metabolites

MDPI

Article

hcapca: Automated Hierarchical Clustering and Principal Component Analysis of Large Metabolomic Datasets in R

Shaurya Chanana, Chris S. Thomas, Fan Zhang, Scott R. Rajski and Tim S. Bugni *

Pharmaceutical Sciences Division, School of Pharmacy, University of Wisconsin, Madison, WI 53705, USA;
schanana@wisc.edu (S.C.); csthomas4@wisc.edu (C.S.T.); fzhang83@wisc.edu (F.Z.);
scott.rajski@wisc.edu (S.R.R.)

* Correspondence: tim.bugni@wisc.edu; Tel.: +1-608-263-2519

Received: 19 June 2020; Accepted: 18 July 2020; Published: 21 July 2020

Abstract: Microbial natural product discovery programs face two main challenges today: rapidly prioritizing strains for discovering new molecules and avoiding the rediscovery of already known molecules. Typically, these problems have been tackled using biological assays to identify promising strains and techniques that model variance in a dataset such as PCA to highlight novel chemistry. While these tools have shown successful outcomes in the past, datasets are becoming much larger and require a new approach. Since PCA models are dependent on the members of the group being modeled, large datasets with many members make it difficult to accurately model the variance in the data. Our tool, ***hcapca***, first groups strains based on the similarity of their chemical composition, and then applies PCA to the smaller sub-groups yielding more robust PCA models. This allows for scalable chemical comparisons among hundreds of strains with thousands of molecular features. As a proof of concept, we applied our open-source tool to a dataset with 1046 LCMS profiles of marine invertebrate associated bacteria and discovered three new analogs of an established anticancer agent from one promising strain.

Keywords: metabolites; genomics; PCA; HCA; dendrogram; variance; open source; LCMS

1. Introduction

Natural product drug discovery programs continue to provide new and bio-medically relevant pharmacophores [1]. Among the potential sources of natural products, bacteria have proven to be a particularly prolific resource; for example, the genus *Streptomyces* is responsible for an unrivaled 80% of known actinomycete natural products [2]. Many natural products discovery programs rely heavily on collecting source organisms from diverse ecological niches in an attempt to harness the biological and chemical diversity stemming from these living systems. Given the time-span of a traditional natural product discovery pipeline [3–5], it is important to minimize the chemical redundancy and maximize chemical diversity in an environmental collection in order to minimize rediscovery. As such, tools to effectively survey sources of natural products prior to employing their chemistry for drug discovery are critical for effective discovery programs. Without effective tools, many downstream assay hits invariably result from similar or identical chemotypes, drastically increasing the number of resources required to discover new high-value leads. Although we have previously demonstrated that liquid chromatography mass spectroscopy (LCMS)-based metabolomics help to partly address this problem, we also found that there are fundamental limits to scaling these methods [6,7]. Specifically, there are no good tools to handle large LCMS-based untargeted metabolomics datasets that aligned with drug discovery goals. To meet this need we developed a tool called ***hcapca*** that enables rapid assessment of chemical diversity using low cost LCMS-based untargeted metabolomics.

As an alternative to LCMS techniques, Clark et al. recently demonstrated an excellent method to compare functional chemistry between closely related environmental isolates using in situ matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) based proteomics and metabolomics. They were able to discriminate between freshwater *Micromonospora* isolates that were more than 99% similar by 16S rRNA sequencing [8]. While MALDI clearly holds promise, many natural product discovery programs including ours, have chosen to utilize LCMS-based untargeted metabolomics techniques [3,9–17]. Compared to MALDI, reduced ion suppression [18], increased sensitivity, better metabolite coverage, and the ability to separate complex mixtures based on retention time all make LCMS a very attractive low-cost option for strain dereplication.

Given the appeal of LCMS, there have been a number of recent chemoinformatic advancements to better utilize its power, setting the stage for systems-level metabolomic investigations [19–32]. XCMS [26] and MZmine2 [32] are open source tools written to generate spectral tables that incorporate high-throughput peak detection and retention time correction. Tools such as GNPS [33] and MS-DIAL [24] aim to dereplicate molecular features using tandem MS data (MS2) while more recent techniques such as MASST [21] and Qemistree [34] enable one to search and classify those features based on publicly available molecular databases. However, many of these technologies, though amenable to large datasets, rely on MS2 data which provide rich sub-structure information but often focus only on the most intense ions. Although these innovative tools and techniques are extremely powerful, we believe that as a first step, ***hcapca*** can provide the necessary strain dereplication and help prioritize promising strains. Thereafter, optimization and additional scrutiny can be performed on the selected strains using MS2-based tools yielding even more chemical information.

While the tools used to identify strains with diverse chemistry are immensely important, the chemical diversity is also dependent on the environment where samples originate. New and diverse chemistry can be accessed by exploring non-traditional environmental niches such as caves and insect symbionts, making new sources for lead compounds available [35–48]. Although this increases the likelihood of finding new chemical entities, it is still difficult to identify these elusive molecules because a majority of the molecules produced are often shared, even across different genera. For example, Doroghazi et al. showed that related strains of actinobacteria—a phylum known for their prolific natural product potential [37,38]—shared 80% of their nonribosomal peptide synthetases (NRPSs) and 73% of their type II polyketide synthase (PKS) gene cluster families (GCFs) [49]. Additionally, Ziemert et al. and Jensen et al. showed that while there was a core set of metabolites within each of three species of *Salinospora*, unusual gene clusters were more random in occurrence [2,50]. In short, the microbial potential for chemical diversity does exist within organisms but is difficult to capture without specialized tools. ***hcapca*** employs principal component analysis (PCA), an unsupervised learning technique, which can highlight this hidden chemical diversity in an LCMS dataset. ***hcapca*** models variance in the data, collapses common metabolites, and highlights molecules that account for the greatest overall variance i.e., are likely to be "interesting". Thus, ***hcapca*** enables users to access the diversity even within their large datasets to discover new chemistry.

Unsupervised techniques have been used in the past on metabolomics data [51–59] including hierarchical clustering analysis (HCA) [8,60]. Indeed, even within the realm of natural products discovery, we and others have successfully used PCA to identify novel chemical scaffolds [6,51,52,60–67]. However, we believe we are the first to rigorously integrate HCA with PCA. Thus, we present ***hcapca***—a general and highly effective algorithm designed to enable untargeted strain prioritization for drug discovery from large metabolomics datasets. As a proof of concept, we analyzed 1046 LCMS extracts from marine invertebrate associated bacteria resulting in 71,000+ molecular features. Using ***hcapca***, we rapidly organized this large dataset into 90 clusters. Upon examining one of the 90 clusters, we discovered three previously unknown analogs of lomaiviticin [68,69], an anticancer compound.

2. Results

2.1. PCA

Effective prioritization of samples for natural product drug discovery requires the identification of samples with unique chemistry. Unique or interesting chemistry can be identified by finding the sources of chemical variance amongst our samples. In our tool, we use PCA to model the variance in a dataset. PCA is a powerful algorithm that reorients data along the principal axes of variance in a dataset thus enabling the identification of interesting samples and subsequently, novel molecules [70,71]. PCA is agnostic of sample metadata such as species of the strains in the dataset or biological activity of the metabolites. This enables the discovery of chemically significant outliers [6,7,36,47,72–74] and important overarching patterns [17,70,71,75] even in datasets with little to no metadata i.e., datasets composed of samples from niche (often underexplored) environments.

