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

Abstract.

Using sequence data to understand function in microbial communities presents challenges, in part due to the complex intricacies of processes such as transcription and translation. Here, we sample the metabolome directly using untargeted metabolomics, and combine it with metagenomic sequencing to harness a multi-omics approach. We focus on a diverse sample set representing Earth's habitats to describe the distribution of metabolites, and their relationships with microbial communities and the environment.

Microbes produce an array of secondary metabolites that perform diverse functions from communication to defense[1]. These metabolites have been used to benefit human health and sustainability[2]. To begin to understand the metabolic capabilities of microbial communities, metagenomic sequence data can be profiled for genes related to certain products, allowing interpretation of metabolic potential[3]. Although metagenome mining is a powerful way to annotate biosynthetic gene clusters (BCGs), chemical evidence is required to confirm the presence of metabolites and comprehensively address this fundamental hypothesis, as metagenomic data only identify metabolic potential. To describe the Earth's metabolome, we use an integrated omics approach: the direct survey of metabolites associated with microbial communities spanning diverse environments using untargeted metabolomics coupled with metagenome analysis. We show that the presence of certain classes of secondary metabolites can be limited or amplified by the environment. 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, and that their richness and

composition in any given sample do not directly reflect those of co-occurring microbial communities, but rather vary with the environment.

From a genomics perspective, information regarding metabolic potential is obtained through detection and classification of BGCs, the genomic loci underlying the production of secondary metabolites and their precursors[4]. The most sensitive approaches amplify BGC-specific marker sequences by PCR[5], but only metagenomic methods can link BGCs to their genomes (i.e., metagenome-assembled genomes, or MAGs) of origin and detect BGCs in novel MAGs. Along these lines, a recent large-scale analysis of putative BGC regions from microbial MAGs surprisingly concluded that, although the main classes of secondary metabolites are enriched in particular microbial taxa, the relative distribution of secondary metabolite biosynthetic potential across environments was conserved, implying that most classes of secondary metabolites are not limited or amplified by the environment. One caveat of such analyses is that annotated BGCs often have incomplete sequences, potentially impacting annotation and quantification[3]. More importantly, gene-level data about BGCs inferred from MAGs cannot offer information about actual synthesis (e.g., gene expression), creating uncertainty about the distribution of secondary metabolites across environments[6–9]. 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 the underlying genomic origins[10]. Furthermore, quantifying metabolite diversity from such 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 [11]. Finally, shotgun metagenomics does not

capture BGCs from low-abundance MAGs efficiently, as shown from comparative studies of targeted sequencing approaches[5].

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[12]. Although LC-MS/MS metabolomics has generally suffered from a low metabolite annotation rate when applied to non-model organisms, recent computational advances can systematically classify metabolites using their fragmentation spectra[13]. Untargeted metabolomics provides the relative abundance (i.e., intensity) of each metabolite detected across samples rather than just counts of unique structures (e.g., Fig. 1a vs. 1b), and thus provides a direct readout of the surveyed environment, a result that is difficult to achieve with a purely genomics approach. Whereas there is a clear need for the use of untargeted metabolomics to quantify the metabolic activities of microbiota, this methodology has been limited by the challenge of discriminating the secondary metabolites produced by microbes from tens-of-thousands of metabolites detected in the environment. 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[14] using the GNPS web-platform[15]. These annotations were then queried against microbial metabolite reference databases (i.e., Natural Products Atlas[16] and MIBiG[17]), and molecular networking[18] 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)[13].

We used this methodology to quantify microbial secondary metabolites from diverse microbial communities that span 20 major environments from the Earth Microbiome Project 500 (EMP500) dataset (Fig. S1, Table S1).To avoid confusion with terminology, we define 'microbial community' as consisting of members of Bacteria and Archaea. With this dataset, we show 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 limited or amplified by the environment, especially when considering more resolved chemical class ontology levels (Fig. 1). 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 enriched in certain environments (Fig. 2a,c, Table S2, 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, and fungal samples (Fig. 2a). 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). We also identified specific metabolites that could classify particular environments with 68.9% accuracy in machine-learning analysis (Fig. 3a, Table S4), and found further support for the importance of particular metabolites in distinguishing environments, including the putative cholane steroid above (i.e., $C_{24}H_{38}O_4$), and three metabolites enriched in non-saline soil and plant corpus samples (Fig. 2c, Fig. 3a, Table S4).

In addition to showing that the relative abundances of microbially-related metabolites distinguish environments, our results highlight the advantages of using a multi-omics approach to interpret and predict the contributions of microbes and their environments to chemical profiles in nature. Moreover, our approach illustrates that recent advances in computational annotation tools offer a powerful toolbox to interpret untargeted metabolomics data[13]. With these observations, we hypothesized 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 begin to explore these relationships, we analyzed our shotgun metagenomics data and found strong correlations 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 corpus (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. 2b; Table S5). We also found similarity in the clustering of samples by environment between datasets (Fig. 2c,d), and a strong correlation between sample-sample distances based on microbially related metabolites vs. microbial taxa (Table S6). Using machine-learning, we determined that specific microbial taxa and their functions could classify particular environments with 72.08% and 68.19% accuracy, respectively (Figs. S2 and S3). In addition, we examined correlations between microbe-metabolite co-occurrences learned from shotgun metagenomic profiles and (1) log-fold changes of metabolites across environments, and (2) global distributions of metabolites, and found strong relationships with each (Figure 3b). In particular, the abundances of microbially-related metabolites in aquatic samples had a strong correlation with microbe-metabolite co-occurrences (Figure 3b). In addition to highlighting such environments as

unique vs. other free-living and host-associated samples, this demonstrates that microbes and metabolites can be classified by- and co-occur among environments.

We further generated additional data for EMP500 samples, including gas chromatography-mass spectrometry (GC-MS) and amplicon sequence (i.e., 16S, 18S, ITS, and full-length rRNA operon) data that also supported a strong relationship between bacterial- and archaeal communities and metabolic profiles (Table S6). 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.

Methods. (1,000 words max)

DATASET DESCRIPTION

Sample collection. Samples were contributed by 34 principal investigators (PIs) of the Earth Microbiome Project 500 (EMP500) Consortium. To achieve more even coverage across microbial environments, we devised an ontology of sample types (i.e., microbial environments), the EMP Ontology (EMPO) (http://www.earthmicrobiome.org/protocols-and-standards/empo/)[19] and selected samples to fill out EMPO categories as broadly as possible. Samples were collected following the EMP sample submission guide[20]. 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 (e.g.,

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, although should also be avoided if possible. All sample types were stored at –80°C if possible, otherwise –20°C. Additional details are described in the Supplementary Material.

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[21] and the Qiita Database (https://qiita.ucsd.edu)[22]. The metadata guide provides templates and instructions for each MIxS environmental package (i.e., sample type). Relevant information describing each PIs submission, or study, was organized into a separate study metadata file (Table S1).

METABOLOMICS

LC-MS/MS. We used liquid chromatography with untargeted tandem mass spectrometry (LC-MS/MS) to perform untargeted metabolomic profiling of all samples. Details of sample extraction and preparation, sample analysis (i.e., running the samples), data processing, data annotation, and data analysis are described in the Supplementary Material. Briefly, mass spectrometry data were normalized and converted from raw to the *m/z* extensible markup language format (.mzML) using ProteoWizard (ver. 3.0.19, MSConvert tool)[23]. The mzML files were then processed with MZmine toolbox[24]. Data are available in the MassIVE dataset MSV000083475. The results files of MZmine were uploaded to GNPS (http://gnps.ucsd.edu)[15] and analyzed with the feature-based molecular networking workflow[18].

GC-MS. We used gas chromatography to perform additional untargeted metabolomic profiling of volatile compounds from all fecal, sediment, and soil samples. Details of sample extraction and preparation, data processing and annotation, and data analysis are described in the Supplementary Material. Briefly, data were converted from raw format to the .mzML format as above, and are available in the MassIVE dataset MSV000083743. Files were uploaded to GNPS[15] and processed using the GC-MS data analysis workflow[25].

METAGENOMICS

DNA extraction. For each round of 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[26], following Marotz et al. (2017)[27], Minich et al. (2018)[28], and Minich et al. (2019)[29], and using the QIAGEN® MagAttract® PowerSoil® DNA KF Kit (384) (i.e., optimized for KingFisher). Additional details are described in the Supplementary Material.

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[30]. 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)[31]. Additional details are described in the Supplementary Material.

Amplicon data analysis. For multi-omic analyses including amplicon sequence data, we processed each dataset for comparison of beta-diversity. 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). Subsequent processing to generate feature-tables, merge data across preparations, and generate distance matrices were performed in Qiita; all other analyses were performed with QIIME2[32]. Additional details are described in the Supplementary Material.

Shotgun metagenomic sequencing. One round of DNA extraction was performed 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[33]. The final pool was sequenced at Illumina on a NovaSeq6000 using S2 flow cells and 2x150-bp chemistry (Illumina, San Diego, California, USA). Additional details are described in the Supplementary Material.

Shotgun data analysis. Raw sequence data were converted from bcl to fastq, demultiplexed, and trimmed with Atropos (v1.1.24)[34] to remove universal adapter sequences and poly-G tails introduced by the NovaSeq instrument (i.e., from use of two-color chemistry). Trimmed reads were mapped to the Web of Life database of microbial genomes[35] to produce alignments that were used for both taxonomic and functional analysis of microbial communities. Additional details are described in the Supplementary Material.

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, Faith's PD, and weighted Faith's PD[36]), we performed correlation analysis to better understand their relationship. We used the function *multilevel* available in the R package *correlation*[37] to perform Spearman correlations for each environment (i.e., based on EMPO), treating study as a random effect, and adjusting for multiple comparisons using the Benjamini-Hochberg correction.

Machine-learning analyses. For machine-learning analyses of LC-MS/MS and shotgun metagenomic taxonomic- and functional data, additional samples were filtered from feature-tables such to exclude environments (i.e., based on EMPO) with relatively low sample representation (<10 samples).

For each feature-table, we trained a Random Forest classifier for predicting environments with tuned hyperparameters with a 20-time repeated group 5-fold cross-validation approach using the R package *caret*[38]. 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. Additional details are described in the Supplementary Material.

Metabolite-microbe co-occurrence analysis. For co-occurrence analyses of LC-MS/MS metabolites and microbial genomes profiled from shotgun metagenomic data, feature-tables were further filtered to only retain the samples found in both datasets. We obtained co-occurrence

probabilities using *mmvec*, which we used to predict metabolites based on microbial taxa from neural-network, compositionally-robust modeling[39]. 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[40] 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[40]. We also visualized the relationships between metabolite-microbe co-occurrences and (i) by ordinating the first three PCs and coloring metabolites by their beta values (e.g., Fig. 3c). Additional details are described in the Supplementary Material.

