Multi-omics profiling of Earth's biomes reveals that microbial and metabolite composition are shaped by the environment

Justin P. Shaffer[1,#], Louis-Félix Nothias[2,3,#], Luke R. Thompson[4,5,#], Jon G. Sanders[6], Rodolfo A. Salido[7], Sneha P. Couvillion[8], Asker D. Brejnrod[3], Shi Huang[1,9], Franck Lejzerowicz[1,9], Holly L. Lutz[1,10], Qiyun Zhu[11,12], Cameron Martino[9,13], James T. Morton[14], Smruthi Karthikeyan[1], Mélissa Nothias-Esposito[2,3], Kai Dührkop[15], Sebastian Böcker[15], Hyun Woo Kim[10], Alexander A. Aksenov[2,3], Wout Bittremieux[2,3,16], Jeremiah J. Minich[10], Clarisse Marotz[1], MacKenzie M. Bryant[1], Karenina Sanders[1], Tara Schwartz[1], Greg Humphrey[1], Yoshiki Vásquez-Baeza[9], Anupriya Tripathi[1,3], Laxmi Parida[17], Anna Paola Carrieri[18], Niina Haiminen[17], Kristen L. Beck[19], Promi Das[1,10], Antonio González[1], Daniel McDonald[1], Søren M. Karst[20], Mads Albertsen[21], Gail Ackermann[1], Jeff DeReus[1], Torsten Thomas[22], Daniel Petras[2,10,23], Ashley Shade[24], James Stegen[8], Se Jin Song[9], Thomas O. Metz[8], Austin D. Swafford[9], Pieter C. Dorrestein[2,3], Janet K. Jansson[8], Jack A. Gilbert[1,10], Rob Knight[1,7,9,25,*], and the Earth Microbiome Project 500 (EMP500) Consortium[δ]

[1]Department of Pediatrics, School of Medicine, University of California San Diego, La Jolla, California, USA.
[2]Collaborative Mass Spectrometry Innovation Center; University of California San Diego; La Jolla, CA 92093; USA
[3]Skaggs School of Pharmacy and Pharmaceutical Sciences; University of California San Diego; La Jolla CA 92093; USA
[4]Northern Gulf Institute, Mississippi State University, Mississippi State, Mississippi, USA
[5]Ocean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, Miami, Florida, USA
[6]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York, USA.
[7]Department of Bioengineering, University of California San Diego, La Jolla, California, USA.
[8]Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA
[9]Center for Microbiome Innovation, Jacobs School of Engineering, University of California San Diego, La Jolla, California, USA.
[10]Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA.
[11]School of Life Sciences, Arizona State University, Tempe, AZ 85281, USA
[12]Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University, Tempe, AZ 85281, USA
[13]Bioinformatics and Systems Biology Program, Jacobs School of Engineering, University of California San Diego, La Jolla, California, USA.
[14]Center for Computational Biology, Flatiron Institute, Simons Foundation

[15]Chair for Bioinformatics, Friedrich Schiller University, Jena, Germany
[16]Department of Computer Science, University of Antwerp, Antwerp, Belgium
[17]IBM Research, T.J. Watson Research Center, Yorktown Heights, NY, USA
[18]IBM Research Europe - Daresbury, UK
[19]IBM Research, Almaden Research Center, San Jose, CA, USA
[20]Department of Virus and Microbiological Special Diagnostics, Statens Serum Institute, Copenhagen, Denmark
[21]Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark
[22]Centre for Marine Science and Innovation, School of Biological, Earth and Environmental Science, The University of New South Wales, Sydney, 2052, Australia
[23]Interfaculty Institute of Microbiology and Infection Medicine, University of Tübingen, Tübingen, Baden-Württemberg 72076, Germany
[24]Department of Microbiology and Molecular Genetics, Michigan State University East Lansing MI USA
[25]Department of Computer Science and Engineering, Jacobs School of Engineering, University of California San Diego, La Jolla, California, USA.
[#]Co-first author
*Corresponding author
[δ]A list of authors and their affiliations appears at the end of the paper

ABSTRACT

As our understanding of the structure and diversity of the microbial world grows, interpreting how it functions is of critical interest for understanding and managing the many systems microbes influence. Despite advances in sequencing, lack of standardization challenges comparisons among studies that could provide global insight. Here, we present a multi-omics, meta-analysis of a novel, diverse set of microbial community samples collected for the Earth Microbiome Project. Standardized protocols and analytical methods for characterizing microbial communities, including assessment of function using untargeted metabolomics, facilitate identification of shared microbial and metabolite features, permitting us to explore diversity at extraordinary scale. In addition to a reference database for metagenomic and metabolomic data, we provide a framework for incorporating additional studies, enabling the expansion of existing knowledge in the form of a community resource that will become more valuable with time. To provide an example of applying this database, we tested the hypothesis that metabolic potential is not limited or enriched by the environment. Our results show that the relative abundances of microbially-related metabolites vary across habitats, and highlight the power of certain chemistry – in particular alkaloids – in distinguishing Earth's environments.

A major goal in microbial ecology is to understand the functional potential of microbial communities, and how this relates to their taxonomic and phylogenetic composition. Microbes produce diverse secondary metabolites that perform vital functions from communication to defense[1-3] and can benefit human health and environmental sustainability[4-10]. To begin to understand the metabolic capabilities of microbial communities, metagenomic sequence data can be profiled for genes related to functional products[11-18]. For example, recent work describing metabolic potential across environments via the detection and classification of biosynthetic gene clusters (BGCs) in microbial genomes (i.e., metagenome-assembled genomes, or MAGs) concluded that specific chemistry is not limited or enriched by the environment, such that most classes of secondary metabolites have cosmopolitan distributions[19]. Metagenome mining is a powerful way to discover function, however chemical evidence is required to confirm the presence of metabolites and accurately describe their distribution across the Earth. Here, we present an approach that directly assesses the presence and relative abundance of metabolites that can provide a more accurate description of metabolic potential in microbial communities across the Earth's environments.

Importantly, gene-level data about BGCs inferred from MAGs fundamentally cannot provide information about actual synthesis (e.g., gene expression, enzyme activity and substrate availability), and cannot be used to infer the distribution of secondary metabolites in natural communities[20-23]. Even with high-coverage gene expression data, currently lacking for most environments, the complex structural- and modular nature of many secondary metabolites prevents their accurate association with underlying genomic elements[24]. Furthermore, quantifying metabolite diversity from metatranscriptomic and/or metaproteomic data – also

lacking for most environments – is problematic due to a suite of post-translational processes that can dissociate the level of gene transcription from the abundance of gene products[25]. Similarly, gene products can be inhibited, otherwise altered, or consumed rapidly, preventing their accumulation in the environment. Finally, shotgun metagenomics does not capture BGCs from lowly abundant MAGs efficiently, as shown by comparison with targeted sequencing approaches[18].

An approach to surmount these issues is to complement metagenomics with a direct survey of secondary metabolites using untargeted metabolomics. Liquid chromatography with untargeted tandem mass spectrometry (LC-MS/MS) is a versatile method that detects tens-of-thousands of metabolites in biological samples[26]. Although LC-MS/MS metabolomics has historically suffered from low metabolite annotation rates when applied to non-model organisms, recent computational advances can systematically assign chemical classes to metabolites using their fragmentation spectra[27]. Untargeted metabolomics provides the relative abundance (i.e., intensity) of each metabolite detected across samples rather than just counts of unique structures (i.e., presence/absence data), and thus provides a direct readout of the surveyed environment, complementing a purely genomics-based approach. Although there is a clear need to use untargeted metabolomics to quantify the metabolic activities of microbiota, this methodology has been limited by the challenge of distinguishing the secondary metabolites produced exclusively by microbes from other metabolites detected in the environment (e.g., those produced by multicellular hosts). To resolve this bottleneck, we devised a computational method for recognizing and annotating putative secondary metabolites of microbial origin from fragmentation spectra. The annotations were first obtained from spectral library matching and *in*

silico annotation[28] using the GNPS web-platform[29]. These annotations were then queried against microbial metabolite reference databases (i.e., Natural Products Atlas[30] and MIBiG[31]), and molecular networking[32] was used to propagate the annotation to similar metabolites. Finally, a global chemical classification of these metabolites was achieved using a state-of-the-art annotation pipeline (i.e., SIRIUS)[27].

We used this methodology to quantify microbial secondary metabolites from diverse microbial communities from the Earth Microbiome Project (EMP, http://earthmicrobiome.org). The EMP was founded in 2010 to sample Earth's microbial communities at unprecedented scale, in part to advance our understanding of biogeographic processes that shape community structure. To avoid confusion with terminology, we define 'microbial community' as consisting of members of Bacteria and Archaea. To build on the first meta-analysis of the EMP archive focused on profiling bacterial and archaeal 16S ribosomal RNA (rRNA)[33], we crowdsourced a new set of roughly 900 samples from the scientific community specifically for multi-omics analysis. As a result of expanding the scalable framework of the EMP to include standardized methods for shotgun metagenomic sequencing and untargeted metabolomics for cataloging microbiota globally, we provide a rich resource for addressing outstanding questions and to serve as a benchmark for acquiring new data. To provide an example for using this resource, we present a multi-omics, meta-analysis of this new sample set, tracking not just individual sequences but also genomes and metabolites, across diverse studies with sample types classified using an updated and standardized environmental ontology, describing large-scale ecological patterns, and exploring important hypotheses in microbial ecology.

RESULTS

An invaluable resource for a meta-analytical- and multi-omics approach to microbial ecological research. Here, we generated data for 880 environmental samples that span 20 major environments contributed by 34 principal investigators of the Earth Microbiome Project 500 (EMP500), a novel sample set for multi-omics protocol development and data exploration (Figure 1; Table S1). To normalize sample collection for this- and future studies, while following standardized protocols for sample collection, sample tracking, sample metadata curation, sample shipping, and data release, we updated and followed the existing Earth Microbiome Project (EMP) Sample Submission Guide (https://earthmicrobiome.org/protocols-and-standards/emp500-sample-submission-guide/)[34], which we highlight here to encourage its use. Importantly, we updated the previous EMP Metadata Guide to accommodate the EMP500 sampling design as well as updates to other standardized ontologies (see Online Methods), including our own application ontology, the Earth Microbiome Project Ontology (EMPO). In addition to the environments previously described by EMPO, EMPO (version 2) now recognizes an important split within host-associated samples representing saline and non-saline environments (Fig. 1a) not detected in the EMP's previous analysis of 16S rRNA from a separate set of <23,000 samples[33].

For the majority of samples examined here, we successfully generated data for bacterial and archaeal 16S ribosomal RNA (rRNA), eukaryotic 18S rRNA, internal transcribed spacer (ITS) 1 of the fungal ITS region, bacterial full-length rRNA operon, shotgun metagenomics, and untargeted metabolomics (i.e., LC-MS/MS and gas chromatography coupled with mass spectrometry [GC-MS]). A summary of sample representation among each data layer, including

read counts for sequence data, is presented in Table S2. To foster exploration of this novel dataset, we have made the raw sequence- and metabolomics data publicly available through Qiita (https://qiita.ucsd.edu; study ID: 13114)[35] and GNPS (https://gnps.ucsd.edu; MassIVE IDs: MSV000083475, MSV000083743)[29], respectively. We also provide complete protocols for laboratory- and computational workflows for both metagenomics and metabolomics data for use by the broader community, available on GitHub (https://github.com/biocore/emp/blob/master/methods/methods_release2.md). We hope that the dataset and workflows presented here serve as useful tools for others, in addition to providing a podium for launching additional future studies. To provide an example of the utility of the dataset for addressing important questions in microbial community ecology, we present an analysis of microbially-related metabolites and microbe-metabolite co-occurrences across the Earth's environments. Specifically, we explore the hypothesis that metabolic potential can be limited or enriched by the environment, and that although most groups of metabolites have cosmopolitan distributions[19], the relative abundance of metabolites will vary among environments. Thus, we predict that whereas the presence/absence of metabolites alone will show profiles that are relatively uniform across environments, their relative intensities will provide great power in distinguishing among environments. Furthermore, we expect variation in metabolite profiles among environments to be in part driven by variation in microbial community composition.

Relative abundance vs. presence/absence highlights enrichment of certain metabolites in particular environments. In total, we generated untargeted metabolomics data (i.e., LC-MS/MS)

for 618 of 880 samples (Table S2), resulting in 52,496 unique molecular structures, or metabolites, across all samples. We then refined that dataset to include only putative, microbially-related metabolites, resulting in 6,588 metabolites across all samples, which represent 12.55% of the original data. Focusing on this subset, we found that although the presence/absence of major classes of microbially-related metabolites is indeed relatively conserved across habitats, their relative abundance reveals specific chemistry that is lacking or enriched in particular environments, especially when considering more resolved chemical class ontology levels (Fig. 2).

Importantly, when considering differences in the relative abundances of all microbially-related metabolites, profiles among environments were so distinct that we could identify particular metabolites whose abundances were significantly enriched in certain environments (Fig. 3a, Table S3). For example, microbially-related metabolites associated with the carbohydrate pathway were especially enriched in aquatic samples, whereas those associated with the polyketide- and shikimate and phenylpropanoid pathways enriched in sediment, soil, plant detritus, and fungal samples (Fig. 3a). These patterns are likely driven in part by inherent differences in overall alpha-diversity of microbially-related metabolites across environments, and we note that whereas the aquatic sediment samples are most rich (also true for microbial taxon richness), non-saline samples from plant surfaces are uniquely depauperate in metabolite diversity (Fig. 3b). When also considering the identity and relative intensity of each metabolite, we observed a separation of samples based in part on salinity and host-association, with specific environments clustering in ordination space and certain metabolite features identified as globally important (Fig. 3c, Table S4). Somewhat unexpectedly, we observed the majority of saline

environments to cluster between two groups of non-saline ones when considering the first two axes, separating the soil from the 'Animal distal gut' samples (Fig. 3c). Similarly, whereas the split in salinity is strong for host-associated environments, it is less so for free-living ones, with both water and sediment samples exhibiting similarity with respect to metabolite composition (Fig. 3c).

Metabolite- and microbial diversity exhibit strong but environment-specific relationships. With these observations, we predicted that the differences in the relative abundances of particular metabolites among environments were due in part to underlying differences in microbial community composition and diversity. To explore these relationships, we found strong correlations in our shotgun metagenomic data between microbially-related metabolite richness and microbial taxon richness for certain environments (i.e., Animal proximal gut (saline) r = 0.73, p-value < 0.01; Plant detritus (non-saline) r = 0.74, p-value < 0.001; Sediment (non-saline) r = 0.42., p-value = 0.05; Water (saline) r = 0.57, p-value = 0.01) (Fig. 3b; Table S5). As expected, sediment samples had the highest alpha-diversity of both metabolites and microbial taxa, with sediments from saline environments exhibiting relatively greater metabolite diversity than those from non-saline environments (Fig. 3b). Correlations between metabolite richness and phylogenetically informed metrics for estimating alpha-diversity from our shotgun metagenomics data were weaker (Table S5).

When considering the identity and relative abundance (i.e., intensity for metabolites), we also found similarity in the clustering of samples by environment between datasets (Fig. 3c,d), and a strong correlation between sample-sample distances based on microbially related

metabolites vs. microbial taxa (Table 1). With the exception of certain animal-associated samples (e.g., Animal corpus [non-saline], Animal proximal gut [non-saline], the distribution of environments in ordination space is nearly identical between the metabolomics- and microbial taxon data (Fig. 3c,d). There also appears to be a more consistent split between the saline and non-saline samples in the microbial taxon data as compared to the metabolomics data (Fig. 3c,d). Similarly, the first two principal components of the metabolite ordination explain nearly three-times as much variation as the first two principal coordinates of the microbial taxon ordination (Fig. 3c,d). For both datasets, we observed a decay in the separation of samples across environments as the resolution of EMPO becomes more fine, with the largest effect sizes exhibited by differences based on host-association (i.e., EMPO 1, Fig. 3c,d).

