AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

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

Victoria M. Anderson,[a,b,c] Karen L. Wendt,[a,b,c] Fares Z. Najar,[c,d] Laura-Isobel McCall,[c,e,f] Robert H. Cichewicz[a,b,c]

[a]Natural Products Discovery Group, University of Oklahoma, Norman, Oklahoma, USA
[b]Institute for Natural Products Applications and Research Technologies, University of Oklahoma, Norman, Oklahoma, USA
[c]Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma, USA
[d]Chemistry and Biochemistry Bioinformatics Core, University of Oklahoma, Norman, Oklahoma, USA
[e]Department of Microbiology and Plant Biology, University of Oklahoma, Norman, Oklahoma, USA
[f]Laboratories of Molecular Anthropology and Microbiome Research, University of Oklahoma, Norman, Oklahoma, USA

ABSTRACT The success of natural product-based drug discovery is predicated on having chemical collections that offer broad coverage of metabolite diversity. We propose a simple set of tools combining genetic barcoding and metabolomics to help investigators build natural product libraries aimed at achieving predetermined levels of chemical coverage. It was found that such tools aided in identifying overlooked pockets of chemical diversity within taxa, which could be useful for refocusing collection strategies. We have used fungal isolates identified as *Alternaria* from a citizen-science-based soil collection to demonstrate the application of these tools for assessing and carrying out predictive measurements of chemical diversity in a natural product collection. Within *Alternaria*, different subclades were found to contain nonequivalent levels of chemical diversity. It was also determined that a surprisingly modest number of isolates (195 isolates) was sufficient to afford nearly 99% of *Alternaria* chemical features in the data set. However, this result must be considered in the context that 17.9% of chemical features appeared in single isolates, suggesting that fungi like *Alternaria* might be engaged in an ongoing process of actively exploring nature's metabolic landscape. Our results demonstrate that combining modest investments in securing internal transcribed spacer (ITS)-based sequence information (i.e., establishing gene-based clades) with data from liquid chromatography-mass spectrometry (i.e., generating feature accumulation curves) offers a useful route to obtaining actionable insights into chemical diversity coverage trends in a natural product library. It is anticipated that these outcomes could be used to improve opportunities for accessing bioactive molecules that serve as the cornerstone of natural product-based drug discovery.

IMPORTANCE Natural product drug discovery efforts rely on libraries of organisms to provide access to diverse pools of compounds. Actionable strategies to rationally maximize chemical diversity, rather than relying on serendipity, can add value to such efforts. Readily implementable biological (i.e., ITS sequence analysis) and chemical (i.e., mass spectrometry-based feature and scaffold measurements) diversity assessment tools can be employed to monitor and adjust library development tactics in real time. In summary, metabolomics-driven technologies and simple gene-based specimen barcoding approaches have broad applicability to building chemically diverse natural product libraries.

KEYWORDS natural products, LC-MS metabolomics, chemical diversity, drug discovery, fungi, library design, metabolomics

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

Citation Anderson VM, Wendt KL, Najar FZ, McCall L-I, Cichewicz RH. 2021. Building natural product libraries using quantitative clade-based and chemical clustering strategies. mSystems 6:e00644-21. https://doi.org/10.1128/mSystems.00644-21.

Editor Marcelino Gutierrez, INDICASAT

Copyright © 2021 Anderson et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Robert H. Cichewicz, rhcichewicz@ou.edu, or Laura-Isobel McCall, lmccall@ou.edu.

Received 25 May 2021
Accepted 16 September 2021
Published 26 October 2021

of many chemical collections used in modern drug screening and molecular probe development throughout academia, industry, and government. Small-molecule libraries based on organic compounds of various sizes (e.g., <900 Da for most synthetic libraries but ranging up to around ~2,000 Da for some natural product collections) play a dominant role in such efforts, with many collections accommodating a variety of screening and discovery methodologies (e.g., fragment based, target focused, diversity oriented, combinatorial, DNA encoded, repurposed, and virtual) (1–6).

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 optimal collection sizes, are largely driven by adherence to dogma or convenience rather than evidence-based reasoning. Such questions grow increasingly relevant, as opinions influencing the last 4 decades of library design have shifted tremendously, with the large collections of the 1980s and 1990s (e.g., combinatorial chemistry [7]) being replaced by smaller tailored collections in the early 2000s (e.g., "focused" collections [8, 9]) and moving toward megascale libraries in recent years (e.g., DNA encoded libraries [10–15]).

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 and plants). Many ideas have emerged related 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 (16–19). Despite the tremendous ingenuity and effort that have 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 is 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 (20, 21).

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 (22, 23). These factors can make the rational design of natural product libraries challenging. Therefore, methods to perform chemical diversity measurements have the potential to aid and inform 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) (24). 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 based on the use of standard bioassay-driven screening procedures, 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, Letzel and colleagues carried out a survey of natural product biosynthetic gene cluster diversity represented in 119 *Salinispora* sp. genomes (25). 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 new natural products (26), 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 that exhibit gene-level equivalencies being assigned different binomial names (27). In other instances, the high degree of genetic diversity 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 (28, 29). Complicating these matters, the regional variation and global distribution of most fungal taxa remain poorly defined, which has given rise to unresolved questions about the true extent of biological and chemical diversity throughout the fungal kingdom. Here, 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 (30, 31) known to produce many types of metabolites (32–37). 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.

RESULTS AND DISCUSSION

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 (38, 39), 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 (40–42) to *Alternaria* type strain data available in GenBank and defined by Woudenberg et al. (31). Upon plating, all strains exhibited colony morphologies consistent with the genus *sensu stricto*.

Alternaria exemplifies 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 characteristics, telemorphic states, various marker genes, and more to delineate this group and its allied genera (28, 31, 43). 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* exhibits tremendous morphological and genetic plasticity. Recognizing that 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 liquid chromatography-mass spectrometry (LC-MS) metabolome profiling data. This bifunctional approach offered insights into the relationship between phylogeny and chemistry, which enabled (i) assessment of natural product chemical diversity within species complexes, (ii) identification of prospective pools of under- and oversampled secondary-metabolite scaffolds, and (iii) 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 (44–46), we view our approach as a departure from prior schemes, considering its quantitative aspects that we now explore.

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, we can use certain low-cost and

minimally intensive tools to gain functional insights concerning their phylogenetic relationships. For fungi, the ITS barcode system serves as one such tool, offering an efficient way to establish a working set of phylogenetic associations among environmental isolates (29). The phylogenetic analysis of *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 (27, 29).

