Title

Minimizing Taxonomic and Natural Product Redundancy in Microbial Libraries Using MALDI-TOF MS and the Bioinformatics Pipeline IDBac

Journal
Journal of Natural Products, 82(8)

ISSN
0163-3864

Authors
Costa, Maria S
Clark, Chase M
Ómarsdóttir, Sesselja
et al.

Publication Date
2019-08-23

DOI
10.1021/acs.jnatprod.9b00168

Peer reviewed

Published in final edited form as:
J Nat Prod. 2019 August 23; 82(8): 2167–2173. doi:10.1021/acs.jnatprod.9b00168.

Minimizing Taxonomic and Natural Product Redundancy in Microbial Libraries using MALDI-TOF MS and the bioinformatics pipeline IDBac.

Maria S. Costa[#†,‡], **Chase M. Clark**[#‡], **Sesselja Ómarsdóttir**[†], **Laura M. Sanchez**[‡], **Brian T. Murphy**[‡]

[†]Faculty of Pharmaceutical Sciences, University of Iceland, Hagi, Hofsvallagata 53, IS-107 Reykjavík, Iceland [‡]Department of Medicinal Chemistry and Pharmacognosy, College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street (MC 781), Room 539, Chicago, IL, United States

[#] These authors contributed equally to this work.

Abstract

Libraries of microorganisms have been a cornerstone of drug discovery efforts since the mid-1950s, but strain duplication in some libraries has resulted in unwanted natural product redundancy. In the current study, we implemented a workflow that minimizes both the natural product overlap and the total number of bacterial isolates in a library. Using a collection expedition to Iceland as an example, we purified every distinct bacterial colony (1,616 total) off isolation plates derived from 86 environmental samples. We employed our mass spectrometry (MS) based IDBac workflow on these isolates to form groups of taxa based on protein MS fingerprints (3-15 kDa), and further distinguished taxa subgroups based on their degree of overlap within corresponding natural product spectra (0.2-2 kDa). This informed the decision to create a library of 301 isolates spanning 54 genera. This process required only 25 hours of data acquisition and 2 hours of analysis. In a separate experiment, we reduced the size of an existing library based on the degree of metabolic overlap observed in natural product MS spectra of bacterial colonies (from 833 to 233 isolates, a 72.0% reduction). Overall, our pipeline allows for the reduction of library size and costs associated with library generation, and minimizes natural product redundancy entering into downstream biological screening efforts.

Libraries of microorganisms have been a cornerstone of drug discovery efforts since the middle of the 1950s. Natural products (including their semi-synthetic derivatives) isolated from these libraries have afforded us more than 170 cancer drugs and greater than half of marketed anti-infective drugs.[1–4] In particular, discoveries of pyocyanase, penicillin, and tyrothricin (gramicidin and tyrocidine) ushered in an unprecedented global effort to mine the environment for new microbial natural products.[5–8] This effort was highly successful, and was driven by sampling expeditions whose aim was to amass libraries of cultivable microorganisms from the environment. Many method innovations in pharma involving automation, miniaturization, cultivation, and biological screening, were responsible for these successes.[9,10] Despite this, by the end of the “Golden Age” of antibiotic discovery in roughly the 1970s, the re-isolation of known natural products became (and remains) a major problem.[10,11] One cause of this was the high degree of strain duplication in these libraries, which resulted in unwanted natural product redundancy and lower coverage of chemical space.[12–15] This redundancy is the result of a few circumstances: 1) strain libraries were often created on the basis of visual inspection of colony morphology, which does not fully account for natural product production;[13,14,16] 2) bacterial isolation efforts focused heavily on spore forming bacteria such as Actinobacteria and Firmicutes, which have a history of producing biologically active natural products.[2] Diminishing returns on this investment and a shift toward combinatorial chemistry approaches, among several other factors, led most of the pharmaceutical industry to divest from microbial-based natural product drug discovery.[2,13,17–20]

In order to overcome the high rate of compound re-discovery, it is necessary to reduce the degree of overlapping inter- and intra-species chemical space in a microbial library. We previously developed a freely available high-throughput matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) data acquisition and bioinformatics pipeline – IDBac – that allows researchers to rapidly collect putative phylogenetic and natural product information from single colonies of unknown environmental bacteria.[21] From each colony, we generate two data sets: 1) protein fingerprints (3,000-15,000 Da, primarily ribosomal)[22,23] that are used to assign each isolate into putative genus/species groupings, and 2) natural product fingerprints (200-2,000 Da) of each colony to compare inter- and intra-species differences/similarities in natural product production. We recently published a video protocol detailing this procedure.[24] A few reviews and more recent studies highlight attempts to prioritize bacterial strains in a library with regard to either taxonomic identity or natural product potential; this list is not comprehensive.[12,13,15,16,25–30]

Herein, we provide a template for the generation of a diverse microbial library using MALDI-TOF MS and IDBac. We broadly define natural product diversity as natural products that exhibit different chemical structures. To demonstrate the effectiveness of our method, we applied IDBac to samples collected during an expedition to Iceland in 2017. From 1,616 total bacterial isolates, we created a library of 301 isolates that span 54 genera and exhibited minimal natural product overlap, based on metabolic profiling. This pipeline allowed the creation of a diverse library while requiring limited resources (toothpick, bacterial colonies, access to a MALDI-TOF MS), and can be heavily tailored toward creating custom collections of individual taxa, targeting the isolation of understudied taxa, and studying patterns of pseudo-phylogeny and natural product production within large groups of microbial isolates. Use of this method provides an alternative to DNA sequencing, liquid fermentation, extraction, or chromatographic analyses that would be used to generate similar taxonomic and natural product information.

RESULTS AND DISCUSSION

Collection and processing of environmental samples.

In total, 86 samples from 30 locations, consisting of sediment (46), marine invertebrates (25), algae (12), and water column (3), were collected (see Table S2 and Figure S1, Supporting Information). Each environmental sample was processed using common techniques that select for spore-forming and non-spore-forming bacteria, and plated onto five different agar-based nutrient media.[31–33] After incubation of up to 60 days, every distinct bacterial colony growing on nutrient media was isolated (those appearing to exist as a single phenotype; Figure S2, Supporting Information). In general, a relatively low number of bacterial colonies were observed partially due to the following: biofilms were excluded from analysis, media were not optimized for the geothermal nature of some collection sites, and high nutrient media typically afforded higher bacterial density, and as a result fewer distinct colonies. When all 1,616 isolates were obtained, each was re-plated in individual wells of a 48-well agar plate onto high nutrient A1 media to facilitate comparison on the same time scale (marine and freshwater-derived datasets were analyzed separately). Single colonies were then transferred to a MALDI target plate, data were acquired on an Autoflex Speed LRF MALDI-TOF mass spectrometer, and analyzed using IDBac. The final step of re-plating, incubating, and acquiring data on bacterial isolates was performed in a three-week span in order to allow for simultaneous growth and triplicate data acquisition for each colony (4,848 total MALDI spots). With advanced instrumentation and automation (bacterial isolation and sample preparation), this entire process could conceivably be improved to a few days to one week.