The presence of certain metabolites and microbes is predictive of particular environments. Based on the strong relationships between metabolites, microbial taxa, and the environment, we next tested the hypothesis that specific metabolites, microbial taxa, or microbial functional products (i.e., Gene Ontology terms) could be used to classify samples among environments. Using a machine-learning approach, we were able to identify specific metabolites that could classify particular environments with 68.9% accuracy (Fig. 4a, Fig. S1, Table S6). We also found strong support among methods for the importance of particular metabolites in distinguishing environments. For example, one putative cholane steroid (i.e., $C_{24}H_{38}O_4$; ID: 25552) identified in this analysis also stood out in our analysis of differential abundance and our ordination. Similarly, three metabolites enriched in non-saline soil and plant detritus samples – one phenylpropanoid (ID: 12599), one pseudoalkaloid (ID: 58021), and one histidine alkaloid (ID:

34551), also stood out in our ordination (Fig. 3c, Fig. 4a, Table S4, Table S6). Interestingly, distinct analytical approaches identified specific metabolites as particularly important for distinguishing aquatic samples ($C_{28}H_{58}O_{15}$, pathway: carbohydrates, superclass: glycerolipids), non-saline plant surface samples ($C_{13}H_{10}O$, pathway: shikimates and phenylpropanoids, superclass: flavonoids, class: chalcones), and non-saline animal distal gut samples ($C_{24}H_{38}O_4$, pathway: terpenoids, superclass: steroids, class: cholane steroids) (Fig. 3a,c, Table S3, Table S4). We explore these relationships further in our multi-omics analyses.

Using the same machine learning approach on our metagenomic sequence data, we identified specific microbial taxa and microbial functional products that could classify particular environments with 72.1% and 68.2% accuracy, respectively (Figs. 5, Fig. S1). With respect to microbial taxa, we note that the presence of *Conexibacter woesei* (Actinobacteria) was highly predictive of non-saline soil samples, as were several members of *Bradyrhizobium* (Alphaproteobacteria), including *B. lablabi* and *B. valentinum* (Fig. 5). Interestingly, we recovered sequences mapping to *Candidatus* Entotheonella TSY1 (Tectomicrobia), a known filamentous symbiont of marine sponges, and which was highly predictive of marine sponges (i.e., Animal corpus (saline) and Animal secretion (saline) (Fig. 5). We also note that taxa most predictive of aquatic environments – both saline and non-saline – included *Ca.* Pelagibacter ubique (Alphaproteobacteria), as well as members of the Rhodobacteriaceae (Alphaproteobacteria), the latter of which is widely known for its chemo- and photoautotrophic members (Fig. 5). With respect to microbial functions, we note that the majority of predictive functional products were GO terms for molecular function as opposed to those for biological processes, which were second most common, followed lastly by those for cellular components

(Fig. 5). Across the board, the GO term for DNA-templated transcription was the strongest predictor, and was most associated with non-saline soils and sediments, followed by saline soils and sediments (Fig. 5). In general, whereas both the prevalence and mean relative abundance of functional products across environments appears somewhat uniformly high, which is unique from the patterns for metabolites and microbial taxa, a few particular environments are severely lacking in each product across the board (Fig. 5).

Multi-omics analysis reveals strong correlations between metabolites, microbes, and the environment. In addition to exploring relationships between metabolite and microbial diversity, we sought to explicitly quantify metabolite-microbe co-occurrence patterns, especially associations between metabolites and the environment (e.g., Fig. 3a,c), while also considering each metabolite's co-occurrence with all microbes in the dataset. In that regard, we first generated metabolite-microbe co-occurrences learned from both LC-MS/MS metabolomics- and shotgun metagenomic profiles across all samples, for a cross-section of 6,501 microbially-related metabolites and 4,120 microbial genomes (i.e., OGUs). Then, using metabolite-metabolite distances based on global co-occurrence profiles considering all microbial taxa, we ordinated metabolites in microbe space. We then examined correlations between metabolite loadings on each principal coordinate and (1) log-fold changes of metabolites across environments, and (2) global distributions of metabolites, and found strong relationships with each (Figure 4b). In particular, the abundances of microbially-related metabolites in plant surface (non-saline), sediment (saline), and aquatic samples (i.e., those from saline and non-saline water) had strong correlations with microbe-metabolite co-occurrences (Figure 4b,c). Visualizing the differential

abundance of metabolites with respect to water (non-saline) samples in the broader context of these co-occurrences highlights the especially unique community profiles in that environment (Fig. 4c). In addition to highlighting such environments vs. other free-living and host-associated habitats, this demonstrates that microbes and metabolites can be classified by- and co-occur among environments.

Harnessing additional data generated for EMP500 samples, including GC-MS and amplicon sequence (i.e., 16S, 18S, ITS, and full-length rRNA operon) data, we compared sample-sample distances between each pair of datasets (Table 1). The weakest relationships were between the GC-MS metabolomics profiles and those from eukaryotic (i.e., 18S rRNA and fungal ITS) sequence data (Table 1). The strongest correlations we observed were between sequence data from Bacteria and Archaea (Table 1). Interestingly, the relationship between 16S rRNA profiles and those from (1) full-length rRNA operons, and (2) shotgun metagenomics, were equally strong, followed closely by the relationship between shotgun metagenomics profiles and those from full-length rRNA operons (Table 1). Importantly, we found further support for a strong relationship between bacterial- and archaeal communities and metabolic profiles (Table 1). Interestingly, we found the relationships between the metabolomics data (i.e., LC-MS/MS and GC-MS) and the sequence data from Bacteria and Archaea (i.e., shotgun metagenomics, 16S, and full-length rRNA operon) to be much stronger than those between metabolomics data and sequence data from eukaryotes (i.e., 18S and ITS). Excluding relationships between any two sets of microbial sequence data, the strongest relationship we observed was that between microbially-related metabolites (i.e., LC-MS/MS data) and microbial genomes (i.e., shotgun metagenomics data for taxonomy) (Table 1). In addition to supporting

that specific chemistry is limited or enriched by the environment, these results highlight the strong relationship between metabolic profiles and microbial taxonomic composition across habitats spanning the globe.

DISCUSSION

Here, we produced as a resource a novel, multi-omics dataset comprising 880 samples that span 19 major environments, contributed by 34 principal investigators for the EMP500 (Fig. 1, Table S1). While we updated EMPO to include these 19 environments, we recognize that certain environments are represented here by only a handful of samples (Fig. 1), and note that we had to exclude them from some of our analyses (e.g., machine learning and co-occurrence analyses). In that regard, we recommend that future similar efforts focus on these environments, in order to further generalize our findings to those habitats. Similarly, we hope to expand sampling geographically to expand our scope of inference, as many important environments and locations could not be included here (or, indeed, in the EMP's 27,000 sample dataset[33]). To foster this research activity, we expanded upon the widely-adopted set of the EMP's standardized protocols for guiding microbiome research – from sample collection to data release[34] – with new protocols for performing untargeted metabolomics and shotgun metagenomics across a diversity of sample types (Fig. 1a; Online Methods).

Across all 880 samples, we generated eight layers of data, including untargeted metabolomics and shotgun metagenomics (Table S2), providing a valuable resource for both multi-omics and meta-analyses of microbiome data. As expected, sample dropout in any one data layer was non-negligible (Table S2), reducing the number of samples in any one layer to at most

roughly 500 samples (hence the EMP500). For future similar studies, we note that considering multi-omics applications during the experimental design and/or sample collection phases of studies is crucial, in part because certain metabolomics approaches are not amenable to samples stored in particular storage solutions commonly used for metagenomics (e.g., RNA later).

We also included an example of how to apply this dataset towards addressing important questions in microbial ecological research, by describing the Earth's metabolome using an integrated ‘omics approach. Our results show that the presence of certain classes of secondary metabolites can be limited- or enriched by the environment (Fig. 3a,c, Table S3). Importantly, our data indicate that considering the relative abundances of secondary metabolites (i.e., rather than only presence/absence) strengthens differences in metabolite profiles across environments (Fig. 2). We note that the uniform distributions of metabolite classes across environments based on presence/absence data (Fig. 2a,c) are similar to those previously observed based on BGC annotation of a global dataset of MAGs[19], suggesting that presence/absence data alone are not able to accurately depict chemistry in nature. Importantly, the groups of metabolites driving the most obvious changes in overall patterns when considering intensities are those that appear in few samples (i.e., those at low prevalence from a presence/absence perspective; e.g., carbohydrates, alkaloids, Fig. 2a,b). Many metabolite superclasses — including apocarotenoids, fatty acids and conjugates, glycerolipids, steroids, and polyketides — differed in distribution between environments, suggesting that changes in metabolite intensity adds information above and beyond that gained by presence/absence data alone. Interestingly, we note that the environments with the most unannotated metabolites include terrestrial animal cadavers, ruminant-associated bioreactors, and freshwater and marine habitats (Fig. 2b). These

environments merit further attention with respect to functional description, and represent valuable opportunities for the discovery of novel metabolites.

We also showed that the richness and composition of metabolites in any given sample vary with the surrounding environment (Fig. 3b,c,d, Table S5), and only partially reflect those of co-occurring microbial communities. To our knowledge, this is the first assessment of the metabolic alpha-diversity in marine sediments as it relates to microbial diversity in those samples in a global context including other diverse environments, and we note that in addition to confirming that indeed marine sediments are one of the most microbially diverse environments on Earth, they also appear to be the most metabolically diverse (Fig. 3b). When considering beta-diversity, we note a much weaker influence of salinity in separating samples based on metabolites vs. microbes, indicating that although microbial taxa respond strongly to this gradient, microbial communities in saline vs. non-saline habitats have similar potential (Fig 3c,d). This is supported by our finding that the differential abundance (i.e., intensity) of important pathways and superclasses is highly similar between freshwater and marine environments (Fig 3a).

We also identified several metabolites, microbial taxa, and microbial functions that could accurately classify samples among environments (Fig. 4a, Fig. 5, Table S6). Across the three datasets, we observed much greater variation in the overall prevalence and mean relative abundance of predictive metabolites and microbial taxa across environments as compared to microbial functions (Fig. 4a, Fig. 5). Similarly, the overall predictive accuracy for microbial taxa was greater than for microbial functions. These observations support previous findings that redundancy of microbial functions within and across ecosystems is greater than for microbial

taxa[36-38]. We note that when considering the top 32 most important metabolites regarding predictive power, metabolites classified as carbohydrates and polyketides are not present (Fig. 4a), although the former were enriched- and differentially intense in both freshwater and marine samples (Fig. 2a, Fig 3a). Rather, the most abundant pathway among those highly predictive metabolites were alkaloids, highlighting the importance of this group of BGCs in distinguishing Earth's environments (Fig. 4a).

In addition to showing that the relative abundances of microbially-related metabolites varies across environments, our results highlight the advantages of using a multi-omics approach including metabolomics to interpret and predict the contributions of microbes and their environments to chemical profiles in nature. Whereas similar approaches have been used to gain novel insight previously, to our knowledge this is its first application to a global dataset encompassing a diversity of environmental sample types. Furthermore, the majority of studies reported as performing multi-omics combine two or more sequencing based approaches, such as metagenomics with metatranscriptomics[39-41], with few incorporating metaproteomics[42-44], and even fewer untargeted metabolomics[45-47]. We argue that using only sequence-based approaches to interpret functional potential can be misleading, as the presence of genomic loci and/or transcripts does not equate to the presence of a functional product in the environment. For example, a recent analysis of BGCs annotated from a global database of microbial MAGs concluded that the relative distribution of secondary metabolite biosynthetic potential across environments was conserved, implying that most classes of secondary metabolites are not limited- or enriched by the environment[19]. Using our presence/absence data for metabolites, we observed a similar trend (Fig. 2a,c). However, when taking into account the relative intensities of

metabolites – to date only possible using proteomics or metabolomics – we observed significant differences in the distribution of particular groups of metabolites across the Earth's environments (Fig. 2b,d, Fig. 3a). This emphasizes the utility and importance of directly measuring functional products in the environment, rather than estimating their potential from underlying genomic elements.

Moreover, our approach illustrates that recent advances in computational annotation tools offer a powerful toolbox to interpret untargeted metabolomics data[27]. We anticipate that advances in genome-mining will improve the discovery and classification of BGCs from MAGs and provide additional insight into these findings, and by making these data publicly available in Qiita and GNPS our data will provide an important resource for continued collaborative investigations. In the same manner, the development of novel instrumentation and computational methods for metabolomics will expand the depth of metabolites surveyed in microbiome studies.

Tables.

Table 1. Mantel test results comparing data layers generated for the EMP500 samples. Note the strong relationships between the metabolomics data (i.e., LC-MS/MS and GC-MS) and the sequence data from Bacteria and Archaea (i.e., shotgun metagenomics, 16S, and full-length rRNA operon)as compared to between metabolomics data and sequence data from eukaryotes (i.e., 18S and ITS). There are also strong relationships between difference sequence data from Bacteria and Archaea (rho > 0.2 in bolded font; > 0.4 in bolded, italics; > 0.5 additionally underlined).

Dataset 1	Dataset 2	*n*	Spearman rho	p-value
LC-MS/MS	GC-MS	401	0.13	0.001
	Metagenomics (taxa)	454	***0.43***	0.001
	Metagenomics (function)	378	**0.24**	0.001
	16S	477	**0.27**	0.001
	18S	356	0.06	0.034
	ITS	359	0.08	0.001
	full-length rRNA operon	181	**0.34**	0.001
GC-MS	Metagenomics (taxa)	331	0.07	0.002
	Metagenomics (function)	279	0.05	0.029
	16S	349	**0.22**	0.001
	18S	284	-0.01	0.641
	ITS	258	0.02	0.404
	full-length rRNA operon	168	0.11	0.001
Metagenomics (taxa)	Metagenomics (function)	508	**0.24**	0.001
	16S	538	***0.51***	0.001
	18S	378	-0.04	0.159
	ITS	408	0.05	0.023
	full-length rRNA operon	235	***0.48***	0.001
Metagenomics (function)	16S	449	0.11	0.001
	18S	305	0.12	0.001
	ITS	337	0.14	0.001
	full-length rRNA operon	219	**0.21**	0.001
16S	18S	430	0.04	0.117
	ITS	447	0.09	0.001
	full-length rRNA operon	215	***0.51***	0.001
18S	ITS	387	**0.20**	0.001
	full-length rRNA operon	174	0.06	0.06
ITS	full-length rRNA operon	166	0.09	0.001

Figures.

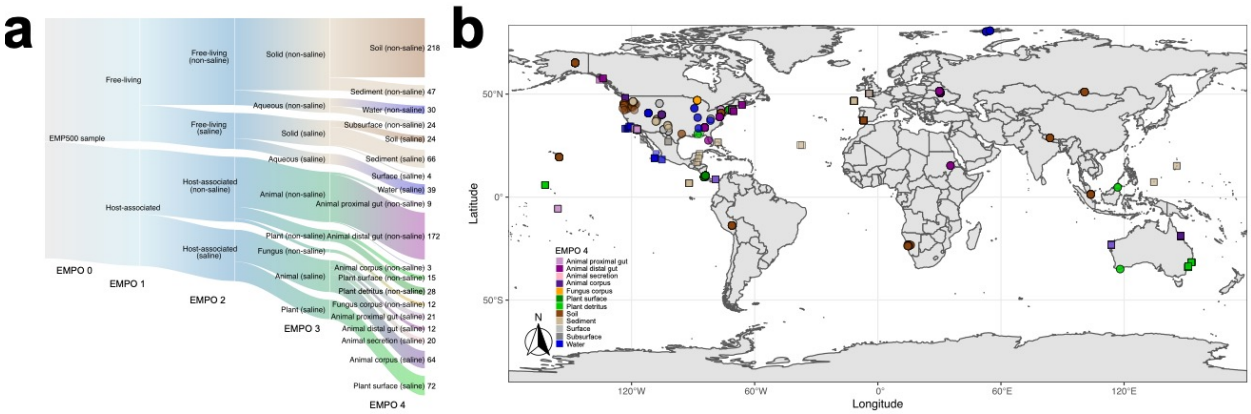

Figure 1 | a, Distribution of samples among the Earth Microbiome Project Ontology (EMPO version 2) categories. This version of EMPO is updated to include the important distinction between saline- and non-saline samples from host-associated environments that was not detected in previous analyses of 16S rRNA alone. **b**, Geographic distribution of samples with points colored by EMPO (version 2, level 4). Points are transparent to highlight cases where multiple samples derive from a single location. We note here that our intent was to sample across environments rather than geography, in part as we previously showed that microbial community composition is more influenced by the former vs. the latter[33], but also to motivate finer-grained geographic exploration as sample analyses decrease in cost. Extensive information about each sample set is described in Table S1.