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 taxonomic), we used QIIME2's *diversity* plugin[32] to perform Mantel tests using Spearman correlations. Input distance matrices are described in the Supplementary Material.

Figures.

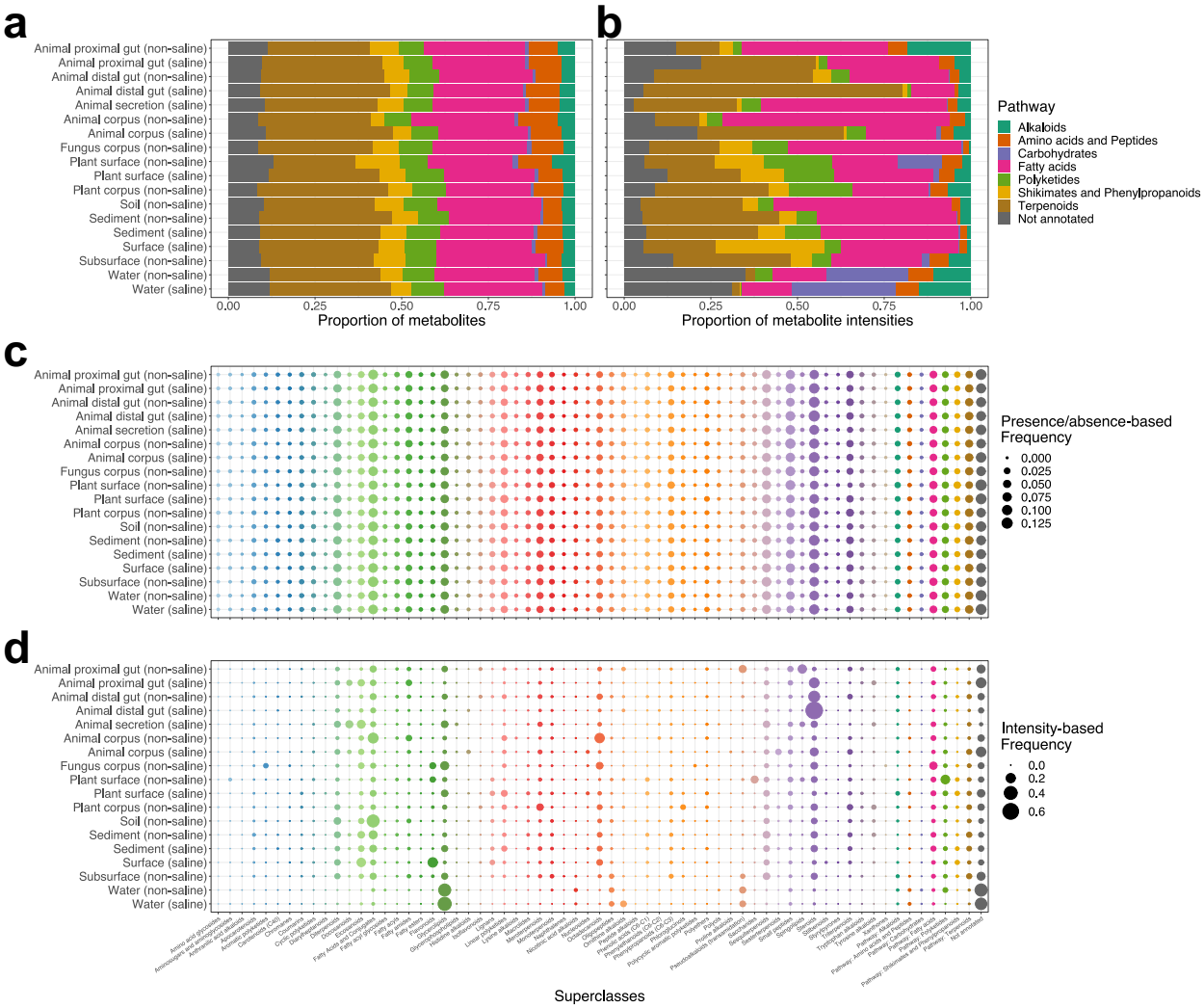

Fig. 1 | Distribution of microbially-related secondary metabolites chemical pathways among environments described using the Earth Microbiome Project Ontology (EMPO version 2, level 4). Both chemical pathway and chemical superclass annotations are shown based on presence/absence (**a, c**) and relative 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.

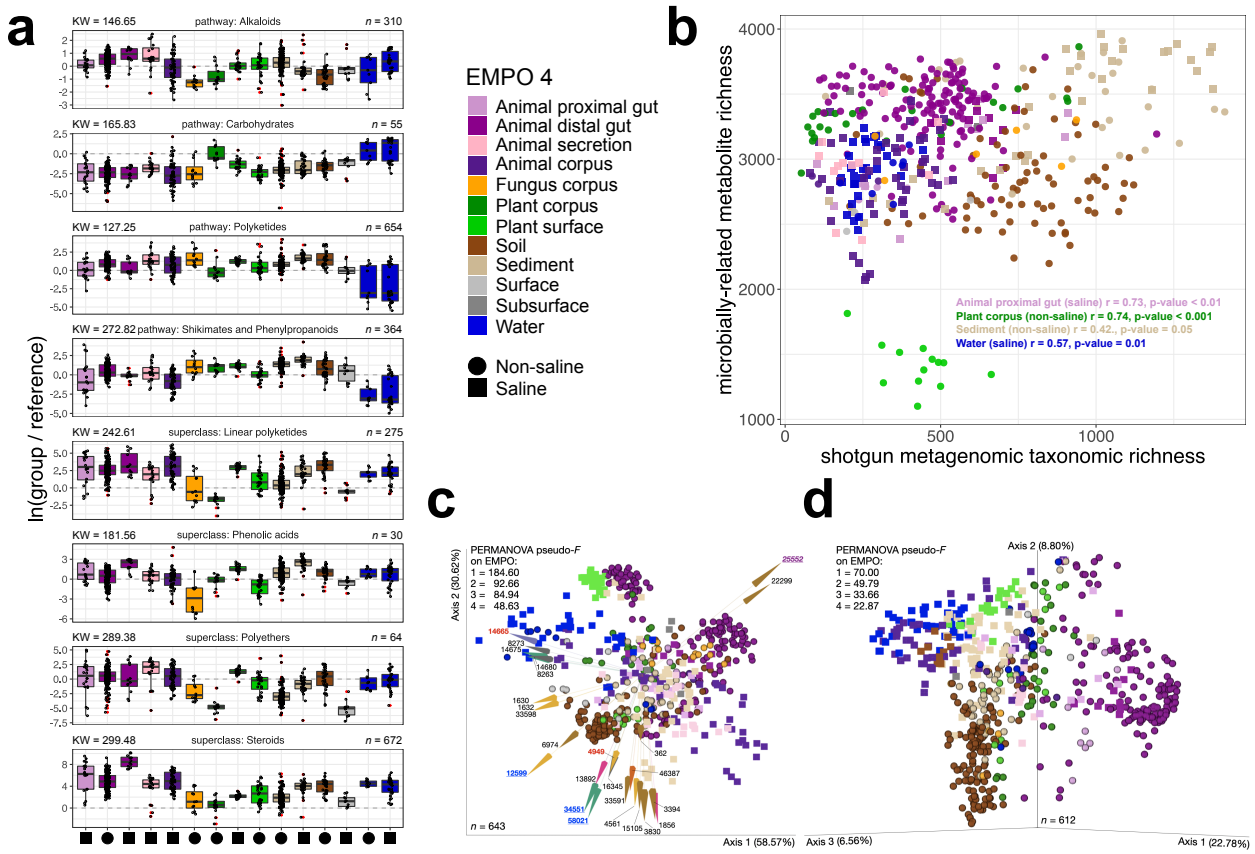

Fig. 2 | Structural-level associations between microbially-related secondary metabolites and specific environments described using the Earth Microbiome Project Ontology (EMPO version 2, level 4). **a,** Differential abundance of molecular features across environments, highlighting four example pathways 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., *Amino Acids and Peptides*, $n = 615$, for pathways and *Flavonoids*, $n = 42$, for superclasses). 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 features 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 pathways. 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 are those 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 from mapping metagenomic reads to the Web of Life database. 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 corpus'), to soils and sediments along Axis 2. Results from PERMANOVA for each level of EMPO are shown (i.e., all tests had *p*-value = 0.001).