Targeting of all distinct colonies represents a major deviation from some past library generation practices that employed visual inspection of morphology and other selective colony picking techniques to build libraries. The ease of data collection and throughput of this pipeline (sample preparation, data collection and visualization of 384 samples in under 4 hours) allowed us to incorporate high numbers of colonies that otherwise would not have been included using our previous protocols, which followed common practices.

Grouping and analysis of bacterial isolates into putative taxa based on MALDI-TOF MS spectra (3,000-15,000 Da).

MALDI-TOF MS is a well-established technique used to identify and group microbial isolates based on analysis of MS spectra in the 3,000-15,000 Da range.[34] Since resulting MS fingerprints are largely due to the high copy number of ionized ribosomal proteins,[22,23] pseudo-phylogenetic relationships can be inferred, often achieving species and sub-species level resolution. Application of this technique has been thoroughly reviewed elsewhere.[22,35,36] In relevance to the current study, Dieckmann *et al.* were among the first to use MALDI-TOF MS to dereplicate bacterial isolates from marine sponges, though only protein spectra were taken into consideration, limiting isolate dereplication to taxonomic groupings as opposed to natural product production.[37]

The 1,616 bacterial isolates from Iceland were first separated into two sets: those isolated from marine and those from freshwater locations. Each set was grown on media supplemented with and without artificial ocean salts, respectively. In our experience, a colony often exhibits different MS profiles in the presence of synthetic ocean salt in nutrient agar compared to its salt-free counterpart. Colonies were transferred to steel target plates and MALDI-TOF MS data were acquired and subsequently processed in IDBac, as previously described.[21,24] IDBac offers several settings when creating hierarchical clustering plots based on protein MS spectra. We suggest pairing cosine-distance and average-linkage-clustering settings for library creation efforts. Average-linkage conserves the 0 to 1 scale of the cosine similarity score, allowing the resulting dendrogram to be more easily interpreted (Figure 2). To confirm that the resulting MALDI MS protein groupings were representative of phylogenetic similarity, we subsampled 361 isolates across both dendrograms (27.5% of freshwater and 17.8% of marine isolates) and sequenced their 16S rRNA genes (Table S2, Supporting Information). We observed that isolate groupings formed as a result of 16S rRNA phylogenetic analysis closely resembled those formed by MALDI-TOF/IDBac analysis (Figure S3, Supporting Information). Our results corroborate many previous efforts that document the use of MALDI-TOF MS to facilitate the grouping and/or taxonomic identification of environmental isolates.[35,38–40]

It is important to note that while isolates often group at the species-level, the dendrograms created in IDBac are not phylogenetic trees. We noticed in our dendrograms that MALDI-TOF MS analysis of bacterial proteins generally results in groupings that represent the genus- taxonomic level and below but lose fidelity of clustering at higher-taxonomic levels. As recently shown by Seuylemezian *et al.*, while highly similar isolates group together based on MALDI protein spectra similarity, the overall topology of clusters was slightly different than those found in a phylogenetic tree constructed using the 16S rRNA gene.[41] In our data, while genera such as *Shewanella*, *Bradyrhizobium*, *Thiomonas*, and *Salinococcus* (among others) form single groupings in the dendrogram, some larger, more diverse genera such as *Bacillus*, *Paenibacillus*, *Streptomyces*, and *Micromonospora* form multiple groupings distributed throughout each dendrogram. To confirm this wasn't due to outside variables (date spectra were acquired, batch of growth media used to purify colonies, etc.), we seeded the freshwater dataset with protein spectra from previously characterized strains. Each of three *Micromonospora* strains from our in-house collection (*Micromonospora aurantiaca* D053 and *M. humi* D077 from marine sediment collected off the coast of Massachusetts, and *M. aurantiaca* B017 from freshwater sediment in Lake Michigan) matched with a corresponding *Micromonospora* grouping in both freshwater and marine datasets (Figure S4–S5, Supporting Information). Similar coherence was observed when seeding our freshwater and marine datasets with characterized *Bacillus* strains (*Bacillus subtilis* 3610; *B. subtilis* NRRL B14596; Figure S4–S5, Supporting Information). The spread of these genera across the dendrogram reflects, in some cases, the higher degree of resolution achieved by MALDI-TOF MS, since signals represent ionized proteins from more than just ribosomal origin. It also highlights the limitations of methods used to define bacterial taxonomy, since two isolates may share high 16S rRNA sequence homology but low genomic DNA homology, as explained previously.[41] In the presence of a large reference database of protein spectra from characterized environmental microorganisms, the rapid taxonomic identification of unknown isolates is possible. A few studies describe the creation of custom MALDI-TOF protein MS databases to assist in the identification of specific bacterial taxa, however the absence of an extensive database is a major roadblock to the use of MALDI-TOF MS as a means to identify bacteria isolated from the environment.[41–43] Currently, IDBac allows the user to build custom databases that can be readily searched, modified, and shared.

In the context of creating a diverse bacterial library for natural product discovery, it is not essential that hierarchical clustering perfectly mirrors phylogenetic groupings. Clustering isolates based on protein MS fingerprints primarily serves as a data reduction strategy for the following step, comparison of natural product production within closely-related groups of isolates. Occasionally an isolate "mismatches" to a cluster that contains other genera of bacteria (e.g. a *Bacillus* isolate is paired with a group of *Streptomyces* isolates); this is generally due to the absence of close relatives in the dataset and a number of shared peaks emanating from proteins that are not of ribosomal origin. Regardless, in our experience it will stand as an obvious outlier when comparing natural product overlap between isolates within that cluster.

In the case of the Iceland protein MS dataset, a threshold of 0.75 was used to create 72 distinct freshwater and 101 distinct marine isolate groupings. Groups contained anywhere between 1 and 229 isolates. Each group was then manually selected within IDBac to generate a metabolite association network. These networks were used as the basis for selecting isolates with minimal overlapping natural product production for addition to the library. Strategies to analyze natural product overlap within these groups are discussed in the next section.

Use of Metabolite Association Networks (MANs) to reduce inter- and intra-species natural product redundancy in a bacterial library.