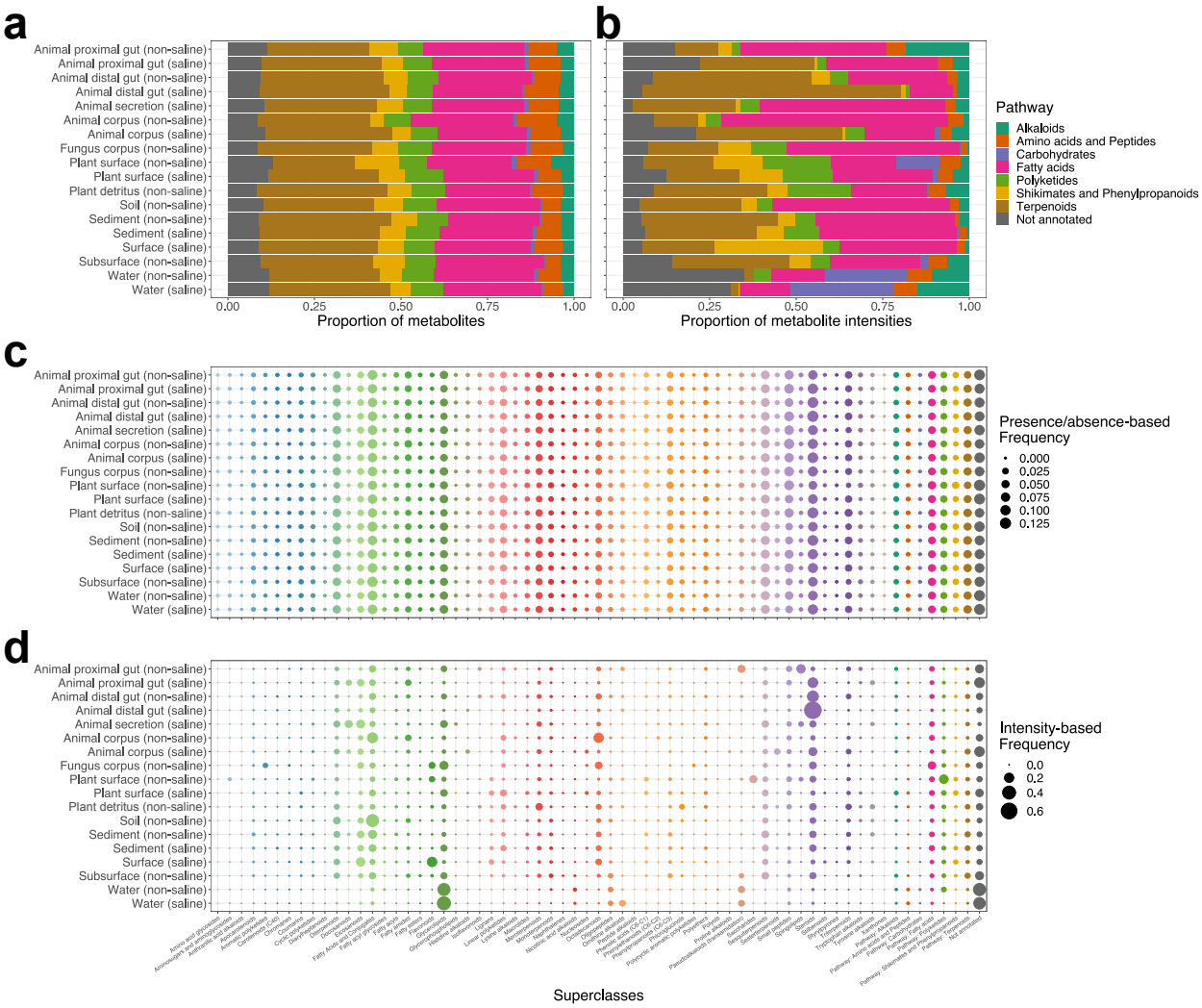

Fig. 2 | Distribution of microbially-related secondary metabolite pathways and superclasses among environments described using the Earth Microbiome Project Ontology (EMPO version 2, level 4). Individual metabolites are represented by their higher level classifications. Both chemical pathway and chemical superclass annotations are shown based on presence/absence (**a, c**) and relative intensities (**b, d**) of molecular features, respectively. For superclass annotations in panels **c** and **d**, we included pathway annotations for metabolites where superclass annotations were not available, when possible.

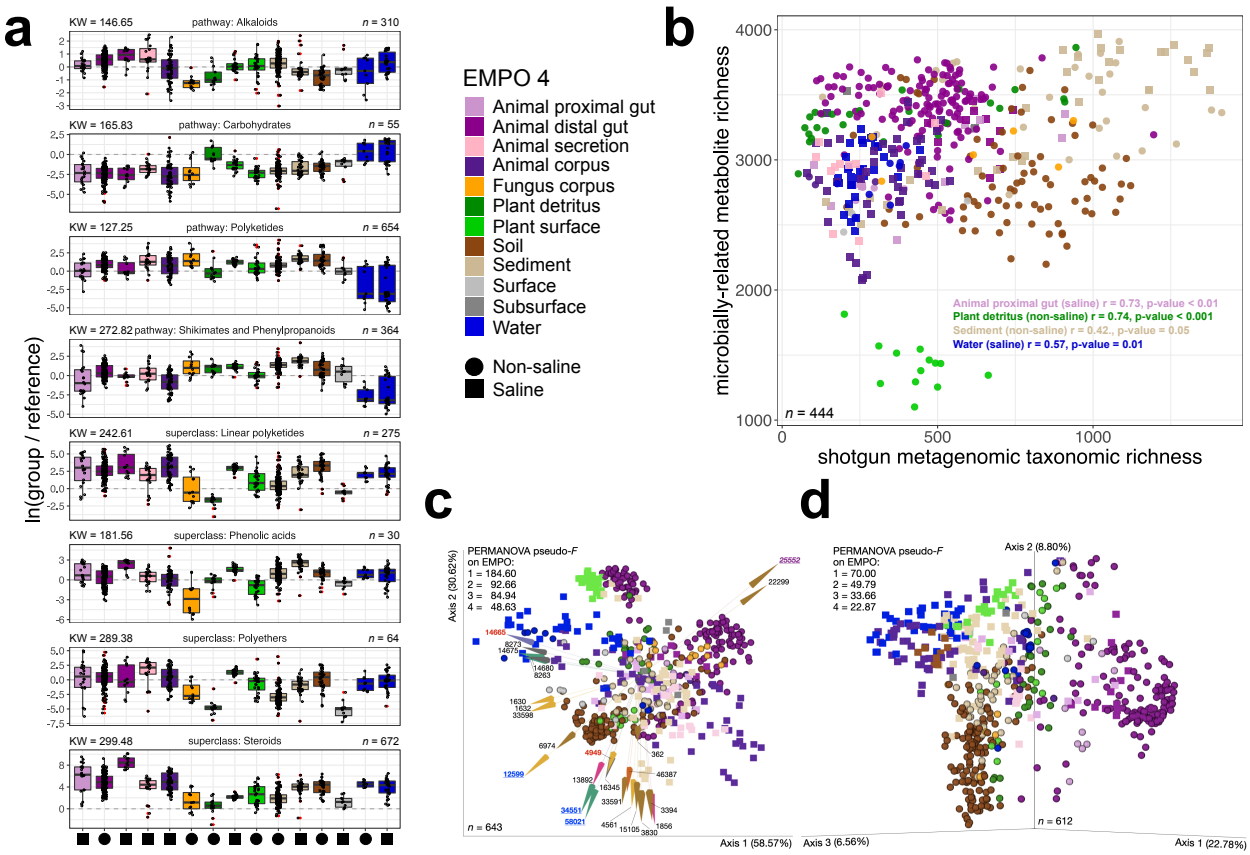

Fig. 3 | Structural-level associations between microbially-related secondary metabolites and specific environments described using the Earth Microbiome Project Ontology (EMPO version 2, level 4). **a,** Differential abundance of molecular features across environments, highlighting four pathways and four superclasses in separate panels. For each panel, the y-axis represents the natural log-ratio of the intensities of metabolites annotated as the listed ingroup divided by the intensities of metabolites annotated as the reference group (i.e., pathway reference: *Amino Acids and Peptides*, $n = 615$; superclass reference: *Flavonoids*, $n = 42$). The number of metabolites in each ingroup is shown, as well as the chi-squared statistic from a Kruskal-Wallis rank sum test for differences in log-ratios across environments (i.e., each test had p-value < 2.2E-16). Each test included 606 samples. Outliers from boxplots are colored red to highlight that they are also represented in the overlaid, jittered points. Associations between molecular features and

environments were identified using Songbird multinomial regression (model: *composition* = EMPO version 2, level 4; pseudo-Q^2 = 0.21). Additional information about each feature is described in Table S2. **b,** Relationship between microbially-related metabolite richness and microbial taxon richness across samples and environments, with significant relationships noted. Correlations with metabolite richness were weaker when using Faith's PD and weighted Faith's PD for quantifying microbial alpha-diversity (Table S5). **c,** Turnover in composition of microbially-related secondary metabolites across environments, visualized using Robust Aitchison Principal Components Analysis (RPCA) showing samples separated based on LC-MS/MS spectra. Shapes represent samples and are colored and shaped by EMPO. Arrows represent metabolites, and are colored by chemical pathway. The direction and magnitude of each arrow corresponds to the strength of the correlation between the relative abundance (i.e., intensity) of the metabolite it represents and the RPCA axes. Samples close to arrow heads have strong, positive associations with respective features, whereas samples at and beyond arrow origins strong, negative associations. The 25 most important metabolites are shown and are described in Table S3. Features annotated in red were also identified in our multinomial regression analysis as among the top 10 ranked metabolites per environment (Tables S2), those in blue also separated environments in machine-learning analysis (Table S4), and those in purple identified as important in all three analyses. **d,** Turnover in composition of microbial taxa across environments, visualized using Principal Coordinates Analysis (PCoA) of weighted UniFrac distances. Distances are based on counts of microbial genomes (i.e., OGUs) from mapping metagenomic reads to the Web of Life database. Sample sizes in **a** indicate metabolites, and all other panels indicate samples. Note the similarities between panels **c** and **d** with respect to the

separation of free-living (e.g., 'Water') and host-associated (e.g., 'Animal distal gut') environments along Axis 1, and a gradient from living hosts (e.g., 'Plant surface'), to detritus (e.g., 'Plant detritus), to soils and sediments along Axis 2. Results from PERMANOVA for each level of EMPO are shown (all tests had *p*-value = 0.001).

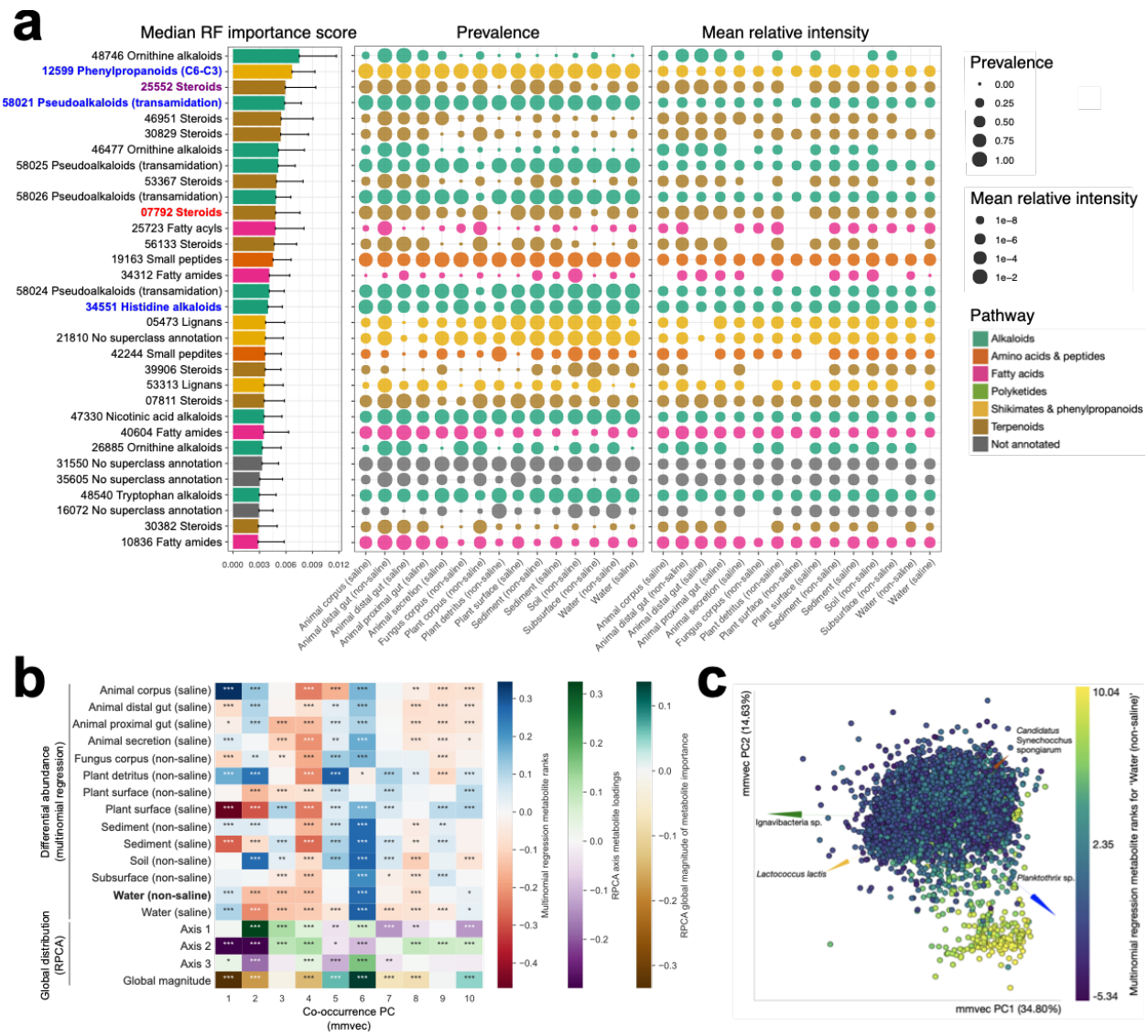

Fig. 4 | Microbially-related metabolites classify environments with a high accuracy and co-occur with specific microbial taxa. a, The random forest (RF) importance score, environment-wide prevalence, and mean relative intensity for the top 32 most important microbially-related

metabolites contributing to the separation of environments. Metabolites are further described in Table S4. Those in red are those also identified in our multinomial regression analysis as among the top 10 ranked metabolites per environment (Tables S2), those in blue also identified to be strongly associated with RPCA axes (Fig. 3c, Table S3). **b,** Co-occurrence analysis results showing correlation between *mmvec* principal coordinates (PCs) and (i) multinomial regression betas for metabolite intensities across environments, (ii) principal components from the RPCA biplot in Fig. 3d corresponding to clustering of samples by environment (i.e., EMPO version 2, level 4) based on metabolite profiles, and (iii) a vector representing the global magnitude of microbial taxon abundances from the RPCA biplot in Fig. 3d. Values are Spearman correlation coefficients. Asterisks indicate significant correlations (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). **c,** The relationship between multinomial regression betas for metabolites with respect to 'Water (non-saline)' and the first three *mmvec* PCs shown as a multi-omics biplot of metabolite-microbe co-occurrences learned from microbial profiles. Points represent metabolites separated by their co-occurrences with microbial taxa. Metabolites are colored based on multinomial regression betas for their intensities with respect to 'Water (non-saline)'. Vectors represent specific microbial taxa strongly associated with ordination axes. The model predicting metabolite-microbe co-occurrences was more accurate than one representing a random baseline, with a pseudo-Q^2 value of 0.18, indicating much reduced error during cross-validation.