However, for very large datasets with more complex chemistry, PCA alone is insufficient to reveal clear trends. Ultimately, the model is dependent on the set of strains chosen for the dataset [7]. As shown in Figure 1, as we decrease the number of samples being modeled (numbers at the top right corner for subplots Figure 1a–d), trends become increasingly clearer. The amount of variance being explained by each principal component or PC (shown in parentheses on the axes of each subplot) also increases since there is less variation in the dataset. In Figure 1a, there are 1046 samples (and thus 1046 PCs) present and the first two PCs represent only 5.7% of the total variance in the dataset. For comparison, useful PCA models typically have far fewer PCs and an order of magnitude higher explained variance in PCs 1 and 2 than in the dataset being modeled [7,70,71,75].

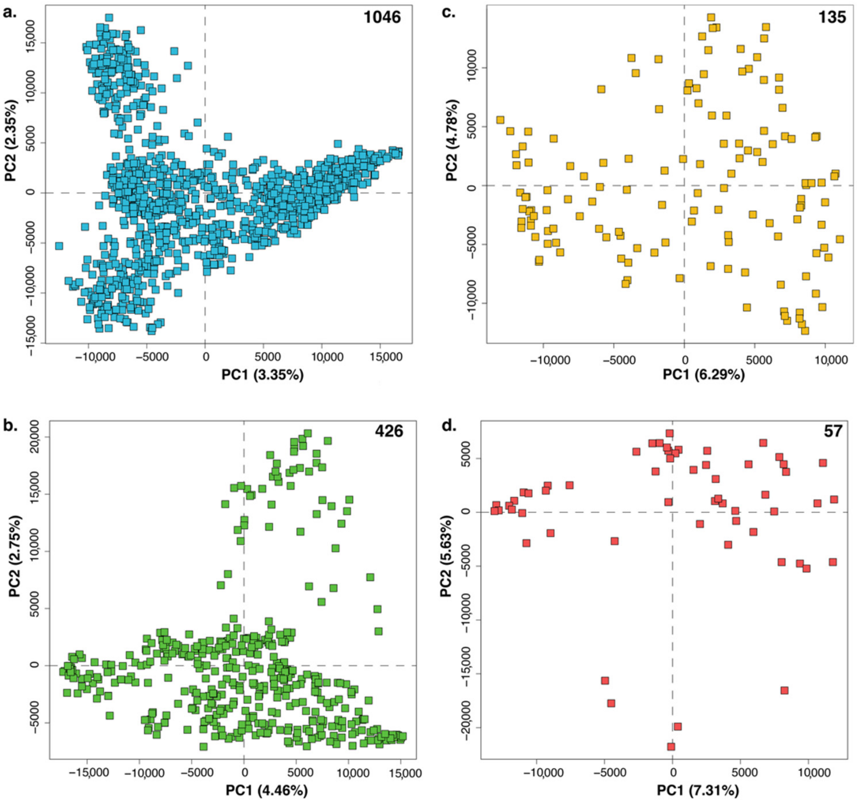

Figure 1. Subplots (**a**–**d**) show the PCA scores plot for four datasets. The number of samples in each dataset is shown in the top right corner of each plot. The total variance explained by a principal component (PC) is shown in parentheses next to the axis labels on each subplot. As the number of samples in a PCA decreases, the variance explained by each PC increases due to a combination of fewer samples and lesser overall variance in the dataset.

2.2. HCA

An overabundance of data in a single PCA model can be avoided by using HCA to split the dataset into smaller subgroups and then subjecting each of those subgroups to PCA. Analogous to clustering gene expression data [53], metabolite-based HCA assumes that samples with similar metabolic profiles are chemically related and should be grouped together. Performed alone, HCA is able to organize samples into a tree based on the similarity between microbial metabolomes. An example of such a tree, also called a dendrogram, is shown in Figure 2. The next step is to pick smaller clusters from this tree and subject each of them to PCA. However, while the human eye can pick out a few clusters or groups, the problem of choosing appropriate clusters quickly becomes intractable. Additionally, for very large datasets such as ours, dendrograms are extremely large and difficult to both visualize and interpret by human eyes alone. In fact, the dendrogram shown in Figure 2 is only a small portion (123 samples) of the tree representing all 1046 samples. Representations of the full tree can be found in Figures S1 and S3.

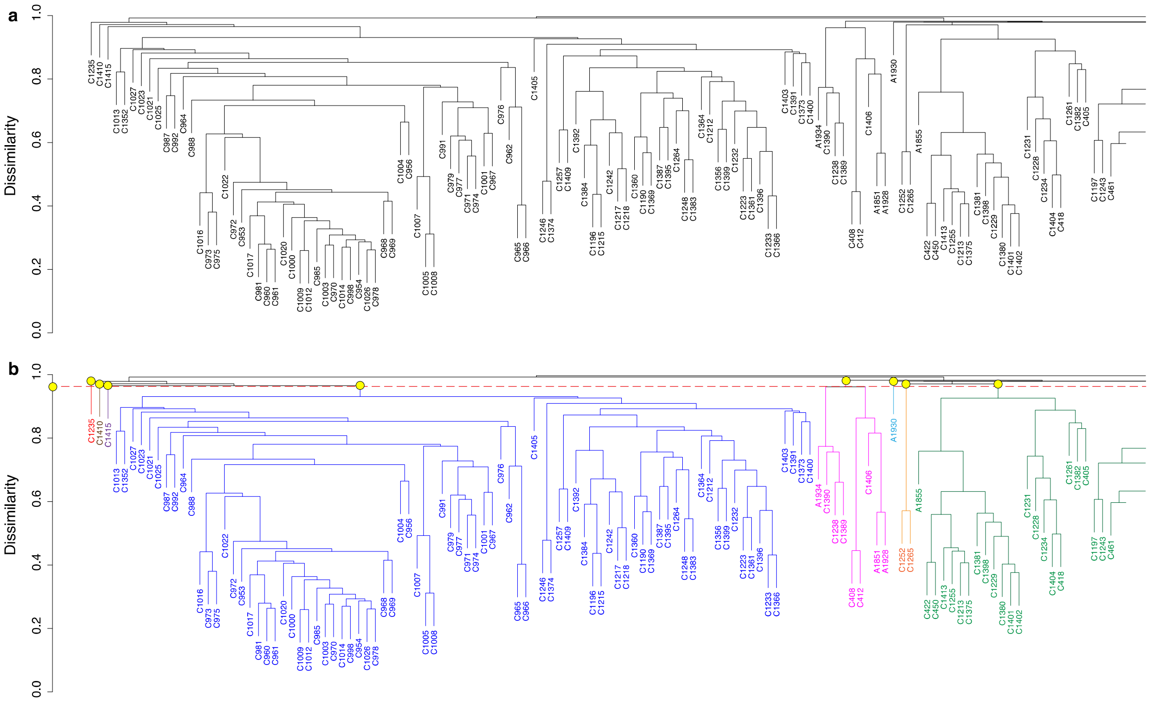

Figure 2. (**a**) Partial dendrogram generated from an HCA of all 1046 samples. The scale on the left denotes dissimilarity i.e., the closer to the bottom a pair of samples are, the more similar they are to each other. Only a small subset of the figure is shown for clarity; the original complete dendrogram may be found in the Supplementary Information as Figures S1 (linear display) and S3 (circular display). (**b**) Arbitrary dissimilarity cutoff choice of 0.95 results in eight different groups being formed. The groups have been colored accordingly. The eight groups have been colored as red, brown, grey, blue, magenta, teal, orange, and green. The yellow dots indicate the point at which the tree branch diverges to form each respective colored group of samples.