It has been demonstrated that natural product production, broadly speaking, follows taxa-specific trends at the genus level.[44–46] Conversely, genera may also exhibit extensive natural product biosynthetic gene diversity at the genus level.[47,48] Further, it is well known that on a species-level, isolates with highly similar phenotypes can display a differential capacity to produce natural products.[45,47,49] We previously demonstrated a similar phenomenon in a group of *Micromonospora* strains isolated from the same 1 cm^3 of sediment.[21] In the context of IDBac, grouping bacteria by taxa is an effective means by which the user can detect these more subtle genus/species-level differences in natural product production, and account for molecules that deviate from taxa-specific patterns.

As shown in Figures 3 and 4B, larger nodes represent individual bacterial isolates, while smaller nodes represent *m/z* values detected in their corresponding MS spectra (200-2,000 Da). Matrix and media ions were automatically subtracted from the resulting analysis. We previously demonstrated that the majority of *m/z* values in a MAN between two *Bacillus* strains (33 of 42; 78.6%) represented natural products, giving confidence in the ability of IDBac to filter out matrix signals.[21] It is worth nothing that the mass resolving power of MALDI-TOF MS does not allow for level-1 or −2 identifications of natural products;[50] further validation is necessary to elucidate a structure.

Since it is difficult to observe meaningful intra-species natural product production relationships in a MAN of all 1,616 isolates, it is recommended to view MANs in smaller subsets based on hierarchical clustering results (*e.g.*, groups that range from 10-50 isolates). Based on our data, groups between 1 and 50 isolates were optimal, though this range may vary depending on several factors within a dataset and user needs. It may be possible to mitigate MANs that appear as a single cluster as a result of viewing large numbers of isolates, with high-resolution, high-mass accuracy MALDI-TOF MS instruments.

The purpose of using MANs is to quickly assess the natural product capacity of groups of bacterial colonies that share similar phylogenetic origin and make an informed judgement of which to add to a microbial library. They allow researchers to address fundamental questions that have immense downstream effects on microbial drug discovery efforts including: how different is natural product production between colonies with similar morphology from the same/different collection sites? Figure 3 depicts four relationships that would be expected to occur in the environment from any pair of isolates that share >99% 16S rRNA gene sequence identity. Neither visual inspection nor taxonomic classification strategies would inform the researcher about which scenario is occurring. IDBac provides this information and allows a researcher to make an educated choice of which bacterial isolates they will add to a library.

We built our microbial library based on the analysis of MANs generated from 72 freshwater and 101 marine isolate protein MS groupings. Each dendrogram grouping and corresponding MAN is provided in the Supporting Information (Figure S8–S9). It is up to the individual researcher and available resources to determine roughly how many isolates they wish to add to a library, and what constitutes significant overlap. Figure 4 depicts a grouping of 16 isolates (estimated to be *Bacillus* spp. from 16S rRNA gene sequencing of 116H-8 and 116F-10). The nine isolates clustered in pink exhibited a high degree of natural product overlap, therefore two isolates (116G-4 and 116H-1) were selected for addition to the library. Alternatively, isolates 116F-10 and 116H-8 showed a high degree of unique natural product signatures (non-overlapping with other isolates in the group), so they were also included in the library. Isolates 116G-1 and 122A-5, as well as 117D-5 and 117D-2 are a more ambiguous case. Although some natural product overlap exists, all isolates in these groups contain a set of *m/z* values that are unique, which may represent a molecule(s) that exists exclusively in one bacterium over another. The researcher is now empowered to make an informed decision of whether to include one, both, or neither of the pair.

After approximately 25 hours of data acquisition (data analysis time is researcher dependent), 301 isolates were selected for addition to the library out of a total 1,616. The 301 isolates span approximately 54 genera. In particular, included within the library are 25 isolates from 16 understudied genera of the phylum Actinobacteria, that in total have an estimated 11 publications to date that report a natural product (Table S3, Supporting Information). Many of these understudied genera present colony morphologies that are atypical of traditional actinomycetes, which are often identified by their hard, leathery appearance, and presence of spores. Unbiased selection of all colonies in the early stages of bacterial isolation is credited with the inclusion of these in the library.

Use of IDBac to reduce redundancy in existing bacterial libraries.

We performed three previous expeditions to Iceland in 2013, 2014, and 2015. After each expedition, traditional morphology-guided colony picking strategies were employed. In total, the three previous expeditions yielded 833 bacterial isolates. In order to assess the degree of redundancy in this library, the MALDI-TOF/IDBac pipeline was applied. Using the parameters and cutoffs described in the previous section and discretion of the researcher as to what constituted overlap, 600 isolates were discarded due to taxonomic and natural product redundancy, reducing the library to 233 isolates, a 72.0% reduction in library size. This reduction has significant downstream cost savings. There are several major obstacles to creating a microbial library for drug discovery, and these have been well documented.[2,27,51] Generally speaking, therapeutic discovery pipelines have relied on libraries of natural product extracts or fractions, some of which were sourced from libraries of microorganisms. Each isolate in a library is processed through purification, liquid fermentation, extraction, and in some cases chromatographic separation to generate natural product material for the screening library. This process is costly and time consuming, as it may require tens to hundreds of thousands of dollars of investment to generate.[52] The current work provides a relatively rapid, cost-effective avenue for microbial library generation. It also precludes the need for DNA sequencing of isolates to determine phylogenetic similarity or chromatographic analyses of extracts or fractions to assess natural product overlap.

Visualization of sample relationships based on user defined categories.

IDBac was also designed to facilitate post-hoc analyses of field collections by providing visualizations across the protein MS dendrograms.[24] These visualizations occur in two forms, either in coloring the labels of the dendrogram or by graphing marks beside the dendrogram (Figures S6–S7, Supporting Information). In either case, users are able to associate metadata about isolates (*e.g.* geographic coordinates, sample source, water temperature, isolation media, etc.) to the dendrogram in order to quickly discern patterns across tens to thousands of isolates. For example, following the Iceland expedition we were able to quickly evaluate genus and species-level patterns of occurrence as a function of media formulation (Figures S6–S7, Supporting Information). This allows us to design more targeted bacterial isolation procedures for future collection expeditions.

Limitations of IDBac for library creation.

Limitations of the MALDI-TOF MS/IDBac method to external factors such as fluctuations in temperature and growth medium, and unintended microbial contamination, among other factors, were discussed previously.[21] In regard to the creation of microbial libraries, other limitations exist. The choice of MALDI matrix will affect which metabolites are ionized. The α-cyano-4-hydroxycinnamic acid matrix (CHCA) used in this study has been shown to ionize a variety of molecules such as PKS and PKS-NRPS-derived natural products,[53,54] alkaloids (200-400 Da), biosurfactants, and siderophores.[55] Regardless, MS profiles are not comprehensive of the full natural product producing capacity of each bacterial isolate. Cultivation of each isolate under multiple growth conditions followed by analysis in this pipeline will afford more comprehensive metabolic profiles. This is particularly feasible if the analysis is performed in multiwell plates, however one must be cognizant that gas phase natural products may diffuse into other wells and differentially influence natural product production. Regardless, it is recommended that each laboratory develop a standardized growth protocol to facilitate comparison of data across different strains, time points, and researchers.

