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

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

[1]Department of Pediatrics, School of Medicine, University of California San Diego, La Jolla, California, USA.

[2]Collaborative Mass Spectrometry Innovation Center; University of California San Diego; La Jolla, CA 92093; USA

[3]Skaggs School of Pharmacy and Pharmaceutical Sciences; University of California San Diego; La Jolla CA 92093; USA

[4]Northern Gulf Institute, Mississippi State University, Mississippi State, Mississippi, USA

[5]Ocean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, Miami, Florida, USA

[6]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York, USA.

[7]Department of Bioengineering, University of California San Diego, La Jolla, California, USA.

[8]Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA

[9]Center for Microbiome Innovation, Jacobs School of Engineering, University of California San Diego, La Jolla, California, USA.

[10]Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA.

[11]School of Life Sciences, Arizona State University, Tempe, AZ 85281, USA

[12]Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University, Tempe, AZ 85281, USA

[13]Bioinformatics and Systems Biology Program, Jacobs School of Engineering, University of California San Diego, La Jolla, California, USA.

[14]Center for Computational Biology, Flatiron Institute, Simons Foundation

[15]Chair for Bioinformatics, Friedrich Schiller University, Jena, Germany

[16]Department of Computer Science, University of Antwerp, Antwerp, Belgium

[17]IBM Research, T.J. Watson Research Center, Yorktown Heights, NY, USA

[18]IBM Research Europe - Daresbury, UK

[19]IBM Research, Almaden Research Center, San Jose, CA, USA

[20]Department of Virus and Microbiological Special Diagnostics, Statens Serum Institute, Copenhagen, Denmark

[21]Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark

[22]Centre for Marine Science and Innovation, School of Biological, Earth and Environmental Science, The University of New South Wales, Sydney, 2052, Australia

[23]Interfaculty Institute of Microbiology and Infection Medicine, University of Tübingen, Tübingen, Baden-Württemberg 72076, Germany

[24]Department of Microbiology and Molecular Genetics, Michigan State University East Lansing MI USA

[25]Department of Computer Science and Engineering, Jacobs School of Engineering, University of California San Diego, La Jolla, California, USA.

[#]Co-first author

Abstract: 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]. In their analysis of the Genomes from Earth's Microbiomes (GEM) catalog[3], Nayfach and co-authors observed that, whereas genes coding for certain classes of secondary metabolites are limited or enriched in certain microbial taxa, "*specific chemistry is not limited or amplified by the environment, and that most classes of secondary metabolites can be found nearly anywhere*". 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, in contrast to Nayfach and co-authors, 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 biosynthetic gene clusters (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. Nayfach and co-authors uncovered 104,211 putative BGC regions from 52,515 microbial MAGs. Surprisingly, their analysis showed 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. The authors acknowledged that most of their annotated BGCs had incomplete sequences, potentially impacting annotation and quantification, but that this was consistent with previous studies. 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. While 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 SIRIUS/CANOPUS[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 (Extended Data Fig. 1, Table S1). With this dataset, we show that although the presence/absence (i.e., occurrence) 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 at 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 confirmed 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) (Figs. 2c, Tables S2, S3). We also identified specific metabolites that could classify specific environments with 68.9% accuracy in machine-learning analysis (Fig. 3a, Extended Data Fig. 3, 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 (Extended Data Table 1). Using machine-learning, we determined that specific microbial taxa and their functions could classify environments with 72.08% and 68.19% accuracy, respectively (Extended Data Figs. 2 and 3). 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,c). 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 (Extended Data Table 1). 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.

Figures.