Principal-coordinate analysis (PCoA) 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 (see Fig. S1 in the supplemental material).

The results generated from the ITS barcode and metabolomics data sets were overlaid, demonstrating a high degree of consensus between the two models (Fig. 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 (clades) to reveal otherwise unexpected pockets of chemical divergence within genetic groups. A few 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 (47) and/or genomic/epigenome-scale events that resulted in the loss of chemical scaffolds (48, 49), which served to differentiate clusters 1, 3, 4, 5, and 6 from cluster 2. Analyses in this report were conducted in parallel on both clade and cluster models, with the chemical cluster model generating results similar to those of the clade model (Fig. S2B and C and Fig. S3, S5, and S7).

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 (Fig. 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 detected only 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 artifacts.

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 (Fig. 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 supplemental material.

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 (Fig. 3A). This observation held true ($P < 0.0001$) after subsampling of the clades to alleviate potential errors introduced due to sample size nonequivalence (Fig. S2A). Relatively few outliers were detected within the genetic clades, indicating high levels of consistency for the metabolic output of the

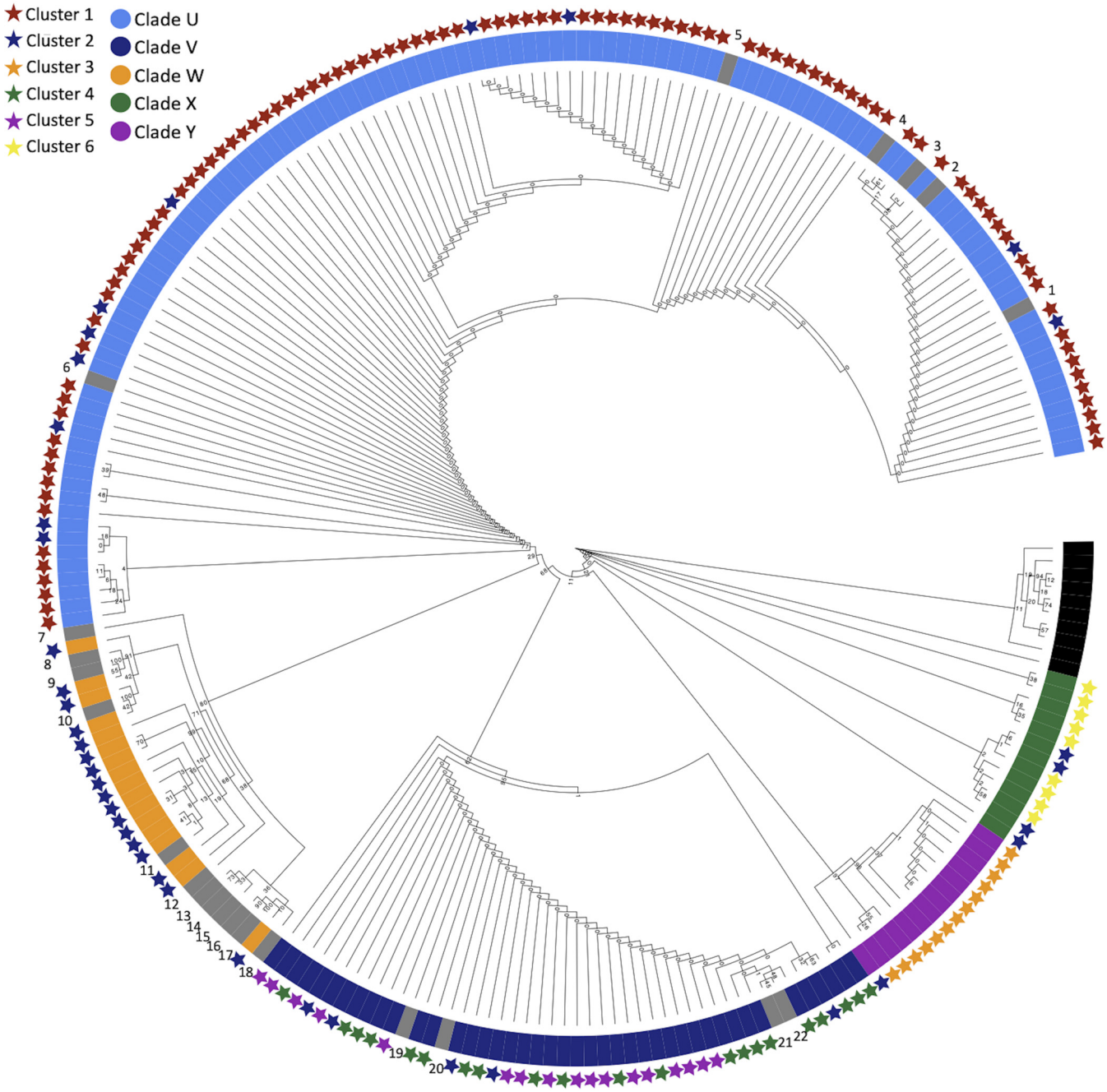

FIG 1 Genetic and chemical clustering of *Alternaria*. ITS phylogeny of *Alternaria* isolates is shown. Inner ring indicates the clade, while color-coded stars represent the chemical cluster. The clades and clusters show remarkable overlap but also reveal a hidden chemical cluster within clade V. Numbers indicate type strain data from GenBank (Table S2).

isolates in each group. Clades V, W, and X were found to have significantly fewer features than clade U (Tukey's honestly significant difference [HSD] of analysis of variance [ANOVA], $P < 0.0001$ in all cases), suggesting that clade U is chemically more diverse than the other clades.

