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.

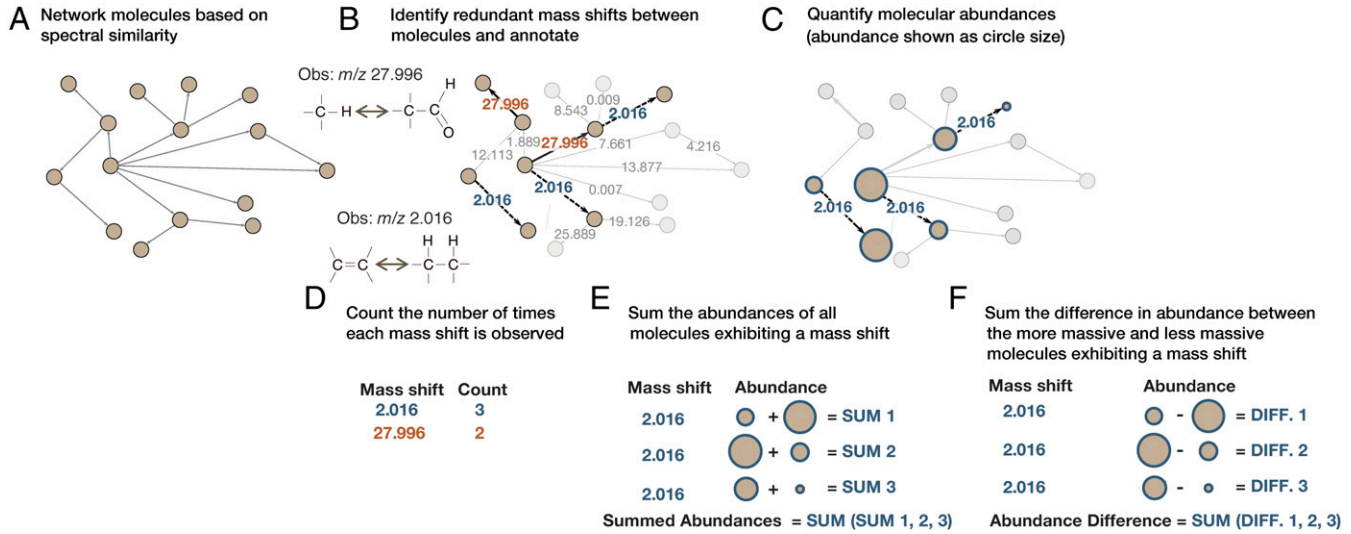

Fig. 1. Data processing and generation based on a simplified molecular network and two redundant mass shifts. (*A*) GNPS used MS/MS fragmentation spectra to elucidate molecular similarities and network similar molecules (i.e., related molecules). (*B*) Redundant mass shifts between related molecules 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.

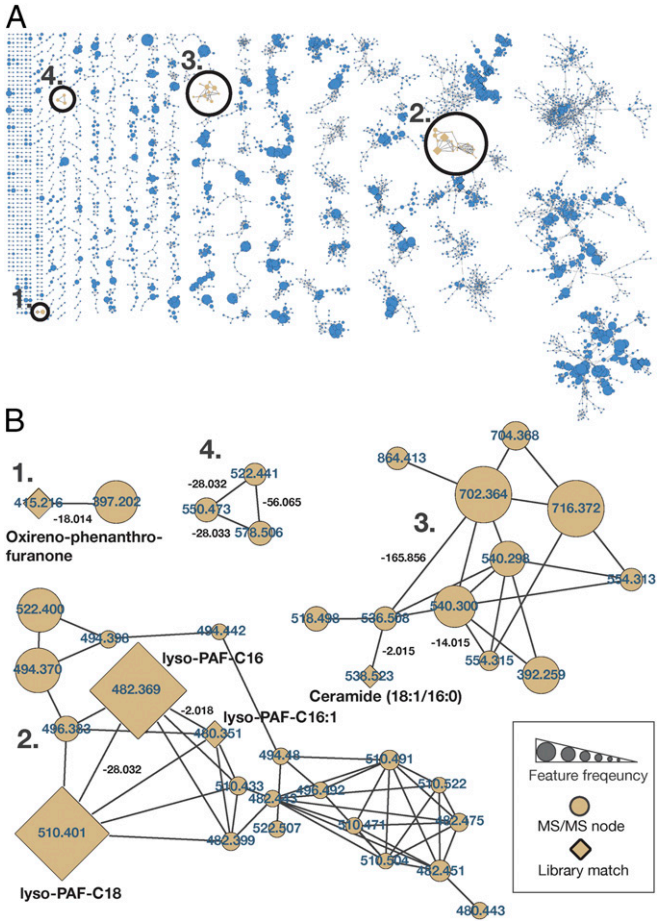

Fig. 2. Molecular network of the reef holobiont MS/MS dataset. (*A*) The global molecular networks of MS/MS spectra from the coral reef holobiont metabolomic dataset. Each node represents a unique or set of identical spectra (i.e., molecular feature). Lines connecting the nodes represent their spectral 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

