molecules

MDPI

Communication

Automated Microbial Library Generation Using the Bioinformatics Platform IDBac

Chase M. Clark [1], Linh Nguyen [2], Van Cuong Pham [2], Laura M. Sanchez [3] and Brian T. Murphy [1,*]

[1] Department of Pharmaceutical Sciences, Center for Biomolecular Sciences, College of Pharmacy, University of Illinois at Chicago, 833 S. Wood St., Chicago, IL 60612, USA; cmclark8@wisc.edu

[2] Institute of Marine Biochemistry, Vietnam Academy of Science and Technology, Nghiado, Caugiay, Hanoi 10000, Vietnam; nguyenthuylinhorc@gmail.com (L.N.); phamvc@imbc.vast.vn (V.C.P.)

[3] Department of Chemistry and Biochemistry, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA; lmsanche@ucsc.edu

[*] Correspondence: btmurphy@uic.edu; Tel.: +1-(312)-413-9057

Abstract: Libraries of microorganisms have served as a cornerstone of therapeutic drug discovery, though the continued re-isolation of known natural product chemical entities has remained a significant obstacle to discovery efforts. A major contributing factor to this redundancy is the duplication of bacterial taxa in a library, which can be mitigated through the use of a variety of DNA sequencing strategies and/or mass spectrometry-informed bioinformatics platforms so that the library is created with minimal phylogenetic, and thus minimal natural product overlap. IDBac is a MALDI-TOF mass spectrometry-based bioinformatics platform used to assess overlap within collections of environmental bacterial isolates. It allows environmental isolate redundancy to be reduced while considering both phylogeny and natural product production. However, manually selecting isolates for addition to a library during this process was time intensive and left to the researcher's discretion. Here, we developed an algorithm that automates the prioritization of hundreds to thousands of environmental microorganisms in IDBac. The algorithm performs iterative reduction of natural product mass feature overlap within groups of isolates that share high homology of protein mass features. Employing this automation serves to minimize human bias and greatly increase efficiency in the microbial strain prioritization process.

Keywords: IDBac; MALDI; bioinformatics; drug discovery; microorganisms; natural products

check for updates

Citation: Clark, C.M.; Nguyen, L.; Pham, V.C.; Sanchez, L.M.; Murphy, B.T. Automated Microbial Library Generation Using the Bioinformatics Platform IDBac. *Molecules* **2022**, *27*, 2038. https://doi.org/10.3390/molecules27072038

Academic Editor: José Rubén Tormo

Received: 21 February 2022
Accepted: 16 March 2022
Published: 22 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Recently, several steps in the microbial drug discovery pipeline have undergone significant advances including the analytical detection of natural products (NPs) [1,2], bioinformatic-driven NP dereplication [3], NP structure elucidation [4], genome mining [5], and biological activity screening [6,7]. However, the process of selecting culturable bacteria for addition to a strain library has seen relatively few widely applied conceptual changes since the 1940s [8]. To date, culturable isolates entering into a bacterial library are often selected based on common morphological phenotypes or 16S rRNA gene sequencing, which do not fully account for NP production [9–12]. These practices have simultaneously resulted in significant phylogenetic and NP overlap in strain libraries while failing to capture subtle differences in NP capacity between highly similar strains [13,14]. This results in a major obstacle facing microbial NP drug discovery efforts—the consistent re-discovery of known NPs [15,16].

To address this obstacle, we previously developed IDBac, an informatics pipeline that facilitates the data-driven prioritization of environmental bacterial isolates based on matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) analyses of cell material taken directly from a bacterial colony on a Petri dish [17,18]. In short, IDBac clusters spectra that represent ionized small proteins (2000–15,000 Daltons) of each isolate into a dendrogram, resulting in the creation of pseudo-phylogenetic groups. The program then allows users to assess the degree of NP overlap within each group of isolates through visualization of metabolite association networks (MANs, 200–2000 Daltons). Both datasets are acquired successively from a single sample preparation, and evaluation of both pseudo-phylogenetic and NP data affords researchers an opportunity to create a diverse microbial library while incorporating fewer environmental isolates. IDBac has the potential to significantly reduce costs associated with downstream library-based discovery, and towards community adoption; as of 1 March 2022, IDBac has been downloaded 1137 times from 36 countries.

