Meta-mass shift chemical profiling of metabolomes from coral reefs

Aaron C. Hartmann[a,b,1], Daniel Petras[c], Robert A. Quinn[c], Ivan Protsyuk[d], Frederick I. Archer[e], Emma Ransome[b,f], Gareth J. Williams[g], Barbara A. Bailey[h], Mark J. A. Vermeij[i,j], Theodore Alexandrov[c,d], Pieter C. Dorrestein[c], and Forest L. Rohwer[a]

[a]Department of Biology, San Diego State University, San Diego, CA 92182; [b]Invertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560; [c]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093; [d]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany; [e]Southwest Fisheries Science Center, National Oceanic and Atmospheric Administration, La Jolla, CA 92037; [f]Department of Life Sciences, Imperial College London, Ascot SL5 7PY, United Kingdom; [g]School of Ocean Sciences, Bangor University, Anglesey LL59 5AB, United Kingdom; [h]Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182; [i]CARMABI Foundation, Willemstad, Curaçao; and [j]Aquatic Microbiology, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, 1098 XH Amsterdam, The Netherlands

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved September 12, 2017 (received for review June 9, 2017)

Untargeted metabolomics of environmental samples routinely detects thousands of small molecules, the vast majority of which cannot be identified. Meta-mass shift chemical (MeMSChem) profiling was developed to identify mass differences between related molecules using molecular networks. This approach illuminates metabolome-wide relationships between molecules and the putative chemical groups that differentiate them (e.g., H_2, CH_2, $COCH_2$). MeMSChem profiling was used to analyze a publicly available metabolomic dataset of coral, algal, and fungal mat holobionts (i.e., the host and its associated microbes and viruses) sampled from some of Earth's most remote and pristine coral reefs. Each type of holobiont had distinct mass shift profiles, even when the analysis was restricted to molecules found in all samples. This result suggests that holobionts modify the same molecules in different ways and offers insights into the generation of molecular diversity. Three genera of stony corals had distinct patterns of molecular relatedness despite their high degree of taxonomic relatedness. MeMSChem profiles also partially differentiated between individuals, suggesting that every coral reef holobiont is a potential source of novel chemical diversity.

untargeted metabolomics | molecular networking | small molecules | coral reefs

Untargeted tandem mass spectrometry is a powerful tool for wide-scale analysis of small molecules. The resulting metabolomes are potential treasure troves of previously unidentified molecules and chemistries, but a major problem in realizing this potential is that most detected molecules cannot be identified (1–5). One possible solution is to use spectral fragmentation similarity to identify relatives of known molecules to generate annotations (6–8). These approaches have rapidly expanded reference databases, but remain inherently limited by the number of known molecules. Therefore, there is a need for analyses that do not rely upon molecular reference libraries (9).

The online platform Global Natural Products Social Molecular Networking [GNPS (5)] uses spectral fragmentation patterns to compare tens of thousands of molecular features and create networks of structurally similar molecules. Here we expand the analysis of GNPS networks to identify chemical differences between related molecules (Fig. 1). This approach is called meta-mass shift chemical (MeMSChem) profiling, and uses the mass differences (or mass shifts) between related molecules to identify and annotate known chemical groups such as H_2, CH_2, $COCH_2$, and so forth. Annotating molecules based on their mass shifts facilitates correlations between metabolomics, biochemistry, and genomics, which could help pinpoint sites of molecular modifications resulting from known and unknown enzymatic activities.

Coral reefs are noted sources of commercially useful compounds (10). Reef holobionts [e.g., corals, sponges, and algae with their associated viral and microbial communities (11)] have distinct metabolomes, with a high degree of within-holobiont similarity (12, 13). The positive relationship between taxonomic and molecular diversity is evident at the ecosystem level, but mechanisms explaining how high molecular diversity is generated remain missing. To address this question, MeMSChem profiling was applied to an existing dataset (12) composed of seven coral reef holobiont types collected in the Line Islands, which are some of the most remote and pristine coral reefs in the world (14, 15). MeMSChem profiling showed that molecular mass shift patterns differ significantly between holobionts, offering insights into why high molecular diversity is found on coral reefs.

Results

Identifying Redundant Mass Shifts in Metabolomes of Coral Reef Holobionts. The dataset used as the basis for creating MeMSChem profiles was previously published in ref. 12 and can be found on the Mass spectrometry Interactive Virtual Environment (MassIVE) at https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp with accession no. MSV000078598. This dataset was derived from an LC-MS/MS analysis of three genera of scleractinian coral (*Montipora* spp., *Pocillopora* spp., and *Porites* spp.), two coralline algae [crustose

Significance

Coral reef taxa produce a diverse array of molecules, some of which are important pharmaceuticals. To better understand how molecular diversity is generated on coral reefs, tandem mass spectrometry datasets of coral metabolomes were analyzed using a novel approach called meta-mass shift chemical (MeMSChem) profiling. MeMSChem profiling uses the mass differences between molecules in molecular networks to determine how molecules are related. Interestingly, the same molecules gain and lose chemical groups in different ways depending on the taxa it came from, offering a partial explanation for high molecular diversity on coral reefs.