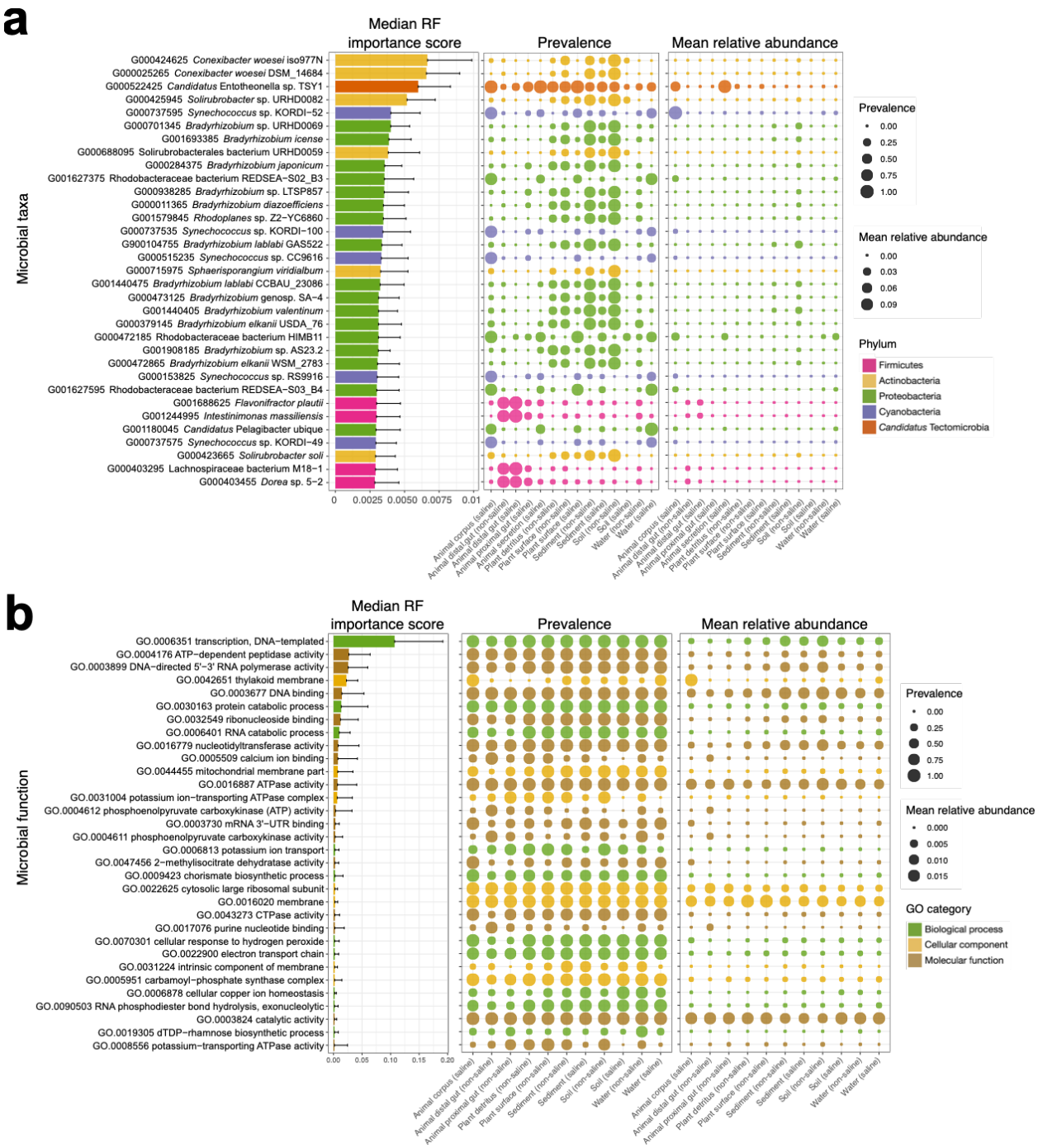

Figure 5 | Machine-learning analysis of shotgun metagenomics data highlighting the most important microbial taxa and functions. **a,** The random-forest (RF) importance score, environment-wide prevalence, and mean relative abundance for the top 32 most important microbial taxa contributing to the separation of environments. **b,** The RF importance score,

environment-wide prevalence, and mean relative abundance for the top 32 most important microbial functions (i.e., GO Terms) contributing to the separation of environments. For both analyses, environments are described by the Earth Microbiome Project Ontology (EMPO version 2, level 4). All samples were analyzed using 20-time repeated group five-fold cross validation, where 'group' indicates the study identifier that a sample belongs to. In this approach, we split the data by 'group' such that all samples from the same study identifier can be assigned to either train or test data sets.

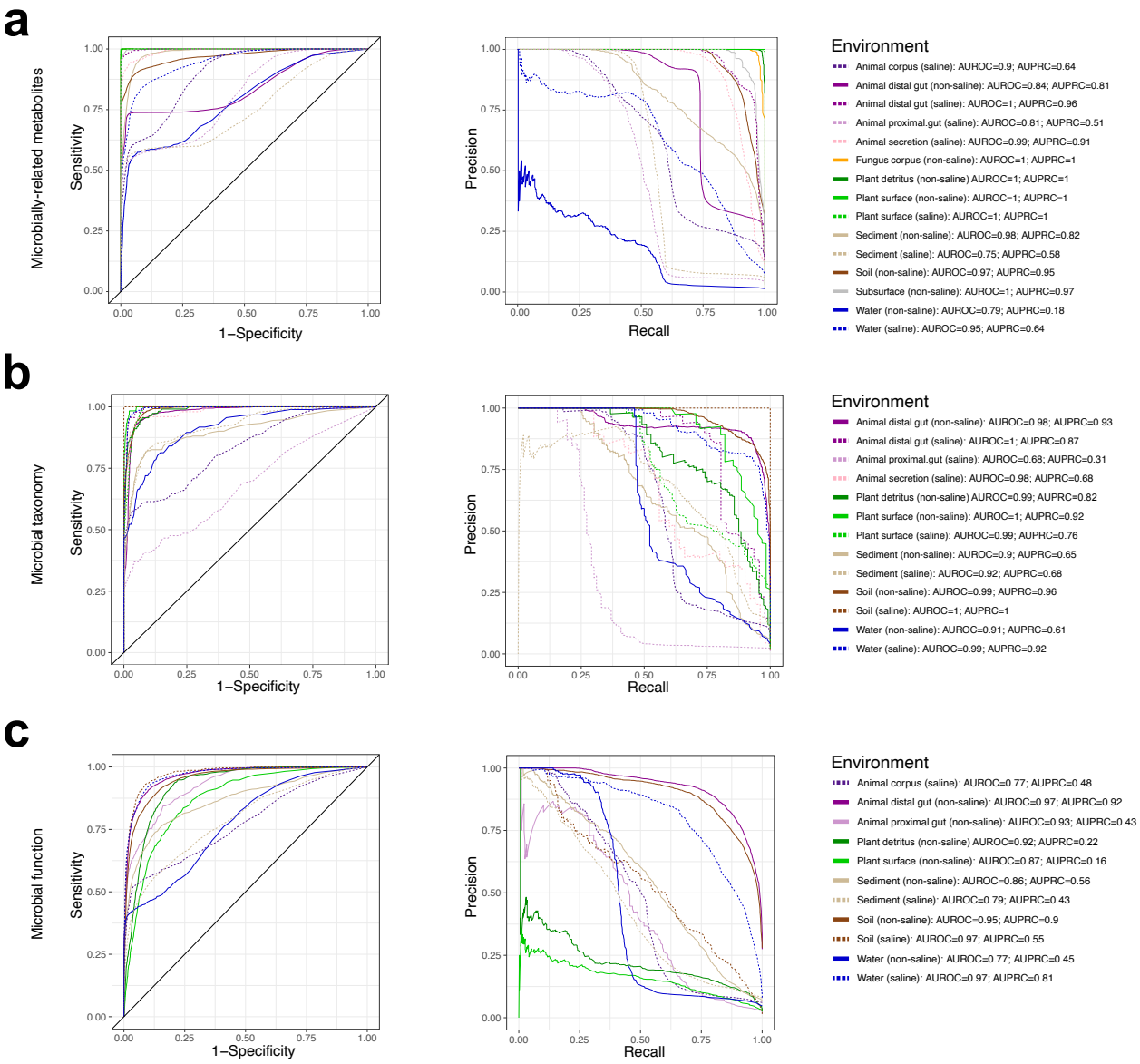

Figure 6 | Machine learning analysis of LC-MS/MS metabolomics and shotgun metagenomics data highlighting per-environment classification accuracy. a, Receiver operating characteristic curves (and AUROC) and precision-recall curves (and AUPRC) illustrating classification accuracy of the random forest model across all environments based on **(a)** microbially-related metabolite profiles, **(b)** microbial taxonomic (i.e., OGU) profiles, and **(c)** microbial functional (i.e., GO terms) profiles. For all three analyses, environments are described by the Earth Microbiome Project Ontology (EMPO version 2, level 4). All samples were

analyzed using 20-time repeated group five-fold cross validation, where ‘group’ indicates the study identifier that a sample belongs to. In this approach, we split the data by ‘group’ such that all samples from the same study identifier can be assigned to either train or test data sets.

References.

1. Davies, D. G. *et al.* The involvement of cell-to-cell signals in the development of a bacterial biofilm. *Science* **280**, 295-298. (1998).
2. Hibbing, M. E. *et al.* Bacterial competition: surviving and thriving in a microbial jungle. *Nat. Rev. Microbiol.* **8**, 15-25 (2010).
3. Davies, J. Specialized microbial metabolites: functions and origins. *J. Antibiot.* **66**, 361–364 (2013).
4. Gunatilaka, A. A. L. Natural products from plant-associated microorganisms: distribution, structural diversity, bioactivity, and implications of their occurrence. *J. Nat. Prod.* **69**, 509-526 (2006).
5. Kelly, C. R. *et al*. Fecal microbiota transplant for treatment of *Clostridium difficile* infection in immunocompromised patients. *Am. J. Gastroenterol.* **109**, 1065-1071 (2014).
6. Louis, P. *et al.* The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol.* **12**, 661-672 (2014).
7. Bell, T. H. *et al*. A diverse soil microbiome degrades more crude oil than specialized bacterial assemblages obtained in culture. *Appl. Environ. Microbiol.* **82**, 5530-5541 (2016).
8. Bokulich, N. A. *et al*. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* **8**, 1-13 (2016).
9. Tang, W. H. W. *et al.* Gut microbiota in cardiovascular health and disease. *Circ. Res.* **120**, 1183-1196 (2017).
10. Pham, J. V. *et al.* A Review of the Microbial Production of Bioactive Natural Products and Biologics. *Front. Microbiol.* **10**, 1404 (2019).

11. Dinsdale, E. A., *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452**,629-632 (2008).

12. Louca, S. *et al.* Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272-1277 (2016).

13. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61-66 (2017).

14. Quince, C., *et al.* Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833-844 (2017).

15. Franzosa, E. A., *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods* **15**, 962-968 (2018).

16. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81-W87 (2019).

17. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes - a review. *Nat. Prod. Rep.* **33**, 988–1005 (2016).

18. Libis, V. *et al.* Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence network analysis of targeted sequences. *Nat. Commun.* **10**, 3848 (2019).

19. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0718-6.

20. Hugerth, L. W. *et al.* Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biol.* **16**, 279 (2015).

21. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data* **5**, 170203 (2018).

22. Stewart, R. D. *et al.* Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961 (2019).
23. Van Goethem, M. W. *et al.* Long-read metagenomics of soil communities reveals phylum-specific secondary metabolite dynamics. *Cold Spring Harbor Laboratory* 2021.01.23.426502 (2021) doi:10.1101/2021.01.23.426502.
24. Amos, G. C. A. *et al.* Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E11121–E11130 (2017).
25. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).
26. Aksenov, A. A., da Silva, R., Knight, R., Lopes, N. P. & Dorrestein, P. C. Global chemical analysis of biology by mass spectrometry. *Nature Reviews Chemistry* **1**, 0054 (2017).
27. Dührkop, K. *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0740-8.
28. Mohimani, H. *et al.* Dereplication of microbial metabolites through database search of mass spectra. *Nat. Commun.* **9**, 4035 (2018).
29. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
30. van Santen, J. A. *et al.* The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Cent Sci* **5**, 1824–1833 (2019).
31. Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).

32. Nothias, L.-F. *et al.* Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* **17**, 905–908 (2020).

33. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

34. Thompson, L. *et al.* EMP Sample Submission Guide v1. *protocols.io* (2018) doi:10.17504/protocols.io.pfqdjmw.

35. Gonzalez, A. *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**, 796–798 (2018).

36. Allison, S. D. & Martiny, J. B. H. Resistance, resilience, and redundancy in microbial communities. *Proc. Nat'l. Acad. Sci. USA* **105**, 11512-11519.

37. Louca, S. *et al.* Function and functional redundancy in microbial communities. *Nat. Ecol. Evol.* **2**, 936-943 (2018).

38. Barnes, E. M. *et al.* Predicting microbiome function across space is confounded by strain-level differences and functional redundancy across taxa. *Frontiers Microbiol.* **11**, 101 (2020).

39. Lim, Y. W. *et al.* Metagenomics and metatranscriptomics: windows on CF-associated viral and microbial communities. *J. Cystic Fibrosis* **12**, 154-164 (2013).

40. Bikel, S. *et al.* Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a system-level understanding of human microbiome. *Computation. Struct. Biotechnol. J.* **13**, 390-401 (2015).

41. Schirmer, M. *et al.* Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nature Microbiology* **3**, 337-346 (2018).

42. Erickson, A. R. *et al.* Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS ONE* **7**, e49138 (2012).

43. Williams, T. J. *et al.* The role of planktonic *Flavobacteria* in processing organic matter in coastal East Antarctica revealed using metagenomics and metaproteomics. *Environ. Microbiol.* **15**, 1302-1317 (2013).

44. Leary, D. H. *et al.* Integrated metagenomic and metaproteomic analyses of marine biofilm communities. *Biofueling* **30**, 1211-1223 (2014).

45. Lu, K. *et al.* Arsenic exposure perturbs the gut microbiome and its metabolic profile in mice: an integrated metagenomics and metabolomics analysis. *Environ. Health Perspectives* **122**, 284-291 (2014).

46. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem of inflammatory bowel diseases. *Nature* **569**, 655-662 (2019).

47. Califf, K. J. *et al.* Multi-omics analysis of periodontal pocket microbial communities pre- and posttreatment. *mSystems* **2**, e00016-17 (2017).

Acknowledgements.

