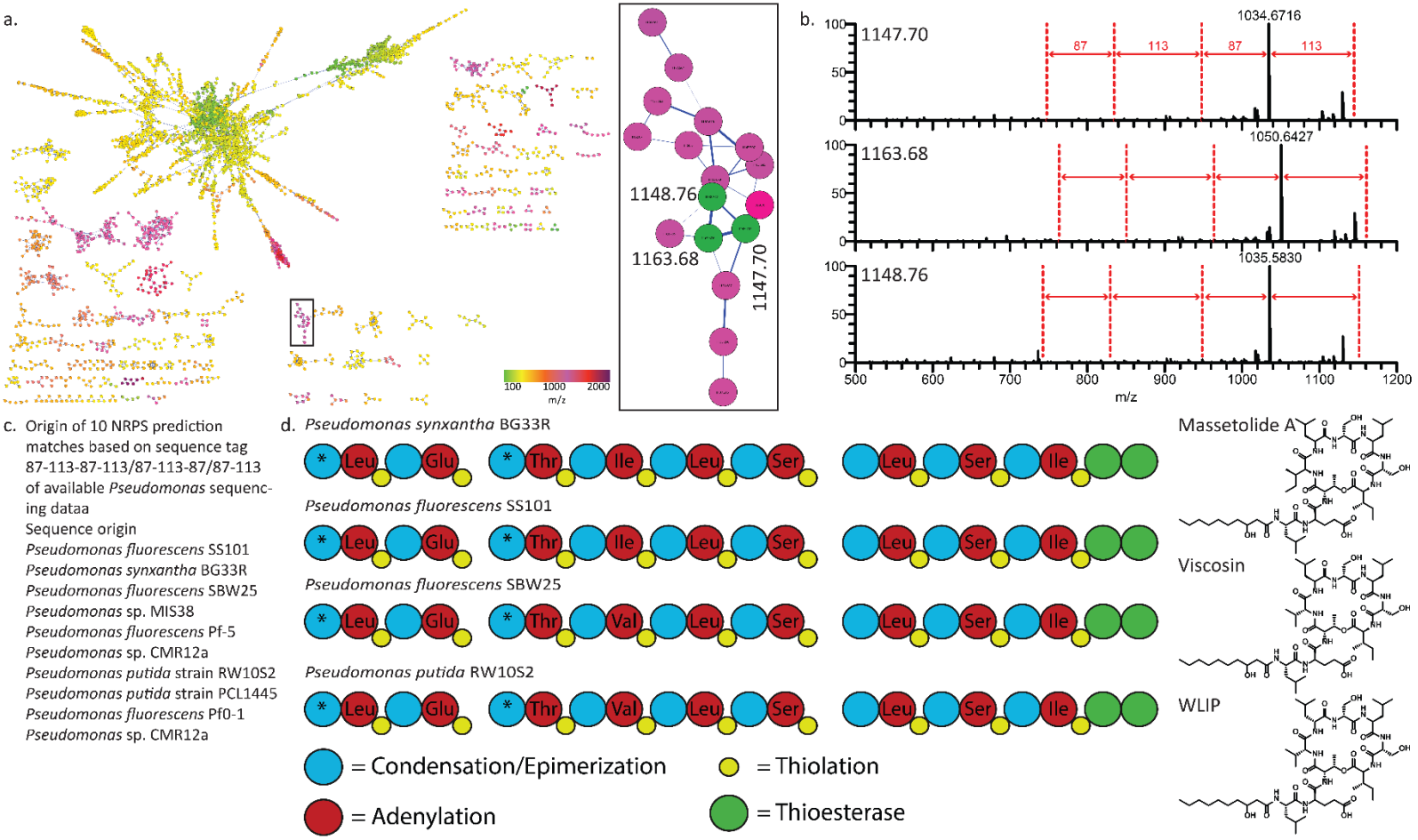


Fig 2-5. Identification of the viscosin GCF-MF pair through the combination of peptidogenomics analysis with publicly available pseudomonad genomes. (A) Molecular network with viscosin MS/MS cluster (boxed and englarged). **(B)** Random node selection for MS/MS raw spectra analysis and sequence tag generation. **(C)** Genome mining took place on publicly available pseudomonad genomes. **(D)** Matching generated sequence tags to the corresponding viscosin NRPS gene clusters.

2.5.4 Dereplication and validation of GCF-MF correlations

When making such correlations, there are at least four possible outcomes. First, the gene cluster and mass spectrometry data for a given molecule may match perfectly with a known molecule-gene cluster pair already described in the literature, as was the case for the bacilli and pseudomonad examples (Figure 2-2). Second, a known molecule is successfully associated to a gene cluster or family of gene clusters where this pair had no prior example in the literature. Third, a family of gene clusters may be associated with a known molecule, based on the MS, MS/MS, and GCF analysis. Finally, there is the possibility that, using the dereplication strategies we employed here, a newly discovered gene cluster may match a newly discovered molecule not previously associated with any molecule already described in the literature. This may indicate either a new molecular entity or an incorrect match.

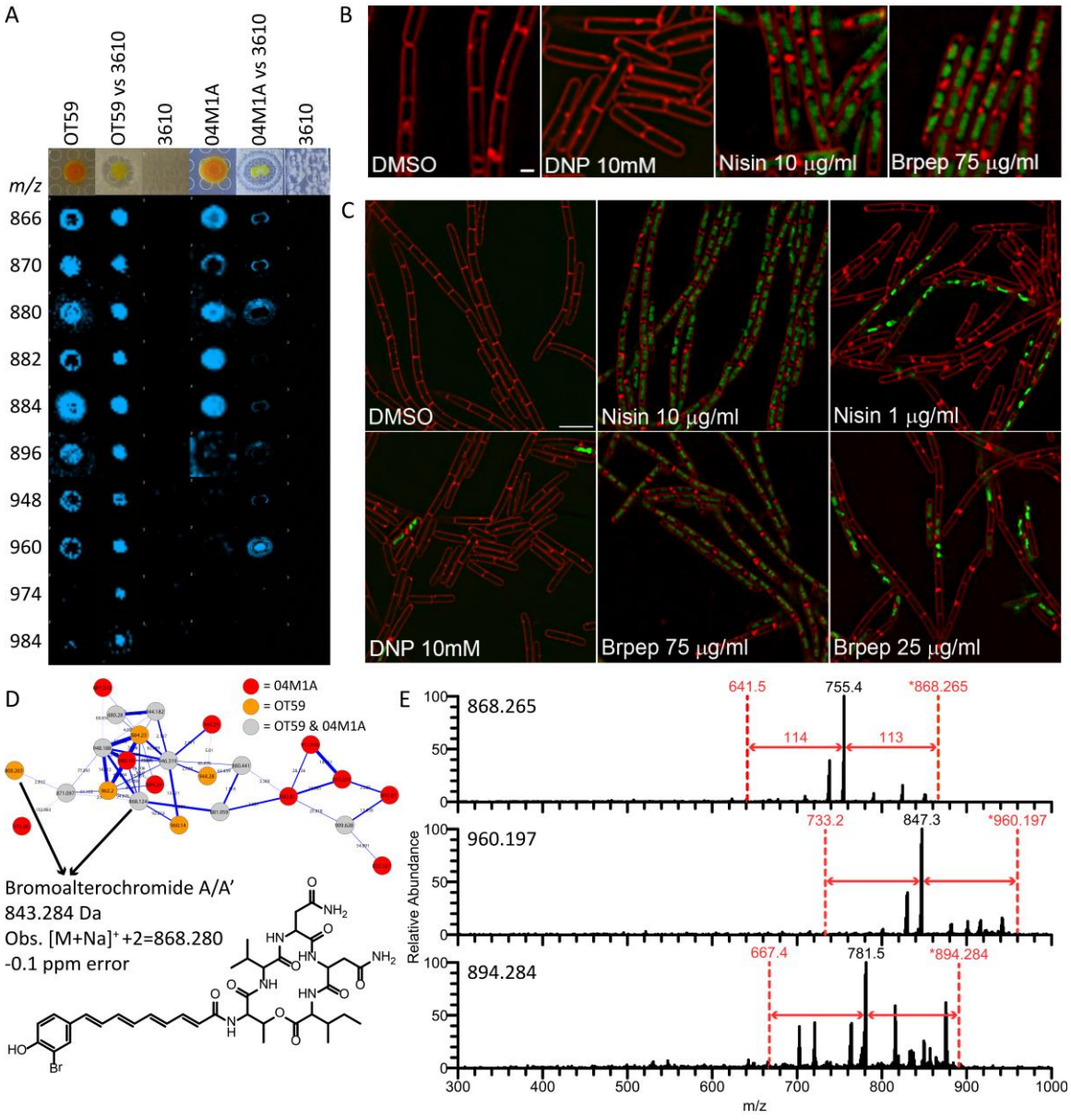
How can we validate any of these correlations? At this stage we cannot unless additional confirmation is acquired. There are many avenues to perform such verifications. These include matching the MS/MS data to MS/MS data published in the literature, comparative *in silico* dereplication to databases such as NORINE, isolation and NMR analysis of the molecule, creating knock-outs of the gene cluster, or (partially) sequencing the gene cluster to verify it is indeed present within the genome of the unsequenced organism. Because the costs for validation are so significant, it is our opinion that these approaches should only be performed when the molecules are of great biological or chemical interest. However, if the effort required for verification of a single molecule takes place, this verification can be extended to the entire GCF-MF correlation and will never have to be carried out again. Furthermore, any new data from molecules or organisms that are added to a network that also cluster to an existing GCF-MF pair can have the original verification extended to it. Should such additional verification fail, however, then the GCF-MF association is incorrect. Thus far, we have not encountered such a scenario with our data. GCF-MF pairs that could not be verified by the literature, or where no literature is found, should only be

considered putative associations. Even though the above analysis was not an exhaustive search, the data provided sufficient evidence that the methodology works and is providing the foundation for future algorithmic development.