IDBac is currently limited to qualitative manual assessment of MANs. A major limitation in past library creation methods was the implementation of visual bias while manually selecting colonies from petri dishes.[13] In the case of the current study, although a degree of bias remains, a researcher is basing the decision to keep/discard isolates based on both pseudo-phylogenetic groupings and natural product production capacity. This facilitates a more rigorous decision making process that can be experimentally verified across labs, based on spectral profiles. Further improvements include the development of algorithms to prioritize isolates directly from hierarchical clustering/MANs data. Lastly, it is possible that phylogenetically similar bacterial isolates produce isobaric species, giving the false appearance of natural product overlap. This can be improved through use of instruments with higher resolving power capabilities such as MALDI-FT-ICR or MALDI Orbitrap, though increased data size and storage needs must be addressed. Particularly when processing high numbers of bacterial isolates, this pipeline will result in some degree of redundancy added to a library, or in the removal of isolates that may harbor unique natural product biosynthetic potential. Despite these shortcomings, a library constructed using the methodology described here is more likely to harbor a higher degree of natural product diversity compared to one formed on the basis of 16S rRNA gene sequencing or morphology-guided practices.

Nearly two decades ago, Goodfellow and colleagues wrote extensively on the importance of ensuring taxonomic diversity in microbial libraries. In summarizing the methods needed to accomplish this, they noted the potential of MS techniques (pyrolysis MS in particular) to ensure rapid intraspecies pseudo-taxonomic dereplication of colonies growing directly on isolation plates. They opined that such a method would permit “the rational collection of colonies” and reduce “the requirement for time-consuming laboratory testing to distinguish duplicate colonies.”[51] Nearly 70 years have passed since researchers began to rely on microorganisms to produce antibiotics for use in the clinic, and despite advances such as colony picking automation, cultivation practices, and DNA sequencing, among other innovations that led or could lead to the discovery of new antibiotics,[9,10,13,56–58] researchers continue to create libraries that are plagued with redundancy because there has been no cost effective way to rapidly and simultaneously assess both the bacterial identity and the natural product capacity of colonies in high-throughput. In the current work, we employed our freely available pipeline to group thousands of environmental bacterial isolates based on similarities in protein MS spectra, and coupled this with global comparison of their natural product profiles. This pipeline allows for the rapid generation of diverse microbial ‘smart’ libraries that do not necessitate high numbers of entries.[52,59] Rather, it minimizes entries into the library, reduces natural product redundancy entering into downstream biological screening efforts, and significantly decreases costs associated with library generation.

EXPERIMENTAL SECTION

Sample collection and processing.

Sponges, macroalgae, marine and freshwater sediment samples were collected by SCUBA. Samples were processed immediately after collection. Approximately 1 cm^3 portions of all samples were submitted to heat in sterile water for 8 minutes at 60 °C. A similar portion of select samples was also plated on nutrient media without heat treatment. All samples were vortexed, diluted 1:10 in sterile, de-ionized water, and 10 μL used to inoculate five different agar-based nutrient media containing 28 μM cycloheximide (to inhibit the growth of fungi).

Nutrient plates were sealed with Parafilm® and left at room temperature. After 30-60 days, all distinguishable colonies were purified on 60 mm petri dishes containing high nutrient media (A1).

MALDI-TOF MS analysis.

From the petri dishes described in the previous section, 1,616 bacterial isolates were re-plated onto 48-well plates containing A1 media, and supplemented with or without synthetic ocean salts, depending on the sample source. After 7 days of growth three technical replicates of each bacterial isolate were applied, as a thin smear, to a 384-spot MALDI target plate (Bruker Daltonics, Billerica, MA) using a sterile toothpick. The number and type of replicates will vary depending on the researcher’s specific needs, however both biological and technical replicates are recommended as MALDI-TOF MS takes seconds to acquire and can be averaged within the IDBac pipeline. After transferring the colonies to the MALDI target plate, 1 μL of 70% formic acid (Optima, Fisher Chemical) was overlaid by pipette and allowed to dry, followed by overlay and drying of 1 μL of 10 mg/mL α-cyano-4-hydroxycinnamic acid (CHCA; recrystallized from the 98% pure Sigma-Aldrich, part-C2020) matrix prepared in 50% acetonitrile, 47.5% water, and 2.5% trifluoroacetic acid. All solvents were MS grade.

MALDI-TOF MS analysis were performed using an Autoflex Speed LRF mass spectrometer (Bruker Daltonics) equipped with a smartbeam™-II laser (355 nm). Automated data acquisitions were performed using flexControl software v. 3.4.135.0 (Bruker Daltonics) and flexAnalysis software v. 3.4. Protein spectra were recorded in positive linear mode (1200 shots; RepRate: 1000; delay: 29793 ns; ion source 1 voltage: 19.5 kV; ion source 2 voltage: 18.2 kV; lens voltage: 7.5 kV; mass range: 1.9 kDa to 2.1 kDa, matrix suppression cutoff: 1.5 kDa). Protein spectra were corrected with external Bruker Daltonics bacterial test standard (BTS). Natural product spectra were recorded in positive reflectron mode (5000 shots; RepRate: 2000 Hz; delay: 9297 ns; ion source 1 voltage: 19 kV; ion source 2 voltage:16.55 kV; lens voltage: 8.3 kV; mass range: 50 Da to 2,700 Da, matrix suppression cutoff: 50 Da). Natural Product spectra were corrected with external Bruker Daltonics peptide calibration standard and CHCA $[2M+H]^+$ (379.0930 Da). Detailed experimental settings are described in Clark and Costa *et al.*[21]

16S rRNA sequencing.

DNA from 361 bacterial isolates was extracted using a DNeasy UltraClean Microbial Kit (Qiagen). The 16S rRNA gene was amplified using 27F (5′-CAGAGTTTGATCCTGGCT-3′) and 1492R (5′-AGGAGGTGATCCAGCCGCA-3′)[60] primers using polymerase chain reaction (PCR) under the following conditions: initial denaturation at 95 °C for 5 minutes; followed by 35 cycles of denaturation at 95 °C for 15 seconds, annealing at 60 °C for 15 seconds, and extension at 72 °C for 30 seconds; and a final extension step at 72 °C for 2 minutes. PCR products were purified using a QIAquick PCR Purification kit from Qiagen and the amplicons sequenced by Sanger sequencing. Data were analyzed by Geneious V11.1.4 software. Using the SILVA Alignment, Classification and Tree (ACT) Service, all 361 isolates were aligned and a phylogenetic tree using

Fastree[61] was created, using default settings.[62] All the sequences were deposited in the NCBI nucleotide database (Table S2, Supporting Information).

