UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

METABOLOMIC AND PHYLOGENETIC ANALYSES TO BETTER INFORM THE CONSTRUCTION OF NATURAL PRODUCTS DISCOVERY LIBRARIES

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

VICTORIA ANDERSON
Norman, Oklahoma
2021

METABOLOMIC AND PHYLOGENETIC ANALYSES TO BETTER INFORM THE CONSTRUCTION OF NATURAL PRODUCTS DISCOVERY LIBRARIES

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY

Dr. Robert H. Cichewicz, Chair

Dr. Charles V. Rice

Dr. Susan J. Schroeder

Dr. Si Wu

Dr. Bradley S. Stevenson

Acknowledgements

This work would have been impossible without the support and assistance of many individuals. First among these is Dr. Robert Cichewicz, I thank him for his support and encouragement. I have grown immeasurably because Boss pushed and challenged me making me a more a more determined, persistent, and insightful scientist.

I would also like to thank my committee Dr. Charles Rice, Dr. Susan Schroeder, Dr. Si Wu, and Dr. Bradley Stevenson for their time and support over the years. Additionally, I would like to thank Dr. Fares Najar for his endless optimism and assistance. Even my silliest questions were answered with a smile. I would also like to thank Dr. Laura-Isobel McCall for introducing me to metabolomics and for her assistance collecting data that was essential for my gradate research.

I would not be where I am today without my NPDG family. I have loved this group of people from the very first and cannot imagine completing my degree anywhere else. I owe so much to them for sharing their knowledge with me and their endless patience and encouragement. The friendship, training, humor, knowledge, adventures, tea, and frustrations that we shared are very dear to me. Specifically, I would like to thank former NPDG members Yun-seo, Wentao, Lin, Hongyan, Adam and Shengxin. I would also like to thank Karen and Jin Woo for being the pillars of NPDG with me during this past year.

My projects would not have been possible without the Citizen Science Soil Collection program, so I issue a heartfelt thank you to Candace for coordinating this program as well as the Zoo team members who have isolated my fungi for me. I would also like to thank the many undergraduate students I have mentored during my time in NPDG: their enthusiasm and insights were a constant inspiration to me.

Finally, I would like to thank my family for their endless support and encouragement. Through the highs and lows of graduate school my family has been a constant pillar of support. Their support and love have been unshakable, and I thank them from the bottom of my heart.

Table of Contents

List of Figures

Abstract

Natural product discovery programs rely on diverse libraries of organisms to provide access to a diverse pool of compounds. The quality and chemical diversity of these libraries are therefore of utmost importance to the discovery efforts. Despite the heavy dependence on these libraries, very little work has been done to determine the best way to construct these libraries. Conventional approaches to library building assumes that the larger the library is the more diverse it will be and the higher the probability of finding novel bioactive compounds. However, these large libraries are unwieldly to both manage and screen. The field of drug discovery would benefit from tools that were able to assess the diversity and to direct future directions for library construction.

Metabolomics is a sophisticated field that attempts to quantify the entire metabolic output of an organism. The development of new metabolomics techniques has allowed the expansion of metabolomics into many different fields, including drug discovery. Drug discovery libraries are at the very heart of discovery efforts and the application of metabolomic tools to drug discovery is an ideal way to investigate the diversity of drug discovery libraries. In an effort to address the lack of concrete guidance about the construction of natural products libraries, we have tested two aspects that are very important to the design of a drug discovery library: 1) the sampling depth required for maximal chemical diversity, and 2) the collection strategy that results in improves the library chemical diversity

While the normal course of drug discovery involves screening extract libraries and then purifying compounds from active extracts based on bioassay data, this does not address the concern of the library's chemical diversity. Using metabolomic tools instead of those typical in dis-

covery labs may be a more appropriate strategy for answering basic questions about the construction of libraries. The combination of metabolomic tools with phylogenetic analysis allows for an indirect extrapolation of chemical diversity in the library while making use of commonly used tools that are used in library building. In Chapter 3, *Alternaria* was used to showcase a method to build a library that encapsulates maximum levels of chemical diversity and suggests a strategy to expand into previously unavailable chemistry. This study revealed that chemical diversity is distributed within a genus in an unbalanced manner. Predictive analysis revealed that 99% of *Alternaria* chemical features would be detected if the collection consisted of 195 isolates. Feature and scaffold accumulation analysis allows an observable level of the chemistry expected from a group of organisms as well as identifying the contribution of new chemistry afforded by including more individuals and leads to building a comprehensive library. These methods can improve the chemical diversity of libraries that are the backbone of natural-product-based drug discovery.

Drug discovery efforts have in the past have emphasized the exploration of diverse environments in the search for novel bioactive compounds. While this strategy has provided new compounds, there is little evidence that new compounds cannot also be locally sourced. The use of metabolomics tools has allowed the examination and comparison of the chemical diversity of groups of organisms that were present in accessible as well as inaccessible environments. The use of traditional metabolomics methods were used to compare the chemical profile of extracts directly, while ecological methods were used to examine which scaffolds were present in both populations or unique to one or the other population. In chapter 4, we use three species of *Penicillium* that were present in both the sediments of Lake Michigan and soils from the states surrounding this lake to investigate if the origin of the organism confers different chemical production. The metabolic profile of each species showed marked overlap despite the different origin

of organisms. The analysis of scaffolds revealed that between 78% and 83% of total scaffolds were shared depending upon species. The community composition of these environments was examined to determine how much overlap is present and while the two environments do share some diversity, there is a community of fungi that is unique to the single environment. The results of this experiment suggest that the value of sampling diverse environments will be predominantly in those organisms that are unique to each space.

Chapter 1: An Introduction to Metabolomics and its Uses

1.1 Metabolomics: a definition

Metabolomics is the study of the complete or near complete metabolic output of an organism,[1-2] and is therefore characterized by detecting analytes at low concentrations in an extremely complex matrix.[3] Metabolomics is the youngest of the "omics", following in the footsteps of genomics, transcriptomics, and proteomics.[4] An "omics" field is one which takes a comprehensive view on the field of study.[3-4] These metabolites are the summation of the genome, transcriptome, and proteome.[2] Metabolites are the end products of many biosynthetic pathways; hence metabolomics provides a downstream view of these cellular processes.[5]

As such, metabolomics can be used as a way to directly study the phenotype, but also has larger opportunities and promise.[1, 5] Many metabolomics experiments focus on those compounds involved with central carbon metabolism, this is likely due to the central importance of this metabolism with downstream functions such as synthesis of nucleotides.[6] Metabolomics may be able to provide answers to questions that reach across species boundaries than would genomics because while the gene-structure may be very different from species to species, basic primary metabolites are conserved and in most cases have similar function.[7]

In contrast to the older "omics", metabolomics provides a direct snapshot of the influences of genetics, environment, stage of development, etc. because it is most related to the phenotype of the organism.[2-4, 8] A metabolomics study therefore shows what is actually happening chemically in an organism as it reacts with its environment or in response to stimuli.[2-3]

The complete collection of all small molecules (less than 1,000 Da) of an organism is referred to as the metabolome.[6] A wide range of techniques are currently in use for the analysis of the metabolome. Because of the complexity of these samples (wide range both of compound type and concentrations) the selection of a technique must be made by considering the focus of the experiment.[6] The successful application of metabolomics study therefore requires the use of sensitive and efficient instrumentation. The diversity of the metabolome means that metabolomics can be applied to a wide range of fields.

Metabolomics likely took longer to reach the prominence of genomics and proteomics because the instrumentation, data analytics, and computing power necessary to analyze the massive amount of data to characterize the entirety of the detectible metabolome were slow to be developed.[4] This is understandable considering the many challenges inherent in metabolomics studies: the vast number of metabolites produced, the variability of production of metabolites, the lack of sufficient reference spectra to allow identification of metabolites, and variation in experimental conditions that can cause inconsistencies in the data (instrument error or similar not related to biology).[8] Many of these challenges are due to the complexity of organisms, and others are due to limitations in the tools required to fully characterize that complexity.[8]

1.2 Targeted vs Untargeted metabolomics

Metabolomics studies are typically divided into either targeted (includes metabolite profiling) or untargeted (metabolic fingerprinting) analyses.[1] A targeted metabolomics study involves the investigation and quantification of a one or several target metabolites to the exclusion of all other signals.[1, 9-10] This method is the most straightforward of the metabolomics experiments and is often used for hypothesis driven analyses.[2] Targeted analyses are often the validation of an untargeted study that identifies the metabolites of interest.[10-12]

This type of analysis is appropriate if the compound of interest is known and can be expected in samples at a detectable level given appropriate sample preparation techniques.[1] However, targeted metabolomics analysis largely ignores signals which are not associated with the target compound, which leaves the vast majority of chemistry unstudied.[1] A variant of targeted metabolomics is metabolite profiling. This technique quantifies the level of a suite of metabolites (often related to a particular metabolic pathway).[1] The assumption of metabolite profiling is that there would be an observable difference in the metabolite levels despite a non-observable phenotypic change in response to stimulus.[1]

1.2.1 Targeted Metabolomics

Targeted metabolomics can also be used to observe the specific consequence of a disease, condition, or environmental exposure on the metabolites of interest. Depending upon the experiment these metabolites could be involved in central metabolism[11, 13], or could be more targeted[10]. Targeted metabolomics can also be used in the study and diagnosis of disease states because they involve the quantification of a known metabolite or metabolites. A targeted study of 19 neurotransmitters in patient plasma was conducted with the intention of developing an objective diagnostic for major depressive disorder by identifying diagnostic biomarkers for the condition.[10] While more validation is required, the model was able to distinguish between plasma of a patient with major depressive disorder and bipolar disorder.[10]

The identification of plasma based biomarkers of early stage gastric cancer would lead to greatly improved prognosis for patients.[13] In the course of a targeted metabolomics study, both phenylalanine and tryptophan and their associated pathways were found to be perturbed in the serum of patients of gastric cancer.[13] While further work is necessary with larger sample sizes to confirm that these are associated with the disease state and are similarly perturbed when analyzed by other hands, this finding is useful in the context of understanding the disease.[13]

Targeted metabolomics analysis can also be used in environmental contexts. This type of analysis is useful to track the exposure of organisms overtime as well as develop risk assessments associated with contamination of aquatic environments with pharmaceuticals or other environmental pollutant.[11] A targeted metabolomics method was developed that was able to detect the selected pharmaceuticals concentrations as low as ng/L for river water or ng/g in freshwater crustaceans which are often used in biomonitoring studies.[11] In response to the most toxic pharmaceutical tested, there were significant differences in 19 of 29 metabolites, and the pattern of fold-change in metabolites differed depending upon the pharmaceutical tested suggesting that different pathways were disrupted.[11]

1.2.2 Untargeted Metabolomics

In contrast, untargeted metabolomics seeks to detect patterns in feature intensity that correlate with experimental conditions.[14] Untargeted metabolomics experiments are often considered to be exploratory and hypothesis generating.[2] For instance, extensive untargeted metabolomics studies have shown that many neurotransmitters are perturbed in patients with major depressive disorder.**[10]** These studies informed the targeted study that has suggested some diagnostic criteria for patients with this disorder which is currently being diagnosed based on subjective interpretation of symptoms.[10] The identification of biomarkers often begin with untargeted analyses which are then validated by targeted analysis.

While targeted analyses ignore all signal that are not associated with the target compounds, untargeted metabolomics experiments considers all metabolites above the limit of detection, with the goal of identifying which metabolites are altered by the experimental conditions.[9] This technique is often referred to as metabolic fingerprinting because it observes the overall profile of metabolites present in a sample without trying to quantify or identify all of these signals.[1] This technique is often applied to the discovery of metabolites that are markers of disease states.[1]

Untargeted analyses, such as fingerprinting, provide a more complete picture of the metabolic output of the organism.[15] Metabolic fingerprinting is a technique which can and has been used to determine the influence of a change on organisms. This change could be something as organic as developmental changes as an organism matures or the differences between species, but could also be used to probe the differences caused by environmental changes or the metabolic consequences of a disease.[8] It is a technique which is easily adapted to a high throughput system allowing the rapid analysis of the many small samples.[8]

Untargeted methods can be applied to the analysis of environmental and ecological samples. The growing conditions (such as altitude, climate, soil, temperature, harvest time, etc.) have an influence on the metabolomic profile of plants.[16-17] Untargeted metabolomics can be used to investigate taxonomic plasticity within loosely described phylogenies, such as those found in several species of coffee.[15] The analysis of leaf samples of 9 different species revealed that despite having a unique metabolic profile associated with each species, that all species varied similarly when samples were taken at different times throughout the growing season.[15] This suggests that identification of species might be possible with only a leaf sample to analyze. This could potentially indicate that common pathways are responsible for the seasonal changes observed.

Untargeted metabolomics experiments have been used to determine the ecotoxicological repercussions of the application of common fungicidal compounds on the metabolic profile of earthworms.[18] Despite a lack of significant change in body weight of worms, PCA showed significant differences in the metabolomes of worms exposed to every tested dose of fungicide.[18] This is one of the most valuable aspects of metabolomics: detecting changes before the stimuli causes visible change in the phenotype of the organism.[18] Further, the two enantiomers of metalaxyl showed different changes in the metabolome of treated worms: metalaxyl-M (97% R-enantiomer) was found to have a lesser impact on the earthworm urea cycle than did the S-enantiomer alone.[18]

Untargeted metabolomics can also be applied to the analysis of healthy vs diseased samples.[10, 19] Because there are quite distinct metabolic differences between healthy cells and malignant breast cancer cells, metabolomics is a potentially valuable tool to differentiate between cell types and provide early diagnostic, which could save lives.[19] The use of untargeted NMR-based metabolomics revealed a metabolic profile that showed near-complete separation between samples from early patients as opposed to patients with metastatic breast cancer.[19] However, examination of a larger cohort is necessary to validate these results.[19] In an additional study, untargeted metabolomic fingerprinting was able to distinguish between clinical strains of Bacillus cereus and non-virulent laboratory strains.[20] This is important in the profiling of this organism because there were no observable genetic differences to predict the pathotype of strains.[20]

While both targeted and untargeted metabolomics have their individual strengths and weaknesses, the most powerful use of metabolomics is to combine the quantitative aspects of targeted metabolomics and the wide qualitative aspects of untargeted metabolomics, thereby identifying previously unknown metabolites as well as tracking the level of such metabolites in response to given experimental conditions.[21] This is the case for the investigation of neurotransmitters associated with major depressive disorder[10].

1.3 Analytical tools used in Metabolomics

Metabolomics can make use of a number of analytical tools, but the most popular are NMR[16, 18-20] and mass spectrometry (MS)[5-6, 15, 22-23]. NMR is a tantalizing option for metabolomics analysis because it is a non-destructive technique, which will allow the reuse of sample.[1, 21] Additionally, NMR-based metabolomics requires minimal sample preparation of biofluids.[3, 8, 19] NMR is a relatively quick technique which is particularly suited for analysis of amino acids and carbohydrates.[18] However, the interpretation of the NMR spectra of complex mixtures is very difficult and less appropriate for metabolites at low abundance.[1, 11, 17, 21] Despite this, NMR-based metabolomics have been used to investigate the metabolic consequence of environmental contamination by fungicides[18], the differential effect of environmental conditions on green tea leaves[16], identification of pathogenicity in a bacterial species[20], and the differentiation of cancer stage in plasma samples[19].

1.3.1 Mass Spectrometry

MS is a sensitive technique appropriate to detect a wide range of metabolites, even at low concentrations.[1] MS works by generating ions which are then separated based on their mass-to-charge ratio (*m/z*) and detected by a mass detector.[24] While the direct injection into the mass spectrometer is a rapid method, which can be used for both metabolite fingerprinting or profiling, the complexity of the data is problematic because ionization efficiency is reduced and there is no way to distinguish between isomers which share a mass.[1, 21] To improve the accuracy of data, MS is often best utilized when coupled to a variety of front-ends which provides a separation of compounds prior to mass analysis.[1] The inclusion of a separation step minimizes situations where isobaric metabolites are co-eluted and detected by the MS jointly.[9] This allows for the deconvolution and separation of different metabolites which share a mass.[21]This step is important because metabolites are not generally sufficiently unique to be useful in a direct injection experiment.[8]

Hyphenated MS techniques detect a large number of metabolite "peaks" which are referred to as "features" and are defined by the unique combination of retention time and mass-to-charge *(m/z)* ratio.[25] The term feature is used in preference of compound or metabolite in refereeing to MS-based signals because in the course of ionization each compound may produce a number of different ions: adducts, dehydrations, dimers, dimers of adducts etc.[26] A mass analyzer which is capable of MS^2 or MS^n is the most helpful in metabolomics studies because it provides some structural information based on the loss of characteristic mass losses or more largely characteristic fragmentation patterns.[21] But despite this, most features cannot be identified from MS-based metabolomics experiments without additional and extensive experimentation.[25] The lack of annotation of LCMS features is a major road block in the interpretation of metabolomics data.[14, 25, 27]

MS is a valuable tool for metabolomics, not the least because of the adaptability of front end systems.[3] Hyphenated MS is a marked improvement on direct injection because it allows the separation based on polarity in addition to mass thus simplifying the composition of the mobile phase when it arrives at the ionization source limiting co-suppression.[1, 14, 28] While there is no one separation technique that will be uniformly acceptable for all metabolomics studies, the consideration of the goals of the experiment and the available technology will allow allows an almost infinite adaptability for separation. Hyphenated MS techniques include capillary electrophoresis mass spectrometry (CEMS), gas chromatography mass spectrometry (GCMS), and liquid chromatography mass spectrometry (LCMS). These techniques are variably appropriate for the analysis of a wide range of analytes from a wide range of samples.

1.3.1.1 Hyphenated MS techniques: Capillary Electrophoresis Mass Spectrometry

One possible front-ends to supply a separation prior to MS analysis is capillary electrophoresis (CE). CE is a technique that separates by their size to charge ratio, making it appropriate for the analysis of polar and charged compounds[6, 12], thus it is a good candidate for the analysis of many biomolecules.[29] Depending upon the coating of the capillary, CE can be used to separate cationic or anionic molecules.[6] CEMS instruments are occasionally built in house, which makes for creating the standardization that is required for metabolomics analyses difficult.[3] This has most likely contributed to the impression in the community that CE is not sufficiently reproducible and sensitive to be useful when coupled to MS in the context of metabolomics experiments.[12] There have however been some advances that suggest that CEMS could be a contender in the metabolomics field.[12]

CE can separate compounds in an aqueous system, making it a valuable tool for the analysis of bio-fluids such as urine.[3] There is additionally minimal sample preparation and is a technique that can be adapted to use minimal sample volumes (nL-range in some cases) .[3] In some cases CE able to analyze the contents of a single cell.[29] These characteristics prompted tests to discover the utility of this method to the field of metabolomics. This method was tested by analyzing plasma samples that had been spiked with a series of isotope labeled biomarkers at several concentrations.[12] This method was able to detect the differences in samples which were characterized by presence/absence differences as well as differences in concentration of spiked in biomarkers.[12]

1.3.1.2 Hyphenated MS techniques: Gas Chromatography Mass Spectrometry

Gas chromatography (GC) is another potential front end which provides chromatographic separation prior to MS analysis.[30] GCMS/MS is appropriate for the analysis of biological samples and using MS databases can provide putative identification of carbohydrates, alcohols, amino acids, organic acids, and fatty acids.[30-31] However, GC occasionally requires derivatization (a process that converts less volatile compounds to compounds that will be detected in the gaseous state) prior to separation.[6, 31] Depending upon the compounds of interest the derivatization will be different.[6] GC is not appropriate for compounds that are not stable at high temperatures.[6] GCMS has been used in a wide range of metabolomics experiments including, but not limited to, analysis of changes in potato metabolome depending upon cultivar and length of storage[31], taxonomy of fungal species which had previously been possible using traditional mycological techniques[30], pathway analysis of the response of components added to growth medium[22], and quantifying neurotransmitters associated with major depressive disorder[10].

1.3.1.3 Hyphenated MS techniques: Liquid Chromatography Mass Spectrometry

LC is one of the most common front ends for MS-based metabolomics studies.[5, 10-11, 13, 15, 23, 27, 32-33] LCMS has limit of detection that is significantly lower than those found in CEMS or GCMS metabolomics experiments allowing the detection of features at low concentrations.[6, 10] The combination of GCMS and LCMS/MS was used to quantify both high and low abundance neurotransmitters.[10] The use of LCMS/MS allowed detection of neurotransmitters at concentrations that were more than 2 orders of lower than those detected by GCMS in addition to overcoming challenges associated with each method individually.[10]

LCMS is considered to be particularly useful for the analysis of secondary metabolites—especially those from plants—because these classes of metabolites are typically semi-polar.[17] LCMS is not typically appropriate for compounds that are highly polar because they are difficult to separate on the most typical stationary phase columns.[3] LCMS is a good candidate for metabolomics experiments because of the ability to separate a wide range of metabolites of varying polarity, high reproducibility from run to run, and the simplicity of mobile phases which is ideal for electrospray ionization (ESI).[9] LCMS metabolomics has been used for a very wide range of analyses including: the analysis of environmental samples and model organisms to asses the inpact of contamination with pharmaceuticals[11], the analysis of plants to determine the genetic and environmental influences behing metabolic change[15, 32], investigation of metabolic consequences of human diseases such as cancer[5, 13] and depression[10], and investigate the secondary metabolite profiles of fungi[23, 27].

1.4 Metabolomics Workflow

The general steps of a metabolomics experiment based in mass spectrometry will include the following basic steps data generation which consists of sample collection, preparation, and sample analysis (including separation via LC and mass detection via MS); data processing which takes the raw data files and creates an aligned peak list; and data analysis which includes a variety of statistical techniques in addition to data interpretation.[21]

1.4.1 Data generation

1.4.1.1 Sample selection and preparation for metabolomics experiment

The Metabolomics Standards Initiative has a stated goal of creating a “minimum reporting standards” for metabolomics experiments.[34] As part of this work, this august group has made a variety of recommendations for the design of metabolomics experiments, including the recommendation that at least 3 but preferably 5 biological replicates be used for metabolomics analysis.[12, 18, 34] Despite these recommendations, a power analysis is the best way to determine the appropriate number of samples for the specific experimental conditions.[9] A very typical conclusion from a metabolomics experiment is that the study should be repeated with a larger number of samples to validate the results.[13, 19, 30-31]

After the number of samples is determined, the next step is sampling and preparing the samples. Because metabolite levels can differ based on when in the day they are sampled, consistency is important when sampling, so that sampling artefacts are not interpreted as differences between treatments.[15] Samples are often flash frozen in liquid nitrogen to prevent metabolic conversion which would skew the "metabolic snapshot".[17] Similarly, the extraction procedure should be conducted such that enzymatic activity is minimized, usually by processing sample while maintaining cold temperature.[11, 17, 25, 35] The best sample preparation steps should be fairly simple and be applied universally to all samples of the study.[21]

To correct for these internal differences that are common in biological samples, it is good practice to take several samples from each patient which when analyzed will give an idea of the base-line variation to be expected in that patient.[19] One of the challenges of metabolomics is the innate variability in biological systems: it is conceivable that significant metabolic variation as a result of patient age, gender, etc. .[19]

To eliminate signal drift, a quality control sample that was generated from pooling experimental samples should be run periodically with the experimental samples which were randomized to reduce any confounding effects.[15, 36] Additionally, it is wise to analyze data in randomized order to minimize the effect of subtle changes in the column or ionization source that occur as time progresses.[17] It is also possible to adjust the retention/migration time using internal standards.[6] The use of an internal standard also allows for more accurate quantification.[1, 9]

1.4.2 Data processing

The peak list that is the output of a metabolomics study is a huge amount of data that is then processed and transformed into manageable data which can be used to draw conclusions.[12, 19]The processing of data is critical to avoid false conclusions from metabolomics studies.[37] This often begins with the removal of signals from the data set that originate from the extraction process or media components allowing the focus of the analysis to be on the signals that are real and distinct based on the treatments.[37] Removal of spurious features is an important step in metabolomics analysis: typically features that are present in blanks are not likely to be experimentally useful, nor are feature that are present in only a single sample.[23] The addition of a MS2 fragmentation pattern to these details can be used to dereplicate features.[21]

Not all variations in feature intensity are related to the biological differences between samples.[33] Errors in sample preparation, sample injection, or instrument performance can cause differences in peak intensity.[33] Normalization is used to correct for these errors, and allow the biological differences to become apparent.[6, 16, 33] Normalization to the total ion current is the most typical type of normalization, in which the intensity of each peak is normalized to the total sum of intensities in each chromatograph.[21] Depending upon the questions of the study, the data may be normalized to the internal standard allowing more accurate quantification of features.[38] This technique is most successful if at least one isotope labeled internal standard is used.[38] An alternative method is to transform the data to indicate presence/absence rather than the intensity of each peak. This allows metabolites to have equal weight regardless of if they are high or low intensity peaks.[21] Ultimately, this data will be analyzed and visualized using both supervised and unsupervised methods to determine the overlap profiles and define boundaries between groups.[12]

1.4.3 Data analysis

Metabolomics has seen a renaissance in recent years as the computational tools required to efficiently handle the massive amounts of data that are produced in a typical metabolomics study are becoming more accessible.[39] The objective of many metabolomics studies is to determine how the metabolome of one population compares to different populations.[9, 13, 33] This is accomplished by adopting the null hypothesis that there is no variation of metabolomes between populations.[13,

[33] After conducting a series of tests using either univariate or multivariate methods, the null hypothesis may be rejected if the type I error is below the previously designated threshold (often at 5%).[13, 36] Many studies often then identify those metabolites that are responsible for the divergence of profiles.[10, 12-13, 18, 36]

Univariate methods are used in metabolomics when specific features are being compared between groups.[9, 36] Univariate techniques are particularly well suited to targeted metabolomics experiments because they consider the intensity of a specific feature across the sample set.[5, 10-11, 22] These methods often include ANOVA or student t-tests, and each feature must be tested by an ANOVA or t-test individually.[9] While it is manageable to test features individually up to a point, it becomes very difficult to examine the entire metabolome, which may be made up of thousands of features on an individual basis.[9] The complexity of metabolomics data often requires the use of multivariate statistical models to reveal the trends hidden in the masses of data.[13] This is where multivariate techniques come into shine. Multivariate techniques analyze multiple variables simultaneously, so are very well suited to analyzing metabolic profiling/fingerprinting experiments.[18, 22, 30, 32] Depending upon the objective of the experiment, a strategic combination of multivariate and univariate analyses can be beneficial to reducing the number of features that need to be analyzed.[36]

Both supervised (PLS-DA, RF, etc.) and unsupervised (PCA, PCoA, etc.) methods of multivariate analysis are important tools for the analysis of metabolomics studies. The primary difference between supervised and unsupervised methods is the blinding of the data in unsupervised methods.[40-41] Supervised methods analyze for metabolic differences between classifications the investigator stipulates. These classifications would align with the experimental groups: diseased/disease-free, drug treated/vehicle control, mutant/wild-type etc.[9, 41] While unsupervised methods look for patterns in variables without prior knowledge of the groups.[41]

Principal component analysis (PCA) is an unsupervised technique often used to define trends in data by creating groupings that are not defined by metadata.[18] PCA only shows differentiation if the within group variation is not much lower than the between group variation.[13] Following PCA, a supervised technique such as partial least squares-discriminant analysis (PLS-DA) can be used to determine the metabolites that differentiate the treatments.[10, 12-13, 18, 36]

Unsupervised methods examine the data simply based on the variations within the data, without investigator classifications and are therefore unbiased analyses.[9] These methods are particularly useful when expected differences are small or when inter-sample variation is high because supervised methods are only able to distinguish differences between defined groups only when the out-of-group variation is higher than the in-group variation.[8-9] Unsupervised methods only consider the simultaneous relationships between the presence and intensity of the features and is therefore an unbiased analysis of the chemical profile of the sample.[9] One of the most commonly used unsupervised metabolomics analytical tool is principal component analysis (PCA) because this method is equally effect in either untargeted or semi-targeted studies.[21]

Both PCA and PLS work by focusing the analysis on those variables (in the case of metabolomics features) that are diagnostic of the groups that are designated by the experiment, either as determined by the investigator or on the merit of the data alone.[8] PCA uses linear transformations to reduce the dimensionality of the data while maintaining as much variance as possible.[8] This method, because it is unsupervised, requires that groups be rather homogenous, this makes it appropriate for the analysis of untargeted metabolomics datasets.[8] PLS is another method of reducing the dimensionality of a data set, but it is a supervised method in which groups are designated by the investigator, which make it a good choice for hypothesis driven untargeted metabolomics experiments.[8]

While the data from supervised methods is useful, it is subject to being overfitted.[9] Overfitting occurs when the model is generated from an excessive number of variables.[36, 42] The model therefore appears to fit the existing data, but will fail to be validated by a repeated study.[42] This is most likely the reason that despite considerable effort, there has been so little success in the discovery of biomarkers for a variety of disease states: models were overfitted and therefore could not be validated given a different sample pool.[12] Overfitting of data is more likely when samples are at low concentration, emphasizing yet again the importance of appropriate sample preparation.[37] Overfitting can be avoided by dividing the data into a training dataset, a validation dataset, and a test dataset.[9] If these datasets are assigned randomly and the trends remain the same through several iterations of analysis, the results are likely not due to overfitting of the data and the use of them will be a valid outcome of the experiment.[9]

Selection of the appropriate method must come from the experimental design. Because PCA is an unsupervised method, it is generally best suited for analysis of exploratory experiments with an aim to remain unbiased when testing the homogeneity of the groups.[8] The results of these methods are scores which can be plotted in a scatterplot to visualize the differentiation of groups.[8] However additional metrics must be used to determine the statistical relevance of the resultant distributions.[8]

1.5 Challenges associated with metabolomics (emphasis on MS)

The unparalleled sensitivity of MS achieves very low detection limits, so when used for MS-based metabolomics results in a large number of low-intensity signals which are likely unknown and are below the level of detection for most tools necessary to determine the structure (i.e, NMR).[21] While it is tempting to consider each signal/peak found in a MS-based study as an individual compound, this is not the case.[9] Peaks in the chromatogram are generally referred to as "features" rather than compounds or metabolites because a single metabolite might ionize in a number of ways, form a variety of adducts, or fragment in the source resulting in multiple features originating from a single metabolite.[9]

These features are identified by a unique combination of retention time (RT) and mass to charge ratio (*m/z*).[25, 36] Each feature is exported along with its intensity, determined by calculating the area under the curve of the peak.[36] This value is a measure of the relative intensity of the feature's abundance.[36] In many cases, features are typically matched by their UV and RT to an libraries for dereplication and identification of potential new compounds, but this is not necessarily transferable between instruments or laboratories.[43] Therefore, one of the most challenging aspects of metabolite data analysis is the identification of the features detected in the analysis.[9, 14, 27] Indeed the vast majority of features detected in these experiments require extensive additional experimentation to be fully confirmed.[25]

Despite these challenges, some feature identification can be achieved by matching MS2 spectra to libraries of spectra available either freely or by subscription.[9] These databases will gain in value and accuracy if reference spectra are constantly being updated and expanded. And indeed there are increasing numbers of libraries and databases which can be used to dereplicate features in metabolomics studies based primarily on MS2 fragmentation patterns, but occasionally incorporating retention time and UV pattern.[21] Although these efforts are somewhat impeded by the fact that there will be differences in MS2 fragmentation patterns based on the mass analyzer or the brand of the instrument, to say nothing about the variability associated with chromatographic separations.[21]

However, matches to library spectra can only provide putative feature identification.[34, 36] The confirmation of identity of these putatively identified metabolites is a major bottle neck in the metabolomics workflow.[14, 27] Despite the advances in the field, the unparalleled level of detection of MS has meant that the vast majority of chemistry observed in metabolomics experiments cannot be identified.[44] These identifications should ideally be confirmed by comparing the *m/z* and RT of an authentic standard run under identical experimental conditions to the original experiment.[34, 36] This is particularly problematic for the analysis of secondary metabolites as opposed to primary metabolites because of the wide diversity of secondary metabolites and the lack of commercial standards which would allow the confirmation of feature identity.[17] Because of the time investment involved in making these putative matches, and the vast number of features in a metabolomics analysis, a triage step is necessary to determine which features will be important in the context of the study.[36] This assessment is conducted with the help of either multi or univariate statistics.[36] Multivariate approaches are more efficient for LCMS metabolomics datasets because they analyze multiple variables simultaneously, but the inherent complexity of the dataset and the complexity of analysis there is a substantial risk of overfitting the model to the data such that the model becomes less accurate.[36]

1.6 Applications for metabolomics

Metabolomics analysis can be applied to answer questions in many fields that involves the comparison of metabolite levels between groups.[1, 5, 11, 13, 15-16, 18-19] These include diseased to healthy organisms, mutant to wild type, stress due to abiotic factors to unstressed, mature to immature, to name only a few.[1, 5, 11, 13, 15-16, 18-19] Metabolomics is particularly well suited to identify biomarkers for a disease-state or track changes that occur as a result of a treatment, either by its presence or absence.[5, 13, 19, 24]

Cancer diagnostics is perceived to be a particularly promising target for metabolomic analysis because cancer is known to cause profound shifts in metabolism of cells even at early stages.[5, 13, 19] Ideally these biomarkers can be identified and traced back to a metabolic pathway to shed light on to the metabolic consequence of the experimental conditions.[14] Because there are quite distinct metabolic differences between healthy cells and malignant cancer cells, metabolomics is a potentially valuable tool to differentiate between cell types and provide early diagnostic, which could save lives by allowing earlier detection.[19] However, despite many metabolomics studies and considerable effort dedicated to defining biomarkers of a variety of disease states, this has not translated into the adoption of biomarkers in the clinic.