2.5.5 Associating GCFs with biologically interesting MFs from unsequenced organisms

Having established the methodology of MS/MS-guided genome mining of unsequenced organisms, we wanted to demonstrate the utility of the method to identify a gene cluster for a molecule showing antimicrobial activity from unsequenced organisms. Two *Pseudoalteromonas* species (strains OT59 and 04M1A) were isolated from the Panamanian octocorals *Leptogorgia alba* and *Psammagorgia sp.*, respectively. Both of these strains showed inhibitory activity against *B. subtilis* 3610 when co-cultured on agar. We used microbial MALDI imaging mass spectrometry (MALDI-IMS) to monitor the distribution of metabolites and to determine which of these metabolites are responsible for the observed bioactivity (22). Subjecting a co-culture of *Pseudoalteromonas* and *B. subtilis* 3610 to IMS revealed two MS signals, *m/z* 880 and 960, from strain 04M1A that sit at the zone of inhibition, suggesting these molecules may contribute to antibiotic activity (Figure 2-6A). We isolated these molecules by HPLC and tested the fractions in a conventional disk diffusion assay against *B. subtilis* 3610. This revealed that a molecule with mass of 921.191 Da (protonated vs potassiated adduct at *m/z* 960 Da) as a major contributor to the antibiotic activity. Cytological profiling of both DNA and the cell membrane suggested that the mechanism of action of the 921.191 Da compound is similar to that of nisin; an amphipathic cationic peptide that disrupts cytoplasmic membranes and distinct from that of other membrane active compounds (Figure 2-6B and 2-6C).

Figure 2-6. Imaging mass spectrometry and MS/MS networking of the bromoalterochromide family, and fluorescence microscopy of the 921.191 Da dibromoalterochromide and its effect on *B. subtilis* 3610 cells. (**A**) MALDI-IMS showing bromoalterochromide production by *Pseudoalteromonas* OT59 and 04M1A and its special correlation with *B. subtilis* 3610 inhibition. (**B**) Fluorescence micrographs of growing *B. subtilis* 3610 cells treated with DMSO, 2,4-dinitrophenol (DNP), or the 921.191 Da dibromoalterochromide. Red stain is FM 4–64, a fluorescent membrane stain; green stain is Sytox Green, a DNA stain that is membrane impermeable and shows increased fluorescence in permeabilized cells. (**C**) Fluorescence micrographs of growing *B. subtilis* 3610 cells treated with DMSO, DNP and two concentrations of nisin the 921.191 Da dibromoalterochromide at or below (right panels) the MIC . (**D**) MS/MS networking from nanoDESI MS of *Pseudoalteromonas* OT59 and 04M1A. The monobromoalterochromide gives two major isotopes while the dibromoalterochromide gives three major isotopes. At least one of these isotopes was detected by MS/MS networking. (**E**) Node selection of predicted bromoalterochromide species (possibly present as protonated, sodiated, and potassiated adducts) for raw MS/MS spectra analysis and sequence tag generation. Amino acid mass shifts of 113-114 Da correspond to a sequence tag of Ile-Asn or Leu-Asn.


A
OT59
OT59 vs 3610
3610
04M1A
04M1A vs 3610
3610
m/z
866
870
880
882
884
896
948
960
974
984
B
DMSO
DNP 10mM
Nisin 10 μg/ml
Brpep 75 μg/ml
C
DMSO
Nisin 10 μg/ml
Nisin 1 μg/ml
DNP 10mM
Brpep 75 μg/ml
Brpep 25 μg/ml
D
= 04M1A
= OT59
= OT59 & 04M1A
Bromoalterochromide A/A'
843.284 Da
Obs. [M+Na]+ +2=868.280
-0.1 ppm error
E
868.265
641.5
755.4
*868.265
114
113
960.197
733.2
847.3
*960.197
894.284
667.4
781.5
*894.284
Relative Abundance
m/z

Subjecting colonies of *Pseudoalteromonas* OT59 and 04M1A grown on agar to nanoDESI and MS/MS networking revealed that the mass of 921.191 Da belonged to a cluster of related molecules (Figure 2-6D). Inspecting the MS/MS data from the 921.191 Da and surrounding nodes revealed that the molecules were peptidic and had a sequence tag of Ile/Leu and Asn (Figure 2-6E). A search in the AntiMarin database did not find a match for the 921.191 ion, but there was a match to a related node. The match was to a molecule of mass 843.280 Da named bromoalterochromide A and A', an unusual monobrominated lipopeptide (Figure 2-6E) (50, 51). To confirm that the 921.191 Da compound isolated from both OT59 and 04M1A was indeed a bromoalterochromide, the MS/MS spectra were further examined and an NMR analysis of the purified compound was performed (Table 2-5). This data was compared to that found by Kalinovskaya and coworkers (51). Both the NMR and MS/MS analysis support the assignment of this molecule to the bromoalterochromide family. This highlights how MS/MS networking tools can be used to dereplicate related molecules. Our molecule was 77.911 Da larger than the bromoalterochromide found in AntiMarin, suggesting our molecule was a doubly brominated bromoalterochromide. To find a candidate GCF, every publicly available *Pseudoalteromonas sp.* genome was mined for NRPS or NRPS/PKS hybrid gene clusters (as we had done with the bacilli and pseudomonads) and the A-domain specificities were examined until we found a positive hit within *Pseudoalteromonas piscicida* JCM 20779^{T} (52). The A-domain and gene cluster analysis revealed that the gene cluster from *P. piscicida* contained a halogenase, type II polyketide synthase/type II fatty acid proteins, and NRPS modules that are predicted to load Thr-Val-Asn-Asn-Ile/Leu (Table 2-6). This means we have found a putative GCF-MF pair.

Table 2-5. NMR measurement. The 921.191 Da dibromoalterochromide was dissolved in 50 µL of CD_3OD for NMR acquisition. NMR spectra were recorded on a Bruker Avance III 600 MHz (Manex superconducting magnet, 14.1T) NMR with 1.7 mm Micro-CryoProbe at 298 K, with standard pulse sequence provided by Bruker. Data was analyzed using MestReNova software. Chemical shifts based off HSQC, HMBC, and ^{1}H-NMR.

Unit	**Position**	**^{13}C-NMR**	**^{1}H-NMR**	**Multiplicity**	***J* (Hz)**
Thr	C=O	n/o[a]	--		
	α	56.85	4.74	d	3.4
	β	74.06	4.98	m	
	γ	17.21	1.54	m	
	NH	--	n/o[a]		
Val	C=O	172.53	--		
	α	62.12	3.98	m	
	β	30.29	2.04	m	
	γ	19.18	1.02	m	
	γ'	19.18	1.02	m	
Asn 1	C=O	170.34	--		
	α	53.17	4.5	m	
	β	37.12	2.57	m	
	γ C=O	172.55	--		
	NH	--	n/o[a]		
	NH2	--	n/o[a]		
Asn 2	C=O	170.34	--		
	α	53.17	4.5	m	
	β	37.12	2.57	m	
	γ C=O	172.55	--		
	NH	--	n/o[a]		
	NH2	--	n/o[a]		
Leu	C=O	174.31	--		
	α	58.13	4.54	m	
	β	30.43	2.04	m	
	γ	n/o[a]	n/o[a]		
	δ	14.11	0.9	m	
	δ'	14.11	0.9	m	
	NH	--	n/o[a]		
Ile	C=O	169.12	--		
	α	53.13	4.5	m	
	β	39	2.01	m	

Table 2-5, Continued					
Unit	**Position**	**^{13}C-NMR**	**^{1}H-NMR**	**Multiplicity**	***J* (Hz)**
	γ	30.37	1.29	m	
	δ	14.74	0.93	m	
	γ'	11.49	0.95	m	
	NH	--	n/o[a]		
Ph	1'	132.7	--		
	2'	131.21	7.59	s	
	3'	110.84	--		
	4'	150.23	--		
	5'	110.84	--		
	6'	131.03	7.62	s	
	OH				
CO-(CH=CH)$_4$	1	n/o[a]	6.12-7.01	m	
	2	n/o[a]	6.12-7.01	m	
	3	n/o[a]	6.12-7.01	m	
	4	n/o[a]	6.12-7.01	m	
	5	n/o[a]	6.12-7.01	m	
	6	n/o[a]	6.12-7.01	m	
	7	n/o[a]	6.12-7.01	m	
	8	n/o[a]	6.12-7.01	m	
	9	n/o[a]	6.12-7.01	m	

[a] n/o = not observed

Table 2-6 . Genes in the bromoalterochromide biosynthetic cluster of *P. piscicida* JCM 20779^T (NZ_AHCC01000002.1) and predicted functions based on homology and protein domain analysis.