Data availability.

MALDI-TOF MS data were deposited in MASSIVE (doi:10.25345/C5261K, MASSIVE accession: MSV000083461). The 16S rRNA sequences of environmental bacteria used in this study were deposited in GenBank with the accession numbers MK143106 to MK143377; MK163385 to MK163438; and MK168020 to MK168054 (Table S2 Supporting Information).

Software availability

IDBac is available at https://chasemc.github.io/IDBac and the source code of each release backed-up in permanence at DOI:10.5281/zenodo.1115619

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors would like to acknowledge Erlendur Bogason and his team at the Strytan Dive Center and Eydis Einarsdottir for assistance in sample collection, and Mark Sadek, Ying Gao, and Casey Murphy for assistance with microbial cultivation. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM125943 (BTM and LMS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was also supported by Icelandic Research Fund Grant 152336-051 (BTM and SO) and UIC startup funds (LMS).

REFERENCES

(1). Bérdy J J. Antibiot 2005, 58, 1–26. [PubMed: 15813176]

(2). Katz L; Baltz RH J. Ind. Microbiol. Biotechnol 2016, 43, 155–176. [PubMed: 26739136]

(3). Newman DJ; Cragg GM J. Nat. Prod 2016, 79, 629–661. [PubMed: 26852623]

(4). Wohlleben W; Mast Y; Stegmann E; Ziemert N Microb. Biotechnol 2016, 9, 541–548. [PubMed: 27470984]

(5). The Discovery and Development of Penicillin – American Chemical Society https://www.acs.org/content/acs/en/education/whatischemistry/landmarks/flemingpenicillin.html (accessed Jan 13, 2019).

(6). Fleming A Br. J. Exp. Pathol 1929 10, 226–236.

(7). Gause GF; G. BM. Nature 1944, 154, 703–703.

(8). Emmerich R; Löw O Zeitschrift für Hyg. und Infect 1899, 31, 1–65.

(9). Pereira DA; Williams JA Br. J. Pharmacol 2007, 152, 53–61. [PubMed: 17603542]

(10). Leeds JA; Schmitt EK; Krastel P Expert Opin. Investig. Drugs 2006, 15, 211–226.

(11). Lewis K Nat. Rev. Drug Discov 2013, 12, 371–387. [PubMed: 23629505]

(12). Larsen TO; Smedsgaard J; Nielsen KF; Hansen ME; Frisvad JC Nat. Prod. Rep 2005, 22, 672–695. [PubMed: 16311630]

(13). Genilloud O; González I; Salazar O; Martín J; Tormo JR; Vicente F J. Ind. Microbiol. Biotechnol 2011, 38, 375–389. [PubMed: 20931260]

(14). Okuda T; Ando K; Bills G Chapter 6 from Handbook of Industrial Mycology; CRC Press/Taylor and Francis Group, Boca Raton, FL, 2004.

(15). Fujimori F; Okuda T J. Antibiot 1994, 47, 173–182. [PubMed: 8150713]

(16). Knight V; Sanglier J-J; DiTullio D; Braccili S; Bonner P; Waters J; Hughes D; Zhang L Appl. Microbiol. Biotechnol 2003, 62, 446–458. [PubMed: 12838377]

(17). Fenical W; Jensen PR Nat. Chem. Biol 2006, 2, 666–673. [PubMed: 17108984]

(18). Koehn FE; Carter GT Nat. Rev. Drug Discov 2005, 4, 206–220. [PubMed: 15729362]

(19). Demain AL Nat. Biotechnol 2002, 20, 331. [PubMed: 11923826]

(20). Barrett JF Curr. Opin. Microbiol 2005, 8, 498–503. [PubMed: 16125445]

(21). Clark CM; Costa MS; Sanchez LM; Murphy BT Proc. Natl. Acad. Sci 2018, 115, 4981–4986. [PubMed: 29686101]

(22). Ryzhov V; Fenselau C Anal. Chem 2001, 73, 746–750. [PubMed: 11248887]

(23). Fernández-Álvarez C; Torres-Corral Y; Santos Y J. Proteomics 2018, 170, 59–69. [PubMed: 28939340]

(24). Clark CM; Costa MS; Conley E; Li E; M. SL.; Murphy BT J. Vis. Exp 2019, 147, e59219.

(25). Crüsemann M; O'Neill EC; Larson CB; Melnik AV; Floros DJ; da Silva RR; Jensen PR; Dorrestein PC; Moore BS J. Nat. Prod 2017, 80, 588–597. [PubMed: 28335604]

(26). Hubert J; Nuzillard J-M; Renault J-H Phytochem. Rev 2017, 16, 55–95.

(27). Liu X; Ashforth E; Ren B; Song F; Dai H; Liu M; Wang J; Xie Q; Zhang L J. Antibiot 2010, 63, 415–422. [PubMed: 20606699]

(28). Floros DJ; Jensen PR; Dorrestein PC; Koyama N Metabolomics 2016, 12, 145. [PubMed: 28819353]

(29). Wang M; Carver JJ; Phelan VV ; Sanchez LM; Garg N; Peng Y; Nguyen DD; Watrous J; Kapono CA; Luzzatto-Knaan T; et al. Nat. Biotechnol 2016, 34, 828–837. [PubMed: 27504778]

(30). Hindra; Huang T; Yang D; Rudolf JD; Xie P; Xie G; Teng Q; Lohman JR; Zhu X; Huang Y; et al. J. Nat. Prod 2014, 77, 2296–2303. [PubMed: 25238028]

(31). Jensen PR; Gontang E; Mafnas C; Mincer TJ; Fenical W Environ. Microbiol 2005, 7, 1039–1048. [PubMed: 15946301]

(32). Gontang EA; Fenical W; Jensen PR Appl. Environ. Microbiol 2007, 73, 3272–3282. [PubMed: 17400789]

(33). Joseph SJ; Hugenholtz P; Sangwan P; Osborne CA; Janssen PH Appl. Environ. Microbiol 2003, 69, 7210–7215. [PubMed: 14660368]

(34). Sauer S; Freiwald A; Maier T; Kube M; Reinhardt R; Kostrzewa M; Geider K PLoS One 2008, 3, e2843. [PubMed: 18665227]

(35). Suarez S; Ferroni A; Lotz A; Jolley KA; Guérin P; Leto J; Dauphin B; Jamet A; Maiden MCJ; Nassif X; et al. J. Microbiol. Methods 2013, 94, 390–396. [PubMed: 23916798]