The diagnosis of major depressive disorders is currently a very subjective process, which can often leave patients in a gray area where their actual diagnosis is unclear.[10] The use of metabolomics can perhaps allow a more objective approach to the diagnosis of this disorder and other mental health disorders.[10] By focusing this analysis specifically on neurotransmitters rather than more general metabolites, the authors were able to identify pathways that are perhaps pathogenic of the disorder.[10] This study found that patients with major depressive disorder had a different metabolic profile than did those patients with bipolar disorder, which suggests that there is promise for identifying biomarker that will objectively diagnose metal health diseases.[10] This points towards improvements in accurately diagnosing the condition as well as investigating the mechanics of the condition which might allow a novel treatment approach.[10]

Related to this concept, metabolomics may be applied to monitoring and quantifying the intermediates of metabolic pathways, to give a snapshot of the phenotype of the organism.[21] The examination of metabolomic information in the context of pathway analysis can provide a wealth of information about disease progression in different populations of people[5] or to how organisms respond to environmental contaminants**[11, 18]**. [12] The understanding of the molecules that are altered from a healthy state to a diseased state can help in identifying the metabolic pathways that are influenced by the disease and can in turn suggest targeted therapeutics for the disease.[25]

More broadly, metabolomics experimentation can be used to observe metabolite fluctuations in response to environmental conditions, ecological questions can be answered using metabolomics.[7, 11, 15-16, 18, 24] The investigation of the influence of biotic or abiotic stresses on the metabolic output of plants or microorganisms can lead to better understanding of normal metabolism as well as stress responses.[7, 18, 24] Metabolomics has the potential to expose the subtle changes in metabolism in response to pollutants despite the lack of an observable phenotypic response.[7] This would be useful to help ameliorate the effects of pollution because it gives a better idea of the actual toll pollution is taking on the organism.[7, 18] Metabolomics analysis was also used to discover the impact of pharmaceuticals on a sentinel aquatic organism.[11] While studies with more pharmaceuticals would be beneficial, the observation of the impact of environmental contamination is an important issue that metabolomics can perhaps begin to address.[11] Metabolomics can be used to assess the levels of contamination of the food supply[27] as well as determining the nutrient levels of that food.[32] Other applications of metabolomics to ecology are limited only by the imaginations of investigators provided appropriate experimental design which will allow results to be meaningful in the context of ecology.[7]

Metabolomics also shows great promise for the analysis of complex extracts for natural products discovery.[45] Metabolomics technology has progressed sufficiently, that not only are metabolomics experiments expected to be able to provide putative identifications of natural products, potentially highlighting new metabolites in the process, but some work has been done to highlight the active features in complex mixtures, which promises to define the future of natural products discovery.[45]

Traditional natural products discovery is approached through bioassay guided fractionation, but this strategy is biased towards abundant compounds that are easily detected and isolated.[37] The wide range of natural products makes for a daunting isolation process because natural products are not typically encountered as single molecules, but as sets of compounds representing the total metabolic outputs of organisms.[46] Metabolomics analysis can be used to expedite the discovery process and allow the focus removed to the less abundant metabolites.[37] This is especially true in the hyphenated MS techniques. Modern separation science is able to achieve a high degree of resolution with the help of ultra performance LC systems. This coupled to MS which is one of the most sensitive instruments available gives great power to sift through signals both high and low.[21, 28] The use of MS2 fragmentation patterns can also help to differentiate peaks with similar *m/z* and RT, but different structure.[47] Metabolomics is a valuable tool for focusing natural products discovery on the bioactive metabolites in a complex mixture.[37]

In the context of natural products, molecular networking is potentially a very powerful tool. Natural product families or scaffolds shows higher correlation with the biosynthetic gene clusters than do individual natural products.[48] This suggests that networking analysis based on MS2 pattern is a valid approach to natural products metabolomics.[48] Suggesting that if a single feature is found to be active, then the other features, that share that same active scaffold by merit of being in a shared network, would be good candidates to probe structure activity relationships. Metabolomics has become a valuable tool which can be used to probe many questions relating to natural products.

1.7 Chemical diversity and metabolomics

Metabolomics is a tool that has and can be applied to a wide range of fields. At the heart of metabolomics is the description of chemical diversity. This is often interpreted as a comparison between treatments, but can be applied more broadly to describing the chemical diversity of a group of organisms. Metabolomics has previously been applied to some aspects of natural products discovery, but thus far has not been applied to assessing the chemical diversity of the libraries used to discover natural products. This work represents the first steps in providing evidence-based guidance to inform library building.

Chapter 2: Hypothesis and Chapter Overviews

2.1 Hypothesis

Drug discovery libraries make up the backbone of the natural products discovery pipeline. Despite this reliance upon the library, there is little to no evidence-based research to guide the construction of these libraries. Metabolomics is a valuable tool which can be used to examine the metabolic output of an organism making it a valuable tool in the evaluation of natural product library chemical diversity. With the aim of initiating rational natural products library assessment and design, the hypothesis guiding my research was: **metabolomics is a valuable tool that can be combined with common diversity measurements to assess the diversity and inform the design of drug discovery libraries.** This hypothesis was tested via the following specific aims:

1. Use metabolomic analyses to investigate the chemical and genetic diversity within a fungal genus to determine the appropriate library size to achieve maximal chemical diversity.

2. Use metabolomic analyses to probe the value of niche environments in the development of a natural product extract library.

2.2 Building Natural Product Libraries Using Quantitative Clade-Based and Chemical Clustering Strategies

In Chapter 3, I present the development of a method to assess the chemical diversity of natural products discovery libraries. Natural products libraries are often built upon the assumption that genetic diversity will result in chemical diversity. The degree to which genetic diversity results in chemical diversity is however unknown. The distribution of chemistry within a genetlically similar group is also largely unexplored but is particularly relevant to natural products libraries. Depending upon the chemical homogeneity of strains, library size could be inferred to maximize chemical diversity. This work aims to address this question by combining chemical and genetic analyses. While traditional metabolomics-based data analysis is employed, the chemical diversity is explored using tools adapted from ecology. This hybrid approach to assessing library diversity allows the prioritization of certain strains that would maximize the chemical diversity of the library as a whole. This improvement in library building strategy could result in more successful discovery efforts.

2.3 Assessing Metabolic and Biological Diversity to Support Natural Product Library Assembly

In Chapter 4, I discuss the investigation of the source of organisms included in natural products discovery libraries and the value of these organisms to the chemical diversity of those libraries. In the search for novel natural products, many extreme and remote environments have been surveyed, the assumption being that organisms that have adapted to live in such environments will have chemical production capabilities that are not observed in organisms found in more accessible locations. While there have been cases that extremophiles have been found to produce novel chemistry, the value of these organisms to libraries has not been firmly tested. Does the environment shape any organism that can persist to survive in it to produce novel chemistry, or does the environment select for organisms that are suited to that environment and these organisms are inherently more likely to produce novel chemistry? This is a complex question to answer and will require a different collection strategy based on the answer. In an effort to address the first question, we investigate the first point and examine organisms that are found in both inaccessible and accessible locations to determine if there is any evolutionary plasticity in their chemical output. Answering this question will allow the refocusing of collection efforts with the emphasis on those organisms most likely to increase the library's chemical diversity. This will in turn aid in the search for novel chemistry that will aid in the search that is the heart of natural products discovery.

Chapter 3: Building Natural Product Libraries Using Quantitative Clade-Based and Chemical Clustering Strategies

This chapter was adapted from a paper with the same title that has been submitted to mSystems in May 2021. The authors are Victoria Anderson, Karen Wendt, Fares Z. Najar, Laura-Isobel McCall, and Robert H. Cichewicz. The work presented in this chapter was conducted as follows: Victoria Anderson performed fungal culture, DNA barcoding and phylogenetic analysis, fungal extract preparation, LC-MS-MS data collection, and metabolomics data analysis.

3.1 Introduction

Drug discovery has changed tremendously during the last century, with the process undergoing continuous reinvention to avail itself of new scientific methods and trends. Numerous ideas and tools have been put into practice, resulting in the creation of many chemical collections used in modern drug screening and molecular probe development throughout academia, industry, and government. Small-molecule libraries based upon organic compounds of various sizes (e.g., <900 Da for most synthetic libraries, but ranging up to around ~2,000 Da for some natural products) play a dominant role in such efforts, with collections accommodating a variety of screening and discovery methodologies (e.g., fragment-based, target-focused, diversity-oriented, combinatorial, DNA-encoded, repurposed, virtual, and more).[49-54]

Despite the vast amounts of time, money, and energy poured into building small-molecule screening collections, the answers to many basic questions about their design and development, such as identifying optimal collection size, are largely driven by adherence to dogma or convenience rather than evidence-based reasoning. Such questions grow increasingly relevant as opinions influencing the last four decades of library design have shifted tremendously with the large collections of the 1980s and 1990s (e.g., combinatorial chemistry[55]) being replaced by smaller tailored collections (e.g., “focused” collections[56-57]) in the early 2000s, and now moving toward mega-scale libraries (e.g., encoded libraries[58-60]) in recent years.[61-63]

While such trends are strongly linked to the creation of synthetic chemical collections, a similar set of concerns applies to the construction of libraries assembled from natural sources (e.g., microorganisms, plants, and more). Many ideas have emerged relating to best practices for building natural product libraries with extracts, fractions, and pure compounds defining the three dominant types of chemical complexity encountered in screening collections.[64-67] Despite the tremendous ingenuity and effort that has gone into assessing these and other methods of building natural product libraries, comparatively less consideration has been given to identifying optimal sample sizes needed to construct nature-based screening collections. Answering such questions are important since the degree of chemical diversity in a screening collection is considered a key contributor to the success (or failure) of bioassay screening endeavors.[68-69]

A possible reason for neglecting this problem may stem from the fact that as opposed to synthetic libraries, natural products are encountered not as single molecules, but as compound sets (e.g., metabolomes) representing the total metabolic output of each organism. Given the degree to which natural product biosynthetic gene clusters and their molecular controlling factors are swapped, recombined, and otherwise altered within host organisms, even the metabolomes of low-ranking monophyletic clades (e.g., a species or genera) can exhibit divergent chemical profiles. These factors can make the rational design of natural product libraries challenging. Therefore, methods to perform chemical diversity measurements have the potential to aid in the design of natural product drug screening collections.

Two examples help illustrate the practical need for solving this problem. In an intriguing opinion piece offered by Baltz, various scenarios were offered to overcome the global slowing of antibiotic discovery from actinomycetes (order: Actinomycetales Buchanan, 1917).[70] Based on that analysis, it was concluded that using traditional bioassay-guided antibacterial discovery alone would require testing >10^7 actinomycetes to identify the next, major new class of antibiotic. Although this estimate was highly theoretical and predicated on standard bioassay-driven screening methods, it provided a compelling starting point for considering how the integration of compound diversity measurements into bioassay screening could help serve as a chemically focused approach to assessing real and presumed barriers to natural product discovery. In another case, Jensen and colleagues carried out a survey of natural product biosynthetic gene cluster diversity represented in 119 *Salinispora* sp. genomes.[71] A key takeaway from the study was that despite high levels of global gene conservation among *Salinispora* isolates, roughly half of all the biosynthetic gene clusters detected were found in two or fewer isolates. Thus, deep sampling of this genus was expected to continue yielding new families of natural products. With no end in sight for the sustained emergence of novel natural product scaffolds,[72] questions surrounding how to define, measure, and construct optimally sized natural-product-based chemical libraries take on critical importance.

Fungi epitomize many of the challenges inherent in sourcing natural products, and thus serve as a useful starting point for establishing a quantitative approach to natural product library design. Topmost among the difficulties working with fungi are the complex, and in many cases, poorly resolved taxonomic relationships exhibited by these organisms. For example, many fungi adopt different sexual states that are metabolically and morphologically distinct. Historically, such cases have resulted in fungal isolates exhibiting gene-level equivalencies being assigned different binomial names.[73] In other instances, the high degree of genetic diversity exhibited within certain fungal clades has created taxonomic quagmires that have left some fungi loosely classified into poorly defined species complexes, polyphyletic clades, and paraphyletic groups.[74-75] Complicating these matters, the regional variation and global distribution of most fungal taxa remains poorly defined, which has given rise to unresolved questions about the true extent of biological and chemical diversity throughout the fungal kingdom. Herein, we present a set of guiding principles for combining, quantifying, and assessing chemical and source-organism diversity during the construction of natural product libraries. Our efforts focused on *Alternaria* Ness, which is a cosmopolitan and taxonomically perplexing fungal genus[76-77] known to produce many types of metabolites[78-83]. Although these experiments concentrated on fungi, we expect that the procedures laid out here will be generally applicable to the evaluation of natural products from other source organisms.

3.2 Results and Discussion

3.2.1 Basis for a bifunctional analysis tool to assess *Alternaria* ITS barcode and chemical diversity.

The *Alternaria* isolates used in this study were obtained through the University of Oklahoma, Citizen Science Soil Collection Program,[84-85] which to date has received 9,670 soil samples from across the United States, yielding 78,581 fungal isolates identified by single-read internal transcribed spacer (ITS) sequencing data. A query performed on the ITS barcode data yielded an initial set of 219 candidate *Alternaria* isolates, which was refined to a subset of 198 samples having >90% ITS sequence similarity[86-88] to *Alternaria* type strain data available in GenBank and defined by Woudenberg et al.[77] Upon plating, all strains exhibited colony morphologies consistent with the genus *sensu stricto*.

Alternaria exemplify many of the practical problems and limitations that researchers face when developing natural product libraries. Specifically, *Alternaria* is a taxon in flux, having undergone revisions as mycologists have striven to consider morphological characters, telemorphic states, various marker genes, and more to delineate this group and its allied genera.[74, 89-93] While the outcomes of those efforts have differed, resulting in proposals supporting various combinations of monophyletic species groups and species complexes, they have found agreement on the grounds that *Alternaria* exhibit tremendous morphological and genetic plasticity. Recognizing these problems are common throughout the microbial world, we adopted a hybrid method of library construction focused on assessing the prospective taxonomic affinity of each isolate (preferably to a genus-level taxon using ITS barcode sequence results) in combination with LCMS metabolome profiling data. This bifunctional approach offers insights into the relationship between phylogeny and chemistry, which enables (1) assessment of natural product chemical diversity within species complexes, (2) identification of prospective pools of under- and over-sampled secondary metabolite scaffolds, and (3) application of quantitative metrics to establish and track goals concerning chemical diversity in an existing or growing natural product collection. Whereas numerous tactics have been reported for guiding natural product library development[94-96], we view our approach as a departure from prior schemes, considering its quantitative aspects that we now explore.

3.2.2 Characterizing ITS barcode (clades) and metabolome (clusters) based groups in *Alternaria*.

While achieving a state of perfect knowledge about the evolutionally histories of microorganisms is nearly impossible to achieve, we can use certain low-cost and minimally intensive tools to gain functional insights concerning their phylogenetic relationships. For fungi, the ITS barcoding system serves as one such tool offering an efficient way to establish a working set of phylogenetic associations among environmental isolates.[75] Phylogenetic analysis of the *Alternaria* ITS data revealed five sequence-based clades (Clades U, V, W, X, and Y). Whereas further taxonomic resolution might be achievable using additional genetic markers, ITS provides a reasonable method to identify isolates and draw attention to potential points of evolutionary divergence.[73, 75]

Principal coordinate analysis was performed on the *Alternaria* metabolomics data. The components detected in *Alternaria* metabolomes were treated as chemical features based on a combination of their LC retention times and mass-to-charge ratio. Those efforts resulted in a model that supported the presence of six chemical clusters (Clusters 1, 2, 3, 4, 5, and 6) among the *Alternaria* isolates.

The results generated from the ITS barcode and metabolomics data sets were overlaid demonstrating a high degree of consensus between the two models (Figure. 3.1). The data indicated that Clade U was composed primarily of chemical Cluster 1, Clade W was composed of chemical Cluster 2, Clade X was composed primarily of chemical Cluster 6, and Clade Y was composed of chemical Cluster 3. Notably, Clade V contained both Clusters 4 and 5. This underscored the value of layering chemical data (clusters) on top of genetic data (clade) to reveal otherwise unexpected pockets of chemical divergence within genetic groups. A handful of cases were noted in the principal coordinate analysis, revealing that some members of chemical Cluster 2 were embedded in Clades U, V, and X. Although the reasons behind these cases are uncertain, we speculate that it may be due to culture-dependent effects on metabolite production and/or genomic/epigenome-scale events that resulted in the loss of chemical scaffolds, which served to differentiate Clusters 1, 3, 4, 5, and 6 from Cluster 2.

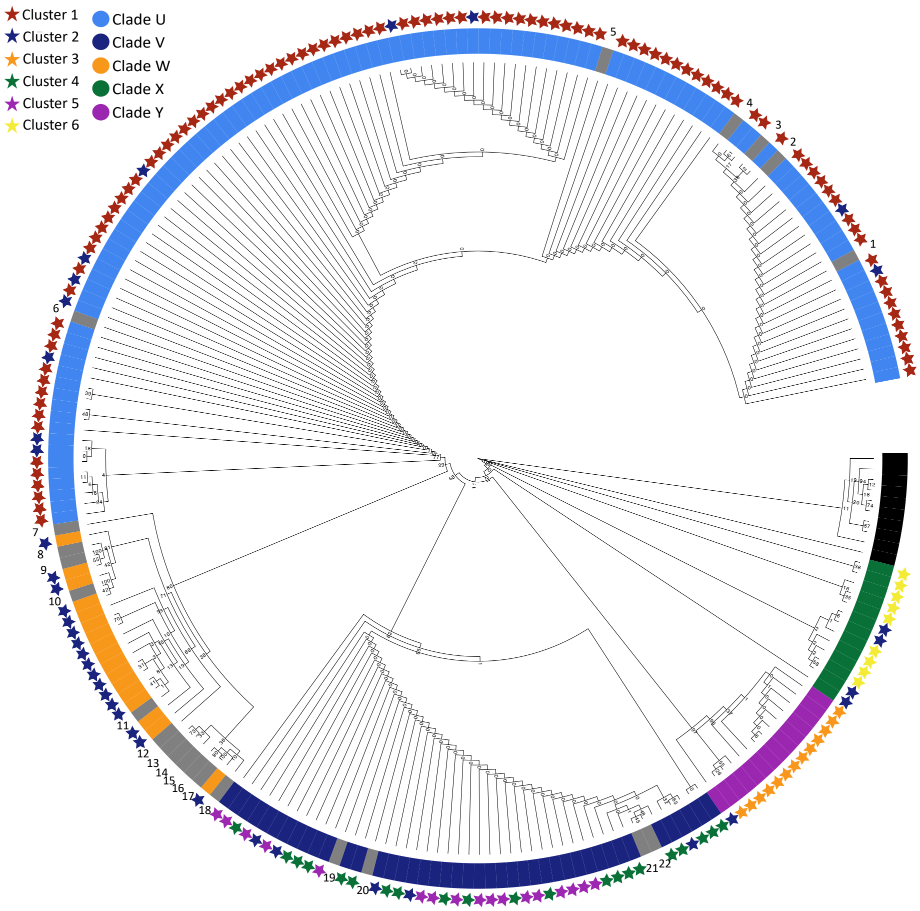


Figure 3.1. Genetic and chemical distribution of Alternaria. ITS phylogeny of Alternaria isolates. Inner ring indicates the clade, while stars indicate the chemical cluster of isolate extracts. The clade and clusters show remarkable overlap, but also reveal a hidden chemical cluster within a single clade. Numbers indicate type strains from Genbank (Supplemental Table 3.1

Considering the geographic scope of the collection, the genetic clade and chemical cluster data were evaluated to determine if their distributions might be limited to certain geographical regions (Figure 3.2). Given the number of samples tested over such a large land mass, we are cautious in interpreting our results; however, we did note that Cluster 5 was only detected in the far western portion of the United States. Additionally, Clusters 3 and 4 were absent from the southeastern portion of the United States. Both observations served to fuel speculation that the occurrence of some *Alternaria* chemical features might be limited to circumscribed geographical ranges. Further investigation will be required to determine if these are veritable patterns or sampling artefacts.

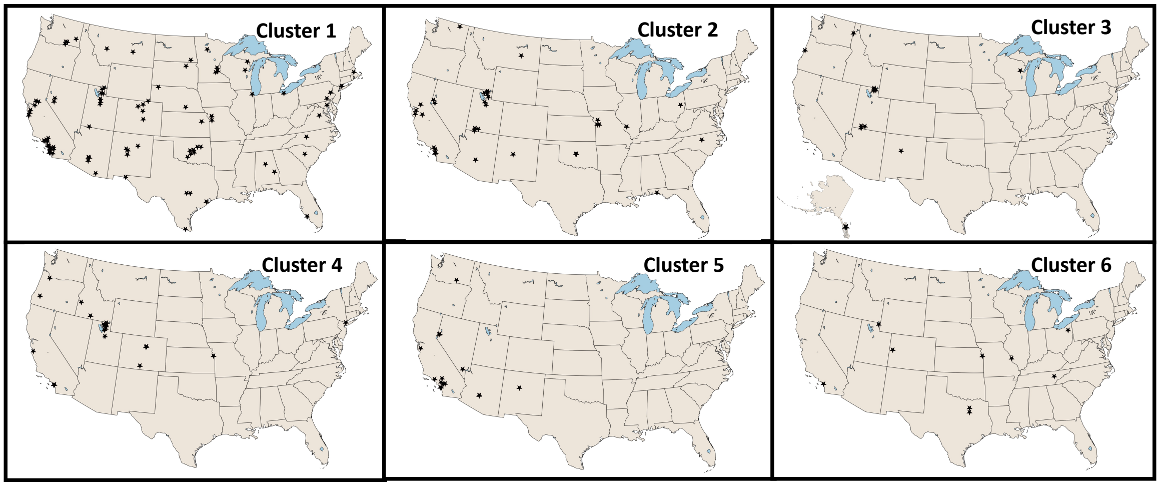


Figure 3.2. Chemical and geographical distribution of Alternaria. Geographic distribution of isolates by chemical cluster. Chemical clusters overlap with Genetic clades with the exception of Cluster 4 & 5 which are embedded in Clade V.

3.2.3 Chemical feature production among genetic clades.

Before proceeding, it is worth noting that in the comparisons presented here and in subsequent sections, the discussion could have been structured around evaluating *Alternaria* isolates according to ITS clades (genetics) or chemical features (metabolomics). Apart from Clade V, our tests demonstrated rather strong agreement between the two models, which indicated that both clustering mechanisms worked well to organize data along seemingly natural divisions. Knowing that taxonomically driven strategies continue to play prominent roles in natural product collection efforts, we have opted to analyze the chemical diversity findings in the context of ITS clades (Figure 3.1). However, we see no reason why a chemistry-centric grouping could not be used, and several examples of parallel tests based on chemical clusters are provided in the Appendix 1.

Median numbers of detected chemical features differed significantly between ITS-based clades ($p< 0.0001$), with Clades U and Y containing isolates that produced the greatest total numbers of chemical features (Figure 3.3A). This observation held true ($p< 0.0001$) after performing sub-sampling of the clades to alleviate potential errors introduced due to sample size non-equivalence (Supplemental Figure 3.1A). Relatively few outliers were detected within the genetic clades indicating high levels of consistency for the metabolic output of the isolates in each group. Clades V, W, and X were found to have significantly fewer features than Clade U (Tukey's HSD of ANOVA $p< 0.0001$ in all cases), suggesting that Clade U is chemically more diverse than the other clades.

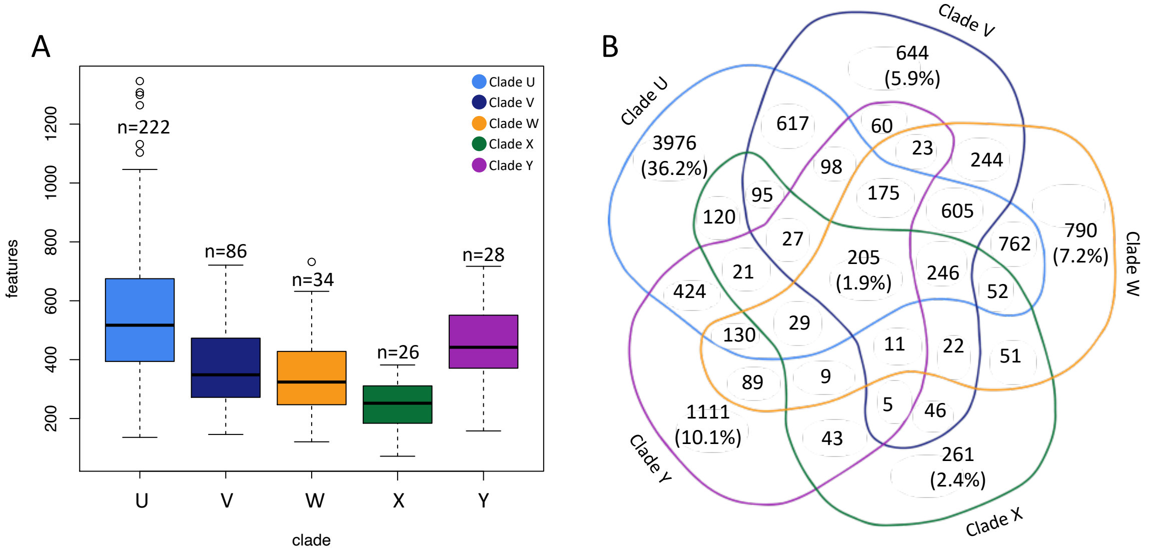


Figure 3.3Examining feature diversity of Alternaria. (A) Alpha diversity of genetic clades. Median number of chemical features differed significantly by clade. (B) Chemical overlap of features by clade.

Only 1.9% of features (205) were detected in all clades, comprising the core metabolome of the *Alternaria* isolates (Figure 3.3B, cluster-based analysis Supplemental Figure 3.2). While up to 40% of chemistry is shared between clades, we found that the bulk of features were limited in occurrence to just a single clade. Progressing from the smallest to the largest number of clade-specific features, 2.4% of features (261) were found only in Clade X, 5.9% of features (644) were present only in Clade V, 7.2% of features (790) were detected only in Clade W, 10.1% of features (1,111) were observed only in Clade Y, and 36.2% of features (3,976) were identified only in Clade U. These results demonstrate that high levels of chemical diversity exist even within the traditionally recognized boundaries that define *Alternaria*.

3.2.4 Making informed library building decisions based on chemical feature diversity.

To monitor and better understand how feature diversity could be used to make informed decisions about constructing natural product libraries, feature accumulation curves were constructed from the metabolomics data (Figure 3.4A). The results showed that despite a large degree of ascribed taxonomic diversity in *Alternaria*, a surprisingly limited number of isolates were required to provide broad chemical coverage of the genus. Indeed, random sampling of the *Alternaria* data found that on average, a set consisting of as few as 23 isolates was expected to provide 50% of the total pool of *Alternaria* features. Expanding on these findings, randomly selected subsets consisting of 57, 104, 142, and 195 isolates were anticipated to provide 75%, 90%, 95%, and 99%, respectively, of *Alternaria* features (Figure 3.4A). Thus, it was determined that feature accumulation data could serve as a useful tool for estimating levels of chemical feature coverage within taxonomic groups.