We thank Lindsay Goldasich and Julia Toronczak for assistance with sample processing for sequencing; Marcus Fedarko, Rachel Diner, Joshua Ladau, Elisha Wood-Charlson, Stephen Nayfach, Daniel Udwary, and Emiley Eloe-Fadrosh for reviewing the manuscript. This work was supported in part by the United States (US) National Institute of Health (NIH) (awards 1RF1-AG058942-01, 1DP1AT010885, R01HL140976, R01DK102932, R01HL134887, U19AG063744, U01AI124316), US Department of Agriculture – National Institute of Food and Agriculture (USDA-NIFA) (award 2019-67013-29137), the US National Science Foundation (NSF) - Center for Aerosol Impacts on Chemistry of the Environment, Crohn's & Colitis Foundation Award (CCFA) (award 675191), US Department of Energy - Office of Science - Office of Biological and Environmental Research - Environmental System Science Program, Semiconductor Research Corporation and Defense Advanced Research Projects Agency (SRC/DARPA) (award GI18518), Department of Defense (award W81XWH-17-1-0589), the Office of Naval Research (ONR) (award N00014-15-1-2809), the Emerald Foundation (award 3022), IBM Research AI through the AI Horizons Network, and the Center for Microbiome Innovation. J.P.S. was supported by NIH/NIGMS IRACDA K12 GM068524. L.F.N. was supported by the NIH (award R01-GM107550). A.D.B. was supported by the Danish Council for Independent Research (DFF) (award 9058-00025B). W.B. was supported by the Research Foundation – Flanders (12W0418N). K.D. and S.B. were supported by Deutsche Forschungsgemeinschaft (BO 1910/20 and 1910/23). P.C.D. was supported by the Gordon and Betty Moore Foundation (award GBMF7622) and the NIH (award R01-GM107550). Metabolomics analyses at Pacific Northwest National Laboratory (PNNL) were supported by the

Laboratory Directed Research and Development program via the Microbiomes in Transition Initiative and performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the US Office of Biological and Environmental Research and located at PNNL. This contribution originates in part from the River Corridor Scientific Focus Area project at PNNL. PNNL is a multiprogram national laboratory operated by Battelle for the Department of Energy (DOE) under contract DE-AC05-76RLO 1830. We thank Eppendorf, Illumina, and Integrated DNA Technologies for in-kind support at various phases of the project.

Author contributions.

J.A.G., J.K.J., and R.K. conceived the idea for the project. P.C.D., and R.K. designed the multi-omics component of the project and provided project oversight. J.P.S. managed the project, performed preliminary data exploration, coordinated data analysis, analyzed data, and provided data interpretation. L.F.N. coordinated and performed LC-MS/MS sample processing, coordinated and performed annotation of LC-MS/MS, analyzed LC-MS/MS data, and provided LC-MS/MS data interpretation. L.R.T. designed the multi-omics component of the project, solicited sample collection, curated sample metadata, processed samples, performed preliminary data exploration, and provided project oversight. J.G.S. designed the multi-omics component, managed the project, developed protocols and tools, coordinated and performed sequencing, and performed preliminary exploration of sequence data. R.A.S. developed protocols and coordinated and performed sequencing. S.P.C. and T.O.M. coordinated and performed GC-MS sample processing and provided interpretation of GC-MS data. A.D.B. conceived the idea for the paper,

performed preliminary data exploration, analyzed data, and provided data interpretation. S.H. performed machine-learning analyses. F.L. performed co-occurrence analysis, multinomial regression analyses, and correlations with co-occurrence data. H.L.L. performed multinomial regression analyses. Q.Z. developed tools and provided interpretation of shotgun metagenomics data. C.Mart. and J.T.M. provided oversight and interpretation of RPCA, multinomial regression, and co-occurrence analyses. S.K. performed preliminary exploration of shotgun metagenomics data. K.D., S.B, and H.W.K. annotated LC-MS/MS data. A.A.A. processed GC-MS data. W.B. provided oversight for machine-learning analyses. C.Maro. processed samples for sequencing. Y.V.B. performed preliminary data exploration and provided oversight for machine-learning analysis. A.T. and D.P. performed preliminary data exploration. L.P., A.P.C., N.H., and K.L.B. performed preliminary exploration of shotgun metagenomic data and performed machine learning analyses. P.D. performed preliminary exploration of shotgun metagenomics data. A.G. developed tools, provided interpretation of shotgun metagenomics data, and analyzed shotgun metagenomics data. G.H. coordinated short-read amplicon and shotgun metagenomics sequencing. M.M.B. and K.S. performed short-read amplicon and shotgun metagenomics sequencing. T.S. assisted with DNA extraction. D.M. coordinated long-read amplicon sequencing, analyzed shotgun metagenomics data, and provided interpretation of the data. S.M.K. and M.A. coordinated and performed long-read amplicon sequencing and long-read sequence data analysis. J.J.M. collected samples, coordinated field logistics, developed protocols, and performed short-read amplicon and shotgun metagenomics sequencing. S.S. collected samples, coordinated field logistics, and provided interpretation of the data. G.L.A. curated sample metadata and organized sequence data. J.D. processed sequence data. A.D.S.

provided project oversight and data interpretation. T.T., A.S., and J.S. collected samples, coordinated field logistics, and provided interpretation of the data. J.P.S. and L.F.N. wrote the manuscript, with contributions from all authors.

Earth Microbiome Project 500 (EMP500) Consortium.

Lars T. Angenant[1], Alison M. Berry[2], Leonora S. Bittleston[3], Jennifer L. Bowen[4], Max Chavarría[5,6], Don A. Cowan[7], Dan Distel[4], Peter R. Girguis[8], Jaime Huerta-Cepas[9], Paul R. Jensen[10], Lingjing Jiang[11], Gary M. King[12], Anton Lavrinienko[13], Aurora MacRae-Crerar[14], Thulani P. Makhalanyane[7], Tapio Mappes[13], Ezequiel M. Marzinelli[15], Gregory Mayer[16], Katherine D. McMahon[17], Jessica L. Metcalf[18], Sou Miyake[19], Timothy A. Mousseau[13], Catalina Murillo-Cruz[5], David Myrold[20], Brian Palenik[10], Adrián A. Pinto-Tomás[5], Dorota L. Porazinska[21], Jean-Baptiste Ramond[7,22], Forest Rowher[23], Taniya RoyChowdhury[24,25], Stuart A. Sandin[10], Steven K. Schmidt[26], Henning Seedorf[19,27], J. Reuben Shipway[28,29], Jennifer E. Smith[10], Frank J. Stewart[30], Karen Tait[31], Yael Tucker[32], Jana M. U'Ren[33], Phillip C. Watts[13], Nicole S. Webster[34,35], Jesse R. Zaneveld[36], Shan Zhang[37]

[1]University of Tuebingen, Tuebingen, Germany. [2]University of California, Davis, Davis, California, USA. [3]Boise State University, Boise, Idaho, USA. [4]Northeastern University, Boston, Massachusetts, USA. [5]University of Costa Rica, San José, Costa Rica. [6]CENIBiot, San José, Costa Rica. [7]University of Pretoria, Pretoria, South Africa. [8]Harvard University, Cambridge, Massachusetts, USA. [9]Universidad Politécnica de Madrid, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Madrid, Spain. [10]University of California San Diego, La Jolla,

California, USA. [11]Janssen Research & Development, San Diego, California, USA. [12]Louisiana State University, Baton Rouge, Louisiana, USA. [13]University of Jyväskylä, Jyväskylä, Finland. [14]University of Pennsylvania, Philadelphia, Pennsylvania, USA. [15]The University of Sydney, Sydney, Australia. [16]Texas Technology University, Lubbock, Texas, USA. [17]University of Wisconsin, Madison, Wisconsin, USA. [18]Colorado State University, Fort Collins, Colorado, USA. [19]Temasek Life Sciences Laboratory, Singapore, Singapore. [20]Oregon State University, Corvallis, Oregon, USA. [21]University of Florida, Gainesville, Florida, USA. [22]Pontificia Universidad Católica de Chile, Santiago, Chile. [23]San Diego State University, San Diego, California, USA. [24]Pacific Northwest National Laboratory, Richland, Washington, USA. [25]University of Maryland, College Park, Maryland, USA. [27]National University of Singapore, Singapore, Singapore. [28]University of Plymouth, Drake Circus, Plymouth, United Kingdom. [29]University of Massachusetts Amherst, Amherst, Massachusetts, USA. [30]Montana State University, Bozeman, Montana, USA. [31]Plymouth Marine Laboratory, Plymouth, United Kingdom. [32]National Energy Technology Laboratory, USA. [33]University of Arizona, Tucson, Arizona, USA. [34]Australian Institute of Marine Science, Townsville, Qld, Australia. [35]University of Queensland, Brisbane, Australia. [36]University of Washington Bothell, Bothell, Washington, USA. [37]University of New South Wales, Sydney, Australia.

Competing interests.

S.B. and K.D. are co-founders of Bright Giant GmbH.

ONLINE METHODS

DATASET DESCRIPTION

Sample collection. Samples were contributed by 34 principal investigators of the Earth Microbiome Project 500 (EMP500) Consortium. To achieve more even coverage across microbial environments, we devised an ontology of sample types (microbial environments), the EMP Ontology (EMPO) (http://earthmicrobiome.org/protocols-and-standards/empo/)[1] and selected samples to fill out EMPO categories as broadly as possible. As we anticipated previously[1], we have updated the number of levels as well as states therein for EMPO (referred to here as EMPO version 2; Figure 1b), based on an important, additional salinity gradient observed among host-associated samples when considering the novel shotgun metagenomic and metabolomic data generated here (Fig. 4c,d).

Samples were collected following the Earth Microbiome Project sample submission guide[2]. Briefly, samples were collected fresh, split into 10 aliquots, and then frozen, or alternatively collected and frozen, and subsequently split into 10 aliquots with minimal perturbation. Aliquot size was sufficient to yield 10–100 ng genomic DNA (approximately 10^7–10^8 cells). To leave samples amenable to chemical characterization (metabolomics), buffers or solutions for sample preservation (e.g., RNAlater) were avoided. Ethanol (50–95%) was allowed as it is compatible with LC-MS/MS though should also be avoided if possible.

Sampling guidance was tailored for four general sample types: bulk unaltered (e.g., soil, sediment, feces), bulk fractionated (e.g., sponges, corals, turbid water), swabs (e.g., biofilms), and filters. Bulk unaltered samples were split fresh (or frozen) sampled into 10 pre-labeled 2-mL screw-cap bead beater tubes (Sarstedt cat. no. 72.694.005 or similar), ideally with at least 200 mg

biomass, and flash frozen in liquid nitrogen (if possible). Bulk fractionated samples were fractionated as appropriate for the sample type, split into 10 pre-labeled 2-mL screw-cap bead beater tubes, ideally with at least 200 mg biomass, and flash frozen in liquid nitrogen (if possible). Swabs were collected as 10 replicate swabs using 5 BD SWUBE dual cotton swabs with wooden stick and screw cap (cat. no. 281130). Filters were collected as 10 replicate filters (47 mm diameter, 0.2 um pore size, polyethersulfone (preferred) or hydrophilic PTFE filters), placed in a pre-labeled 2-mL screw-cap bead beater tubes, and flash frozen in liquid nitrogen (if possible). All sample types were stored at –80 °C if possible, otherwise –20 °C.

To track the provenance of sample aliquots, we employed a QR coding scheme. Labels were affixed to aliquot tubes before shipping, when possible. QR codes had the format "name.99.s003.a05", where "name" is the PI name, "99" is the study ID, "s003" is the sample number, and "a05" is the aliquot number. QR codes (version 2, 25 pixels x 25 pixels) were printed on 1.125" x 0.75" rectangular and 0.437" circular cap Cryogenic Direct Thermal Labels (GA International, part no. DFP-70) using a Zebra model GK420d printer and ZebraDesigner Pro software for Windows. After receipt but before aliquots were stored in freezers, QR codes were scanned into a sample inventory spreadsheet using a QR scanner.

Sample metadata. Environmental metadata was collected for all samples based on the EMP Metadata Guide[2], which combines guidance from the Genomics Standards Consortium MIxS (Minimum Information about any Sequence) standard[3] and the Qiita Database (https://qiita.ucsd.edu)[4]. The metadata guide provides templates and instructions for each MIxS

environmental package (i.e., sample type). Relevant information describing each PI submission, or study, was organized into a separate study metadata file (Table S1).

METABOLOMICS

LC-MS/MS sample extraction and preparation. To profile metabolites among all samples, we used liquid chromatography with untargeted tandem mass spectrometry (LC-MS/MS), a versatile method that detects tens of thousands of metabolites in biological samples[12]. All solvents and reactants used were LC-MS grade. To maximize the biomass extracted from each sample, the samples were prepared depending on their sampling method (e.g., bulk, swabs, filter, and controls). The bulk samples were transferred into a microcentrifuge tube (polypropylene, PP) and dissolved in 7:3 MeOH:H_2O using a volume varying from 600 µL to 1.5 mL, depending on the amounts of sample available, and homogenized in a tissue-lyser (QIAGEN) at 25 Hz for 5 min. Then, the tubes were centrifuged at 15,000 rpm for 15 min, and the supernatant was collected in a 96-well plate (PP). For swabs, the swabs were transferred into a 96-well plate (PP) and dissolved in 1.0 mL of 9:1 EtOH:H_2O. The prepared plates were sonicated for 30 min, and after 12 hours at 4°C, the swabs were removed from the wells. The filter samples were dissolved in 1.5 mL of 7:3 MeOH:H_2O in microcentrifuge tubes (PP) and were sonicated for 30 min. After 12 hours at 4°C, the filters were removed from the tubes. The tubes were centrifuged at 15,000 rpm for 15 min, and the supernatants were transferred to 96-well plates (PP). The process control samples (bags, filters, and tubes) were prepared by adding 3.0 mL of 2:8 MeOH:H_2O and by recovering 1.5 mL after 2 min. After the extraction process, all sample plates were dried with a vacuum concentrator and subjected to solid phase extraction (SPE). The SPE is used to remove

salts that could reduce ionization efficiency during mass spectrometry analysis, as well as the most polar and non-polar compounds (e.g., waxes) that cannot be analyzed efficiently by reversed-phase chromatography. The protocol was as follows: The samples (in plates) were dissolved in 300 µL of 7:3 MeOH:H_2O and put in an ultrasound bath for 20 min. SPE was performed with SPE plates (Oasis HLB, Hydrophilic-Lipophilic-Balance, 30 mg with particle sizes of 30-µm). The SPE beds were activated by priming them with 100% MeOH, and equilibrated with 100% H_2O. The samples were loaded on the SPE beds, and 100% H_2O was used as wash solvent (600 µL). The eluted washing solution was discarded, as it contains salts and very polar metabolites that subsequent metabolomics analysis is not designed for. The sample elution was carried out sequentially with 7:3 MeOH:H_2O (600 µL) and with 100% MeOH (600 µL). The obtained plates were dried with a vacuum concentrator. For mass spectrometry analysis, the samples were resuspended in 130 µL of 7:3 MeOH:H_2O containing 0.2 µM of amitriptyline as an internal standard. The plates were centrifuged at 2,000 rpm for 15 min at 4°C. 100 µL of samples were transferred into a new 96-well plate (PP) for mass spectrometry analysis.

LC-MS/MS sample analysis. The extracted samples were analyzed by ultra-high performance liquid chromatography (UHPLC, Vanquish, Thermo Scientific, Waltham, Massachusetts, USA) coupled to a quadrupole-Orbitrap mass spectrometer (Q Exactive, Thermo Scientific, Waltham, Massachusetts, USA) operated in data-dependent acquisition mode (LC-MS/MS in DDA mode). Chromatographic separation was performed using a Kinetex C_{18} 1.7-µm (Phenomenex, Torrance, California, USA), 100-Å pore size, 2.1-mm (internal diameter) x 50-mm (length)

column with a C_{18} guard cartridge (Phenomenex). The column was maintained at 40°C. The mobile phase was composed of a mixture of (A) water with 0.1% formic acid (v/v) and (B) acetonitrile with 0.1% formic acid. Chromatographic elution method was set as follows: 0.00–1.00 min, isocratic 5% B; 1.00–9.00 min, gradient from 5% to 100% B; 9.00–11.00 min, isocratic 100% B; and followed by equilibration 11.00–11.50 min, gradient from 100% to 5% B; 11.50–12.50 min, isocratic 5% B. The flow rate was set to 0.5 mL/min.