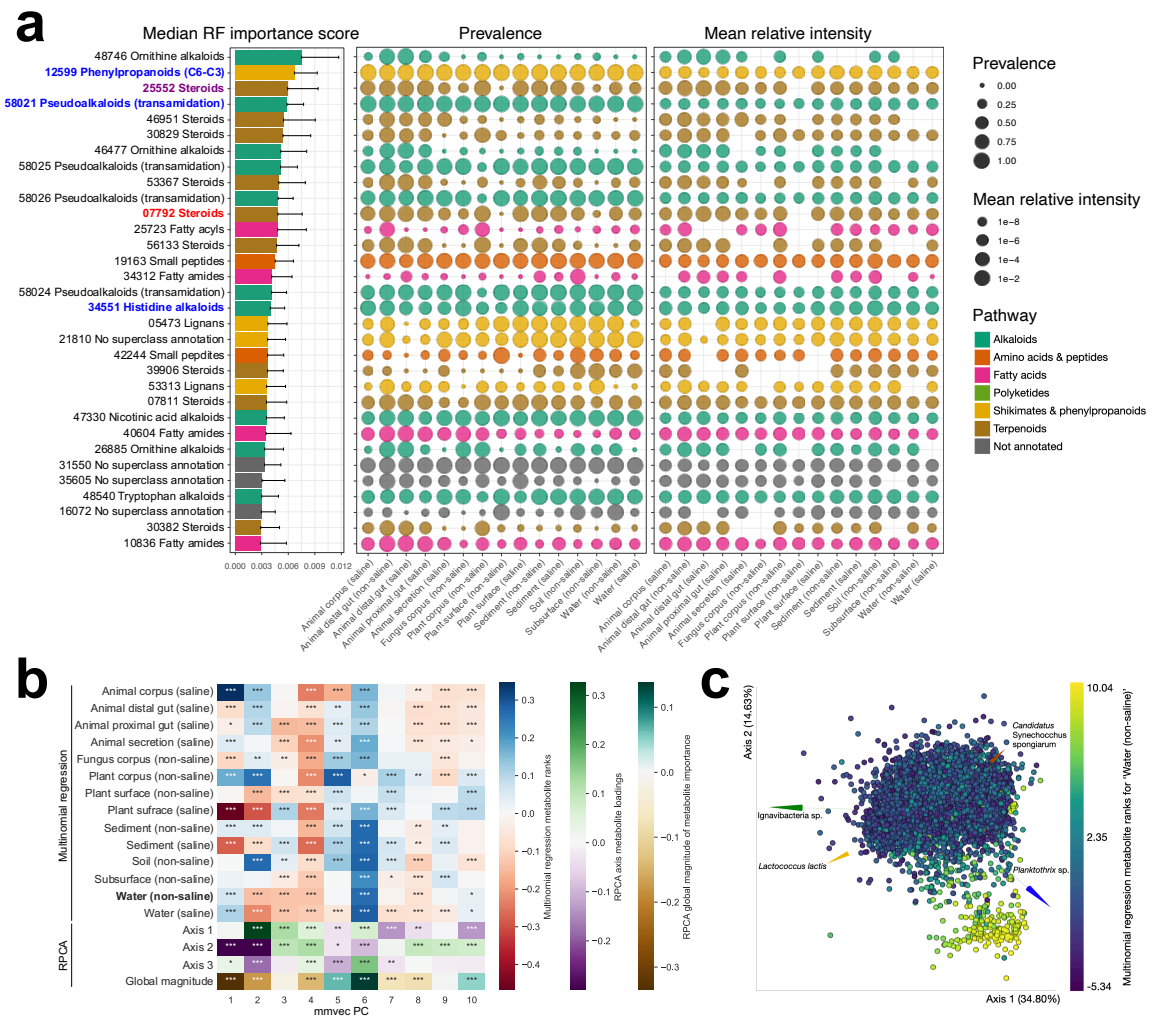

Fig. 3 | Microbially-related metabolites classify environments with a high accuracy and co-occur with specific microbial taxa. a, The random-forest importance score, environment-wide prevalence, and mean relative intensity for the top 32 most important microbially-related metabolites contributing to 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. 2c, Table S3). **b,** Co-occurrence analysis results

showing correlation between *mmvec* principal coordinates (PCs) and (i) multinomial regression betas for metabolite abundances across environments, (ii) axes from the RPCA biplot in Fig. 2d 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 features from the RPCA biplot in Fig. 2d. 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 abundances 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.

References.

1. Davies, J. Specialized microbial metabolites: functions and origins. *J. Antibiot.* **66**, 361–364 (2013).
2. Pham, J. V. *et al.* A Review of the Microbial Production of Bioactive Natural Products and Biologics. *Front. Microbiol.* **10**, 1404 (2019).
3. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0718-6.
4. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes - a review. *Nat. Prod. Rep.* **33**, 988–1005 (2016).
5. 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).
6. Hugerth, L. W. *et al.* Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biol.* **16**, 279 (2015).
7. 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).
8. 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).
9. 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.
10. 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).

11. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).

12. 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).

13. 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.

14. Mohimani, H. *et al.* Dereplication of microbial metabolites through database search of mass spectra. *Nat. Commun.* **9**, 4035 (2018).

15. 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).

16. 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).

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

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

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

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

21. 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).

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

23. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).

24. 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).

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

26. 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.

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

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

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

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

31. 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).

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

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

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

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

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

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

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

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

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

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, U01 AI124316), 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.

Jennifer L. Bowen[1], Max Chavarría[2,3], Don A. Cowan[4], Jaime Huerta-Cepas[5], Paul Jensen[6], Lingjing Jiang[7], Anton Lavrinienko[8], Thulani P. Makhalanyane[4], Tapio Mappes[8], Ezequiel M. Marzinelli[9], Sou Miyake[10], Timothy A. Mousseau[8], Catalina Murillo-Cruz[2], Adrián A. Pinto-Tomás[2], Dorota L. Porazinska[11], Jean-Baptiste Ramond[4,12], Taniya RoyChowdhury[13,14], Henning Seedorf[10,13], J. Reuben Shipway[16,17], Frank J. Stewart[18], Jana M. U'Ren[19], Phillip C. Watts[8], Nicole S. Webster[20,21], Jesse R. Zaneveld[22], Shan Zhang[23]

[1]Northeastern University, Boston, Massachusetts, USA. [2]University of Costa Rica, San José, Costa Rica. [3]CENIBiot, San José, Costa Rica. [4]University of Pretoria, Pretoria, South Africa. [5]Universidad Politécnica de Madrid, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Madrid, Spain. [6]University of California San Diego, La Jolla, California, USA. [7]Janssen Research & Development, San Diego, California, USA. [8]University of Jyväskylä, Jyväskylä, Finland. [9]The University of Sydney, Sydney, Australia. [10]Temasek Life Sciences Laboratory, Singapore, Singapore. [11]University of Florida, Gainesville, Florida, USA. [12]Pontificia Universidad Católica de Chile, Santiago, Chile. [13]Pacific Northwest National Laboratory, Richland, Washington, USA. [14]University of Maryland, College Park, Maryland, USA. [15]National University of Singapore, Singapore, Singapore. [16]University of Portsmouth, Portsmouth, United Kingdom. [17]University of Massachusetts Amherst, Amherst, Massachusetts, USA. [18]Montana State University, Bozeman, Montana, USA. [19]University of Arizona, Tucson, Arizona, USA. [20]Australian Institute of Marine Science, Townsville, Qld, Australia. [21]University

of Queensland, Brisbane, Australia. [22]University of Washington Bothell, Bothell, Washington, USA. [23]University of New South Wales, Sydney, Australia.

Competing interests.

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

Supplementary Material.

SUPPLEMENTARY 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://www.earthmicrobiome.org/protocols-and-standards/empo/)[1] and selected samples to fill out EMPO categories as broadly as possible. 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. 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 two minutes. 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 600 µL 100% H_2O was used as wash solvent. The

eluted washing solution was discarded, as it contains salts and very polar metabolites that subsequent metabolomics analysis is not designed to handle. The sample elution was carried out sequentially with 600 µL 7:3 MeOH:H_2O and with 600 µL 100% MeOH. 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 0.0 to 12.5 min. The automatic gain control (AGC) target and maximum injection time were set at 1 x 10^6 and 160 ms for MS^1 scans, and set at 5 x 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 a 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 *ca.* 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. Briefly, feature detection and deconvolution was performed with the ADAP chromatogram builder[8], and local minimum search algorithm. The isotopologues were regrouped, and the features (i.e., 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 3 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, the DEREPLICATOR tools were used via GNPS[11,12]. The MS^2 spectra were systematically annotated with the SIRIUS computational tools[13] computational annotation tool (v. 4.4.25, headless, Linux). 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 ([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 to 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 feature metadata file. 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 contained 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* to quantify beta-diversity (i.e., robust Aitchison distance)[23] 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 = 1E-05, 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. Two protocols were used for sample extraction, one for the 105 soil samples and another for the 356 fecal and sediment samples which 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 can be found at the following link: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d5c5135a59eb48779216615e8d5cb3ac The library search can be found at the following link: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=59b20fc8381f4ee6b79d35034de81d86

GC-MS data analysis. For multi-omic analyses including GC-MS data, we first removed noisy features by excluding those with GNPS balance scores <50. We then estimated beta-diversity for each dataset using robust Aitchison distances[23]. The final feature table for GC-MS contained 460 samples and 216 features that were used for subsequent analysis.

METAGENOMICS

DNA extraction. For each round of 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) (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), and resulting DNA 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 Crenarachaeota/Thamarchaeota[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] (https://msystems.asm.org/content/1/1/e00009-15). For the barcoded forward primer, constructs included (5' to 3'): the 5' Illumina adapter (AATGATACGGCGACCACCGAGATCTACACGCT), a Golay barcode (variable), 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 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 (variable), 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 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 (variable), 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' (ONT)'. The resulting product is 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 a custom bead purification protocol. 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. 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-omic analyses including amplicon sequence data, we processed each dataset for comparison of beta-diversity. 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 SATé-Enabled Phylogenetic Placement (SEPP)[56], to produce a phylogeny for diversity analyses. Reads (i.e., sOTUs) not placed during SEPP were removed from each feature-table. We then merged feature-tables across sequencing runs, excluded singleton sOTUs, rarefied to 5,000 features per sample, and estimated beta-diversity using unweighted UniFrac distances, using QIIME2's[22] *diversity* plugin. The final feature-table for 16S beta-diversity analysis had 681 samples and 93,260 features.

For 18S data, we first merged feature-tables across sequencing runs, and classified taxonomy for each sOTU via pre-fitted machine-learning classifiers[57], using the SILVA 138 reference database[58]. We then excluded singleton sOTUs and samples with a total frequency <3,000 features, and estimated beta-diversity for each dataset using robust Aitchison distances[23]. The final feature table for 18S beta-diversity analysis had 496 samples and 40,587 features.

For fungal ITS data separately, we first merged feature-tables across sequencing runs, and classified taxonomy for each sOTU as above but using UNITE 8 reference database[59]. We then excluded singleton sOTUs and samples with a total frequency <500 features, and estimated beta-diversity for each dataset using robust Aitchison distances[23]. The final feature table for fungal ITS beta-diversity analysis had 488 samples and 10,821 features.

For full-length rRNA operon data, per-sample fasta files were re-formatted for importing to QIIME2 as *SampleData[Sequences]* (i.e., with each header as '>{sample_identifier}_{sequence_identifier}), concatenated into a single fasta file, and imported. We then used QIIME2'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 excluded singleton OTUs and samples with a total frequency <500 features, and estimated beta-diversity for each dataset using robust Aitchison distances[23]. The final feature table for full-length rRNA operon beta-diversity analysis had 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 were added 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 diversity analyses contained 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 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 excluded singleton features and rarefied the data to 5,000 sequences per sample. The final feature-table for functional analysis 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, Faith's PD, and weighted Faith's PD), we performed correlation analysis to better understand their relationship. 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. For machine-learning analyses of LC-MS/MS and shotgun metagenomic taxonomic- and functional data, additional samples were filtered from feature-tables such to exclude environments (i.e., based on EMPO version 2, level 4) with relatively low sample representation (<10 samples).

For the LC-MS/MS feature-table, we excluded samples in the three EMPO environments (i.e., "Animal corpus (non-saline)", "Surface (saline)", "Animal proximal gut (non-saline)"), and the final feature-table included 606 samples (covering 15 EMPO environments), and 6,588 secondary metabolites.

For the shotgun metagenomic feature-table for taxonomic analysis, we used the rarefied table described above, but further excluded samples in the five lowest-representation environments (based on EMPO version 2, level 4); the final feature-table included 598 samples (representing 14 environments), and 8,567 features (i.e., OGUs).