2.3. HCA and PCA in Combination

hcapca solves the problem of sub-cluster generation and readily enables visualization of the sub-trees generated. Our aim is to create a tree based on the similarity between strains and then divide that tree into smaller sub-groups. Upon generating a tree such as Figure 2a, a decision must be made on how to divide this tree into smaller sub-groups. Typically, this is done by choosing a dissimilarity value on the Y-axis and drawing a line straight across the entire tree. The branches below the intersections of the straight line are considered separate sub-groups as shown in Figure 2b. Our next step would be to model the variance in these sub-groups using PCA. However, there are two problems that we must first solve. Firstly, there exist many values of dissimilarity cutoffs that would lead to groups with only one sample. Figure 2b shows how, if we chose a dissimilarity cutoff of 0.95, we are able to draw a line straight across (red dashed line) which intersects the tree branches at various points (colored yellow). These branches and the samples within are each treated as a separate group. Notice that the "groups" colored red, brown, purple (first three samples on the left in Figure 2b) and teal contain one sample each. PCA models are not possible for single samples. Secondly, the blue colored group in Figure 2b still contains 88 samples; far too many to allow for a robust PCA model. One simple solution is to regenerate a new tree from just the blue group and again choose an arbitrary dissimilarity cutoff to form sub-groups. While this solves our problem of having too many samples in a sub-group, our dissimilarity cutoff decision is still arbitrary and dependent on the choice of samples [7]. This is the same phenomenon we have observed with PCA—the choice of samples changes the PCA (or HCA) model completely [7].

Herein lies the innovation of ***hcapca***; instead of choosing an arbitrary dissimilarity cutoff value for making sub-groups, we can use the variance explained by the PCs (from a PCA model) to decide the cutoff for us. First, as before, a distance matrix of all the samples is generated and a large tree is made (Figure 3a). Next, the tree is partitioned at the point where it first branches (point of first divergence in Figure 3a–c), and the sum of the variance is explained by the first two PCs (SoV_{12}) is simultaneously calculated from PCA models of each sub-cluster. If the SoV_{12} is smaller than the preset cutoff value (25%), the cluster is re-partitioned (this condition is denoted in red color below each cluster). If the SoV_{12} is greater than or equal to the preset cutoff value, the cluster is no longer partitioned (this condition is denoted in green color below each cluster) and a PCA model of the samples in that sub-cluster is calculated (the squares at the bottom of Figure 3c).

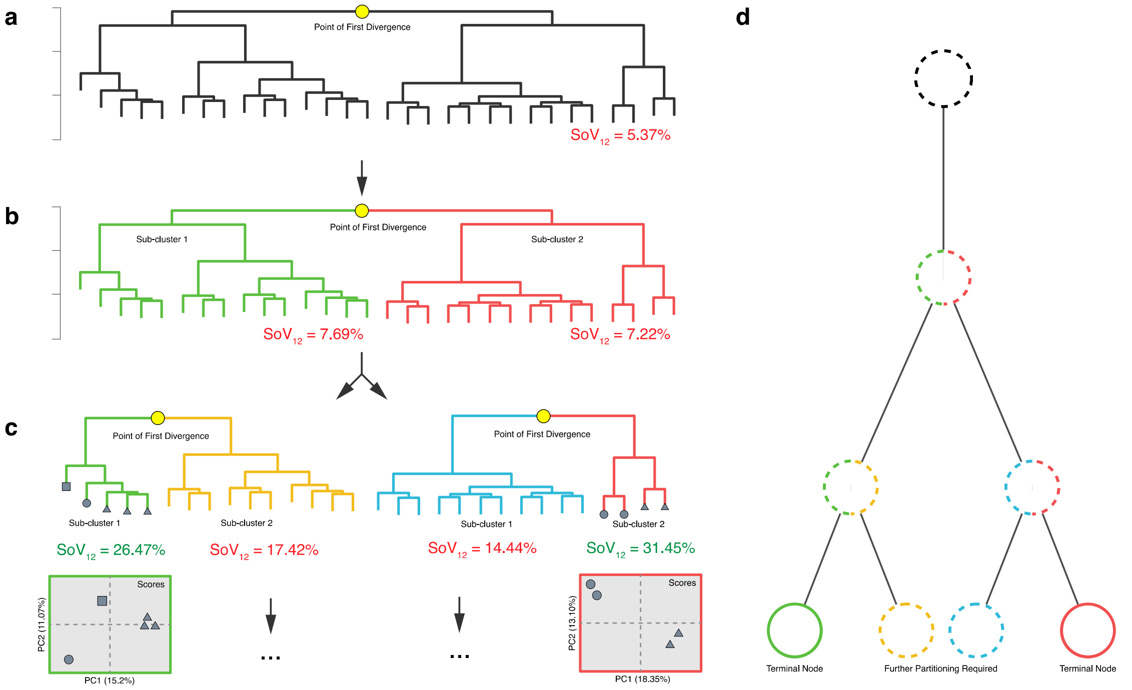

Figure 3. Scheme depicting ***hcapca*** logic. Note also that a small (35 sample) example of the walk through of ***hcapca*** processing and interactive visualization is depicted in Supplementary Information Figures S4–S13. (**a**) The first tree is partitioned into two smaller sub-clusters. (**b**) Since the SoV_{12} for the two sub-clusters does not meet the cutoff value (25%), they are further split into smaller groups (**c**). The SoV_{12} of the red and green clusters is more than the cutoff value and so their partitioning stops and PCA models are made (red/green squares). The green and blue sub-clusters have SoV_{12}s lower than

the cutoff so they are split further as indicated by the ellipsis. (**d**) The overall structure of this schema results in a "tree-of-trees". The circles represent the various nodes being formed and are colored as per the trees (from a, b, and c) that they represent. Dashed borders indicate nodes that need to be partitioned further while solid lines denote nodes that can no longer be split.

Using explained variance as a user-set cutoff for determining cluster composition and R Shiny [76–78] based interactive visualizations, ***hcapca*** thus helps to visualize the breakdown of the large tree into smaller and smaller sub-trees based on the chemical similarity within the data.

In effect, the overall process yields a large "tree-of-trees" where each node represents a smaller tree. This representation is shown in Figure 3d where the overall tree is drawn parallel to each corresponding part of Figure 3a–c, and colored based on the splitting. Nodes that can be further partitioned are indicated by dashed borders and ones that will no longer be split are indicated by solid lines. In Figure 3d, two of the nodes reached the cutoff threshold (solid red and green borders); the other two did not and consequently were partitioned further (dashed blue and yellow circles). Thus, by combining HCA with PCA and recursive partitioning of the tree, we can obtain small, chemically similar sample groupings that yield more informative PCA models.

2.4. Identification of Novel Chemistry

It is important to reiterate our hypothesis that "outliers" in metabolomic data are more likely to be novel. Using PCA to model the chemical variance in a dataset, identify promising strains and their metabolites, and thereby discover novel molecules is a credible approach backed up by both our group's own work [6,7,17,36,47,48,73,74], as well as by genomic studies [49,50] done by other groups. The addition of HCA serves to broaden the scope to allow large scale analyses and offer a more robust method of identifying promising strains. To demonstrate and utilize this algorithm for strain prioritization and drug discovery, we examined PCA models of the terminal nodes more closely and identified bacterium WMMA1901 (henceforth A1901) as a possible producer of novel chemistry (Figure 4). Figure S2 shows the location of the node in the overall tree. The HCA for this node contained eight samples (Figure 4a) and the PCA model for the node showed that the strain harbored interesting chemistry (Figure 4b). The highlighted red square on the Scores plot shows how A1901 is pulled out in the PCA. Since the scores and loadings plots are algebraically and geometrically related, the red squares on the loadings plot highlight the metabolites of interest that likely originate from strain A1901. Subsequent traditional isolation and purification by HPLC and structure elucidation by NMR revealed three new molecules (Figure 4c). Compounds **FZ4-23-1**, **FZ4-22-1**, and **FZ4-22-2**, are analogs of lomaiviticin, a class of anticancer compounds [68,69].