Author contributions: A.C.H. and F.L.R. designed research; A.C.H., D.P., R.A.Q., M.J.A.V., and F.L.R. performed research; I.P., F.I.A., G.J.W., B.A.B., T.A., and P.C.D. contributed new reagents/analytic tools; A.C.H., D.P., R.A.Q., I.P., F.I.A., E.R., G.J.W., and B.A.B. analyzed data; and A.C.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The molecular spectra used here are available on the Mass Spectrometry Interactive Virtual Environment (MassIVE) data repository (accession no. MSV000078598).

[1]To whom correspondence should be addressed. Email: aaron.hartmann@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1710248114/-/DCSupplemental.

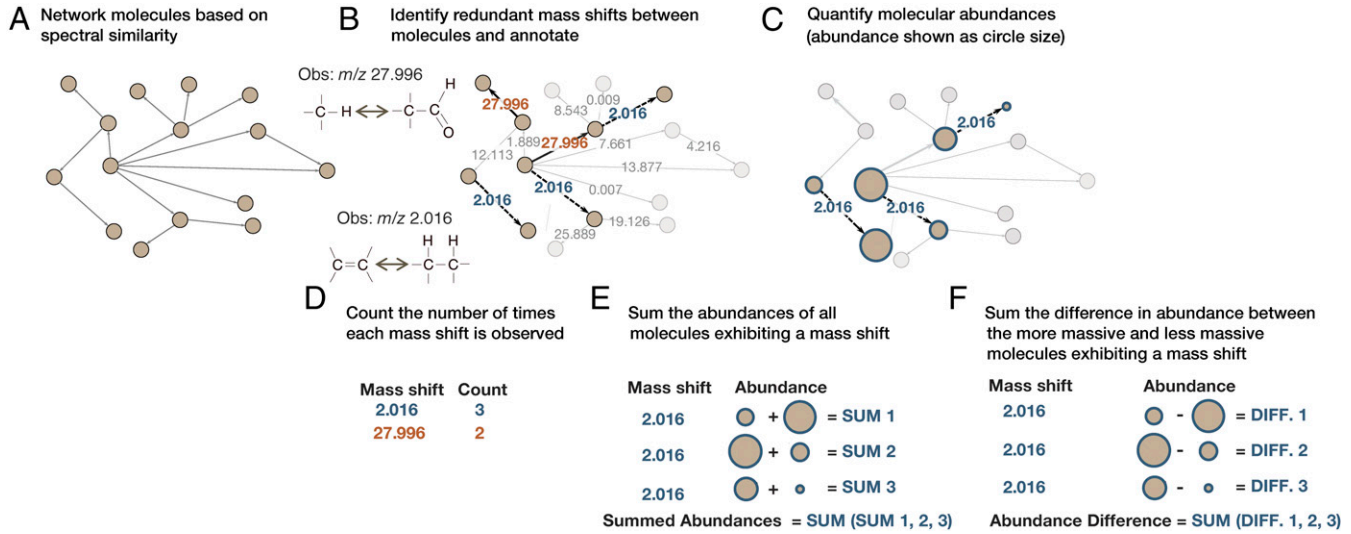

Fig. 1. Data processing and generation based on a simplified molecular network and two redundant mass shifts. (*A*) GNPS used MS/MS fragmentation spectra to elucidate molecular similarities and network similar molecules (i.e., related molecules). (*B*) Redundant mass shifts between related molecules were identified and annotated to known chemical groups when possible. Two annotated mass shifts are shown here, *m/z* 2.016 in blue with dashed lines and *m/z* 27.996 in orange with solid lines. (*C*) Molecular features that differed by a redundant mass shift were quantified based on MS. (*D–F*) Data were generated for (*D*) the number of times each redundant mass shift was observed across all networks, (*E*) the summed abundances of all molecules exhibiting each redundant mass shift, and (*F*) the sum of the differences in abundances between the more massive and less massive molecules for all pairs of molecules connected by a mass shift.

coralline algae (CCA) and *Halimeda* sp.], two noncalcifying algae (macroalgae and turf algae), and a fungal mat.

The online platform GNPS [gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp (5); Fig. S1] was used to cluster identical MS/MS spectra into nodes and identify the degree to which each node was structurally similar (i.e., related) to other nodes (henceforth referred to as "molecular features") based on a cosine score of spectral similarity. All pairs of molecular features receiving a cosine score above a threshold of 0.6 were considered to be related and connected in the network (see *SI Materials and Methods* for more details regarding the cosine score threshold). Each mass shift between network connections was then mined for multiple (i.e., redundant) occurrences (Fig. 1*B*). When the mass shifts of five or more molecular pairs differed by $<m/z$ 0.001, the mass shift was counted. All molecular features comprising the pairs with this mass shift were assigned to a bin (Fig. 1 *C–E*; see *SI Materials and Methods* for more details).

MeMSChem profiling identified 62 mass shifts that passed the filter of five or more mass shifts within *m/z* 0.001 (Table S1). Among these mass shifts, 10 were annotated to known adducts and artifacts and were removed before further analyses (Tables S1 and S2). The remaining mass shifts were annotated to known chemical groups involving hydrogen, carbon, and oxygen where possible, leading to the annotation of 13 of the 62 mass shifts identified (Table 1 and Table S1). This represents a conservative list of annotations, and the additional mass shifts identified here may be annotatable in future investigations.