The UHPLC was interfaced to the orbitrap using a heated electrospray ionization (HESI) source with the following parameters: ionization mode: positive, spray voltage, +3496.2 V; heater temperature, 363.90 °C; capillary temperature, 377.50 °C; S-lens RF, 60 (arb. units); sheath gas flow rate, 60.19 (arb. units); and auxiliary gas flow rate, 20.00 (arb. units). The MS^1 scans were acquired at a resolution (at *m/z* 200) of 35,000 in the *m/z* 100–1500 range, and the fragmentation spectra (MS^2) scans at a resolution of 17,500 from 0 to 12.5 min. The automatic gain control (AGC) target and maximum injection time were set at 1.0×10^6 and 160 ms for MS^1 scans, and set at 5.0×10^5 and 220 ms for MS^2 scans, respectively. Up to three MS^2 scans in data-dependent mode (Top 3) were acquired for the most abundant ions per MS^1 scans using the apex trigger mode (4 to 15 s), dynamic exclusion (11 s), and automatic isotope exclusion. The starting value for MS^2 was *m/z* 50. Higher-energy collision induced dissociation (HCD) was performed with a normalized collision energy of 20, 30, 40 eV in stepped mode. The major background ions originating from the SPE were excluded manually from the MS^2 acquisition. Analyses were randomized within plate and blank samples analyzed every 20 injections. A QC mix sample assembled from 20 random samples across the sample-types was injected at the beginning, the middle, and the end of each plate sequence. The chromatographic shift observed

throughout the batch is estimated as less than 2 s, and the relative standard deviation of ion intensity was 15% per replicates.

LC-MS/MS data processing. The mass spectrometry data were centroided and converted from the proprietary format (.raw) to the *m/z* extensible markup language format (.mzML) using ProteoWizard (ver. 3.0.19, MSConvert tool)[5]. The mzML files were then processed with MZmine toolbox[6] using the ion-identity networking modules[7] that allows advanced detection for adduct/isotopologue annotations. The MZmine processing was performed on Ubuntu 18.04 LTS 64-bits workstation (Intel Xeon E5-2637, 3.5 GHz, 8 cores, 64 Gb of RAM) and took ~3 d. The MZmine project, the MZmine batch file (.XML format), and results files (.MGF and .CSV) are available in the MassIVE dataset MSV000083475. The MZmine batch file contains all the parameters used during the processing. In brief, feature detection and deconvolution was performed with the ADAP chromatogram builder[8], and local minimum search algorithm. The isotopologues were regrouped, and the features (peaks) were aligned across samples. The aligned peak list was gap filled and only peaks with an associated fragmentation spectrum and occurring in a minimum of three files were conserved. Peak shape correlation analysis grouped peaks originating from the same molecule, and to annotate adduct/isotopologue with ion-identity networking[7]. Finally the feature quantification table results (.CSV) and spectral information (.MGF) were exported with the GNPS module for feature-based molecular networking analysis on GNPS[9] and with SIRIUS export modules.

LC-MS/MS data annotation. The results files of MZmine (.MGF and .CSV files) were uploaded to GNPS (http://gnps.ucsd.edu)[10] and analyzed with the feature-based molecular networking workflow[9]. Spectral library matching was performed against public fragmentation spectra (MS^2) spectral libraries on GNPS and the NIST17 library.

For the additional annotation of small peptides, we used the DEREPLICATOR tools available on GNPS[11,12]. We then used SIRIUS[13] (v. 4.4.25, headless, Linux) to systematically annotate the MS^2 spectra. Molecular formulas were computed with the SIRIUS module by matching the experimental and predicted isotopic patterns[14], and from fragmentation trees analysis[15] of MS^2. Molecular formula prediction was refined with the ZODIAC module using Gibbs sampling[16] on the fragmentation spectra (chimeric spectra or had a poor fragmentation were excluded). *In silico* structure annotation using structures from biodatabase was done with CSI:FingerID[17]. Systematic class annotations were obtained with CANOPUS[18] and used the NPClassifier ontology[19].

The parameters for SIRIUS tools were set as follows, for SIRIUS: molecular formula candidates retained (80), molecular formula database (ALL), maximum precursor ion *m/z* computed (750), profile (orbitrap), *m/z* maximum deviation (10 ppm), ions annotated with MZmine were prioritized and other ions were considered (i.e., [M+H3N+H]+, [M+H]+, [M+K]+,[M+Na]+, [M+H-H2O]+, [M+H-H4O2]+, [M+NH4]+); for ZODIAC: the features were split into 10 random subsets for lower computational burden and computed separately with the following parameters: threshold filter (0.9), minimum local connections (0); for CSI:FingerID: *m/z* maximum deviation (10 ppm) and biological database (BIO).

To establish putative microbially-related secondary metabolites, we collected annotations from spectral library matching and the DEREPLICATOR tools and queried them against the largest microbial metabolite reference databases (Natural Products Atlas[20] and MIBiG[21]). Molecular networking[9] was then used to propagate the annotation of microbially-related secondary metabolites throughout all molecular families (i.e., the network component).

LC-MS/MS data analysis. We combined the annotation results from the different tools described above to create a comprehensive metadata file describing each metabolite feature observed. Using that information, we generated a feature-table including only secondary metabolite features determined to be microbially-related. We then excluded very low-intensity features introduced to certain samples during the gap-filing step described above. These features were identified based on presence in negative controls that were universal to all sample types (i.e., bulk, filter, and swab), and by their relatively low per-sample intensity values. Finally, we excluded features present in positive controls for sampling devices specific to each sample type (i.e., bulk, filter, or swab). The final feature-table included 618 samples and 6,588 putative microbially-related secondary metabolite features that were used for subsequent analysis.

We used QIIME 2's[22] *diversity* plugin to quantify alpha-diversity (i.e., feature richness) for each sample, and *deicode*[23] to quantify beta-diversity (i.e., robust Aitchison distances) between each pair of samples. We parameterized our robust Aitchison Principal Components Analysis (RPCA)[23] to exclude samples with fewer than 500 features, and features present in fewer than 10% of samples. We used the *taxa* plugin to quantify the relative abundance of microbially-related secondary metabolite pathways and superclasses (i.e., based on NPClassifier)

within each environment (i.e., for each level of EMPO, version 2 level 4), and *songbird*[24] to identify sets of microbially-related secondary metabolites whose abundances were associated with certain environments. We parameterized our *songbird* model as follows: epochs = 1,000,000, differential prior = 0.5, learning rate = 1.0×10^{-5}, summary interval = 2, batch size = 400, minimum sample count = 0, and training on 80% of samples at each level of EMPO version 2, level 4 using 'Animal proximal gut (non-saline)' as the reference environment.

We used the RPCA biplot and EMPeror[25] to visualize differences in composition among samples, as well as the association with samples of the 25 most influential microbially-related secondary metabolite features (i.e., those with the largest magnitude across the first three principal component loadings). We tested for significant differences in metabolite composition across the various levels of EMPO (version 2) using permutational multivariate analysis of variance (PERMANOVA), implemented with QIIME 2's *diversity* plugin[22]. In parallel, we visualized significant differences in the relative abundances of specific microbially-related secondary metabolite pathways and superclasses within each environment by first selecting features and calculating log-ratios using *qurro*[26], and then plotting using the *ggplot2* package[27] in R[28]. We tested for significant differences in relative abundances across environments using Kruskal–Wallis tests, implemented using the base *stats* package in R[28].

GC-MS sample extraction and preparation. To profile volatile small molecules among all samples in addition to what was captured with LC-MS/MS, we used gas chromatography coupled with mass spectrometry (GC-MS). All solvents and reactants were GC-MS grade. Two protocols were used for sample extraction, one for the 105 soil samples and second for the 356

fecal and sediment samples that were treated as biosafety level 2. The 105 soil samples were received at the Pacific Northwest National Laboratory and processed as follows. Each soil sample (1 g) was weighed into microcentrifuge tubes (Biopur Safe-Lock, 2.0 mL, Eppendorf, Hamburg, Germany). 1 mL of H_2O and one scoop (~0.5 g) of a 1:1 (v/v) mixture of garnet (0.15-mm, Omni International, Kennesaw, Georgia, USA) and stainless steel (SS) (0.9 – 2.0-mm blend, Next Advance, Troy, New York) beads and one 3-mm SS bead (Qiagen, Hilden, Germany) was added to each tube. Samples were homogenized in a tissue lyser (Qiagen, Hilden, Germany) for 3 min at 30 Hz and transferred into 15-mL polypropylene tubes (Olympus, Genesee Scientific, San Diego, California, USA). Ice-cold water (1 mL) was used to rinse the smaller tube and combined into the 15-mL tube. 10 mL of 2:1 (v/v) chloroform:methanol added and samples were rotated at 4°C for 10 min followed by cooling at -70°C for 10 min and centrifuged at 4,000 rpm for 10 min to separate phases. The top and bottom layers were combined into 40 mL glass vials and dried using a vacuum concentrator. 1 mL of 2:1 chloroform:methanol was added to each large glass vial and the sample was transferred into 1.5-mL tubes and centrifuged at 12,000 g. The supernatant was transferred into glass vials and dried for derivatization.

The remaining 356 samples that were received from UCSD that included fecal and sediment samples were processed as follows: 100 µL of each sample was transferred to a 2 mL microcentrifuge tube using a scoop (MSP01, Next Advance, Tustin, California, USA). The final volume of sample was brought to 1.5 mL ensuring the solvent ratio is 3:8:4 H_2O:$CHCl_3$:MeOH by adding the appropriate volumes of H_2O, MeOH, and $CHCl_3$. After transfer, one 3-mm SS bead (QIAGEN), 400 µL of methanol, and 300 µL of H2O were added to each tube and the

samples were vortexed for 30 s. Then, 800 µL of chloroform was added and samples were vortexed for 30 s. After centrifuging at 4,000 rpm for 10 min to separate phases, the top and bottom layers were combined in a vial and dried for derivatization.

The samples were derivatized for GC-MS analysis as follows: 20 µL of a methoxyamine solution in pyridine (30 mg/mL) was added to the sample vial and vortexed for 30 s. A bath sonicator was used to ensure the sample was completely dissolved. Samples were incubated at 37°C for 1.5 h while shaking at 1,000 rpm. 80 µL of N-methyl-N-trimethylsilyltrifluoroacetamide and 1% trimethylchlorosilane (MSTFA) solution was added and samples were vortexed for 10 s, followed by incubation at 37°C for 30 min with 1000 rpm shaking. The samples were then transferred into a vial with an insert.

An Agilent 7890A gas chromatograph coupled with a single quadrupole 5975C mass spectrometer (Agilent Technologies, Santa Clara, California, USA) and a HP-5MS column (30-m × 0.25-mm × 0.25-µm; Agilent Technologies, Santa Clara, California, USA) was used for untargeted analysis. Samples (1 µL) were injected in splitless mode, and the helium gas flow rate was determined by the Agilent Retention Time Locking function based on analysis of deuterated myristic acid (Agilent Technologies, Santa Clara, California, USA). The injection port temperature was held at 250°C throughout the analysis. The GC oven was held at 60°C for 1 min after injection, and the temperature was then increased to 325°C by 10°C/min, followed by a 10 min hold at 325°C. Data were collected over the mass range of *m/z* 50–600. A mixture of FAMEs (C8–C28) was analyzed each day with the samples for retention index alignment purposes during subsequent data analysis.

GC-MS data processing and annotation. The data were converted from vendor's format to the .mzML format and processed using GNPS GC-MS data analysis workflow (https://gnps.ucsd.edu)[29]. The compounds were identified by matching experimental spectra to the public libraries available at GNPS, as well as NIST 17 and Wiley libraries. The data are publicly available at the MassIVE depository (https://massive.ucsd.edu); dataset ID: MSV000083743. The GNPS deconvolution is available on GNPS (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d5c5135a59eb48779216615e8d5cb3ac), as is the library search (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=59b20fc8381f4ee6b79d35034de81d86).

GC-MS data analysis. For multi-omics analyses including GC-MS data, we first removed noisy (i.e., suspected background contaminant- and artifactual) features by excluding those with GNPS-balance scores <50. We then used QIIME 2's *deicode*[23] plugin to estimate beta-diversity for each dataset using robust Aitchison distances. The final feature table for GC-MS beta-diversity analysis included 460 samples and 216 features.

METAGENOMICS

DNA extraction. For each round of DNA extractions described below for both amplicon and shotgun metagenomic sequencing, a single aliquot of each sample was processed for DNA extraction. DNA was extracted following the EMP 96-sample, magnetic bead-based DNA extraction protocol[30], following Marotz et al. (2017)[31], Minich et al. (2018)[32], and Minich et al. (2019)[33], and using the QIAGEN® MagAttract® PowerSoil® DNA KF Kit (384-sample) (i.e.,

optimized for KingFisher). Importantly, material from each sample was added to a unique bead tube (containing garnet beads) for single-tube lysis, which has been shown to reduce sample-to-sample contamination common in plate-based extractions[33]. For bulk samples, 0.1 to 0.25 g of material was added to each well; for filtered samples, one entire filter was added to each well; for swabbed samples, one swab head was added to each well. The lysis solution was dissolved at 60°C before addition to each tube, then capped tubes were incubated at 65°C for 10 min prior to mechanical lysis at 6000 rpm for 20 min using a MagNA Lyser (Roche Diagnostics, California, USA). Lysate from each bead tube was then randomly assigned and added to wells of a 96-well plate, and then cleaned-up using the KingFisher Flex system (Thermo Scientific, Waltham, Massachusetts, USA). Resulting DNA was stored at –20°C for sequencing. We note that whereas QIAGEN does not offer a 'hybrid' extraction kit allowing for single-tube lysis and plate-based clean-up, the Thermo MagMAX Microbiome Ultra kit does, and was recently shown to be comparable to the EMP protocol used here[34].

Amplicon sequencing. We generated amplicon sequence data for variable region four (V4) of the bacterial and archaeal 16S ribosomal RNA (rRNA) gene, variable region nine (V9) of the eukaryotic 18S rRNA gene, and the fungal internal transcribed spacer one (ITS1). For amplifying and sequencing all targets, we used a low-cost, miniaturized (i.e., 5-µL volume), high-throughput (384-sample) amplicon library preparation method implementing the Echo 550 acoustic liquid handler (Beckman Coulter, Brea, California, USA)[35]. The same protocol was modified with different primer sets and PCR cycling parameters depending on the target. Two rounds of DNA extraction and sequencing were performed for each target to obtain greater

coverage per sample. For a subset of 500 samples, we also generated high-quality sequence data for full-length bacterial rRNA operons following the protocol described by Karst et al. (2021)[36], which is outlined briefly below.

The protocol for 16S is outlined fully in Caporaso et al. (2018)[37]. To target the V4 region, we used the primers 515F (Parada) (5'-GTGYCAGCMGCCGCGGTAA-3') and 806R (Apprill) (5'-GGACTACNVGGGTWTCTAAT-3'). These primers are updated from the original EMP 16S-V4 primer sequences[38,39] in order to (1) remove bias against Crenarchaeota/Thaumarchaeota[40] and the marine freshwater clade SAR11 (Alphaproteobacteria)[41], and (2) enable the use of various reverse primer constructs (e.g., the V4-V5 region using the reverse primer 926R[42]) by moving the barcode/index to the forward primer[40]. We note that whereas we previously named these updated primers "515FB" and "806RB" to distinguish them from the original primers, the "B" may be misinterpreted to indicate "Barcode". To avoid ambiguity, we now use the original names suffixed with the lead author name (i.e., "515F (Parada)", "806R (Apprill)", and "926R (Quince)". We highly recommend to always check the primer sequence in addition to the primer name. For Qiita users, studies with "library_construction_protocol" as "515f/806rbc" used the original primers, whereas "515fbc/806r" indicates use of updated primers, where "bc" refers to the location of barcode.