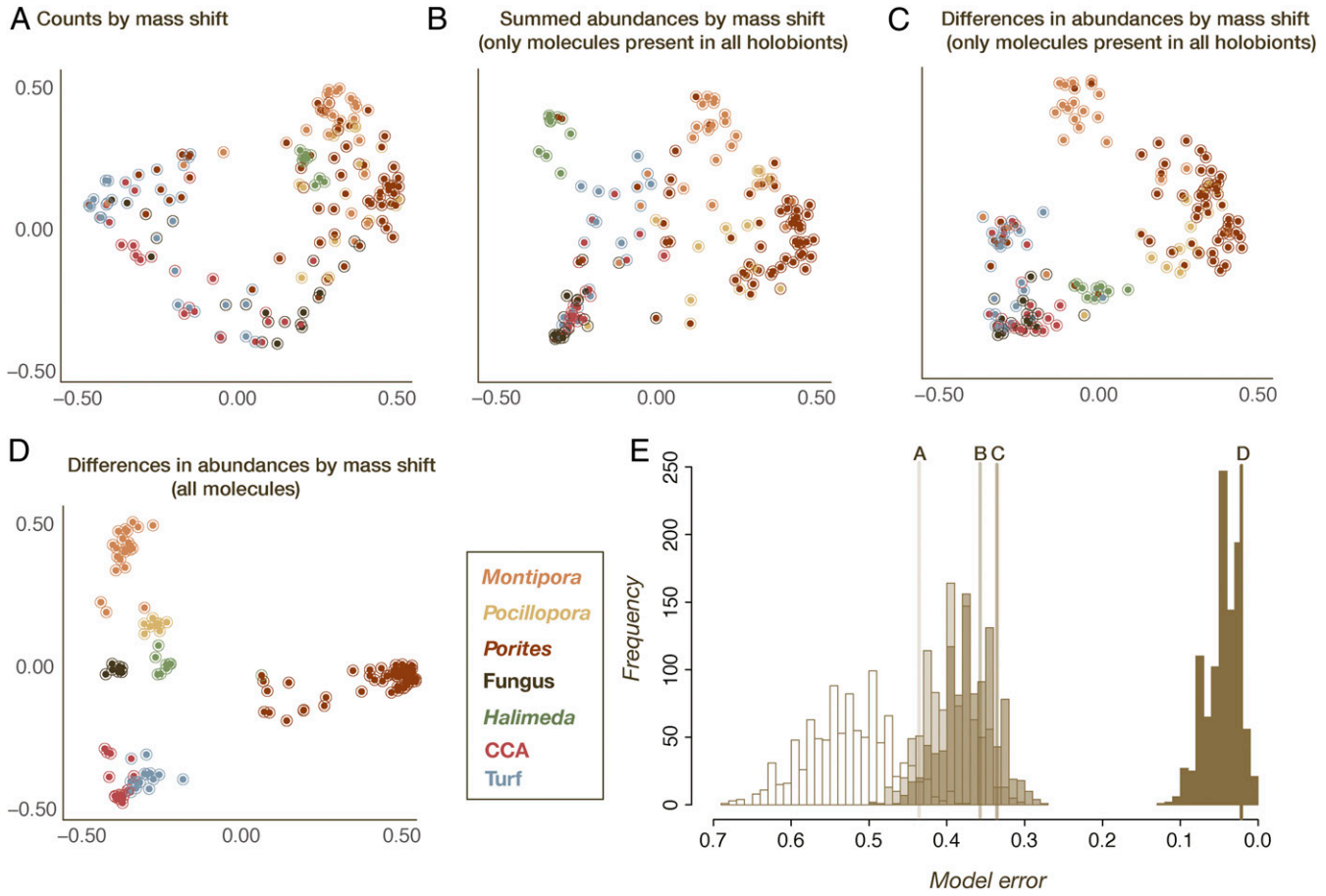

Fig. 3. Results of tests measuring the extent to which holobionts were resolved by MeMSChem profiling. (*A*) A visualization of the first two dimensions of a Random Forests proximity matrix of the number of times that each redundant mass shift was identified (counts data). The color of the filled circle 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

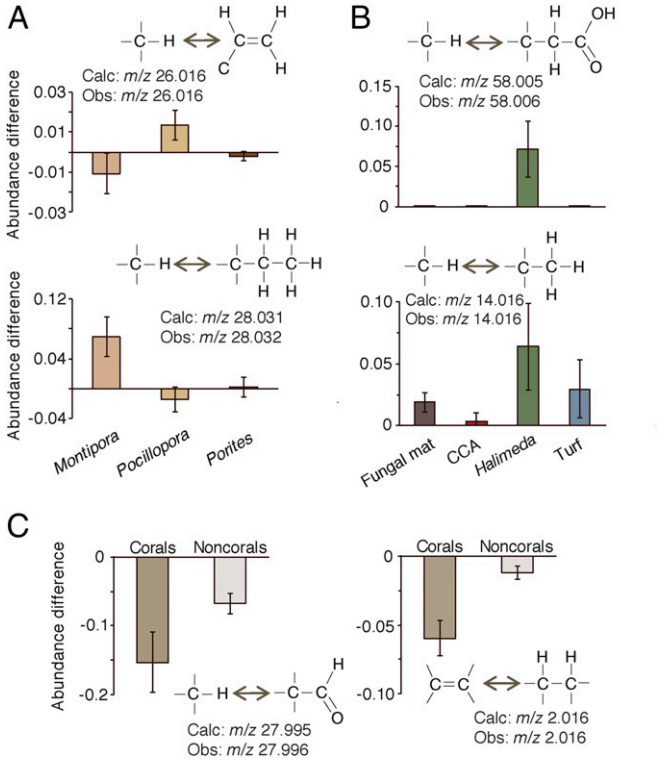

Fig. 4. Annotated mass shifts that best differentiated each holobiont type. (*A*) The annotated mass shifts that best distinguish between coral genera based on the mean decrease accuracy of a supervised Random Forests model. (*B*) The annotated mass shifts that best distinguish between the noncoral holobiont 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.

Supporting Information

Hartmann et al. 10.1073/pnas.1710248114

SI Materials and Methods

LC-MS/MS Data Collection. As described in Quinn et al. (12), samples of coral, algae, and a fungal mat were extracted in 70% methanol and analyzed with a Thermo Fisher Scientific UltiMate 3000 Dionex UHPLC coupled to a Bruker Daltonics maXis qTOF mass spectrometer. An internal standard of glycocholic acid was used for normalization across samples. Lock mass internal calibration was achieved using a wick within the source that was saturated with hexakis phosphazene ions (1H,1H,3H-tetrafluoropropoxy, *m/z* 922.0098; SynQuest Laboratories) to allow for constant infusion of the calibrant into the instrument. An injection volume of 30 μL was used and samples were separated on a Kinetex 2.6-μm C18 (30 × 2.10 mm) UHPLC column. A linear water/acetonitrile gradient of +0.1% formic acid was used for the mobile phase at a flow rate of 0.5 $mL \cdot min^{-1}$. Acquisition was carried out for MS spectra in the mass range of *m/z* 50 to 2,000, and the 10 most intense ions per scan were fragmented using collision-induced dissociation at 35 eV for +1 ions and 25 eV for +2 ions. The instrument was operated in positive-ion mode and data-dependent acquisition. Ions were ignored for MS/MS after being selected in three consecutive scans but were refragmented when their intensity increased to greater than two and a half times that of the previous fragmentation scan in which it was detected.

Molecular Networking. Raw MS data were recalibrated using the lock masses of an internal standard constantly infused in the ESI source, normalized to an internal glycocholic acid standard, and converted to mzXML files. The files reported in ref. 12 were submitted for molecular network analysis using the online workflow at GNPS (5) (Fig. S1). These files can be found on the Mass spectrometry Interactive Virtual Environment (MassIVE) at https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp with accession no. MSV000078598. In GNPS, the similarity of spectral fragmentation patterns (i.e., similarity of two molecular features) was compared in a pairwise fashion across all spectra using the cosine score as a reporter of relatedness. Network parameters are chosen by the user in the online interface before network generation, and detailed documentation is available on GNPS to guide the user through this process. The following GNPS parameters were used here: Data were filtered by removing all MS/MS peaks within *m/z* 17 of the precursor *m/z*. MS/MS spectra were window-filtered by choosing only the top six peaks in an *m/z* ±50 window throughout the spectrum. The data were then clustered with MSCluster (24) with a parent mass tolerance of *m/z* 0.02 and an MS/MS fragment ion tolerance of *m/z* 0.02 to create consensus spectra (referred to as "nodes" in GNPS and "molecular features" here). Molecular features that contained fewer than two spectra were discarded.