Mass shifts of 0 were abundant in the networks and may represent isomers. These mass shifts were removed due to the likelihood that two isomers were merged into a single molecular feature or that the same molecular feature was split into two molecules during networking, due to the high degree of spectral similarity or difference in the number of observable fragments, respectively. An approach using retention time differences or

Table 1. All mass shifts for which the mass difference between network pairs was within the error of known chemical groups

Obs. mass	Calc. mass	% mass shifts	Putative element or group
2.016	2.016	14.48	H_2
3.955	3.995	0.62	$CH_2 \leftrightarrow H_2O$
12.000	12.000	1.95	C
14.016	14.016	11.09	CH_2
26.016	26.016	4.21	C_2H_2
28.032	28.031	8.73	C_2H_4
56.064	56.063	5.65	C_4H_8
15.995	15.995	1.23	O
18.010	18.011	2.36	H_2O
27.996	27.996	1.64	CO
42.009	42.010	0.62	$COCH_2$
56.025	56.026	0.62	C_3H_4O
58.006	58.005	0.92	CO_2CH_2

Mass shifts are shown separately based upon whether they putatively involve oxygen (blue) or only carbon and hydrogen (red). Reported are the mass shifts observed in the real data (Obs. mass), the calculated mass shift of the known mass shift (Calc. mass), the percentage of all mass shifts representing that mass shift (% mass shift), and the putative element or group composition.

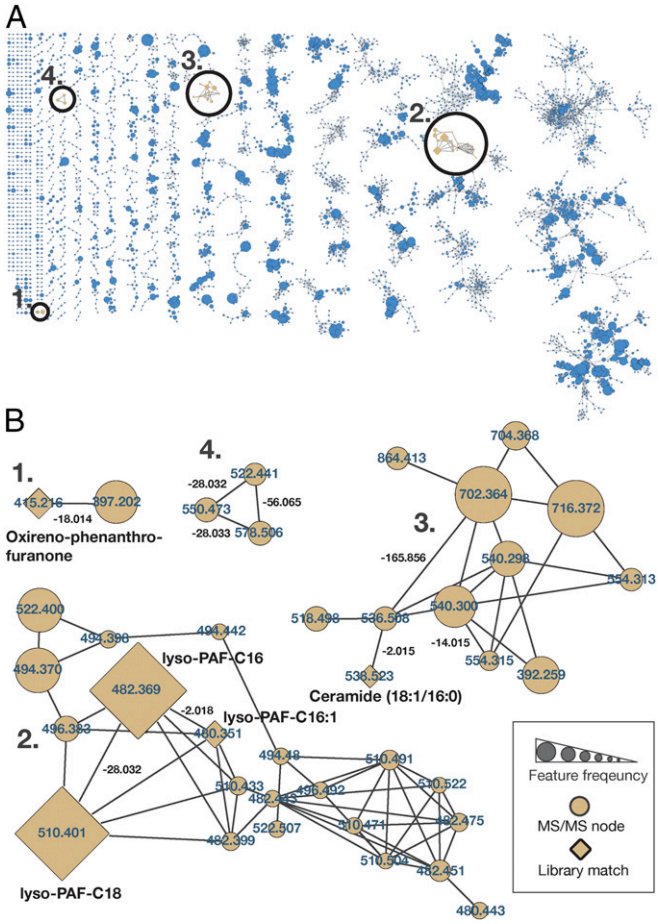

Fig. 2. Molecular network of the reef holobiont MS/MS dataset. (*A*) The global molecular networks of MS/MS spectra from the coral reef holobiont metabolomic dataset. Each node represents a unique or set of identical spectra (i.e., molecular feature). Lines connecting the nodes represent their spectral similarity, creating subnetworks that can be considered molecular families. Circles indicate zoomed-in regions of selected subnetworks shown in *B*. (*B*) Node labels represent parent masses, and line labels between the connected nodes represent the mass shift between related molecular features. Nodes with diamond shapes had a spectrum library match (e.g., lyso-PAF, as identified by ref. 12) and are further labeled with the molecular names. The size of the nodes indicates the sample frequency in which the spectra were found.

chiral separation columns should be employed to separate isomers in future applications of MeMSChem profiling.

Quantifying Mass Shifts in Holobionts. MS/MS-based molecular features associated with redundant mass shifts were quantified from the MS scan of the parent molecule using Optimus software (https://github.com/MolecularCartography/Optimus; Fig. 1*C*). A molecular feature filter was applied to remove features that were not detected in all samples. Consequently, only the features present in all samples were quantified. This filter allowed us to determine whether holobionts exhibited different mass shifts associated with the same molecules (cf. different mass shifts associated with molecules that are only found in that holobiont; Fig. 1 *E* and *F*).

Three forms of metabolome-wide data were generated for each sample (Fig. 1 *A*–*C*). First, all instances where a redundant mass shift was observed in the network were tabulated for each sample. These "counts" data provided a metric of the commonness and rarity of each mass shift in each sample (Fig. 1*D*). Second, the abundance of every molecular feature was summed by mass shift regardless of whether that feature was the higher or lower mass feature in a network pair. These "summed abundances" data provided a metric for the overall occurrence of each mass shift throughout each sample (Fig. 1*E*; see *SI Materials and Methods* for equations). Third, for each network pair, the difference in abundance between the more and less massive feature was calculated, and then these values were summed by mass shift for each sample (Fig. 1*F*; see *SI Materials and Methods* for equations). These "differences in abundances" data reflected whether, metabolome-wide, molecules were more likely to gain or lose a given mass, potentially reflecting active interconversion or branching of largely shared biosynthetic pathways. All resultant data are provided in Dataset S1. Among the redundant mass shifts, 7 of the 10 most common mass shifts were putatively annotated to known chemical groups, constituting nearly 50% of the network pairs isolated from the networks. These mass shifts included *m/z* 2.016, 14.016, 28.032, 56.064, 26.016, 18.010, and 12.000, which were putatively annotated as H_2, CH_2, C_2H_4, C_4H_8, C_2H_2, H_2O, and C, respectively.