To facilitate sequencing on Illumina platforms, the following primer constructs were used to integrate adapter sequences during amplification[38,39,43]. For the barcoded forward primer, constructs included (5' to 3'): the 5' Illumina adapter (AATGATACGGCGACCACCGAGATCTACACGCT), a Golay barcode (12-bp variable

sequence), a forward primer pad (TATGGTAATT), a forward primer linker (GT), and the forward primer (515F [Parada]) (GTGYCAGCMGCCGCGGTAA). For the reverse primer, constructs included (5' to 3'): The reverse complement of 3' Illumina adapter (CAAGCAGAAGACGGCATACGAGAT), a reverse primer pad (AGTCAGCCAG), a reverse primer linker (CC), and the reverse primer (806R [Apprill]) (GGACTACNVGGGTWTCTAAT).

For each 25-µL reaction, we combined 13 µL PCR-grade water (Sigma St. Louis, MO, USA, cat. no. W3500 or QIAGEN, Hilden, Germany, cat. no. 17000-10), 10 µL Platinum Hot Start PCR Master Mix (2X) (Thermo Scientific, Waltham, Massachusetts, USA, cat. no. 13000014), 0.5 µL of each primer (10 µM), and 1 µL of template DNA. The final concentration of the master mix in each 1X reaction was 0.8X and that of each primer was 0.2 µM. Cycling parameters for a 384-well thermal cycler were as follows: 94°C for 3 min; 35 cycles of 94°C for 1 min, 50°C for 1 min, and 72°C for 105 s; and 72°C for 10 min. For a 96-well thermal cycler, we recommend the following: 94°C for 3 min; 35 cycles of 94°C for 45 s, 50°C for 1 min, and 72°C for 90 s; and 72°C for 10 min.

We amplified each sample in triplicate (i.e., each sample was amplified in three replicate 25-µL reactions), and pooled products from replicate reactions for each sample into a single volume (75 µL). We visualized expected products between 300-350 bp on agarose gels, and note that whereas low-biomass samples may yield no visible bands, instruments such as a Bioanalyzer or TapeStation (Agilent, Santa Clara, California, USA) can be used to confirm amplification. We quantified amplicons using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Scientific, Waltham, MA, USA, cat. no. P11496), following the manufacturer's instructions. To pool samples, we combined an equal amount of product from each sample (240 ng) into a single tube,

and cleaned the pool using the UltraClean PCR Clean-Up Kit (QIAGEN, Hilden, Germany, cat. no. 12596-4), following the manufacturer's instructions. We checked DNA quality using a Nanodrop (Thermo Scientific, Waltham, Massachusetts, USA), confirming A260/A280 ratios were between 1.8-2.0.

For sequencing, the following primer constructs were used. Read 1 constructs included (5' to 3'): a forward primer pad (TATGGTAATT), a forward primer linker (GT), and the forward primer (515F [Parada]) (GTGYCAGCMGCCGCGGTAA). Read 2 constructs included (5' to 3'): a reverse primer pad (AGTCAGCCAG), a reverse primer linker (CC), and the reverse primer (806R [Apprill]) (GGACTACNVGGGTWTCTAAT). The index primer sequence was AATGATACGGCGACCACCGAGATCTACACGCT, which we highlight as having an extra GCT at the 3' end compared to Illumina's index primer sequence, in order to increase the T_m for read 1 during sequencing.

The protocol for 18S is outlined fully in Amaral-Zettler et al. (2018)[44]. To target variable region nine (V9), we used the primers 1391f (5'-GTACACACCGCCCGTC-3') and EukBr (5'-TGATCCTTCTGCAGGTTCACCTAC-3'). These primers are based on those of Amaral-Zettler et al. (2009)[45] and Stoek et al. (2010)[46], and are designed for use with Illumina platforms. The forward primer is a universal small-subunit primer, whereas the reverse primer favors eukaryotes but with mismatches can bind and amplify Bacteria and Archaea. In addition to deviations from the 16S protocol above with respect to primer construct sequences and PCR cycling parameters, we included a blocking primer that reduces amplification of vertebrate host DNA for host-associated samples, based on the strategy outlined by Vestheim et al. (2008)[47]. We note that the

blocking primer is particularly useful for host-associated samples with a low biomass of non-host eukaryotic DNA.

The following primer constructs were used to integrate adapter sequences during amplification. For the barcoded forward primer, constructs included (5' to 3'): the 5' Illumina adapter (AATGATACGGCGACCACCGAGATCTACAC), a forward primer pad (TATCGCCGTT), a forward primer linker (CG), and the forward primer (Illumina_Euk_1391f) (GTACACACCGCCCGTC). For the reverse primer, constructs included (5' to 3'): The reverse complement of 3' Illumina adapter (CAAGCAGAAGACGGCATACGAGAT), a Golay barcode (12-bp variable sequence), a reverse primer pad (AGTCAGTCAG), a reverse primer linker (CA), and the reverse primer (806R [Apprill]) (TGATCCTTCTGCAGGTTCACCTAC). The construct for the blocking primer is as such and is formatted for ordering from IDT (Coralville, Iowa, USA):
"GCCCGTCGCTACTACCGATTGG/ideoxyI//ideoxyI//ideoxyI//ideoxyI//ideoxyI/TTAGTGAG GCCCT/3SpC3/".

Reaction mixtures without the blocking primer (i.e., those for non-vertebrate hosts or free-living sample types as defined by EMPO) were prepared as described for 16S. For reactions including the blocking primer, we combined 9 µL PCR-grade water, 10 µL master mix, 0.5 µL of each primer (10 µM), 4 µL of blocking primer (10 µM), and 1 µL of template DNA. The final concentration of the master mix in each 1X reaction was 0.8X, that of each primer was 0.2 µM, and that of the blocking primer was 1.6 µM. Without blocking primers, cycling parameters for a 384-well thermal cycler were as follows: 94°C for 3 min; 35 cycles of 94°C for 45 s, 57°C for 1 min, and 72°C for 90 s; and 72°C for 10 min. With blocking primers, cycling parameters for a

384-well thermal cycler were as follows: 94°C for 3 min; 35 cycles of 94°C for 45 s, 65°C for 15 s, 57°C for 30 s, and 72°C for 90 s; and 72°C for 10 min. Expected bands ranged between 210-310 bp.

For sequencing, the following primer constructs were used. Read 1 constructs (Euk_illumina_read1_seq_primer) included (5' to 3'): a forward primer pad (TATCGCCGTT), a forward primer linker (CG), and the forward primer (1391f) (GTACACACCGCCCGTC). Read 2 constructs (Euk_illumina_read2_seq_primer) included (5' to 3'): a reverse primer pad (AGTCAGTCAG), a reverse primer linker (CA), and the reverse primer (EukBr) (TGATCCTTCTGCAGGTTCACCTAC). The index primer construct (Euk_illumina_index_seq_primer) included (5' to 3'): the reverse complement of the reverse primer (EukBr) (GTAGGTGAACCTGCAGAAGGATCA), the reverse complement of the reverse primer linker (TG), and the reverse complement of the reverse primer pad (CTGACTGACT).

The protocol for ITS is outlined fully in Smith et al. (2018)[48]. To target the fungal internal transcribed spacer (ITS1), we used the primers ITS1f (5'-CTTGGTCATTTAGAGGAAGTAA-3') and ITS2 (5'-GCTGCGTTCTTCATCGATGC-3'). These primers are based on those of White et al. (1990) [49], and we note that primer ITS1f used here binds 38 bp upstream of ITS1 reported in that study.

The following primer constructs were used to integrate adapter sequences during amplification. For the barcoded forward primer, constructs included (5' to 3'): the 5' Illumina adapter (AATGATACGGCGACCACCGAGATCTACAC), a forward primer linker (GG), and the forward primer (ITS1f) (CTTGGTCATTTAGAGGAAGTAA). For the reverse primer,

constructs included (5' to 3'): The reverse complement of 3' Illumina adapter (CAAGCAGAAGACGGCATACGAGAT), a Golay barcode (12-bp variable sequence), a reverse primer linker (CG), and the reverse primer (ITS2) (GCTGCGTTCTTCATCGATGC).

Reaction mixtures without the blocking primer were prepared as described for 16S. Cycling parameters for a 384-well thermal cycler were as follows: 94°C for 1 min; 35 cycles of 94°C for 30 s, 52°C for 30 s, and 68°C for 30 s; and 68°C for 10 min. Expected bands ranged between 250-600 bp[50,51].

For sequencing, the following primer constructs were used. Read 1 sequencing primer constructs included (5' to 3'): a forward primer segment (TTGGTCATTTAGAGGAAGTAA), and a region extending into the amplicon (AAGTCGTAACAAGGTTTCC). Read 2 sequencing primer constructs included (5' to 3'): a reverse primer segment (CGTTCTTCATCGATGC), and a region extending into the amplicon (VAGARCCAAGAGATC). The index sequencing primer construct included (5' to 3'): the reverse complement of the region extending into the amplicon (TCTC), the reverse complement of the reverse primer (GCATCGATGAAGAACGCAGC), and the reverse complement of the linker (CG).

The protocol for generating bacterial full-length rRNA operon data is described in Karst et al. (2021)[36]. The method uses a unique molecular identifier (UMI) strategy to remove PCR errors and chimeras, resulting in a mean error rate of 0.0007% and a chimera rate of 0.02% of the final amplicon data. Briefly, the bacterial rRNA operons were targeted with an initial PCR using tailed versions of 27f (AGRGTTYGATYMTGGCTCAG)[52] and 2490r (GACGGGCGGTGWGTRCA)[53]. The primer tails contained synthetic priming sites and 18-bp long patterned UMIs (NNNYRNNNYRNNNYRNNN). The PCR reaction (50-µL) contained 1-2

ng DNA template, 1U Platinum SuperFi DNA Polymerase High Fidelity (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and a final concentration of 1× SuperFi buffer, 0.2mM of each dNTP, 500nM of each tailed 27f and tailed 2490r. The PCR program consisted of initial denaturation (3 min at 95°C) and two cycles of denaturation (30 s at 95°C), annealing (30 s at 55°C) and extension (6 min at 72°C). The PCR product was purified using a custom bead purification protocol 'SPRI size selection protocol for >1.5–2 kb DNA fragments' (Oxford Nanopore Technologies)'. The resulting product consists of uniquely tagged rRNA operon amplicons. The uniquely tagged rRNA operons were amplified in a second PCR, where the reaction (100-µL) contained 2U Platinum SuperFi DNA Polymerase High Fidelity (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and a final concentration of 1X SuperFi buffer, 0.2 mM of each dNTP, 500 nM of each forward and reverse synthetic primer targeting the tailed primers from above. The PCR program consisted of initial denaturation (3 min at 95°C) and then 25-35 cycles of denaturation (15 s at 95°C), annealing (30 s at 60°C) and extension (6 min at 72°C) followed by final extension (5 min at 72°C). The PCR product was purified using the custom bead purification protocol above. Batches of 25 amplicon libraries were barcoded and sent for PacBio Sequel II library preparation and sequencing (Sequel II SMRT Cell 8M and 30 h collection time) at the DNA Sequencing Center at Brigham Young University. Circular consensus sequencing (CCS) reads were generated using CCS v.3.4.1 (https://github.com/PacificBiosciences/ccs) using default settings. UMI consensus sequences were generated using the longread_umi pipeline (https://github.com/SorenKarst/longread_umi) using the following command: longread_umi pacbio_pipeline -d ccs_reads.fq -o out_dir -m 3500 -M 6000 -s 60 -e 60 -f CAAGCAGAAGACGGCATACGAGAT -F

AGRGTTYGATYMTGGCTCAG -r AATGATACGGCGACCACCGAGATC -R CGACATCGAGGTGCCAAAC -U '0.75;1.5;2;0' -c 2.

Amplicon data analysis. For multi-omics analyses including amplicon sequence data, we processed each dataset for comparison of beta-diversity. For all amplicon data except that for bacterial full-length rRNA amplicons, raw sequence data were converted from bcl to fastq, and then multiplexed files for each sequencing run uploaded as separate preparations to Qiita (study: 13114). For each sequencing run, data were then demultiplexed, trimmed to 150-bp, and denoised using Deblur[54] to generate a feature-table of sub-operational taxonomic units (sOTUs) per sample.

For each 16S sequencing run, we placed denoised reads into the GreenGenes 13_8 phylogeny[55] via fragment insertion using QIIME 2's[22] SATé-Enabled Phylogenetic Placement (SEPP)[56] plugin, to produce a phylogeny for diversity analyses. Reads (i.e., sOTUs) not placed during SEPP were removed from each feature-table. We then used QIIME 2's *feature-table* plugin to merge feature-tables across sequencing runs, exclude singleton sOTUs, and rarefy the data to 5,000 features per sample. We then used QIIME 2's[22] *diversity* plugin to estimate beta-diversity using unweighted UniFrac distances. The final feature-table for 16S beta-diversity analysis included 681 samples and 93,260 features.

For 18S data, we used QIIME 2[22] to first merge feature-tables across sequencing runs, and then classify taxonomy for each sOTU via pre-fitted machine-learning classifiers[57] and the SILVA 138 reference database[58]. We then used QIIME 2's[22] *feature-table* plugin to exclude singleton sOTUs and samples with a total frequency <3,000 features, and the *deicode*[23] plugin to

estimate beta-diversity for each dataset using robust Aitchison distances[23]. The final feature table for 18S beta-diversity analysis included 496 samples and 40,587 features.

For fungal ITS data separately, we used QIIME 2[22] to first merge feature-tables across sequencing runs, and then classify taxonomy for each sOTU as above but using the UNITE 8 reference database[59]. We then used QIIME 2's *feature-table* plugin to exclude singleton sOTUs and samples with a total frequency <500 features, and the *deicode*[23] plugin to estimate beta-diversity for each dataset using robust Aitchison distances[23]. The final feature table for fungal ITS beta-diversity analysis included 488 samples and 10,821 features.

For full-length rRNA operon data, per-sample fasta files were re-formatted for importing to QIIME 2 as *SampleData[Sequences]* (i.e., with each header as '>{sample_identifier}_{sequence_identifier}), concatenated into a single fasta file, and imported. We then used QIIME 2's *vsearch* plugin[60] to dereplicate sequences and then cluster them at 65% similarity (i.e., due to rapid evolution at bacterial ITS regions). The 65% OTU feature-table had 365 samples and 285 features. The concatenated fasta file and 65% OTU feature-table were uploaded to Qiita as a distinct preparations (study: 13114). We then used QIIME 2's[22] *feature-table* plugin to exclude singleton OTUs and samples with a total frequency <500 features, and the *deicode*[23] plugin to estimate beta-diversity for each dataset using robust Aitchison distances[23]. The final feature table for full-length rRNA operon beta-diversity analysis included 242 samples and 196 features.

Shotgun metagenomic sequencing. One round of DNA extraction was performed as above for shotgun metagenomic sequencing. Sequencing libraries were prepared using a high-throughput

version of the HyperPlus library chemistry (Kapa Biosystems) miniaturized to approximately 1:10 reagent volume and optimized for nanoliter-scale liquid-handling robotics[61]. An exhaustive, step-by-step protocol and accompanying software can be found in Sanders et al. (2019)[61]. Briefly, DNA from each sample was transferred to a 384-well plate and quantified using the Quant-iT PicoGreen dsDNA Assay Kit, and then normalized to 5 ng in 3.5 µL of molecular-grade water using an Echo 550 acoustic liquid-handling robot (Labcyte, San Jose, CA, USA). For library preparation, reagents for each step (i.e., fragmentation, end repair and A-tailing, ligation, and PCR) were added in 1:10 the recommended volumes using a Mosquito HTS micropipetting robot (SPT Labtech, Tokyo, Japan). Fragmentation was performed at 37°C for 20 min and A-tailing at 65°C for 30 min.