Two features were connected in a network when their spectral similarity had a cosine score above 0.6 (henceforth referred to as "network pairs"), when both features had at least four matching spectral peaks, and when each of the features was among each other's top 10 most similar features. The scoring algorithm is based on matching MS/MS fragments between two spectra, allowing an equal mass delta between MS/MS fragments as there is between the precursors. This process primarily identifies analogs that differ via one modification, or through multiple modifications at one specific side of the molecule. Compounds differing by two or more modifications at distinct sides might be connected in the network through a single modified analog.

With these settings applied (0.6 cosine score), 11,623 nodes (consensus spectra or features) were generated in GNPS, 4,597 of which were networked to at least one other molecular feature (39.6%). When the default GNPS cosine score of 0.7 was applied [i.e., stricter limits on whether two features were networked (5)], 11,619 nodes were generated and 4,525 of those nodes networked, 38.9% of the total, or 0.7% fewer nodes than with the 0.6 cosine score. While the increased number of network pairs was small, networking with the 0.6 cosine score was used to develop MeMSChem profiling to include more features and allow molecular features with slightly more disparate spectra, but that could still differ by a redundant mass shift, to be considered.

The spectra in the network were then searched against the GNPS spectral libraries using the default settings. To focus only on the most high-quality library hits, only spectrum library matches with precursor mass errors <20 ppm were considered for further analysis (Fig. 2). A total of 0.74% of features were annotated to library IDs. This low degree of identification is common in other nontargeted metabolomic datasets (25–27), confirming that there is much information still to be gained from preexisting LC-MS/MS datasets.

Identifying Aggregations of Mass Shifts in Network Pairs. Connections between two features in any network were used to identify aggregations of mass shifts (i.e., repetitive or redundant mass differences between network pairs). When a feature was networked to more than one other feature it was considered to exhibit multiple potential mass shifts. All network pairs were downloaded from GNPS as a comma separated values (CSV) file list that included the GNPS ID of each feature in every pair as well as the exact mass shift between the two features in every pair (henceforth referred to as "network pair mass shift"). Features associated with contaminants were removed from the network pair dataset before identifying redundant mass shifts (Table S2). Contaminants were identified in two ways: (*i*) when a molecular feature was matched to a library ID known to be a contaminant, and (*ii*) when a mass shift known to arise from a contaminant was found, even if the molecules were not identified. Included in removals were a number of molecular features that were double-charged, which yielded half-mass-unit shifts (e.g., *m/z* 3.5XX). Mass shift *m/z* 3.008 was merged with *m/z* 2.016, as the former likely resulted from molecules with the heavy isotope of hydrogen. All network pairs containing at least one contaminant molecule or separated by a known contaminant mass shift (e.g., sodium formate clusters) were removed before further analysis, which led to the removal of 384 molecular features.

To identify redundancies, the network pairs were searched for aggregations around precise mass differences between related molecules. Doing this first required determining the optimal mass resolution for identifying aggregations of network pair mass shifts. This was done by determining how well different filtering criteria separated the similar masses of CO and C_2H_4 (*m/z* 27.995 and *m/z* 28.031, respectively; Fig. S6). First, network pairs were rank-ordered by mass shift (i.e., smallest to largest mass shift between two spectra), and then a sliding window was passed sequentially across the network pairs, one network pair at a time (Fig. S1*D*). A window was flagged as a potential aggregation when the range of mass shifts (largest mass shift − smallest mass shift within a window) was equal to or below a threshold value. Multiple mass shift thresholds were tested including *m/z* 0.04, 0.02, 0.01, and 0.001 using 5 versus 10 network pairs within a window (Fig. S6). Any network pairs that did not fall within a flagged window were discarded to isolate only the network pairs exhibiting aggregations/redundancies. Windows with a width of *m/z* 0.001 that contained five network pairs best separated the known mass shifts of CO and C_2H_4, and thus these criteria were applied to the entire dataset (Fig. S6).

Because the sliding window increased sequentially across every network pair but included five network pairs, a single network pair could fall within multiple flagged windows. To account for this redundancy and precisely identify the center of aggregation around specific mass shifts, the most abundant mass shifts remaining after filtering were identified (visualized in a histogram in Fig. S1*E*). Network pairs m/z ±0.001 of the "peaks" in mass shifts in network pairs were included in the final list of redundant mass shifts and used to bin/annotate network pairs by the mass shift they exhibited. Also considered was whether the absolute variance changed over the range of parent masses, given that the relative variance decreases as the parent mass increases. Little evidence was found that absolute variance was related to parent mass (Fig. S7), though the change in relative variance across parent masses could be considered in a more advanced binning approach. While these parameters were employed here, more or less stringent criteria can be applied (e.g., fewer network pairs per window), or this approach could be used to find mass shifts of interest a priori, depending on the research goal.

Tabulating Redundant Mass Shifts Found in Each Sample. The GNPS networks were created using features found in all samples, though not all features were present in all samples. As a result, mass shifts had to be identified on a per-sample basis. To generate these data, features found in all redundant mass shift network pairs had to be isolated from all features found in the networks and then searched for in each sample. To generate the lists of features to isolate, network pairs were split into two lists of features—one of the more massive feature ("gaining" dataset) and the other of the less massive feature ("losing" dataset) for each network pair. A unique network pair ID, GNPS feature ID, and mass shift of the network pair were maintained in the gaining and losing feature lists. When features were networked to more than one other feature, they were included redundantly in the dataset such that all their network pairs were represented.

The list of all molecular features identified by MSCluster, the samples they were found in, and all associated information were downloaded from GNPS. This overall feature list was then reduced to only the features found in the gaining and losing feature lists, which were then used to tabulate the presence or absence of each feature in each sample in a binary bucket table (1, present; 0, absent). The gaining and losing counts bucket tables were then combined into a single binary bucket table that tabulated whether both features in a network pair were found in each sample. This step was necessary because both features in a network pair were not always present in every sample. In this combined dataset, the network pair mass shift was appended in place of the molecule ID numbers to annotate each network pair by the mass of the mass shift it exhibits. Network pair mass shifts were then binned for each sample by tabulating the binary values (i.e., 1 or 0) of the presence or absence of a network pair, yielding the "counts" of mass shifts by sample (Dataset S1, Counts tab).

The counts data were used to identify holobiont-specific mass shift patterns based on how common or rare mass shifts were among holobiont types (Fig. 3*A*). For these analyses, the counts data were not normalized for two reasons: (*i*) The dataset was predominated by zeros, and (*ii*) the network pairs in the redundant mass shift datasets represent a small subset of all network pairs found in the overall networks generated by GNPS, and thus to normalize to, for example, the total number of network pairs in a sample would normalize to a value that does not reflect the truncated dataset isolated during MeMSChem profiling. On the other hand, the network pairs in the overall networks include many network pairs that do not reflect redundant mass shifts (hence the need to identify them here), and thus the total number of network connections does not represent a useful normalizing value either.