Protein ID	Size [aa]	Predicted Function	Homologous Gene Function	Accession Number	Protein Identity/Protein Similarity
ZP_10287793.1	539	Ammonia Lyase	phenylalanine ammonia-lyase/histidase family protein [Pseudomonas syringae pv. oryzae str. 1_6]	ZP_04586587.1	41/59
ZP_10287794.1	97	Acyl Carrier Protein	acyl carrier protein [Vibrio tubiashii ATCC 19109]	ZP_08738232.1	43/65
ZP_10287795.1	498	Acyl-CoA Synthetase	AMP-dependent synthetase and ligase [Hirschia baltica ATCC 49814]	YP_003059702.1	26/43
ZP_10287796.1	727	3-oxoacyl-(Acyl-carrier-protein) synthase	3-oxoacyl-(Acyl-carrier-protein) synthase II [Bacillus thuringiensis serovar pulsiensis BGSC 4CC1]	ZP_04081953.1	33/54
ZP_10287797.1	255	ABC Transporter	ABC transporter ATPase [Candidatus Koribacter versatilis Ellin345]	YP_589363.1	53/74
ZP_10287798.1	371	ABC Transporter	ABC-2 type transporter family [Stigmatella aurantiaca DW4/3-1]	ZP_01465758.1	29/49
ZP_10287799.1	132	Hydroxyacyl ACP Dehydratase	beta-hydroxyacyl-(acyl-carrier-protein) dehydratase FabZ [Synergistes sp. 3_1_syn1]	ZP_09362329.1	42/61
ZP_10287800.1	243	3-oxoacyl ACP reductase	3-oxoacyl-(Acyl-carrier protein) reductase [Bacillus thuringiensis serovar pulsiensis BGSC 4CC1]	ZP_04081955.1	40/61
ZP_10287801.1	235	--	hypothetical protein bthur0012_56500 [Bacillus thuringiensis serovar pulsiensis BGSC 4CC1]	ZP_04081956.1	24/50

Table 2-6, Continued					
Protein ID	**Size [aa]**	**Predicted Function**	**Homologous Gene Function**	**Accession Number**	**Protein Identity/Protein Similarity**
ZP_10287802.1	251	Thioesterase	thioesterase [Streptomyces virginiae]	BAF50718.1	33/47
ZP_10287803.1	1517	NRPS 1: Thr	Non-ribosomal peptide synthase [Xenorhabdus nematophila ATCC 19061]	YP_003712913.1	36/53
ZP_10287804.1	2615	NRPS 2: Val-Asn	Non-ribosomal peptide synthase [Xenorhabdus nematophila ATCC 19061]	YP_003712913.1	40/57
ZP_10287805.1	3280	NRPS 3: Asn-Leu/Ile	amino acid adenylation domain-containing protein [Nostoc punctiforme PCC 73102]	YP_001866468.1	33/52
ZP_10287806.1	412	Halogenase	flavoprotein/dehydrogenase [Bizionia argentinensis JUB59]	ZP_08820287.1	47/65

2.5.6 Confirming the GCF-MF pairing of the bromoalterochromides from *Pseudoalteromonas*

Several complementing approaches were employed to verify that this molecule was indeed a member of the bromoalterochromide family of NRPS-derived molecules. First, if our prediction of the gene cluster-molecule family pair was indeed correct, then the sequenced organisms *P. piscicida* JCM 20779^{T} should also produce the bromoalterochromides, even though this has not been described in the literature. To confirm the production of the bromoalterochromides from *P. piscicida* JCM 20779^{T}, the organism was obtained and subjected to nanoDESI. MS/MS data generated from *P. piscicida* JCM 20779^{T} was merged with the MS/MS networking data from OT59 and 04M1A, revealing that *P. piscicida* JCM 20779^{T} does indeed produce compounds that fall within the bromoalterochromide cluster. This suggests that the biosynthetic machinery is present

in *P. piscicida* JCM 20779^T as in OT59 and 04M1A (Figure 2-7). Interestingly, *P. piscicida* JCM 20779^T only produces the 843.280 Da monobromoalterochromide and not the 921.191 Da dibromoalterochromide, whereas OT59 and 04M1A produce both. These results provide additional confirmation that the GCF-MF pair has correctly been identified and that it is possible to connect MFs from unsequenced organisms to GCFs in publicly available sequencing data. Having candidate molecules and the gene cluster in hand, it is now possible to evaluate if the molecule and biosynthetic gene cluster match as well. The gene cluster contains all of the biosynthetic components needed to produce the bromoalterochromides and was identified bioinformatically (Figure 2-7 and SI Table 2-6). The condensation and epimizerization domains were subjected to phylogenetic analysis with starter C domains, $^{L}C_{L}$ and $^{D}C_{L}$ condensation domains, dual C/E domains, and standalone E domains as described by Rausch and colleagues (59). Interestingly, the epimerization domains from *P. piscicida* JCM 20779^T do not clade tightly with the epimerization domains from *Bacillus cereus*, *B. licheniformis*, or *B. subtilis*, most likely due to phylogenetic divergence of the organisms, which also inhabit different ecological niches (Figure 2-8). Still, the epimerization domains are in the correct locations to encode for D-Thr, D-Val, L-Asn, D-Asn, D-Leu/Ile, as previously described (50, 51). The polyketide portion of the biosynthetic pathway is missing the enoyl reductase, as one would predict based on the structure of the molecule. Lastly, the pathway contains a flavin-dependent halogenase. The only other candidate brominating flavin-dependent halogenase described to date is found in the jamaicamide pathway (60). Again, the gene cluster analysis matches perfectly with the expected biosynthesis of the bromoalterochromides, supporting the notion that this molecule family, including 921.191 Da, belong to the bromoalterochromide family. Finally, to further confirm this finding of the bromoalterochromide GCF-MF pair, we set out to show that similar biosynthetic genes exist in our strains. For this reason, OT59 and 04M1A were subjected to partial genome sequencing using Illumina sequencing, which revealed that the same NRPS genes are present with 96% identity to the genes found in *P. piscicida*

JCM 20779^T based on blastn sequence alignments (Figure 2-9). The extensive tasks ranging from NMR analysis of the purified bromoalterochromide to the gene cluster analysis and partial sequencing of OT59 and 04M1A were carried out to verify, with complete certainty, that we had the correct molecule as well as the correct GCF-MF pairing. All of this information that was dedicated towards the analysis of the 921.191 Da dibromoalterochromide but can now be extended to the various family members of this molecule. Additionally, if any other data are added to this network, and cluster to this MF, the steps for verification of the molecule and GCF-MF pairing do not need to be carried out again, thus increasing the speed and reducing the cost of studying these molecules. Combined, these data demonstrate that MS/MS networking peptidogenomics analysis enables the mapping of observed MFs to already available sequenced genomes and that this can lead to the discovery of the previously unidentified GCFs, as we demonstrated for the 921.191 Da antimicrobial agent dibromoalterchromide from unsequenced *Pseudoalteromonas* species.