Examining Known Mass Shifts Associated with Library-Identified Molecular Features. Instances in which known mass shifts were associated with identified molecules provided conformational evidence that mass shifts were correctly annotated. Four examples are highlighted in Fig. 2*B*, as follows. (*i*) A feature identified as phenanthro-furanone with a mass shift of *m/z* 18.014 (H_2O; Fig. 2*B*, example 1 and Fig. S2). (*ii*) A subnetwork with three forms of lyso-platelet activating factor (lyso-PAF) and related compounds (Fig. 2*B*, example 2 and Fig. S3). The identification of one molecular feature, lyso-PAF-C16, in these samples was previously confirmed using a reference standard by ref. 12. This subnetwork is particularly informative, because the three identified compounds were networked to one another, showing that the mass shifts truly correspond to a desaturation and elongation of a fatty acid chain, *m/z* 2.018 (H_2) and *m/z* 28.032 (C_2H_4). (*iii*) A subnetwork of ceramide-related compounds (Fig. 2*B*, example 3 and Fig. S4) with mass shifts of *m/z* 2.015 (H_2), *m/z* 14.015 (CH_2), and *m/z* 165.057 ($C_6H_{10}O_5$; glycosylation). A coral-associated ceramide was recently identified (16) with one additional desaturation relative to the ceramide identified here, and this newly identified ceramide has an extremely similar mass (*m/z* 536.504) to the unknown feature (*m/z* 536.508) networked to the ceramide here. The newly identified ceramide also differs in mass from the identified ceramide by *m/z* 2.015, consistent with one fewer saturation. (*iv*) A subnetwork of three unidentified molecules with mass shifts of *m/z* 28.032 (C_2H_4), *m/z* 28.033 (C_2H_4), and *m/z* 56.065 (C_4H_8) (Fig. 2*B*, example 4 and Fig. S5).

Differences in Mass Shift Profiles Between Types of Holobionts. To determine how well MeMSChem profiling resolved each holobiont type, Random Forests classification (17) was used to generate an out-of-bag error (henceforth referred to as a "model error"), which reflects the extent to which the model correctly categorized every sample (i.e., whether *Halimeda* sp. samples were correctly placed into the model's *Halimeda* group). Random Forests offers exceptional classification performance and is robust to nonnormally distributed data and correlated predictors (18), both of which were present in this dataset (Dataset S1).

The usefulness of recategorizing molecules by their mass shifts was first evaluated based on the number of times that each mass shift was observed (counts data described above). The model error of the Random Forests model classifying holobiont types using the counts data was 0.44, which indicates that 44% of the time samples were assigned to the incorrect holobiont type. The resolution gained from the observed counts data (i.e., actual data) was compared with that from 1,000 permutations of the data in which pairs were randomly binned and counted while keeping the original proportions consistent (Dataset S2). The observed counts data outperformed 95% of the randomly generated datasets, suggesting that the counts of redundant mass shifts aided in differentiating between holobiont types despite the relatively high model error (Fig. 3*A*).

Molecular abundance data were then incorporated into the analysis and compared against the holobiont resolution gained