Despite the efficiency of this pipeline, manual inspection of MANs was still required to determine areas of significant phylogenetic and NP overlap between isolates, resulting in potential researcher-introduced bias in the informatics pipeline [17]. This was particularly apparent when analyzing data from large sample collection expeditions. Typically, greater than 30 collected environmental samples plated on multiple media types resulted in up to thousands of environmental isolates to be processed and analyzed [19]. To mitigate potential bias, we designed and employed an algorithm to automate the prioritization of environmental bacterial isolates within each pseudo-phylogenetic group in the dendrogram. The algorithm seeks to iteratively select isolates until a user-specified percentage of unique m/z features is accumulated within each MAN. The result is the automated creation of a library of custom-size, maximized to capture unique m/z features. This process removes a degree of bias that accompanies manual library curation and serves to diminish the probability for human error while significantly increasing the speed and efficiency of the isolate selection process.

2. Results and Discussion

2.1. Collection and Processing of Environmental Samples

In collaboration with partners from the Vietnam Academy of Science and Technology and with the appropriate permission and permits, 130 marine samples (sediments, sponges, algae) were collected from several locations around the island of Phu Quoc in Southern Vietnam using SCUBA. Samples were plated on eight different media types to maximize bacterial diversity. After an incubation period, all distinguishable bacterial colonies were isolated and purified. A total of 819 bacterial isolates were purified and MALDI-TOF MS analysis was carried out using three technical replicates per isolate, acquiring data in protein and NP regions, as previously described [18,19]. The resulting data were processed in IDBac to generate a dendrogram of 819 isolates (Figure 1), grouped by protein spectrum similarity. An arbitrary cut-height designed to approximate species-level phylogenetic branching points was applied (see red line in Figure 1) resulting in the generation of 83 pseudo-phylogenetic subgroups (and therefore 83 MANs), which can be evaluated for NP overlap.

Since each MAN is composed of closely related isolates, analysis of m/z overlap within the MANs and subsequent selection of isolates with minimally overlapping mass features is critical to the creation of a streamlined bacterial library (see Figure 1b for an example MAN). However, this process relies on the judgement of an individual researcher as to what constitutes significant feature overlap, which can jeopardize the quality of a library due to varying levels of user bias. To automate and thus streamline this process as well as mitigate user-to-user variation, an algorithm was created and embedded into the IDBac code.