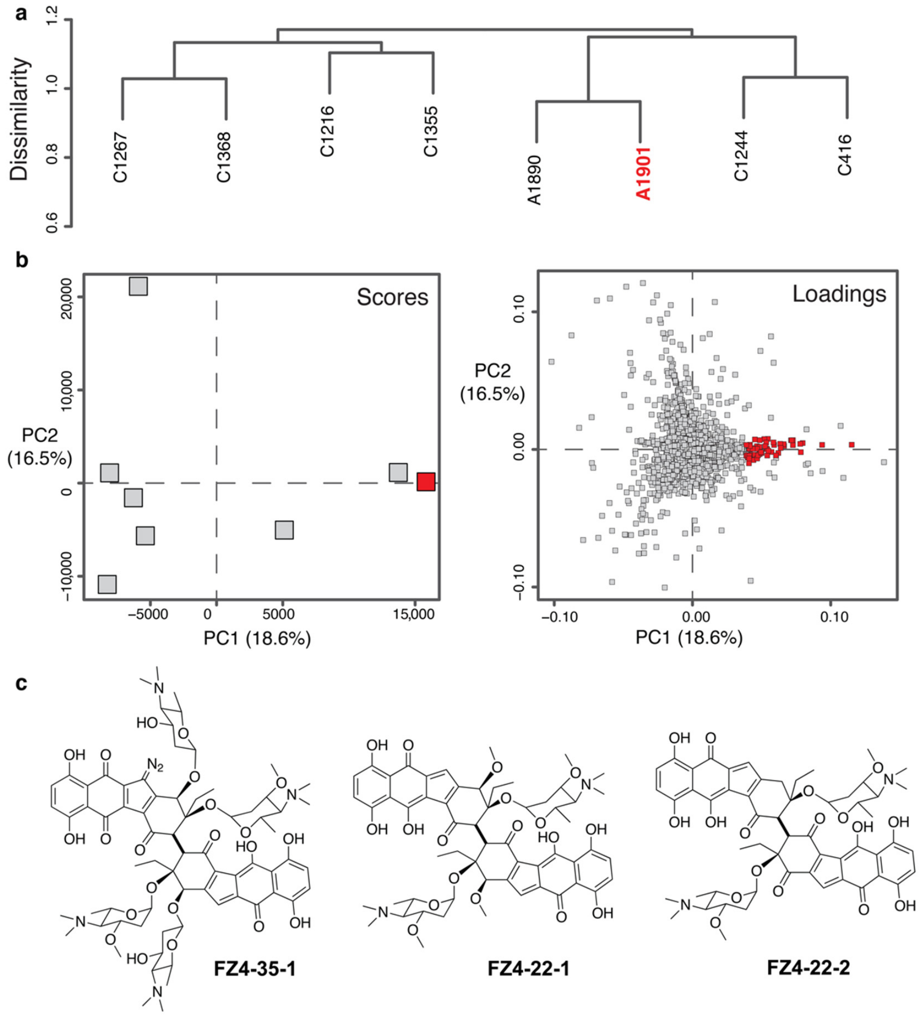

Figure 4. A1901 was identified from the PCA of node 'fj' shown in Figure S2. (**a**) The dendrogram of the node 'fj' contains eight strains in total. (**b**) PCA scores and loadings plots of the node containing A1901 with red squares highlighting the strain and its corresponding metabolites respectively are also shown. (**c**) Structures of the new lomaiviticin congeners.

To further illustrate the utility of our tool, we highlighted the position of A1901 (see Figure 5 below) in each of the four PCA plots shown previously in Figure 1. It is immediately apparent that, without ***hcapca***, neither A1901 nor its unique metabolites would be discovered so expediently. To further illustrate the utility of our tool, we highlighted the position of A1901 (see Figure 5 below) in each of the four PCA plots shown previously in Figure 1. It is immediately apparent that, without ***hcapca***, neither A1901 nor its unique metabolites would be discovered so expediently.