Sequencing adapters and barcode indices were added in two steps[62]. First, the Mosquito HTS robot was used to add universal adapter "stub" adapters and ligase mix to the end-repaired DNA, and the ligation reaction performed for 20°C for 1 h. Adapter-ligated DNA was then cleaned-up using AMPure XP magnetic beads and a BlueCat purification robot (BlueCat Bio, Concord, Massachusetts, USA) by adding 7.5 µL magnetic bead solution to the total sample volume, washing twice with 70% EtOH, and resuspending in 7 µL molecular-grade water. Then, the Echo 550 robot was used to add individual i7 and i5 indices to adapter-ligated samples without repeating any barcodes, and by iterating the assignment of i7 to i5 indices such to minimize repeating unique i7:i5 pairs. Cleaned, adapter-ligated DNA was then amplified by adding 4.5 µL of each sample to 5.5 µL PCR master mix and running for 15 cycles, and then purified again using magnetic beads and the BlueCat robot. Each sample was eluted into 10 µL water, and then transferred to a 384-well plate using the Mosquito HTS robot. Each library was

quantified using qPCR and then pooled to equal molar fractions using the Echo 550 robot. The final pool was sequenced at Illumina on a NovaSeq6000 using S2 flow cells and 2x150-bp chemistry (Illumina, San Diego, California, USA). To increase sequence coverage for certain samples, libraries were re-pooled and a second sequencing run performed as above.

Shotgun data analysis. Raw sequence data were converted from bcl to fastq and demultiplexed to produce per-sample fastq files before processing with Atropos (v1.1.24)[63] to trim universal adapter sequences, and poly-G tails introduced by the NovaSeq instrument (i.e., from use of two-color chemistry), from reads. Trimmed reads were then mapped to the Web of Life database of microbial genomes[64] to produce alignments that were used for both taxonomic and functional analysis of microbial communities.

For the taxonomic analysis, we generated a feature-table of counts of operational genomic units (OGUs) for each sample, using Woltka's (https://github.com/qiyunzhu/woltka) *classify* function, with per-genome alignments and default parameters. For diversity analyses, to best normalize sampling effort per sample while maintaining ≥75% of samples representative of the Earth's environments, we rarefied the OGU feature-table to 6,550 reads per sample. The final feature-table for analyses of shotgun metagenomic taxonomic diversity included 612 samples and 8,692 OGUs.

For alpha-diversity, we quantified three metrics, in part to see which had the strongest correlations with microbially-related metabolite richness. We used the R package *geiger*[65] to quantify weighted Faith's PD for each sample following the method of Swensen[66]. We used

QIIME 2's *diversity* plugin[22] to quantify richness and Faith's PD (i.e., unweighted), as well as beta-diversity (i.e., using weighted UniFrac distance) between each pair of samples. We performed PERMANOVA on that distance matrix to test for significant differences in microbial community composition across the various levels of EMPO (version 2). We then used Principal Coordinates Analysis (PCoA) and EMPeror[25] to visualize differences in microbial community composition among samples.

For the functional analysis, we generated a feature-table of counts of Gene Ontology (GO) Terms (i.e., for biological process, molecular function, and cellular compartment) for each sample using Woltka's *collapse* function, inputting per-gene alignments and with default parameters for mapping to GO Terms through MetaCyc. For subsequent analysis, we used QIIME 2's[22] *feature-table* plugin to exclude singleton features and rarefy the data to 5,000 sequences per sample. The final feature-table for analysis of shotgun metagenomic functional diversity included 517 samples and 3,776 features.

MULTI-OMICS

Alpha-diversity correlations. Using the alpha-diversity metrics for LC-MS/MS (i.e., richness) and shotgun metagenomic taxonomic data (i.e., richness, unweighted Faith's PD, and weighted Faith's PD), we performed correlation analysis to better understand relationships therein. We used the function *multilevel* available in the R package *correlation*[67] to perform Spearman correlations for each environment (i.e., based on EMPO version 2, level 4), treating study as a random effect, and adjusting for multiple comparisons using the Benjamini-Hochberg correction.

Machine-learning analyses. To better understand community composition of microbes and metabolites across environments and specifically which features are predictive of certain habitats, we performed machine-learning. For analyses of LC-MS/MS and shotgun metagenomic taxonomic- and functional data, additional samples were filtered from the feature-tables noted previously such to exclude environments with relatively low sample representation (i.e., <10 samples). For the LC-MS/MS feature-table, we excluded samples in the three EMPO environments (i.e., "Animal corpus (non-saline)", "Animal proximal gut (non-saline)", and "Surface (saline)"). The final feature-table included 606 samples (representing 15 environments), and 6,588 microbially-related secondary metabolites. For the shotgun metagenomic feature-table for taxonomic analysis, we excluded samples in the five most lowly-sampled environments (based on EMPO version 2, level 4). The final feature-table included 598 samples (representing 14 environments), and 8,567 OGUs. For the shotgun metagenomic feature-table for functional analysis, we excluded samples in the seven most lowly sampled environments (based on EMPO version 2, level 4). The final feature-table included 485 samples (representing 12 environments) and 2,436 GO terms.

For each feature-table, we trained a random forest classifier for predicting environments (based on EMPO version 2, level 4) with tuned hyperparameters and a 20-time repeated group 5-fold cross-validation approach using the R package *caret*[68]. Each dataset was split into a training set (80%) and a testing set (20%), with similar environmental distributions in each fold iteration for the classification of samples. To assess any batch effects on the classification of environments (i.e., potential confounding from overrepresentation of environments by single studies), the data-split in our approach was such that all samples from the same study were

assigned to either train- or test data sets. With this approach, we can report the study-level prediction accuracy of the classifiers. We next evaluated the prediction performance via overall accuracy quantified from a confusion matrix based on all predictions in a 100-fold cross-validation. We evaluated how well each classifier could predict each environment by quantifying the area under the receiver operating characteristic (AUROC) and the area under the precision-recall curve (AUPRC). To identify the most important features contributing to the classification, we visualized the importance score and environment-wide prevalence and relative abundance (i.e., intensity for metabolites) of the top 32 most important features in each dataset.

Metabolite-microbe co-occurrence analysis. To begin to explore co-occurrences between microbes and metabolites across environments, we implemented an approach that generates co-occurrence probabilities between all metabolite- and microbial features, clusters metabolites based on their co-occurrence with the microbial community, and highlights individual microbial features driving global patterns in metabolite distribution in this space. For co-occurrence analyses of LC-MS/MS metabolites and genomes profiled from shotgun metagenomic data, feature-tables were further filtered to only retain the 434 samples found in both datasets. For the LC-MS/MS feature-table of microbially-related secondary metabolites, we excluded 172 samples lacking shotgun metagenomics data, resulting in a final set of 6,501 microbially-related metabolites. For the shotgun metagenomics feature-table for taxonomy, we excluded 150 samples lacking LC-MS/MS data, resulting in a final set of 4,120 OGUs.

Specifically, we obtained co-occurrence probabilities and ordinated metabolites in microbial taxon space using *mmvec*, the former of which we used to predict metabolites based on

microbial taxa from neural-network, compositionally-robust modeling[69]. The model was trained on 80% of the 434 samples, which were selected to balance environments (i.e., EMPO version 2, level 4), and used the following parameters: epochs = 200, batch-size = 165, learning-rate = 1e-5, summary-interval = 1, and with 'equalize-biplot'. We also filtered for training and testing to retain only those features present in at least 10 samples (i.e., min-feature-count = 10), and restricted decomposition of the co-occurrence matrix to 10 principal components (PCs) (i.e., latent-dim = 10). We then calculated the Spearman correlation between the loadings of metabolites on each PC vs. (i) betas for metabolite abundances across environments from Songbird multinomial regression, (ii) loadings for metabolites on the first three axes from the RPCA ordination corresponding to clustering of samples by environment, and (iii) a vector representing the global magnitude of metabolite importance across all three axes from the RPCA biplot. We also visualized the relationships between metabolite-microbe co-occurrences and (i) by considering the first three PCs of the co-occurrence ordination and coloring metabolites by their beta values (e.g., Fig. 4c).

Mantel correlations between datasets. To explore the relationships between sample-to-sample distances for any two datasets (e.g., LC-MS/MS vs. shotgun metagenomic for taxonomy), we used QIIME 2's *diversity* plugin[22] to perform Mantel tests on all pairings of the datasets using Spearman correlations. Input distance matrices are those described above for each dataset.

References.

1. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
2. Thompson, L. *et al.* EMP Sample Submission Guide v1 (protocols.io.pfqdjmw). *protocols.io* (2018) doi:10.17504/protocols.io.pfqdjmw.
3. Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **29**, 415–420 (2011).
4. Gonzalez, A. *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**, 796–798 (2018).
5. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
6. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).
7. Schmid, R. *et al.* Ion Identity Molecular Networking in the GNPS Environment. *Cold Spring Harbor Laboratory* 2020.05.11.088948 (2020) doi:10.1101/2020.05.11.088948.
8. Du, X., Smirnov, A., Pluskal, T., Jia, W. & Sumner, S. Metabolomics Data Preprocessing Using ADAP and MZmine 2. *Methods Mol. Biol.* **2104**, 25–48 (2020).
9. Nothias, L.-F. *et al.* Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* **17**, 905–908 (2020).

10. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
11. Mohimani, H. *et al.* Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* **13**, 30–37 (2017).
12. Mohimani, H. *et al.* Dereplication of microbial metabolites through database search of mass spectra. *Nat. Commun.* **9**, 4035 (2018).
13. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
14. Böcker, S., Letzel, M. C., Lipták, Z. & Pervukhin, A. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* **25**, 218–224 (2009).
15. Böcker, S. & Dührkop, K. Fragmentation trees reloaded. *J. Cheminform.* **8**, 5 (2016).
16. Ludwig, M. *et al.* Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. *Nature Machine Intelligence* **2**, 629–641 (2020).
17. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 12580–12585 (2015).
18. Dührkop, K. *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0740-8.
19. Kim, H. *et al.* NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J. Nat. Prod.* (2021) doi: 10.1021/acs.jnatprod.1c00399
20. van Santen, J. A. *et al.* The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Cent Sci* **5**, 1824–1833 (2019).

21. Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).

22. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).

23. Martino, C. *et al.* A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *mSystems* **4**, (2019).

24. Morton, J. T. *et al.* Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **10**, 2719 (2019).

25. Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience* **2**, 16 (2013).

26. Fedarko, M. W. *et al.* Visualizing ’omic feature rankings and log-ratios using Qurro. *NAR Genom Bioinform* **2**, lqaa023 (2020).

27. Wilkinson, L. Ggplot2: Elegant graphics for data analysis by WICKHAM, H. *Biometrics* **67**, 678–679 (2011).

28. Team, R. C. & Others. R: A language and environment for statistical computing. (2013).

29. Aksenov, A. A. *et al.* Auto-deconvolution and molecular networking of gas chromatography-mass spectrometry data. *Nat. Biotechnol.* **39**, 169–173 (2021).

30. Marotz, L. *et al.* Earth Microbiome Project (EMP) high throughput (HTP) DNA extraction protocol v1 (protocols.io.pdmdi46). *protocols.io* (2018) doi:10.17504/protocols.io.pdmdi46.

31. Marotz, C. *et al.* DNA extraction for streamlined metagenomics of diverse environmental samples. *Biotechniques* **62**, 290–293 (2017).

32. Minich, J. J. *et al.* KatharoSeq Enables High-Throughput Microbiome Analysis from Low-Biomass Samples. *mSystems* **3**, (2018).

33. Minich, J. J. *et al.* Quantifying and Understanding Well-to-Well Contamination in Microbiome Research. *mSystems* **4**, (2019).

34. Shaffer, J. P. *et al.* A comparison of DNA/RNA extraction protocols for high-throughput sequencing of microbial communities. *Biotechniques* **70**, 149–159 (2021).

35. Minich, J. J. *et al.* High-Throughput Miniaturized 16S rRNA Amplicon Library Preparation Reduces Costs while Preserving Microbiome Integrity. *mSystems* **3**, (2018).

36. Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat. Methods* **18**, 165–169 (2021).

37. Greg, J. *et al.* EMP 16S Illumina Amplicon Protocol v1 (protocols.io.nuudeww). *protocols.io* (2018) doi:10.17504/protocols.io.nuudeww.

38. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 1**, 4516–4522 (2011).

39. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).

40. Parada, A. E., Needham, D. M. & Fuhrman, J. A. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.* **18**, 1403–1414 (2016).

41. Apprill, A., McNally, S., Parsons, R. & Weber, L. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat. Microb. Ecol.* **75**, 129–137 (2015).

42. Quince, C., Lanzen, A., Davenport, R. J. & Turnbaugh, P. J. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12**, 38 (2011).

43. Walters, W. *et al.* Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. *mSystems* **1**, (2016).

44. Linda *et al.* EMP 18S Illumina Amplicon Protocol v1 (protocols.io.nuvdew6). *protocols.io* (2018) doi:10.17504/protocols.io.nuvdew6.

45. Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W. & Huse, S. M. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One* **4**, e6372 (2009).

46. Stoeck, T. *et al.* Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.* **19 Suppl 1**, 21–31 (2010).

47. Vestheim, H. & Jarman, S. N. Blocking primers to enhance PCR amplification of rare sequences in mixed samples - a case study on prey DNA in Antarctic krill stomachs. *Front. Zool.* **5**, 12 (2008).

48. Dylan P, S. *et al.* EMP ITS Illumina Amplicon Protocol v1 (protocols.io.pa7dihn). *protocols.io* (2018) doi:10.17504/protocols.io.pa7dihn.

49. White, T. *et al.* Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. PCR protocols: a guide to methods and applications. https://www.scienceopen.com/document?vid=36d59e39-6250-4a7f-b5fe-7155abbb4e03.

50. Hoggard, M. *et al.* Characterizing the Human Mycobiota: A Comparison of Small Subunit rRNA, ITS1, ITS2, and Large Subunit rRNA Genomic Targets. *Front. Microbiol.* **9**, 2208 (2018).

51. Bokulich, N. A. & Mills, D. A. Improved selection of internal transcribed spacer-specific primers enables quantitative, ultra-high-throughput profiling of fungal communities. *Appl. Environ. Microbiol.* **79**, 2519–2526 (2013).

52. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucl. Acids Res.* **41**, e1 (2012).

53. Hunt, D. E. *et al.* Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. *Appl. Environ. Microbiol.* **72**, 2221-2225 (2006).

54. Amir, A. *et al.* Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* **2**:e00191-16 (2017).

55. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analysis of bacteria and archaea. *ISME J.* **6**, 610-618 (2012).

56. Janssen, S. *et al.* Phylogenetic placement of exact amplicon sequences improves associations with clinical information. *mSystems* **3**:e00021-18 (2018).

57. Bokulich, N. A. *et al.* Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90 (2018).

58. Yilmaz, P. *et al.* The SILVA and "All-species Living Tree Project (LPT)" taxonomic frameworks. *Nucl. Acids Res.* **42**, D643-D648 (2014).

59. Nilsson, R. H. *et al.* 2018. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucl. Acids Res.* **47**, D259-D264. (2018).

60. Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. Vsearch: a versatile open source tool for metagenomics. *PeerJ*, **4**:e2584 (2016).

61. Sanders, J. G. *et al.* Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biol.* **20**, 226 (2019).

62. Glenn, T. C. *et al.* Adapterama I: Universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext). *bioRxiv* 049114 (2019) doi:10.1101/049114.

63. Didion, J. P., Martin, M. & Collins, F. S. Atropos: specific, sensitive, and speedy trimming of sequencing reads. *PeerJ* **5**, e3720 (2017).

64. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).

65. Pennell, M. W. *et al.* geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* **30**, 2216-2218 (2014).

66. Swenson, N. G. *Functional and Phylogenetic Ecology in R*. (Springer, New York, NY, 2014). doi:10.1007/978-1-4614-9542-0.

67. Makowski, D., *et al*. Methods and Algorithms for Correlation Analysis in R. *Journal of Open Source Software*, **5**, 2306 (2019).

68. Kuhn, M. Classification and Regression Training [R package caret version 6.0-86]. (2020).

69. Morton, J. T. *et al.* Learning representations of microbe-metabolite interactions. *Nat. Methods* **16**, 1306–1314 (2019).