Notably, the counts data yielded relatively few counts per mass shift per sample (mean <1 per redundant mass shift). This apparent scarcity of mass shifts was amplified by the assignment of very similar features (i.e., potentially the same molecule) to different features in GNPS. When this happened, it reduced the likelihood that both features in a network pair were found in any given sample, thus reducing the number of mass shifts (i.e., counts data) overall. This issue was ameliorated by incorporating another feature-finding tool, Optimus, which independently identified features using the list of features generated by GNPS using MS scans (i.e., parent molecules). Features were combined when Optimus identified multiple GNPS features as a single feature and vice versa, consolidating and corroborating the molecular features represented in the dataset.

Quantifying Feature Abundances. The Optimus workflow was used to quantify molecular features exhibiting redundant mass shifts based on MS data (https://github.com/MolecularCartography/Optimus). Among its many functionalities, Optimus uses m/z and retention time feature lists to search mass spectra data, returning feature abundances based on MS (i.e., parent molecule). This search was used to narrow down the list of all detected MS-based features to MS/MS-based feature identifications and networks generated with GNPS. Optimus employs a feature occurrence rate filter, which reduces the number of "missed" features in the dataset, thereby aiding in statistical comparisons. To quantify features involved in redundant mass shifts, all features in the gaining and losing datasets were input into Optimus along with their GNPS IDs, which allowed Optimus-quantified features to be mapped back to the GNPS features and mass shift network pairs. Three feature abundance datasets were generated using Optimus: (*i*) the abundances of only the features involved in redundant mass shifts that were present in all files/holobionts, (*ii*) the abundances of only the features involved in redundant mass shifts that were present in each holobiont type (i.e., each group was analyzed in separate batches and then combined), and (*iii*) the abundances of all molecules (i.e., molecules involved in redundant mass shifts and molecules not involved). The data in (*iii*) were used to normalize the abundance data collected in (*i*) and (*ii*) by dividing the abundance of each molecule by the sum of the abundance of all molecules in the sample, as follows:

Molecular feature (GNPS ID)	Raw abundance (feature)	Total raw abundance (sample)	Normalized abundance
100	8,836	121,566,027	8,836/121,566,027 = 0.000073
101	8,032	121,566,027	8,032/121,566,027 = 0.000066
102	188,208	121,566,027	188,208/121,566,027 = 0.001548
103	689,740	121,566,027	689,740/121,566,027 = 0.005674

The data in (*i*) represented only the molecules present in all holobionts to assess whether different holobiont types exhibit unique mass shift patterns among a pool of molecules that are present/shared across all holobionts. The data in (*ii*) allow for mass shifts among features found in all holobionts and features that are only found in one or a few holobionts. Among only the molecules found in all holobionts (*i*), MeMSChem profiling isolated 512 molecular features involved in redundant mass shifts.

Among all molecules in the entire dataset (*ii*), MeMSChem profiling isolated 728 molecular features involved in redundant mass shifts.

The feature abundances quantified by Optimus were mapped back to their respective GNPS features. Abundance of features identified in blanks were not removed, due to the fact that Optimus identifies the abundance of all features in all samples (i.e., the number of features found in blanks can be artificially elevated). Three forms of MeMSChem mass shift data were generated, as described in *Results*: (*i*) the number of times each mass shift was observed in each sample (counts data); (*ii*) the summed abundances of all molecules exhibiting each mass shift in a sample (summed abundances data); and (*iii*) the difference in abundances of each network pair (abundance of the more massive molecule − abundance of the less massive molecule) for each mass shift, summed by mass shift (differences in abundances data). All data are presented in Dataset S1 and the calculations were carried out as shown below:

Network pair (GNPS ID of both molecular features)	Mass shift	Count	Summed abundances	Differences In abundance
101–102	− *m/z* 2.016	1	0.000073 + 0.000066 = 0.000139	0.000073 − 0.000066 = 0.000007
102–103	− *m/z* 2.016	2	0.001548 + 0.005674 = 0.007222	0.001548 − 0.005674 = −0.004126
. . .	− *m/z* 2.016	. . .*n*	. . .sum_*n*	. . .differences_*n*
	m/z 2.016 BIN	*n*	0.000139 + 0.007222 + . . .sum_*n*	0.000007 + −0.004126 + . . .differences_*n*

The differences in abundances data provided a conservative estimate of molecular abundance changes, because taking the difference between feature abundances can lead to small values when both features have similar abundances and because summing negative and positive values draws the summed value toward zero. Therefore, the detection of directional feature abundances (positive or negative) serves as a robust indicator of systemic processes.

Data Analysis Using Random Forests. MeMSChem data (counts, summed abundances, differences in abundances) were analyzed using the ensemble machine learning algorithm Random Forests (17) with the R packages randomForest (28) and rfPermute (29). A Random Forests consists of an ensemble of decision trees for classification. Each tree in the ensemble is grown using a different bootstrap sample of the original data. In addition, when growing each tree, a small random subset of the candidate variables available for splitting at each node of the tree is selected (in the present analysis, four variables were selected). For highly correlated data, the set of randomly selected variables tends to decorrelate the trees and produces more diverse trees. For the final prediction, the class majority vote from all of the trees is used. Replicate samples of the seven holobiont types were used as classifiers with MeMSChem data as predictors in an rfPermute Random Forests model. The number of samples per classifier used in each tree was equilibrated due to unequal sample sizes between holobiont types and run for 10,000 trees. The equilibration was equal to half the sample size of the most sample-depauperate holobiont type.

The first two dimensions of the proximity matrix from Random Forests models that were run for each data type are projected in Fig. 3. The proximity matrix is the measure of the "closeness" for each pair of observations in the data. After the Random Forests is grown, all of the data (both training and out-of-bag) are put down each tree. If two observations end up in the same terminal node, then their proximity is increased by one. At the end, the proximity is normalized by dividing by the number of trees. The out-of-bag error of the Random Forests model (OOB) reflects the extent to which the model placed samples back into their correct classes (i.e., holobiont types) and was referred to here with the general term "model error." OOB was used as a model response metric in two ways. First, OOB was used to gain biologically relevant information across different types of mass shift data: counts, summed abundances, and differences in abundances. For example, differences among holobiont types based on the counts data may be less than that in the mass shift abundance data, which would be reflected in high and low OOB values, respectively. Second, the Random Forests class error and confusion matrix, reflecting the model's ability to classify each holobiont type and the classifiers into which samples were placed, were used to assess how similar or distinct each holobiont type was from every other.

Random Forests Permutation Tests. Permutation tests were used to determine how well the MeMSChem data differentiated the seven holobiont types. In these tests, the model error was used to measure model accuracy based on correctly classifying holobionts. To test how well the samples were classified when binned by counts, summed abundances, or differences in abundances, the Random Forests model error of the actual data was compared with the model error of 1,000 permutations of the data in which pairs were randomly binned while keeping the original proportions consistent (Dataset S2). The ability of the actual data to differentiate the seven holobiont types was then determined based upon the quantile into which it fell in the distribution of model errors in the randomly generated datasets.