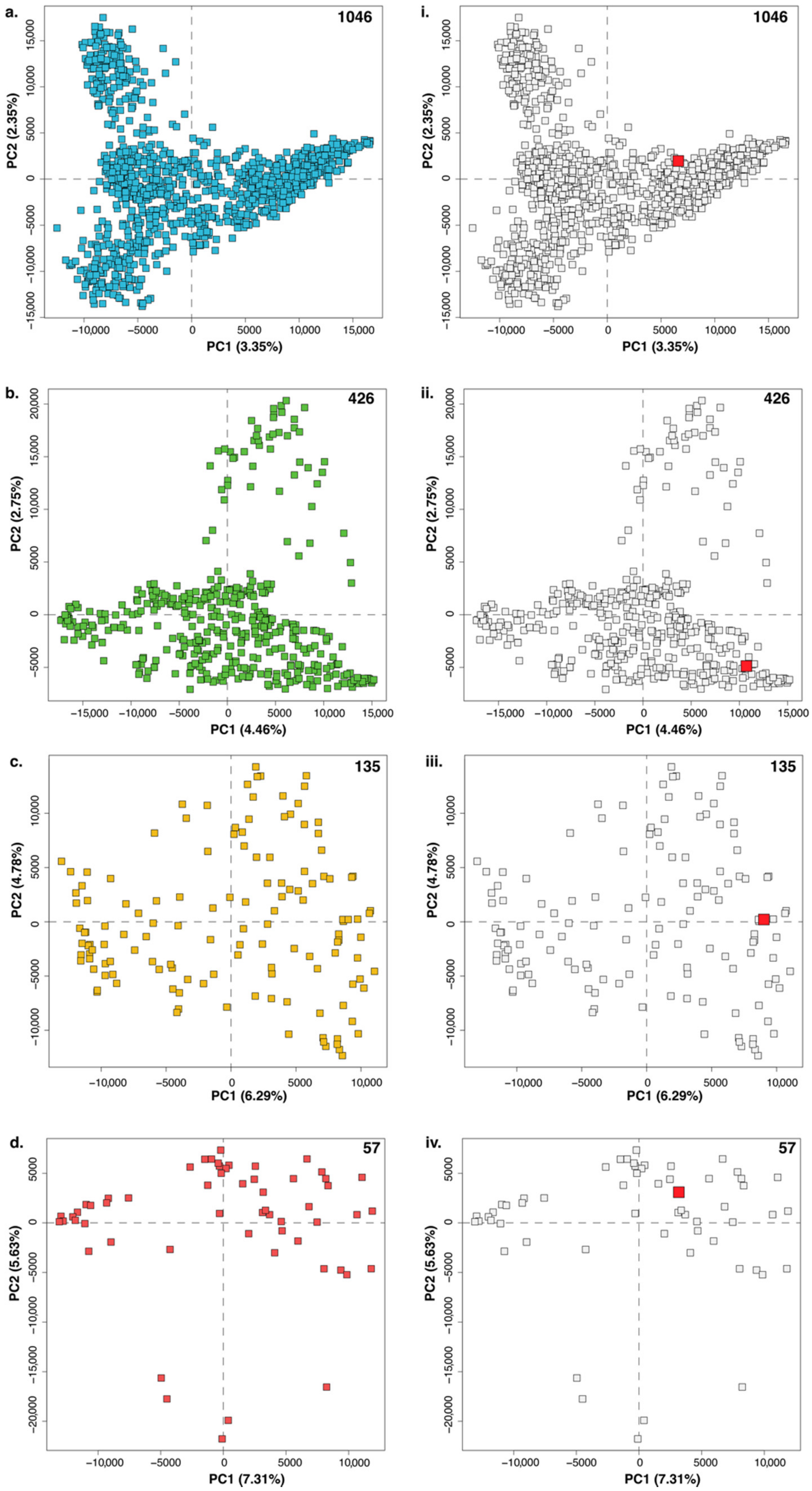

Figure 5. (**a**–**d**) represent the PCA models for the nodes mw, yq, ss, and bm from Figure S2, respectively. Sub-plots (**i**–**iv**) correspond to (**a**–**d**), respectively, highlighting the position of A1901 using a red dot while de-emphasizing other points in the plot by making them grey. Without the utilization of ***hcapca***, the discovery of the new anticancer compounds would not have been possible.

3. Materials and Methods

Previously generated LCMS data from our library were used for this analysis. The library itself was created through isolation of bacteria from sponge and ascidian specimens, cultivation in solid or liquid media, extraction using solvents, and finally a UPLC-HRMS analysis to generate LCMS profiles. Please refer to our previous publications for the details [6,7,17] and to Supplementary Information Table S1 for all media employed during fermentations.

3.1. Generation of Spectral Intensity Tables

3.1.1. Profile Analysis

Bruker Compass ProfileAnalysis 2.3 (Billerica, MA, USA). Find Molecular Features was applied to LCMS data under these parameters: S/N threshold, 5; correlation coefficient threshold, 0.7; minimum compound length, 10; smoothing width, 1. The LCMS datasets were evaluated in a time range from 2 to 14 min and in a mass range from *m/z* 150 to 1500. Advanced bucketing was employed using ΔRT = 0.33 min and Δ*m/z* = 4 ppm as parameters.

3.1.2. MZmine2

We generated mass lists (detected ions) for each scan using Mass Detection (cutoff of 1E3), detected chromatograms using Chromatogram Builder (Δ*m/z* = 4 ppm, ΔRT = 0.33 min), and separated individual peaks using the Chromatogram Deconvolution module (using ADAP module: an S/N threshold of 5, peak duration of 0 to 0.33 min, and RT wavelet range of 0 to 10 min). Isotopes were removed using the Isotopic peak grouper module and alignment was performed using both RANSAC and Join aligners. Finally, the data was exported to a CSV file and separated into 4 parts as specified in the Data format section below.

3.1.3. Data Format

The script expects data in four different files: Analyses.dat contains the sample names separated by new line characters, Variables_m.dat contains the *m/z* values separated by spaces all on one line, Variables_t.dat contain retention time values separated by spaces all on one line, and Table.dat contains the spectral intensity values separated by spaces with one line for each sample. Each row of Table.dat contains the intensity for the corresponding analysis name in the same order as the Variables_m and Variables_t values. It is important to realize that, even though the script was written to accommodate LCMS data, ***hcapca*** can be adapted to other kinds of data. For example, if a user does not have data for both *m/z* and retention time, they can fill one of the tables with zeroes and it has no bearing on the shape of the tree, placement of the nodes, etc.; i.e., the results are subject to contextual interpretation. Examples of the table formatting are available at—https://github.com/chanana/hcapca#table-format.

3.1.4. Hierarchical Clustering Analysis (HCA)

Using the spectral intensity table (Table.dat), a distance matrix was first calculated using the Euclidean distance between each sample along its vector of *m/z*-rt values.

$$d(\boldsymbol{p},\boldsymbol{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{1}$$

where, d is the distance between points $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ and $\boldsymbol{q} = (q_1, q_2, \ldots, q_n)$. Here $\boldsymbol{p}$ can be thought of a sample in $\mathbb{R}^n$ with the *m/z*-rt pair $\boldsymbol{p}_i$ representing a coordinate in each dimension.

Then, clustering was performed by employing the unweighted pair group method using the arithmetic mean (UPGMA) using correlation as a distance measure. Subsequently, the dataset was partitioned into the first two clusters that formed. This process was repeated on the resulting two clusters until the user-specified variance cutoff was met.

3.1.5. Principal Component Analysis (PCA)

Once the entire tree had been built using the procedure described above, the clusters at the ends of the trees (the terminal nodes) were subjected to PCA. The data at each terminal node was Pareto scaled [79] and mean subtracted before PCA was performed. To demonstrate visually what the entire tree looks like prior to the construction of more refined elements of the dendrogram we refer readers to Figures S1 and S3 of the Supporting Information.

3.1.6. Displaying Results

The results of ***hcapca*** data processing were displayed using our custom-made Shiny app [76], an interactive web design package written for the R programming language [78]. This app can be accessed once the code is run based on the instructions located at https://github.com/chanana/hcapca. The app runs on the local system of the user and is specific to each analysis the user performs.

3.1.7. Source Code and Instructions

https://github.com/chanana/hcapca

4. Conclusions

Unique molecules within a large dataset generally have an increased likelihood of being new or novel. We have demonstrated that ***hcapca*** is able to leverage this property by being able to differentiate the A1901 metabolome from all other samples in its subgroup leading to the discovery of the lomaiviticin analogs. Our tool is open source, written in R [78], and encompasses all steps from the HCA, partitioning of the tree, and subsequent PCA of the terminal nodes requiring only a table of LCMS spectral intensities and a cutoff variance as input. Importantly, ***hcapca*** is available as a Docker image allowing it to be run on Windows, macOS, and Linux (Ubuntu) operating systems. Both proprietary and open-source software such as MZmine2 [32] can be used for upstream processing of raw LCMS files to generate the spectral tables required as the input; with appropriately formatted data tables (please refer to the Github repository or the Data Format section of the Methods for details). The installation instructions, as well as the entire source code for ***hcapca*** is available to the public for free at https://github.com/chanana/hcapca. As exemplified here with the discovery of new secondary metabolites from A1901, ***hcapca*** represents an important technology enabling the rapid identification of unique data points from very large datasets and is virtually unlimited in its potential applications to assorted scientific fields.

Supplementary Materials: The following are available online at http://www.mdpi.com/2218-1989/10/7/297/s1, Figure S1: Full dendrogram of all 1046 samples obtained by HCA. The figure's actual size is 192 × 9" making it extremely difficult to visualize at a detailed level, Figure S2: Depiction of the tree-of-trees i.e., the processing of all samples via ***hcapca***, Figure S3: The large tree (of 1046 samples seen in Figure S1) represented as a circular dendrogram using iToL, Figure S4: Screen shot of first step in analyzing HCA and PCA results following ***hcapca*** of a dataset, Figure S5: Screen shot of tree tab (with instructions) of ***hcapca*** results, Figure S6: Example of how to view a given dendrogram for a specific node viewable by drop-down menu, Figure S7: Screen shot of window in which PCA tab showing results of overall tree can be used to view the node and its "parent" nodes, Figure S8: How to navigate the PCA tab of results from ***hcapca*** processing of the example dataset provided, Figure S9: Demonstration of how the Scores plot (from Figure S8 steps) gets drawn and can be used to identify samples with greatest variance, Figure S10: Screen shot of the Loadings plot and how metabolite samples within a specific microbial producer vary from each other, Figure S11: Depiction of how ***hcapca*** leads to the plot showing individual and cumulative variances for all principal components in a given sample, Figure S12: Screen shot of first "log off" page enabling one to exit the ***hcapca*** application, Figure S13: Screen shot of page indicating completion of exit from hcapca application, Table S1: Comprehensive listing of all media ingredients used to generate data herein for the 1046 sample set.