For the shotgun metagenomic feature-table for functional analysis, we used the rarefied table described above, but further excluded the seven lowest-representation environments; the final feature-table included 485 samples (representing 12 environments) and 2,436 features (i.e., GO terms).

For each feature-table, we trained a Random Forest classifier for predicting environments with tuned hyperparameters with 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 Random Forest classifiers. We next evaluated the prediction performance with overall accuracy quantified from a confusion matrix based on all predictions in a 100-fold cross-validation. The overall accuracy of the classifiers was calculated from the confusion matrix based on all predictions in the 100-fold cross-validation. We evaluated how well the 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. 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 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.

We obtained co-occurrence probabilities using *mmvec*, 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 ordinating the first three PCs and coloring metabolites by their beta values (e.g., Fig. 3c).

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 taxonomic), we used QIIME2's *diversity* plugin[22] to perform Mantel tests using Spearman correlations. Input distance matrices are those described above for each dataset.

SUPPLEMENTARY FIGURES

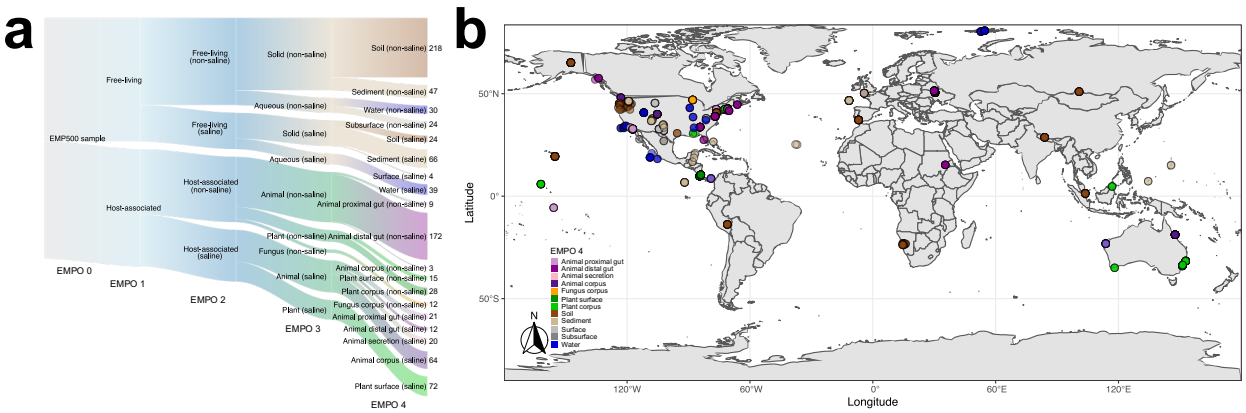

Figure S1 | a, Distribution of samples among the Earth Microbiome Project Ontology (EMPO version 2) categories. **b**, Geographic distribution of samples with points colored by EMPO (version 2, level 4). Extensive information about each sample set contributed is described in Table S1.

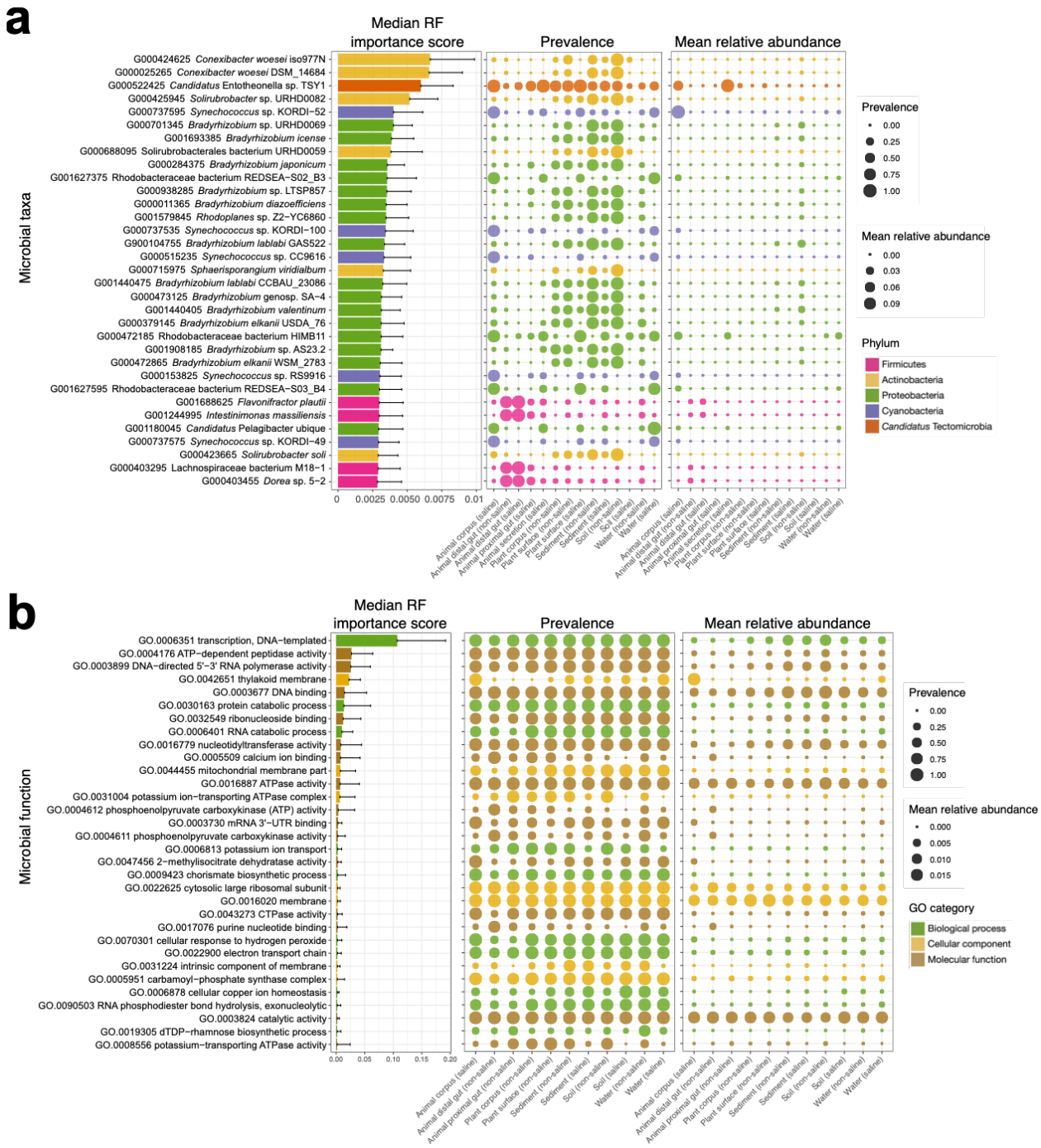

Figure S2 | Machine-learning analysis of shotgun metagenomics data highlighting the most important microbial taxa and functions. **a,** The random-forest (RF) importance score, environment-wide prevalence, and mean relative abundance for the top 32 most important microbial taxa contributing to the separation 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.

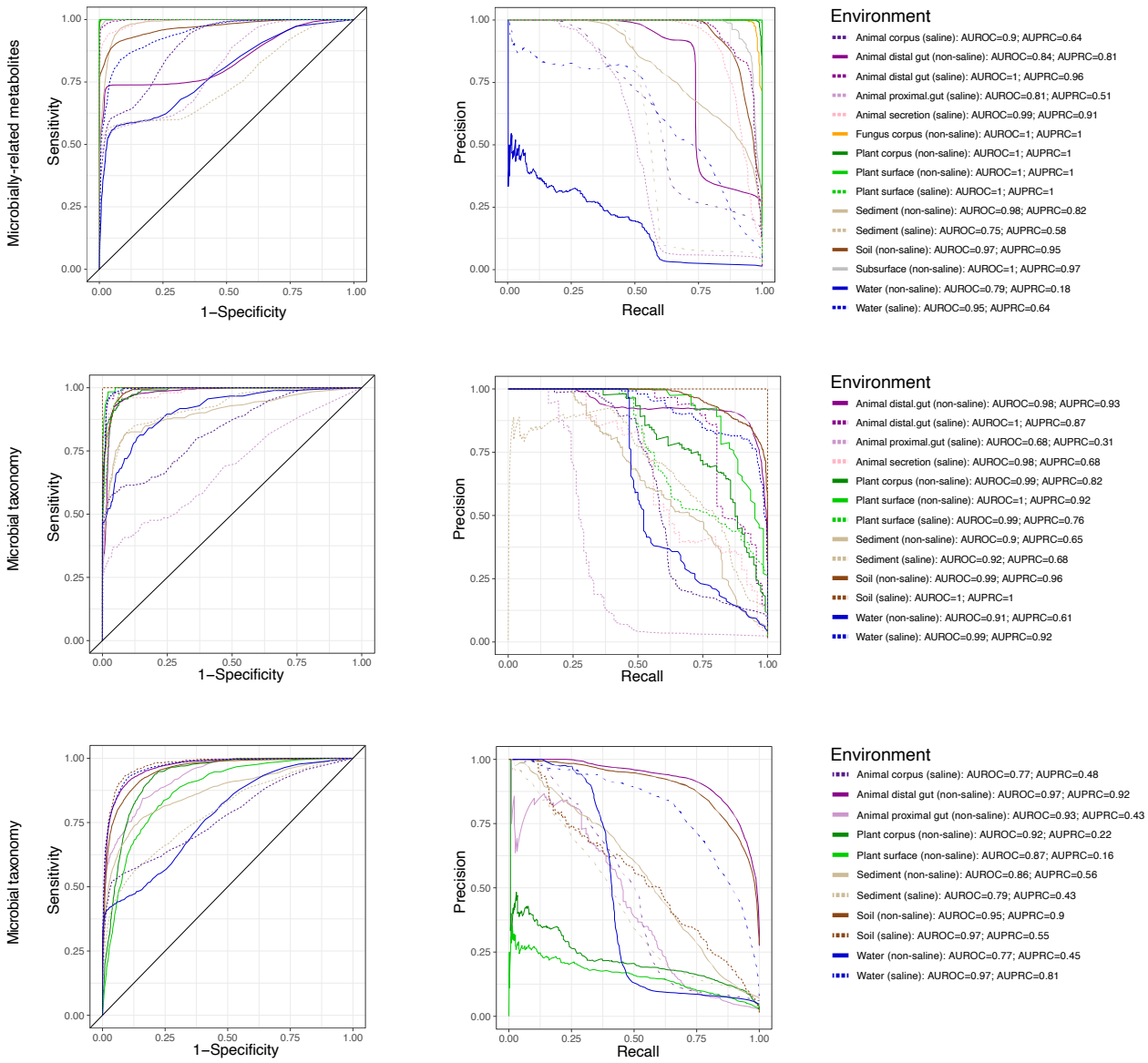

Figure S3 | Machine-learning analysis of LC-MS/MS metabolomics and shotgun metagenomics data highlighting per-environment classification accuracy. a, Receiver operating characteristic curves (and AUROC) and precision-recall curves (and AUPRC) illustrating classification accuracy of the random forest model across all environments based on microbial taxonomic (i.e., OGU) profiles. **b,** 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 microbial functional (i.e., GO terms)

profiles. 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.