Identifying the Mass Shifts That Best Differentiate Holobiont Types. The extent to which each redundant mass shift aids the Random Forests model in correctly assigning samples by holobiont type was evaluated across all holobionts based on the mean decrease accuracy score and feature importance score (for each holobiont type). Implementation of the rfPermute package in R (29) generated statistical *P* values for each mean decrease accuracy/feature importance score by randomly permuting the response variable and comparing the observed importance scores with the null distribution from the permutations. These comparisons were run based on various groupings of holobiont types to determine which mass shifts best differentiated, for example, all corals vs. all noncorals, coral genera from each other, and noncoral holobionts from each other.

Data Accessibility, Processing Sources, and Scripts. All LC-MS/MS data can be found on the Mass spectrometry Interactive Virtual Environment (MassIVE) at https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp with accession no. MSV000078598. Molecular networking was performed using the online GNPS platform [gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp (5)]. The results of the network analysis with a cosine score of 0.6 can be found at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=faaf7ea1ff8c4942afcd819b621d871e, while the analysis at a cosine score of 0.7 can be found at gnps.ucsd.edu//ProteoSAFe/status.jsp?task=06f164e107404b239ea2f04f710f8c40. MS

features were quantified using Optimus (https://github.com/MolecularCartography/Optimus) in KNIME 3.2.0 (https://www.knime.org). Identifying redundant mass shifts was carried out in Microsoft Excel. The R code used for the rfPermute data analyses can be found in Dataset S2.

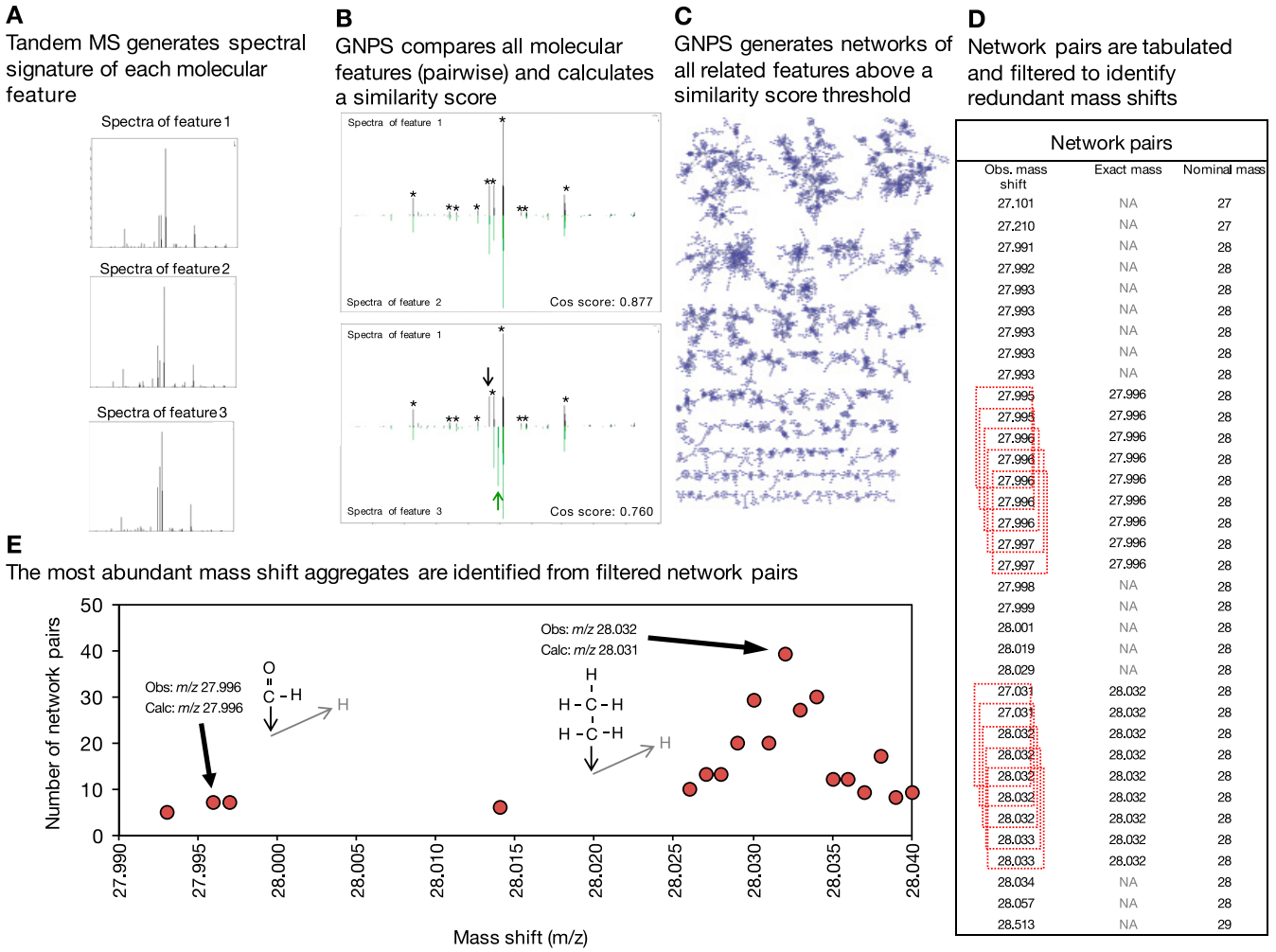

Fig. S1. Analytical pipeline for identifying redundant mass shifts. (*A*) Tandem mass spectrometry precisely weighs molecular features by their mass-to-charge ratio, after which a subset of features are fragmented to generate spectral signatures for each feature. The spectra of three molecular features are shown as an example. (*B*) GNPS, a publicly available molecular networking platform, compares the spectra of all molecular features with that of all others in a pairwise fashion. Each of these comparisons generates a cosine similarity score, denoting the similarity of the two molecular features. Here, feature 1 is compared with features 2 and 3. Asterisks denote shared fragments, and arrows denote major fragments that are not shared between the features. (*C*) Represented here are the GNPS networks of all related molecular features based upon fragment similarity patterns. Molecular relatedness is determined by a threshold cosine score, above which all molecular features are included in the network. (*D*) All connections in the network were tabulated and sorted by the precise mass difference between two related/networked molecules (Obs. mass difference). Redundant mass shifts were identified when a window (red box) contained five network pairs within a mass range less than or equal to *m/z* 0.001. Thus, "NA" denotes that network pair mass did not fall in a window that met these criteria. (*E*) Network pairs were further filtered to identify the most abundant mass shifts (e.g., *m/z* 28.032). The differentiability of two similar, known mass shifts of *m/z* 27.996 and *m/z* 28.031 is depicted as an example to illustrate the importance of identifying putative mass shifts to a relatively high degree of mass precision.