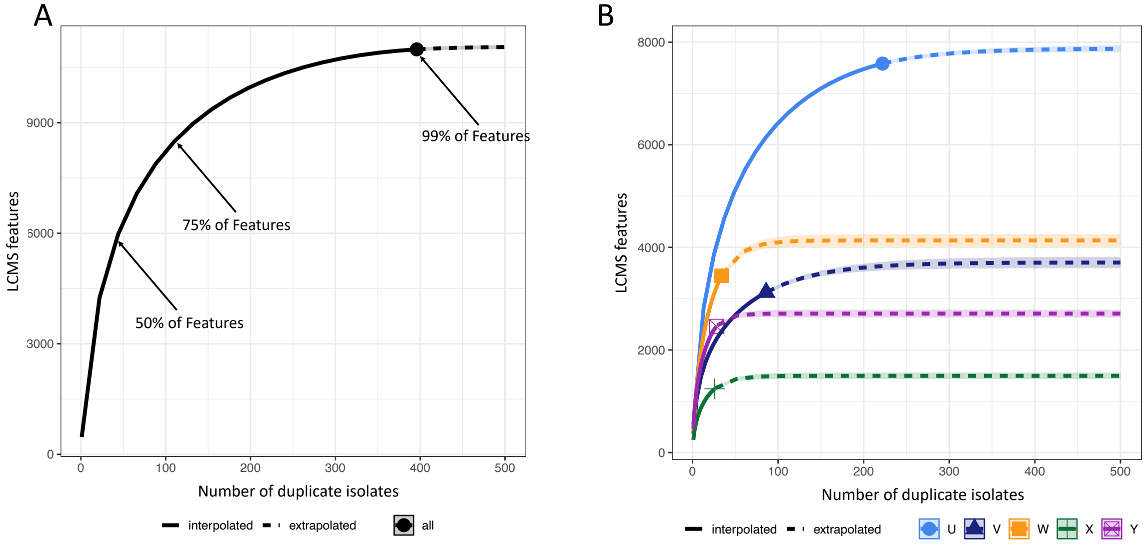


Figure 3.4 Extrapolating feature diversity of Alternaria. (A) Extrapolated rarefaction curve of Alternaria. (B) Extrapolated rarefaction curves of clades within Alternaria. Clades are both genetically and chemically distinct

Whereas the genus-based amalgamation of feature data provided useful insights into the chemical diversity of *Alternaria*, a more granular exploration of feature accumulation results by sub-genus clades has the potential to afford a complimentary perspective for library design. Clade-based feature accumulation curves (Figure 3.4B) showed that feature coverage levels of 99% were achievable in Clades U (contained the most feature-rich isolates, Figure 3.3A) and X (contained the most feature-poor isolates, Figure 3.3A) with 170 and 51 total isolates, respectively. In contrast to the rank order of the median numbers of features per isolate, the point at which 99% feature saturation occurred followed a different pattern for Clades V, W and Y. Clade Y, which contained the second highest level of features per isolate (Figure 3.3A), was found to require the fewest number of isolates (39 isolates) to achieve a level of 99% feature coverage. Clade V contained the third highest level of features per isolate (Figure 3.3A), while also needing the second highest number of isolates (141 isolates) to achieve a level of 99% feature accumulation. These results are likely due to the presence of two chemical clusters embedded in Clade V. Clade W, contained the second lowest number of features per isolate (Figure 3.3A), but was predicted to require the third highest number of isolates (66 isolates) to achieve a level of 99% feature accumulation. Thus, feature accumulation curves utilizing ITS-based clades offer a useful method for identifying and monitoring genetically-defined groups of organisms that are likely to require increased efforts (i.e., more isolates) to achieve pre-specified levels of feature accumulation coverage. Related to these efforts, the rarefaction curve slopes were plotted in relationship to the number of samples representing each clade (Supplemental Figure 3.3). The results of that analysis revealed an inverse relationship existed between the slopes of interpolated rarefaction curves and the number of samples surveyed within a clade supporting the idea that in this data set, the larger ITS-based clades tended to approach saturation of feature coverage.

3.2.5 Probing chemical scaffolds distribution and diversity in *Alternaria*.

Whereas the analysis of chemical features offers a straightforward approach to comparing LC-MS data from different natural product sources, such results can be prone to misrepresenting underlying chemical diversity trends. Specifically, the output from natural product biosynthetic pathways tend to occur as assemblages of structurally related metabolites rather than as single products due to several factors related to the *in situ* formation of natural products, including substrate promiscuity, competing actions of multifarious tailoring enzymes, and more.[46, 97-98] Consolidating chemical features that share underlying structural similarities into groups referred to as scaffolds is one approach to account for this phenomenon. Molecular networking [47, 99-101] is an approach that has gained widespread use to build scaffold-level relationships in the field of natural products.[39, 102-104]

Using molecular networking to identify structurally related metabolites from *Alternaria*, the 10,991 molecular features were condensed into 5,754 of scaffolds (Figure 3.5A). Upon removing singleton scaffolds (4,193) from the dataset, 17.2% of the scaffolds (285) were found to be shared by all five ITS-based clades (Figure 3.5B and Supplemental Figure 3.4). These shared scaffolds represented the core metabolome of the *Alternaria* encountered in this study. We also found that 32.5% (539) of the non-singleton scaffolds were detected in just a single clade. Clade U contained the largest number of unique chemical scaffolds (19.6%, 326 unique scaffolds) followed by Clades Y (5.1%; 84 unique scaffolds), W (3.6%; 59 unique scaffolds), V (2.9%; 48 unique scaffolds), and X (1.3%; 22 unique scaffolds). The rank order of the scaffolds detected in a clade mirrored the respective levels of chemical features observed in each group (Figure 3.3A). Thus, we speculate that the relative quantities of chemical features detected within taxa might serve as a surrogate measure for predicting comparative levels of relative scaffold diversity in other taxa. These results also highlighted the need to differentiate scaffold versus feature diversity goals when establishing parameters for natural product library design since 17.2% of scaffolds were found to be shared by all clades of *Alternaria*, but only 1.9% of features were shared by all clades. Furthermore 61.7% of chemical features were found to be unique to a single clade, but this held true for only 32.5% scaffolds, which is not surprising given that scaffolds are more highly conserved across *Alternaria* isolates.

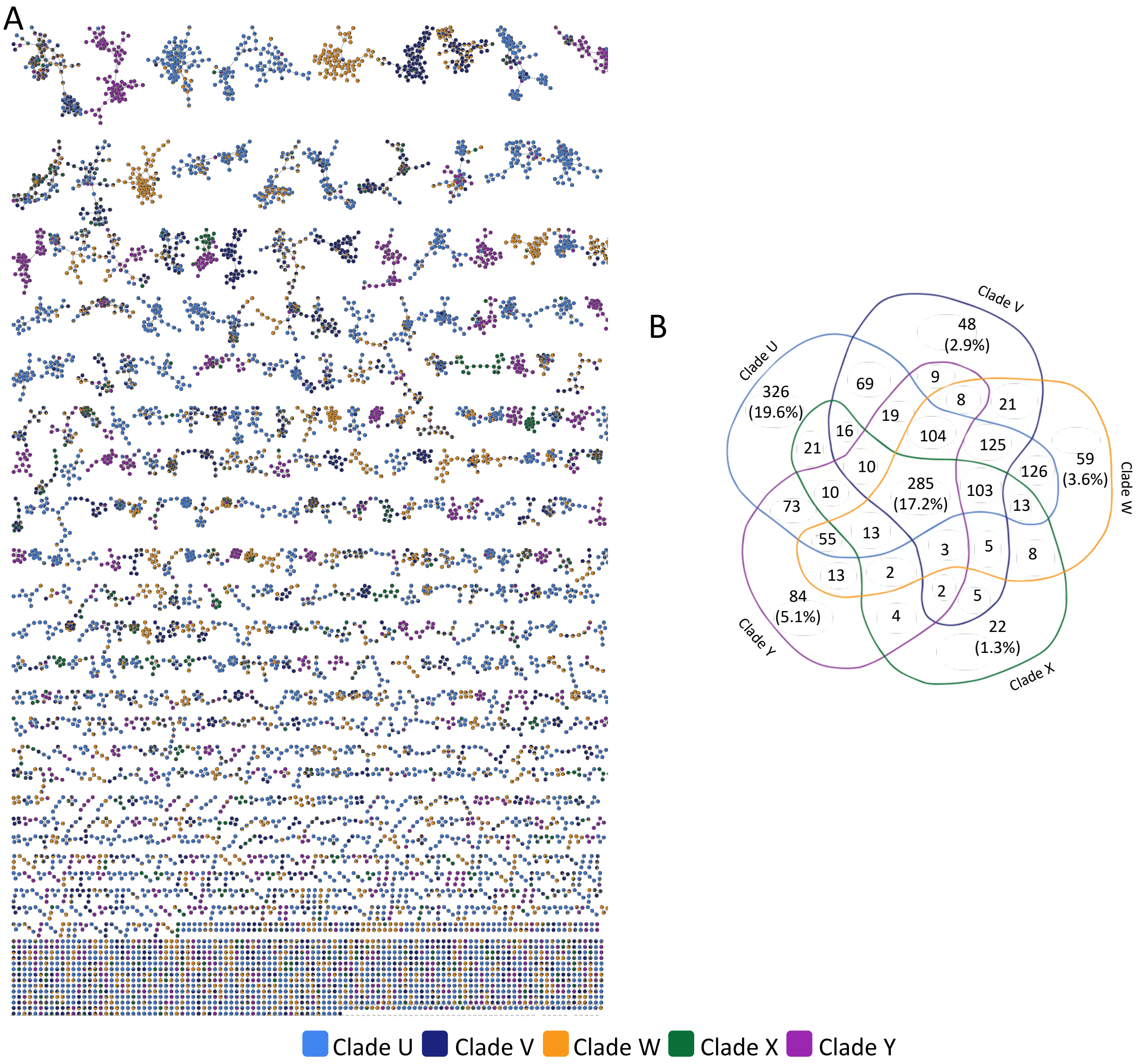


Figure 3.5 Examining scaffold diversity. (A) Molecular network of extracts showing 5,754 sub-net-works/scaffolds. Nodes are colored by Clade. (B) Overlap of chemical scaffolds by clade.

3.2.6 Applying clade and cluster data to assess progress toward goals for natural product library coverage.

Considering the entwined functions that phylogeny and chemistry play in natural product library development, we explored how less abundant taxa might contribute to the overall chemical diversity within a screening library. Such models could be useful for understanding how rigorous efforts to include less abundant taxa, or purposeful endeavors to exclude highly abundant groups of organisms, might impact the representation of chemical scaffolds in a collection. We first examined how forming a library by exclusively focusing on only the most abundant taxon, Clade U, would affect the chemical diversity of a collection (Figure 3.6A and Supplemental Figure 3.6). The accumulation curves revealed that the 111 isolates in Clade U were capable of providing access to 80.1% of all *Alternaria* scaffolds, while the remaining, less abundant Clades V, W, X, and Y added just 7.0%, 5.4%, 1.7%, and 5.7%, respectively, of additional chemical families (note that the order in which Clades V, W, X, and Y were added was arbitrarily chosen). In contrast, when the scaffold accumulation data were examined with the focus placed on sampling just the less abundant taxa, it was found that the 87 isolates representing Clades V, W, X, and Y afforded access to 78.3% of total scaffolds encountered from *Alternaria* (Figure 3.6B). This result was unanticipated with near-equivalent percentages of unique scaffolds afforded via these contrasting approaches. We realize that most real-world library-building efforts are unlikely to engage in such restrictive collection practices; however, these results could have practical implications for cases in which searching out less abundant (i.e., rare taxa) or difficult to culture organisms may add undue cost or time to building a natural products drug screening library. Thus, modeling scaffold (or chemical feature) accumulation can help researchers focus on achieving desired levels of chemical coverage in natural product libraries, as well as monitoring whether collection efforts have led to oversaturation or under-sampling of the theoretical chemical diversity within a given taxon.

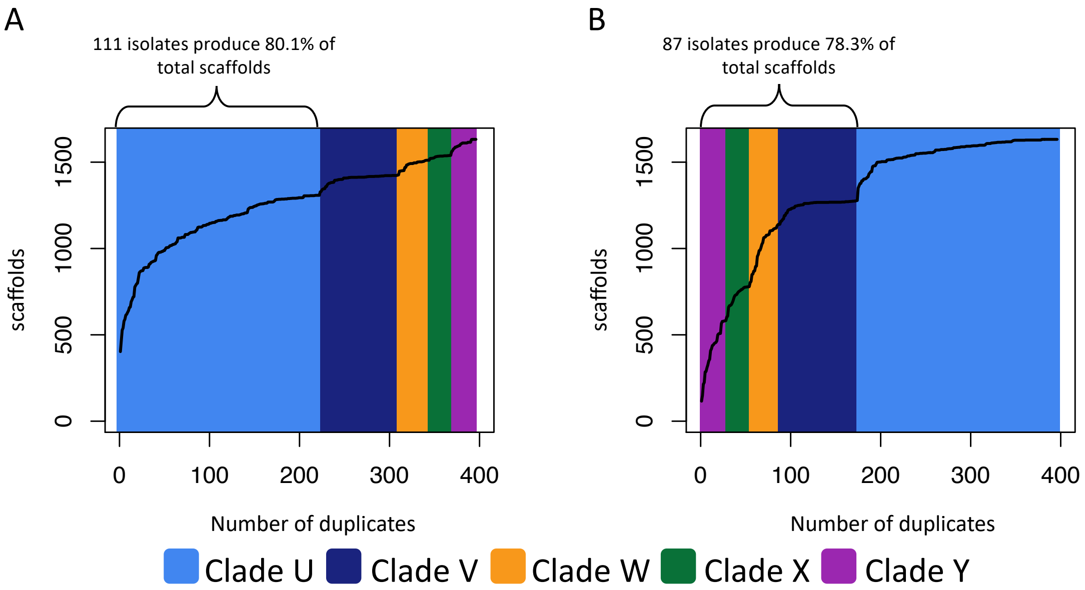


Figure 3.6 Examining scaffold accumulation of Alternaria. (A) Feature accumulation curve ordered by clade (U, V, W, X, Y). (B) Feature accumulation curve ordered by clade (Y, X, W, V, U).

3.3 Conclusions and Future Directions: Putting the pieces together to create natural product chemical collections.

It is our opinion that to date, many efforts to construct natural products libraries have been based largely on opportunism and subjective reasoning rather than founded on data-driven goals and assessment. Whereas tremendous room exists to plot customized paths for building collections of secondary metabolites based on different parameters (e.g., genetic clades versus chemical clusters, features versus scaffolds), the best routes are likely to rely upon well-balanced sample collection strategies that combine appropriate amounts of chemical breadth and depth in the resultant libraries. The purpose of our effort to measure natural product diversity was to give researchers opportunities to establish goals and provide the means for assessing progress toward those goals during library development. However, such goals should also be considered in the context of bioactive compound discovery, which in many ways is a heroic game of chance. To this point, we noted that within the *Alternaria* isolates studied here, 17.9% of metabolite features were found in only a single culture. Thus, overly stringent measures aimed at simply capturing just the core metabolome of genetic clades or chemical clusters risk missing outstanding pools of unique chemical matter that may prove critical for the success of a drug discovery program. We hope that these methods will help researchers set library building goals that are not only economical, but are also well poised to deliver the chemical matter needed to drive fruitful drug discovery operations.

3.4 Materials and Methods

3.4.1 General sample selection and culture.

A subset of 198 fungal isolates from the University of Oklahoma, Citizen Science Soil Collection that had been identified as *Alternaria* were used in this study (Supplemental Table 3.2). A map illustrating the sites where the isolates were obtained was generated in qGIS v 3.10. The fungal isolates were identified based on BLASTN[105] comparisons of their ITS sequence data to the sequences of *Alternaria* type strains deposited in GenBank[105]. When cultured on Petri plates containing a modified potato dextrose agar, all isolates were determined to be consistent with the gross morphological features of *Alternaria* spp. For metabolomics experiments, the isolates were cultured for 3 weeks in duplicate, on a solid-state medium composed of Cheerios® breakfast cereal supplemented with a 0.3% sucrose solution containing 0.005% chloramphenicol[106].

3.4.2 PCR and phylogenetic tree building.

Fungal cell lysates were prepared by removing fresh mycelium from each isolate and placing the samples in microcentrifuge tubes containing 200 μL Tris-EDTA buffer (10 mM Tris-

HCl, 1 mM disodium EDTA, pH 8.0) and a 1:1 mixture of 1 mm and 0.5 mm zirconium oxide bead. Samples were homogenized using a BulletBlender® (Next Advantage) set at maximum speed for 5 minutes. The 5.8S-ITS region was amplified by PCR using primers ITS1 5′-TCCG-TAGGTGAACCTGCGG-3′ and ITS4 5′-TCCTCCGCTTATTGATATGC-3'[107]. Amplification and confirmation of PCR product formation was performed using a LightCycler 480 Instrument II (Roche) operated under the following conditions: 1 cycle of denaturation at 94 °C for 2 minutes followed by 40 cycles of denaturation at 94 °C for 1 minute, annealing at 50 °C for 1 minute, and extension at 72 °C for 1 minute. Samples were submitted to Genewiz for Sanger sequencing and forward and reverse reads were assembled using PhredPhrap (release #29) (minimum phred score 50)[108-109]. Sequences were used for phylogenetic analysis using MEGA-X[110]. ITS sequences for *Alternaria* type strains were obtained from the NCBI database (Supplemental Table 3.1)[105]. An outgroup consisting of five *Penicillium* spp. and five *Clonostachys* spp. isolates retrieved from the University of Oklahoma, Citizen Science Soil Collection were used for tree rooting. Sequences were aligned using clustalW in Mega X. Neighbor joining tree analysis was carried out with 500 bootstraps using Kimura2+G algorithm[110-111].

3.4.3 Metabolite sample preparation.

Samples for fungal metabolome analysis were prepared on an automated platform that combined both extraction and partitioning steps. Fungal cultures prepared in 16 × 100 mm borosilicate tubes were placed on a Tecan Freedom EVO® platform and 3 mL of ethyl acetate was added to each sample. After extracting for 4 hours, 3 mL of water were added to each tube to facilitate the partitioning process. Aliquots consisting of 2 mL of the upper ethyl acetate layers were transferred to deep-well 96 well plates. While the ethyl acetate was being removed from the samples *in vacuo*, the fungal culture tubes were each charged with an additional 3 mL of ethyl acetate to continue the partitioning process. The plates were returned to the liquid handler platform at which point a second set of 2 mL aliquots of ethyl acetate was removed from the tubes and deposited into the deep-well 96 well plates. The organic solvent was removed *in vacuo* and the remaining organic residues were stored at -20 °C for liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis.

3.4.4 LC-MS/MS analysis.

Extracts were resuspended in 135 μL of 9:1 methanol-water spiked with 0.5 μM sulfadimethoxine, which served as an internal standard. Samples were analyzed on a Thermo Fisher Scientific Vanquish Flex Binary LC system, coupled to a Thermo Fisher Q Exactive Plus hybrid quadrupole-orbitrap mass spectrometer, using a C_{18} LC column (Kinetex, 50 x 2.1 mm, 1.7 μm particle size, 100 Å pore size, Phenomenex, Torrance, USA). The mobile phase consisted of LCMS-grade acetonitrile and water (Fisher Optima; both eluents contained 0.1% formic acid). Sample elution was performed using a gradient system starting with 5% acetonitrile (held for 1 minute), which was increased to 100% acetonitrile over 8 minutes, and held at 100% acetonitrile for 2 minutes. Between samples, the eluent was returned to 5% acetonitrile over 30 seconds and held for 1 minute before the next injection occurred. The column compartment and autosampler were held at 40 °C and 10 °C, respectively, for the duration of the analysis. Sample injection volumes of 5 μL were used, and samples were introduced in random order. Blanks and pooled quality control samples were interspersed throughout the analysis after every 12 samples. Electrospray conditions and data acquisition parameters are detailed in Supplemental Table 3.3.

3.4.5 Data processing and analyses.

Data were processed using MZmine v2.33 with the parameters provided in Table S4[112]. Data for the aligned peaks were exported from MZmine. All features identified as occurring in controls (blanks) and test samples were removed, and the remaining features were normalized to the total ion current (TIC) in the R statistical package. Principal coordinate analysis (PCoA) and hierarchical clustering were performed on normalized tabulated data with QIIME1[113] using a Bray-Curtis distance metric[114]. The selection of 6 clusters was determined to be optimal based on a silhouette plot. Results were visualized using Emperor[115]. Silhouette analysis is used to determine how similar a data point, in this case each extract, is to the other datapoints within its own cluster as compared to other clusters. The closer the silhouette score is to 1, the better the model fits the data. In this case, the silhouette analysis was applied to the data considering between 2 and 13 clusters. The peak average silhouette score was highest when the data was grouped into 6 clusters, thus the selection of that model.

Feature accumulation curves were made in Vegan using binarized tabulated data[116], and plots were generated using a standard x-axis representing the whole data set. Extrapolated rarefaction curves were generated in iNEXT with an endpoint of 500 duplicates.[117-119] Alpha diversity (observed chemical richness) was calculated using the Python package Scikit-Bio (version 0.2.0, http://scikit-bio.org) and analyzed using a one-way ANOVA and Tukey's HSD test in R[120]. To ensure that the differences in sample size did not skew analyses, balanced sets of randomly generated sample were analyzed for alpha diversity. Venn analyses were conducted using http://bioinformatics.psb.ugent.be/webtools/Venn/ and InteractiVenn[121]. GNPS feature-based molecular networking was performed[47, 99] using output from MZmine2[112] with the parameters described in Table S5. These parameters were modeled on those used by McCall et al. which used the same instrument and method in their study.[122] The network was then used to condense the features into scaffolds. This was accomplished by collapsing each subnetwork so it could be considered as a whole rather than an assemblage of features. The code may be found on GitHub.

3.4.6 Data availability.

LC-MS/MS data were deposited in MassIVE under accession number MSV000083002. The feature-based molecular networking method is accessible at: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f0608e9f1e0f4f3cb4d67bf16308e897. Sequencing data were deposited in GenBank under accession numbers MW729050 - MW729257. Codes for other analysis methods can be accessed on GitHub at https://github.com/NPDG/Alternaria.

3.5 Acknowledgments

We thank citizen scientists P. Allford, T. Amundson, R. Anderson, C. Bookless, M. Borden, C. Burleson, D. Clinkenbeard, S. Dayley, L. Dettman, H. Dong, C. England, B. Estep, A. Field, H. Frye, R. Fu, L. Gartz, G. Guilinger, T. Hall, A. Henrie, S. Kyle, S. Leak, J. Lee, L. Libero, V. Massaro, M. McDougal, W. Meyers, S. Morford, D. Nyer, M. Oras, A. Pool, E. Robinson, L. Sandell, C. Schlote, L. Stephenson, L. Tyson, R. Utley, A. Vaish, N. Valencia, A. Wilhoite, J. Williamson, C. Winchester, and others for their generous contributions of soil samples used to obtain the *Alternaria* isolates examined in this study. We appreciate the help of C. Coker for citizen science program and data management. The authors acknowledge R. da Silva (University of São Paulo) for sharing the ClusterApp software.

Chapter 4. Assessing Metabolic and Biological Diversity to Support Natural Product Library Assembly

This chapter was adapted from a manuscript which is currently being prepared for submission in 2021. The authors are Victoria Anderson, Karen Wendt, Fares Z. Najar, James B. Caughron, Hagan Matlock, Nitin Rangu, Andrew N. Miller, Mark R. Luttenton, and Robert H. Cichewicz. The work presented in this chapter was conducted as follows: Victoria Anderson fungal culture, fungal extract preparation, LC-MS-MS data collection, mentoring of students who performed growth analysis, and metabolomics data analysis.

4.1 Introduction

The search for bioactive natural products has brought researchers to virtually every part of the globe.[123-127] While these quests have yielded many pharmaceutical agents, they have also provided an incredible wealth of knowledge about the structures, functions, and formation of natural products. The past and ongoing successes of translating natural products into drug leads have helped continue fueling new discovery efforts, which today include microbes[128], plants[129], and marine life[130] from nearly every major environmental system around the globe.

Reflecting on the range of efforts applied to sampling organisms from locations far and wide, it would seem that such exertions would be based on rigorous scientific knowledge demonstrating that previously unsampled locations offer unique opportunities for accessing new natural products. Whereas such arguments are verified for many types of macroscale flora (e.g., trees, lianas, and more) and fauna (e.g., sponges, tunicates, and more) that live in circumscribed environments, the same cannot be readily said for many microorganisms. In many ways, the lifestyles of fungi and bacteria remain poorly understood with the natural ranges of most species not defined.[131-133] Further complicating these matters, is that fact that the biosynthetic genes responsible for producing many types of natural products are swapped among microorganisms enabling some natural product scaffolds to encircle the globe, even though their host organisms occupy more restricted geographical ranges.[134-137] Thus, for many free-living microorganisms, it is difficult to predict where they might exist and which natural products they produce.

Addressing such questions is important to the field of microbial natural products research and drug discovery because one of the prevailing ideas within the field is that free-living microbes obtained from different environmental sources are assumed to produce distinctive types of natural products. This type of logic has been used to justify the pursuit of microbes and their natural products from many types of curious and extraordinary locations, but in most situations, the rationale supporting those decisions remains untested.

To help develop an evidence-based understanding whether microorganisms obtained from dissimilar environments generate different sets of natural products, we compared fungi from two distinct ecosystems, sediments from the Great Lakes, USA, and soils from the surrounding states. Additionally, our study examined how fungal biodiversity compared between these two systems for the purpose of identifying candidate fungi that differentiate microbial communities. These results are expected to help enhance the efficiency of microbial natural products library building and drug discovery efforts through the judicious exploration of the microbial communities that occupy dissimilar ecosystems.

4.2 Results and Discussion

4.2.1 Study Design and Selection of Fungal Isolates.

To assess the role that the environment might play in the process of influencing the selection of metabolomes that reflect adaption to specific ecological niches, we needed to identify a suitable set of juxtaposed ecological systems that had experienced a sustained period of stabile partitioning. The Great Lakes of North America and the land surrounding them (generally characterized as the Eastern Temperate Forest ecoregion) were deemed to be a fitting study site (Figure 4.1) since they represent ecologically divergent environments that have persisted for 10,000-12,000 years following the retreat of the Laurentide Ice Sheet. An examination of isolates from the University of Oklahoma Citizen Science Soil Collection Program[84-85] (source of "terrestrial" samples) and sediments from Lake Michigan[138] (source of "aquatic" samples) enabled us to identify several candidate species that co-occurred in the two locations. We ultimately identified a subset of 79 isolates, which based on ITS taxonomic analysis (Supporting Information Figure 4.1), were consistent with type strains and sequences reported in GenBank for *Penicillium brevicompactum* (12 terrestrial and 13 aquatic), *Penicillium expansum* (13 terrestrial and 13 aquatic), and *Penicillium oxalicum* (14 terrestrial and 14 aquatic) (Figure 4.1 and Supporting Information Figure 4.1, Table S1).

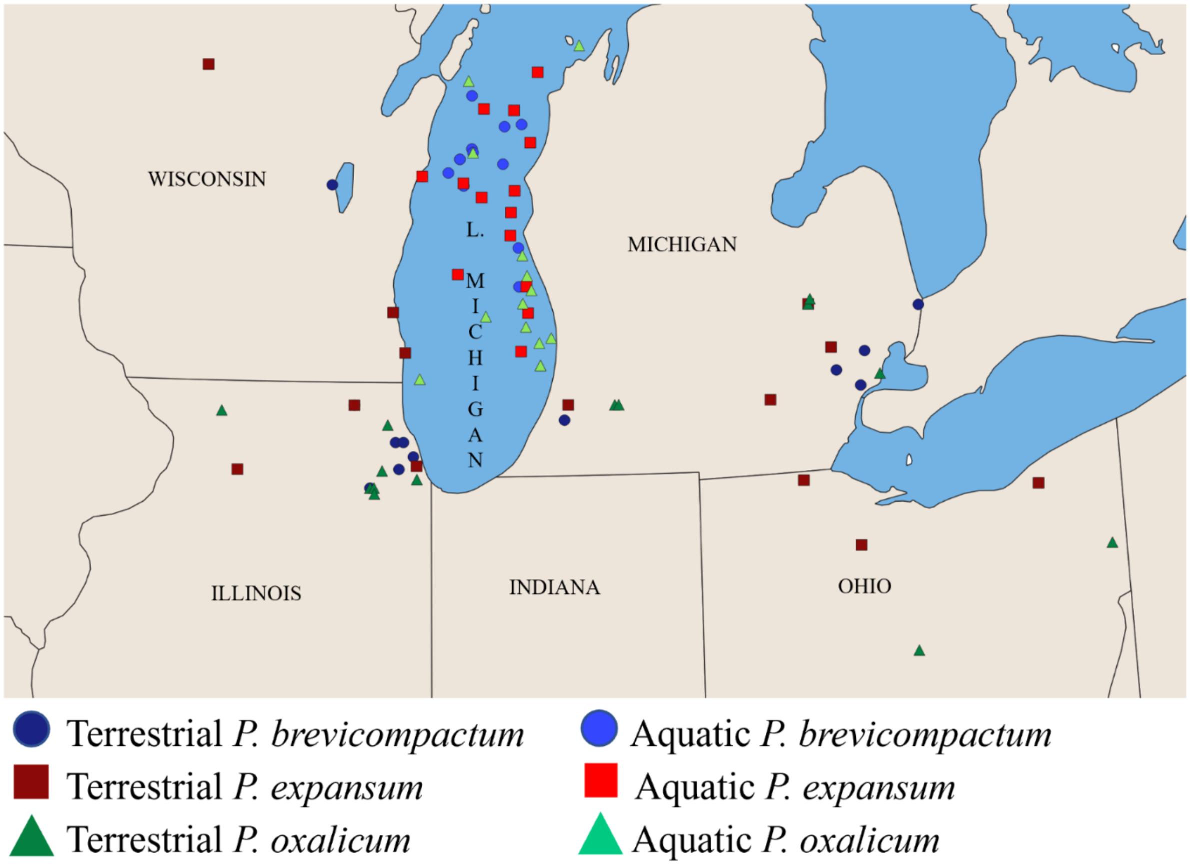

WISCONSIN
MICHIGAN
L. MICHIGAN
ILLINOIS
INDIANA
OHIO
● Terrestrial *P. brevicompactum*
● Aquatic *P. brevicompactum*
■ Terrestrial *P. expansum*
■ Aquatic *P. expansum*
▲ Terrestrial *P. oxalicum*
▲ Aquatic *P. oxalicum*

Figure 4.1. Location of sampling sites for Penicillium isolates from Lake Michigan and the states sur-rounding Lake Michigan. Species are indicated by marker shape and color (P. brevicompactum: blue circles, P. expansum: red squares, P. oxalicum: green triangle) and environmental sources are indicated by color saturation (darker = terrestrial, lighter = aquatic).

4.2.2 Phenotype Assessment of Fungal Isolates.

The question was raised whether fundamental differences in physiological characteristics may exist resulting from adaptive changes within the fungal populations occupying these distinctive environments. Looking at the gross morphological characteristics of the fungi, no intraspecific variation was observed within each of the three species groups (data not shown). To further probe the fungi for evidence of phenotypic variation, we focused on a key quantifiable variable, temperature, and its relationship to growth rate. This was accomplished by monitoring the colony sizes of triplicate samples prepared from each of the isolates grown at 4 °C (mimicking the average temperature of the benthic environment in Lake Michigan) and 20 °C (representing average summer temperatures of soils in portions of the southern Lake Michigan basin region). Two of the isolate sets achieved significant differences in colony diameter: *P. brevicompactum* isolates from the Great Lakes grew to greater colony diameters compared to terrestrial samples at 20 °C and *P. expansum* isolates from the Great Lakes grew to greater colony diameters compared to terrestrial samples at 4 °C (Figure 4.2). No significant differences were observed for the *P. oxalicum* sample set at either temperature. It is notable that the results of the test with *P. brevicompactum* proved to be contrary to expectations that fungi from the Great Lakes might be better able to grow at colder temperatures; however, we suspect that our surprise might be the product of naïve assumptions, as well as the need to disentangle the multifaceted influence of a single environmental variable (i.e., temperature) on a complex fungal physiological process (i.e., colony diameter). Nevertheless, the results of this experiment hinted at possible physiological divergence occurring within two of the three fungal species used in this study.