Only 1.9% of features (205) were detected in all clades, comprising the core metabolome of the *Alternaria* isolates (Fig. 3B). While up to 40% of chemistry is shared between two or more 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

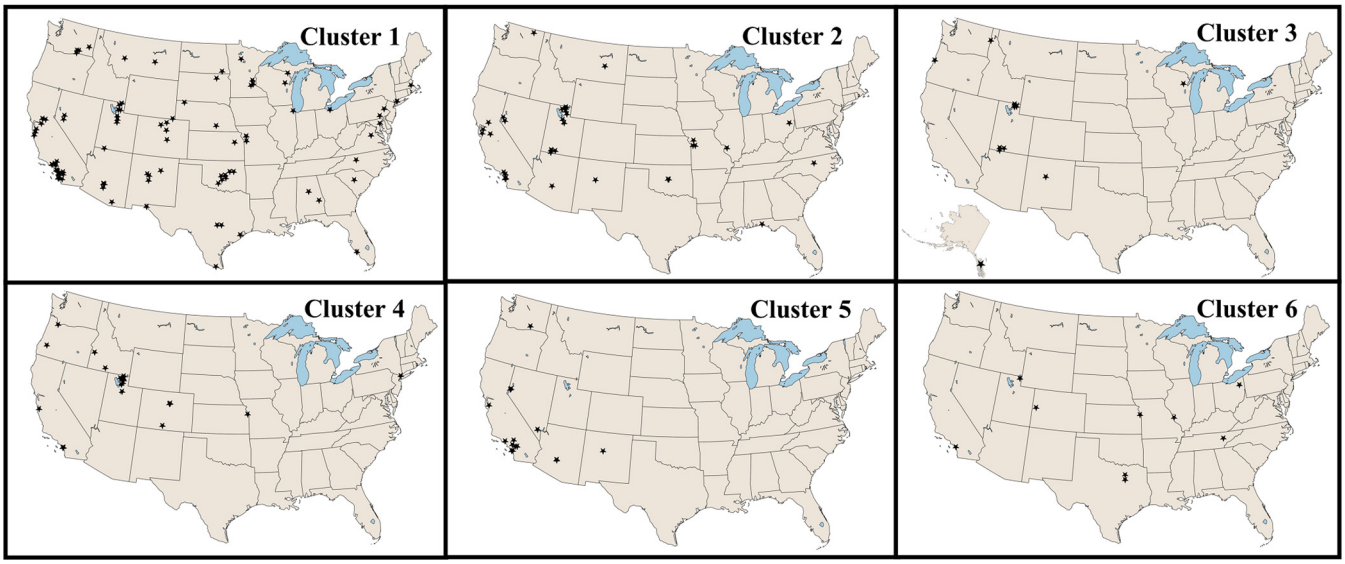

FIG 2 Chemical and geographical distribution of *Alternaria*. Shown is the geographical distribution of isolates by chemical cluster. Whereas clusters 1, 2, and 6 are well distributed throughout the study area, clusters 3, 4, and 5 occupy more limited ranges.

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*.

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 (Fig. 4A). The results show that despite a large degree of ascribed taxonomic diversity in *Alternaria*, a surprisingly limited number of isolates are 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,

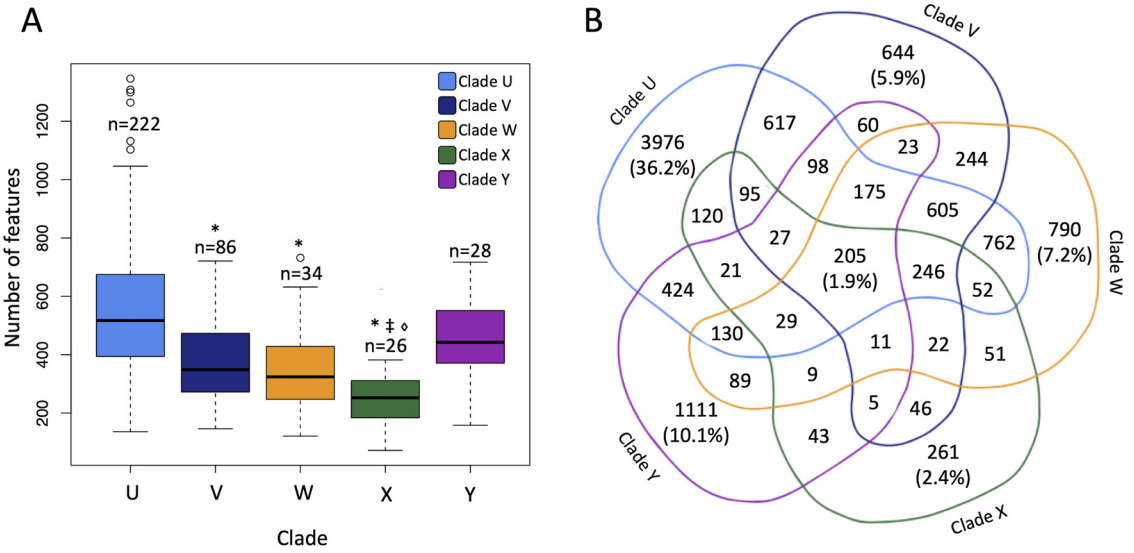

FIG 3 Summary of feature diversity in *Alternaria*. (A) Alpha diversity of genetic clades. The median numbers of chemical features differed significantly by clade. The asterisk indicates a statistically significant difference from clade U. The double dagger indicates a statistically significant difference from clade V. The diamond indicates a statistically significant difference from clade Y. (B) Venn diagram of features by clade.