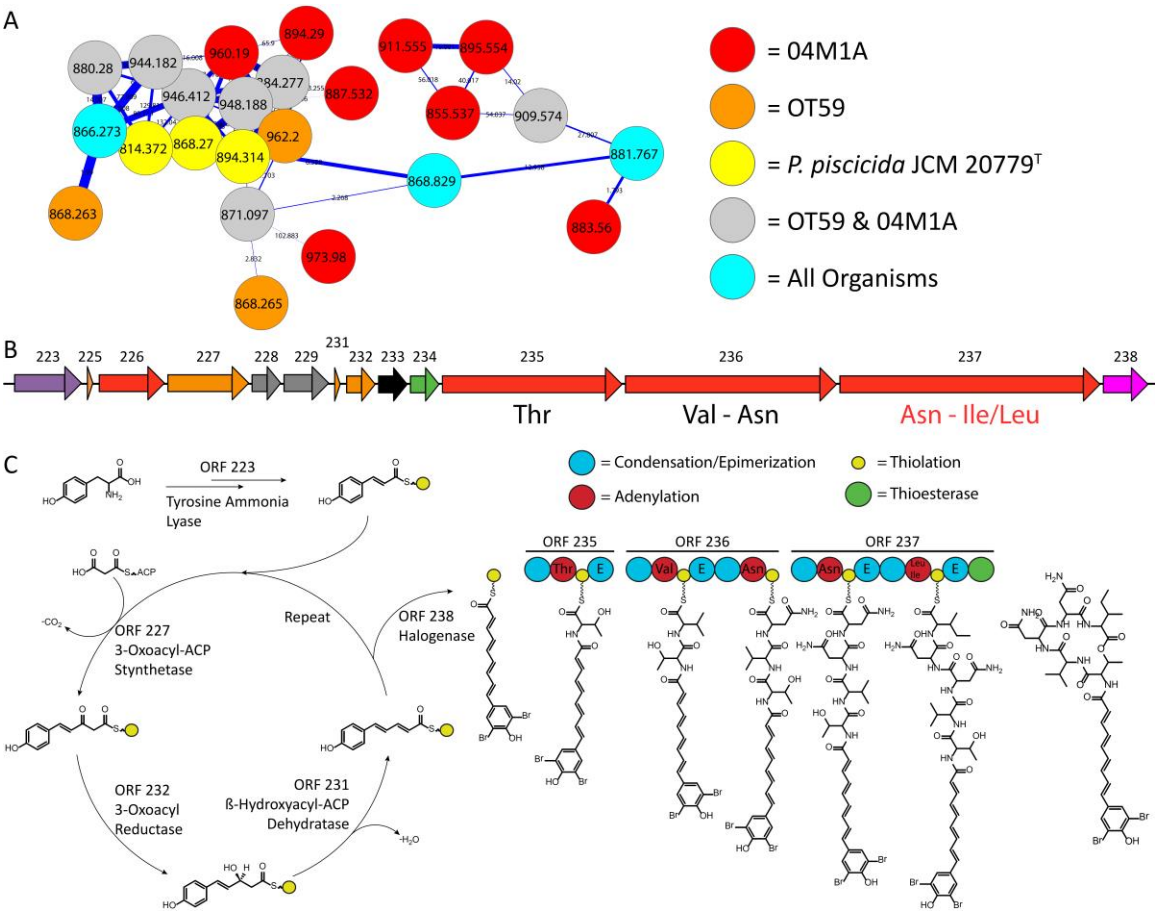


Figure 2-7. MS/MS networking of *Pseudoalteromonas* OT59, 04M1A, and *piscicida* JCM 20779^T. (**A**) MS/MS cluster of bromoalterochromide (red, orange, yellow, grey, and teal nodes are 04M1A, OT59, *P. piscicida* JCM 20779^T, OT59/04M1A, or metabolites originating from all organisms, respectively). (**B**) Bromoalterochromide gene cluster. Purple – ammonia lyase, orange – type II PKS/fatty acid synthase, red – NRPS, grey – transporters, green – thioesterase, pink – halogenase, and black – hypothetical proteins. (**C**) Proposed biosynthetic pathway with speculative double halogenation by ORF 238 after acyl chain elongation.

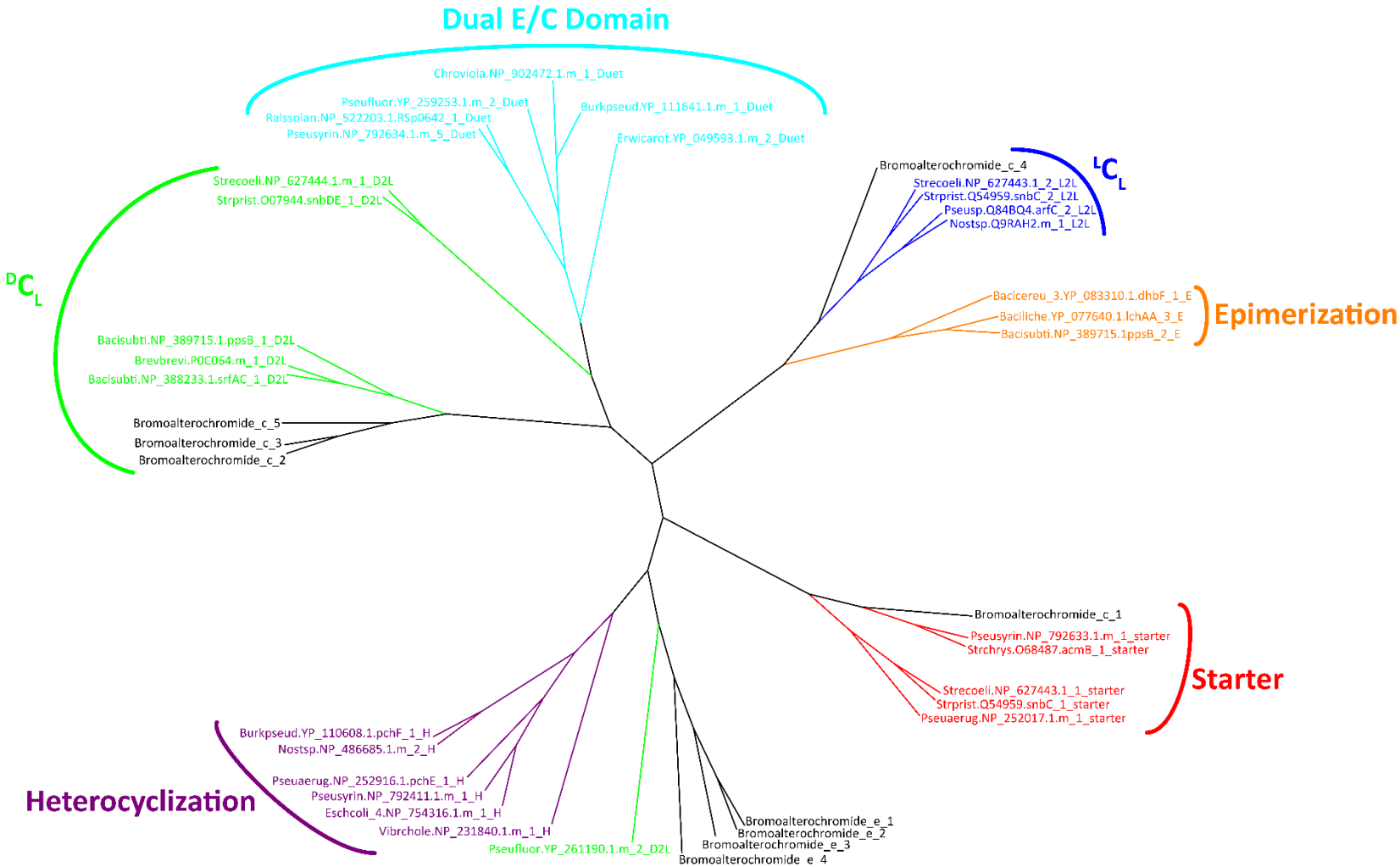


Figure 2-8. Phylogenetic analysis of E and C domains from bromoalterochromide NRPS gene cluster. The first C domain clusters amongst the other starter C domains; C domains 2, 3, and 5 cluster with the $^{D}C_{L}$ domains; C domain 4 clusters with the $^{L}C_{L}$ domains suggesting the first asparagine is an L-Asn and the second a D-Asn. The bromoalterochromide epimerization domains clade alone.