(36). Sandrin TR; Goldstein JE; Schumaker S Mass Spectrom. Rev 2013, 32, 188–217. [PubMed: 22996584]

(37). Dieckmann R; Graeber I; Kaesler I; Szewzyk U; von Döhren H Appl. Microbiol. Biotechnol 2005, 67, 539–548. [PubMed: 15614563]

(38). Holland RD; Wilkes JG; Rafii F; Sutherland JB; Persons CC; Voorhees KJ; Lay JO Rapid Commun. Mass Spectrom 1996, 10, 1227–1232. [PubMed: 8759332]

(39). Popović NT; Kazazić SP; Strunjak-Perović I; Čož-Rakovac R Environ. Res 2017, 152, 7–16. [PubMed: 27741451]

(40). Croxatto A; Prod'hom G; Greub G FEMS Microbiol. Rev 2012, 36, 380–407. [PubMed: 22092265]

(41). Seuylemezian A; Aronson HS; Tan J; Lin M; Schubert W; Vaishampayan P Front. Microbiol 2018, 9, 780. [PubMed: 29867782]

(42). Strejcek M; Smrhova T; Junkova P; Uhlik O Front. Microbiol 2018, 9, 1294. [PubMed: 29971049]

(43). Rahi P; Prakash O; Shouche YS Front. Microbiol 2016, 7, 1359. [PubMed: 27625644]

(44). Hoffmann T; Krug D; Bozkurt N; Duddela S; Jansen R; Garcia R; Gerth K; Steinmetz H; Müller R Nat. Commun 2018, 9, 803. [PubMed: 29476047]

(45). Grubbs KJ; Bleich RM; Santa Maria KC; Allen SE; Farag S; AgBiome Team A; Shank EA; Bowers AA mSystems 2017, 2, e00040–17. [PubMed: 29152584]

(46). Adamek M; Alanjary M; Sales-Ortells H; Goodfellow M; Bull AT; Winkler A; Wibberg D; Kalinowski J; Ziemert N BMC Genomics 2018, 19, 426. [PubMed: 29859036]

(47). Letzel A-C; Li J; Amos GCA; Millán-Aguiñaga N; Ginigini J; Abdelmohsen UR; Gaudêncio SP; Ziemert N; Moore BS; Jensen PR Environ. Microbiol 2017, 19, 3660–3673. [PubMed: 28752948]

(48). Klementz D; Döring K; Lucas X; Telukunta KK; Erxleben A; Deubel D; Erber A; Santillana I; Thomas OS; Bechthold A; et al. Nucleic Acids Res. 2016, 44, D509–D514. [PubMed: 26615197]

(49). Seipke RF PLoS One 2015, 10, e0116457. [PubMed: 25635820]

(50). Sumner LW; Amberg A; Barrett D; Beale MH; Beger R; Daykin CA; Fan TW-M; Fiehn O; Goodacre R; Griffin JL; et al. Metabolomics 2007, 3, 211–221. [PubMed: 24039616]

(51). Bull AT; Ward AC; Goodfellow M Microbiol. Mol. Biol. Rev 2000, 64, 573–606. [PubMed: 10974127]

(52). Silver LL Clin. Microbiol. Rev 2011, 24, 71–109. [PubMed: 21233508]

(53). Spraker JE; Sanchez LM; Lowe TM; Dorrestein PC; Keller NP ISME J. 2016, 10, 2317–2330. [PubMed: 26943626]

(54). Moree WJ; McConnell OJ; Nguyen DD; Sanchez LM; Yang Y-L; Zhao X; Liu W-T; Boudreau PD; Srinivasan J; Atencio L; et al. ACS Chem. Biol 2014, 9, 2300–2308. [PubMed: 25058318]

(55). Phelan VV; Moree WJ; Aguilar J; Cornett DS.; Koumoutsi A; Noble SM; Pogliano K; Guerrero CA; Dorrestein PC. J. Bacteriol 2014, 196, 1683–1693. [PubMed: 24532776]

(56). Ling LL; Schneider T; Peoples AJ; Spoering AL; Engels I; Conlon BP; Mueller A; Schäberle TF; Hughes DE; Epstein S; et al. Nature 2015, 517, 455–459. [PubMed: 25561178]

(57). Berdy B; Spoering AL; Ling LL; Epstein SS Nat. Protoc 2017, 12, 2232–2242. [PubMed: 29532802]

(58). Mahler L; Wink K; Beulig RJ; Scherlach K; Tovar M; Zang E; Martin K; Hertweck C; Belder D; Roth M Sci. Rep 2018, 8, 13087. [PubMed: 30166560]

(59). Bull AT; Stach JE; Ward AC; Goodfellow M Antonie Van Leeuwenhoek 2005, 87, 65–79.

(60). Lane J,D. Nucleic Acid Tech. Bact. Syst 1991, 115–175.

(61). Price MN; Dehal PS; Arkin AP Mol. Biol. Evol 2009, 26, 1641–1650. [PubMed: 19377059]

(62). Pruesse E; Peplies J; Glöckner FO Bioinformatics 2012, 28, 1823–1829. [PubMed: 22556368]