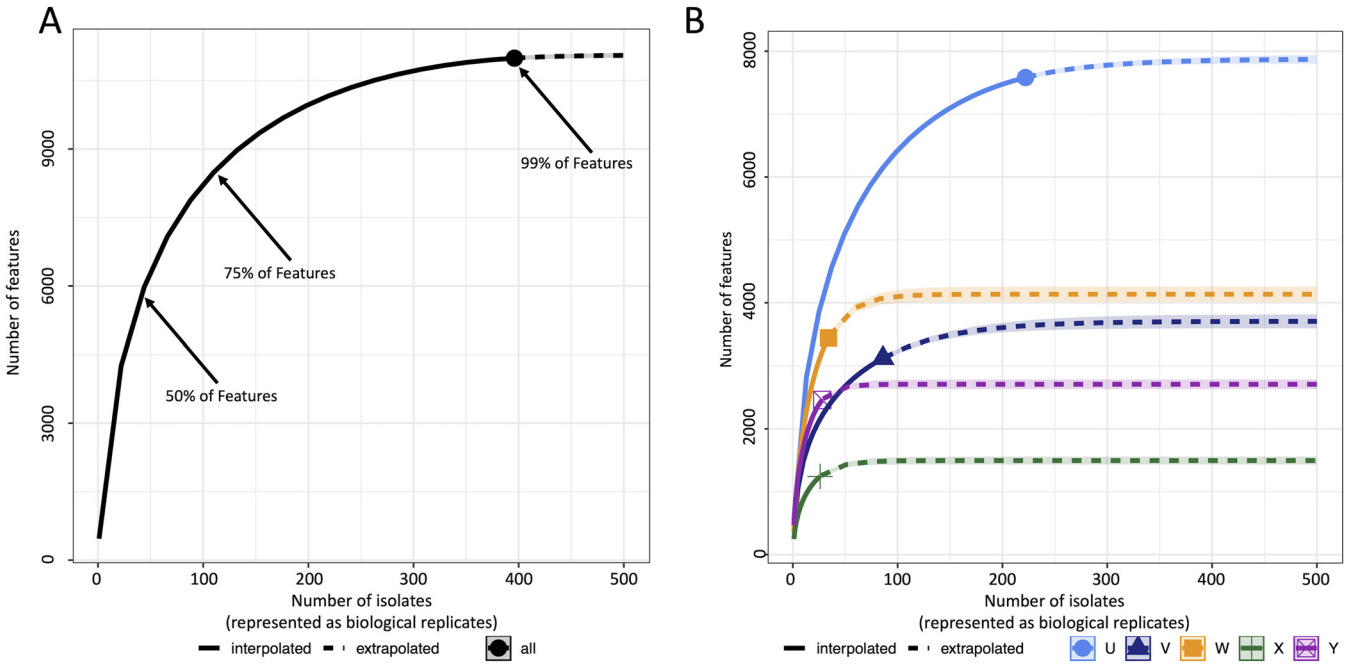

FIG 4 Chemical diversity curves and data extrapolation. (A) Rarefaction curve of chemical features within *Alternaria*. (B) Rarefaction curves for each ITS-based clade within *Alternaria*.

142, and 195 isolates were anticipated to provide 75%, 90%, 95%, and 99%, respectively, of *Alternaria* features (Fig. 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.

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 subgenus clades has the potential to afford a complementary perspective for library design. Clade-based feature accumulation curves (Fig. 4B) showed that feature coverage levels of 99% were achievable in clades U (contained the most feature-rich isolates [Fig. 2A]) and X (contained the most feature-poor isolates [Fig. 2A]), 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 (Fig. 2A), was found to require the lowest number of isolates (39 isolates) to achieve a level of 99% feature coverage. Clade V contained the third highest level of features per isolate (Fig. 2A), 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 being embedded in clade V. Clade W contained the second lowest number of features per isolate (Fig. 2A) 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 prespecified levels of feature accumulation coverage. Related to these efforts, rarefaction curve slopes were plotted in relationship to the number of samples representing each clade (Fig. S4). The results of that analysis revealed that 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.

Probing of chemical scaffold 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 tends 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 (47, 50, 51). Consolidating chemical features that share underlying structural similarities into groups referred to as scaffolds is one approach to account for this phenomenon. Molecular networking (52–55) is a method that has gained widespread use to build scaffold-level relationships in the field of natural products (56–59).

Using molecular networking to identify structurally related metabolites from *Alternaria*, the 10,991 molecular features were combined into 5,754 scaffolds (Fig. 5A). Upon removing singleton scaffolds (4,193) from the data set, 17.2% of the scaffolds (285) were found to be shared by all five ITS-based clades (Fig. 5B). These shared scaffolds represented the core metabolome of the *Alternaria* encountered in this study. We also found that 32.5% (539) of the nonsingleton 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 (Fig. 2A). Thus, we speculate that the relative quantities of chemical features detected within taxa might serve as a surrogate measure for predicting their comparative levels of relative scaffold diversity, although further analysis will be necessary to explore this. 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% of scaffolds, which indicates that many chemical scaffolds are conserved among *Alternaria* isolates.

Applying clade and cluster data to assess progress toward goals for natural product library coverage. Considering the entwined functions that phylogeny and chemistry have 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 outcome of a collection (Fig. 6A and Fig. S6). The accumulation curves revealed that the 111 isolates in clade U could provide 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 all scaffolds encountered from *Alternaria* (Fig. 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 product 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 undersampling of the theoretical chemical diversity within a given taxon.

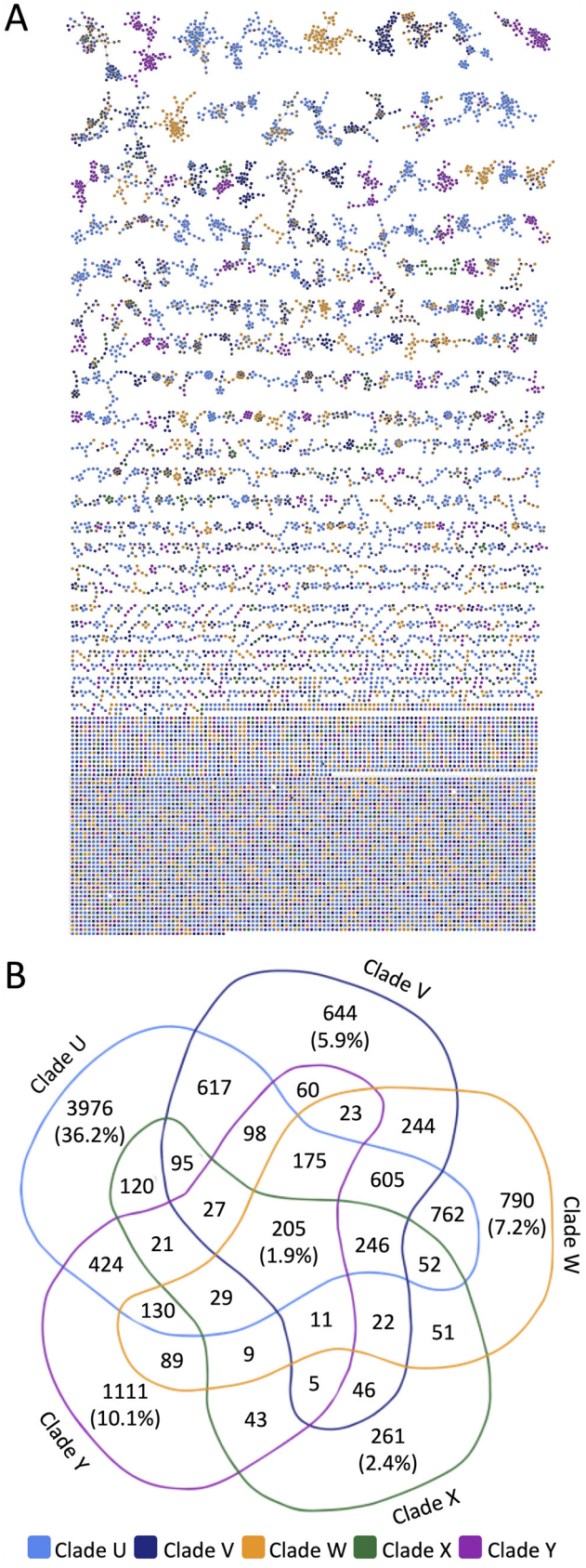

FIG 5 Scaffold diversity in *Alternaria*. (A) Results from molecular networking analysis constructed from LC-MS data reveal 5,754 subnetworks/scaffolds. Nodes are colored by clade. (B) Venn diagram illustrating chemical scaffolds by clade.