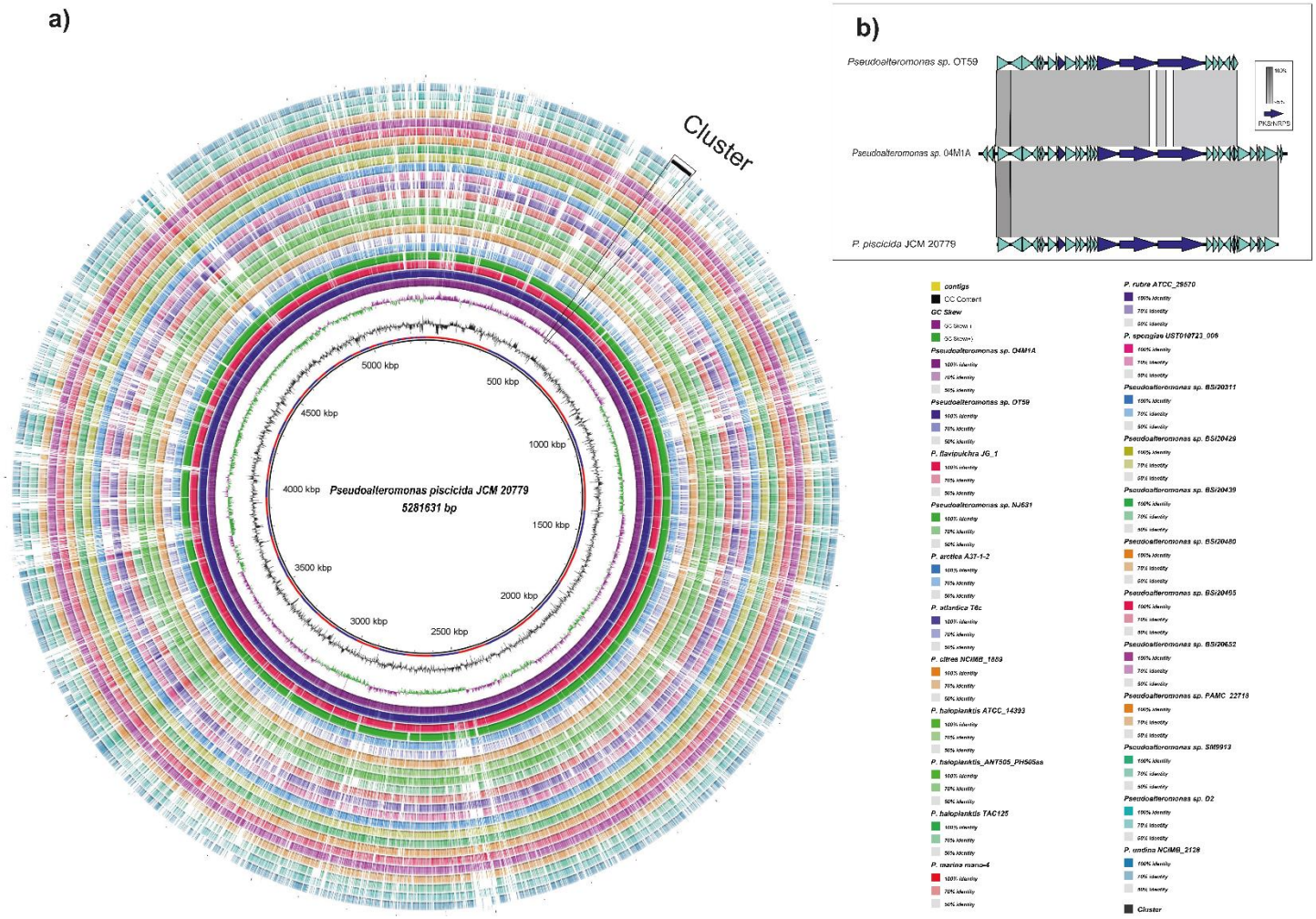


Figure 2-9. **Comparison of OT59 and 04M1A sequence to publicly available *Pseudoalteromonas* genomes. a)** Circular representation of 73 contigs of *Pseudoalteromonas piscicida* JCM 20779^{T} in comparison with 23 other *Pseudoalteromonas spp.* **b)** Bromoalterochromide biosynthetic gene cluster in *P. piscicida* JCM 20779^{T}, 04M1A, and OT59.

2.6 Conclusion

One of the major bottlenecks in genome mining is that it takes a significant amount of time and money to connect a molecule to its biosynthetic signature. Here we have highlighted one example of how we can increase the speed and reduce the cost of the analysis of specialized metabolites from a large cohort of organisms using sequence information already available in public databases. We targeted NRPS systems with our MS/MS network-based genome mining strategy to demonstrate that genome mining can be accomplished for unsequenced organisms by borrowing an already sequenced genome of a related organism. We expect, however, that through the creative adaptation of this approach, additional small molecule classes such as isoprenoids, polyketides, oligosaccharides, glycolipids, and lipids can be mined as well. Connecting the gene cluster families and molecular families is additive as it provides an approach to correlate molecules to genetic information that, once linked in a manner that is analogous to annotations in sequence repositories, can be extrapolated quickly to new samples, especially as more genome sequences become available. This saves time and effort and so far we have not yet reached a limitation in terms of the number of samples that can be compared. Such an approach (or related approaches) will become the first step in the molecular characterization of unsequenced microbes, even in field collected samples, and especially as mass spectrometers interfaced with ambient ionization are becoming cheaper and even portable (39, 40). Our approach could serve nicely as a strain pre-selection strategy for therapeutic discovery applications and it is now beginning to be used in our laboratories to mine metagenomics data rather than strictly full genome sequencing data. With this approach one can begin to create GCF-MF associations for molecules analyzed directly from environmental and personal microbial communities, such as the ones found on our skin, in our gut, in soil, in coral reefs, or on plant roots—something not commonly attempted with today's genome mining technologies and thereby avoiding the requirement of producing viable cultures in the laboratory.

2.7 Materials and Methods

Bacilli, pseudomonad, OT59, 04M1A, and *P. piscicida* JCM20779^T culture conditions

The 42 bacilli and 18 pseudomonads are listed in Table 2-1. All strains were grown in LB broth (Fischer Scientific) overnight, shaken at 28°C, transferred to ISP2 agar (1 L of medium containing 4 g yeast extract (Sigma Aldrich), 10 g malt extract (Sigma-Aldrich), 4 g dextrose (EMD), and 10 g agar (Sigma-Aldrich)) and incubated for 48 hours at 30°C. OT59, 04M1A, and *P. piscicida* JCM 20779^T grown on M1 agar (500 mL of medium contained 9 g agar (Sigma-Aldrich), 5 g potato starch (Sigma-Aldrich), 2 g yeast extract (Sigma-Aldrich), 1 g peptone (Sigma-Aldrich), and 14 g of aquarium salt (Aquatic Systems, Instant Ocean)).

Sample preparation for Matrix-assisted laser ionization (MALDI)

After the strains were grown individually for 48 hours on ISP2 agar, approximately 1 µl of cells were scraped directly from the live colony and transferred to an MSP 96 MALDI anchor plate. The cells were then covered with 1 µL of saturated matrix solution (35 mg/mL of Universal MALDI Matrix (1:1 mixture of DHB and α-cyano-4-hydroxy-cinnamic acid, Sigma-Aldrich) in 78% acetonitrile and 1% formic acid), until proper crystal formation. The MALDI plate was inserted into an Autoflex Bruker Daltonics mass spectrometer and data was recorded in reflectron positive mode. Data analysis was carried out using ClinProTools to generate heat maps to analyze chemical profiles of the strains simultaneously.

Live colony nanoDESI MS/MS Data Acquisition

Overnight cultures of bacilli and pseudomonads were prepared as stated above. Four cultures (0.5 µL each) were spotted on an ISP2 agar plate and grown for 48 hours at 30°C. Colonies of OT59, 04M1A, and *P. piscicida* JCM20779^T were grown as stated above. NanoDESI was carried out as described by Watrous et al. using a solvent mixture of 65:35 acetonitrile:water with 0.05%

formic acid for the bacilli and pseudomonads and 50:50 MeOH:Water with 1% formic acid for the *Pseudoalteromonas* (7). Spray voltage was kept between 2.0 and 3.0 kV. Data was collected using a data-dependent MS/MS method on a hybrid 6.4T LTQ-FT (Thermo Electron, North America) mass spectrometer. In this method, an MS^1 scan of 50-1600 m/z was followed by MS/MS of the four most intense ions (2 m/z isolation width, a normalized collision energy of 35%, and an activation time of 30 ms), which were then added to an exclusion list, allowing for another MS^1 scan followed by MS/MS of the next four most intense ions.

MS/MS networking and sequence tagging

The MS/MS data of 42 bacilli, 18 pseudomonads, OT59, 04M1A, and P. piscicida JCM 20779^T were clustered as described by Watrous et al. (34). Algorithms assumed a precursor mass tolerance of 1.0 Da and a fragment mass tolerance of 0.3 Da with the cosine threshold set at 0.7. Two plugins were used for aid with data visualization. The FM3 layout was used to organize and align the nodes within the network and the HiderSlider plugin was used to hide or show nodes within the network to determine whether the origin of the node was bacilli or pseudomonad in origin. Once clusters of specific molecules were located, individual nodes were selected and the MS/MS spectra were examined for sequence tags.

Peptidogenomics and genome mining

All available genome sequences for the bacilli and pseudomonads were gathered from the National Center for Biotechnology Information (NCBI), U.S. Department of Energy Joint Genome Institute (DOE JGI), and PseudoDB. *Pseudoalteromonas* spp. genomes used for peptidogenomics and genome mining are described by Xie BB and coworkers (52). Targeted nucleotide sequences were subject to antiSMASH, NP.searcher, NRPSpredictor2, and PKS/NRPS Analysis to determine the amino acid specificity of the NRPS A-domain (44, 16, 43, 44, 57). NRPS and NRPS/PKS hybrid

gene clusters were screened with the Ile/Leu-Asn sequence tag obtained from molecular networking. After obtaining potential NRPS gene clusters, the protein sequences were pulled out and A-domain accuracy was examined by PKS/NRPS Analysis to eliminate NRPS genes that were unlikely to produce the bromoalterochromides. For all remaining gene clusters, BLAST analysis was performed on the NRPS-surrounding genes to determine their functions.