Author Contributions: Conceptualization, S.C. and C.S.T.; chemical isolation, purification, structure elucidation, F.Z.; software development, S.C.; data curation, S.C., C.S.T., and T.S.B.; writing—original draft preparation, S.C.; writing—review and editing, S.C., C.S.T., F.Z., S.R.R., and T.S.B.; funding acquisition, T.S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Institutes of Health (NIH), through the administration of NIGMS Grants R01GM104192 and U19TW009872 (also funded by FIC, NCCIH, NCI).

Acknowledgments: The authors would like to acknowledge Deepa Acharya for her help in evaluating the manuscript draft from a non-developer's perspective. The authors also wish to thank Jason Kwan for lending the use of his computational server for testing a portion of the script.

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

References

1. Newman, D.J.; Cragg, G.M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83*, 770–803. [CrossRef] [PubMed]
2. Jensen, P.R.; Moore, B.S.; Fenical, W. The marine actinomycete genus *Salinispora*: A model organism for secondary metabolite discovery. *Nat. Prod. Rep.* **2015**, *32*, 738–751. [CrossRef] [PubMed]
3. Shen, B. A new golden age of natural products drug discovery. *Cell* **2015**, *163*, 1297–1300. [CrossRef]
4. Harvey, A.L.; Edrada-Ebel, R.; Quinn, R.J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **2015**, *14*, 111–129. [CrossRef]
5. Koehn, F.E. High impact technologies for natural products screening. *Nat. Compd. Drugs Vol. I* **2008**, *65*, 175–210. [CrossRef]
6. 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. [CrossRef]
7. Chanana, S.; Thomas, C.; Braun, D.; Hou, Y.; Wyche, T.; Bugni, T. Natural product discovery using planes of principal component analysis in R (PoPCAR). *Metabolites* **2017**, *7*, 34. [CrossRef]
8. 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]
9. Baker, M. Metabolomics: From small molecules to big ideas. *Nat. Methods* **2011**, *8*, 117–121. [CrossRef]
10. Astarita, G.; Langridge, J. An emerging role for metabolomics in nutrition science. *Lifestyle Genom.* **2013**, *6*, 181–200. [CrossRef]
11. Gibbons, H.; O'Gorman, A.; Brennan, L. Metabolomics as a tool in nutritional research. *Curr. Opin. Lipidol.* **2015**, *26*, 30–34. [CrossRef] [PubMed]
12. Wikoff, W.R.; Anfora, A.T.; Liu, J.; Schultz, P.G.; Lesley, S.A.; Peters, E.C.; Siuzdak, G. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 3698–3703. [CrossRef]
13. Nicholson, J.K.; Holmes, E.; Kinross, J.; Burcelin, R.; Gibson, G.; Jia, W.; Pettersson, S. Host-gut microbiota metabolic interactions. *Science* **2012**, *336*, 1262–1267. [CrossRef]
14. Demain, A.L.; Fang, A. The natural functions of secondary metabolites. In *History of Modern Biotechnology I*; Fiechter, A., Ed.; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–39. ISBN 978-3-540-44964-5.
15. Newman, D.J.; Cragg, G.M. Endophytic and epiphytic microbes as "sources" of bioactive agents. *Front. Chem.* **2015**, *3*, 34. [CrossRef]
16. Newman, D.J.; Cragg, G.M. Plant endophytes and epiphytes: Burgeoning sources of known and "unknown" cytotoxic and antibiotic agents? *Planta Med.* **2020**. [CrossRef] [PubMed]
17. Ellis, G.A.; Thomas, C.S.; Chanana, S.; Adnani, N.; Szachowicz, E.; Braun, D.R.; Harper, M.K.; Wyche, T.P.; Bugni, T.S. Brackish habitat dictates cultivable *Actinobacterial* diversity from marine sponges. *PLoS ONE* **2017**, *12*, e0176968. [CrossRef]
18. Ishihama, Y. Proteomic LC–MS systems using nanoscale liquid chromatography with tandem mass spectrometry. *J. Chromatogr. A* **2005**, *1067*, 73–83. [CrossRef]
19. Thomas, T.; Moitinho-Silva, L.; Lurgi, M.; Björk, J.R.; Easson, C.; Astudillo-García, C.; Olson, J.B.; Erwin, P.M.; López-Legentil, S.; Luter, H.; et al. Diversity, structure and convergent evolution of the global sponge microbiome. *Nat. Commun.* **2016**, *7*, 11870. [CrossRef]

20. Blin, K.; Shaw, S.; Steinke, K.; Villebro, R.; Ziemert, N.; Lee, S.Y.; Medema, M.H.; Weber, T. antiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **2019**, *47*, W81–W87. [CrossRef]
21. Wang, M.; Jarmusch, A.K.; Vargas, F.; Aksenov, A.A.; Gauglitz, J.M.; Weldon, K.; Petras, D.; da Silva, R.; Quinn, R.; Melnik, A.V.; et al. Mass spectrometry searches using MASST. *Nat. Biotechnol.* **2020**, *38*, 23–26. [CrossRef]
22. Nothias, L.F.; Petras, D.; Schmid, R.; Dührkop, K.; Rainer, J.; Sarvepalli, A.; Protsyuk, I.; Ernst, M.; Tsugawa, H.; Fleischauer, M.; et al. Feature-based molecular networking in the GNPS analysis environment. *bioRxiv* **2019**. [CrossRef]
23. Röst, H.L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; et al. OpenMS: A flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **2016**, *13*, 741–748. [CrossRef] [PubMed]
24. Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **2015**, *12*, 523–526. [CrossRef] [PubMed]
25. Lai, Z.; Tsugawa, H.; Wohlgemuth, G.; Mehta, S.; Mueller, M.; Zheng, Y.; Ogiwara, A.; Meissen, J.; Showalter, M.; Takeuchi, K.; et al. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat. Methods* **2018**, *15*, 53–56. [CrossRef]
26. Smith, C.A.; Want, E.J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006**, *78*, 779–787. [CrossRef]
27. Jarmusch, A.K.; Wang, M.; Aceves, C.M.; Advani, R.S.; Aguire, S.; Aksenov, A.A.; Aleti, G.; Aron, A.T.; Bauermeister, A.; Bolleddu, S.; et al. Repository-scale co- and re-analysis of tandem mass spectrometry data. *bioRxiv* **2019**. [CrossRef]
28. van der Hooft, J.J.J.; Wandy, J.; Young, F.; Padmanabhan, S.; Gerasimidis, K.; Burgess, K.E.V.; Barrett, M.P.; Rogers, S. Unsupervised discovery and comparison of structural families across multiple samples in untargeted metabolomics. *Anal. Chem.* **2017**, *89*, 7569–7577. [CrossRef]
29. Wandy, J.; Zhu, Y.; van der Hooft, J.J.J.; Daly, R.; Barrett, M.P.; Rogers, S. Ms2lda.org: Web-based topic modelling for substructure discovery in mass spectrometry. *Bioinformatics* **2018**, *34*, 317–318. [CrossRef]
30. van der Hooft, J.J.J.; Wandy, J.; Barrett, M.P.; Burgess, K.E.V.; Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 13738–13743. [CrossRef]
31. Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D.S.; Xia, J. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.* **2018**, *46*, W486–W494. [CrossRef]
32. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinform.* **2010**, *11*, 395. [CrossRef]
33. Wang, M.; Carver, J.J.; Phelan, V.V.; Sanchez, L.M.; Garg, N.; Peng, Y.; Nguyen, D.D.; Watrous, J.; Kapono, C.A.; Luzzatto-Knaan, T.; et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **2016**, *34*, 828–837. [CrossRef]
34. Tripathi, A.; Vazquez-Baeza, Y.; Gauglitz, J.M.; Wang, M.; Duhrkop, K.; Esposito-Nothias, M.; Acharya, D.; Ernst, M.; van der Hooft, J.J.J.; Zhu, Q.; et al. Chemically-informed analyses of metabolomics mass spectrometry data with qemistree. *bioRxiv* **2020**. [CrossRef]
35. Van Arnam, E.B.; Ruzzini, A.C.; Sit, C.S.; Horn, H.; Pinto-Tomás, A.A.; Currie, C.R.; Clardy, J. Selvamicin, an atypical antifungal polyene from two alternative genomic contexts. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 12940–12945. [CrossRef]
36. Wyche, T.P.; Piotrowski, J.S.; Hou, Y.; Braun, D.; Deshpande, R.; McIlwain, S.; Ong, I.M.; Myers, C.L.; Guzei, I.A.; Westler, W.M.; et al. Forazoline A: Marine-derived polyketide with antifungal in vivo efficacy. *Angew. Chem. Int. Ed.* **2014**, *53*, 11583–11586. [CrossRef]
37. Abdelmohsen, U.R.; Bayer, K.; Hentschel, U. Diversity, abundance and natural products of marine sponge-associated actinomycetes. *Nat. Prod. Rep.* **2014**, *31*, 381–399. [CrossRef]
38. Abdelmohsen, U.R.; Yang, C.; Horn, H.; Hajjar, D.; Ravasi, T.; Hentschel, U. Actinomycetes from red sea sponges: Sources for chemical and phylogenetic diversity. *Mar. Drugs* **2014**, *12*, 2771–2789. [CrossRef]