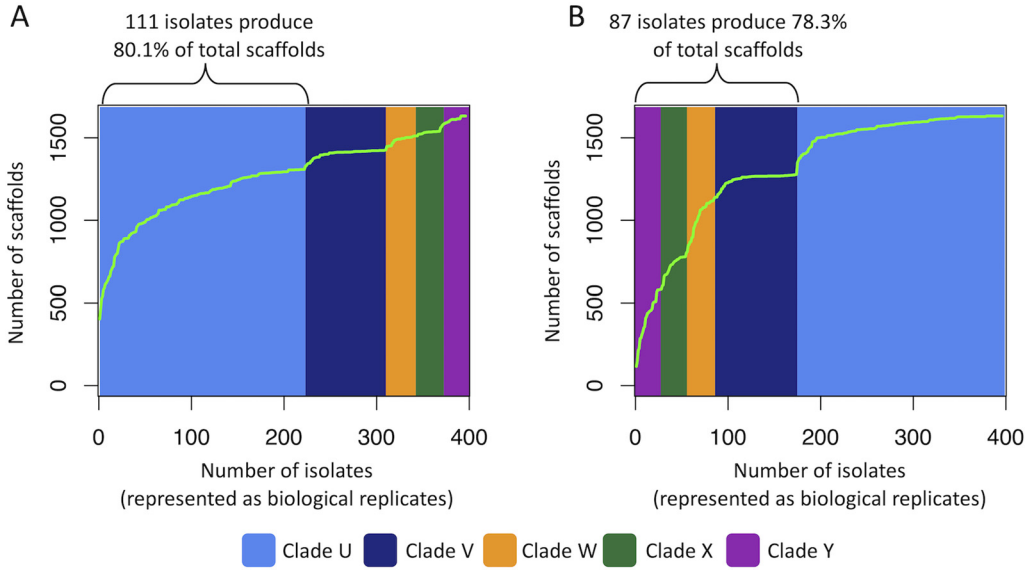

FIG 6 Visualization of scaffold accumulation models. (A) Scaffold accumulation curve generated starting with the most abundant clade (clade U) before adding isolates from the less abundant clades. (B) Scaffold accumulation curve generated by starting with less abundant clades (clades Y, X, W, and V) before introducing isolates from the most abundant clade.

Putting the pieces together to create natural product chemical collections. It is our opinion that many efforts to construct natural product 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 or features versus scaffolds), the best routes are likely to rely upon well-balanced sample collection strategies that combine appropriate amounts of chemical breadth in the resultant libraries. The purpose of our effort to measure natural product diversity was to afford researchers opportunities to establish library development goals and provide the means for assessing progress toward those targets. 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, 17.9% of metabolite features were found in only a single isolate. Thus, overly stringent measures aimed at simply capturing only 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 also well poised to deliver the chemical matter needed to drive fruitful drug discovery operations.

MATERIALS AND METHODS

General sample selection and culture. A cohort of 198 fungal isolates from the University of Oklahoma, Citizen Science Soil Collection, that had been identified as *Alternaria* were used in this study (Table S1). The map illustrating the sites where the isolates were obtained (Fig. 2) was generated in qGIS v3.10. The fungal isolates were identified based on BLASTN (60) comparisons of their ITS sequence data to the sequences of *Alternaria* type strains deposited in GenBank (60). 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 (61).

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 of 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 beads. Samples were homogenized using a BulletBlender (Next Advantage) set at maximum speed

for 5 min. The 5.8S-ITS region was amplified by PCR using primers ITS1 (5′-TCCGTAGGTGAACCTGCGG-3′) and ITS4 (5′-TCCTCCGCTTATTGATATGC-3′) (62). Amplification and confirmation of PCR product formation were performed using a LightCycler 480 Instrument II (Roche) operated under the following conditions: 1 cycle of denaturation at 94°C for 2 min followed by 40 cycles of denaturation at 94°C for 1 min, annealing at 50°C for 1 min, and extension at 72°C for 1 min. Samples were submitted to Genewiz for Sanger sequencing with forward and reverse reads assembled using PhredPhrap (release 29) (minimum phred score: 50) (63, 64). Sequences were prepared for phylogenetic analysis using MEGA-X (65). ITS sequences for *Alternaria* type strains were obtained from the NCBI database (Table S2) (60). An outgroup consisting of five *Penicillium* spp. and five *Clonostachys* species isolates retrieved from the University of Oklahoma, Citizen Science Soil Collection, were used for tree rooting. Sequences were aligned using Clustal W in Mega X. Neighbor-joining tree analysis was carried out with 500 bootstraps using the Kimura2+G algorithm (65, 66).

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- by 100-mm borosilicate tubes were placed on a Tecan Freedom EVO platform and 3 ml of ethyl acetate was added to each sample. After extraction for 4 h, 3 ml of water was 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.

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 by 2.1 mm, 1.7-μm particle size, 100-Å pore size; Phenomenex, Torrance, CA). The mobile phase consisted of LC-MS-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 min), which was increased to 100% acetonitrile over 8 min and held at 100% acetonitrile for 2 min. Between samples, the eluent was returned to 5% acetonitrile over 30 s and held for 1 min 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 provided in Table S3 (part A).

Data processing and analyses. Data were processed using MZmine v2.33 with the parameters provided in Table S3 (part B) (67). 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 (68) using a Bray-Curtis distance metric (69). The selection of 6 clusters was determined to be optimal based on a silhouette plot. Results were visualized using Emperor (70). Feature accumulation curves were made in vegan using binarized tabulated data (71), 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 (72, 73). 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 (74). 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 (75). Global Natural Products Social Molecular Networking (GNPS) feature-based molecular networking was performed (52, 53) using output from MZmine2 (67) with the parameters described in Table S3 (part C).

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 to MW729257. Codes for other analysis methods can be accessed on GitHub at https://github.com/NPDG/Alternaria.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, TIF file, 0.8 MB.