SUPPLEMENTARY TABLES

Table S1. Study metadata, including study identifiers, principal investigator, title, sampling method, number of samples, environmental package, EMPO level 3 (version 1), and EMPO level 4 (version 2) for each study.

emp500_study_id	emp500_principal_investigator	emp500_title	sampling_method	number_of_samples	environmental_package	empo_3	empo_v2_4
1	Kshtrika	Biota oil	bulk	14	misc environment	Subsurface (non-saline)	Subsurface (non-saline)
2	Berry	Postglacial pond sediment profiles	bulk	20	sediment	Sediment (non-saline)	Sediment (non-saline)
3	MacRae-Crerar	Mongolia soils (NSF PIRE)	bulk	12	soil	Soil (non-saline)	Soil (non-saline)
5	Myrold	Tree-associated soils	bulk	12	soil	Soil (non-saline)	Soil (non-saline)
9	Zaneveld	Scleractinian corals	fractionated	22	host-associated, water	Animal secretion, Water (saline)	Animal secretion, Water (saline)
18	Thomas	Ocean sponges	fractionated	57	host-associated	Animal corpus	Animal corpus (saline)
19	Thomas	Australian algae	swab	64	host-associated	Plant surface	Plant surface (saline)
21	McMahon	Bioreactors for wastewater and anammox	bulk	17	wastewater/sludge	Water (non-saline)	Water (non-saline)
23	Shade	Centralia soil	bulk	10	soil	Soil (non-saline)	Soil (non-saline)
26	Stewart	Marine oxygen minimum zone water	filter	8	water	Water (saline)	Water (saline)
27	King	Puhimau Hawai'i volcanic soil	bulk	21	soil	Soil (non-saline)	Soil (non-saline)
33	Mayer	West Texas playa	bulk	7	sediment	Sediment (non-saline)	Sediment (non-saline)
34	Mayer	Animas watershed	bulk	10	sediment	Sediment (non-saline)	Sediment (non-saline)
36	Stegen	Permafrost soil	bulk	15	soil	Soil (non-saline)	Soil (non-saline)
37	Stegen	Hyporheic sediment	bulk	10	sediment	Sediment (non-saline)	Sediment (non-saline)
38	Stegen	Active layer soil	bulk	18	soil	Soil (non-saline)	Soil (non-saline)
39	Smith	Palmyra algae	bulk or swab	8	host-associated	Plant surface	Plant surface (saline)
40	Metcalf	Mouse decomposition soil	bulk	14	soil	Soil (non-saline)	Soil (non-saline)
42	Palenik	San Diego seawater	filter	20	water	Water (saline)	Water (saline)
43	Jensen	Ocean sediments	bulk	19	sediment	Sediment (saline)	Sediment (saline)
45	Roy-Chowdhury	Spanish hypersaline sandy soils	bulk	24	soil	Soil (non-saline)	Soil (non-saline)
46	Makhalanyane	Namibia gravel plain gradient soils	bulk	10	soil	Soil (non-saline)	Soil (non-saline)
47	Makhalanyane	Namibia dune aridity gradient soils	bulk	9	soil	Soil (non-saline)	Soil (non-saline)
50	Girguis	Whale feces	bulk	12	host-associated	Animal distal gut	Animal distal gut (saline)
51	Song	Captive terrestrial mammal feces	bulk	20	host-associated	Animal distal gut	Animal distal gut (non-saline)
52	Song	Captive herp feces	bulk	15	host-associated	Animal distal gut	Animal distal gut (non-saline)
53	Song	Captive bird feces	bulk	15	host-associated	Animal distal gut	Animal distal gut (non-saline)
54	Sandin	Fish gut contents	bulk	15	host-associated	Animal proximal gut	Animal proximal gut (saline)
56	Schmidt	Peru and Nepal alpine soils	bulk	26	soil	Soil (non-saline)	Soil (non-saline)
58	Tucker	Coal and coal-associated water	filter	16	misc environment	Subsurface (non-saline)	Subsurface (non-saline)
59	Myrold	Oregon soils	bulk	22	soil	Soil (non-saline)	Soil (non-saline)
62	Pinto	Costa Rican leaf-cutter ant soils	bulk	24	soil	Soil (non-saline)	Soil (non-saline)
63	Pinto	Costa Rican leaf-cutter ant and bee material	bulk	28	host-associated	Plant corpus	Plant corpus (non-saline)
65	Angenent	Rumen animals and bioreactor	bulk	9	host-associated	Animal proximal gut	Animal proximal gut (non-saline)
68	Bittleston	Pitcher plant fluids	bulk	15	plant-associated	Plant surface	Plant surface (non-saline)
71	Seedorf	Singapore plant-associated soils	bulk	14	soil	Soil (non-saline)	Soil (non-saline)
72	Distel	Shipworm tissue and wood	fractionated	15	host-associated	Animal corpus, Animal proximal gut, Surface (non-saline)	Animal corpus, Animal proximal gut, Surface (non-saline)
74	Bowen	Massachusetts salt marsh sediments	bulk	16	sediment	Sediment (saline)	Sediment (saline)
75	Song	Captive aquatic mammal feces	bulk	5	host-associated	Animal distal gut	Animal distal gut (non-saline)
76	Minich	Livestock feces	bulk	6	host-associated	Animal distal gut	Animal distal gut (non-saline)
77	Tait	English Channel water	filter	9	water	Water (saline)	Water (saline)
78	Tait	English Channel sediment	bulk	3	sediment	Sediment (saline)	Sediment (saline)
81	Metcalf	Mouse decomposition skin	bulk	3	host-associated	Animal corpus	Animal corpus (non-saline)
82	Uren	Michigan forest lichen	bulk	12	plant-associated	Fungus corpus	Fungus corpus (non-saline)
84	Rohwer	Franz Josef Land glacier soils and water	bulk	15	soil, water	Soil (non-saline), Water (non-saline)	Soil (non-saline), Water (non-saline)
85	Rohwer	Iberia and Guatemala deep-sea sediments	bulk	27	sediment	Sediment (saline)	Sediment (saline)
86	Rohwer	Mexico microbial mats	bulk	3	microbial mat/biofilm, sediment	Surface (saline), Sediment (saline)	Surface (saline), Sediment (saline)
88	Mousseau	Chernobyl vole feces	bulk	112	host-associated	Animal distal gut	Animal distal gut (non-saline)

Note: Scientific justification and metadata columns for each study are included in the version of the table here and will be included with the final Supplementary Material, but were excluded here due to space limitations.

Table S2. The top 10 ranked microbially-related metabolites for each environment, described using the Earth Microbiome Project Ontology (EMPO version 2, level 4). Data are from Songbird multinomial regression (model: *composition* = EMPO version 2, level 4; pseudo-Q^2 = 0.21). Metabolites in bold font are those also identified to be strongly associated with RPCA axes (Fig. 2c, Table S3), those in bold, underlined font also strongly separated environments in machine-learning analysis (Table S4), and those in bold, underlined, italic font identified as important in all three analyses.