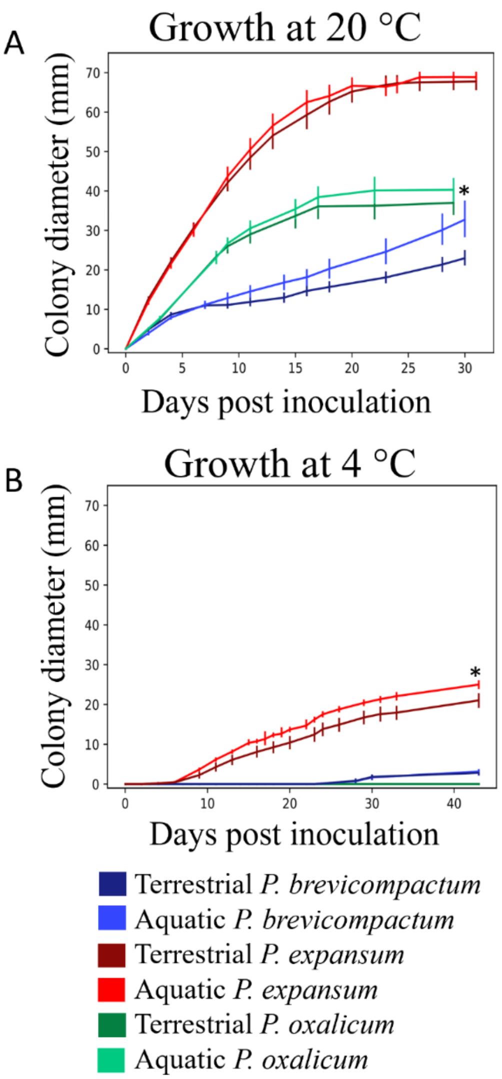


*Figure 4.2. Growth curves for Penicillium isolates, grouped by species and source. Error bars indicate 95% confidence intervals. Species are indicated by color (P. brevicompactum: blue, P. expansum: red, P. oxalicum: green) and environmental sources are indicated by color saturation (darker = terrestrial, lighter = aquatic). (A) Colony diameters of Penicillium isolates at 20°C. * p<0.0001 between aquatic and terrestrial P. brevicompactum (B) Colony diameters of Penicillium isolates at 4 °C. * p<0.0001 between aquatic and terrestrial P. expansum at 20 °C.*

4.2.3 Probing Metabolomics Feature Data.

Metabolomics provides a snapshot of the global small-molecule output representing the results of physiological and biochemical processes occurring in living organisms. Those processes are dependent on the complex influence of biotic and abiotic factors (e.g., genetics, evolution, environment, and life history) on organisms. One of the quantifiable factors that we examined concerning the fungal metabolomes was chemical richness. Measuring chemical richness can provide insight into the diversity of metabolites present in a population; however, it does not account for potential differences in their relative abundances. To evaluate chemical richness, the feature data for terrestrial and aquatic isolates from each fungal species were compared. No significant differences were detected between the sediment and soil derived isolates within *P. brevicompactum* (*p*=0.062), *P. expansum* (*p*=0.257), and *P. oxalicum* (*p*=0.361) indicating that neither environmental source produced a statistically greater number of metabolites. These results are intriguing because one might anticipate that greater metabolic heterogeneity would be found among fungi from a seasonally varying terrestrial environment as compared to the more constant conditions experience in the benthic habitat of Lake Michigan.

Whereas chemical richness is a useful tool for comparing the numbers of metabolites between sample cohorts, it does not address whether the types of metabolites in those groups are similar or different. Therefore, we performed a non-metric dimensional scaling analysis of the LC-MS/MS-derived feature data representing the metabolomes of the aquatic and terrestrial isolates to determine if their metabolic outputs varied based on the location from which the isolates were obtained (Figure 4.3). The result showed no significant intraspecific resolution occurred based on the different environmental sources for *P. brevicompactum* and *P. oxalicum* (PERMANOVA r^2 = 4.21% and 4.23% respectively, p value = 0.26 and 0.17 respectively), while *P. expansum* showed a correlation between metabolic profile and isolation source (PERMANOVA r^2 = 5.84%, p value = 0.018). Although statistical significance was reached in the case of *P. expansum*, the r^2 value indicates less than 6% of the variance in metabolic profiles can be attributed to the environmental conditions. A visual inspection of the overlap of the features of these extracts (Supporting Information Figure 4.2) reveal that while each environment does produce organisms that produce different chemistry, a majority of the chemistry is shared between the environments. In the case of *P. expansum*, isolates from the terrestrial environment cover over 85% of the features detected and aquatic, over 80% (Supporting Information Figure 4.2). This suggests that the amount of unique chemistry available in differing environments is not as high as often assumed.

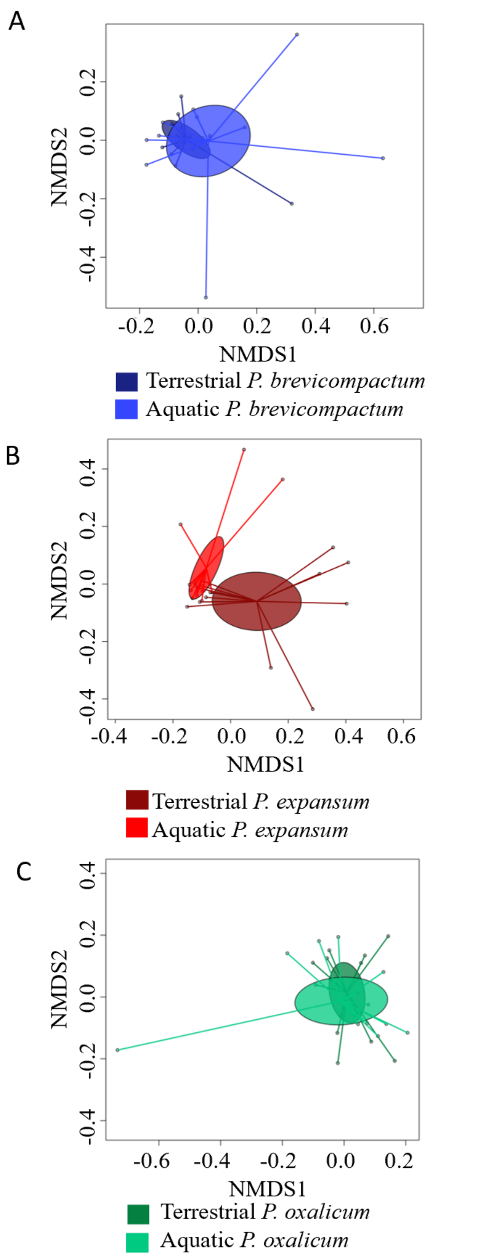


Figure 4.3. Non-metric dimensional scaling (NMDS) analysis of metabolomic features detected in aquatic and terrestrial Penicillium spp. using a Bray-Curtis matrix. The centroid of each group is indicated by the point where the lines converge while the circle delineates the 95% confidence interval of the SE. (A) Chemical profile of P. brevicompactum*. (B) Chemical profile of* P. expansum*. (C) Chemical profile of* P. oxalicum

A key hypothesis heading into this study had been the environmental disparities between the sampling sites would generate different adaptive pressures resulting in the selection for traits leading to dissimilarities in the metabolic outputs among the different isolate sets. However, the lack of difference in metabolic output for two of the three species studied here demonstrates that a more nuanced approach to sourcing isolates for natural product library building is necessary (ie. the assumption that different environments produce different chemistry is not applicable to all fungal species). Several reasons may explain the metabolic homogeneity between the groups such as a lack of substantive adaptive pressures on genetic traits within fungal populations, higher than anticipated retention of metabolic plasticity, greater than expected mobility of isolates between samples sites, the possibility that some fungi persist in benthic settings only as viable propagules, and more. Further investigation will be required to identify the factors contributing to the processes shaping the metabolomes of the isolates and the roles that divergent source environments play in influencing the chemical output of these fungi. While the reasons behind the metabolic consistency exhibited by the majority of the fungi from the two environmental niches remains unknown, we noted that the results conflict with some aspects of conventional wisdom used to justify strategies for sourcing organisms to expand chemical diversity in natural product collections. For example, it has been suggested that fungi from unusual or niche environments (e.g., acidic lakes[139-143], mines[144-145], caves[146], marine[125, 147] and freshwater[148] sediments, and more[123, 149-150]) offered value in the form of access to distinctive sets of natural products. While our results are limited in scope, they do indicate that such ideas may not be readily generalizable to all fungi.

4.2.4 Scaffold-Based Data to Informs Chemical Library Design.

Natural products are frequently encountered as sets of analogues that share underlying chemical structures called scaffolds. In general, scaffolds may be considered the principal products of coordinated sets of biosynthetic process (e.g., natural product biosynthetic gene clusters) with the contributing effects of accessory[151-152] or tailoring enzymes[153-154], kinetically[155] or thermodynamically[156] favorable organic chemical processes, biosynthetic 'stutter'[157] or off-loading[158], and more[159-160], contributing to the generation of structurally divergent analogues. Whereas such analogues may afford evolutionary advantages to host organisms[161-162] and natural product chemists intent on identifying new bioactive compounds[163-164], the multitudinous presentation of metabolic products derived from just a handful of biosynthetic pathways can create an unbalanced understanding of the actual scaffold-level chemical diversity within organisms. For these reasons, we shifted to using scaffold-based measurements as a complementary means for assessing and comparing the metabolomes of fungi from terrestrial and aquatic systems.

Venn diagrams were created for each of the *Penicillium* spp. scaffold-based datasets revealing high levels of chemical overlap between the isolates obtained from the two environmental sources (Figure 4.4). A total of 83%, 81%, and 78% of scaffolds were shared by terrestrial and aquatic isolates of *P. brevicompactum*, *P. expansum*, and *P. oxalicum*, respectively. In all case, the terrestrial isolates produced slightly elevated levels of uniqe scaffolds compared to aquatic isolates for *P. brevicompactum* (10% of scaffolds were unique to terrestrial isolates versus 7% unique to aquatic isolates), *P. expansum* (11% of scaffolds were unique to terrestrial isolates versus 8% unique to aquatic isolates), and *P. oxalicum* (14% of scaffolds were unique to terrestrial isolates versus 8% unique to aquatic isolates).

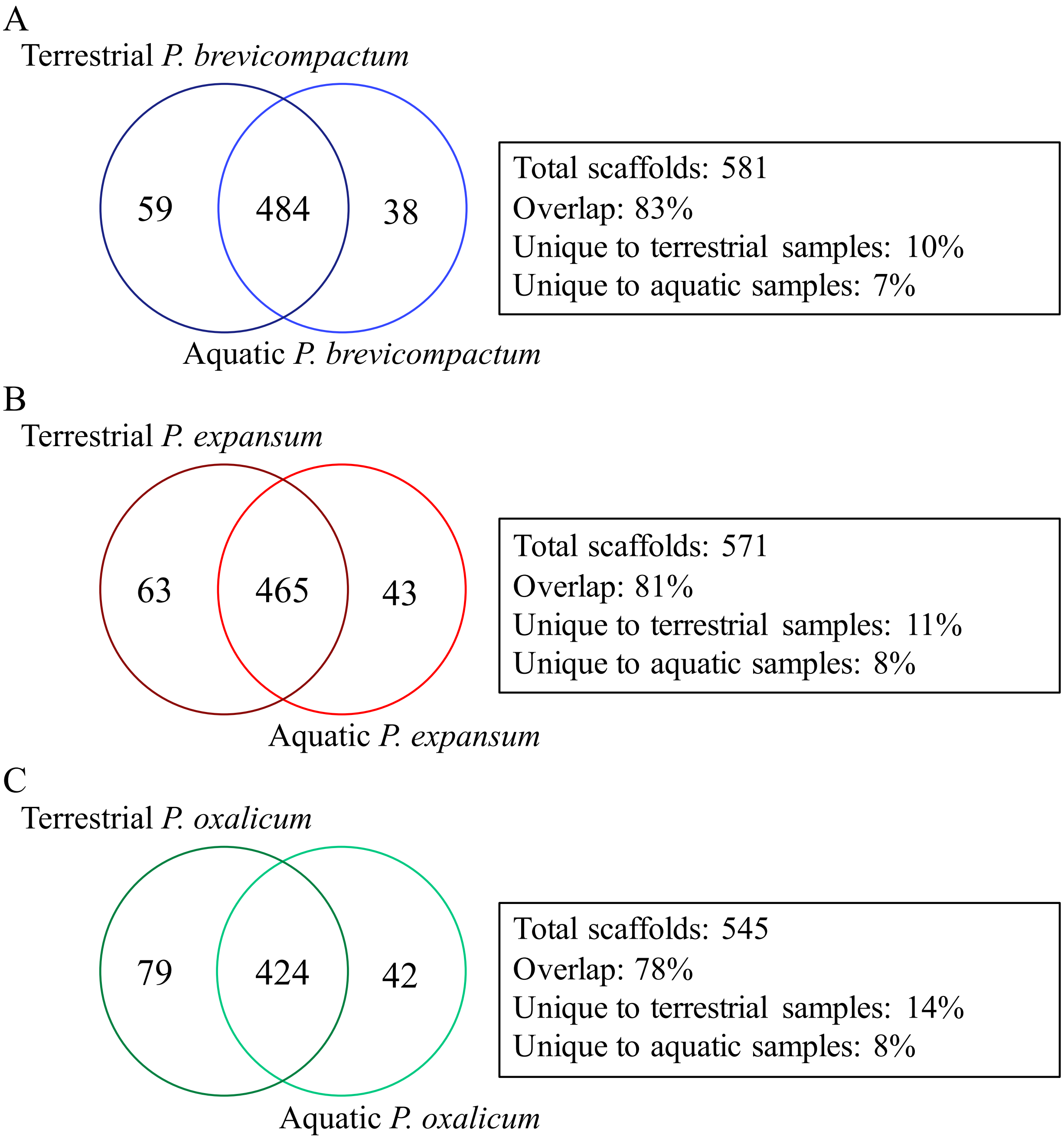

A
Terrestrial P. brevicompactum
59
484
38
Aquatic P. brevicompactum
Total scaffolds: 581
Overlap: 83%
Unique to terrestrial samples: 10%
Unique to aquatic samples: 7%
B
Terrestrial P. expansum
63
465
43
Aquatic P. expansum
Total scaffolds: 571
Overlap: 81%
Unique to terrestrial samples: 11%
Unique to aquatic samples: 8%
C
Terrestrial P. oxalicum
79
424
42
Aquatic P. oxalicum
Total scaffolds: 545
Overlap: 78%
Unique to terrestrial samples: 14%
Unique to aquatic samples: 8%

Figure 4.4. Examination of overlap of chemical scaffolds. Venn diagrams of scaffolds detected in each species collected from two source environments.

The scaffold data were further analyzed using collector's curves to model the effects of what might happen if a natural product library were constructed using isolates from just a single source environment. Whereas the conditions surrounding the theoretical need to limit collections to a single environment may appear enigmatic, the practical challenges of dealing with real and apparent barriers (e.g., costs of travel and collecting, limits imposed by geopolitical boarders, and more) can in certain situations limit the breadth of natural product exploration. For this reason, the collector's curves were used to understand how mining single environmental sources might impact the scaffold diversity of metabolites. It was observed that most of the scaffold diversity within each of the three fungal species was accessible through samples taken from just a single source (Figure 4.5). This was surprising given that conventional wisdom suggested that efforts to procure samples from alternative environments were justifiable based on the need to access pools of otherwise untapped chemical diversity. Currently, we do not know if this trend holds true for other fungi and organism types; however, these data do support the idea that large proportions of metabolite scaffold diversity may be attainable through the exploration of a single environment.

These results suggest that laboratory culture of organisms on a consistent medium from different environments result in similar chemical profiles. This trend may be disrupted if the culture conditions were altered to mimic the environment from which these organisms originated. As this is one of the first studies that compare the chemical profiles across environments, a common and easily implemented culture method was selected. Further study would be necessary to probe the influence of culture conditions on the chemical profiles across environments.

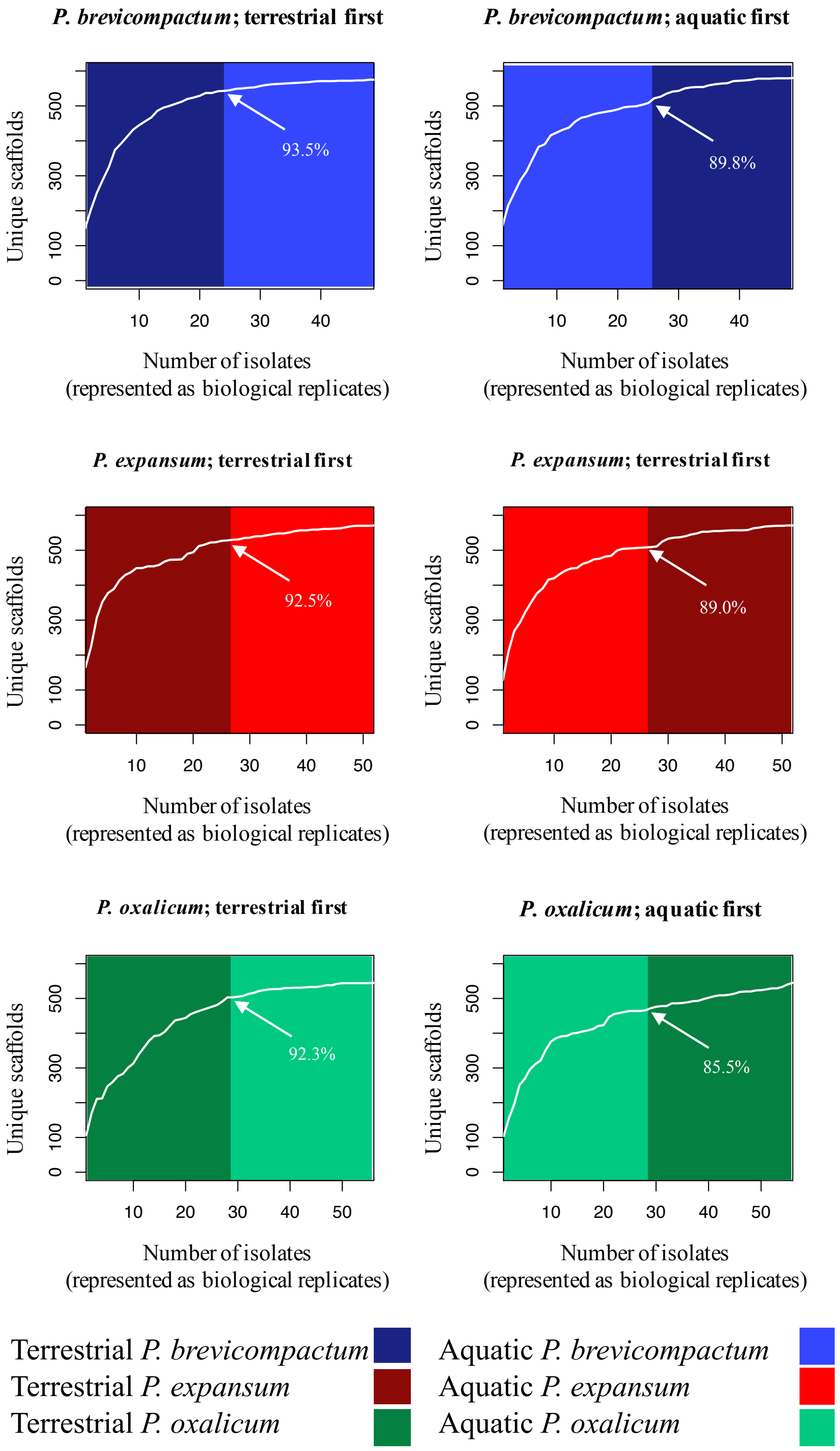


Figure 4.5. Examination of accumulation of scaffolds. Scaffold accumulation curves of each species with terrestrial samples appearing first (left column). Scaffold accumulation curves of each species with aquatic samples appearing first (right column).

4.2.5 Fungal Biological Diversity in Aquatic and Terrestrial Environments.

The analysis of chemical features and scaffolds among the three *Penicillium* spp. from terrestrial and aquatic environments showed a high degree of overlap suggesting that species-level fungal taxa capable of occupying both systems might not offer the most favorable opportunities for accessing unique or niche-specific metabolites. This led to the idea that a method aimed at identifying the organisms that are unique relative to another environmental niche might serve as a better option for increasing the likelihood of encountering new chemical scaffolds. A total of 3196 and 4183 isolates from the aquatic and terrestrial environments, respectively, were included in this analysis. To initiate that search, bar graphs illustrating the percentages of fungi associated with different classes indicated a remarkable degree of similarity in the community structure for both the aquatic and terrestrial system (Supporting Information, Figure 4.3, Table 4.1). However, when the data were further analyzed at the order level, it uncovered substantial disparities between the fungi occupying the two environments (Figure 4.6, Supporting Information Table 4.2).

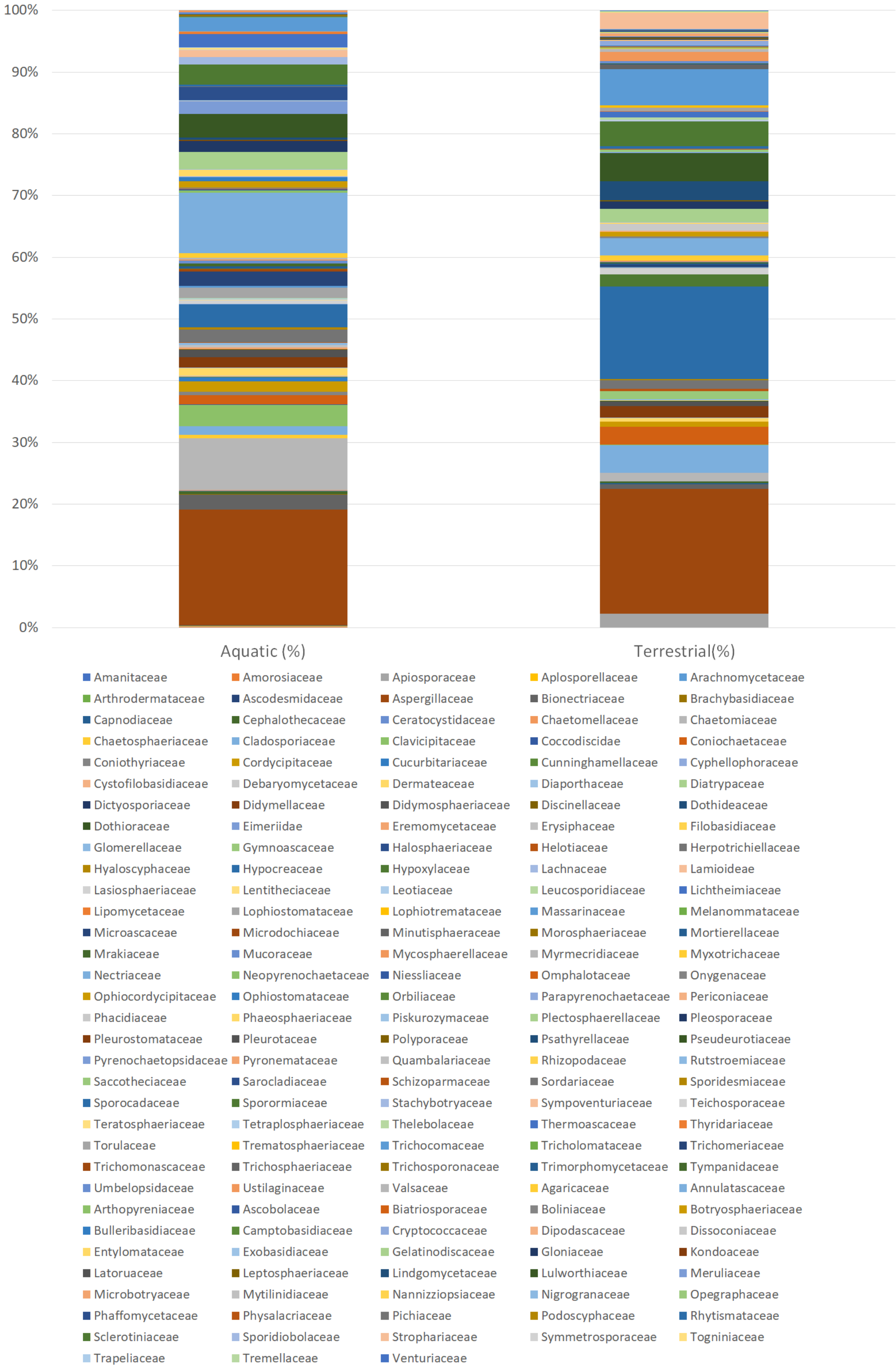


Figure 4.6. Summary of fungal families by environmental source: aquatic and terrestrial. Fungi that could not be identified at the family level were removed (58 out of 317 genera from the aquatic environment and 40 out of 328 genera from the terrestrial environment).

To determine which fungi served as the drivers of this disparity, the genus-level assignments for all isolates were analyzed using a modified volcano plot, which enabled the identification of fungal specimens that served to strongly differentiate the two environments (Figure 4.7). This method uncovered several fungal genera that numerically dominated their respective isolate pools, and served to differentiate the culturable fungal communities of the Great Lakes and the surrounding terrestrial areas. Substantially greater numbers of *Trichoderma* isolates were obtained from aquatic sediments followed by *Talaromyces, Pseudeurotium, Cladosporium, Preussia, Coprinellus, Arthrinum, Hypoxylon, Gymnoascus,* and *Philota*. In comparison, isolates from *Penicillium, Fusarium, Pseudogymnoascus, Acremonium, Humicola, Aspergillus, Metarhizium, Pyrenochaetopsis, Sporomia*, and *Chaetomium* served as the dominate culturable species recovered from soil samples originating from the surrounding terrestrial settings. Although these fungal genera and their constitutive species were not found exclusively in the locations referenced above (Supporting Information, Table 4.3), these results help draw attention to the types of fungi that exhibit higher levels of taxonomic diversity in a particular environment. Such results could be used to help selectively mine for organisms that may potentially harbor niche-specific compounds or elevated levels of metabolic diversity and thereby help improved the chemical coverage of natural product libraries.

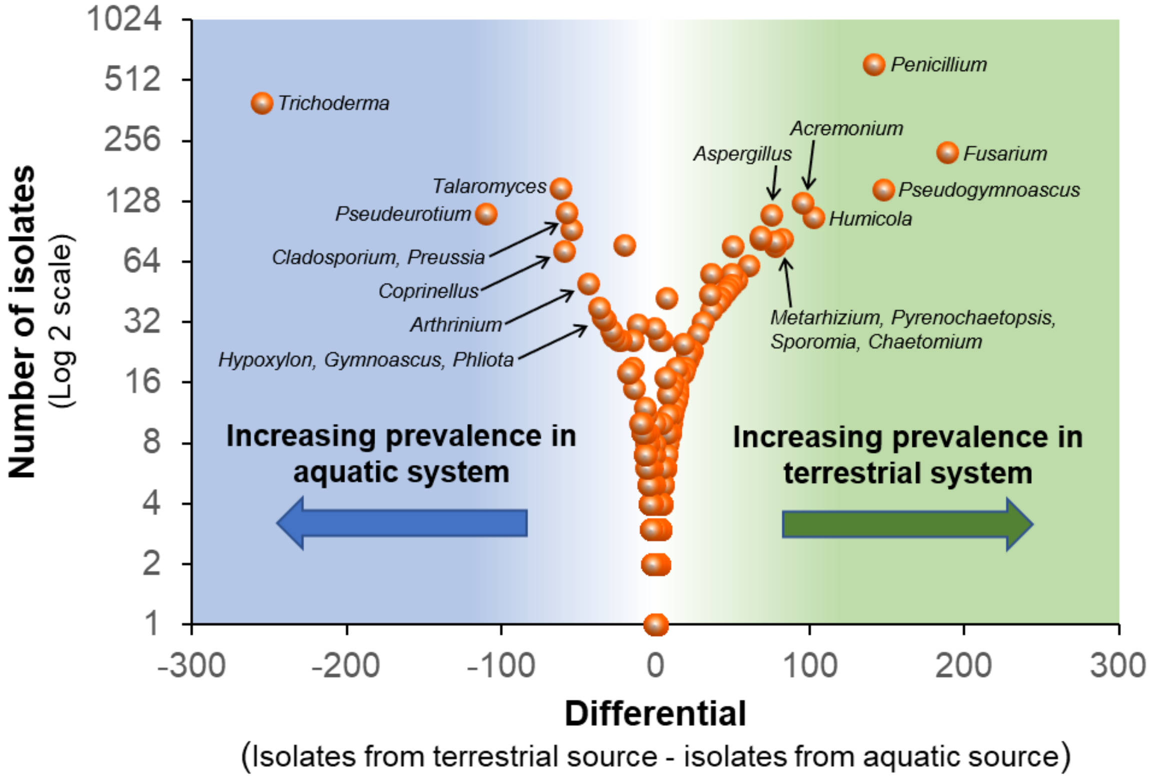


Figure 4.7. Prevalence of isolates in either aquatic or terrestrial system. Genera of fungi from aquatic and terrestrial environments are plotted with the difference between environment on the x-axis and the number of isolates (log2) on the y-axis. Isolates at the center of the plot are represented equally or near equally in both environments. Genera identified at the upper left and upper right are more prevalent in the aquatic or terrestrial system, respectively.

4.3 Conclusions and further directions

From these results, the emphasis on ubiquitous organisms from different environments does not seem to hold the most value in maximizing the chemical diversity of the library. If the goal is maximizing over all chemical diversity, a combination of a scaffold level analysis and focusing on organisms not found in other environments may be appropriate. Despite the similarity of the chemical profiles of ubiquitous organisms, the community of culturable organisms from these environments provide an alternative strategy for identifying additional chemical diversity in libraries generated from multiple environments.

4.3 Materials and Methods

4.3.1 Fungal isolates.

The “aquatic” fungi used in this study were collected from sediment samples collected in Lake Michigan, USA. The corresponding “terrestrial” fungi were procured from soil samples obtained through the University of Oklahoma, Citizen Science Soil Collection. The fungi were identified based on BLASTN[105] comparisons of their ITS-sequences to type strain data for *Penicillium brevicompactum*, *Penicillium expansum*, and *Penicillium oxalicam* that are available in GenBank.[105] A list of the isolates used in this study along with their identification codes, source location data, and GenBank accession numbers is provided in the Supporting Information (Table 4.4). A map illustrating the sites from which the fungi were obtained was generated in qGIS v 3.10 and is shown in Figure 4.1.