FIG S2, TIF file, 1.5 MB.

FIG S3, TIF file, 2 MB.

FIG S4, TIF file, 0.6 MB.

FIG S5, TIF file, 1.9 MB.

FIG S6, TIF file, 0.5 MB.

FIG S7, TIF file, 1.2 MB.

TABLE S1, DOCX file, 0.04 MB.
TABLE S2, DOCX file, 0.01 MB.
TABLE S3, DOCX file, 0.02 MB.

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 the citizen science program and data management. We acknowledge R. da Silva (University of São Paulo) for sharing the ClusterApp software.

V.M.A. and R.H.C. designed research; V.M.A., L.-I.M., and K.L.W. performed research; F.Z.N. contributed bioinformatics tools; V.M.A., K.L.W., L.-I.M., and R.H.C. analyzed data; and V.M.A., L.-I.M., and R.H.C. wrote the paper.

REFERENCES

1. Dandapani S, Rosse G, Southall N, Salvino JM, Thomas CJ. 2012. Selecting, acquiring, and using small molecule libraries for high-throughput screening. Curr Protoc Chem Biol 4:177–191. https://doi.org/10.1002/9780470559277.ch110252.
2. Franzini RM, Neri D, Scheuermann J. 2014. DNA-encoded chemical libraries: advancing beyond conventional small-molecule libraries. Acc Chem Res 47:1247–1255. https://doi.org/10.1021/ar400284t.
3. Boldt GE, Dickerson TJ, Janda KD. 2006. Emerging chemical and biological approaches for the preparation of discovery libraries. Drug Discov Today 11:143–148. https://doi.org/10.1016/S1359-6446(05)03697-4.
4. Webb TR. 2005. Current directions in the evolution of compound libraries. Curr Opin Drug Discov Devel 8:303–308.
5. Liu R, Li X, Lam KS. 2017. Combinatorial chemistry in drug discovery. Curr Opin Chem Biol 38:117–126. https://doi.org/10.1016/j.cbpa.2017.03.017.
6. Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, Johnston SE, Vrcic A, Wong B, Khan M, Asiedu J, Narayan R, Mader CC, Subramanian A, Golub TR. 2017. The Drug Repurposing Hub: a next-generation drug library and information resource. Nat Med 23:405–408. https://doi.org/10.1038/nm.4306.
7. Mario Geysen H, Schoenen F, Wagner D, Wagner R. 2003. Combinatorial compound libraries for drug discovery: an ongoing challenge. Nat Rev Drug Discov 2:222–230. https://doi.org/10.1038/nrd1035.
8. Spear KL, Brown SP. 2017. The evolution of library design: crafting smart compound collections for phenotypic screens. Drug Discov Today Technol 23:61–67. https://doi.org/10.1016/j.ddtec.2017.05.001.
9. Lenci E, Trabocchi A. 2019. Smart design of small-molecule libraries: when organic synthesis meets cheminformatics. Chembiochem 20:1115–1123. https://doi.org/10.1002/cbic.201800751.
10. Goodnow RA, Dumelin CE, Keefe AD. 2017. DNA-encoded chemistry: enabling the deeper sampling of chemical space. Nat Rev Drug Discov 16:131–147. https://doi.org/10.1038/nrd.2016.213.
11. Song M, Hwang GT. 2020. DNA-encoded library screening as core platform technology in drug discovery: its synthetic method development and applications in DEL synthesis. J Med Chem 63:6578–6599. https://doi.org/10.1021/acs.jmedchem.9b01782.
12. Favalli N, Bassi G, Scheuermann J, Neri D. 2018. DNA-encoded chemical libraries—achievements and remaining challenges. FEBS Lett 592:2168–2180. https://doi.org/10.1002/1873-3468.13068.
13. Gong Z, Hu G, Li Q, Liu Z, Wang F, Zhang X, Xiong J, Li P, Xu Y, Ma R, Chen S, Li J. 2017. Compound libraries: recent advances and their applications in drug discovery. Curr Drug Discov Technol 14:216–228. https://doi.org/10.2174/1570163814666170425155154.
14. Busby SA, Carbonneau S, Concannon J, Dumelin CE, Lee Y, Numao S, Renaud N, Smith TM, Auld DS. 2020. Advancements in assay technologies and strategies to enable drug discovery. ACS Chem Biol 15:2636–2648. https://doi.org/10.1021/acschembio.0c00495.
15. Noah JW. 2010. New developments and emerging trends in high-throughput screening methods for lead compound identification. Int J High Throughput Screen 1:141–149. https://doi.org/10.2147/IJHTS.S8683.
16. Koehn FE, Carter GT. 2005. The evolving role of natural products in drug discovery. Nat Rev Drug Discov 4:206–220. https://doi.org/10.1038/nrd1657.
17. Wagenaar MM. 2008. Pre-fractionated microbial samples—the second generation natural products library at Wyeth. Molecules 13:1406–1426. https://doi.org/10.3390/molecules13061406.
18. Thornburg CC, Britt JR, Evans JR, Akee RK, Whitt JA, Trinh SK, Harris MJ, Thompson JR, Ewing TL, Shipley SM, Grothaus PG, Newman DJ, Schneider JP, Grkovic T, O'Keefe BR. 2018. NCI Program for Natural Product Discovery: a publicly-accessible library of natural product fractions for high-throughput screening. ACS Chem Biol 13:2484–2497. https://doi.org/10.1021/acschembio.8b00389.
19. Eldridge GR, Vervoort HC, Lee CM, Cremin PA, Williams CT, Hart SM, Goering MG, O'Neil-Johnson M, Zeng L. 2002. High-throughput method for the production and analysis of large natural product libraries for drug discovery. Anal Chem 74:3963–3971. https://doi.org/10.1021/ac025534s.
20. Huggins DJ, Venkitaraman AR, Spring DR. 2011. Rational methods for the selection of diverse screening compounds. ACS Chem Biol 6:208–217. https://doi.org/10.1021/cb100420r.
21. Koutsoukas A, Paricharak S, Galloway WRJD, Spring DR, Ijzerman AP, Glen RC, Marcus D, Bender A. 2014. How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. J Chem Inf Model 54:230–242. https://doi.org/10.1021/ci400469u.
22. Lind AL, Wisecaver JH, Lameiras C, Wiemann P, Palmer JM, Keller NP, Rodrigues F, Goldman GH, Rokas A. 2017. Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. PLoS Biol 15:e2003583. https://doi.org/10.1371/journal.pbio.2003583.
23. Wisecaver JH, Rokas A. 2015. Fungal metabolic gene clusters—caravans traveling across the genomes and environments. Front Microbiol 6:161. https://doi.org/10.3389/fmicb.2015.00161.
24. Baltz RH. 2006. Marcel Faber Roundtable: is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration? J Ind Microbiol Biotechnol 33:507–513. https://doi.org/10.1007/s10295-005-0077-9.
25. Letzel AC, Li J, Amos GCA, Millán-Aguiñaga N, Ginigini J, Abdelmohsen UR, Gaudêncio SP, Ziemert N, Moore BS, Jensen PR. 2017. Genomic insights into specialized metabolism in the marine actinomycete *Salinispora*. Environ Microbiol 19:3660–3673. https://doi.org/10.1111/1462-2920.13867.
26. Atanasov AG, Zotchev SB, Dirsch VM, Supuran CT, the International Natural Product Sciences Taskforce. 2021. Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov 20:200–216. https://doi.org/10.1038/s41573-020-00114-z.