MALDI-IMS Screening of *Pseudoalteromonas* OT59 and 04M1A against *B. subtilis* 3610

B. subtilis 3610 was grown to an OD of 0.2-0.5 in M1 liquid media and 20 uL were spotted onto an M1 agar plate (described above) and spread into a lawn using glass beads. The cultures were allowed to dry, at which point 2 μL of OT59 or 04M1A stock, frozen at an OD of 1.0 in 20% glycerol in M1 liquid media, was spotted at the center of the plate. The cultures were incubated at 30°C in the dark for 48 hours. The interactions of OT59 with B. subtilis 3610 were prepared as stated above. These interactions, as well as the corresponding controls, were excised out of the agar and placed on a Bruker MSP 96 stainless steel target plate. A film of Universal MALDI Matrix (Sigma-Aldrich) was applied to the surface of the excised agar using a 53 μm sieve (Hogentogler & Co., Inc). The target plate was dried at 37°C until the agar pieces had dried completely and adhered to the target plate. The samples were subjected to MALDI-IMS using reflectron positive mode on a Bruker Microflex with Compass 1.2 software suite containing flexImaging 2.0, flexControl 3.0, and flexAnalysis 3.0.

Extraction and Isolation of the 921.191 Da dibromoalterochromide

2 μL of OT59 stock, frozen at an OD of 1.0-1.2, was inoculated onto an M1 agar plate. The colonies were grown for 48 hours at 30°C at which point the colonies were excised from the agar plate and extracted with methanol. The crude extract was fractionated on a Sephadex LH-20 column (MeOH) at a flow rate of 1.5 mL/4 min where upon the fractions were tested for bioactivity against

B. subtilis 3610. Lawns of *B. subtilis* 3610 were prepared as stated above. The fractions were dried down and resuspended in 10-100 μL of methanol. 2-8 μL were spotted onto a paper disk, allowed to dry, and then placed onto the newly prepared lawn of *B. subtilis* 3610. The cultures were incubated in the dark at 30°C for 48 hours. Bioactive fractions were analyzed by MALDI for molecular signature. Bioactive sephadex fractions containing the bromoalterochromides were pooled together and finally purified by HPLC, to obtain the 921.191 Da dibromoalterochromide. Purification was performed on an Agilent 1260 HPLC equipped with a Discovery reverse phase C18 5 μm, 180 A, 25 cm x 10 mm column (Supelco) using the water/acetonitrile gradient listed below. Solvent A is HPLC grade water (J. T. Baker) with 0.1% TFA (Sigma-Aldrich), and solvent B is HPLC grade acetonitrile (J. T. Baker) with 0.1% TFA. The gradient was run at a flow rate of 2 mL/min.

Minute	0	1	31	37.5	38	39	40
% A	60	60	35	0	0	60	60
% B	40	40	65	100	100	40	40

NMR measurements of the 921.191 Da dibromoalterochromide

NMR data was acquired at the UCSD Skaggs School of Pharmacy and Pharmaceutical Sciences NMR Facility. Proton and 2D NMR of the purified 921.191 Da dibromoalterochromide were measured on a 600 MHz NMR (Manex superconducting magnet, 14.1 T) fitted with a 1.7 mm cryoprobe, Bruker Avance console, and operated using Bruker TopSpin software.

Fluorescence microscopy

B. subtilis PY79 was grown in LB to an OD_{600} of 0.3, centrifuged, resuspended in 1/10 the volume and 14.25 μL of concentrated cells were added to a 1.7 mL microcentrifuge tubes. At t = 0, 0.75 μL of 100% DMSO and appropriately diluted samples of DNP, nisin or the 921.191 Da dibromopeptide (in 100% DMSO) were added to cell aliquots. The tubes were capped and

incubated at 37°C in a roller. After 20 minutes, 3 µL of cells were added to 0.75 µL of stain mix containing 30 µg/mL FM 4-64, 2.5 µM Sytox Green and 1 µg/mL DAPI (4',6-diamidino-2-phenylindole) prepared in 1X T-base and placed on a 10% LB 1.0% agarose pad containing 0.375 µg/mL FM 4-64 and 0.025 µg/mL DAPI.

De novo assembly for OT59 and O4M1A

Paired-end Illumina reads generated on Illumina GAIIx were used for the de novo assembly of the genomes of O4M1A and OT59. The quality-trimmed reads were assembled de novo using the assembler Velvet (v1.2.07) with hash length parameter of 25 (PMID:18349386). This resulted in an assembly of 5.4Mb for O4M1A comprised of 5793 contigs with an N50 of 1.8 Kb. The OT59 genome was assembled into 1484 contigs corresponding to 5.26 Mb with an N50 of 17.5Kb. Similar to earlier de novo assembly approaches (PMID: 20544019), the assembled genome was annotated using the RAST server (PMID:18261238) with default parameters. This resulted in 4231 predicted coding sequences in O4M1A and 4530 predicted coding sequences in OT59, respectively.

2.8 Acknowledgements

We acknowledge the Government of Panama (ANAM, ARAP) for granting permission to make the collections of the corals for the isolation of strains OT59 and 04M1A. Financial support was provided by the National Institute of Health GM097509 (BSM, NB, PCD), GM094802 (PCD), AI095125 (KP, PCD), and GM098105 (BP) the Fogarty International Center's International Cooperative Biodiversity Groups program TW006634 (WHG, MG), MHM was supported by the Dutch Technology Foundation (STW), which is the applied-science division of The Netherlands Organisation for Scientific Research (NWO) and the Technology Programme of the Ministry of Economic Affairs (grant STW 10463). RDM acknowledges support from the KU Leuven Research

Council (grant GOA/011/2008). Instrumentation used in this study is supported by Bruker Therapeutic Discovery Mass Spectrometry Center and NIH GM S10RR029121 (PCD).

2.9 REFERENCES

1. Corre C & Challis GL (2009) New natural product biosynthetic chemistry discovered by genome mining. *Natural product reports* 26(8):977-986.

2. Kersten RD, Yang YL, Xu Y, Cimermancic P, Nam SJ, Fenical W, Fischbach MA, Moore BS, & Dorrestein PC (2011) A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nature chemical biology* 7(11):794-802.

3. de Bruijn I, de Kock MJ, Yang M, de Waard P, van Beek TA, & Raaijmakers JM (2007) Genome-based discovery, structure prediction and functional analysis of cyclic lipopeptide antibiotics in Pseudomonas species. *Molecular microbiology* 63(2):417-428.

4. Li MH, Ung PM, Zajkowski J, Garneau-Tsodikova S, & Sherman DH (2009) Automated genome mining for natural products. *BMC bioinformatics* 10:185.

5. Bode HB & Muller R (2005) The impact of bacterial genomics on natural product research. *Angew Chem Int Ed Engl* 44(42):6828-6846.

6. Maksimov MO, Pelczer I, & Link AJ (2012) Precursor-centric genome-mining approach for lasso peptide discovery. *Proceedings of the National Academy of Sciences of the United States of America* 109(38):15223-15228.

7. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, van der Voort M, Pogliano K, Gross H, Raaijmakers JM, Moore BS, Laskin J, Bandeira N, & Dorrestein PC (2012) Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences of the United States of America* 109(26):E1743-1752.

8. Shendure J & Ji H (2008) Next-generation DNA sequencing. *Nature biotechnology* 26(10):1135-1145.

9. Medigue C & Moszer I (2007) Annotation, comparison and databases for hundreds of bacterial genomes. *Research in microbiology* 158(10):724-736.

10. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, & Kyrpides NC (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic acids research* 40(Database issue):D115-122.

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

12. Finking R & Marahiel MA (2004) Biosynthesis of nonribosomal peptides1. *Annual review of microbiology* 58:453-488.

13. Donadio S, Monciardini P, & Sosio M (2007) Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. *Natural product reports* 24(5):1073-1109.

14. Caboche S, Pupin M, Leclere V, Fontaine A, Jacques P, & Kucherov G (2008) NORINE: a database of nonribosomal peptides. *Nucleic acids research* 36(Database issue):D326-331.

15. Mossialos D, Ochsner U, Baysse C, Chablain P, Pirnay JP, Koedam N, Budzikiewicz H, Fernandez DU, Schafer M, Ravel J, & Cornelis P (2002) Identification of new, conserved, non-ribosomal peptide synthetases from fluorescent pseudomonads involved in the biosynthesis of the siderophore pyoverdine. *Molecular microbiology* 45(6):1673-1685.

16. Rausch C, Weber T, Kohlbacher O, Wohlleben W, & Huson DH (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic acids research* 33(18):5799-5808.

17. Grunewald J & Marahiel MA (2006) Chemoenzymatic and template-directed synthesis of bioactive macrocyclic peptides. *Microbiology and molecular biology reviews : MMBR* 70(1):121-146.

18. Chen Y, Ntai I, Ju KS, Unger M, Zamdborg L, Robinson SJ, Doroghazi JR, Labeda DP, Metcalf WW, & Kelleher NL (2012) A proteomic survey of nonribosomal peptide and polyketide biosynthesis in actinobacteria. *Journal of proteome research* 11(1):85-94.

19. Rounge TB, Rohrlack T, Nederbragt AJ, Kristensen T, & Jakobsen KS (2009) A genome-wide analysis of nonribosomal peptide synthetase gene clusters and their peptides in a Planktothrix rubescens strain. *BMC genomics* 10:396.