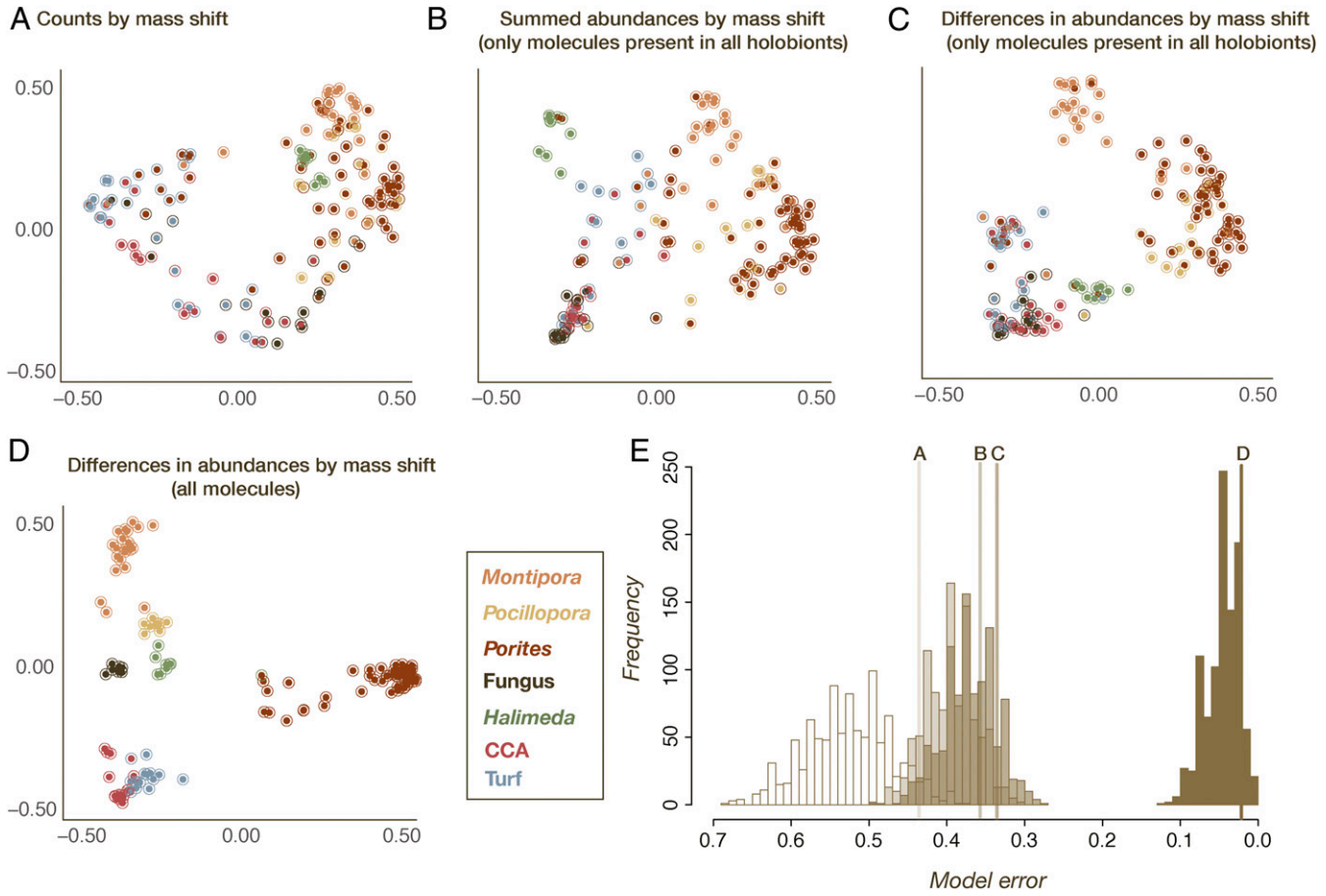

Fig. 3. Results of tests measuring the extent to which holobionts were resolved by MeMSChem profiling. (*A*) A visualization of the first two dimensions of a Random Forests proximity matrix of the number of times that each redundant mass shift was identified (counts data). The color of the filled circle represents the holobiont type of the sample, while the color of the halo around each circle corresponds to the holobiont type it was placed in by the Random Forests model (i.e., if the circle and halo are different colors, the model incorrectly categorized the sample). (*B*) An analogous representation of *A* for the summed abundances of molecules grouped by the mass shifts they exhibit among only the molecular features present in all holobionts. (*C*) An analogous representation of *A* using the difference in abundances of molecules "gaining" minus "losing" a mass, summed by the mass shift they exhibit among only the molecular features present in all holobionts. (*D*) An analogous representation of *A* using the difference in abundances of molecules gaining minus losing a mass, summed by the mass shift they exhibit among all of the molecules in the dataset. (*E*) A histogram of the permutation tests from randomly generated datasets used to determine how well MeMSChem profiling resolves each holobiont type based on the model error. Letters above each line correspond to the model error of the actual data in the figure panel matching that letter. The histograms reflect the model errors of 1,000 permutations of the actual data in which pairs were randomly binned while keeping the original proportions consistent. This was repeated for the data in *A* to *D*, the distributions for which are shown in order and darkening color of counts, summed abundances, differences in abundances in molecules present in all holobionts, and differences in abundances in the entire molecular dataset.

from the counts data. When the summed abundances of each mass shift among molecules present in all holobionts were considered, the model error from the abundance data was 0.36 (Fig. 3*B*). Therefore, incorporating feature abundance data improved the accuracy of the model by 8% when resolving between holobiont types. The value of summing feature abundances by mass shift was also tested by comparing its accuracy with the models of 1,000 permutations of the data in which network pairs were randomly binned and summed while keeping the original proportions consistent (as was done for the counts data above). Among only the molecular features present in all holobionts, summing of feature abundances by mass shift resolved holobiont types better than 90% of the datasets generated from random summing of feature abundances (Fig. 3*B*). Thus, binning abundance data by redundant mass shifts categorizes molecules in a nonrandom manner. Molecular abundances binned by mass shifts also reflected differences among holobiont types better than when holobionts were compared with data that lack any feature abundance information (i.e., counts of the number of mass shifts).

To determine whether mass shifts may reflect active sites of molecular interconversions, as would be expected if a molecular modification had occurred, the summed abundances were compared with the differences in abundances between molecular pairs by mass shift. This is akin to one molecule being the reactant and the other the product. The model error of the differences in abundances data was 0.34, demonstrating that organizing the data by the differences in abundances slightly outperformed the summed abundances data (model error, 34 and 36%, respectively; Fig. 3*C*). Compared with 1,000 random permutations of the actual data, the differences in abundances data outperformed 86% of the random models.