4.3.2 Growth measurements.

Isolates were cultured on Petri plates containing MEA medium (malt extract 10 g, yeast extract 1 g, gellan gum 7.5 g, $CaCl_2$ 0.5 g, H_2O 1 L) in triplicate. Plates were incubated in the dark at either 4 °C or 20 °C. Colony diameters were measured with a ruler and the data plotted in Python using the Seaborn package[165] with ANOVA and Tukey’s HSD calculated in R.[120]

4.3.3 PCR and phylogenetic tree building.

Fungal cell lysates were generated by adding a small quantity of mycelium from each isolate to a microcentrifuge tube with 200 μL PBS buffer (137 mM NaCl, 2.7 mM KCl, 10 mM

Na_2HPO_4, 1.8 mM KH_2PO_4) and a 1:1 (vol:vol) mixture of 1 mm and 0.5 mm zirconium oxide bead. Samples were homogenized using a BulletBlender® (Next Advantage) at maximum speed for 5 minutes. The ITS1-5.8S-ITS2 region was PCR-amplified and sequenced using primers ITS1 5′-TCCGTAGGTGAACCTGCGG-3′ and ITS4 5′-TCCTCCGCTTATTGATATGC-3'[107]. Amplification and confirmation of PCR product formation was performed in a LightCycle 480 Instrument II (Roche) using the following conditions: 1 cycle of denaturation at 94 °C for 2 minutes followed by 40 cycles of denaturation at 94 °C for 1 minute, annealing at 50 °C for 1 minute, and extension at 72 °C for 1 min. After treatment with EXOSAPit, samples were then submitted for Sanger sequencing by GENEWIZ. Sequences were used for phylogenetic analysis using MEGA-X.[110] ITS sequences for *P. brevicompactum, P. expansum, and P. oxalicum* type strains (accession numbers NR_121299.1, NR_077154.1, NR_121232.1 respectively)and three *Beauveria* species (NR_077147.1, NR_151832.1, NR_111595.1) were obtained from the NCBI database to root the tree.[105] Sequences were aligned using clustalW in Mega X. Maximum likelihood tree analysis was carried out with 500 bootstraps using Kimura2+G algorithm.[110-111] The tree then was visualized in Evolview.[166]

4.3.4 Metabolite sample preparation.

For metabolomics experiments, isolates were cultured for 3 weeks in duplicate, in borosilicate test tubes (16 × 100 mm) on a solid-state medium composed of Cheerios® breakfast cereal supplemented with a 0.3% sucrose solution containing 0.005% chloramphenicol.[106] Cultures were extracted with 3 mL ethyl acetate for 4 hours before being partitioned against 3 mL water on a Tecan Freedom EVO® platform. Aliquots consisting of 2 mL of the upper ethyl acetate layer were transferred to deep well 96 well plates. The ethyl acetate was removed from the samples *in vacuo*, and each of the cultures were subjected to a second round of partitioning following the addition of an additional 3 mL of ethyl acetate. After 2 hours, 2 mL aliquots of the ethyl acetate layer from second round of partitioning were removed and transferred to 96-well plates. The solvent was removed *in vacuo* and the samples stored in a freezer at -20 °C until liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis was performed.

4.3.5 LC-MS/MS analysis.

Samples were suspended in 200 μL of 9:1 methanol-water and sonicated. The plates were centrifuged for 5 minutes at 16,000 rpm and the supernatants were transferred to new 96-well plates. Samples were analyzed on a Thermo Fisher Scientific Vanquish Flex Binary LC system using a C_{18} column (Accucore, 100 x 2.1 mm, 1.5 μM particle size, 80 Å pore size, Thermo Fisher Scientific Inc., Waltham, USA), which was coupled to a Thermo Fisher LTQ mass spectrometer. Mobile phases were LC-MS-grade acetonitrile and water (both with 0.1% formic acid). Gradient elution was performed as follows: 10% acetonitrile held for 0.5 minutes before increasing to 95% acetonitrile over the course of 7 minutes, and again, increasing to 100% acetonitrile over 0.5 minutes. The gradient was held for 0.5 minutes at 100% acetonitrile before returning to 10% acetonitrile over 0.5 minutes. The column was held at 10% acetonitrile for a 2 minute equilibration period before the next injection. The column compartment and autosampler were maintained at 40 °C and 10 °C, respectively, for the duration of the analysis. Samples were injected (5 μL aliquots) in a randomly assigned order. Control samples consisting of culture medium and MeOH blanks, as well as pooled quality control samples were run after every 12 samples. Electrospray conditions and data acquisition parameters are provided in the Supporting Information (Table 4.5).

4.3.6 Data processing and analyses.

Data were exported and processed using MZmine2.33[112] with the parameters described in Supporting Information Table 4.6. Features identified as appearing in controls (medium only) and solvent blanks were removed from the sample data sets. The remaining features were sorted by source, species, and species-source for conversion to a presence-absence data matrix. Bray-Curtis distance matrices were constructed from the tabulated data for each species group with the function “vegdist” in Vegan.[116-167] Matrices were used to perform non-metric dimensional scaling with the metaMDS function and visualized with the ordihull function in Vegan.[116] Feature accumulation curves were prepared in Vegan using tabulated data[116], and plots were generated using a uniform x-axis representing the whole data set. Alpha diversity (observed chemical richness) was calculated using the Python package Scikit-Bio (version 0.2.0, http://scikit-bio.org) and analyzed using a one-way ANOVA and Tukey’s HSD test in R.[120] GNPS feature-based molecular networking[47, 99] was performed on the peak list and MS^2 data derived from MZmine2[112] using the parameters described in the Supporting Information, Table 4.7.

4.3.7 Data availability.

LC-MS/MS data were deposited in MassIVE under accession number MSV000087143. Feature-based molecular networking methods are accessible at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=572d43f9657f451e92b930e2ddecd62a. Sequencing data were deposited under accession number MZ362513 - MZ362590. Codes created for data analysis are accessible on GitHub at https://github.com/NPDG/GreatLakes. .

4.4 Acknowledgements

We thank citizen scientists for their generous contributions of soil samples used to obtain the *Penicillium* isolates examined in this study. We appreciate the help of C. Coker for citizen science program and data management.

References

1. Shulaev, V., Metabolomics technology and bioinformatics. *Briefings in bioinformatics* **2006,** *7* (2), 128-139.
2. Wang, X.; Li, L., Mass Spectrometry for Metabolome Analysis. *Mass Spectrometry Letters* **2020,** *11* (2), 17-24.
3. Ullsten, S.; Danielsson, R.; Bäckström, D.; Sjöberg, P.; Bergquist, J., Urine profiling using capillary electrophoresis-mass spectrometry and multivariate data analysis. *Journal of Chromatography A* **2006,** *1117* (1), 87-93.
4. Dettmer, K.; Hammock, B. D., Metabolomics--a new exciting field within the" omics" sciences. *Environmental health perspectives* **2004,** *112* (7), A396-A397.
5. Shen, J.; Yan, L.; Liu, S.; Ambrosone, C. B.; Zhao, H., Plasma Metabolomic Profiles in Breast Cancer Patients and Healthy Controls: By Race and Tumor Receptor Subtypes. *Translational Oncology* **2013,** *6* (6), 757-765.
6. Timischl, B.; Dettmer, K.; Kaspar, H.; Thieme, M.; Oefner, P. J., Development of a quantitative, validated Capillary electrophoresis ‐ time of flight ‐ mass spectrometry method with integrated high ‐ confidence analyte identification for metabolomics. *Electrophoresis* **2008,** *29* (10), 2203-2214.
7. Jones, O. A.; Maguire, M. L.; Griffin, J. L.; Dias, D. A.; Spurgeon, D. J.; Svendsen, C., Metabolomics and its use in ecology. *Austral Ecology* **2013,** *38* (6), 713-720.
8. Worley, B.; Powers, R., Multivariate analysis in metabolomics. *Current Metabolomics* **2013,** *1* (1), 92-107.
9. Gertsman, I.; Barshop, B. A., Promises and pitfalls of untargeted metabolomics. *Journal of inherited metabolic disease* **2018,** *41* (3), 355-366.
10. Pan, J.-X.; Xia, J.-J.; Deng, F.-L.; Liang, W.-W.; Wu, J.; Yin, B.-M.; Dong, M.-X.; Chen, J.-J.; Ye, F.; Wang, H.-Y., Diagnosis of major depressive disorder based on changes in multiple plasma neurotransmitters: a targeted metabolomics study. *Translational psychiatry* **2018,** *8* (1), 1-10.
11. Gómez-Canela, C.; Miller, T. H.; Bury, N. R.; Tauler, R.; Barron, L. P., Targeted metabolomics of Gammarus pulex following controlled exposures to selected pharmaceuticals in water. *Science of The Total Environment* **2016,** *562*, 777-788.
12. Zhang, W.; Segers, K.; Mangelings, D.; Van Eeckhaut, A.; Hankemeier, T.; Vander Heyden, Y.; Ramautar, R., Assessing the suitability of capillary electrophoresis ‐ mass spectrometry for biomarker discovery in plasma ‐ based metabolomics. *Electrophoresis* **2019,** *40* (18-19), 2309-2320.
13. Lario, S.; Ramírez-Lázaro, M. J.; Sanjuan-Herráez, D.; Brunet-Vega, A.; Pericay, C.; Gombau, L.; Junquera, F.; Quintás, G.; Calvet, X., Plasma sample based analysis of gastric cancer progression using targeted metabolomics. *Scientific reports* **2017,** *7* (1), 1-10.
14. Kaever, A.; Landesfeind, M.; Feussner, K.; Mosblech, A.; Heilmann, I.; Morgenstern, B.; Feussner, I.; Meinicke, P., MarVis-Pathway: integrative and exploratory pathway analysis of non-targeted metabolomics data. *Metabolomics* **2015,** *11* (3), 764-777.
15. Souard, F.; Delporte, C.; Stoffelen, P.; Thévenot, E. A.; Noret, N.; Dauvergne, B.; Kauffmann, J.-M.; Van Antwerpen, P.; Stevigny, C., Metabolomics fingerprint of coffee species determined by untargeted-profiling study

using LC-HRMS. *Food chemistry* **2018,** *245*, 603-612.

16. Lee, J.-E.; Lee, B.-J.; Chung, J.-O.; Hwang, J.-A.; Lee, S.-J.; Lee, C.-H.; Hong, Y.-S., Geographical and climatic dependencies of green tea (Camellia sinensis) metabolites: a 1H NMR-based metabolomics study. *Journal of agricultural and food chemistry* **2010,** *58* (19), 10582-10589.

17. De Vos, R. C.; Moco, S.; Lommen, A.; Keurentjes, J. J.; Bino, R. J.; Hall, R. D., Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature protocols* **2007,** *2* (4), 778.

18. Zhang, R.; Zhou, Z., Effects of the chiral fungicides metalaxyl and metalaxyl-M on the earthworm Eisenia fetida as determined by 1H-NMR-based untargeted metabolomics. *Molecules* **2019,** *24* (7), 1293.

19. Oakman, C.; Tenori, L.; Claudino, W. M.; Cappadona, S.; Nepi, S.; Battaglia, A.; Bernini, P.; Zafarana, E.; Saccenti, E.; Fornier, M.; Morris, P. G.; Biganzoli, L.; Luchinat, C.; Bertini, I.; Di Leo, A., Identification of a serum-detectable metabolomic fingerprint potentially correlated with the presence of micrometastatic disease in early breast cancer patients at varying risks of disease relapse by traditional prognostic methods. *Annals of Oncology* **2011,** *22* (6), 1295-1301.

20. Bundy, J. G.; Willey, T. L.; Castell, R. S.; Ellar, D. J.; Brindle, K. M., Discrimination of pathogenic clinical isolates and laboratory strains of Bacillus cereus by NMR-based metabolomic profiling. *FEMS microbiology letters* **2005,** *242* (1), 127-136.

21. Dettmer, K.; Aronov, P. A.; Hammock, B. D., Mass spectrometry ‐ based metabolomics. *Mass spectrometry reviews* **2007,** *26* (1), 51-78.

22. Hu, X.; Li, H.; Tang, P.; Sun, J.; Yuan, Q.; Li, C., GC–MS-based metabolomics study of the responses to arachidonic acid in Blakeslea trispora. *Fungal Genetics and Biology* **2013,** *57*, 33-41.

23. Maciά ‐ Vicente, J. G.; Shi, Y. N.; Cheikh ‐ Ali, Z.; Grün, P.; Glynou, K.; Kia, S. H.; Piepenbring, M.; Bode, H. B., Metabolomics ‐ based chemotaxonomy of root endophytic fungi for natural products discovery. *Environmental microbiology* **2018,** *20* (3), 1253-1270.

24. Dunn, W. B.; Ellis, D. I., Metabolomics: current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry* **2005,** *24* (4), 285-294.

25. Pirhaji, L.; Milani, P.; Leidl, M.; Curran, T.; Avila-Pacheco, J.; Clish, C. B.; White, F. M.; Saghatelian, A.; Fraenkel, E., Revealing disease-associated pathways by network integration of untargeted metabolomics. *Nature methods* **2016,** *13* (9), 770-776.

26. Nielsen, K. F.; Smedsgaard, J., Fungal metabolite screening: database of 474 mycotoxins and fungal metabolites for dereplication by standardised liquid chromatography–UV–mass spectrometry methodology. *Journal of Chromatography A* **2003,** *1002* (1-2), 111-136.

27. Uka, V.; Moore, G. G.; Arroyo-Manzanares, N.; Nebija, D.; De Saeger, S.; Diana Di Mavungu, J., Secondary metabolite dereplication and phylogenetic analysis identify various emerging mycotoxins and reveal the high intra-species diversity in Aspergillus flavus. *Frontiers in microbiology* **2019,** *10*, 667.

28. Zhang, A.; Sun, H.; Wang, P.; Han, Y.; Wang, X., Modern analytical techniques in metabolomics analysis. *Analyst* **2012,** *137* (2), 293-300.

29. Aerts, J. T.; Louis, K. R.; Crandall, S. R.; Govindaiah, G.; Cox, C. L.; Sweedler, J. V., Patch Clamp Electrophysiology and Capillary Electrophoresis–Mass Spectrometry Metabolomics for Single Cell Characterization. *Analytical Chemistry* **2014,** *86* (6), 3203-3208.

30. Aliferis, K. A.; Cubeta, M. A.; Jabaji, S., Chemotaxonomy of fungi in the Rhizoctonia solani species complex performing GC/MS metabolite profiling. *Metabolomics* **2013,** *9* (1), 159-169.

31. Uri, C.; Juhász, Z.; Polgár, Z.; Bánfalvi, Z., A GC–MS-based metabolomics study on the tubers of commercial potato cultivars upon storage. *Food Chemistry* **2014,** *159*, 287-292.

32. Han, A.-R.; Hong, M. J.; Nam, B.; Kim, B.-R.; Park, H. H.; Baek, I.; Kil, Y.-S.; Nam, J.-W.; Jin, C. H.; Kim, J.-B., Comparison of Flavonoid Profiles in Sprouts of Radiation Breeding Wheat Lines (Triticum aestivum). *Agronomy* **2020,** *10* (10), 1489.

33. Di Guida, R.; Engel, J.; Allwood, J. W.; Weber, R. J.; Jones, M. R.; Sommer, U.; Viant, M. R.; Dunn, W. B., Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics* **2016,** *12* (5), 93.

34. Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J. L., Proposed minimum reporting standards for chemical analysis. *Metabolomics* **2007,** *3* (3), 211-221.

35. Wisselink, H. W.; Cipollina, C.; Oud, B.; Crimi, B.; Heijnen, J. J.; Pronk, J. T.; Van Maris, A. J., Metabolome, transcriptome and metabolic flux analysis of arabinose fermentation by engineered Saccharomyces cerevisiae. *Metabolic engineering* **2010,** *12* (6), 537-551.

36. Vinaixa, M.; Samino, S.; Saez, I.; Duran, J.; Guinovart, J. J.; Yanes, O., A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. *Metabolites* **2012,** *2* (4), 775-795.

37. Caesar, L. K.; Kellogg, J. J.; Kvalheim, O. M.; Cech, N. B., Opportunities and limitations for untargeted mass spectrometry metabolomics to identify biologically active constituents in complex natural product mixtures. *Journal of natural products* **2019,** *82* (3), 469-484.

38. Boysen, A. K.; Heal, K. R.; Carlson, L. T.; Ingalls, A. E., Best-matched internal standard normalization in liquid chromatography–mass spectrometry metabolomics applied to environmental samples. *Analytical chemistry* **2018,** *90* (2), 1363-1369.

39. Kang, K. B.; Woo, S.; Ernst, M.; van der Hooft, J. J.; Nothias, L.-F.; da Silva, R. R.; Dorrestein, P. C.; Sung, S. H.; Lee, M., Assessing specialized metabolite diversity of Alnus species by a digitized LC–MS/MS data analysis workflow. *Phytochemistry* **2020,** *173*, 112292.

40. Crüsemann, M.; O'Neill, E. C.; Larson, C. B.; Melnik, A. V.; Floros, D. J.; da Silva, R. R.; Jensen, P. R.; Dorrestein, P. C.; Moore, B. S., Prioritizing natural product diversity in a collection of 146 bacterial strains based on growth and extraction protocols. *Journal of natural products* **2017,** *80* (3), 588-597.

41. Chen, C.; Gonzalez, F. J.; Idle, J. R., LC-MS-based metabolomics in drug metabolism. *Drug metabolism reviews* **2007,** *39* (2-3), 581-597.

42. Zhang, Z., Too much covariates in a multivariable model may cause the problem of overfitting. *Journal of thoracic disease* **2014,** *6* (9), E196-E197.

43. Floros, D. J.; Jensen, P. R.; Dorrestein, P. C.; Koyama, N., A metabolomics guided exploration of marine natural product chemical space. *Metabolomics* **2016,** *12* (9), 145.

44. Tuttle, R. N.; Demko, A. M.; Patin, N. V.; Kapono, C. A.; Donia, M. S.; Dorrestein, P.; Jensen, P. R., Detection of natural products and their producers in ocean sediments. *Applied and environmental microbiology* **2019,** *85* (8).

45. Caesar, L. K.; Nogo, S.; Naphen, C. N.; Cech, N. B., Simplify: a mass spectrometry metabolomics approach to identify additives and synergists from complex mixtures. *Analytical chemistry* **2019,** *91* (17), 11297-11305.

46. Pan, R.; Bai, X.; Chen, J.; Zhang, H.; Wang, H., Exploring Structural Diversity of Microbe Secondary Metabolites Using OSMAC Strategy: A Literature Review. *Frontiers in Microbiology* **2019,** *10* (294).

47. Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W.-T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya P, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P.-M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.; Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. Ø.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016,** *34* (8), 828-837.

48. Schmidt, R.; Ulanova, D.; Wick, L. Y.; Bode, H. B.; Garbeva, P., Microbe-driven chemical ecology: past, present and future. *The ISME journal* **2019,** *13* (11), 2656-2663.

49. Dandapani, S.; Rosse, G.; Southall, N.; Salvino, J. M.; Thomas, C. J., Selecting, Acquiring, and Using Small Molecule Libraries for High-Throughput Screening. *Curr Protoc Chem Biol* **2012,** *4*, 177-191.

50. Franzini, R. M.; Neri, D.; Scheuermann, J., DNA-encoded chemical libraries: advancing beyond conventional small-molecule libraries. *Accounts of chemical research* **2014,** *47* (4), 1247-55.

51. Boldt, G. E.; Dickerson, T. J.; Janda, K. D., Emerging chemical and biological approaches for the preparation

of discovery libraries. *Drug discovery today* **2006,** *11* (3-4), 143-8.

52. Webb, T. R., Current directions in the evolution of compound libraries. *Current opinion in drug discovery & development* **2005,** *8* (3), 303-8.

53. Liu, R.; Li, X.; Lam, K. S., Combinatorial chemistry in drug discovery. *Curr Opin Chem Biol* **2017,** *38*, 117-126.

54. Corsello, S. M.; Bittker, J. A.; Liu, Z.; Gould, J.; McCarren, P.; Hirschman, J. E.; Johnston, S. E.; Vrcic, A.; Wong, B.; Khan, M.; Asiedu, J.; Narayan, R.; Mader, C. C.; Subramanian, A.; Golub, T. R., The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat Med* **2017,** *23* (4), 405-408.

55. Mario Geysen, H.; Schoenen, F.; Wagner, D.; Wagner, R., Combinatorial compound libraries for drug discovery: an ongoing challenge. *Nature Reviews Drug Discovery* **2003,** *2* (3), 222-230.

56. Spear, K. L.; Brown, S. P., The evolution of library design: crafting smart compound collections for phenotypic screens. *Drug discovery today. Technologies* **2017,** *23*, 61-67.

57. Lenci, E.; Trabocchi, A., Smart Design of Small-Molecule Libraries: When Organic Synthesis Meets Cheminformatics. *Chembiochem : a European journal of chemical biology* **2019,** *20* (9), 1115-1123.

58. Goodnow, R. A.; Dumelin, C. E.; Keefe, A. D., DNA-encoded chemistry: enabling the deeper sampling of chemical space. *Nature Reviews Drug Discovery* **2017,** *16* (2), 131-147.

59. Song, M.; Hwang, G. T., DNA-Encoded Library Screening as Core Platform Technology in Drug Discovery: Its Synthetic Method Development and Applications in DEL Synthesis. *Journal of medicinal chemistry* **2020,** *63* (13), 6578-6599.

60. Favalli, N.; Bassi, G.; Scheuermann, J.; Neri, D., DNA-encoded chemical libraries - achievements and remaining challenges. *FEBS letters* **2018,** *592* (12), 2168-2180.

61. Gong, Z.; Hu, G.; Li, Q.; Liu, Z.; Wang, F.; Zhang, X.; Xiong, J.; Li, P.; Xu, Y.; Ma, R.; Chen, S.; Li, J., Compound Libraries: Recent Advances and Their Applications in Drug Discovery. *Current drug discovery technologies* **2017,** *14* (4), 216-228.

62. Busby, S. A.; Carbonneau, S.; Concannon, J.; Dumelin, C. E.; Lee, Y.; Numao, S.; Renaud, N.; Smith, T. M.; Auld, D. S., Advancements in Assay Technologies and Strategies to Enable Drug Discovery. *ACS Chemical Biology* **2020,** *15* (10), 2636-2648.

63. exerts reported by William Downey, C. L. a. D. J. o. H. H. B. D. i. D. D. W. *High Throughput Screening 2010: Effective Strategies Innovative Technologies, and Use of Better Assays*; HighTech Business Decisions: HighTech Business Decisions, 2010.

64. Koehn, F. E.; Carter, G. T., The evolving role of natural products in drug discovery. *Nature reviews. Drug discovery* **2005,** *4* (3), 206-20.

65. Wagenaar, M. M., Pre-fractionated microbial samples--the second generation natural products library at Wyeth. *Molecules (Basel, Switzerland)* **2008,** *13* (6), 1406-26.

66. Thornburg, C. C.; Britt, J. R.; Evans, J. R.; Akee, R. K.; Whitt, J. A.; Trinh, S. K.; Harris, M. J.; Thompson, J. R.; Ewing, T. L.; Shipley, S. M.; Grothaus, P. G.; Newman, D. J.; Schneider, J. P.; Grkovic, T.; O'Keefe, B. R., NCI Program for Natural Product Discovery: A Publicly-Accessible Library of Natural Product Fractions for High-Throughput Screening. *ACS Chem Biol* **2018,** *13* (9), 2484-2497.

67. Eldridge, G. R.; Vervoort, H. C.; Lee, C. M.; Cremin, P. A.; Williams, C. T.; Hart, S. M.; Goering, M. G.; O'Neil-Johnson, M.; Zeng, L., High-throughput method for the production and analysis of large natural product libraries for drug discovery. *Analytical chemistry* **2002,** *74* (16), 3963-71.

68. Huggins, D. J.; Venkitaraman, A. R.; Spring, D. R., Rational Methods for the Selection of Diverse Screening Compounds. *ACS Chemical Biology* **2011,** *6* (3), 208-217.

69. Koutsoukas, A.; Paricharak, S.; Galloway, W. R. J. D.; Spring, D. R.; Ijzerman, A. P.; Glen, R. C.; Marcus, D.; Bender, A., How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space. *Journal of Chemical Information and Modeling* **2014,** *54* (1), 230-242.

70. Baltz, R. H., Marcel Faber Roundtable: is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration? *Journal of industrial microbiology & biotechnology* **2006,** *33* (7), 507-13.

71. Letzel, A. C.; Li, J.; Amos, G. C. A.; Millán-Aguiñaga, N.; Ginigini, J.; Abdelmohsen, U. R.; Gaudêncio, S. P.; Ziemert, N.; Moore, B. S.; Jensen, P. R., Genomic insights into specialized metabolism in the marine actinomycete Salinispora. *Environmental microbiology* **2017,** *19* (9), 3660-3673.

72. Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Linington, R. G., Retrospective analysis of natural products provides insights for future discovery trends. *Proceedings of the National Academy of Sciences* **2017,** *114* (22), 5601-5606.

73. Raja, H. A.; Miller, A. N.; Pearce, C. J.; Oberlies, N. H., Fungal identification using molecular tools: a primer for the natural products research community. *Journal of natural products* **2017,** *80* (3), 756-770.

74. Lawrence, D. P.; Gannibal, P. B.; Peever, T. L.; Pryor, B. M., The sections of Alternaria: formalizing species-group concepts. *Mycologia* **2013,** *105* (3), 530-546.

75. Schoch, C. L.; Seifert, K. A.; Huhndorf, S.; Robert, V.; Spouge, J. L.; Levesque, C. A.; Chen, W.; Consortium, F. B., Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences* **2012,** *109* (16), 6241-6246.

76. Lawrence, D. P.; Rotondo, F.; Gannibal, P. B., Biodiversity and taxonomy of the pleomorphic genus Alternaria. *Mycological Progress* **2016,** *15* (1), 3.

77. Woudenberg, J.; Groenewald, J.; Binder, M.; Crous, P., Alternaria redefined. *Studies in mycology* **2013,** *75*, 171-212.

78. Egidi, E.; Delgado-Baquerizo, M.; Plett, J. M.; Wang, J.; Eldridge, D. J.; Bardgett, R. D.; Maestre, F. T.; Singh, B. K., A few Ascomycota taxa dominate soil fungal communities worldwide. *Nature communications* **2019,** *10* (1), 1-9.

79. Van der Waals, J.; Korsten, L.; Aveling, T.; Denner, F., Influence of environmental factors on field concentrations ofAlternaria solani conidia above a South African potato crop. *Phytoparasitica* **2003,** *31* (4), 353-364.

80. Cai, S.; King, J. B.; Du, L.; Powell, D. R.; Cichewicz, R. H., Bioactive sulfur-containing sulochrin dimers and other metabolites from an Alternaria sp. isolate from a Hawaiian soil sample. *Journal of natural products* **2014,** *77* (10), 2280-2287.

81. Zwickel, T.; Kahl, S. M.; Rychlik, M.; Müller, M. E., Chemotaxonomy of mycotoxigenic small-spored Alternaria fungi–do multitoxin mixtures act as an indicator for species differentiation? *Frontiers in microbiology* **2018,** *9*, 1368.

82. Carter, A. C.; King, J. B.; Mattes, A. O.; Cai, S.; Singh, N.; Cichewicz, R. H., Natural-Product-Inspired Compounds as countermeasures against the liver carcinogen aflatoxin b1. *Journal of natural products* **2019,** *82* (6), 1694-1703.

83. Kim, M.-Y.; Sohn, J. H.; Ahn, J. S.; Oh, H., Alternaramide, a cyclic depsipeptide from the marine-derived fungus Alternaria sp. SF-5016. *Journal of natural products* **2009,** *72* (11), 2065-2068.

84. Du, L.; Robles, A. J.; King, J. B.; Powell, D. R.; Miller, A. N.; Mooberry, S. L.; Cichewicz, R. H., Crowdsourcing natural products discovery to access uncharted dimensions of fungal metabolite diversity. *Angewandte Chemie* **2014,** *126* (3), 823-828.

85. Jewett, M. C.; Hofmann, G.; Nielsen, J., Fungal metabolite analysis in genomics and phenomics. *Current opinion in biotechnology* **2006,** *17* (2), 191-197.

86. Morris, M. H.; Smith, M. E.; Rizzo, D. M.; Rejmánek, M.; Bledsoe, C. S., Contrasting ectomycorrhizal fungal communities on the roots of co‐occurring oaks (Quercus spp.) in a California woodland. *New Phytologist* **2008,** *178* (1), 167-176.

87. Izzo, A.; Agbowo, J.; Bruns, T. D., Detection of plot‐level changes in ectomycorrhizal communities across years in an old‐growth mixed‐conifer forest. *New Phytologist* **2005,** *166* (2), 619-630.

88. Nilsson, R. H.; Tedersoo, L.; Abarenkov, K.; Ryberg, M.; Kristiansson, E.; Hartmann, M.; Schoch, C. L.; Nylander, J. A.; Bergsten, J.; Porter, T. M., Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. *MycoKeys* **2012,** *4*, 37.

89. Woudenberg, J. H. C.; Groenewald, J. Z.; Binder, M.; Crous, P. W., Alternaria redefined. *Studies in mycology* **2013,** *75* (1), 171-212.

90. Andrew, M.; Peever, T. L.; Pryor, B. M., An expanded multilocus phylogeny does not resolve morphological species within the small-spored Alternaria species complex. *Mycologia* **2009,** *101* (1), 95-109.

91. Lawrence, D. P.; Gannibal, P. B.; Peever, T. L.; Pryor, B. M., The sections of Alternaria: formalizing species-group concepts. *Mycologia* **2013,** *105* (3), 530-46.

92. Woudenberg, J. H.; Groenewald, J. Z.; Binder, M.; Crous, P. W., Alternaria redefined. *Stud Mycol* **2013,** *75* (1), 171-212.

93. Andersen, B.; Sørensen, J. L.; Nielsen, K. F.; Gerrits van den Ende, B.; de Hoog, S., A polyphasic approach to the taxonomy of the Alternaria infectoria species-group. *Fungal Genet Biol* **2009,** *46* (9), 642-56.

94. Costa, M. S.; Clark, C. M.; Ómarsdóttir, S.; Sanchez, L. M.; Murphy, B. T., Minimizing taxonomic and natural product redundancy in microbial libraries using MALDI-TOF MS and the bioinformatics pipeline IDBac. *Journal of natural products* **2019,** *82* (8), 2167-2173.

95. Harvey, A. L., Natural products in drug discovery. *Drug discovery today* **2008,** *13* (19-20), 894-901.

96. Breinbauer, R.; Vetter, I. R.; Waldmann, H., From protein domains to drug candidates—natural products as guiding principles in the design and synthesis of compound libraries. *Angewandte Chemie International Edition* **2002,** *41* (16), 2878-2890.

97. Walsh, C. T., A chemocentric view of the natural product inventory. *Nature Chemical Biology* **2015,** *11* (9), 620-624.

98. Bode, H. B.; Bethe, B.; Höfs, R.; Zeeck, A., Big effects from small changes: possible ways to explore nature's chemical diversity. *ChemBioChem* **2002,** *3* (7), 619-627.

99. Nothias, L.-F.; Petras, D.; Schmid, R.; Dührkop, K.; Rainer, J.; Sarvepalli, A.; Protsyuk, I.; Ernst, M.; Tsugawa, H.; Fleischauer, M.; Aicheler, F.; Aksenov, A. A.; Alka, O.; Allard, P.-M.; Barsch, A.; Cachet, X.; Caraballo-Rodriguez, A. M.; Da Silva, R. R.; Dang, T.; Garg, N.; Gauglitz, J. M.; Gurevich, A.; Isaac, G.; Jarmusch, A. K.; Kameník, Z.; Kang, K. B.; Kessler, N.; Koester, I.; Korf, A.; Le Gouellec, A.; Ludwig, M.; Martin H, C.; McCall, L.-I.; McSayles, J.; Meyer, S. W.; Mohimani, H.; Morsy, M.; Moyne, O.; Neumann, S.; Neuweger, H.; Nguyen, N. H.; Nothias-Esposito, M.; Paolini, J.; Phelan, V. V.; Pluskal, T.; Quinn, R. A.; Rogers, S.; Shrestha, B.; Tripathi, A.; van der Hooft, J. J. J.; Vargas, F.; Weldon, K. C.; Witting, M.; Yang, H.; Zhang, Z.; Zubeil, F.; Kohlbacher, O.; Böcker, S.; Alexandrov, T.; Bandeira, N.; Wang, M.; Dorrestein, P. C., Feature-based molecular networking in the GNPS analysis environment. *Nat Methods* **2020,** *17* (9), 905-908.

100. Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X.; Boudreau, P. D.; Bruns, N.; Glukhov, E.; Wodtke, A.; De Felicio, R.; Fenner, A., Molecular networking as a dereplication strategy. *Journal of natural products* **2013,** *76* (9), 1686-1699.

101. Ramos, A. E. F.; Evanno, L.; Poupon, E.; Champy, P.; Beniddir, M. A., Natural products targeting strategies involving molecular networking: Different manners, one goal. *Natural product reports* **2019,** *36* (7), 960-980.

102. van Der Hooft, J. J.; Mohimani, H.; Bauermeister, A.; Dorrestein, P. C.; Duncan, K. R.; Medema, M. H., Linking genomics and metabolomics to chart specialized metabolic diversity. *Chemical Society Reviews* **2020,** *49* (11), 3297-3314.

103. da Silva, R. R.; Wang, M.; Nothias, L.-F.; van der Hooft, J. J.; Caraballo-Rodríguez, A. M.; Fox, E.; Balunas, M. J.; Klassen, J. L.; Lopes, N. P.; Dorrestein, P. C., Propagating annotations of molecular networks using in silico fragmentation. *PLoS computational biology* **2018,** *14* (4), e1006089.

104. Wandy, J.; Zhu, Y.; van der Hooft, J. J.; Daly, R.; Barrett, M. P.; Rogers, S., Ms2lda. org: web-based topic modelling for substructure discovery in mass spectrometry. *Bioinformatics* **2018,** *34* (2), 317-318.

105. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *Journal of molecular biology* **1990,** *215* (3), 403-410.

106. King, J. B.; Carter, A. C.; Dai, W.; Lee, J. W.; Kil, Y.-S.; Du, L.; Helff, S. K.; Cai, S.; Huddle, B. C.; Cichewicz, R. H., Design and application of a high-throughput, high-content screening system for natural product inhibitors of the human parasite Trichomonas vaginalis. *ACS infectious diseases* **2019,** *5* (8), 1456-1470.

107. White, T. J.; Bruns, T.; Lee, S.; Taylor, J., Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR protocols: a guide to methods and applications* **1990,** *18* (1), 315-322.

108. Ewing, B.; Green, P., Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research* **1998,** *8* (3), 186-194.

109. Ewing, B.; Hillier, L.; Wendl, M. C.; Green, P., Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome research* **1998,** *8* (3), 175-185.

110. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K., MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution* **2018,** *35* (6), 1547-1549.

111. Kimura, M., A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution* **1980,** *16* (2), 111-20.

112. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M., MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010,** *11* (1), 395.

113. Caporaso, J. G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F. D.; Costello, E. K.; Fierer, N.; Peña, A. G.; Goodrich, J. K.; Gordon, J. I.; Huttley, G. A.; Kelley, S. T.; Knights, D.; Koenig, J. E.; Ley, R. E.; Lozupone, C. A.; McDonald, D.; Muegge, B. D.; Pirrung, M.; Reeder, J.; Sevinsky, J. R.; Turnbaugh, P. J.; Walters, W. A.; Widmann, J.; Yatsunenko, T.; Zaneveld, J.; Knight, R., QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **2010,** *7* (5), 335-336.

114. Bray, J. R.; Curtis, J. T., An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs* **1957,** *27* (4), 325-349.

115. Vázquez-Baeza, Y.; Pirrung, M.; Gonzalez, A.; Knight, R., EMPeror: a tool for visualizing high-throughput

microbial community data. *GigaScience* **2013,** *2* (1).

116. Dixon, P., VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* **2003,** *14* (6), 927-930.

117. Hsieh, T.; Ma, K.; Chao, A.; Hsieh, M. T., Package 'iNEXT'. *URL* http://chao.stat.nthu.edu.tw/wordpress/software_download/(accessed *228 2017)* **2016**.

118. Chao, A.; Gotelli, N. J.; Hsieh, T.; Sander, E. L.; Ma, K.; Colwell, R. K.; Ellison, A. M., Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological monographs* **2014,** *84* (1), 45-67.

119. Hsieh, T.; Ma, K.; Chao, A.; Hsieh, M. T., Package 'iNEXT'. **2020**.

120. R Development Core Team *R: A language and environment for statistical computing*, R Foundation for Statistical Computing: Vienna, Austria, 2019.

121. Heberle, H.; Meirelles, G. V.; da Silva, F. R.; Telles, G. P.; Minghim, R., InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC bioinformatics* **2015,** *16* (1), 1-7.

122. McCall, L.-I.; Anderson, V. M.; Fogle, R. S.; Haffner, J. J.; Hossain, E.; Liu, R.; Ly, A. H.; Ma, H.; Nadeem, M.; Yao, S., Characterization of the workplace chemical exposome using untargeted LC-MS/MS: a case study. *bioRxiv* **2019**, 541813.

123. Hagestad, O. C., Bioprospecting of marine fungi from the High Arctic: A study of high latitude marine fungi from understudied taxa; bioactivity potential, taxonomy and genomics. **2021**.

124. Forcina, G. C.; Castro, A.; Bokesch, H. R.; Spakowicx, D. J.; Legaspi, M. E.; Kucera, K.; VIllota, S.; Narvaez-Trujillo, A.; McMahon, J. B.; Gustafson, K. R.; Strobel, S. A., Stelliosphareols A and B, Sesquiterpene-Polyol Conjugates from an Ecuadorian Fungal Endophyte. *Journal of Natural Products* **2015,** *78*, 3005-3010.

125. Uchoa, P. K. S.; Pimenta, A. T.; Braz-Filho, R.; de Oliveira, M. d. C. F.; Saraiva, N. N.; Rodrigues, B. S.; Pfenning, L. H.; Abreu, L. M.; Wilke, D. V.; Florêncio, K. G., New cytotoxic furan from the marine sediment-derived fungi Aspergillus niger. *Natural product research* **2017,** *31* (22), 2599-2603.

126. Singh, V. P.; Yedukondalu, N.; Sharma, V.; Kushwaha, M.; Sharma, R.; Chaubey, A.; Kumar, A.; Singh, D.; Vishwakarma, R. A., Lipovelutibols A-D: Cytotoxic Lipopeptaibols from the Himalayan Cold Habitat Fungus Trichoderma velutinum. *Journal of Natural Products* **2018,** *81*, 219-226.

127. Schmidt, S. K.; Gendron, E. M. S.; Vincent, K.; Solon, A. J.; Sommers, P.; Schubert, Z. R.; Vimercati, L.; Porazinska, D. L.; Darcy, J. L.; Sowell, P., LIfe at extreme elevations on Atacama volcanoes: the closest thing to Mars on Earth? *Antonie van Leeuwenhoek* **2017,** *111*, 1389-1401.

128. El-Elimat, T.; Raja, H. A.; Figueroa, M.; Al Sharie, A. H.; Bunch, R. L.; Oberlies, N. H., Freshwater Fungi as a Source of Chemical Diversity: A Review. *Journal of Natural Products* **2021,** *84*, 898-916.

129. Ramalhete, C.; Mansoor, T. A.; Mulhovo, S.; Molnar, J.; Ferreira, M.-J. U., Cucurbitane-Type Triterpenoids from the African Plant Momordica balsamina. *Journal of Natural Products* **2009,** *72*, 2009-2013.

130. Khushi, S.; Salim, A. A.; Elbanna, A. H.; Nahar, L.; Bernhardt, P. V.; Capon, R. J., Dysidealactams and Dysidealactones: Sesquiterpene Glycinyl-Lactams, Imides, and Lactones from a Dysidea sp. Marine Sponge Collected in Southern Australia. *Journal of Natural Products* **2020,** *83*, 1577-1584.

131. Tedersoo, L.; Bahram, M.; Põllme, S.; Kõljalg, U.; Yorou, N. S.; Wijesundera, R.; Ruiz, L. V.; Vasco-Palacios, A. M.; Thu, P. Q.; Suija, A.; Smith, M. E.; Sharp, C.; Saluveer, E.; Saitta, A.; Rosas, M.; Riit, T.; Ratkowsky, D.; Pritsch, K.; Põldmaa, K.; Piepenbring, M.; Phosri, C.; Peterson, M.; Parts, K.; Pärtel, K.; Otsing, E.; Nouhra, E.; Njouonkou, A. L.; Nilsson, R. H.; Morgado, L. N.; Mayor, J.; May, T. W.; Majuakim, L.; Lodge, D. J.; Lee, S. S.; Larsson, K.-H.; Kohout, P.; Hosaka, K.; Hiiesalu, I.; Henkel, T. W.; Harend, H.; Guo, L.; Greslebin, A.; Grelet, G.; Geml, J.; Gates, G.; Dunstan, W.; Bunck, C.; Drenkhan, R.; Dearnaley, J.; De Kesel, A.; Dang, T.; Chen, X.; Buegger, F.; Brearley, F. Q.; Bonito, G.; Anslan, S.; Abell, S.; Abarenkov, K., Global diversity and geography of soil fungi. *Science* **2014,** *346* (6213).

132. Crowther, T. W.; van den Hoogen, J.; Wan, J.; Mayes, M. A.; Keiser, A. D.; Mo, L.; Averill, C.; Maynard, D. S., The global soil community and its influence on biogeochemistry. *Science* **2019,** *365*, 772.

133. Delgado-Baquerizo, M.; Oliverio, A. M.; Brewer, T. E.; Benavent-Gonzalez, A.; Eldridge, D. J.; Bardgett, R. D.; Maestre, F. T.; Singh, B. K.; Fierer, N., A global atlas of the dominant bacteria found in soil. *Science* **2018,** *359*, 320-325.

134. Hu, J.; Chen, C.; Peever, T.; Dang, H.; Lawrence, C.; Mitchell, T., Genomic characterization of the conditionally dispensible chromosome in Alternaria arborescens provides evidence for horizontal gene transfer. *BMC Genomics* **2012,** *13*.

135. Khaldi, N.; Collemare, J.; Lebrun, M.-H.; Wolfe, K. H., Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. *Genome Biology* **2008,** *9*.

136. Sheinman, M.; Arkhipova, K.; Arndt, P. F.; Dutilh, B. E.; Hermsen, R.; Massip, F., Identical sequences found in distant genomes reveal frequent horizontal transfer across the bacterial domain. *eLife* **2021,** *10*.

137. Nongkhlaw, F. M. W.; Joshi, S. R., Horizontal Gene Transfer of the Non-ribosomal Peptide Synthetase Gene Among Endophytic and Epiphytic Bacteria Associated with Ethnomedicinal Plants. *Current Microbiology* **2016,** *72*, 1-11.

138. Wahl, H. E.; Raudabaugh, D. B.; Bach, E. M.; Bone, T. S.; Luttenton, M. R.; Cichewicz, R. H.; Miller, A. N., What lies beneath? Fungal diversity at the bottom of Lake Michigan and Lake Superior. *Journal of Great Lakes research* **2018,** *44* (2), 263-270.

139. Stierle, D. B.; Stierle, A. A.; Hobbs, J. D.; Stokken, J.; Clardy, J., Berkeleydione and berkeleytrione, new bioactive metabolites from an acid mine organism. *Organic Letters* **2004,** *6* (6), 1049-1052.

140. Stierle, D. B.; Stierle, A. A.; Patacini, B.; McIntyre, K.; Girtsman, T.; Bolstad, E., Berkeleyones and related meroterpenes from a deep water acid mine waste fungus that inhibit the production of interleukin 1-β from induced inflammasomes. *Journal of natural products* **2011,** *74* (10), 2273-2277.

141. Stierle, A. A.; Stierle, D. B.; Kelly, K., Berkelic acid, a novel spiroketal with selective anticancer activity from an acid mine waste fungal extremophile. *The Journal of organic chemistry* **2006,** *71* (14), 5357-5360.

142. Stierle, A. A.; Stierle, D. B.; Goldstein, E.; Parker, K.; Bugni, T.; Baarson, C.; Gress, J.; Blake, D., A novel 5-HT receptor ligand and related cytotoxic compounds from an acid mine waste extremophile. *Journal of natural products* **2003,** *66* (8), 1097-1100.

143. Stierle, A. A.; Stierle, D. B.; Kemp, K., Novel sesquiterpenoid matrix metalloproteinase-3 inhibitors from an acid mine waste extremophile. *Journal of natural products* **2004,** *67* (8), 1392-1395.

144. Park, H. B.; Kwon, H. C.; Lee, C.-H.; Yang, H. O., Glionitrin A, an antibiotic− antitumor metabolite derived from competitive interaction between abandoned mine microbes. *Journal of natural products* **2009,** *72* (2), 248-252.

145. Rusman, Y.; Held, B. W.; Blanchette, R. A.; Wittlin, S.; Salomon, C. E., Soudanones A–G: antifungal isochromanones from the ascomycetous fungus Cadophora sp. isolated from an iron mine. *Journal of natural products* **2015,** *78* (6), 1456-1460.

146. Belyagoubi, L.; Belyagoubi-Benhammou, N.; Jurado, V.; Dupont, J.; Lacoste, S.; Djebbah, F.; Ounadjela, F. Z.; Benaissa, S.; Habi, S.; Abdelouahid, D. E., Antimicrobial activities of culturable microorganisms (actinomycetes and fungi) isolated from Chaabe Cave, Algeria. *International Journal of Speleology* **2018,** *47* (2), 8.

147. Grunwald, A. L.; Cartmell, C.; Kerr, R. G., Auyuittuqamides A–D, Cyclic Decapeptides from Sesquicillium microsporum RKAG 186 Isolated from Frobisher Bay Sediment. *Journal of Natural Products* **2020**.

148. Abdelwahab, M. F.; Fouad, M. A.; Kamel, M. S.; Özkaya, F. C.; Kalscheuer, R.; Müller, W. E.; Lin, W.; Liu, Z.; Ebrahim, W.; Daletos, G., Tanzawaic acid derivatives from freshwater sediment-derived fungus Penicillium sp. *Fitoterapia* **2018,** *128*, 258-264.

149. Liu, C.-C.; Zhang, Z.-Z.; Feng, Y.-Y.; Gu, Q.-Q.; Li, D.-H.; Zhu, T.-J., Secondary metabolites from Antarctic marine-derived fungus Penicillium crustosum HDN153086. *Natural product research* **2019,** *33* (3), 414-419.

150. Ding, T.; Zhou, Y.; Qin, J.-j.; Yang, L.-j.; Zhang, W.-d.; Shen, Y.-h., Chemical constituents from wetland soil fungus Penicillium oxalicum GY1. *Fitoterapia* **2020,** *142*, 104530.

151. Hai, Y.; Jenner, M.; Tang, Y., Complete Stereoinversion of L-Tryptophan by a Fungal Single-Module Nonribosomal Peptide Synthetase. *Journal of the American Chemical Society* **2019,** *141*, 16222-16226.

152. Wheadon, M. J.; Townsend, C. A., Evolutionary and functional analysis of an NRPS condensation domain integrates B-lactam, D-amino acid, and dehydroamino acid synthesis. *PNAS* **2021,** *118*.

153. Patel, H. M.; Walsh, C. T., In Vitro Reconsitution of the Pseudomonas aeruginosa Nonribosomal Peptide Synthesis of Pyochelin: Characterization of Backbone Tailoring Thiazoline Reductase and N-Methyltransferase Activities. *Biochemistry* **2001,** *40*, 9023-9031.

154. Walsh, C. T., Polyketide and Nonribosomal Peptide Antibiotics: Modularity and Versatility. *Science* **2004,** *303*, 1805-1810.

155. Pfeiffer, E.; Hildebrand, A. A.; Becker, C.; Schnattinger, C.; Baumann, S.; Rapp, A.; Goesmann, H.; Syldatk, C.; Metzler, M., Identification of an Aliphatic Epoxide and the Corresponding Dihydrodiol as Novel Congeners of Zearalenone in Cultures of Fusarium graminearum. *Journal of Agricultural and Food Chemistry* **2010,** *58*, 12055-12062.

156. Elbanna, A. H.; Khalil, Z. G.; Bernhardt, P. V.; Capon, R. J., Neobulgarones Revisited: Anti and Syn Bianthrones

form an Australian Mud Dauber Wasp Nest-Associated Fungus, Penicillium sp. CMB-MD22. *Journal of Natural Products* **2021,** *84*, 762-770.

157. He, J.; Hertweck, C., Iteration as Programmed Event during Polyketide Assembly; Molecular Analysis of the Aureothin Biosynthesis Gene Cluster. *Chemistry & Biology* **2003,** *10*, 1225-1232.

158. Mullowney, M. W.; McClure, R. A.; Robey, M. T.; Kelleher, N. L.; Thomson, R. J., Natural products from thioester reductase containing biosynthetic pathways. *Natural Product Reports* **2018,** *35*, 847-878.

159. Beck, B. J.; Yoon, Y. J.; Reynolds, K. A.; Sherman, D. H., The Hidden Steps of Domain Skipping: Macrolactone Ring Size Determination in the Pikromycin Modular Polyketide Synthase. *Chemistry & Biology* **2002,** *9*, 575-583.

160. Wei, X.; Chen, X.; Chen, L.; Yan, D.; Wang, W.-G.; Matsuda, Y., Heterologous Biosynthesis of Tetrahydroxanthose Dimers: Determination of Key Factors for Selective or Divergent Synthesis. *Journal of Natural Products* **2021,** *84*, 1544-1549.

161. Sudheeran, P. K.; Ovadia, R.; Galsarker, O.; Maoz, I.; Sela, N.; Maurer, D.; Feygenberg, O.; Shamir, M. O.; Alkan, N., Glycoslated flavonoids: fruit's concealed antifungal arsenal. *New Phytologist* **2019,** *225*, 1788-1798.

162. SM, C.; JF, R.; T, E.; J, M., Mushroom chemical defense : Pungent sesquiterpenoid dialdehyde antifeedant to opossum. *Journal of Chemical Ecology* **1983,** *9*, 1439-1947.

163. Lee, J. W.; Collins, J. E.; Wendt, K. L.; Chakrabarti, D.; Cichewicz, R. H., Leveraging Peptaibol Biosynthetic Promiscuity of Next-Generation Antiplasmodial Therapeutics. *Journal of Natural Products* **2021,** *84*, 503-517.

164. King, J. B.; Carter, A. C.; Dai, W.; Lee, J. W.; Kil, Y.-S.; Du, L.; Helff, S. K.; Cai, S.; Huddle, B. C.; Cichewicz, R. H., Design and Application of a High-Throughput, High-Content Screening System for Natural Product Inhibitors of the Human Parasite Trichomonas vaginalis. *ACS Infectious Diseases* **2019,** *5*, 1456-1470.

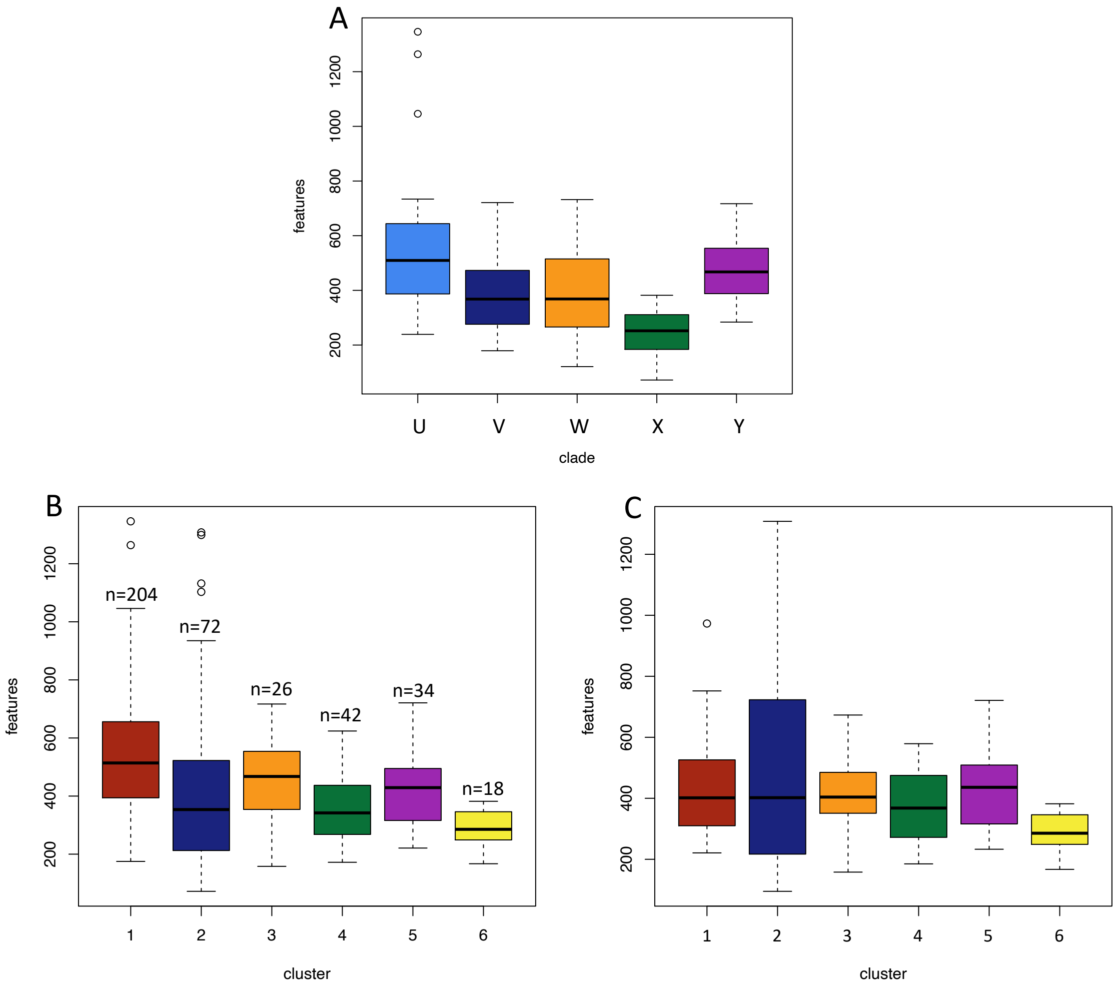
165. Waskom, M.; Botvinnik, O.; Gelbart, M.; Lukauskas, S.; Hobson, P.; Gemperline, D. C.; Augspurger, T.; Halchenko, Y.; Cole, J. B.; Warmenhoven, J.; de Ruiter, J.; Pye, C.; Hoyer, S.; Vanderplas, J.; Villalba, S.; Kunter, G.; Quintero, E.; Bachant, P.; Martin, M.; Meyer, K.; Swain, C.; Miles, A.; Brunner, T.; O'Kane, D.; Yarkoni, T.; Williams, M. L.; Evans, C.; Fitzgerald, C.; Brian. *mwaskom/seaborn: v0.10.1 (April 2020)*, v0.10.1; Zenodo: 2020.

166. Subramanian, B.; Gao, S.; Lercher, M. J.; Hu, S.; Chen, W.-H., Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Research* **2019,** *47*, W270-W275.

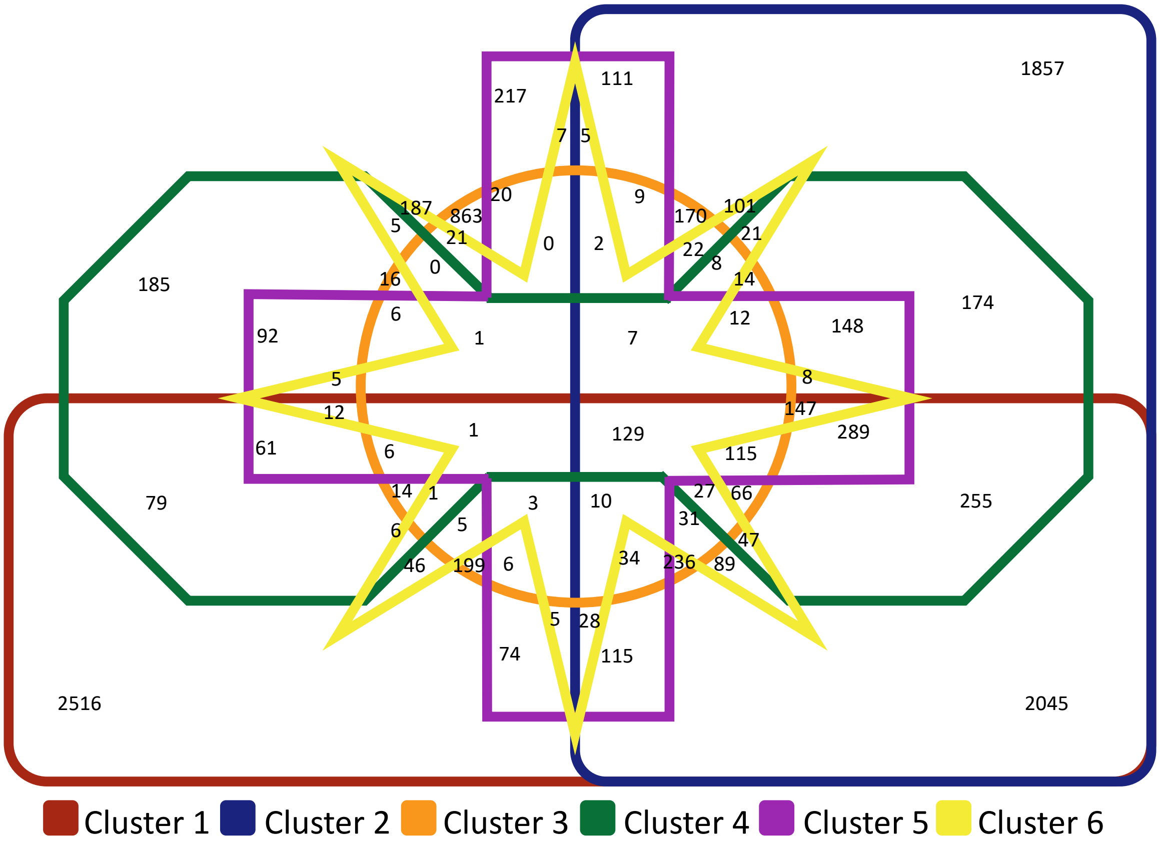
167. Bray, J. R.; Curtis, J. T., An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs* **1957,** *27* (4), 326-349.

Appendix A: Supporting Information for Chapter 3

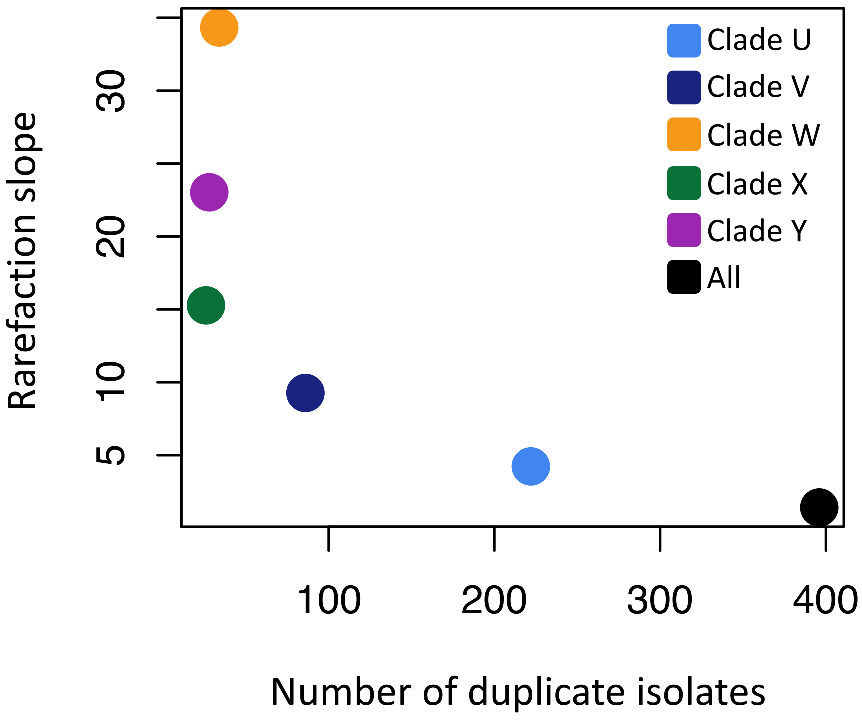
Appendix Table of Contents



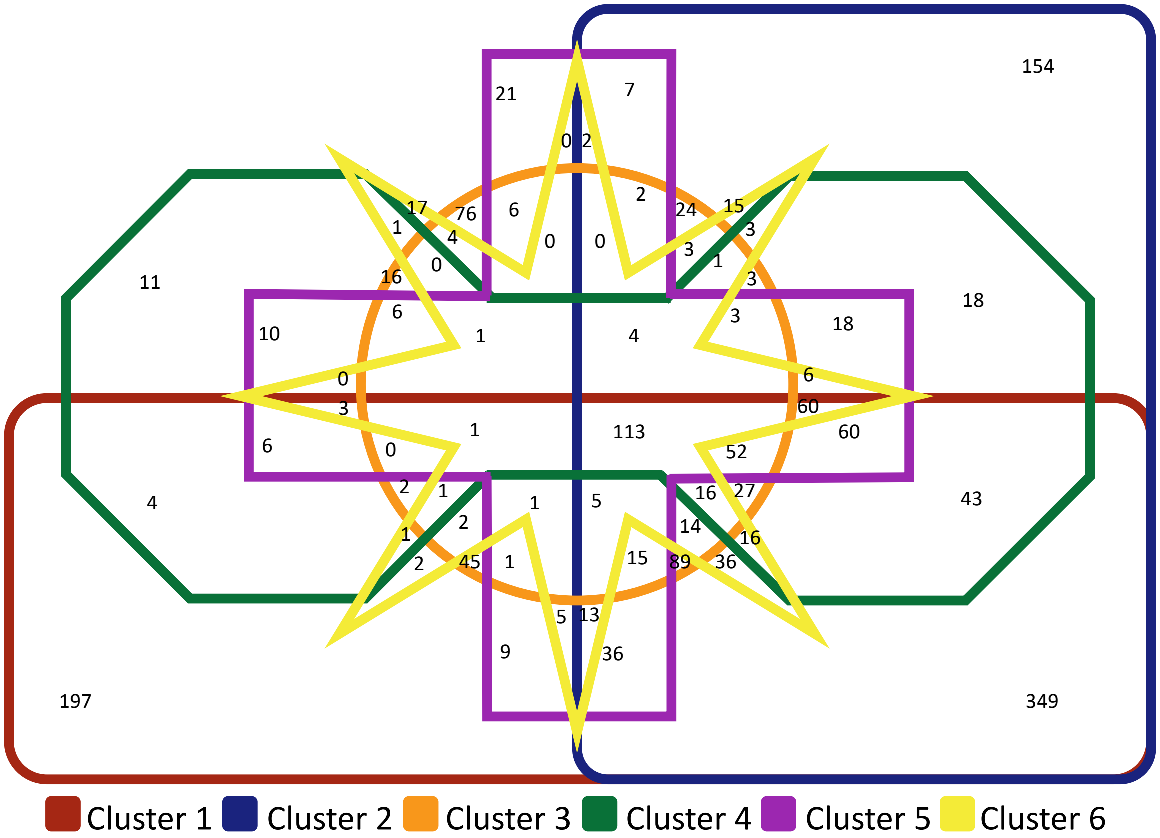
Supplemental Figure 3.1 Feature richness and diversity of Alternaria. (A) Feature count with random selection of isolates from larger clades (n=26). Significant differences in the chemical richness of clades persisted even when the sample size was sub-sampled to achieve a balanced dataset (p<0.001). (B) Feature count by chemical cluster. Chemical clusters also showed significant differences in chemical richness both when analyzed as a whole (p<0.001). (C) Feature count with random selection of isolates from larger clusters (n=18). Chemical richness of a balanced dataset (n=18) yielded significant differences between chemical clusters (p=0.0338).