Environment	Feature rank	Feature ID	Molecular formula	Pathway	Superclass	Class
Animal corpus (saline)	1	23235	C27H40O2	Terpenoids	Sesterterpenoids	Not annotated
Animal corpus (saline)	2	41006	Not annotated	Not annotated	Not annotated	Not annotated
Animal corpus (saline)	3	28910	C27H38	Terpenoids	Steroids	Cholestane steroids
Animal corpus (saline)	4	23136	C25H32O3	Terpenoids	Cyclic polyketides	Not annotated
Animal corpus (saline)	5	21140	C29H46O4	Terpenoids	Triterpenoids	Not annotated
Animal corpus (saline)	6	3559	Not annotated	Not annotated	Not annotated	Not annotated
Animal corpus (saline)	7	23517	Not annotated	Not annotated	Not annotated	Not annotated
Animal corpus (saline)	8	41001	Not annotated	Not annotated	Not annotated	Not annotated
Animal corpus (saline)	9	3955	Not annotated	Not annotated	Not annotated	Not annotated
Animal corpus (saline)	10	39515	Not annotated	Not annotated	Not annotated	Not annotated
Animal distal gut (non-saline)	1	25552	C24H38O4	Terpenoids	Steroids	Cholane steroids
Animal distal gut (non-saline)	2	7792	C24H38O4	Terpenoids	Steroids	Cholane steroids
Animal distal gut (non-saline)	3	25862	C27H42O7	Terpenoids	Steroids	Ecdysteroids
Animal distal gut (non-saline)	4	51039	C27H44O3	Terpenoids	Steroids	Cholestane steroids
Animal distal gut (non-saline)	5	2081	C15H10O4	Shikimates and Phenylpropanoids	Isoflavonoids	Isoflavones
Animal distal gut (non-saline)	6	1236	C18H34O5	Fatty acids	Octadecanoids	Other Octadecanoids
Animal distal gut (non-saline)	7	50823	C26H45NO6S	Terpenoids	Steroids	Cholane steroids
Animal distal gut (non-saline)	8	22299	C29H46O2	Terpenoids	Meroterpenoids	Prenyl quinone meroterpenoids
Animal distal gut (non-saline)	9	2826	C19H34O4	Fatty acids	Octadecanoids	Other Octadecanoids
Animal distal gut (non-saline)	10	12107	C24H41NO3	Terpenoids	Steroids	Cholane steroids
Animal distal gut (saline)	1	29157	Not annotated	Not annotated	Not annotated	Not annotated
Animal distal gut (saline)	2	38117	C40H52O4	Terpenoids	Carotenoids (C40)	Carotenoids (C40, β-β)
Animal distal gut (saline)	3	7942	C17H24N2O4	Alkaloids	Ornithine alkaloids	Not annotated
Animal distal gut (saline)	4	53059	C15H18BrO	Terpenoids	Sesquiterpenoids	Not annotated
Animal distal gut (saline)	5	30947	C7H9NO2	Shikimates and Phenylpropanoids	Not annotated	Shikimic acids and derivatives
Animal distal gut (saline)	6	20600	Not annotated	Not annotated	Not annotated	Not annotated
Animal distal gut (saline)	7	50368	C27H31N3O3	Alkaloids	Tryptophan alkaloids	Not annotated
Animal distal gut (saline)	8	52381	Not annotated	Not annotated	Not annotated	Not annotated
Animal distal gut (saline)	9	50979	C24H36O2S	Terpenoids	Steroids	Cholane steroids
Animal distal gut (saline)	10	12107	C24H41NO3	Terpenoids	Steroids	Cholane steroids
Animal proximal gut (saline)	1	46142	C19H39NO2S	Fatty acids	Fatty Acids and Conjugates	Not annotated
Animal proximal gut (saline)	2	50500	Not annotated	Not annotated	Not annotated	Not annotated
Animal proximal gut (saline)	3	46161	C19H39NO2S	Fatty acids	Fatty Acids and Conjugates	Not annotated
Animal proximal gut (saline)	4	50865	Not annotated	Not annotated	Not annotated	Not annotated
Animal proximal gut (saline)	5	8409	C22H35NO2	Fatty acids	Fatty amides	N-acyl ethanolamines (endocannabinoids)
Animal proximal gut (saline)	6	15710	Not annotated	Not annotated	Not annotated	Not annotated
Animal proximal gut (saline)	7	19189	Not annotated	Not annotated	Not annotated	Not annotated
Animal proximal gut (saline)	8	44429	Not annotated	Not annotated	Not annotated	Not annotated
Animal proximal gut (saline)	9	47998	Not annotated	Not annotated	Not annotated	Not annotated
Animal proximal gut (saline)	10	49293	C18H36N4O4	Amino acids and Peptides	Small peptides	Not annotated
Animal secretion (saline)	1	24213	C25H40O3	Fatty acids	Glycerolipids	Monoacylglycerols
Animal secretion (saline)	2	46948	C20H35N5O13	Carbohydrates	Polyols	Amino cyclitols
Animal secretion (saline)	3	24185	C20H33ClO4	Polyketides	Not annotated	Not annotated
Animal secretion (saline)	4	10248	C24H41NO5	Not annotated	Steroids	Not annotated
Animal secretion (saline)	5	16936	C33H52O13	Fatty acids	Glycerolipids	Glycosyldiacylglycerols
Animal secretion (saline)	6	24147	C25H40O3	Terpenoids	Steroids	Cholane steroids
Animal secretion (saline)	7	40413	C22H46NO8P	Fatty acids	Glycerophospholipids	Glycerophosphates
Animal secretion (saline)	8	44580	C11H21N3O	Alkaloids	Small peptides	Dipeptides
Animal secretion (saline)	9	22091	C28H50O9	Polyketides	Polyethers	Not annotated
Animal secretion (saline)	10	46073	C10H17N	Terpenoids	Monoterpenoids	Not annotated
Fungus corpus (non-saline)	1	20438	C18H10O7	Shikimates and Phenylpropanoids	Flavonoids	Flavonols
Fungus corpus (non-saline)	2	16650	C18H12O8	Shikimates and Phenylpropanoids	Flavonoids	Flavonols
Fungus corpus (non-saline)	3	16117	C25H43NO9	Fatty acids	Glycerolipids	Glycosyldiacylglycerols
Fungus corpus (non-saline)	4	16952	Not annotated	Not annotated	Not annotated	Not annotated
Fungus corpus (non-saline)	5	16752	C18H30Cl2O2	Fatty acids	Fatty Acids and Conjugates	Not annotated
Fungus corpus (non-saline)	6	16681	C18H12O6	Shikimates and Phenylpropanoids	Flavonoids	Not annotated
Fungus corpus (non-saline)	7	16632	C17H10O6	Polyketides	Xanthones	Methyl xanthones
Fungus corpus (non-saline)	8	16615	C18H12O7	Shikimates and Phenylpropanoids	Not annotated	Not annotated
Fungus corpus (non-saline)	9	16743	C18H12O8	Shikimates and Phenylpropanoids	Flavonoids	Flavonols
Fungus corpus (non-saline)	10	15767	C20H34O2	Fatty acids	Fatty Acids and Conjugates	Not annotated

Table S2. (continued)

Environment	Feature rank	Feature ID	Molecular formula	Pathway	Superclass	Class
Fungus corpus (non-saline)	1	20438	C18H10O7	Shikimates and Phenylpropanoids	Flavonoids	Flavonols
Fungus corpus (non-saline)	2	16650	C18H12O8	Shikimates and Phenylpropanoids	Flavonoids	Flavonols
Fungus corpus (non-saline)	3	16117	C25H43NO9	Fatty acids	Glycerolipids	Glycosyldiacylglycerols
Fungus corpus (non-saline)	4	16952	Not annotated	Not annotated	Not annotated	Not annotated
Fungus corpus (non-saline)	5	16752	C18H30Cl2O2	Fatty acids	Fatty Acids and Conjugates	Not annotated
Fungus corpus (non-saline)	6	16681	C18H12O6	Shikimates and Phenylpropanoids	Flavonoids	Not annotated
Fungus corpus (non-saline)	7	16632	C17H10O6	Polyketides	Xanthones	Methyl xanthones
Fungus corpus (non-saline)	8	16615	C18H12O7	Shikimates and Phenylpropanoids	Not annotated	Not annotated
Fungus corpus (non-saline)	9	16743	C18H12O8	Shikimates and Phenylpropanoids	Flavonoids	Flavonols
Fungus corpus (non-saline)	10	15767	C20H34O2	Fatty acids	Fatty Acids and Conjugates	Not annotated
Plant corpus (non-saline)	1	24670	C33H40O4	Polyketides	Meroterpenoids	Not annotated
Plant corpus (non-saline)	2	24678	C13H16BNO7	Amino acids and Peptides	Not annotated	Not annotated
Plant corpus (non-saline)	3	24692	C33H48O4	Polyketides	Meroterpenoids	Not annotated
Plant corpus (non-saline)	4	7395	C22H30O3	Terpenoids	Meroterpenoids	Cannabinoids
Plant corpus (non-saline)	5	24684	Not annotated	Not annotated	Not annotated	Not annotated
Plant corpus (non-saline)	6	24677	C24H26O4	Polyketides	Phloroglucinols	Acyl phloroglucinols
Plant corpus (non-saline)	7	30111	C27H42O6	Terpenoids	Steroids	Ecdysteroids
Plant corpus (non-saline)	8	24866	C19H18O5	Polyketides	Phloroglucinols	Not annotated
Plant corpus (non-saline)	9	26285	C28H42	Terpenoids	Steroids	Vitamin D2 and derivatives
Plant corpus (non-saline)	10	31735	C24H39N5O8	Amino acids and Peptides	Small peptides	Tripeptides
Plant surface (non-saline)	1	11326	C24H41NO4	Terpenoids	Steroids	Cholane steroids
Plant surface (non-saline)	2	36400	C32H54O4	Polyketides	Meroterpenoids	Not annotated
Plant surface (non-saline)	3	20775	C20H41NO4	Fatty acids	Not annotated	Other Octadecanoids
Plant surface (non-saline)	4	4949	C13H10O	Shikimates and Phenylpropanoids	Flavonoids	Chalcones
Plant surface (non-saline)	5	30751	C9H15NO4	Polyketides	Cyclic polyketides	2-pyrone derivatives
Plant surface (non-saline)	6	5118	C21H24O3	Terpenoids	Diterpenoids	Tetracyclic diterpenoids
Plant surface (non-saline)	7	20098	C7H8NO4	Amino acids and Peptides	Small peptides	Aminoacids
Plant surface (non-saline)	8	45169	C26H41NO8	Alkaloids	Not annotated	Not annotated
Plant surface (non-saline)	9	38866	Not annotated	Not annotated	Not annotated	Not annotated
Plant surface (non-saline)	10	21815	C30H50N6O4	Amino acids and Peptides	Oligopeptides	Lipopeptides
Plant surface (saline)	1	55614	C8H13N3O	Alkaloids	Nicotinic acid alkaloids	Pyridine alkaloids
Plant surface (saline)	2	55975	C22H41N3O8	Polyketides	Not annotated	Not annotated
Plant surface (saline)	3	56485	Not annotated	Not annotated	Not annotated	Not annotated
Plant surface (saline)	4	56324	C21H37N3O7	Polyketides	Not annotated	Not annotated
Plant surface (saline)	5	34407	C10H19N3O2	Alkaloids	Not annotated	Primary amides
Plant surface (saline)	6	56318	C10H17N3O	Alkaloids	Monoterpenoids	Pyridine alkaloids
Plant surface (saline)	7	34652	C10H19N3O2	Alkaloids	Not annotated	Primary amides
Plant surface (saline)	8	56548	C23H40O8	Fatty acids	Linear polyketides	Open-chain polyketides
Plant surface (saline)	9	56014	C22H39N3O7	Amino acids and Peptides	Not annotated	Not annotated
Plant surface (saline)	10	56496	C21H41N3O7	Polyketides	Linear polyketides	Open-chain polyketides
Sediment (non-saline)	1	15024	C13H24N6O7	Carbohydrates	Not annotated	Not annotated
Sediment (non-saline)	2	4702	C9H16O5	Fatty acids	Glycerolipids	Not annotated
Sediment (non-saline)	3	28221	C36H54O8	Terpenoids	Diterpenoids	Tigliane diterpenoids
Sediment (non-saline)	4	8030	C20H30	Terpenoids	Steroids	Not annotated
Sediment (non-saline)	5	21379	C25H26O5	Shikimates and Phenylpropanoids	Coumarins	Not annotated
Sediment (non-saline)	6	5882	C9H6O2	Shikimates and Phenylpropanoids	Phenylpropanoids (C6-C3)	Cinnamic acids and derivatives
Sediment (non-saline)	7	1721	C27H38O5	Terpenoids	Not annotated	Not annotated
Sediment (non-saline)	8	11099	C19H34O5	Fatty acids	Linear polyketides	Open-chain polyketides
Sediment (non-saline)	9	25351	Not annotated	Not annotated	Not annotated	Not annotated
Sediment (non-saline)	10	15086	C12H18O6	Polyketides	Not annotated	Not annotated
Sediment (saline)	1	12617	C27H47NO13	Fatty acids	Not annotated	Not annotated
Sediment (saline)	2	42202	Not annotated	Not annotated	Not annotated	Not annotated
Sediment (saline)	3	10361	C42H56O5	Terpenoids	Carotenoids (C40)	Carotenoids (C40, β-β)
Sediment (saline)	4	15118	C19H35NO9	Fatty acids	Not annotated	Not annotated
Sediment (saline)	5	46984	C21H34O3	Terpenoids	Steroids	Cholane steroids
Sediment (saline)	6	33113	C32H57NO9	Polyketides	Linear polyketides	Open-chain polyketides
Sediment (saline)	7	12537	C40H58O5	Terpenoids	Carotenoids (C40)	Carotenoids (C40, β-β)
Sediment (saline)	8	12862	C21H34O3	Fatty acids	Eicosanoids	Not annotated
Sediment (saline)	9	12623	C28H49NO13	Fatty acids	Not annotated	Not annotated
Sediment (saline)	10	12750	Not annotated	Not annotated	Not annotated	Not annotated