Classification was further improved by incorporating the full molecular dataset, and thus the molecules that were present in all holobionts and the molecules that were only found in one or a few holobionts. When these molecules were included, the model error was 0.02. This reflects a 32% decrease in the model error relative to when only molecules found in all holobionts were considered and was nearly perfect in assigning samples to their correct holobiont type. The real data outperformed 92% of the randomly generated datasets (Fig. 3*D* and summarized in Fig. 3*E*). These results suggest

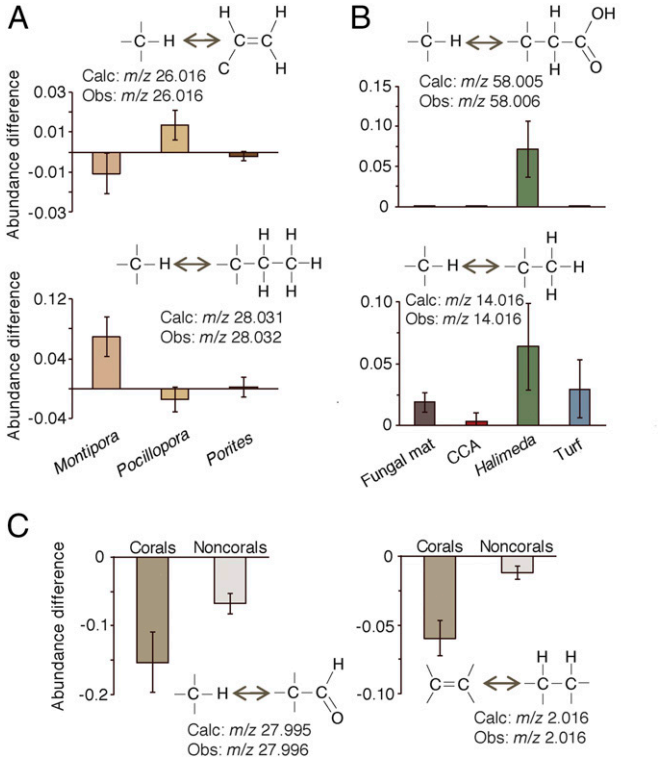

Fig. 4. Annotated mass shifts that best differentiated each holobiont type. (*A*) The annotated mass shifts that best distinguish between coral genera based on the mean decrease accuracy of a supervised Random Forests model. (*B*) The annotated mass shifts that best distinguish between the noncoral holobiont types. (*C*) The annotated mass shifts that best distinguish the coral holobionts from the noncoral holobionts. Bars represent the 95% confidence interval around the mean.

that the highest level of holobiont resolution was achieved when (*i*) molecular features were binned by the redundant mass shifts they exhibited, (*ii*) molecular abundances were included as the difference in abundance between molecules in a network pair, and (*iii*) molecules/pairs that were only found in certain holobionts were included in addition to those molecules present in all holobionts.

Mass Shifts That Best Distinguish Each Holobiont Type. Among the molecular features present in all holobionts, coral genera were best differentiated from one another by mass shifts corresponding to two carbons that were either saturated (*m/z* 28.032, C_2H_4 or 2^*CH_2) or unsaturated (*m/z* 26.016 C_2H_2) ($P < 0.01$ for both; Fig. 4*A*). The three coral genera exhibited distinct patterns between these two mass shifts: Molecular features of *Montipora* exhibited the addition of C_2H_4 and loss of C_2H_2, while *Pocillopora* exhibited the opposite pattern. *Porites*-associated molecules did not gain or lose either mass shift. Putative CH_2 and CH_2OOH mass shifts best differentiated the noncoral holobionts ($P < 0.01$ for both; Fig. 4*B*). *Halimeda* features predominantly gained CH_2, as did turf algae, the fungal mat, and all of the corals, though to a lesser degree than *Halimeda*. Additions of CH_2OOH were unique to *Halimeda*. Corals were best differentiated from noncorals based on larger losses of CO and H_2, the latter suggesting a dehydrogenated state perhaps due to higher concentrations of unsaturated lipids.

Discussion

MeMSChem profiling provides an approach to identify mass shifts between related molecules and annotate them to known chemical groups in complex metabolomes. Seven coral reef holobiont types were well-resolved by MeMSChem profiling. Even among molecular features detected in all holobionts, mass shift profiles differed among holobiont types, suggesting that each type of holobiont is modifying the same molecules in different ways. The chemical differences between holobionts was much more apparent when all molecules were considered (i.e., molecules only produced by certain holobionts were also incorporated), suggesting that disparate mass shift patterns between holobionts play a role in generating molecular diversity in this ecosystem. Shifts in the abundance of molecules exhibiting each mass shift better resolved holobiont types than the number of times each mass shift was detected. Together, these findings suggest that holobionts differ more in their patterns of molecular abundance changes (akin to gene expression) than in the diversity of mass shifts they can carry out (akin to genomic diversity).

Mass Shifts Associated with Holobionts Reflect Differences in Molecular Diversity. By focusing on the differences in mass shift profiles between related molecules, MeMSChem profiling expands metabolomic analysis beyond molecular matches in reference libraries to systemic insights into holobiont biochemistry. If annotated mass shifts reflect single types or classes of molecular modifications catalyzed by enzymes, then disparate mass shift patterns among holobionts may arise for multiple reasons. Holobionts for which the hosts have large genomic differences, such as stony corals and turf algae, may merely possess different biochemical pathways. Among closely related holobiont types such as the three stony coral genera, the distinct patterns of molecular relatedness may arise from differential expression of shared genes. However, the largest disparity among coral holobionts was found by including the mass shifts of molecules that are unique to each holobiont. This suggests that the mass shifts of holobiont-specific molecules largely generate each coral holobiont's unique biochemical profile despite the high degree of taxonomic relatedness among these corals.