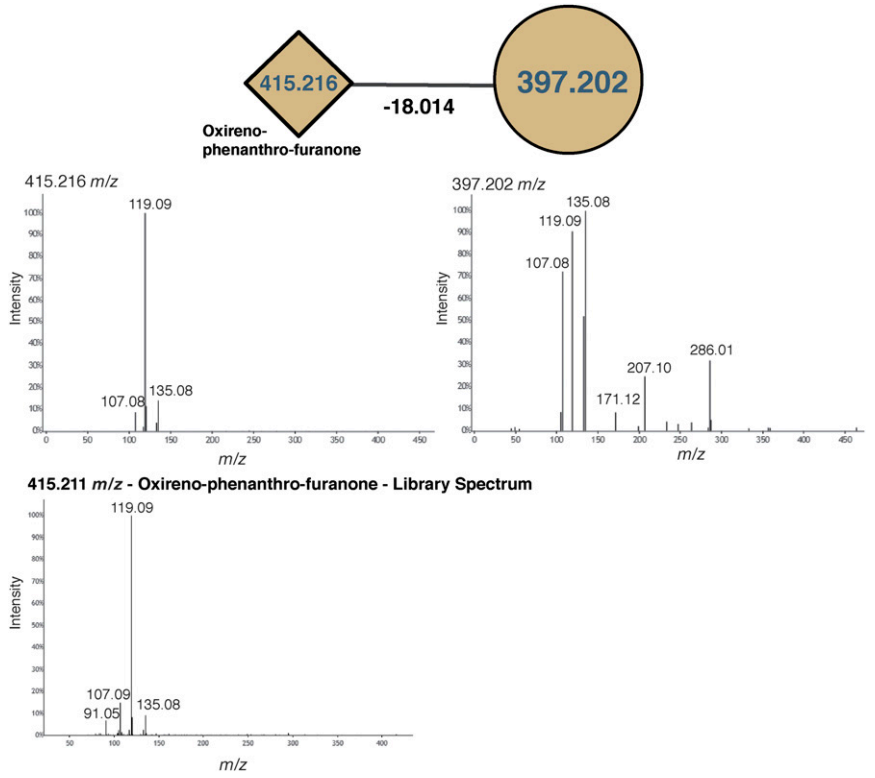

Fig. S2. Spectra of the molecular features shown in Fig. 2*B*, example 1.

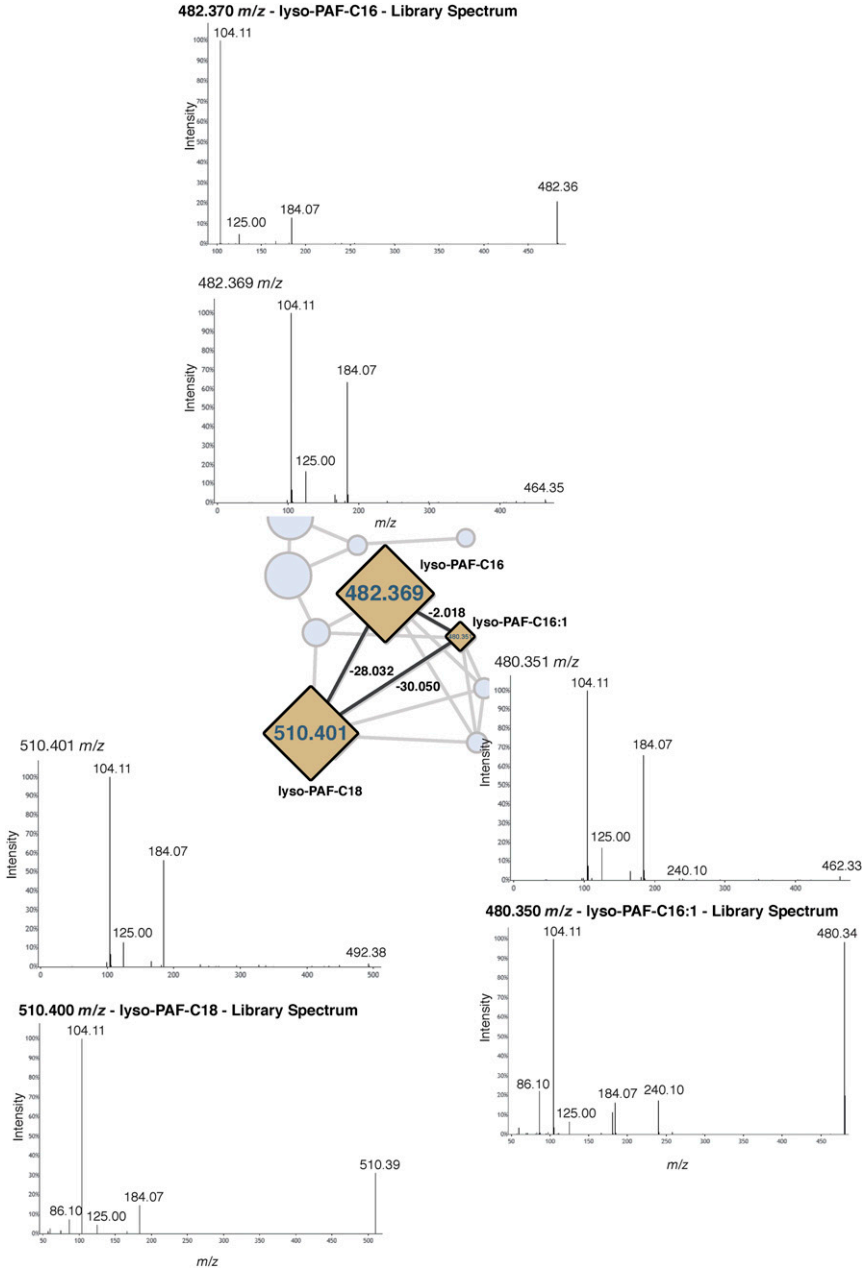

Fig. S3. Spectra of the molecular features shown in Fig. 2*B*, example 2.