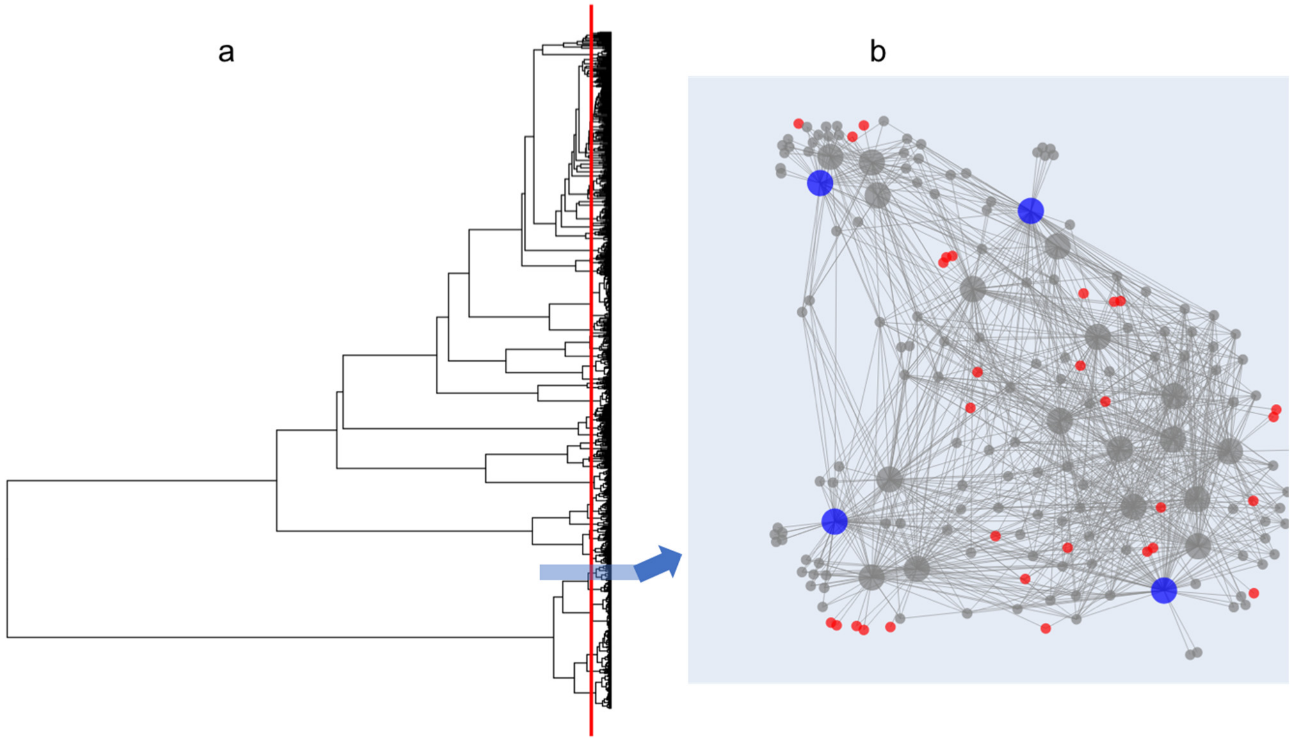

Figure 1. Dendrogram (**a**) created from 819 bacterial isolates (2000–15,000 Da; protein spectra run in triplicate for each isolate). An example Metabolite Association Network (MAN) is displayed in (**b**), where large circles represent bacterial isolates that are connected to smaller circles representing m/z features. Colored circles represent isolates and mass features chosen by the algorithm, discussed further in Figure 2. The pseudo-phylogenetic grouping of these (**b**) isolates is depicted by the blue box on the dendrogram.