39. Yang, Q.; Franco, C.M.M.; Zhang, W. Sponge-associated actinobacterial diversity: Validation of the methods of actinobacterial DNA extraction and optimization of 16S rRNA gene amplification. *Appl. Microbiol. Biotechnol.* **2015**, *99*, 8731–8740. [CrossRef]
40. Edlund, A.; Loesgen, S.; Fenical, W.; Jensen, P.R. Geographic distribution of secondary metabolite genes in the Marine Actinomycete *Salinispora arenicola*. *Appl. Environ. Microbiol.* **2011**, *77*, 5916–5925. [CrossRef]
41. Nam, S.-J.; Kauffman, C.A.; Jensen, P.R.; Moore, C.E.; Rheingold, A.L.; Fenical, W. Actinobenzoquinoline and Actinophenanthrolines A-C, unprecedented alkaloids from a Marine Actinobacterium. *Org. Lett.* **2015**, *17*, 3240–3243. [CrossRef]
42. Leutou, A.S.; Yang, I.; Kang, H.; Seo, E.K.; Nam, S.-J.; Fenical, W. Nocarimidazoles A and B from a marine-derived Actinomycete of the genus *Nocardiopsis*. *J. Nat. Prod.* **2015**, *78*, 2846–2849. [CrossRef] [PubMed]
43. Shaaban, K.A.; Saunders, M.A.; Zhang, Y.; Tran, T.; Elshahawi, S.I.; Ponomareva, L.V.; Wang, X.; Zhang, J.; Copley, G.C.; Sunkara, M.; et al. Spoxazomicin D and Oxachelin C, potent Neuroprotective Carboxamides from the Appalachian coal fire-associated isolate *Streptomyces* sp. RM-14-6. *J. Nat. Prod.* **2017**, *80*, 2–11. [CrossRef]
44. Wang, X.; Zhang, Y.; Ponomareva, L.V.; Qiu, Q.; Woodcock, R.; Elshahawi, S.I.; Chen, X.; Zhou, Z.; Hatcher, B.E.; Hower, J.C.; et al. Mccrearamycins A–D, Geldanamycin-derived Cyclopentenone Macrolactams from an Eastern Kentucky abandoned coal mine microbe. *Angew. Chem. Int. Ed.* **2017**, *56*, 2994–2998. [CrossRef] [PubMed]
45. Wang, X.; Elshahawi, S.I.; Cai, W.; Zhang, Y.; Ponomareva, L.V.; Chen, X.; Copley, G.C.; Hower, J.C.; Zhan, C.-G.; Parkin, S.; et al. Bi- and tetracyclic Spirotetronates from the coal mine fire isolate *Streptomyces* sp. LC-6-2. *J. Nat. Prod.* **2017**, *80*, 1141–1149. [CrossRef]
46. Derewacz, D.K.; McNees, C.R.; Scalmani, G.; Covington, C.L.; Shanmugam, G.; Marnett, L.J.; Polavarapu, P.L.; Bachmann, B.O. Structure and stereochemical determination of hypogeamicins from a cave-derived actinomycete. *J. Nat. Prod.* **2014**, *77*, 1759–1763. [CrossRef] [PubMed]
47. Beemelmanns, C.; Ramadhar, T.R.; Kim, K.H.; Klassen, J.L.; Cao, S.; Wyche, T.P.; Hou, Y.; Poulsen, M.; Bugni, T.S.; Currie, C.R.; et al. Macrotermycins A-D, glycosylated macrolactams from a termite-associated *Amycolatopsis* sp. M39. *Org. Lett.* **2017**, *19*, 1000–1003. [CrossRef]
48. Wyche, T.P.; Ruzzini, A.C.; Beemelmanns, C.; Kim, K.H.; Klassen, J.L.; Cao, S.; Poulsen, M.; Bugni, T.S.; Currie, C.R.; Clardy, J. Linear peptides are the major products of a biosynthetic pathway that encodes for cyclic depsipeptides. *Org. Lett.* **2017**, *19*, 1772–1775. [CrossRef]
49. Doroghazi, J.R.; Albright, J.C.; Goering, A.W.; Ju, K.; Haines, R.R.; Tchalukov, K.A.; Labeda, D.P.; Kelleher, N.L.; Metcalf, W.W. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **2014**, *10*, 963–968. [CrossRef]
50. Ziemert, N.; Lechner, A.; Wietz, M.; Millan-Aguinaga, 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–E1139. [CrossRef] [PubMed]
51. Goodwin, C.R.; Sherrod, S.D.; Marasco, C.C.; Bachmann, B.O.; Schramm-Sapyta, N.; Wikswo, J.P.; McLean, J.A. Phenotypic mapping of metabolic profiles using self-organizing maps of high-dimensional mass spectrometry data. *Anal. Chem.* **2014**, *86*, 6563–6571. [CrossRef]
52. Goodwin, C.R.; Covington, B.C.; Derewacz, D.K.; McNees, C.R.; Wikswo, J.P.; McLean, J.A.; Bachmann, B.O. Structuring microbial metabolic responses to multiplexed stimuli via self-organizing metabolomics Maps. *Chem. Biol.* **2015**, *22*, 661–670. [CrossRef] [PubMed]
53. Altman, N.; Krzywinski, M. Points of significance: Clustering. *Nat. Methods* **2017**, *14*, 545–546. [CrossRef]
54. Frank, A.M.; Bandeira, N.; Shen, Z.; Tanner, S.; Briggs, S.P.; Smith, R.D.; Pevzner, P.A. Clustering millions of tandem mass spectra. *J. Proteome Res.* **2008**, *7*, 113–122. [CrossRef]
55. Meinicke, P.; Lingner, T.; Kaever, A.; Feussner, K.; Göbel, C.; Feussner, I.; Karlovsky, P.; Morgenstern, B. Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. *Algorithms Mol. Biol.* **2008**, *3*, 9. [CrossRef] [PubMed]
56. Broeckling, C.D.; Afsar, F.A.; Neumann, S.; Ben-Hur, A.; Prenni, J.E. RAMClust: A novel feature clustering method enables spectral-matching-based annotation for metabolomics data. *Anal. Chem.* **2014**, *86*, 6812–6817. [CrossRef]