27. Raja HA, Miller AN, Pearce CJ, Oberlies NH. 2017. Fungal identification using molecular tools: a primer for the natural products research community. J Nat Prod 80:756–770. https://doi.org/10.1021/acs.jnatprod.6b01085.
28. Lawrence DP, Gannibal PB, Peever TL, Pryor BM. 2013. The sections of *Alternaria*: formalizing species-group concepts. Mycologia 105:530–546. https://doi.org/10.3852/12-249.
29. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Consortium FB, Fungal Barcoding Consortium Author List. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for *Fungi*. Proc Natl Acad Sci U S A 109:6241–6246. https://doi.org/10.1073/pnas.1117018109.
30. Lawrence DP, Rotondo F, Gannibal PB. 2016. Biodiversity and taxonomy of the pleomorphic genus *Alternaria*. Mycol Progress 15:3. https://doi.org/10.1007/s11557-015-1144-x.
31. Woudenberg J, Groenewald J, Binder M, Crous P. 2013. *Alternaria* redefined. Stud Mycol 75:171–212. https://doi.org/10.3114/sim0015.
32. Egidi E, Delgado-Baquerizo M, Plett JM, Wang J, Eldridge DJ, Bardgett RD, Maestre FT, Singh BK. 2019. A few Ascomycota taxa dominate soil fungal communities worldwide. Nat Commun 10:2369. https://doi.org/10.1038/s41467-019-10373-z.
33. Van der Waals J, Korsten L, Aveling T, Denner F. 2003. Influence of environmental factors on field concentrations of *Alternaria solani* conidia above a South African potato crop. Phytoparasitica 31:353–364. https://doi.org/10.1007/BF02979806.
34. Cai S, King JB, Du L, Powell DR, Cichewicz RH. 2014. Bioactive sulfur-containing sulochrin dimers and other metabolites from an *Alternaria* sp. isolate from a Hawaiian soil sample. J Nat Prod 77:2280–2287. https://doi.org/10.1021/np5005449.
35. Zwickel T, Kahl SM, Rychlik M, Müller ME. 2018. Chemotaxonomy of mycotoxigenic small-spored *Alternaria* fungi—do multitoxin mixtures act as an indicator for species differentiation? Front Microbiol 9:1368. https://doi.org/10.3389/fmicb.2018.01368.
36. Carter AC, King JB, Mattes AO, Cai S, Singh N, Cichewicz RH. 2019. Natural-product-inspired compounds as countermeasures against the liver carcinogen aflatoxin B1. J Nat Prod 82:1694–1703. https://doi.org/10.1021/acs.jnatprod.9b00290.
37. Kim M-Y, Sohn JH, Ahn JS, Oh H. 2009. Alternaramide, a cyclic depsipeptide from the marine-derived fungus *Alternaria* sp. SF-5016. J Nat Prod 72:2065–2068. https://doi.org/10.1021/np900464p.
38. Du L, Robles AJ, King JB, Powell DR, Miller AN, Mooberry SL, Cichewicz RH. 2014. Crowdsourcing natural products discovery to access uncharted dimensions of fungal metabolite diversity. Angew Chem 126:823–828. https://doi.org/10.1002/ange.201306549.
39. Jewett MC, Hofmann G, Nielsen J. 2006. Fungal metabolite analysis in genomics and phenomics. Curr Opin Biotechnol 17:191–197. https://doi.org/10.1016/j.copbio.2006.02.001.
40. Morris MH, Smith ME, Rizzo DM, Rejmánek M, Bledsoe CS. 2008. Contrasting ectomycorrhizal fungal communities on the roots of co-occurring oaks (*Quercus* spp.) in a California woodland. New Phytol 178:167–176. https://doi.org/10.1111/j.1469-8137.2007.02348.x.
41. Izzo A, Agbowo J, Bruns TD. 2005. Detection of plot-level changes in ectomycorrhizal communities across years in an old-growth mixed-conifer forest. New Phytol 166:619–630. https://doi.org/10.1111/j.1469-8137.2005.01354.x.
42. Nilsson RH, Tedersoo L, Abarenkov K, Ryberg M, Kristiansson E, Hartmann M, Schoch CL, Nylander JAA, Bergsten J, Porter TM, Jumpponen A, Vaishampayan P, Ovaskainen O, Hallenberg N, Bengtsson-Palme J, Eriksson KM, Larsson K-H, Larsson E, Kõljalg U. 2012. Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. MycoKeys 4:37–63. https://doi.org/10.3897/mycokeys.4.3606.
43. Andrew M, Peever TL, Pryor BM. 2009. An expanded multilocus phylogeny does not resolve morphological species within the small-spored *Alternaria* species complex. Mycologia 101:95–109. https://doi.org/10.3852/08-135.
44. Costa MS, Clark CM, Ómarsdóttir S, Sanchez LM, Murphy BT. 2019. Minimizing taxonomic and natural product redundancy in microbial libraries using MALDI-TOF MS and the bioinformatics pipeline IDBac. J Nat Prod 82:2167–2173. https://doi.org/10.1021/acs.jnatprod.9b00168.
45. Harvey AL. 2008. Natural products in drug discovery. Drug Discov Today 13:894–901. https://doi.org/10.1016/j.drudis.2008.07.004.
46. Breinbauer R, Vetter IR, Waldmann H. 2002. From protein domains to drug candidates—natural products as guiding principles in the design and synthesis of compound libraries. Angew Chem Int Ed 41:2878–2890. https://doi.org/10.1002/1521-3773(20020816)41:16<2878::AID-ANIE2878>3.0.CO;2-B.
47. Pan R, Bai X, Chen J, Zhang H, Wang H. 2019. Exploring structural diversity of microbe secondary metabolites using OSMAC strategy: a literature review. Front Microbiol 10:294. https://doi.org/10.3389/fmicb.2019.00294.
48. Du L, King JB, Cichewicz RH. 2014. Chlorinated polyketide obtained from a *Daldinia* sp. treated with the epigenetic modifier suberoylanilide hydroxamic acid. J Nat Prod 77:2454–2458. https://doi.org/10.1021/np500522z.
49. Khaldi N, Collemare J, Lebrun M-H, Wolfe KH. 2008. Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. Genome Biol 9:R18. https://doi.org/10.1186/gb-2008-9-1-r18.
50. Walsh CT. 2015. A chemocentric view of the natural product inventory. Nat Chem Biol 11:620–624. https://doi.org/10.1038/nchembio.1894.
51. Bode HB, Bethe B, Höfs R, Zeeck A. 2002. Big effects from small changes: possible ways to explore nature's chemical diversity. Chembiochem 3:619–627. https://doi.org/10.1002/1439-7633(20020703)3:7<619::AID-CBIC619>3.0.CO;2-9.
52. Nothias L-F, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, Protsyuk I, Ernst M, Tsugawa H, Fleischauer M, Aicheler F, Aksenov AA, Alka O, Allard P-M, Barsch A, Cachet X, Caraballo-Rodriguez AM, Da Silva RR, Dang T, Garg N, Gauglitz JM, Gurevich A, Isaac G, Jarmusch AK, Kameník Z, Kang KB, Kessler N, Koester I, Korf A, Le Gouellec A, Ludwig M, Martin H C, McCall L-I, McSayles J, Meyer SW, Mohimani H, Morsy M, Moyne O, Neumann S, Neuweger H, Nguyen NH, Nothias-Esposito M, Paolini J, Phelan VV, Pluskal T, Quinn RA, Rogers S, Shrestha B, Tripathi A, van der Hooft JJJ, et al. 2020. Feature-based molecular networking in the GNPS analysis environment. Nat Methods 17:905–908. https://doi.org/10.1038/s41592-020-0933-6.
53. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, et al. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837. https://doi.org/10.1038/nbt.3597.
54. Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, Glukhov E, Wodtke A, de Felicio R, Fenner A, Wong WR, Linington RG, Zhang L, Debonsi HM, Gerwick WH, Dorrestein PC. 2013. Molecular networking as a dereplication strategy. J Nat Prod 76:1686–1699. https://doi.org/10.1021/np400413s.
55. Ramos AEF, Evanno L, Poupon E, Champy P, Beniddir MA. 2019. Natural products targeting strategies involving molecular networking: different manners, one goal. Nat Prod Rep 36:960–980. https://doi.org/10.1039/c9np00006b.
56. Kang KB, Woo S, Ernst M, van der Hooft JJ, Nothias L-F, da Silva RR, Dorrestein PC, Sung SH, Lee M. 2020. Assessing specialized metabolite diversity of *Alnus* species by a digitized LC–MS/MS data analysis workflow. Phytochemistry 173:112292. https://doi.org/10.1016/j.phytochem.2020.112292.
57. van Der Hooft JJ, Mohimani H, Bauermeister A, Dorrestein PC, Duncan KR, Medema MH. 2020. Linking genomics and metabolomics to chart specialized metabolic diversity. Chem Soc Rev 49:3297–3314. https://doi.org/10.1039/d0cs00162g.
58. da Silva RR, Wang M, Nothias L-F, van der Hooft JJ, Caraballo-Rodríguez AM, Fox E, Balunas MJ, Klassen JL, Lopes NP, Dorrestein PC. 2018. Propagating annotations of molecular networks using in silico fragmentation. PLoS Comput Biol 14:e1006089. https://doi.org/10.1371/journal.pcbi.1006089.
59. Wandy J, Zhu Y, van der Hooft JJ, Daly R, Barrett MP, Rogers S. 2018. Ms2lda.org: web-based topic modelling for substructure discovery in mass spectrometry. Bioinformatics 34:317–318. https://doi.org/10.1093/bioinformatics/btx582.
60. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
61. King JB, Carter AC, Dai W, Lee JW, Kil Y-S, Du L, Helff SK, Cai S, Huddle BC, Cichewicz RH. 2019. Design and application of a high-throughput, high-content screening system for natural product inhibitors of the human parasite *Trichomonas vaginalis*. ACS Infect Dis 5:1456–1470. https://doi.org/10.1021/acsinfecdis.9b00156.
62. White TJ, Bruns T, Lee S, Taylor JW. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics, p 315–322. *In* Innis MA, Gelfand DH, Sninsky JJ, White TJ (ed), PCR protocols: a guide to methods and applications. Academic Press, Inc, New York, NY.

63. Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194. https://doi.org/10.1101/gr.8.3.186.
64. Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185. https://doi.org/10.1101/gr.8.3.175.
65. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35:1547–1549. https://doi.org/10.1093/molbev/msy096.
66. Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120. https://doi.org/10.1007/BF01731581.
67. Pluskal T, Castillo S, Villar-Briones A, Orešič M. 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11:395. https://doi.org/10.1186/1471-2105-11-395.
68. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336. https://doi.org/10.1038/nmeth.f.303.
69. Bray JR, Curtis JT. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol Monogr 27:325–349. https://doi.org/10.2307/1942268.
70. Vázquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. 2013. EMPeror: a tool for visualizing high-throughput microbial community data. Gigascience 2:16. https://doi.org/10.1186/2047-217X-2-16.
71. Dixon P. 2003. VEGAN, a package of R functions for community ecology. J Veg Sci 14:927–930. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x.
72. Chao A, Gotelli NJ, Hsieh T, Sander EL, Ma K, Colwell RK, Ellison AM. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol Monogr 84:45–67. https://doi.org/10.1890/13-0133.1.
73. Hsieh T, Ma K, Chao A, Hsieh MT. 2020. Package 'iNEXT.'
74. R Development Core Team. 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
75. Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R. 2015. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics 16:169. https://doi.org/10.1186/s12859-015-0611-3.