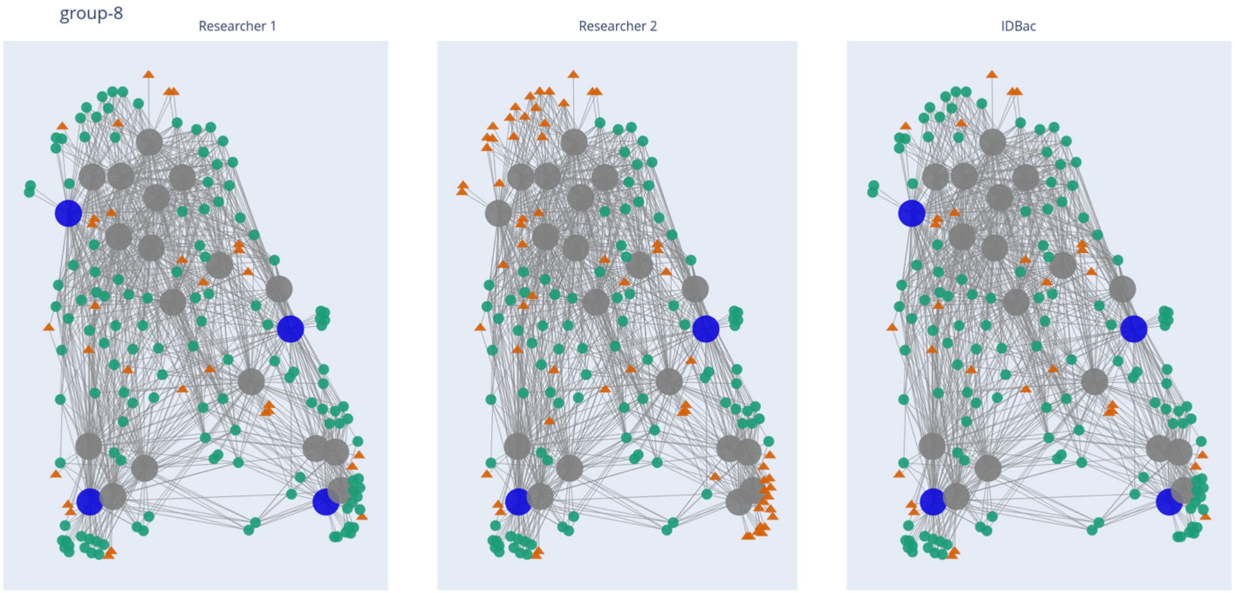

Figure 2. MAN of pseudo-phylogenetic Group 8, depicting the NP coverage resulting from manual and automatic prioritization processes. Large nodes represent bacterial isolates and small nodes represent mass features. Large blue nodes are isolates selected to be added to the library, while small green circles represent mass features "captured" by the selected isolates. Large grey nodes are isolates not selected, while small orange triangles are mass features "missed" by the selected isolates. Links to the other 82 groups may be found in Section Data Availability Statement.

2.2. Application of the Isolate Prioritization Algorithm to a Collection of Unknown Environmental Bacterial Isolates

A prioritization function was added to the IDBac R package (see Data Availability Statement). The function requires a pseudo-phylogenetic dendrogram as input (created using IDBac) and a connection to an IDBac database (created by the IDBac software). The user is required to select the number of pseudo-phylogenetic groups that will be analyzed, either by "cutting" the dendrogram at an arbitrary height or by defining the desired number of groups. The user then defines a threshold percentage of unique small molecule mass features to be kept within each pseudo-phylogenetic group. A small threshold will result in fewer chosen isolates but a higher number of missed mass features while a large threshold will result in a higher number of chosen isolates and lower number of missed mass features. While for this study we chose an arbitrary 75% threshold, we would suggest this number be adjusted, per dataset, until the number of chosen isolates represents the workload a laboratory can handle in downstream studies. The algorithm then proceeds through each pseudo-phylogenetic group, identifies the isolate with the most NP mass features, selects and removes that isolate and its mass features from the group, then iterates this process until the threshold percentage of total mass features within the group is reached.

The automated prioritization algorithm was applied to the MALDI-TOF MS dataset generated for the 819 bacterial isolates. Using a threshold defined to capture 75% of the mass features, the algorithm selected 189 isolates from the 83 pseudo-phylogenetic subgroups. These 189 isolates represented a 77% reduction from the total number of isolates collected from the bacterial diversity plates. The algorithm achieved similar performance when compared with two researchers who were assigned to create a library from the same data set, with instructions to "minimize the number of isolates in the library while capturing approximately 75% of the mass features" (see Supplementary Materials). Using manual curation, Researchers 1 and 2 achieved a 71% and 64% reduction in total number of isolates collected, finishing with a total of 236 and 293 isolates, respectively. While the algorithm captured similar chemical space, it was able to do so using 47 and 104 fewer isolates than Researchers 1 and 2, respectively. This reduction is significant given the high downstream costs of processing these additional isolates through the traditional pipeline of fermentation, extraction, fraction library generation, biological screening, and lead dereplication. Further, Researchers 1 and 2 each required approximately four hours to prioritize isolates from the MANs, while the automated process required only seconds.

The 819 isolates produced 6624 distinct mass features distributed across 83 MANs. It is important to note that multiple mass features can represent a singular chemical structure/entity (see further discussion of isotopologues and adducts in "Study limitations and perspectives"). An example comparison of manual and automated outputs from one pseudo-phylogenetic group is shown in Figure 2. The MANs of Group 8 depict the number of isolates/mass features retained through the library creation process; this was repeated for each of the 82 remaining groups. Both manual and automatic prioritization methods were successful at capturing the majority of m/z features present across the collection: 5455 overlapping features that represented an estimated 82% of the chemical space detected by the MALDI-TOF MS (Figure 3). The algorithm's average across the 83 groups was higher than 75% because it is designed to select isolates continuously from a group until a minimum threshold percent of mass features is accumulated. Because mass features do not accumulate one-by-one but rather by their presence in iteratively chosen isolates, the accumulated mass features can be greater than the input threshold. Regardless, a major advantage of automating the isolate selection process is that the user is guaranteed to reach a threshold mass retention within each pseudo-phylogenetic group.