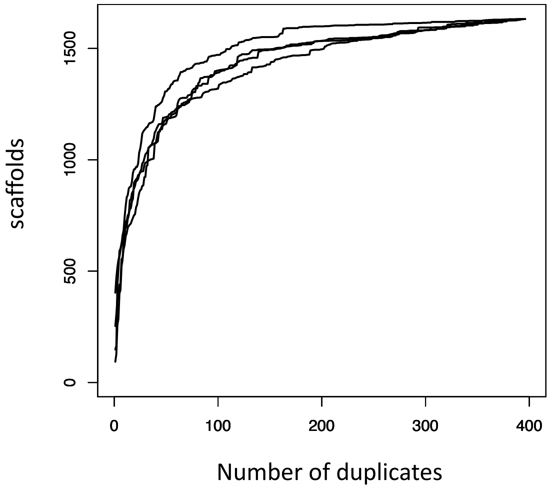
Supplemental Figure 3.2 Venn diagram of features in chemical Clusters 1-6. Feature overlap by chemical cluster is tremendously complex. Clusters were constructed based on hierarchical clustering analysis using a Bray-Curtis distance metric. There is a high degree of overlap between these clusters: 5166 (47.0%) features are shared by at least 2 chemical clusters. The remaining 5825 (53.0%) features are unique to a single cluster: 2516 (22.9%), 1857 (16.9%), 863 (7.9%), 185 (1.7%), 217 (2%), and 187 (1.7%) of features were found to be unique to Cluster 1, 2, 3, 4, 5, and 6 respectively.



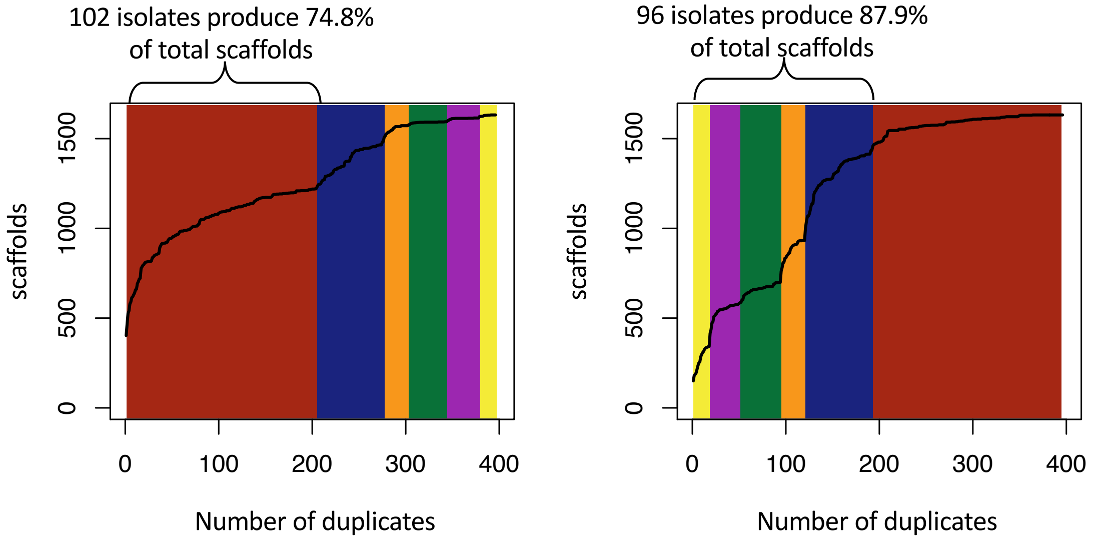
Supplemental Figure 3.3 Relationship between size of clade and proximity to chemical saturation. In addition to using extrapolated rarefaction curves (Figures 3.4A & 3.4B), the slope at the end of interpolated data in rarefaction curves reveals that larger clades have a lower slope indicating that they are closer to saturation (slope=0). Thus, the chemistry of larger clades is more fully described, and investigation of smaller clades may add more new features if sampled more extensively.



Supplemental Figure 3.4 Venn diagram of scaffolds in chemical Clades 1-6. 1185 (71.3%) scaffolds were found to be shared between at least two chemical clusters, while the remaining 476 (28.7%) scaffolds were found to be unique to a single chemical cluster. Of these scaffolds, 197 (11.9%), 154 (9.3%), 76 (4.6%), 11 (0.7%), 21 (1.3%), and 17 (1%) were found to be unique to chemical Clusters 1, 2, 3, 4, 5, and 6 respectively.



Supplemental Figure 3.5 Adaption of collector's curve for metabolomics analysis. In addition to rarefaction curves presented in Figure 3.4A and Figure 3.4B, the use of collector's curve can shed additional light on the accumulation of chemistry. Collector's curves differ from rarefaction curves in that they present the raw data as entered, while the rarefaction analysis creates a model for describing the smooth accumulation of diversity. Because this is raw data, the order of data can vastly change the shape and smoothness of the resulting curve. To illustrate the power of different arrangements of data on this method, the Alternaria dataset was randomized 4 times in Microsoft Excel and scaffold accumulation curves were generated using the collector's method in vegan. These curves were overlaid above. While the beginning and ending point of these curves, the shape between 1000 and 1400 scaffolds are quite different.



Supplemental Figure 3.6 Exploration of scaffold-level diversity within chemical clusters. A library that was constructed exclusively of isolates from the most abundant clade (Clade 1) would provide access to 74.8% of scaffolds. The addition of Clusters 2, 3, 4, 5, and 6 provide an additional 16.1%, 5.5%, 1.3%, 1.3% and 1.0%. However, if the library emphasized the smaller clusters, the 96 isolates that make up Clusters 2-6 provide access to 87.9% of total scaffolds and the addition of Cluster 1 only provides 12.1% of the total scaffolds.

Supplemental Table 3.1 Alternaria type strains identified in Genbank that were used to create ITS-based clades. The Alternaria spp. are identified by number (i.e., Number in tree) in the cladogram shown in Figure 3.1 of the manuscript.

Number in tree	Type strain	Accession number
1	*Alternaria angustiovoidea*	MH861939
2	*Alternaria cerealis*	NR_136117
3	*Alternaria arborescens*	NR_135927
4	*Alternaria daucifolii*	NR_137802
5	*Alternaria alstroemeriae*	NR_163686
6	*Alternaria destruens*	NR_137143
7	*Alternaria tropica*	MH862449
8	*Alternaria infectoria*	NR_131263
9	*Alternaria dactylidicola*	NR_151852
10	*Alternaria rosae*	NR_136017
11	*Alternaria tellustris*	NR_135961
12	*Alternaria molesta*	MH861376
13	*Alternaria lolii*	NR_159632
14	*Alternaria leptinellae*	NR_111866
15	*Alternaria hungarica*	NR_135944
16	*Alternaria hyacinthi*	NR_145168
17	*Alternaria proteae*	NR_135930
18	*Alternaria thalictrigena*	NR_135937
19	*Alternaria zantedeschiae*	NR_160245
20	*Alternaria sorghi*	NR_160246
21	*Alternaria multiformis*	NR_077187
22	*Alternaria terricola*	NR_103600

Supplemental Table 3.2 Source information for Alternaria isolates used in this analysis. Regions are NOAA regions based on the state from which each soil sample was submitted. The number of isolates in each group is indicated by a number in parentheses.

Region	State	City	Sample ID	Full ID	Cryo ID

Alaska (1)	AK (1)	Douglas (1)	106113 (1)	AK06113 RBM-3	200-A5
Central (15)	IL (2)	O'Fallon (1)	107958 (1)	IL07958 RBM+M-4	356-G10
		Oak Park (1)	106098 (1))	IL06098 RBM-1	330-C2
	MO (9)	Blue Springs (6)	104924 (1)	MO04924 GVA-3	360-F7
			104938 (1)	MO04938 RBM-3	307-A3
			104941 (3)	MO04941 CZ-4	365-C8
				MO04941 PFA-8	365-D8
				MO04941 ZMA-1	365-D9
			105221 (1)	MO05221 TV8-3	275-C3
		Lee's Summit (2)	104933 (2)	MO04933 TV8-5	361-E4
				MO04933 SEA-1	361-G5
		Saint Louis (1)	109829 (1)	MO09829 RBM-3	444-A2
	OH (3)	Dennison (1)	12530 (1)	OH2530 CZSW-8	286-G9
		Lakeside Marblehead (1)	12429 (1)	OH2429 TV8-4	197-B1
		Ravenna (1)	13669 (1)	OH3669 PDA-2	354-B6
	TN (1)	Oak Ridge (1)	108832 (1)	TN08832 RBM-2	411-C5
East North Central (8)	MI (1)	Kingsford (1)	106583 (1)	MI06583 RBM+M-4	244-G12
	MN (5)	Andover (1)	101626 (1)	MN01626 TV8-2	175-F4
		Bemidji (1)	11589 (1)	MN1589 TV8-2	253-G9
		Minneapolis (2)	11708 (2)	MN1708 BSA-1	241-F8
				MN1708 BSA-2	241-F9
		Shakopee (1)	15936 (1)	MN5936 TV8-2	170-A3
	WI (2)	Shawano (2)	105148 (2)	WI05148 RBM-4	186-F11
				WI05148 RBM-1	186-F8
Northeast (7)	CT (2)	Stamford (2)	105458 (1)	CT05458 TV8-4	372-F4
			105460 (1)	CT05460 TV8-3	319-E8
	MA (1)	Mattapan (1)	102133 (1)	MA02133 RBM-4	185-D9
	MD (1)	Sparrows Point (1)	108122 (1)	MD08122 BIA-4	282-E10
	NY (1)	Eastchester (1)	101906 (1)	NY01906 GVA-1	335-F9
	PA (2)	Allentown (1)	106591 (1)	PA06591 TV8+M-1	245-E10
		Lancaster (1)	19696 (1)	PA9696 RBM-5	175-C5
Northwest (12)	ID (3)	Post Falls (1)	19935 (1)	ID9935 RBM-1	389-H2
		Star (1)	107855 (1)	ID07855 RBM+M-1	356-F7
		Twin Falls (1)	108003 (1)	ID08003 RBM+M-6	295-C11
	OR (3)	Portland (1)	105493 (1)	OR05493 TV8-3	339-F8

		Roseburg (1)	103007 (1)	OR03007 CZ-1	358-H1
		Yachats (1)	105645 (1)	OR05645 TV8-2	236-G1
	WA (6)	Endicott (1)	106230 (1)	WA06230 TV8+M-1	279-H6
		Republic (1)	104059 (1)	WA04059 TV8-4	323-A8
		West Richland (4)	106432 (4)	WA06432 TV8-2	239-G9
				WA06432 BIA-1	239-H3
				WA06432 BIA-6	239-H7
				WA06432 BIA-4	239-H8
South (18)	KS (1)	Auburn (1)	13211 (1)	KS3211 TV8-3	165-C11
	OK (9)	Marlow (1)	106401 (1)	OK06401 TV8-1	347-D10
		Mounds (2)	105088 (2)	OK05088 RBM-3	342-G4
				OK05088 RBM-4	342-G5
		Oklahoma City (5)	102375 (1)	OK02375 RBM-4	333-E11
			104301 (3)	OK04301 TV8-7	177-E1
				OK04301 RBM-4	177-E5
				OK04301 RBM-5	177-E6
			107080 (1)	OK07080 RBM-4	310-G10
		Tecumseh (1)	10626 (1)	Tucker BIA-1	154-A4
	TX (8)	Alvin (1)	104415 (1)	TX04415 CIT-4	349-G2
		Austin (2)	106180 (2)	TX06180 RBM-1	239-D10
				TX06180 BIA-3	239-D12
		Dallas (2)	103115 (1)	TX03115 RBM-2	325-E11
			103143 (1)	TX03143 PFA-2	369-C2
		El Paso (2)	15878 (1)	TX5878 RBM-4	145-B9
			18357 (1)	TX8357 RBM-1	133-H4
		Weslaco (1)	19737 (1)	TX9737 BIA-2	380-A8
Southeast (8)	AL (2)	Birmingham (1)	12730 (1)	AL2730 BIA-2	258-B8
		Tuskegee (1)	106505 (1)	AL06505 TV8+M-2	337-B2
	FL (2)	Cape Coral (1)	15539 (1)	FL5539 TV8-3	201-E4
		Niceville (1)	105029 (1)	FL05029 TV8-1	331-B9
	NC (2)	Chapel Hill (1)	107859 (1)	NC07859 RBM-3	431-A2
		Clemmons (1)	14376 (1)	NC4376 RBM-5	404-B10
	SC (1)	Columbia (1)	15920 (1)	SC5920 TV8-1	101-H7
	VA (1)	Unionville (1)	19846 (1)	VA9846 TV8-2	175-A7
Southwest (61)	AZ (8)	Phoenix (5)	101714 (1)	AZ01714 RBM+M-5	333-A8

		105667 (3)	AZ05667 TV8-3	339-A11
			AZ05667 RBM-7	340-B7
			AZ05667 TV8-2	346-A6
		106110 (1)	AZ06110 RBM-1	310-A12
	Scottsdale (1)	105966 (1)	AZ05966 RBM-1	306-B5
	Sonoita (1)	107035 (1)	AZ07035 RBM+M-1	337-A5
	Tempe (1)	100157 (1)	AZ00157 RBM-1	199-A7
CO (8)	Alamosa (1)	103739 (1)	CO03739 RBM-5	315-H8
	Arvada (1)	17626 (1)	CO7626 TV8-3	149-A5
	Colorado Springs (1)	106011 (1)	CO06011 RBM-4	360-C1
	Denver (1)	15842 (1)	CO5842 RBM-1	136-B12
	Fort Collins (1)	11330 (1)	CO1330 RBM-3	147-B2
	Golden (1)	15585 (1)	CO5585 RBM+M-4	308-A12
	Grand Junction (1)	108658 (1)	CO08658 RBM-4	421-C11
	Grand Lake (1)	105854 (1)	CO05854 BIA-1	238-B1
NM (7)	Albuquerque (6)	16512 (2)	NM6512 RBM-2	134-D5
			NM6512 RBM-1	147-C6
		16572 (1)	NM6572 TV8-3	139-F1
		16579 (1)	NM6579 TV8-1	139-F2
		101594 (1)	NM01594 RBM-4	179-A9
		104076 (1)	NM04076 SULF-2	390-E1
	Serafina (1)	13634 (1)	NY3634 TV8-2	159-E5
UT (38)	Hyde Park (2)	107209 (2)	UT07209 TV8-3	269-E10
			UT07209 TV8-1	269-E9
	Layton (1)	107902 (1)	UT07902 TV8+M-1	296-H5
	Lindon (1)	107814 (1)	UT07814 RBM+M-1	356-B9
	Logan (15)	106978 (1)	UT06978 RBM-1	264-D2
		107120 (1)	UT07120 TV8-3	267-F12
		107129 (1)	UT07129 RBM-2	269-D2
		107162 (1)	UT07162 TV8-3	267-A5
		107164 (1)	UT07164 TV8-1	267-B12
		107193 (1)	UT07193 TV8-6	264-C2
		107195 (1)	UT07195 RBM-2	268-F1
		107285 (1)	UT07285 TV8-2	355-C2
		107825 (3)	UT07825 RBM-3	334-A8

				UT07825 TV8-6	355-B5
				UT07825 BIA-2	355-B6
			109117 (2)	UT09117 RBM-5	386-A11
				UT09117 SULF-4	386-A2
			109210 (1)	UT09210 TV8-3	383-G3
			109630 (1)	UT09630 RBM-4	444-A9
		Orderville (9)	16905 (2)	UT6905 TV8-4	194-H4
				UT6905 RBM-4	197-D9
			16917 (1)	UT6917 RBM-30	197-E10
			16918 (1)	UT6918 RBM-1	170-F4
			16921 (1)	UT6921 RBM-1	164-D9
			16925 (2)	UT6925 TV8-1	169-E8
				UT6925 TV8-2	169-E9
			16926 (1)	UT6926 RBM-1	157-H2
			16927 (1)	UT6927 RBM-1	190-D4
		Orem (1)	101299 (1)	UT01299 RBM-5	166-G4
		Paradise (2)	109111 (2)	UT09111 RBM-3	386-F1
				UT09111 RBM-4	386-F2
		Payson (1)	102892 (1)	UT02892 RBM-3	331-G7
		Provo (1)	106863 (1)	UT06863 RBM+M-2	247-G9
		Sandy (1)	12290 (1)	UT2290 RBM-2	170-E12
		Tremonton (2)	107838 (2)	UT07838 RBM+M-3	282-B3
				UT07838 TV8+M-2	282-B6
		Wellsville (1)	108880 (1)	UT08880 RBM-4	385-D2
		West Jordan (1)	106991 (1)	UT06991 SULF-3	362-C3
West (58)	CA (50)	Canyon Country (2)	104365 (2)	CA04365 RBM-7	272-A11
				CA04365 RBM-3	272-A8
		Capistrano Beach (3)	106897 (1)	CA06897 TV8-1	253-D6
			106910 (2)	CA06910 TV8-2	260-D9
				CA06910 RBM-5	262-B7
		Dana Point (3)	106912 (1)	CA06912 TV8-5	260-D5
			107639 (1)	CA07639 RBM-1	387-E1
			107649 (1)	CA07649 BIA-1	390-D4
		Dublin (2)	19212 (1)	CA9212 RBM-2	152-B2
			19443 (1)	CA9443 RBM-1	166-H2

Garden Grove (1)	100516 (1)	CA00516 RBM-2	167-A11
La Puente (1)	19633 (1)	CA9633 RBM-1	151-A8
Ladera Ranch (5)	106893 (1)	CA06893 RBM-1	260-C1
	106898 (1)	CA06898 RBM-4	268-B1
	106905 (1)	CA06905 RBM-2	266-C12
	106924 (1)	CA06924 TV8-1	267-G2
	107564 (1)	CA07564 RBM-1	387-C3
Los Alamitos (1)	106077 (1)	CA06077 TV8-4	380-H8
Marina (1)	12503 (1)	CA2503 RBM-2	178-D1
Perris (1)	105759 (1)	CA05759 RBM+M-1	280-A8
Pomona (2)	103709 (2)	CA03709 SULF-5	382-E5
		CA03709 SULF-6	382-E6
Redlands (1)	105688 (1)	CA05688 RBM-4	347-B1
Rio Linda (2)	16630 (1)	CA6630 RBM-5	168-E2
	105902 (1)	CA05902 RBM+M-3	308-G8
Riverside (1)	100382 (1)	CA00382 CEA-1	365-G7
Sacramento (3)	102293 (3)	CA02293 RBM-3	173-C6
		CA02293 RBM-4	180-B10
		CA02293 RBM-2	180-B9
San Clemente (2)	106904 (2)	CA06904 RBM-3	263-C1
		CA06904 TV8-1	263-C2
San Diego (1)	100380 (1)	CA00380 RBM-3	157-F2
San Jose (5)	16130 (2)	CA6130 CGA-1	153-E12
		CA6130 BIA-2	155-E5
	105322 (1)	CA05322 PFA-3	348-B5
	106256 (2)	CA06256 RBM-5	280-B9
		CA06256 TV8-3	307-C11
San Juan Capistrano (5)	106890 (1)	CA06890 TV8-1	387-A9
	106919 (1)	CA06919 RBM-3	268-A6
	106930 (1)	CA06930 RBM-3	266-H2
	106932 (1)	CA06932 RBM-1	268-A8
	107634 (1)	CA07634 TV8-1	386-G8
Santa Ana (1)	105894 (1)	CA05894 RBM-2	352-A7
Simi Valley (5)	100535 (5)	CA00535 MEA-4	359-A3
		CA00535 BFA-2	359-A7

				CA00535 RBM-3	359-B7
				CA00535 RBM-5	359-B9
				CA00535 CZSW-2	359-D5
		Turlock (1)	107860 (1)	CA07860 TV8-2	421-A1
		Yorba Linda (1)	100742 (1)	CA00742 TV8-1	201-B8
	NV (8)	Dayton (1)	17690 (1)	NV7690 TV8-2	158-H6
		Fallon (2)	107695 (2)	NV07695 RBM-3	316-E1
				NV07695 RBM-4	316-E2
		Las Vegas (2)	102048 (1)	NV02048 CIT-1	439-C7
			108352 (1)	NV08352 RBM-5	411-C10
		Reno (2)	107748 (2)	NV07748 CIT-1	423-E8
				NV07748 SULF-5	423-F5
		Sparks (1)	103768 (1)	NV03768 TV8-1	326-C4
West North Central (10)	MT (3)	Helena (1)	13034 (1)	MT3034 RBM-4	148-E3
		Melstone (2)	106089 (2)	MT06089 RBM-3	353-E9
				MT06089 RBM-1	366-D1
	ND (1)	Gwinner (1)	101000 (1)	ND01000 RBM-6	374-F7
	NE (2)	Chadron (1)	101209 (1)	NE01209 TV8-7	185-F7
		Hastings (1)	104278 (1)	NE04278 TV8-8	347-D12
	SD (1)	Aberdeen (1)	16748 (1)	SD6748 TV8-3	183-H4
	WY (3)	Carpenter (1)	14702 (1)	WY4702 RBM-3	396-C10
		Otto (2)	107136 (2)	WY07136 RBM-2	269-B2
				WY07136 RBM-4	269-B4

Supplemental Table 3.3 Data acquisition parameters for LC-MS/MS.

Parameter	**Value**
Data acquisition mode	positive
Scan range	100-1500 m/z
MS1 Resolution	35,000
MS 2 Resolution	17,500
Top N	5

sheath gas	35 L/min
auxiliary gas	10 L/min
sweep gas	0 L/min
auxiliary gas temperature	350 C
spray voltage	3.8 kV
S-lens RF	50 V
capillary temperature	320 C
Maximum injection time (MS1 & MS2)	100 Ms
MS1 AGC target	1E6
MS2 AGC target	5E5
Isolation window	2 *m/z*
Normalized collision energy increments	20%, 30%, 40%
MS2 dynamic exclusion	10 s
Apex trigger	2-8 s
Exclude	Unassigned charges

Supplemental Table 3.4 MZmine data processing parameters.

Process	Parameter	Value
Mass Detection	MS1 Noise Level	4.0E5
	MS2 Noise Level	6.00E+03
	Mass Detector	Centroid
Chromatogram Builder	Minimum Time Span (min)	0.01
	Minimum Height	1E7
	m/z tolerance (ppm)	10
Chromatogram Deconvolution: LOCAL MINIMA algorithm	Chromatographic threshold	20
	Search minimum in RT range (min)	.08
	Minimum relative height	26
	Minimum absolute height	1E7
	Min ratio of peak top/edge	1.19
	Peak duration range (min)	0.01-1.00
	m/z Range for MS2 Scan Pairing (Da)	0.01
	RT Range for MS2 Scan Pairing (min)	0.1
Isotopic Peak Grouper	Retention Time Tolerance (min)	0.1

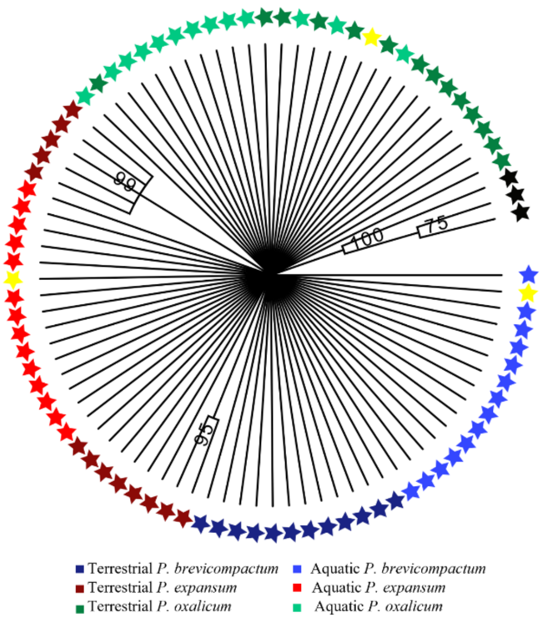
	m/z tolerance (ppm)	10
	Monotonic Shape	Yes
	Maximum Charge	3
	Representative isotope	Lowest *m/z*
Join aligner	*m/z* tolerance (ppm)	15
	m/z to RT weight	1-1
	Retention Time Tolerance (min)	0.25
Row filtering	Retention Time	0.20-12 min
	Keep only peaks with MS2 scan	Enabled
	Minimum peaks in a row	2 (for duplicates)

Supplemental Table 3.5 GNPS parameters.

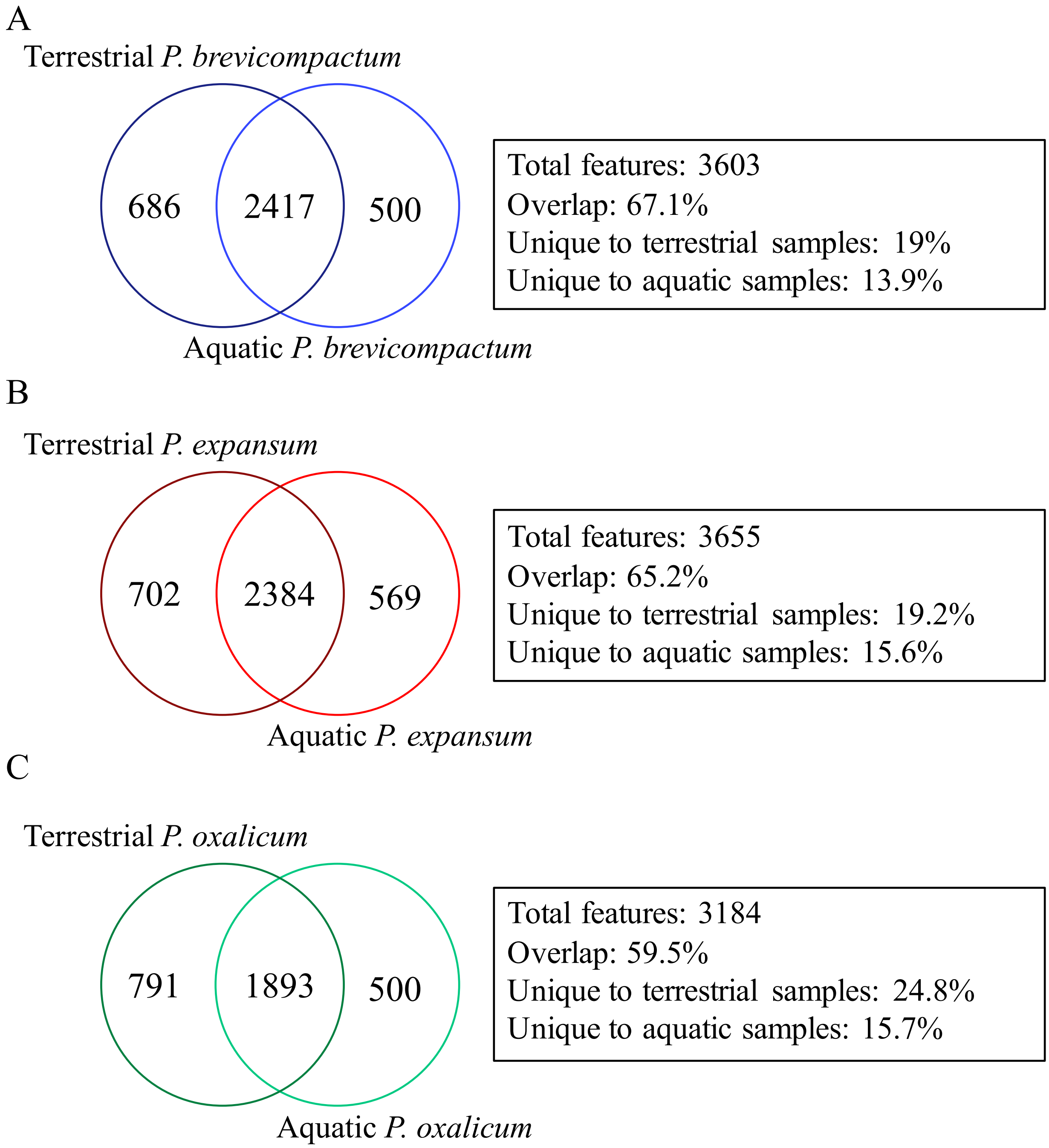
Parameter	**Value**
MS/MS fragment ions filtering	+/- 17 Da of the precursor m/z
MS/MS spectra were window filtered	6 fragment ions in the +/- 50 Da window
precursor ion mass tolerance	0.02 Da
MS/MS fragment ion tolerance	0.02 Da
cosine score	≥0.7
Minimum matched peaks	4
edges between two nodes	10 most similar nodes
maximum size of a molecular family	100
analogue search mode	enabled
MS/MS spectra	200.0
matches kept between network spectra and library spectra	
cosine score	≥0.7
Minimum matched peaks	4

Appendix B: Supporting Information for Chapter 4

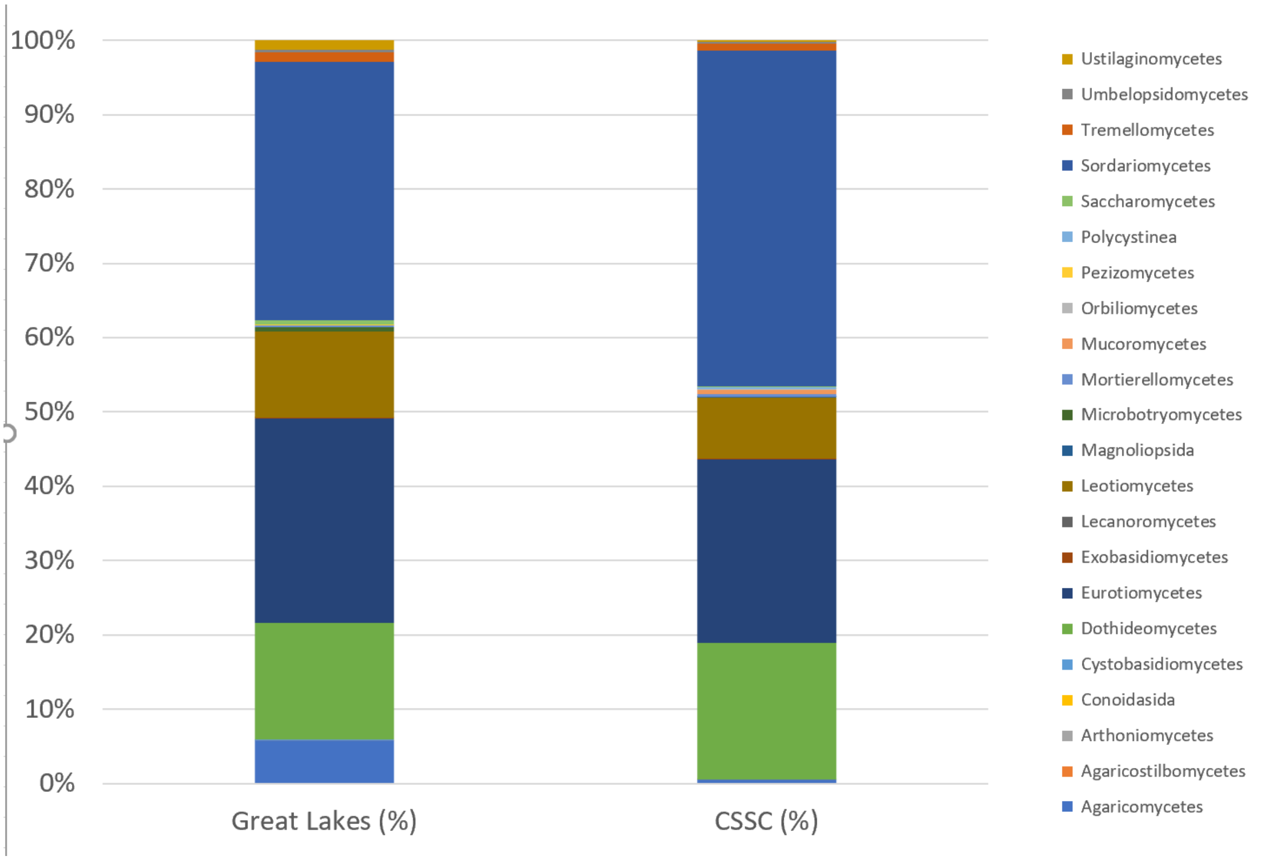
Appendix Table of Contents


99
100
75
95
■ Terrestrial P. brevicompactum
■ Aquatic P. brevicompactum
■ Terrestrial P. expansum
■ Aquatic P. expansum
■ Terrestrial P. oxalicum
■ Aquatic P. oxalicum

*Supplemental Figure 4.1 ITS phylogenetic tree. Type strains indicated by yellow stars (*P. brevicompactum *NR_121299.1,* P. expansum *NR_077154.1*, *and* P. oxalicum *NR_121232.1). Outgroup of Beauveria indicated with black stars (NR_077147.1, NR_151832.1, NR_111595.1). Tree generated with maximum likelihood analysis with 500 bootstraps. Branches are collapsed at 70%.*



Supplemental Figure 4.2 Examination of overlap of chemical features. Venn diagrams of features detected in each species collected from two source environments.