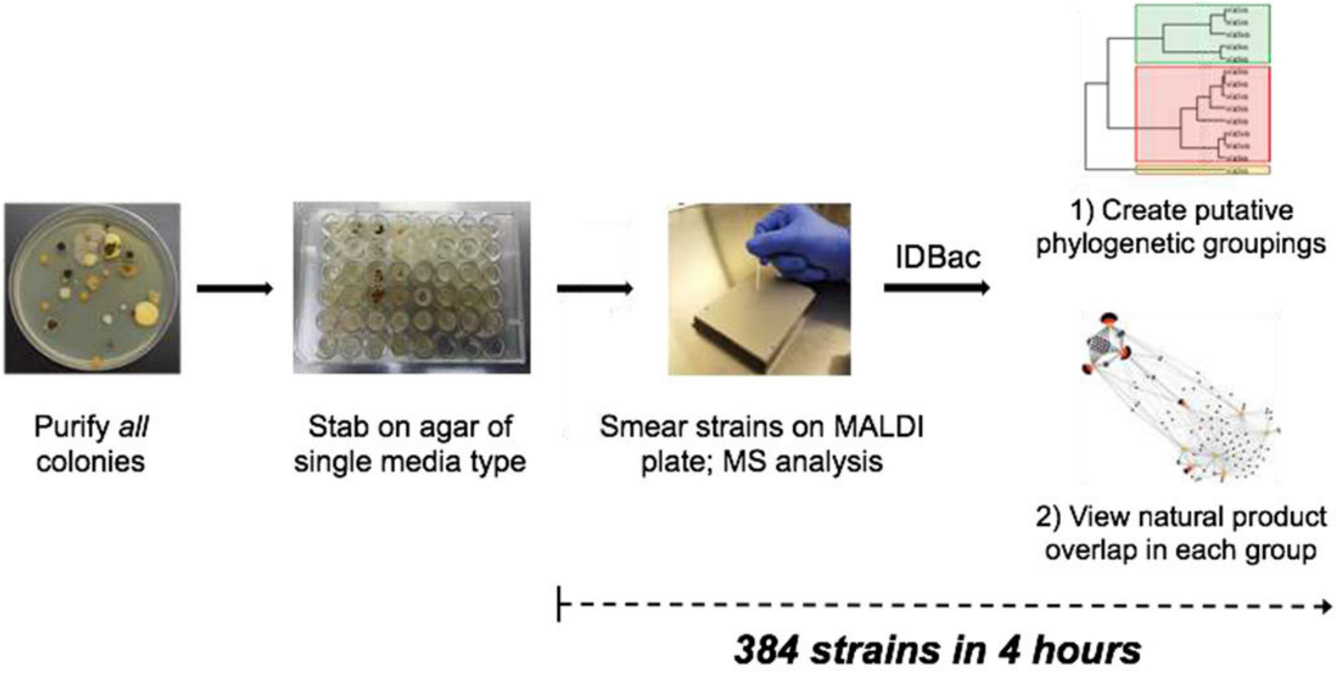

Purify *all* colonies
Stab on agar of single media type
Smear strains on MALDI plate; MS analysis
IDBac
1) Create putative phylogenetic groupings
2) View natural product overlap in each group
384 strains in 4 hours

Figure 1.

Collection trip workflow. Above is the recommended order of procedures for minimizing natural product redundancy in a microbial library. With a few basic laboratory supplies (nutrient agar, multi-well plates, toothpicks) and access to a MALDI-TOF MS, a researcher can build a diverse microbial library that incorporates information from protein and natural product MS analyses, without requiring liquid fermentation, extraction, or chromatographic analyses.

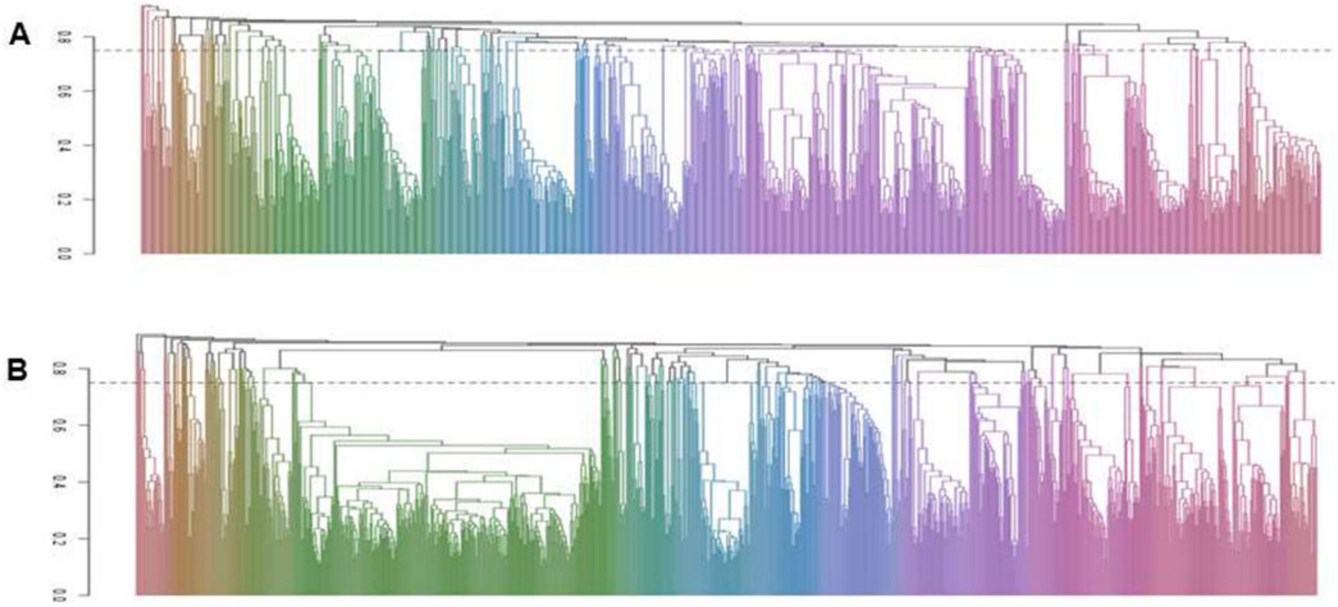
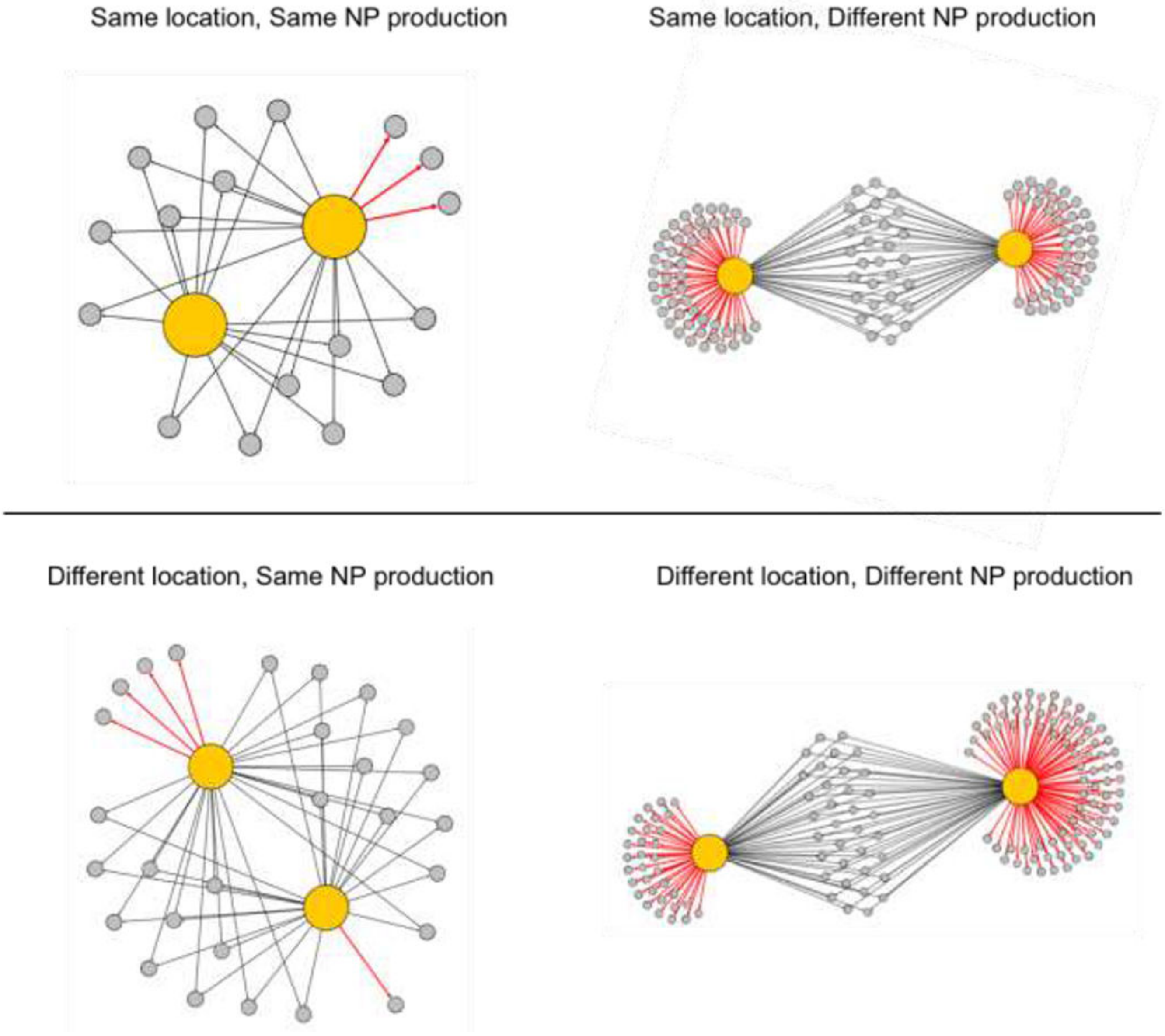