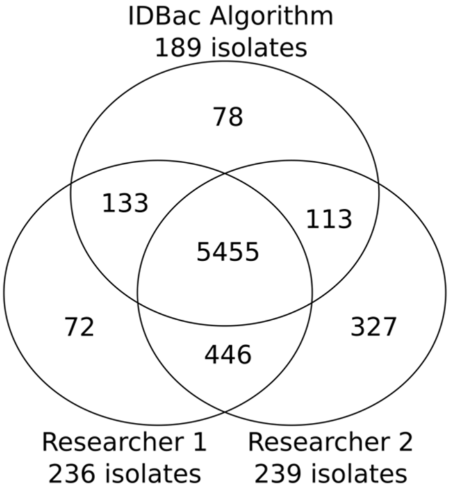

Figure 3. Venn diagram showing the number of mass features "captured" by the two researchers and the IDBac algorithm; 82% were captured by all three, though the algorithm did so with 47 and 104 fewer isolates than Researchers 1 and 2, respectively. The total isolates recovered from each group: Researcher 1: 236; Researcher 2: 293; the IDBac algorithm: 189.

2.3. Study Limitations and Perspectives

For nearly eight decades the process of microbial library generation and curation has been labor intensive and required a significant effort and input from experts, who employed either morphology-guided colony selection from diversity plates, 16S rRNA sequencing of isolates, or both in order to create an isolate library for natural product discovery [8–10]. The current study represents an advance toward an automated microbial library creation pipeline, though some challenges remain. In its current form the algorithm retains the limitations of IDBac. Most relevant is a lack of deisotoping MALDI-TOF MS spectra (i.e., isotopologues are not collapsed into a single mass feature) and no dimensionality to allow for separation of isobars. Future versions of the prioritization algorithm may consider adding awareness of isotopologues and molecular analogues to allow for more high mass accuracy feature selection. Finally, within IDBac, the choice of where to 'cut' a dendrogram to create pseudo-phylogenetic groups remains manual, as automating this step remains a longstanding issue in the bacterial MALDI-TOF MS [20,21] field largely due to a lack of reference spectra from which to train more sophisticated computational models. With the availability of accurate, automated clustering of protein spectra, it is conceivable that the entire front end library generation process could be automated—robotic colony isolation and plating onto a MALDI target plate [22], MALDI-TOF MS data acquisition and processing, isolate prioritization, and isolate distribution into a library.

3. Conclusions

In the current study, we developed an algorithm that works within the existing bioinformatics platform IDBac to automate the prioritization of hundreds to thousands of environmental microorganisms based on MALDI-TOF MS protein and NP spectra. It does so through the iterative reduction of NP mass feature overlap within groups of isolates that share high homology of protein mass features. Employing this automation serves to minimize human bias and greatly increase efficiency during a major front end bottleneck in the drug discovery process. When applied to 819 unknown environmental bacterial isolates clustered into 83 pseudo-phylogenetic groups, the algorithm selected 189 isolates for retention, discarding 630 isolates. The 189 isolates harbored the majority of small molecule m/z features present across the collection, an estimated 82% of the chemical space detected by the MALDI-TOF MS. The algorithm achieved similar performance when compared with two researchers who were assigned to create a library from the same data set, and did so in seconds, compared to multiple hours required for manual curation.