Supplemental Figure 4.3 Summary of fungal classes by environmental source: aquatic and terrestrial. Fungi that could not be identified at the class level were removed (5 genera from the aquatic environment and 5 genera from the terrestrial environment).

Supplemental Table 4.1 Class level taxonomic information for fungi in the aquatic and terrestrial environments. Numbers indicate the number of isolates identified within the respective class.

Class	Aquatic	Terrestrial
Agaricomycetes	177	21
Agaricostilbomycetes	1	0
Arthoniomycetes	1	0
Conoidasida	0	1
Cystobasidiomycetes	2	0
Dothideomycetes	480	771
Eurotiomycetes	842	1036

Class	Aquatic	Terrestrial
Exobasidiomycetes	3	5
Lecanoromycetes	1	0
Leotiomycetes	353	347
Magnoliopsida	0	2
Microbotryomycetes	19	2
Mortierellomycetes	5	15
Mucoromycetes	3	26
Orbiliomycetes	0	1
Pezizomycetes	1	3
Polycystinea	0	6
Saccharomycetes	19	5
Sordariomycetes	1063	1896
Tremellomycetes	39	40
Umbelopsidomycetes	10	9
Ustilaginomycetes	40	11

Supplemental Table 4.2 Family level taxonomic information for fungi in the aquatic and terrestrial environments. Numbers indicate the number of isolates identified within the respective family.

Family	**Aquatic**	**Terrestrial**
Amanitaceae	1	0
Amorosiaceae	2	0
Apiosporaceae	6	59
Aplosporellaceae	1	0
Arachnomycetaceae	1	1
Arthrodermataceae	2	0
Ascodesmidaceae	1	0
Aspergillaceae	728	536
Bionectriaceae	95	19
Brachybasidiaceae	2	0
Capnodiaceae	2	5
Cephalothecaceae	17	6
Ceratocystidaceae	2	0
Chaetomellaceae	4	0
Chaetomiaceae	326	38

Family	**Aquatic**	**Terrestrial**
Chaetosphaeriaceae	22	1
Cladosporiaceae	55	118
Clavicipitaceae	132	1
Coccodiscidae	6	0
Coniochaetaceae	58	78
Coniothyriaceae	19	0
Cordycipitaceae	68	21
Cucurbitariaceae	24	0
Cunninghamellaceae	4	0
Cyphellophoraceae	1	0
Cystofilobasidiaceae	4	0
Debaryomycetaceae	3	1
Dermateaceae	47	15
Diaporthaceae	1	1
Diatrypaceae	1	1
Dictyosporiaceae	2	0
Didymellaceae	63	50
Didymosphaeriaceae	48	19
Discinellaceae	2	0
Dothideaceae	1	1
Dothioraceae	1	1
Eimeriidae	1	0
Eremomycetaceae	9	0
Erysiphaceae	13	0
Filobasidiaceae	1	2
Glomerellaceae	14	6
Gymnoascaceae	1	35
Halosphaeriaceae	1	0
Helotiaceae	4	9
Herpotrichiellaceae	77	38
Hyaloscyphaceae	16	4
Hypocreaceae	144	397
Hypoxylaceae	2	51
Lachnaceae	1	1
Lamioideae	2	0
Lasiosphaeriaceae	21	27
Lentitheciaceae	5	0
Leotiaceae	6	1

Family	**Aquatic**	**Terrestrial**
Leucosporidiaceae	2	0
Lichtheimiaceae	2	0
Lipomycetaceae	1	0
Lophiostomataceae	62	1
Lophiotremataceae	3	0
Massarinaceae	9	2
Melanommataceae	1	2
Microascaceae	91	8
Microdochiaceae	12	0
Minutisphaeraceae	1	0
Morosphaeriaceae	3	0
Mortierellaceae	15	5
Mrakiaceae	18	5
Mucoraceae	19	2
Mycosphaerellaceae	5	5
Myrmecridiaceae	14	0
Myxotrichaceae	26	24
Nectriaceae	382	73
Neopyrenochaetaceae	14	0
Niessliaceae	8	0
Omphalotaceae	3	0
Onygenaceae	5	7
Ophiocordycipitaceae	43	20
Ophiostomataceae	24	0
Orbiliaceae	1	0
Parapyrenochaetaceae	1	0
Periconiaceae	3	5
Phacidiaceae	3	29
Phaeosphaeriaceae	40	7
Piskurozymaceae	1	1
Plectosphaerellaceae	110	58
Pleosporaceae	69	30
Pleurostomataceae	4	0
Pleurotaceae	1	0
Polyporaceae	2	6
Psathyrellaceae	13	81
Pseudeurotiaceae	151	121
Pyrenochaetopsidaceae	77	2

Family	**Aquatic**	**Terrestrial**
Pyronemataceae	2	0
Quambalariaceae	3	0
Rhizopodaceae	1	1
Rutstroemiaceae	2	1
Saccotheciaceae	2	10
Sarocladiaceae	81	2
Schizoparmaceae	2	2
Sordariaceae	3	3
Sporidesmiaceae	1	0
Sporocadaceae	8	9
Sporormiaceae	126	105
Stachybotryaceae	49	3
Sympoventuriaceae	43	1
Teichosporaceae	4	1
Teratosphaeriaceae	12	2
Tetraplosphaeriaceae	1	1
Thelebolaceae	2	9
Thermoascaceae	82	27
Thyridariaceae	15	2
Torulaceae	2	15
Trematosphaeriaceae	3	8
Trichocomaceae	87	156
Tricholomataceae	1	0
Trichomeriaceae	1	1
Trichomonascaceae	1	0
Trichosphaeriaceae	1	18
Trichosporonaceae	12	1
Trimorphomycetaceae	4	2
Tympanidaceae	1	2
Umbelopsidaceae	9	10
Ustilaginaceae	11	40
Valsaceae	1	13
Agaricaceae	0	1
Annulatascaceae	0	2
Arthopyreniaceae	0	3
Ascobolaceae	0	1
Biatriosporaceae	0	3
Boliniaceae	0	1

Family	Aquatic	Terrestrial
Botryosphaeriaceae	0	2
Bulleribasidiaceae	0	1
Camptobasidiaceae	0	1
Cryptococcaceae	0	18
Dipodascaceae	0	3
Dissoconiaceae	0	1
Entylomataceae	0	1
Exobasidiaceae	0	1
Gelatinodiscaceae	0	1
Gloniaceae	0	1
Kondoaceae	0	1
Latoruaceae	0	1
Leptosphaeriaceae	0	5
Lindgomycetaceae	0	3
Lulworthiaceae	0	1
Meruliaceae	0	1
Microbotryaceae	0	14
Mytilinidiaceae	0	1
Nannizziopsiaceae	0	2
Nigrogranaceae	0	2
Opegraphaceae	0	1
Phaffomycetaceae	0	5
Physalacriaceae	0	1
Pichiaceae	0	1
Podoscyphaceae	0	2
Rhytismataceae	0	1
Sclerotiniaceae	0	1
Sporidiobolaceae	0	3
Strophariaceae	0	68
Symmetrosporaceae	0	1
Togniniaceae	0	1
Trapeliaceae	0	1
Tremellaceae	0	4
Venturiaceae	0	3

Supplemental Table 4.3 Genus level taxonomic information for fungi in the aquatic and terrestrial environments. Numbers indicate the number of isolates identified in the indicated genus.

Genus	Aquatic	Terrestrial
Aaosphaeria	0	22
Abortiporus	2	0
Absidia	0	3
Achaetomium	1	0
Acremonium	30	125
Acrocalymma	0	3
Acrodontium	0	12
Acrostalagmus	3	3
Akanthomyces	0	1
Albifimbria	0	18
Alternaria	26	11
Amanita	0	1
Amauroascus	0	1
Amesia	0	3
Annulohypoxylon	1	1
Antennariella	1	2
Anthracocystis	1	0
Aotearoamyces	0	1
Aphanoascus	1	0
Apiosordaria	6	4
Apiotrichum	0	6
Aplosporella	0	1
Apophysomyces	0	2
Aposphaeria	0	1
Arachnomyces	1	1
Arachnotheca	0	1
Arcopilus	0	1
Armillaria	1	0
Arthopyrenia	3	0
Arthrinium	50	6
Arthroderma	0	1
Arthrographis	0	9
Arthropsis	0	1
Articulospora	0	1
Arxiella	0	2

Genus	**Aquatic**	**Terrestrial**
Ascobolus	1	0
Ascochyta	0	2
Ascocoryne	1	0
Ascotricha	0	1
Aspergillus	35	110
Atractium	1	0
Atrocalyx	0	3
Aureobasidium	8	2
Auxarthron	3	3
Barnettozyma	1	0
Bartalinia	0	1
Beauveria	11	17
Biatriospora	3	0
Bipolaris	2	1
Biscogniauxia	4	0
Bisporella	0	3
Blastobotrys	0	1
Boeremia	1	3
Botryotrichum	0	2
Botrytis	1	0
Bulgaria	29	0
Bulleromyces	4	0
Byssochlamys	0	5
Cadophora	26	2
Calonectria	0	1
Camarops	1	0
Candida	9	0
Capnodium	1	0
Caryospora	0	1
Cenococcum	1	0
Cephaliophora	0	1
Cephalotheca	1	1
Cephalotrichum	0	42
Cercophora	5	4
Cercospora	2	1
Chaetomella	0	4
Chaetomium	15	83
Chaetopsina	0	1

Genus	**Aquatic**	**Terrestrial**
Chaetosphaeria	1	0
Cheilymenia	0	1
Chloridium	0	21
Chrysosporium	9	44
Circinella	0	2
Cladophialophora	0	2
Cladorrhinum	0	8
Cladosporium	113	55
Clathrosphaerina	1	0
Clonostachys	17	85
Coccidioides	1	0
Coleophoma	4	2
Colletotrichum	6	14
Coniella	1	2
Coniochaeta	78	58
Coniolariella	1	0
Coniothyrium	0	19
Coniozyma	1	0
Coprinellus	72	13
Coprinus	1	0
Corallomycetella	0	1
Cordana	1	1
Cordyceps	0	40
Cosmospora	6	19
Creosphaeria	2	0
Crocicreas	1	0
Cryptococcus	18	0
Cryptostroma	5	0
Curvularia	1	42
Cutaneotrichosporon	0	2
Cyberlindnera	4	0
Cycasicola	0	1
Cylindrocarpon	1	23
Cylindrocladiella	0	3
Cyphellophora	0	1
Cystodendron	1	0
Cystofilobasidium	0	4
Dactylaria	0	1

Genus	Aquatic	Terrestrial
Dactylonectria	0	3
Daldinia	10	0
Darkera	0	1
Davidiella	4	0
Debaryomyces	1	2
Dendryphion	9	0
Devriesia	2	0
Dialonectria	0	7
Diaporthe	1	1
Diatrype	1	1
Dichotomopilus	0	7
Dictyochaeta	0	1
Dictyosporium	0	1
Didymella	0	23
Didymocyrtis	0	6
Didymosphaeria	0	42
Dimorphospora	3	0
Dipodascopsis	0	1
Discosia	0	2
Dokmaia	0	12
Doratomyces	0	2
Dothichiza	0	1
Dothidea	0	1
Dothiorella	1	0
Edenia	0	1
Emericella	0	1
Emericellopsis	8	3
Entyloma	1	0
Epicoccum	7	25
Eucasphaeria	1	3
Exobasidium	1	0
Exophiala	7	56
Exserohilum	0	1
Fimetariella	8	3
Flammula	2	0
Fusarium	37	226
Fusicolla	1	28
Galactomyces	3	0

Genus	Aquatic	Terrestrial
Galerina	2	0
Ganoderma	5	1
Gemmina	0	1
Geomyces	19	4
Gibellulopsis	35	42
Glaciozyma	1	0
Gliocladium	0	6
Gliomastix	2	9
Gloeopycnis	0	2
Glutinomyces	0	5
Goffeauzyma	2	1
Golovinomyces	0	13
Gongronella	0	1
Gonytrichum	0	1
Graphium	2	1
Gymnoascus	34	0
Halenospora	1	6
Halosarpheia	0	1
Hamigera	1	0
Helicodendron	1	0
Heterosphaeria	1	0
Hirsutella	0	2
Holtermanniella	1	0
Holwaya	2	0
Hongkongmyces	2	0
Hormiactis	0	2
Hormodochis	0	2
Hormonema	1	0
Humicola	4	106
Hyalodendriella	1	3
Hyalopeziza	0	2
Hyaloscypha	0	1
Hymenoscyphus	2	0
Hypholoma	27	0
Hypomyces	1	0
Hypoxylon	38	1
Idriella	0	2
Ijuhya	0	1

Genus	**Aquatic**	**Terrestrial**
Ilyonectria	1	11
Incrucipulum	1	0
Infundichalara	0	2
Isaria	2	6
Jalapriya	0	1
Keissleriella	0	2
Keithomyces	0	1
Kiflimonium	0	2
Knufia	1	1
Kondoa	1	0
Kretzschmaria	2	0
Lachnum	1	0
Lambertella	1	1
Lambiella	1	0
Lasiodiplodia	1	0
Lasiosphaeria	7	0
Lasiosphaeris	0	3
Lecanicillium	0	1
Lectera	0	9
Leptodiscella	2	2
Leptodontidium	2	22
Leptosphaeria	5	0
Leptosphaerulina	7	6
Leptospora	4	0
Leptoxyphium	3	0
Leuconeurospora	0	3
Leucosporidium	0	2
Lomentospora	0	3
Lophiostoma	1	61
Mammaria	0	4
Mariannaea	1	9
Massarina	2	9
Matsushimamyces	1	0
Meira	0	2
Melanomma	2	1
Melanopsichium	1	0
Metapochonia	0	1
Metarhizium	1	83

Genus	**Aquatic**	**Terrestrial**
Meyerozyma	0	1
Microascus	0	5
Microcera	0	1
Microdochium	0	9
Microsphaeropsis	0	1
Microsporum	0	1
Minutisphaera	0	1
Moesziomyces	0	10
Monochaetia	0	1
Monocillium	0	6
Monodictys	1	1
Mortierella	5	15
Mrakia	2	0
Mucor	2	17
Murilentithecium	0	1
Myceliophthora	0	1
Myrmecridium	0	14
Myrothecium	1	15
Mytilinidion	1	0
Nannizziopsis	2	0
Nectria	24	26
Nemania	15	1
Neocucurbitaria	0	5
Neofabraea	1	0
Neopyrenochaeta	0	14
Neosartorya	0	2
Neoscolecobasidium	0	1
Neosetophoma	0	1
Neurospora	2	3
Niesslia	0	2
Nigrograna	2	0
Nigrospora	18	1
Nodulisporium	2	0
Nothophoma	0	1
Ochroconis	1	42
Ogataea	1	0
Oidiodendron	22	26
Ombrophila	1	0

Genus	Aquatic	Terrestrial
Opegrapha	1	0
Ophiostoma	0	1
Ovadendron	0	3
Paecilomyces	27	77
Papulaspora	0	1
Paraconiothyrium	5	0
Paramicrothyrium	2	0
Paraphaeosphaeria	12	5
Paraphoma	2	22
Parasarocladium	0	49
Parascedosporium	0	5
Parastagonospora	1	0
Parathyridaria	0	9
Parvothecium	0	1
Patinella	9	0
Penicillifer	0	2
Penicillium	474	615
Perenniporia	1	1
Periconia	5	3
Pestalotiopsis	9	4
Pezicula	1	0
Pezizella	0	3
Phacidiopycnis	1	0
Phaeosphaeria	3	11
Phaeosphaeriopsis	1	1
Phalangispora	0	2
Phialemoniopsis	0	1
Phialemonium	5	16
Phialocephala	6	1
Phialophora	31	19
Phlebia	1	0
Pholiota	33	0
Phoma	26	3
Phomopsis	6	0
Phragmocamarosporium	0	2
Piskurozyma	0	1
Plectosphaerella	20	56
Pleiochaeta	0	1

Genus	**Aquatic**	**Terrestrial**
Pleospora	0	13
Pleurostoma	0	4
Pleurotus	0	1
Pochonia	0	47
Podocarpomyces	0	2
Podospora	11	5
Pogostemon	0	2
Polyphilus	0	4
Preussia	93	38
Protoventuria	1	0
Psathyrella	9	0
Pseudallescheria	4	1
Pseudeurotium	111	1
Pseudocercosporella	1	0
Pseudogymnoascus	0	147
Pseudombrophila	0	1
Pseudopithomyces	0	1
Pseudozyma	10	1
Psilocybe	4	0
Purpureocillium	1	37
Pyrenochaeta	0	19
Pyrenochaetopsis	0	77
Quambalaria	0	3
Quixadomyces	0	1
Ramichloridium	1	0
Ramularia	0	3
Rasamsonia	1	0
Rhizopus	1	1
Rhodocollybia	0	3
Rhodotorula	3	0
Rollandina	0	1
Roseodiscus	0	1
Roussoella	1	4
Rutstroemia	0	1
Sagenomella	2	1
Saitozyma	2	4
Sarocladium	2	32
Scedosporium	1	2

Genus	Aquatic	Terrestrial
Schizothecium	1	3
Sclerostagonospora	0	4
Scopulariopsis	1	11
Scytalidium	7	5
Selenodriella	0	1
Septoria	1	1
Simplicillium	0	3
Sirastachys	0	1
Solicoccozyma	1	0
Sphacelotheca	7	0
Sphaerostilbella	0	1
Sporidesmium	0	1
Sporormia	4	81
Sporothrix	0	23
Stachybotrys	0	11
Stagonosporopsis	1	0
Staphylotrichum	0	52
Stilbella	0	1
Striatibotrys	0	2
Striaticonidium	0	1
Submersisphaeria	1	0
Subramaniula	0	1
Symmetrospora	1	0
Talaromyces	148	86
Tausonia	3	18
Teichospora	1	4
Testudomyces	1	0
Tetrachaetum	0	1
Tetracladium	30	29
Tetraplosphaeria	0	1
Thelebolus	9	2
Thermomyces	1	0
Thielavia	0	4
Thielaviopsis	0	2
Thyridaria	1	0
Thyridariella	0	1
Thysanophora	1	0
Tilletiopsis	1	0

Genus	Aquatic	Terrestrial
Tolypocladium	19	4
Torula	6	2
Tranzscheliella	7	0
Trematosphaeria	8	3
Tricellula	0	1
Trichocladium	1	23
Trichoderma	396	141
Tricholoma	0	1
Trichopeziza	0	1
Trichosporiella	0	45
Trichosporon	1	4
Trichurus	0	2
Umbelopsis	10	9
Ustilago	1	0
Valsa	7	1
Venturia	2	0
Verticillium	6	10
Vishniacozyma	1	0
Volutella	0	13
Wardomyces	2	17
Westerdykella	8	7
Wojnowiciella	0	1
Xanthothecium	1	0
Xepicula	2	0
Xylaria	9	3
Xylogone	0	6
Xylomelasma	4	2
Yunnania	0	1
Zalerion	1	0
Zopfiella	3	0

Supplemental Table 4.4 Isolates used in the analysis. Values in parentheses indicate the number of isolates at each level.

Species	**Source**	**State or Sediment ID**	**full id**	**cryo**	**link**	**Genbank accession no.**
P. brevicompactum (25)	CSSC (12)	IL (6)	IL02963 TV8-3 (1)	330-C6	https://shareok.org/handle/11244/44943	MZ362515
			IL05106 TV8+M-1 (1)	280-D11	https://shareok.org/handle/11244/44568	MZ362516

Species	Source	State or Sediment ID	full id	cryo	link	Genbank accession no.
			IL08989 BIA-1 (1)	426-B10	https://shareok.org/handle/11244/54424	MZ362518
			IL12323 RBM+LICL-7 (1)	519-A7	https://shareok.org/handle/11244/316382	MZ362525
			IL3098 CEA-1 (1)	344-D2	https://shareok.org/handle/11244/29429	MZ362527
			IL7520 TV8-3 (1)	209-C7	https://shareok.org/handle/11244/28218	MZ362528
		MI (5)	MI05057 RBM-3 (1)	380-F3	https://shareok.org/handle/11244/44567	MZ362572
			MI06134 TV8-3 (1)	250-D12	https://shareok.org/handle/11244/44971	MZ362573
			MI1347 TV8-6 (1)	149-D3	https://shareok.org/handle/11244/30002	MZ362578
			MI4821 RBM-2 (1)	168-G7	https://shareok.org/handle/11244/29699	MZ362580
			MI7656 RBM-1 (1)	194-C11	https://shareok.org/handle/11244/28273	MZ362581
		WI (1)	WI06424 TV8+M-6 (1)	247-F7	https://shareok.org/handle/11244/44500	MZ362588
	GL (13)	LMS 100	LMS 100-20 (1)	GL38-H7		MZ362529
		LMS 57	LMS 57-8 (1)	GL40-A5		MZ362546
		LMS 59	LMS 59-6 (1)	GL39-B6		MZ362548
		LMS 65	LMS 65-1 (1)	GL37-C10		MZ362551
		LMS 68	LMS 68-7 (1)	GL37-E4		MZ362553
		LMS 76(2)	LMS 76-16 (1)	GL36-B8		MZ362555
			LMS 76-18(1)	GL36-B9		MZ362556
		LMS 77	LMS 77-12 (1)	GL38-A5		MZ362558
		LMS 83	LMS 83-4 (1)	GL36-D1		MZ362559
		LMS 86	LMS 86-9 (1)	GL44-F11		MZ362561
		LMS 94	LMS 94-9 (1)	GL36-G2		MZ362563
		LMS 95 (2)	LMS 95-12 (1)	GL36-G9		MZ362564
			LMS 95-4(1)	GL41-B2		MZ362565
P. expansum (26)	CSSC (13)	IL (3)	IL00446 RBM-31 (1)	203-E8	https://shareok.org/handle/11244/28751	MZ362513
			IL12018 RBM-1 (1)	614-D5	https://shareok.org/handle/11244/320904	MZ362521
			IL3098 BFA-4 (1)	293-D6	https://shareok.org/handle/11244/29429	MZ362526
		MI (4)	MI03494 TV8-3 (1)	342-F3	https://shareok.org/handle/11244/29116	MZ362571
			MI06334 TV8-3 (1)	331-D7	https://shareok.org/handle/11244/44474	MZ362574
			MI08762 RBM-1 (1)	383-D9	https://shareok.org/handle/11244/52497	MZ362577
			MI17330 RBM-1 (1)	606-B6	https://shareok.org/handle/11244/320842	MZ362579
		OH (3)	OH00563 TV8-3 (1)	242-B4	https://shareok.org/handle/11244/41911	MZ362582
			OH05830 RBM-5 (1)	329-F11	https://shareok.org/handle/11244/44306	MZ362583
			OH06145 TV8+M-2 (1)	247-B10	https://shareok.org/handle/11244/44419	MZ362584
		WI (3)	WI00343 RBM-4 (1)	180-G9	https://shareok.org/handle/11244/28735	MZ362587
			WI07453 BFA-3 (1)	283-G4	https://shareok.org/handle/11244/47009	MZ362589
			WI08376 ZMA-1 (1)	328-H4	https://shareok.org/handle/11244/51806	MZ362590
	GL (13)	LMS 123	LMS 123-5 (1)	GL76-E1		MZ362531

Species	Source	State or Sediment ID	full id	cryo	link	Genbank accession no.
		LMS 143	LMS 143-1 (1)	GL82-G1		
		LMS 153	LMS 153-4 (1)	GL85-D2		MZ362535
		LMS 182	LMS 182-25 (1)	GL101-F4		MZ362536
		LMS 187	LMS 187-18 (1)	GL102-D8		MZ362537
		LMS 205	LMS 205-10 (1)	GL105-D10		MZ362539
		LMS 49	LMS 49-2 (1)	GL23-F7		MZ362544
		LMS 58	LMS 58-1 (1)	GL39-A12		MZ362547
		LMS 68	LMS 68-6 (1)	GL39-F1		MZ362552
		LMS 71	LMS 71-6 (1)	GL37-G1		MZ362554
		LMS 77	LMS 77-10 (1)	GL40-D9		MZ362557
		LMS 85	LMS 85-6 (1)	GL38-D6		MZ362560
		LMSO 129	LMSO 129-8 (1)	GL79-D7		MZ362567
P. oxalicum (27)	CSSC (14)	IL (7)	IL02902 TV8-5 (1)	343-C8	https://shareok.org/handle/11244/42141	MZ362514
			IL08349 TV8-3 (1)	316-B9	https://shareok.org/handle/11244/51797	MZ362517
			IL11455 PDAT-2 (1)	497-G10	https://shareok.org/handle/11244/301507	MZ362519
			IL11999 RBM-2 (1)	614-B8	https://shareok.org/handle/11244/320884	MZ362520
			IL12295 RBM-5 (1)	518-A5	https://shareok.org/handle/11244/316354	MZ362522
			IL12312 RBM-2 (1)	518-E8	https://shareok.org/handle/11244/316370	MZ362523
			IL12316 TV8-1 (1)	517-H3	https://shareok.org/handle/11244/316375	MZ362524
		MI (5)	MI00176 LBC-4 (1)	399-D8	https://shareok.org/handle/11244/28706	MZ362569
			MI00803 RBM-1 (1)	157-A8	https://shareok.org/handle/11244/28814	MZ362570
			MI08759 TV8-1 (1)	378-F11	https://shareok.org/handle/11244/52494	MZ362575
			MI08761 TV8-1 (1)	385-F9	https://shareok.org/handle/11244/52496	MZ362576
			MI1961 TV8-2 (1)	136-D1	https://shareok.org/handle/11244/29201	
		OH (2)	OH07215 RBM-5 (1)	381-A4	https://shareok.org/handle/11244/46974	MZ362585
			OH9508 RBM-2 (1)	144-A7	https://shareok.org/handle/11244/28586	MZ362586
	GL (14)	LMS 121	LMS 121-10 (1)	GL76-C9		MZ362530
		LMS 13	LMS 13-19 (1)	GL27-A11		MZ362532
		LMS 149	LMS 149-9 (1)	GL82-H5		MZ362533
		LMS 15	LMS 15-23 (1)	GL21-C8		MZ362534
		LMS 197	LMS 197-11 (1)	GL104-A6		
		LMS 2	LMS 2-33 (1)	GL21-A11		MZ362538
		LMS 29 (2)	LMS 29-14 (1)	GL22-C3		MZ362540
			LMS 29-2 (1)	GL26-H7		MZ362541
		LMS 30	LMS 30-9 (1)	GL21-G6		MZ362542
		LMS 35	LMS 35-2 (1)	GL22-D4		MZ362543
		LMS 54	LMS 54-8 (1)	GL23-G10		MZ362545

Species	Source	State or Sediment ID	full id	cryo	link	Genbank accession no.
		LMS 6	LMS 6-4 (1)	GL21-B3		MZ362549
		LMS 94	LMS 94-7 (1)	GL36-F12		MZ362562

Supplemental Table 4.5 Data acquisition parameters for LC-MS/MS.

Parameter	Value
Data acquisition mode	positive
Scan range	180-2000 m/z
Top N	5
sheath gas	40 L/min
auxiliary gas	5 L/min
sweep gas	0 L/min
spray voltage	4.5 kV
S-lens RF	95 V
capillary temperature	270 °C
Normalized collision energy increments	35%

Supplemental Table 4.6 MZmine data processing parameters.

Process	Parameter	Value
Mass Detection	MS1 Noise Level	3.0E3
	MS2 Noise Level	1.5E2
	Mass Detector	Centroid
Chromatogram Builder	Min group size in # of scans	2
	Group intensity threshold	4.0E4

	Min highest intensity	4.0E4
	m/z tolerance (m/z)	1.1
Chromatogram Deconvolution: LOCAL MINIMA algorithm	Chromatographic threshold	30%
	Search minimum in RT range (min)	0.05
	Minimum relative height	20%
	Minimum absolute height	1.2E3
	Min ratio of peak top/edge	1.19
	Peak duration range (min)	0.01-1.00
	m/z Range for MS2 Scan Pairing (Da)	0.01
	RT Range for MS2 Scan Pairing (min)	0.1
Isotopic Peak Grouper	Retention Time Tolerance (min)	0.1
	m/z tolerance (ppm)	15
	Monotonic Shape	Yes
	Maximum Charge	3
	Representative isotope	Lowest *m/z*
Join aligner	*m/z* tolerance (ppm)	15
	m/z to RT weight	1-1
	Retention Time Tolerance (min)	0.25
Row filtering	Keep only peaks with MS2 scan	Enabled
	Minimum peaks in a row	2 (for duplicates)

Supplemental Table 4.7 GNPS parameters.

Parameter	Value
MS/MS fragment ions filtering	+/- 17 Da of the precursor m/z
MS/MS spectra were window filtered	6 fragment ions in the +/- 50 Da window

precursor ion mass tolerance	0.02 Da
MS/MS fragment ion tolerance	0.02 Da
cosine score	≥0.7
Minimum matched peaks	4
edges between two nodes	10 most similar nodes
maximum size of a molecular family	100
analogue search mode	enabled
MS/MS spectra	200.0
cosine score	≥0.7
Minimum matched peaks	4