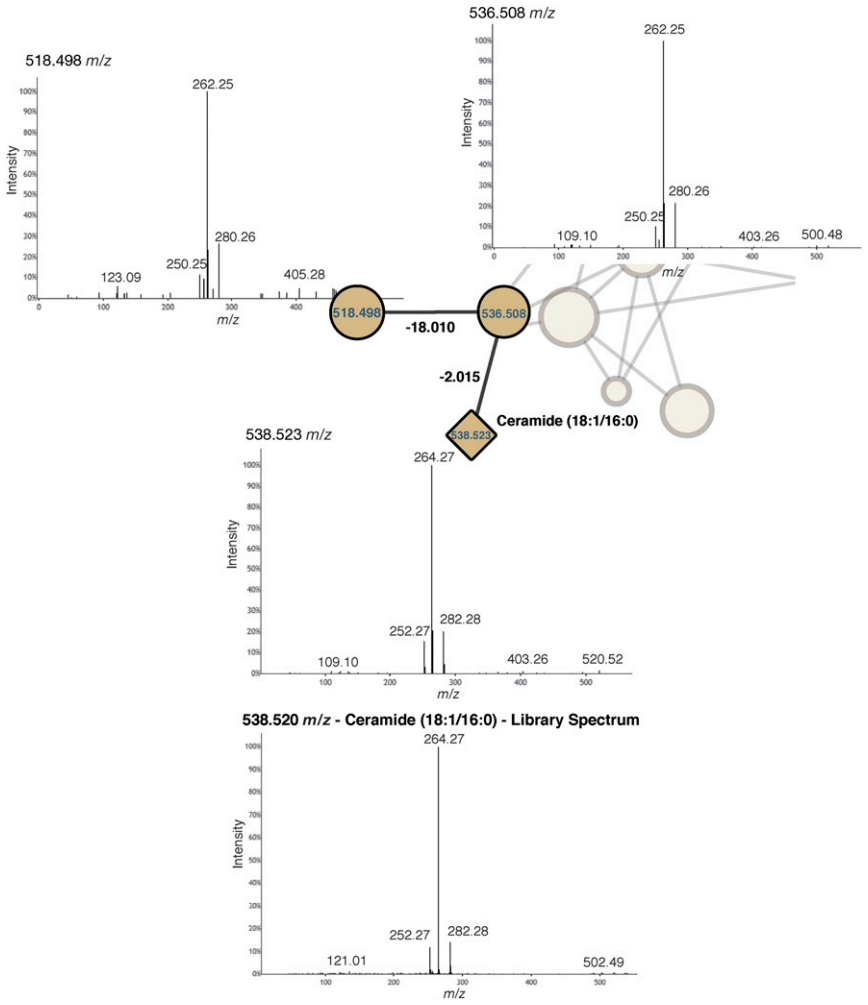

Fig. S4. Spectra of the molecular features shown in Fig. 2*B*, example 3.

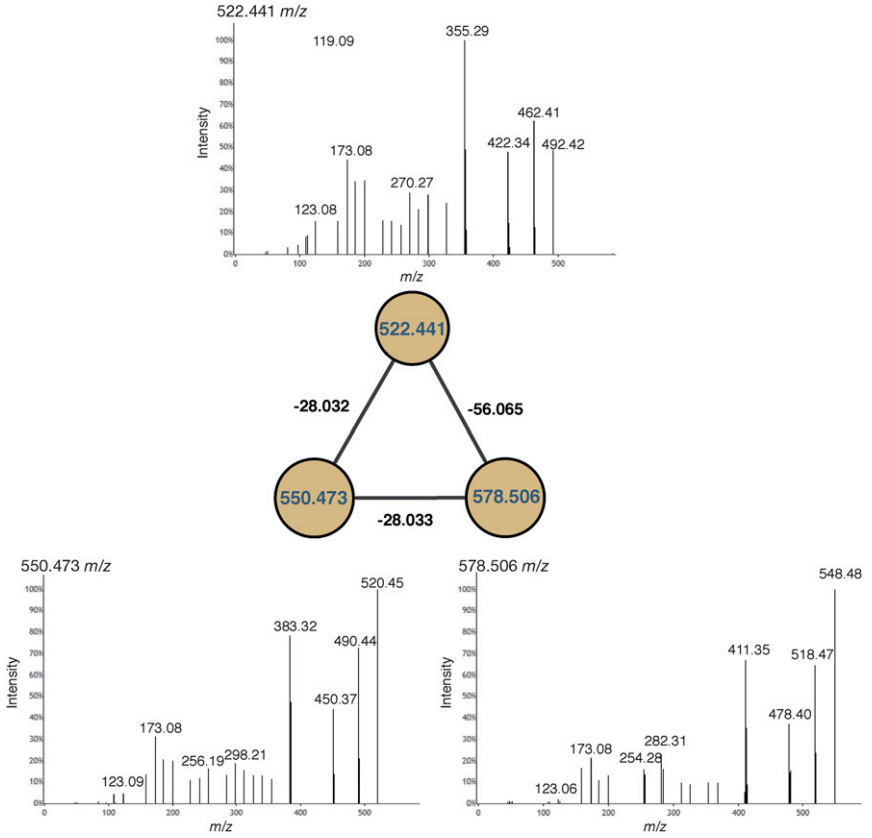

Fig. S5. Spectra of the molecular features shown in Fig. 2*B*, example 4.

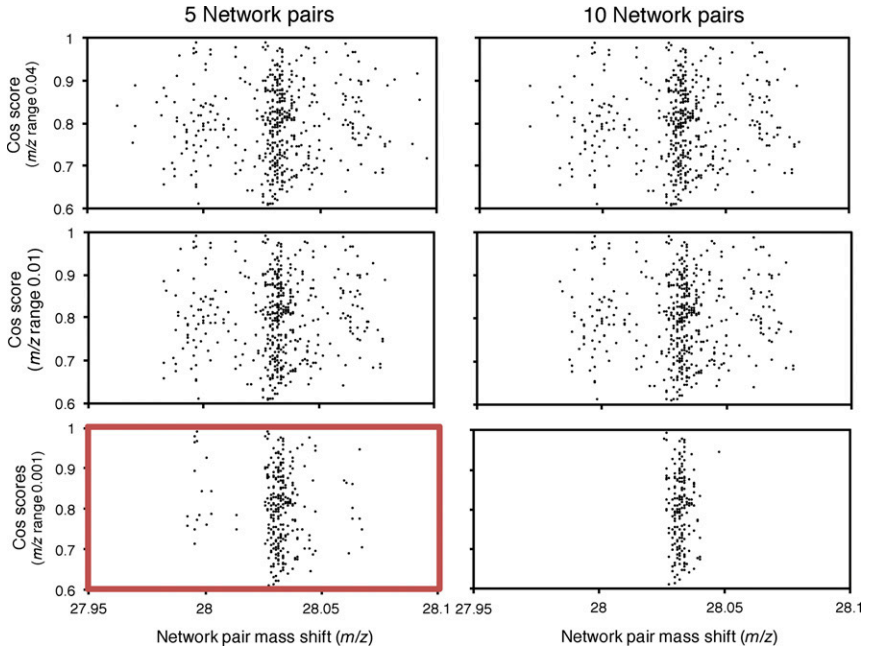

Fig. S6. Identifying the optimal filtering criteria to identify common and redundant mass shifts. Network pair mass shifts between feature pairs were filtered based on six criteria: the mass range (of mass shifts) among network pairs: *m/z* 0.04, 0.01, and 0.001, and the number of network pairs included to generate that range: 5 or 10. The mass shifts between networked pairs are plotted against the cosine scores (cutoff at 0.6). Shown are the known chemical groups of CO and C_2H_4 (calculated *m/z* 27.995 and *m/z* 28.031, respectively) to assess the best filtering approach. Based on these results, criteria of five network pairs and *m/z* 0.001 bins were chosen (red box).