4. Experimental Section

Sample Collection and Processing MALDI-TOF MS

In October of 2018, a total of 130 sediment, sponge, and algae samples were collected from 18 locations in Vietnam and processed immediately following collection. Samples were collected under a long-standing partnership between UIC and the Institute of Marine Biochemistry at VAST, and with the appropriate permits and benefit sharing agreements. Samples were plated on SWA, SCA, M1, A1, PDA, ISP1, ISP2, and NZSG nutrient media (see Supplementary Material for recipes). After incubating up to 60 days, all distinguishable colonies were isolated and purified onto 60 mm Petri dishes containing A1 marine media. A total of 819 marine bacterial isolates were cultivated on high nutrient, marine media and incubated at room temperature for seven days. Following incubation, a small portion of each isolate was transferred to a 384-spot MALDI target plate (Bruker Daltonics, Billerica, MA, USA) in three technical replicates using a sterile toothpick. Then, 1 µL of 70% MS grade formic acid and 1 µL of MALDI matrix (10 mg/mL alpha-cyano-4-hydroxycinnamic acid prepared in 50% acetonitrile, 47.5% water and 2.5% trifluoroacetic acid) was added to each sample spot. MALDI-TOF MS analysis was performed using an Autoflex Speed LRF mass spectrometer (Bruker Daltonics) equipped with a Smartbeam-II laser (355 nm). Further MALDI-TOF MS settings and procedures can be found in our previous IDBac publications [17,19] and detailed protocol [18].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/molecules27072038/s1, Table S1: Media formulations used in this study.

Author Contributions: Conceptualization, C.M.C., B.T.M.; methodology, C.M.C.; software, C.M.C.; validation, C.M.C.; formal analysis, C.M.C.; investigation, C.M.C., L.N., B.T.M.; resources, L.N., V.C.P., L.M.S., B.T.M.; data curation, C.M.C., L.N.; writing—original draft preparation, C.M.C., L.N., B.T.M.; writing—review and editing, C.M.C., L.N., L.M.S., B.T.M.; visualization, C.M.C.; supervision, V.C.P., L.M.S., B.T.M.; project administration, B.T.M.; funding acquisition, C.M.C., L.M.S., V.C.P., B.T.M. All authors have read and agreed to the published version of the manuscript.

Funding: 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 to BTM and LMS; F31AT010419 to CMC; R01AI160696-01 to LN and Tuan Anh Tran; and the National Institute of Allergy and Infectious Diseases/Fogarty International Center under Award Number D43TW010530 to BTM and VCP. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: All data (including MANs), code, and docker instructions necessary to replicate these results have been deposited at Zenodo, doi:10.5281/zenodo.5873982. IDBac is freely available from github.com/chasemc/IDBacApp and doi:10.5281/zenodo.1115619. All MALDI-TOF MS spectra have been deposited in MASSIVE: MSV000089059; doi:10.25345/C5DN40009.

Acknowledgments: The authors would like to acknowledge Vu Thi Quyen, Vu Van Nam, Duong Thu Trang, and Tran Tuan Anh for assistance in sample collection, and Doel Soejarto and Hung Van Nguyen for mentorship and guidance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bouslimani, A.; Sanchez, L.M.; Garg, N.; Dorrestein, P.C. Mass spectrometry of natural products: Current, emerging and future technologies. *Nat. Prod. Rep.* **2014**, *31*, 718–729. [CrossRef] [PubMed]
2. Hou, Y.; Braun, D.R.; Michel, C.R.; Klassen, J.L.; Adnani, N.; Wyche, T.P.; Bugni, T.S. Microbial strain prioritization using metabolomics tools for the discovery of natural products. *Anal. Chem.* **2012**, *84*, 4277–4283. [PubMed]
3. 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.; et al. Molecular networking as a dereplication strategy. *J. Nat. Prod.* **2013**, *76*, 1686–1699. [CrossRef] [PubMed]