Soil (non-saline)	1	2102	C18H37NO2	Fatty acids	Fatty Acids and Conjugates	Unsaturated fatty acids
Soil (non-saline)	2	10177	C32H52N8O5	Amino acids and Peptides	Oligopeptides	Linear peptides
Soil (non-saline)	3	2073	C17H19NO2	Alkaloids	Not annotated	Not annotated
Soil (non-saline)	4	2217	C13H21ClN2O3	Fatty acids	Not annotated	Not annotated
Soil (non-saline)	5	2161	C32H51N5O5	Amino acids and Peptides	Oligopeptides	Cyclic peptides
Soil (non-saline)	6	7661	C20H39NO	Fatty acids	Fatty amides	Primary amides
Soil (non-saline)	7	2135	C18H18N2O	Alkaloids	Tryptophan alkaloids	Simple indole alkaloids
Soil (non-saline)	8	2288	C11H12N2O2	Alkaloids	Tryptophan alkaloids	Simple indole alkaloids
Soil (non-saline)	9	7420	C13H19FOS2	Terpenoids	Monoterpenoids	Not annotated
Soil (non-saline)	10	5153	C20H22O4	Shikimates and Phenylpropanoids	Lignans	Dibenzylbutane lignans
Subsurface (non-saline)	1	14373	C14H19NO	Terpenoids	Sesquiterpenoids	Not annotated
Subsurface (non-saline)	2	14208	C8H12O4	Fatty acids	Fatty Acids and Conjugates	Dicarboxylic acids
Subsurface (non-saline)	3	42569	Not annotated	Not annotated	Not annotated	Not annotated
Subsurface (non-saline)	4	17757	C20H2OO	Terpenoids	Steroids	Not annotated
Subsurface (non-saline)	5	6222	C18H22	Terpenoids	Steroids	Estrane steroids
Subsurface (non-saline)	6	27023	C17H32O5	Fatty acids	Glycerolipids	Monoacylglycerols
Subsurface (non-saline)	7	14337	Not annotated	Not annotated	Not annotated	Not annotated
Subsurface (non-saline)	8	15209	C16H20O2S	Terpenoids	Steroids	Not annotated
Subsurface (non-saline)	9	2225	C14H21NO	Fatty acids	Fatty amides	Phenylalanine-derived alkaloids
Subsurface (non-saline)	10	9506	C17H18O	Terpenoids	Not annotated	Not annotated

Table S2. (continued)

Environment	Feature rank	Feature ID	Molecular formula	Pathway	Superclass	Class
Water (non-saline)	1	14668	C30H62O16	Carbohydrates	Glycerolipids	Not annotated
Water (non-saline)	2	14675	C18H33N7O5	Alkaloids	Pseudoalkaloids (transamidation)	Not annotated
Water (non-saline)	3	14677	C32H66O17	Carbohydrates	Glycerolipids	Not annotated
Water (non-saline)	4	14832	C19H31BN6O4	Alkaloids	Not annotated	Not annotated
Water (non-saline)	5	14725	C18H38O10	Fatty acids	Glycerolipids	Diacylglycerols
Water (non-saline)	6	14693	C22H46O12	Fatty acids	Glycerolipids	Diacylglycerols
Water (non-saline)	7	14821	C14H28O7	Fatty acids	Not annotated	Not annotated
Water (non-saline)	8	14665	C28H58O15	Carbohydrates	Glycerolipids	Not annotated
Water (non-saline)	9	15877	Not annotated	Not annotated	Not annotated	Not annotated
Water (non-saline)	10	14797	Not annotated	Not annotated	Not annotated	Not annotated
Water (saline)	1	14668	C30H62O16	Carbohydrates	Glycerolipids	Not annotated
Water (saline)	2	14677	C32H66O17	Carbohydrates	Glycerolipids	Not annotated
Water (saline)	3	14832	C19H31BN6O4	Alkaloids	Not annotated	Not annotated
Water (saline)	4	14665	C28H58O15	Carbohydrates	Glycerolipids	Not annotated
Water (saline)	5	14693	C22H46O12	Fatty acids	Glycerolipids	Diacylglycerols
Water (saline)	6	14675	C18H33N7O5	Alkaloids	Pseudoalkaloids (transamidation)	Not annotated
Water (saline)	7	15877	Not annotated	Not annotated	Not annotated	Not annotated
Water (saline)	8	14785	Not annotated	Not annotated	Not annotated	Not annotated
Water (saline)	9	14844	C26H55NO13	Carbohydrates	Aminosugars and aminoglycosides	Not annotated
Water (saline)	10	14724	C24H50O13	Fatty acids	Glycerolipids	Diacylglycerols

Table S3. The top 25 microbially-related metabolites across all environments, ranked based on association with RPCA ordination axes (i.e., global magnitude across all axes). Metabolites in bold font are those also identified to be strongly associated with particular environments from multinomial regression (Table S2), those in bold, underlined font also strongly separated environments in machine-learning analysis (Table S4), and those in bold, underlined, italic font identified as important in all three analyses.

Feature ID	Molecular formula	Pathway	Superclass	Class
362	C17H24O3	Terpenoids	Sesquiterpenoids	Not annotated
1630	C9H8O2	Shikimates and Phenylpropanoids	Phenylpropanoids (C6-C3)	Cinnamic acids and derivatives
1632	C9H8O2	Shikimates and Phenylpropanoids	Phenolic acids (C6-C1)	Simple phenolic acids
1856	C22H41NO	Fatty acids	Fatty amides	Primary amides
3394	C28H38	Terpenoids	Steroids	Vitamin D2 and derivatives
3830	C28H38	Terpenoids	Steroids	Vitamin D2 and derivatives
4561	C9H6O2	Shikimates and Phenylpropanoids	Phenylpropanoids (C6-C3)	Cinnamic acids and derivatives
4949	**C13H10O**	**Shikimates and Phenylpropanoids**	**Flavonoids**	**Chalcones**
6974	C14H24O5	Terpenoids	Not annotated	Not annotated
8263	Not annotated	Not annotated	Not annotated	Not annotated
8273	Not annotated	Not annotated	Not annotated	Not annotated
12599	**C10H10O2**	**Shikimates and Phenylpropanoids**	**Phenylpropanoids (C6-C3)**	**Cinnamic acids and derivatives**
13892	C8H18O5	Fatty acids	Glycerolipids	Diacylglycerols
14665	**C28H58O15**	**Carbohydrates**	**Glycerolipids**	**Not annotated**
14675	C18H33N7O5	Alkaloids	Pseudoalkaloids (transamidation)	Not annotated
14680	C22H35N5O5	Alkaloids	Pseudoalkaloids (transamidation)	Not annotated
15105	C26H37NO6	Terpenoids	Diterpenoids	Tigliane diterpenoids
16345	Not annotated	Not annotated	Not annotated	Not annotated
22299	C29H46O2	Terpenoids	Meroterpenoids	Prenyl quinone meroterpenoids
25552	**C24H34O2**	**Terpenoids**	**Steroids**	**Cholane steroids**
33591	C19H30O2	Terpenoids	Diterpenoids	Androstane steroids
33598	C9H8O2	Shikimates and Phenylpropanoids	Phenolic acids (C6-C1)	Simple phenolic acids
34551	**C10H17N3O2**	**Alkaloids**	**Histidine alkaloids**	**Imidazole alkaloids**
46387	C19H35N7O3S	Amino acids and Peptides	Not annotated	Not annotated
58021	**C8H10N4O2**	**Alkaloids**	**Pseudoalkaloids (transamidation)**	**Purine alkaloids**

Table S4. The top 32 most important metabolites across all metabolites contributing to separation of environments, identified in random forest machine-learning analysis. Metabolites in bold font are those also identified to be strongly associated with particular environments from multinomial regression (Table S2), those in bold, underlined font also identified to be strongly associated with RPCA axes (Fig. 2c, Table S3), and those in bold, underlined, italic font identified as important in all three analyses.

Feature ID	Molecular formula	Pathway	Superclass	Class
5473	C20H18O6	Shikimates and Phenylpropanoids	Lignans	Not annotated
7792	C24H36O3	Terpenoids	Steroids	Cholane steroids
7811	C24H36O3	Terpenoids	Steroids	Cholane steroids
10836	C24H45NO3	Fatty acids	Fatty amides	N-acyl amines
12599	**C10H10O2**	**Shikimates and Phenylpropanoids**	**Phenylpropanoids (C6-C3)**	**Cinnamic acids and derivatives**
16072	Not annotated	Not annotated	Not annotated	Not annotated
19163	C10H14N6O4	Amino acids and Peptides	Small peptides	Dipeptides
21810	C18H18O5	Shikimates and Phenylpropanoids	Not annotated	Not annotated
25552	**C24H34O2**	**Terpenoids**	**Steroids**	**Cholane steroids**
25723	C18H39NO4	Fatty acids	Fatty acyls	Fatty alcohols
26885	C11H23N3O2	Alkaloids	Ornithine alkaloids	Polyamines
30382	C27H38O	Terpenoids	Steroids	Cholestane steroids
30829	C27H40O	Terpenoids	Steroids	Not annotated
31550	Not annotated	Not annotated	Not annotated	Not annotated
34312	C24H45NO2	Fatty acids	Fatty amides	Not annotated
34551	**C10H17N3O2**	**Alkaloids**	**Histidine alkaloids**	**Imidazole alkaloids**
35605	Not annotated	Not annotated	Not annotated	Not annotated
39906	C27H39NO	Terpenoids	Steroids	Not annotated
40604	C27H43NO3	Fatty acids	Fatty amides	N-acyl amines
42244	C12H20N2O4	Amino acids and Peptides	Small peptides	Dipeptides
46477	C17H22N2O4	Alkaloids	Ornithine alkaloids	Not annotated
46951	C27H47NO5	Terpenoids	Steroids	Cholestane steroids
47330	C8H14N2	Alkaloids	Nicotinic acid alkaloids	Pyridine alkaloids
48540	C14H18N2O	Alkaloids	Tryptophan alkaloids	Simple indole alkaloids
48746	C17H24N2O4	Alkaloids	Ornithine alkaloids	Not annotated
53313	C18H16O2	Shikimates and Phenylpropanoids	Lignans	Not annotated
53367	C27H44	Terpenoids	Steroids	Cholestane steroids
56133	C27H42O	Terpenoids	Steroids	Cholestane steroids
58021	**C8H10N4O2**	**Alkaloids**	**Pseudoalkaloids (transamidation)**	**Purine alkaloids**
58024	C8H10N4O2	Alkaloids	Pseudoalkaloids (transamidation)	Purine alkaloids
58025	C8H10N4O2	Alkaloids	Pseudoalkaloids (transamidation)	Purine alkaloids
58026	C8H10N4O2	Alkaloids	Pseudoalkaloids (transamidation)	Purine alkaloids

Table S5. Alpha-diversity correlations comparing microbially-related metabolite richness vs. either shotgun metagenomic microbial taxonomic richness, Faith's PD, or weighted Faith's PD. Values are Spearman correlation coefficients, and p-values were adjusted using the Benjamini-Hochberg correction.