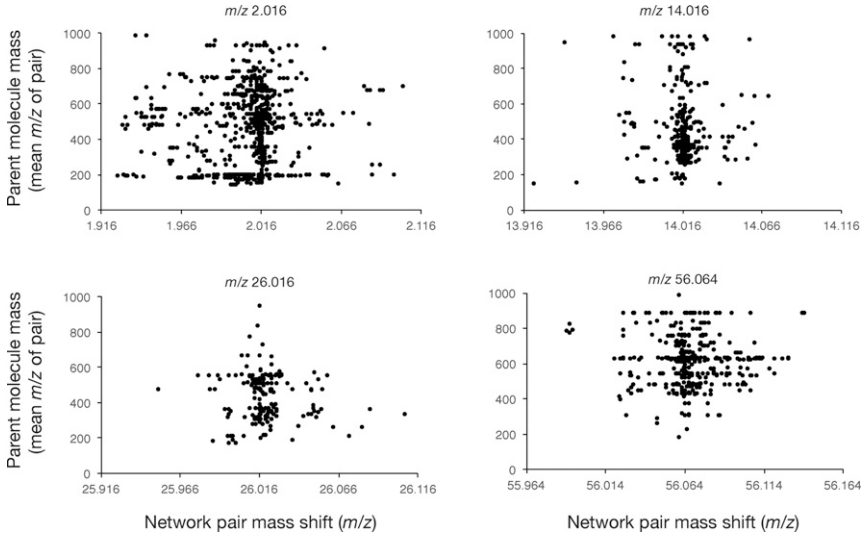

Fig. S7. Relationship between parent molecule mass (mean of both features in the pair) and network pair mass shift.

Table S1. All identified redundant mass shifts

Obs. mass	% mass shift	Putative element or group
0.967	2.05	–
1.007		M+H
1.039	5.34	–
1.089	0.62	–
1.834	1.54	–
1.864	1.64	–
1.947	0.82	–
2.016	14.48	H_2
2.055	0.51	–
2.504		
2.964	3.70	–
3.008		
3.032	0.82	–
3.519		
3.995	0.62	$CH_2 \leftrightarrow H_2O$
4.008	0.72	–
4.010	0.82	–
4.894	0.51	–
4.961		$Na^+ \leftrightarrow NH_4^+$
7.008	1.54	–
12.000	1.95	C
12.003	0.82	–
14.016	11.09	CH_2
14.026	0.82	–
15.995	1.23	O
15.998	1.03	–
16.031	0.72	–
17.026		$M+NH_4^+$
18.010	2.36	H_2O
18.018	0.92	–
21.986		$M+Na^+$
23.015	0.51	–
23.018	0.72	–
24.000	1.95	–
24.002	0.82	–
26.016	4.21	C_2H_2
26.024	1.95	–
27.996	1.64	CO
28.014	0.62	–
28.032	8.73	C_2H_4
28.064	1.23	–
32.026	1.33	–
40.033	0.72	–
42.009	0.62	$COCH_2$
47.997	0.72	–
48.004	0.62	–
48.024	0.62	–
54.049	0.72	–
56.025	0.62	C_3H_4O
56.052	1.03	–
56.064	5.65	C_4H_8
56.073	1.95	–
56.082	0.92	–
56.105	0.92	–
58.006	0.92	CO_2CH_2
60.020		M+IsoProp+H
67.989		$NaHCO_2$
72.023		OH replacement
73.090	1.13	–

Table S1. Cont.

Obs. mass	% mass shift	Putative element or group
86.037	0.82	–
95.075	1.03	–
112.126	0.62	–

Color coding is as follows: putatively involving oxygen (blue), involving only carbon and hydrogen (red), unknown (white), or adducts or artifacts (gray). Reported are the mass shifts observed in the data (Obs. mass), the percentage of all mass shifts representing that mass shift (% mass shift), and the putative annotated element or group composition.

Table S2. Contaminants that were identified during GNPS networking and removed from the dataset before analyses

Contaminant name (in GNPS)
Glycocholate
Glycocholic acid
HMDB:HMDB 01659-1674 acetone
HMDB:HMDB 03366-2272 propanal
HMDB:HMDB 03366-2273 propanal
Massbank:PB 006044 indole-3-acetonitrile|2-(1H-indol-3-yl)acetonitrile
Massbank:PB 006046 indole-3-acetonitrile|2-(1H-indol-3-yl)acetonitrile
MLS001332546-01 glycocholic acid hydrate 475-31-0
MLS001332642-01 sodium glycocholate hydrate 863-57-0
MS_contaminant_sodium_formate_cluster
Sodium formate

See *SI Materials and Methods* for details about how contaminants were identified.

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (PDF)

R code for Random Forests permutation test

```
library(gdata)
library(randomForest)

# mass shift data file, 'original_data'
# column 1: whole number associated with each mass shift;
 'bin_number' (1,2,3,...n)
# column 2: mass shift; 'mass_shift'
# column 3...n: mass shift data for each sample;
 'sample_1'...'sample_n',

# sample names data file; 'sample_names'
# column 1: sample names from original_data; 'sample'
# column 2: type of sample; 'class'


# read original data
orig.df <- read.xls("original_data.xlsx", stringsAsFactors = FALSE)

# extract bin ids into vector
bin.id <- orig.df$bin_number
# extract mass values and create column names
mod.col <- paste("mass_shift", orig.df$mass_shift, sep = "_")

# create transposed matrix of abundances and rename columns
orig.mat <- t(orig.df[, -(1:2)])
colnames(orig.mat) <- mod.col

# function takes a matrix and vector of bins (as long as number of
 columns in 'mat')
# returns matrix of summed values for each bin
sumBins <- function(mat, bins) {
  sums <- tapply(1:ncol(mat), bins, function(i) {
    bin.mat <- orig.mat[, i, drop = FALSE]
    rowSums(bin.mat, na.rm = TRUE)
  })
  do.call(cbind, sums)
}

# this is the empirical matrix of summed values
orig.bin.mat <- sumBins(orig.mat, bin.id)

# this is one matrix of randomly assigned bins (with bin frequency
 kept constant)
ran.bin.mat1 <- sumBins(orig.mat, sample(bin.id))

# read sample-class assignments
sample.names <- read.xls("sample_names.xlsx", stringsAsFactors =
 FALSE)
rownames(sample.names) <- sample.names$sample
row.class <- factor(sample.names[rownames(orig.bin.mat), "class"])

# the class frequencies do not vary, so calculate them here:
```

```
freq = table(row.class)
n = min(ceiling(freq / 2))
n = max(n, 4)
n = rep(n, length(freq))

# setup a randomForest function for your observed and null models
rfFunc <- function(mat, y, n) {
  randomForest(
    mat, y, proximity = TRUE, importance = TRUE,
    ntree = 10000, sampsize = n, replace = FALSE
  )
}

# run your observed randomForest
obs.rf <- rfFunc(orig.bin.mat, row.class, n)
obs.oob <- obs.rf$err.rate[nrow(obs.rf$err.rate), 1]

# use lapply to collect random forest models for your null
 distribution:
null.rf <- lapply(1:10, function(i) {
  cat(i, "\n")
  rfFunc(sumBins(orig.mat, sample(bin.id)), row.class, n)
})

# get whatever you want from models by looping through null.rf
null.oob.dist <- sapply(null.rf, function(x)
 x$err.rate[nrow(x$err.rate), 1])

hist(null.oob.dist)
abline(v = obs.oob, lty = "dashed", col = "red")
```