4. Molinski, T.F. Microscale methodology for structure elucidation of natural products. *Curr. Opin. Biotechnol.* **2010**, *21*, 819–826. [PubMed]
5. Bauman, K.D.; Butler, K.S.; Moore, B.S.; Chekan, J.R. Genome mining methods to discover bioactive natural products. *Nat. Prod. Rep.* **2021**, *38*, 2100–2129.
6. Wilson, B.A.P.; Thornburg, C.C.; Henrich, C.J.; Grkovic, T.; O'Keefe, B.R. Creating and screening natural product libraries. *Nat. Prod. Rep.* **2020**, *37*, 893–918.
7. Potts, M.B.; Kim, H.S.; Fisher, K.W.; Hu, Y.; Carrasco, Y.P.; Bulut, G.B.; Ou, Y.-H.; Herrera-Herrera, M.L.; Cubillos, F.; Mendiratta, S.; et al. Using functional signature ontology (FUSION) to identify mechanisms of action for natural products. *Sci. Signal.* **2013**, *6*, ra90. [CrossRef]
8. Hernandez, A.; Nguyen, L.T.; Dhakal, R.; Murphy, B.T. The need to innovate sample collection and library generation in microbial drug discovery: A focus on academia. *Nat. Prod. Rep.* **2021**, *38*, 292–300. [CrossRef]
9. Liu, X.; Ashforth, E.; Ren, B.; Song, F.; Dai, H.; Liu, M.; Wang, J.; Xie, Q.; Zhang, L. Bioprospecting microbial natural product libraries from the marine environment for drug discovery. *J. Antibiot.* **2010**, *63*, 415–422. [CrossRef]
10. Knight, V.; Sanglier, J.-J.; DiTullio, D.; Braccili, S.; Bonner, P.; Waters, J.; Hughes, D.; Zhang, L. Diversifying microbial natural products for drug discovery. *Appl. Microbiol. Biotechnol.* **2003**, *62*, 446–458. [CrossRef]
11. Okuda, T.; Ando, K.; Bills, G. Fungal germplasm for drug discovery and industrial applications. In *Handbook of Industrial Mycology*; An, Z., Ed.; CRC Press: Boca Raton, FL, USA, 2004; pp. 123–166. [CrossRef]
12. Fujimori, F.; Okuda, T. Application of the random amplified polymorphic DNA using the polymerase chain reaction for efficient elimination of duplicate strains in microbial screening. *J. Antibiot.* **1994**, *47*, 173–182. [CrossRef] [PubMed]
13. Ziemert, N.; Lechner, A.; Wietz, M.; Millán-Aguiñaga, N.; Chavarria, K.L.; Jensen, P.R. Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E1130-9. [CrossRef] [PubMed]
14. Chevrette, M.G.; Carlos-Shanley, C.; Louie, K.B.; Bowen, B.P.; Northen, T.R.; Currie, C.R. Taxonomic and metabolic incongruence in the ancient genus *Streptomyces*. *Front. Microbiol.* **2019**, *10*, 2170. [CrossRef] [PubMed]
15. Baltz, R.H. Marcel Faber Roundtable: Is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration? *J. Ind. Microbiol. Biotechnol.* **2006**, *33*, 507–513. [CrossRef] [PubMed]
16. 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. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 5601–5606. [CrossRef]
17. Clark, C.M.; Costa, M.S.; Sanchez, L.M.; Murphy, B.T. Coupling MALDI-TOF mass spectrometry protein and specialized metabolite analyses to rapidly discriminate bacterial function. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4981–4986. [CrossRef]
18. Clark, C.M.; Costa, M.S.; Conley, E.; Li, E.; Sanchez, L.M.; Murphy, B.T. Using the open-source MALDI TOF-MS IDBac pipeline for analysis of microbial protein and specialized metabolite data. *J. Vis. Exp.* **2019**, *147*, e59219. [CrossRef]
19. 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. *J. Nat. Prod.* **2019**, *82*, 2167–2173. [CrossRef]
20. Strejcek, M.; Smrhova, T.; Junkova, P.; Uhlik, O. Whole-cell MALDI-TOF MS versus 16S rRNA gene analysis for identification and dereplication of recurrent bacterial isolates. *Front. Microbiol.* **2018**, *9*, 1294. [CrossRef]
21. Dumolin, C.; Aerts, M.; Verheyde, B.; Schellaert, S.; Vandamme, T.; Van der Jeugt, F.; De Canck, E.; Cnockaert, M.; Wieme, A.D.; Cleenwerck, I.; et al. Introducing SPeDE: High-throughput dereplication and accurate determination of microbial diversity from matrix-assisted laser desorption-ionization time of flight mass spectrometry data. *mSystems* **2019**, *4*, e00437-19. [CrossRef]
22. Chudejova, K.; Bohac, M.; Skalova, A.; Rotova, V.; Papagiannitsis, C.C.; Hanzlickova, J.; Bergerova, T.; Hrabák, J. Validation of a novel automatic deposition of bacteria and yeasts on MALDI target for MALDI-TOF MS-based identification using MALDI Colonyst robot. *PLoS ONE* **2017**, *12*, e0190038.