The mass shifts that differed among holobiont types included differences putatively assigned to single- and double-bonded carbon and oxygen, likely among phospholipids and their derivatives based upon the molecules identified in this dataset previously (12) and in the current analyses. These data show the expected lower saturation state of corals relative to algae (19, 20) based on the mass shift of *m/z* 2.016 putatively assigned to H_2. Greater fatty acid saturation flexibility can mitigate the damage of elevated temperatures that lead to bleaching in corals (21), suggesting that corals benefit from a higher degree of saturation flexibility and homeoviscous adaptation potential relative to the noncorals studied here. While desaturations in coral molecules generate double bonds between carbons, the shift toward gaining H_2O in coral samples suggests these double bonds may be replaced by hydroxyl groups, either directly or through shifts in the relative abundances of molecules. Hydration of phospholipids can lead to conformational changes that alter membrane surface potential and signaling activity (22), suggesting that the higher abundance of hydroxyl groups in corals reflects systemic changes in cell–cell interactions and cellular signaling pathways.

Applications of MeMSChem Profiling. MeMSChem profiling offers a way to analyze existing LC-MS/MS datasets and provides an approach for identifying system-wide changes in small molecules across metabolomes. Here we analyzed a dataset collected from one of the most remote places in the world. Other researchers may have LC-MS/MS datasets that, like this dataset, cannot be resampled or recreated. Therefore, offering a way to gain information in silico is an attractive proposition for many working with untargeted metabolomic data.

While MeMSChem profiling was applied here to uncover similarities and differences among types of holobionts, it opens doors to answering many more questions. Rather than comparing known groups, MeMSChem profiling may be used to uncover clusters in seemingly homogeneous populations (e.g., individuals of

a coral species in a common environment, human patients suffering from the same disease, cohorts of offspring growing in a shared location). Known mass shifts can also be searched for and quantified, which may be particularly useful when looking for a ubiquitous process such as antioxidant activity.

If molecules of interest are identified, the mass shifts around them may be used to detect putative sites of known modifications or previously unidentified biochemistries. Annotated and unknown mass shifts will require further verification with targeted analyses, such as spiking samples with authentic standards, networking, and examining the mass shifts associated with these standards. Once putative modifications are identified, genetics and molecular biology approaches can be used to confirm or identify the responsible enzyme(s). Such an approach may be particularly useful for tracking molecular changes in time-series samples, a primary need for clinicians (23). Future applications of MeMSChem profiling may also employ a more precise binning approach, taking into account the smaller relative variance at higher masses, changes in MS accuracy across parent masses, and precursor differences. Through this process, the continued application of MeMSChem profiling and the data it generates will produce a wealth of previously uncaptured information from data-rich untargeted metabolomic datasets.

Conclusions

Untargeted metabolomics continues to grow as a tool to examine the complex physiologies of life on Earth. We applied an approach that analyzes untargeted metabolomic data based on the chemical relationships between molecules. An analysis of seven coral reef holobionts demonstrated that the relationships between molecules are diverse and distinct between holobiont types. That different types of holobionts had unique MeMSChem profiles despite high genomic similarity suggests that each possesses physiological capabilities that are not easily identified through traditional genomic approaches. The distinct molecular repertoires identified in each holobiont, coupled with the wide diversity of holobiont types on coral reefs, offer insights into why this ecosystem is an abundant source of chemical diversity.

Materials and Methods

LC-MS/MS Data Collection and Molecular Networking. Samples of seven types of holobionts (hosts and associated viral and microbial communities) including corals, algae, and a fungal mat were extracted in 70% methanol and analyzed with LC-MS/MS [as described in Quinn et al. (12); see *SI Materials and Methods* for data acquisition details]. Files were submitted for molecular network analysis using the online workflow in GNPS (5) (Fig. S1), which compares spectral fragmentation patterns and networks-related molecules (Fig. S1). Molecular spectra were also compared against reference libraries of known molecules in GNPS. Details of the networking parameters can be found in *SI Materials and Methods*.

Identifying Aggregations of Mass Shifts in Network Pairs. Across all pairs, the difference in mass between two networked molecular features (referred to as "network pair mass shifts") was searched for aggregations around precise masses. Criteria for identifying aggregations (i.e., redundancies) were established using the similar masses of CO and C_2H_4 (*m/z* 27.995 and *m/z* 28.031, respectively; Fig. S6; see *SI Materials and Methods* for details). The network pairs involved in aggregations were binned by mass shift and counted per sample (Counts dataset in Dataset S1). All molecular features involved in redundant mass shifts were then quantified using the Optimus workflow (https://github.com/MolecularCartography/Optimus). Optimus was used to quantify features involved in redundant mass shifts that were present in all files/holobionts, features involved in redundant mass shifts that were present in each holobiont type, and all molecular features, including those that were not involved in redundant mass shifts (for normalization of the two former datasets). Molecular abundance data were then used to quantify the molecules exhibiting each mass shift and to quantify the prevailing direction of each mass shift (gaining or losing) in each sample (see *Results* and *SI Materials and Methods* for more details).