Figure 2.
Dendrograms resulting from hierarchical clustering of MALDI-TOF MS protein spectra in IDBac of 733 freshwater isolates (A), and 883 marine isolates (B). The dendrograms were "cut" at a threshold of 0.75 (dashed lines) to create groups of similar isolates. Dendrograms were colored according to groups created by these cuts.



Same location, Same NP production

Same location, Different NP production

Different location, Same NP production

Different location, Different NP production

Figure 3.

When comparing bacterial isolates that are morphologically indistinguishable and share >99% 16S rRNA gene sequence similarity, four relationships are possible. Large orange nodes represent individual bacterial isolates, while smaller gray nodes represent *m/z* values detected in their corresponding MS spectra (200-2,000 Da). “NP” denotes natural product. “Location” can refer to either sample source or geographic location from which the bacterial isolates were derived. For example, the top left and bottom left pane represent isolates that exhibit highly overlapping natural product populations, whereas the top right and bottom right panes represent isolates that exhibit significantly different natural product populations. MANs empower the user to rapidly differentiate natural product production in morphologically indistinguishable isolates.

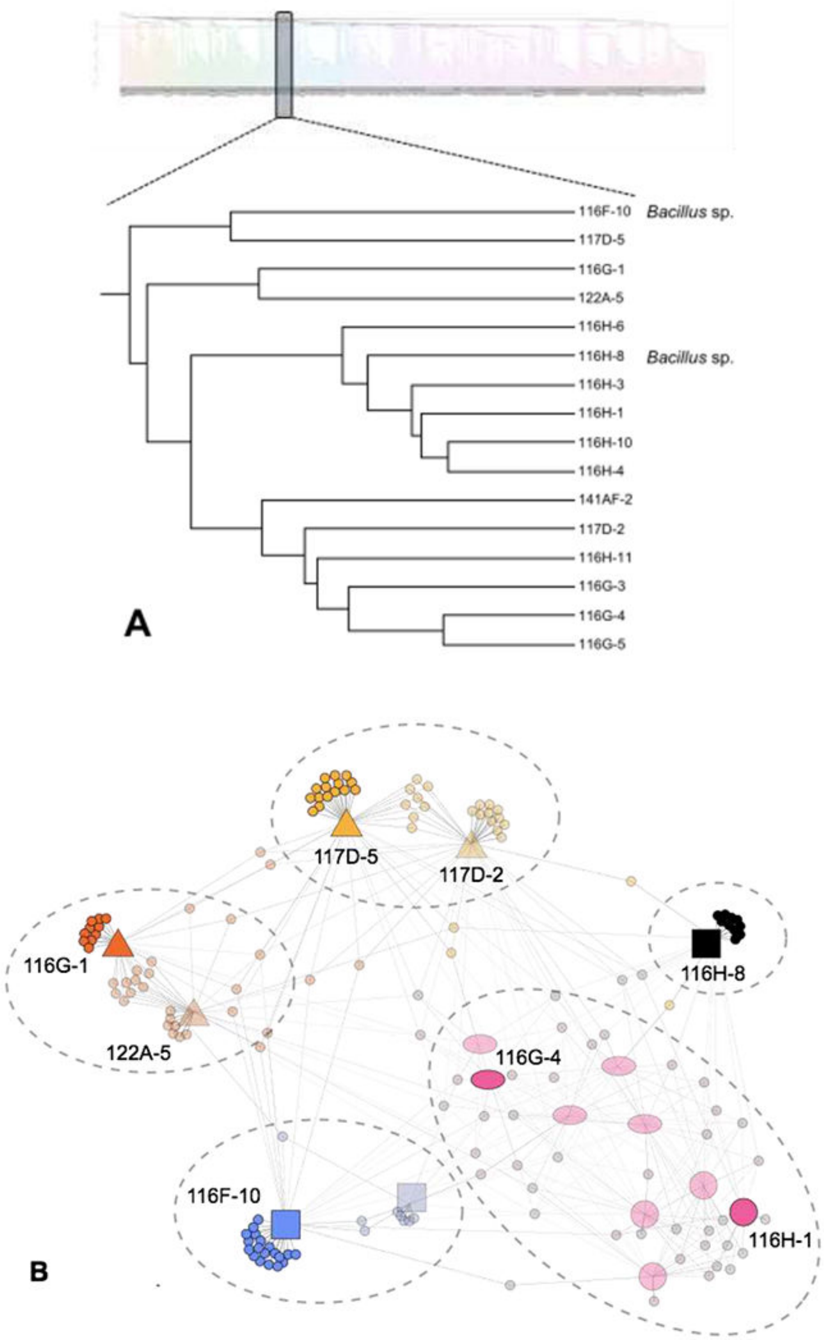


Figure 4.

(A) Protein MS clustering of 16 bacterial isolates (B) MAN generated for the group of selected strains, isolates colored by modularity score. Isolates chosen to be included in the library are colored opaque; transparent strains were discarded.