Environment	*n*	shotgun_metric	*r*	lower 95% CI	upper 95% CI	*t*	*S*	*p*-value
Animal corpus (non-saline)	3	richness	N/A	N/A	N/A	N/A	N/A	N/A
Animal corpus (saline)	44	richness	**0.39**	0.11	0.62	2.74	N/A	**0.01**
Animal distal gut (non-saline)	139	richness	3.43E-04	-0.17	0.17	4.01E-03	N/A	1
Animal distal gut (saline)	11	richness	-0.12	-0.68	0.53	N/A	274.06	0.7
Animal proximal gut (non-saline)	5	richness	-0.21	-0.93	0.84	N/A	24.1	0.7
Animal proximal gut (saline)	12	richness	**0.73**	0.26	0.92	N/A	76	**0.01**
Animal secretion (saline)	12	richness	0.05	-0.55	0.62	N/A	272	0.9
Fungus corpus (non-saline)	6	richness	0.43	-0.61	0.93	N/A	20	0.4
Plant corpus (non-saline)	26	richness	**0.74**	0.49	0.88	N/A	750.63	**0.001**
Plant surface (non-saline)	13	richness	-0.36	-0.77	0.26	N/A	494	0.2
Plant surface (saline)	0	richness	N/A	N/A	N/A	N/A	N/A	N/A
Sediment (non-saline)	31	richness	**0.42**	0.07	0.67	2.47	N/A	**0.05**
Sediment (saline)	31	richness	0.27	-0.09	0.57	1.52	N/A	0.1
Soil (non-saline)	78	richness	0.06	-0.16	0.28	0.54	N/A	0.6
Soil (saline)	0	richness	N/A	N/A	N/A	N/A	N/A	N/A
Subsurface (non-saline)	3	richness	N/A	N/A	N/A	N/A	N/A	**N/A**
Surface (saline)	2	richness	N/A	N/A	N/A	N/A	N/A	N/A
Water (non-saline)	8	richness	0.24	-0.56	0.81	0.6	N/A	0.6
Water (saline)	30	richness	**0.57**	0.26	0.77	3.65	N/A	**0.01**
Animal corpus (non-saline)	3	Faith's PD	N/A	N/A	N/A	N/A	N/A	N/A
Animal corpus (saline)	44	Faith's PD	**0.29**	-0.01	0.54	1.98	N/A	**0.05**
Animal distal gut (non-saline)	139	Faith's PD	0.02	-0.15	0.18	0.19	N/A	0.8
Animal distal gut (saline)	11	Faith's PD	-0.05	-0.64	0.58	N/A	230	0.9
Animal proximal gut (non-saline)	5	Faith's PD	0.1	-0.87	0.91	N/A	18	0.9
Animal proximal gut (saline)	12	Faith's PD	**0.72**	0.23	0.92	N/A	80	**0.01**
Animal secretion (saline)	12	Faith's PD	0.17	-0.46	0.69	N/A	238	0.6
Fungus corpus (non-saline)	6	Faith's PD	0.43	-0.61	0.93	N/A	20	0.4
Plant corpus (non-saline)	26	Faith's PD	**0.74**	0.49	0.88	N/A	760	**0.001**
Plant surface (non-saline)	13	Faith's PD	-0.41	-0.79	0.2	N/A	512	0.2
Plant surface (saline)	0	Faith's PD	N/A	N/A	N/A	N/A	N/A	N/A
Sediment (non-saline)	31	Faith's PD	**0.44**	0.1	0.69	2.65	N/A	**0.05**
Sediment (saline)	31	Faith's PD	0.09	-0.28	0.43	0.47	N/A	0.6
Soil (non-saline)	78	Faith's PD	8.73E-04	-0.22	0.22	7.61E-03	N/A	1
Soil (saline)	0	Faith's PD	N/A	N/A	N/A	N/A	N/A	N/A
Subsurface (non-saline)	3	Faith's PD	N/A	N/A	N/A	N/A	N/A	N/A
Surface (saline)	2	Faith's PD	N/A	N/A	N/A	N/A	N/A	N/A
Water (non-saline)	8	Faith's PD	-0.24	-0.81	0.56	-0.6	N/A	0.6
Water (saline)	30	Faith's PD	**0.6**	0.31	0.79	4.01	N/A	**0.001**
Animal corpus (non-saline)	3	weighted Faith's PD	N/A	N/A	N/A	N/A	N/A	N/A
Animal corpus (saline)	44	weighted Faith's PD	-0.13	-0.41	0.17	-0.87	N/A	0.4
Animal distal gut (non-saline)	139	weighted Faith's PD	0.05	-0.11	0.22	0.64	N/A	0.5
Animal distal gut (saline)	11	weighted Faith's PD	-0.39	-0.81	0.29	N/A	306	0.2
Animal proximal gut (non-saline)	5	weighted Faith's PD	0.1	-0.87	0.91	N/A	18	0.9
Animal proximal gut (saline)	12	weighted Faith's PD	**0.53**	-0.08	0.85	N/A	134	**0.08**
Animal secretion (saline)	12	weighted Faith's PD	0.31	-0.34	0.76	N/A	198	0.3
Fungus corpus (non-saline)	6	weighted Faith's PD	0.6	-0.44	0.95	N/A	14	0.2
Plant corpus (non-saline)	26	weighted Faith's PD	**0.46**	0.07	0.72	N/A	1585	**0.05**
Plant surface (non-saline)	13	weighted Faith's PD	-0.45	-0.81	0.15	N/A	528	0.12
Plant surface (saline)	0	weighted Faith's PD	N/A	N/A	N/A	N/A	N/A	N/A
Sediment (non-saline)	31	weighted Faith's PD	**0.35**	0	0.63	2.02	N/A	**0.05**
Sediment (saline)	31	weighted Faith's PD	-0.19	-0.51	0.17	-0.105	N/A	0.3
Soil (non-saline)	78	weighted Faith's PD	-6.16E+03	-0.23	0.22	-0.05	N/A	1
Soil (saline)	0	weighted Faith's PD	N/A	N/A	N/A	N/A	N/A	N/A
Subsurface (non-saline)	3	weighted Faith's PD	N/A	N/A	N/A	N/A	N/A	N/A
Surface (saline)	2	weighted Faith's PD	N/A	N/A	N/A	N/A	N/A	N/A
Water (non-saline)	8	weighted Faith's PD	0.12	-0.64	0.76	0.29	N/A	0.8
Water (saline)	30	weighted Faith's PD	**0.42**	0.08	0.68	2.48	N/A	**0.05**

Table S6. 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), as well as the 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

SUPPLEMENTARY NOTES

Permits for sample collection.

For samples from Costa Rica, permits were granted by the Institutional Biodiversity Commission of the University of Costa Rica (UCR, resolution number 055-2016) and authorized by the Organization for Tropical Studies (OTS) and the Central Pacific Conservation Area (ACOPAC) of the Ministry of Energy and the Environment (MINAE), Costa Rican government, under UCR project B6-656.

Data availability.

The mass spectrometry method and data (.RAW and .mzML) were deposited on the MassIVE public repository and are available under the dataset accession number MSV000083475. The processing files were also added to the deposition (updates/2019-08-21_lfnothias_7cc0af40/other/1908_EMPv2_INN/). GNPS molecular networking job is available at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=929ce9411f684cf8abd009670b293a33 and was also performed in analogue mode https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=fafdbfc058184c2b8c87968a7c56d7aa. The *DEREPLICATOR* jobs can be accessed here: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ee40831bcc314bda928886964d853a52 and https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1fafd4d4fe7e47dd9dd0b3d8bb0e6606. The SIRIUS results are available on the GitHub repository [`emp/data/metabolomics/FBMN/SIRIUS`]. The notebooks for metabolomics data preparation and microbially-related molecules establishment are available on this repository

(https://github.com/lfnothias/emp_metabolomic). Amplicon and shotgun metagenomic sequence data are submitted to the European Nucleotide Archive under Project: PRJEB42019 (https://www.ebi.ac.uk/ena/browser/view/PRJEB42019). Raw and demultiplexed amplicon and shotgun sequence data, the feature-table for full-length rRNA operon analysis, feature-tables for LC-MS/MS classical molecular networking and feature-based molecular networking, and the feature-table for GC-MS molecular networking data are available for download and analysis through Qiita at qiita.ucsd.edu (study: 13114).

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).

Earth Microbiome Project 500 (EMP500) Consortium*

Lars T. Angenant[1], Alison M. Berry[2], Leonora S. Bittleston[3], Dan Distel[4], Peter R. Girguis[5], Gary M. King[6], Aurora MacRae-Crerar[7], Gregory Mayer[8], Katherine D. McMahon[9], Jessica L. Metcalf[10], David Myrold[11], Brian Palenik[12], Forest Rohwer[13], Stuart A. Sandin[12], Steven K. Schmidt[14], Jennifer E. Smith[12], Karen Tait[15], Yael Tucker[16]

[1]University of Tübingen, Tübingen, Germany. [2]University of California, Davis, Davis, California, USA. [3]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Northeastern University, Boston, Massachusetts, USA. [5]Harvard University, Cambridge, Massachusetts, USA. [6]Louisiana State University, Baton Rouge, Louisiana, USA. [7]University of Pennsylvania, Philadelphia, Pennsylvania, USA. [8]Texas Technology University, Lubbock, Texas, USA. [9]University of Wisconsin, Madison, Wisconsin, USA. [10]Colorado State University, Fort Collins, Colorado, USA. [11]Oregon State University, Corvallis, Oregon, USA. [12]University of California, San Diego, La Jolla, California, USA. [13]San Diego State University, San Diego, California, USA. [14]University of Colorado, Boulder, Colorado, USA. [15]Plymouth Marine Laboratory, Plymouth, United Kingdom. [16]National Energy Technology Laboratory, USA.

*Listed members are EMP500 principal investigators who contributed samples but did not contribute directly to this paper.