Data Analysis Using Random Forests. MeMSChem data were analyzed using the ensemble machine learning algorithm Random Forests (17). The seven holobiont types were used as classifiers, and MeMSChem data were used as predictors. The out-of-bag error (referred to as a model error) indicated how well each holobiont type was resolved by the Random Forests model. Permutation tests were used to determine how well the MeMSChem data differentiated the seven holobiont types. These tests were carried out by comparing the model error of the actual data with a distribution of model errors generated from 1,000 randomizations of the data (see *SI Materials and Methods* for more details). The relative importance of each mass shift in differentiating between holobiont types was determined using the Random Forests mean decrease accuracy score and feature importance score (for each holobiont type).

ACKNOWLEDGMENTS. This work was supported by an NSF Partnerships for International Research and Education Grant (1243541; to F.L.R.) and the Gordon and Betty Moore Foundation (GBMF-3781; to F.L.R.). This work was also supported by the NIH through Grant P41 GM103484 and an NIH grant on the reuse of metabolomic data (R03 CA211211). The European Union's Horizon 2020 Research and Innovation Programme further supported this work under Grant Agreement 634402 (to T.A. and I.P.). We thank the Deutsche Forschungsgemeinschaft for supporting this work with a postdoctoral research fellowship to D.P. (Grant PE 2600/1-1).

1. Nicholson JK, Lindon JC (2008) Systems biology: Metabonomics. *Nature* 455:1054–1056.
2. Cho K, Mahieu NG, Johnson SL, Patti GJ (2014) After the feature presentation: Technologies bridging untargeted metabolomics and biology. *Curr Opin Biotechnol* 28:143–148.
3. da Silva RR, Dorrestein PC, Quinn RA (2015) Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci USA* 112:12549–12550.
4. Pirhaji L, et al. (2016) Revealing disease-associated pathways by network integration of untargeted metabolomics. *Nat Methods* 13:770–776.
5. Wang M, et al. (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* 34:828–837.
6. Heinonen M, Shen H, Zamboni N, Rousu J (2012) Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics* 28:2333–2341.
7. Allen F, Greiner R, Wishart D (2015) Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* 11:98–110.
8. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci USA* 112:12580–12585.
9. van der Hooft JJJ, Wandy J, Barrett MP, Burgess KE, Rogers S (2016) Topic modeling for untargeted substructure exploration in metabolomics. *Proc Natl Acad Sci USA* 113:13738–13743.
10. Simmons TL, et al. (2008) Biosynthetic origin of natural products isolated from marine microorganism-invertebrate assemblages. *Proc Natl Acad Sci USA* 105:4587–4594.
11. Rohwer F, Seguritan V, Azam F, Knowlton N (2002) Diversity and distribution of coral-associated bacteria. *Mar Ecol Prog Ser* 243:1–10.
12. Quinn RA, et al. (2016) Metabolomics of reef benthic interactions reveals a bioactive lipid involved in coral defence. *Proc Biol Sci* 283:20160469.
13. Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG (2017) Retrospective analysis of natural products provides insights for future discovery trends. *Proc Natl Acad Sci USA* 114:5601–5606.
14. Dinsdale EA, et al. (2008) Microbial ecology of four coral atolls in the northern Line Islands. *PLoS One* 3:e1584.
15. Smith JE, et al. (2016) Re-evaluating the health of coral reef communities: Baselines and evidence for human impacts across the central Pacific. *Proc Biol Sci* 283:20151985.
16. Eltahawy NA, et al. (2015) Mechanism of action of antiepileptic ceramide from Red Sea soft coral *Sarcophyton auritum*. *Bioorg Med Chem Lett* 25:5819–5824.
17. Breiman L (2001) Random forests. *Mach Learn* 45:5–32.
18. Berk RA (2006) An introduction to ensemble methods for data analysis. *Sociol Methods Res* 34:263–295.
19. Harland AD, Navarro JC, Davies PS, Fixter LM (1993) Lipids of some Caribbean and Red Sea corals: Total lipid, wax esters, triglycerides and fatty acids. *Mar Biol* 117:113–117.
20. Carballeira NM, Sostre A, Ballantine DL (1999) The fatty acid composition of tropical marine algae of the genus *Halimeda* (Chlorophyta). *Bot Mar* 42:383–387.
21. Tchernov D, et al. (2004) Membrane lipids of symbiotic algae are diagnostic of sensitivity to thermal bleaching in corals. *Proc Natl Acad Sci USA* 101:13531–13535.
22. Mashaghi A, et al. (2012) Hydration strongly affects the molecular and electronic structure of membrane phospholipids. *J Chem Phys* 136:114709.
23. DeBerardinis RJ, Thompson CB (2012) Cellular metabolism and disease: What do metabolic outliers teach us? *Cell* 148:1132–1144.
24. Frank AM, et al. (2008) Clustering millions of tandem mass spectra. *J Proteome Res* 7:113–122.
25. Bouslimani A, et al. (2015) Molecular cartography of the human skin surface in 3D. *Proc Natl Acad Sci USA* 112:E2120–E2129.
26. Petras D, et al. (2016) Mass spectrometry-based visualization of molecules associated with human habitats. *Anal Chem* 88:10775–10784.
27. Floros DJ, et al. (2017) Mass spectrometry based molecular 3D-cartography of plant metabolites. *Front Plant Sci* 8:429.
28. Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2:18–22.
29. Archer F (2016) rfPermute: Estimate Permutation *P*-Values for Random Forest Importance Metrics. R package (Zenodo), Version 2.1.1. Available at doi.org/10.5281/zenodo.60414. Accessed August 23, 2016.