20. Wang X, Chen H, Lee J, & Reilly PT (2012) Increasing the Trapping Mass Range to m/z = 10(9)-A Major Step Toward High Resolution Mass Analysis of Intact RNA, DNA and Viruses. *International journal of mass spectrometry* 328-329:28-35.

21. Denisov E, Damoc, E., Lange, O., Makarov, A. (2012) Orbitrap mass spectrometry with resolving powers above 1,000,000. *International journal of mass spectrometry* 325-327:80-85.

22. Watrous JD & Dorrestein PC (2011) Imaging mass spectrometry in microbiology. *Nature reviews. Microbiology* 9(9):683-694.

23. Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW, Eling WM, Hall N, Waters AP, Stunnenberg HG, & Mann M (2002) Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. *Nature* 419(6906):537-542.

24. Zaia J (2010) Mass spectrometry and glycomics. *Omics : a journal of integrative biology* 14(4):401-418.

25. Herring KD, Oppenheimer SR, & Caprioli RM (2007) Direct tissue analysis by matrix-assisted laser desorption ionization mass spectrometry: application to kidney biology. *Seminars in nephrology* 27(6):597-608.

26. Marvin LF, Roberts MA, & Fay LB (2003) Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry in clinical chemistry. *Clinica chimica acta; international journal of clinical chemistry* 337(1-2):11-21.

27. Claydon MA, Davey SN, Edwards-Jones V, & Gordon DB (1996) The rapid identification of intact microorganisms using mass spectrometry. *Nature biotechnology* 14(11):1584-1586.

28. Zhang Z, Cooks RG, & Ouyang Z (2012) Paper spray: a simple and efficient means of analysis of different contaminants in foodstuffs. *The Analyst* 137(11):2556-2558.

29. Van Berkel GJ, Pasilis SP, & Ovchinnikova O (2008) Established and emerging atmospheric pressure surface sampling/ionization techniques for mass spectrometry. *Journal of mass spectrometry : JMS* 43(9):1161-1180.

30. Ferguson CN, Benchaar SA, Miao Z, Loo JA, & Chen H (2011) Direct ionization of large proteins and protein complexes by desorption electrospray ionization-mass spectrometry. *Analytical chemistry* 83(17):6468-6473.

31. Roach PJ, Laskin J, & Laskin A (2010) Molecular characterization of organic aerosols using nanospray-desorption/electrospray ionization-mass spectrometry. *Analytical chemistry* 82(19):7979-7986.

32. Takats Z, Wiseman JM, Gologan B, & Cooks RG (2004) Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science* 306(5695):471-473.

33. Song Y, Talaty N, Tao WA, Pan Z, & Cooks RG (2007) Rapid ambient mass spectrometric profiling of intact, untreated bacteria using desorption electrospray ionization. *Chem Commun (Camb)* (1):61-63.

34. Pierce CY, Barr JR, Cody RB, Massung RF, Woolfitt AR, Moura H, Thompson HA, & Fernandez FM (2007) Ambient generation of fatty acid methyl ester ions from bacterial whole cells by direct analysis in real time (DART) mass spectrometry. *Chem Commun (Camb)* (8):807-809.

35. Alberici RM, Simas RC, Sanvido GB, Romao W, Lalli PM, Benassi M, Cunha IB, & Eberlin MN (2010) Ambient mass spectrometry: bringing MS into the "real world". *Analytical and bioanalytical chemistry* 398(1):265-294.

36. Demirev PA & Fenselau C (2008) Mass spectrometry for rapid characterization of microorganisms. *Annu Rev Anal Chem (Palo Alto Calif)* 1:71-93.

37. Meetani MA, Shin YS, Zhang S, Mayer R, & Basile F (2007) Desorption electrospray ionization mass spectrometry of intact bacteria. *Journal of mass spectrometry : JMS* 42(9):1186-1193.

38. Takats Z, Wiseman JM, & Cooks RG (2005) Ambient mass spectrometry using desorption electrospray ionization (DESI): instrumentation, mechanisms and applications in forensics, chemistry, and biology. *Journal of mass spectrometry : JMS* 40(10):1261-1275.

39. Keil A, Talaty N, Janfelt C, Noll RJ, Gao L, Ouyang Z, & Cooks RG (2007) Ambient mass spectrometry with a handheld mass spectrometer at high pressure. *Analytical chemistry* 79(20):7734-7739.

40. Soparawalla S, Tadjimukhamedov FK, Wiley JS, Ouyang Z, & Cooks RG (2011) In situ analysis of agrochemical residues on fruit using ambient ionization on a handheld mass spectrometer. *The Analyst* 136(21):4392-4396.

41. Roongsawang N, Washio K, & Morikawa M (2010) Diversity of nonribosomal Peptide synthetases involved in the biosynthesis of lipopeptide biosurfactants. *International journal of molecular sciences* 12(1):141-172.

42. Challis GL, Ravel J, & Townsend CA (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chemistry & biology* 7(3):211-224.

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

44. Rottig M, Medema MH, Blin K, Weber T, Rausch C, & Kohlbacher O (2011) NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nucleic acids research* 39(Web Server issue):W362-367.

45. Prieto C, Garcia-Estrada C, Lorenzana D, & Martin JF (2012) NRPSsp: non-ribosomal peptide synthase substrate predictor. *Bioinformatics* 28(3):426-427.

46. Gross H & Loper JE (2009) Genomics of secondary metabolite production by Pseudomonas spp. *Natural product reports* 26(11):1408-1446.

47. Sansinenea E & Ortiz A (2011) Secondary metabolites of soil Bacillus spp. *Biotechnology letters* 33(8):1523-1538.

48. Raaijmakers JM, De Bruijn I, Nybroe O, & Ongena M (2010) Natural functions of lipopeptides from Bacillus and Pseudomonas: more than surfactants and antibiotics. *FEMS microbiology reviews* 34(6):1037-1062.

49. Ongena M & Jacques P (2008) Bacillus lipopeptides: versatile weapons for plant disease biocontrol. *Trends in microbiology* 16(3):115-125.

50. Speitling M, Smetanina OF, Kuznetsova TA, & Laatsch H (2007) Bromoalterochromides A and A', unprecedented chromopeptides from a marine Pseudoalteromonas maricaloris strain KMM 636T. *The Journal of antibiotics* 60(1):36-42.

51. Kalinovskaya NI, Dmitrenok AS, Kuznetsova TA, Frolova GM, Christen R, Laatsch H, Alexeeva YV, & Ivanova EP (2008) "Pseudoalteromonas januaria" SUT 11 as the source of rare lipodepsipeptides. *Current microbiology* 56(3):199-207.

52. Xie BB, Shu YL, Qin QL, Rong JC, Zhang XY, Chen XL, Shi M, He HL, Zhou BC, & Zhang YZ (2012) Genome sequences of type strains of seven species of the marine bacterium Pseudoalteromonas. *Journal of bacteriology* 194(10):2746-2747.

53. Smoot ME, Ono K, Ruscheinski J, Wang PL, & Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27(3):431-432.

54. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, & Bader GD (2007) Integration of biological networks and gene expression data using Cytoscape. *Nature protocols* 2(10):2366-2382.

55. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, & Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 13(11):2498-2504.

56. Moree WJ, Phelan VV, Wu CH, Bandeira N, Cornett DS, Duggan BM, & Dorrestein PC (2012) Interkingdom metabolic transformations captured by microbial imaging mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* 109(34):13811-13816.

57. Bachmann BO & Ravel J (2009) Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods in enzymology* 458:181-217.

58. Rokni-Zadeh H, Li W, Sanchez-Rodriguez A, Sinnaeve D, Rozenski J, Martins JC, & De Mot R (2012) Genetic and functional characterization of cyclic lipopeptide white-line-inducing principle (WLIP) production by rice rhizosphere isolate Pseudomonas putida RW10S2. *Applied and environmental microbiology* 78(14):4826-4834.

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

60. Edwards DJ, Marquez BL, Nogle LM, McPhail K, Goeger DE, Roberts MA, & Gerwick WH (2004) Structure and biosynthesis of the jamaicamides, new mixed polyketide-peptide neurotoxins from the marine cyanobacterium Lyngbya majuscula. *Chemistry & biology* 11(6):817-833.

Chapter 2, in full, has been submitted for publication of the material as it may appear in the Proceedings of the National Academy of Sciences, 2013, Nguyen, Don D.; Wu, Cheng-Hsuan; Moree, Wilna J; Lamsa, Anne; Medema, Marnix H.; Zhao, Xiling; Gavilan, Ronnie G.; Aparicio, Marystella; Atencio, Librada; Jackson, Chanaye; Ballesteros, Javier; Sanchez, Joel; Watrous, Jeramie D.; Phelan, Vanessa V.; van de Wiel, Corine; Kersten, Roland D.; Mehnaz, Samina; De Mot, René; Shank, Elizabeth A.; Charusanti, Pep; Nagarajan, Harish; Duggan, Brendan M.; Moore, Bradley S.; Bandeira, Nuno. Palsson, Bernhard Ø.; Pogliano, Kit; Gutiérrez, Marcelino; Dorrestein, Pieter C. The thesis author as well as Cheng-Hsuan Wu and Wilna J. Moree were the primary investigators and authors of this paper.