57. Damian, D.; Orešič, M.; Verheij, E.; Meulman, J.; Friedman, J.; Adourian, A.; Morel, N.; Smilde, A.; van der Greef, J. Applications of a new subspace clustering algorithm (COSA) in medical systems biology. *Metabolomics* **2007**, *3*, 69–77. [CrossRef]
58. Li, X.; Lu, X.; Tian, J.; Gao, P.; Kong, H.; Xu, G. Application of fuzzy c-means clustering in data analysis of metabolomics. *Anal. Chem.* **2009**, *81*, 4468–4475. [CrossRef]
59. Depke, T.; Franke, R.; Brönstrup, M. Clustering of MS2 spectra using unsupervised methods to aid the identification of secondary metabolites from *Pseudomonas aeruginosa*. *J. Chromatogr. B* **2017**, *1071*, 19–28. [CrossRef]
60. 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]
61. Krug, D.; Zurek, G.; Revermann, O.; Vos, M.; Velicer, G.J.; Müller, R. Discovering the hidden secondary metabolome of *Myxococcus xanthus*: A study of intraspecific diversity. *Appl. Environ. Microbiol.* **2008**, *74*, 3058–3068. [CrossRef] [PubMed]
62. Krug, D.; Zurek, G.; Schneider, B.; Garcia, R.; Müller, R. Efficient mining of myxobacterial metabolite profiles enabled by liquid chromatography-electrospray ionisation-time-of-flight mass spectrometry and compound-based principal component analysis. *Anal. Chim. Acta* **2008**, *624*, 97–106. [CrossRef]
63. Robertson, V.; Haltli, B.; McCauley, E.; Overy, D.; Kerr, R. Highly variable bacterial communities associated with the Octocoral *Antillogorgia elisabethae*. *Microorganisms* **2016**, *4*, 23. [CrossRef]
64. Forner, D.; Berrué, F.; Correa, H.; Duncan, K.; Kerr, R.G. Chemical dereplication of marine actinomycetes by liquid chromatography-high resolution mass spectrometry profiling and statistical analysis. *Anal. Chim. Acta* **2013**, *805*, 70–79. [CrossRef] [PubMed]
65. Covington, B.C.; McLean, J.A.; Bachmann, B.O. Comparative mass spectrometry-based metabolomics strategies for the investigation of microbial secondary metabolites. *Nat. Prod. Rep.* **2017**, *34*, 6–24. [CrossRef] [PubMed]
66. Derewacz, D.K.; Covington, B.C.; McLean, J.A.; Bachmann, B.O. Mapping microbial response metabolomes for induced natural product discovery. *ACS Chem. Biol.* **2015**, *10*, 1998–2006. [CrossRef] [PubMed]
67. Betancur, L.A.; Naranjo-Gaybor, S.J.; Vinchira-Villarraga, D.M.; Moreno-Sarmiento, N.C.; Maldonado, L.A.; Suarez-Moreno, Z.R.; Acosta-González, A.; Padilla-Gonzalez, G.F.; Puyana, M.; Castellanos, L.; et al. Marine *Actinobacteria* as a source of compounds for phytopathogen control: An integrative metabolic-profiling/bioactivity and taxonomical approach. *PLoS ONE* **2017**, *12*, e0170148. [CrossRef] [PubMed]
68. He, H.; Ding, W.-D.; Bernan, V.S.; Richardson, A.D.; Ireland, C.M.; Greenstein, M.; Ellestad, G.A.; Carter, G.T. Lomaiviticins A and B, potent antitumor antibiotics from *Micromonospora lomaivitiensis*. *J. Am. Chem. Soc.* **2001**, *123*, 5362–5363. [CrossRef]
69. Woo, C.M.; Beizer, N.E.; Janso, J.E.; Herzon, S.B. Isolation of Lomaiviticins C–E, transformation of Lomaiviticin C to Lomaiviticin A, complete structure elucidation of Lomaiviticin A, and structure–activity analyses. *J. Am. Chem. Soc.* **2012**, *134*, 15285–15288. [CrossRef]
70. Lever, J.; Krzywinski, M.; Altman, N. Points of significance: Principal component analysis. *Nat. Methods* **2017**, *14*, 641–642. [CrossRef]
71. Ringnér, M. What is principal component analysis? *Nat. Biotechnol.* **2008**, *26*, 303–304. [CrossRef]
72. Macintyre, L.; Zhang, T.; Viegelmann, C.; Martinez, I.J.; Cheng, C.; Dowdells, C.; Abdelmohsen, U.R.; Gernert, C.; Hentschel, U.; Edrada-Ebel, R. Metabolomic tools for secondary metabolite discovery from Marine Microbial Symbionts. *Mar. Drugs* **2014**, *12*, 3416–3448. [CrossRef] [PubMed]
73. Carr, G.; Poulsen, M.; Klassen, J.L.; Hou, Y.; Wyche, T.P.; Bugni, T.S.; Currie, C.R.; Clardy, J. Microtermolides A and B from termite-associated *Streptomyces* sp. and structural revision of vinylamycin. *Org. Lett.* **2012**, *14*, 2822–2825. [CrossRef] [PubMed]
74. Hou, Y.; Tianero, M.D.B.; Kwan, J.C.; Wyche, T.P.; Michel, C.R.; Ellis, G.A.; Vazquez-Rivera, E.; Braun, D.R.; Rose, W.E.; Schmidt, E.W.; et al. Structure and biosynthesis of the antibiotic bottromycin D. *Org. Lett.* **2012**, *14*, 5050–5053. [CrossRef] [PubMed]
75. Reich, D.; Price, A.L.; Patterson, N. Principal component analysis of genetic data. *Nat. Genet.* **2008**, *40*, 491–492. [CrossRef]

76. Chang, W.; Cheng, J.; Allaire, J.J.; Xie, Y.; McPherson, J. Shiny: Web Application Framework for R. 2020. Available online: https://cran.r-project.org/web/packages/shiny/index.html (accessed on 18 June 2020).
77. Chang, W.; Ribeiro, B.B. Shinydashboard: Create Dashboards with "Shiny". 2018. Available online: https://cran.r-project.org/web/packages/shinydashboard/index.html (accessed on 18 June 2020).
78. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.
79. van den Berg, R.A.; Hoefsloot, H.C.J.; Westerhuis, J.A.; Smilde, A.K.; van der Werf, M.J. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genom.* **2006**, *7*, 142. [CrossRef]