Chapter 3

Future Applications of MALDI-IMS and Molecular Networking

3.1 Summary

As illustrated in Chapter 1, specialized metabolites have proven beneficial to human health (1, 2). Through the discovery of penicillin, the pharmaceutical industry has developed many drugs and drug leads based on specialized metabolites ranging from antibiotics to analgesics. There has, however, been a decline in the development of specialized metabolites due to challenges such as the rediscovery of particular compounds, working with unculturable or unrenewable sources of specialized metabolites, difficulties in working with mixtures of compounds, and the difficulties involved with the dereplication of known compounds to leave only the desirable and bioactive novel molecules (3). Chapter 1 then introduces mass spectrometry based tools that have been developed to aid in the study of specialized metabolites. MALDI-IMS of agar based microbial interactions is a relatively quick way to determine the molecular weight of specialized metabolites involved in metabolic exchange between two or more organisms. Molecular networking allows the visualization of large tandem MS data sets that shows the similarities and differences between molecular ions. This gives some structural information to the observable molecules and very quickly associates which molecules are related to one another and which are not. Chapter 2 demonstrates that MALDI-IMS can then be supplemented by molecular networking to gain insights into the structural features of bioactive specialized metabolites, to determine if there are related compounds to these bioactive substances, but also introduces the concept that families of molecules can be associated to each other based on mass spectral fragmentation data. This is expanded upon, where publicly accessible genomes were used to discover the biosynthetic gene cluster for this family of molecules responsible for the inhibition of *B. subtilis* 3610 by the genus *Pseudoalteromonas*. The concept of connecting molecular families to their gene cluster families came to fruition and was also applied to studying the nonribosomal peptides, not just of the genus *Pseudoalteormonas*, but also of sixty different species of bacilli and pseudomonads. Below are two

descriptions of alternative uses of these mass spectrometry techniques and methodologies that are currently taking place.

3.2 Molecular Networking of Hundreds of *Pseudoalteromonas*

Pseudoalteromonas is a genus of marine gram-negative γ-proteobacteria. These marine microbes are found globally, play an ecological role ranging from influencing biofilm formation of microbes to affecting the germination of invertebrate and alga species to producing a range of high and low molecular weight compounds that have antibacterial, antifungal, anti-fouling, and algacidal effects (4). While circumnavigating the world, researchers on the Galathea 3 vessel isolated over one hundred *Pseudoalteromonas* that showed antibacterial activity against *Vibrio anguillarum* (5). Additionally, *Pseudoalteromonas* have been observed in numerous dairy products, such as cheese, that are colonized with complex microbial ecosystems and can have an impact on the appearance, odor, flavor, and texture of these products (6). Using the techniques described in Chapter 2, we are currently working with researchers from Denmark on mapping the molecular network of the hundreds of *Pseudoalteromonas* collected globally to compare to the coral associated strains OT59 and 04M1A and to the cheese associated *Pseudoalteromonas*. The goals of networking such a large number of *Pseudoalteromonas* strains would be to examine the metabolic similarities and differences of different species of *Pseudoalteromonas*, determine whether geography and specific location plays a role in metabolic output, and to connect the observable molecular families to their gene cluster families with available genomic information. Additionally, several of the molecules produced by strains collected on the Galathea expedition have already been characterized, complete with NMR structural elucidation, and could serve as good starting points for dereplication and determination of molecular families produced by the *Pseudoalteromonas*. In this regards, molecular networking allows us to very quickly examine a large number of organisms, their produced specialized metabolites, connect these metabolites to their biosynthetic origins, and use

existing structural information from known molecules to seed the network and extend this structural information to related nodes within a cluster, which may lead to the discovery of entirely novel compounds with potential therapeutic applications.

3.3 Crop Protection in Algal Biofuels

Lastly, the techniques presented in this thesis are applicable in industrially relevant areas beyond the characterization of specialized metabolites for therapeutic discovery. In collaboration with Susan Golden's Lab in the Division of Biological Sciences at UCSD, we are in the process of demonstrating that MALDI-IMS can be used to examine molecules involved in the grazing activity of predators on photosynthetic microorganisms. Fossil fuels from coal, petroleum, and natural gas make up the largest percentage of the world's energy supply. The demand for fossil based energy is increasing while the reserves for these energy sources are being depleted or becoming too expensive to recover. In addition, continual usage of fossil fuels brings with it increasing CO_2 concentrations in the atmosphere and greenhouse gas-mediated climate change (7). Algal biofuel serves as one possible alternative to fossil fuels, but must first overcome several obstacles before becoming market competitive (8-10). Outdoor raceway pond systems are imperative to cost effective and economical algal biofuel production. However, susceptibility to open conditions, including contamination by other algae species, bacteria, fungi, as well as predators, can lead to the destruction of entire ponds resulting in the loss of thousands of dollars' worth of fuel crops and co-products (8, 11). To better understand the mechanisms of grazing on photosynthetic organisms, Simkovsky and co-workers have established a model system to study the grazing of amoeba HGG1 on the photosynthetic cyanobacteria, *Synechococcus elongatus* PCC 7942 (12). Additionally, they have identified mutations impairing O-antigen synthesis or transport that confer resistance to amoeba grazing. Together we have examined the interaction of HGG1 with wild type *S. elongatus* as well as the *S. elongatus* Wzm O-antigen transporter mutant, 7D3, by MALDI-IMS. Thus far, we

have observed ions that are specific to HGG1, as well as ions that appear only while HGG1 feeds on *S. elongatus*. The remaining steps required will be to identify and structurally characterize these specific signals. Once these signals have been characterized, we hope to expand these studies to determine if the observed ions can be generalized to interactions between other amoeba and cyanobacteria, to test these mass spectrometry based techniques and methodologies in the field on legitimate outdoor algae ponds, and to develop a mass spectrometry workflow for the detection of biological contaminants. If markers for biological contaminants can consistently be identified, then this mass spectrometry based workflow could be used as an early detection system that would allow adequate time for countermeasures to keep algae ponds from crashing, thus aiding in crop protection of algae biofuels.

3.4 REFERENCES

1. Cragg GM & Newman DJ (2013) Natural products: A continuing source of novel drug leads. *Biochim Biophys Acta* 1830(6):3670-3695.

2. Davies J (2011) How to discover new antibiotics: harvesting the parvome. *Curr Opin Chem Biol* 15(1):5-10.

3. Roemer T, Xu D, Singh SB, Parish CA, Harris G, Wang H, Davies JE, & Bills GF (2011) Confronting the challenges of natural product-based antifungal discovery. *Chemistry & biology* 18(2):148-164.

4. Bowman JP (2007) Bioactive compound synthetic capacity and ecological significance of marine bacterial genus pseudoalteromonas. *Marine drugs* 5(4):220-241.

5. Vynne NG, Mansson M, Nielsen KF, & Gram L (2011) Bioactivity, chemical profiling, and 16S rRNA-based phylogeny of Pseudoalteromonas strains collected on a global research cruise. *Mar Biotechnol (NY)* 13(6):1062-1073.

6. Ogier JC, Lafarge V, Girard V, Rault A, Maladen V, Gruss A, Leveau JY, & Delacroix-Buchet A (2004) Molecular fingerprinting of dairy microbial ecosystems by use of temporal temperature and denaturing gradient gel electrophoresis. *Applied and environmental microbiology* 70(9):5628-5643.

7. Brennan L & Owende P (2010) Biofuels from microalgae—A review of technologies for production, processing, and extractions of biofuels and co-products. *Renewable and Sustainable Energy Reviews* 14(2):557-577.

8. Hannon M, Gimpel J, Tran M, Rasala B, & Mayfield S (2010) Biofuels from algae: challenges and potential. *Biofuels* 1(5):763-784.

9. Hunter P (2010) The tide turns towards microalgae. Current research aims to produce traditional biofuels from algae, but their potential to generate sustainable energy might be even greater and more 'natural'. *EMBO reports* 11(8):583-586.

10. Gouveia L & Oliveira AC (2009) Microalgae as a raw material for biofuels production. *Journal of industrial microbiology & biotechnology* 36(2):269-274.

11. Hase R, Oikawa H, Sasao C, Morita M, & Watanabe Y (2000) Photosynthetic production of microalgal biomass in a raceway system under greenhouse conditions in Sendai city. *Journal of Bioscience and Bioengineering* 89(2):157-163.

12. Simkovsky R, Daniels EF, Tang K, Huynh SC, Golden SS, & Brahamsha B (2012) Impairment of O-antigen production confers resistance to grazing in a model amoeba-cyanobacterium predator-prey system. *Proceedings of the National Academy of Sciences of the United States of America* 109(41):16678-16683.