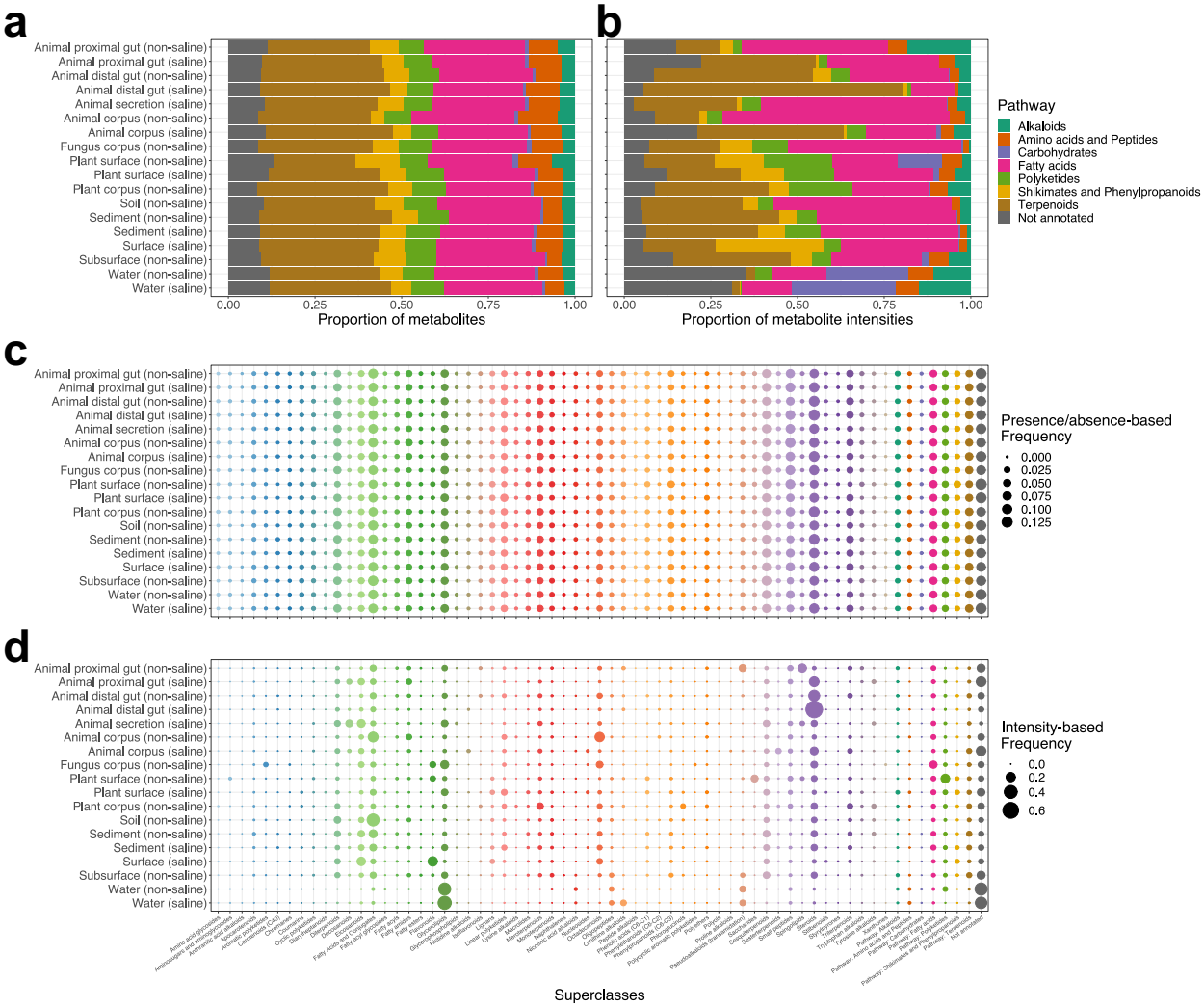


Fig. 1 | Distribution of microbially-related secondary metabolites chemical pathways among environments described using the Earth Microbiome Project Ontology (EMPO version 2, level 4). Both chemical pathway and chemical superclass annotations are shown based on presence/absence (**a, c**) and relative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

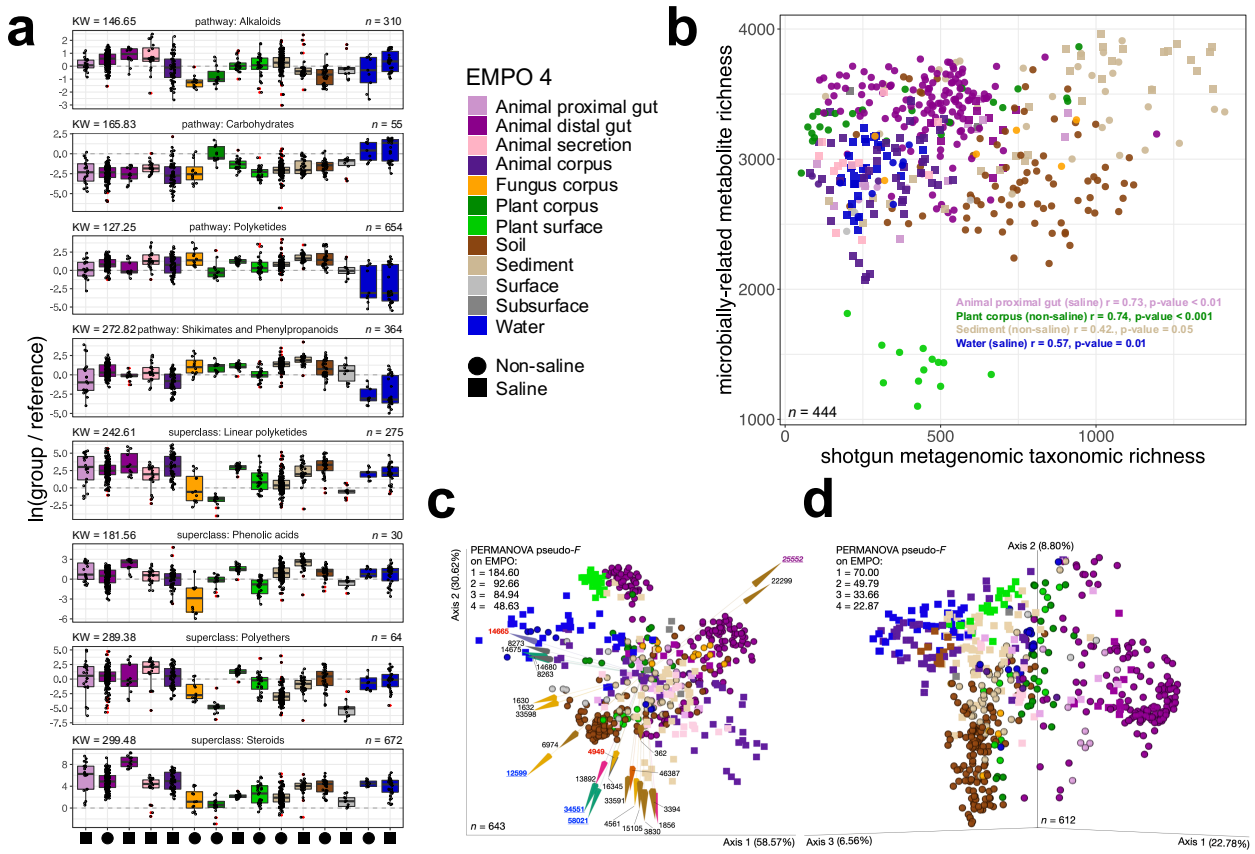


Fig. 2 | Structural-level associations between microbially-related secondary metabolites and specific environments described using the Earth Microbiome Project Ontology (EMPO version 2, level 4). **a,** Differential abundance of molecular features across environments, highlighting four example pathways 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).

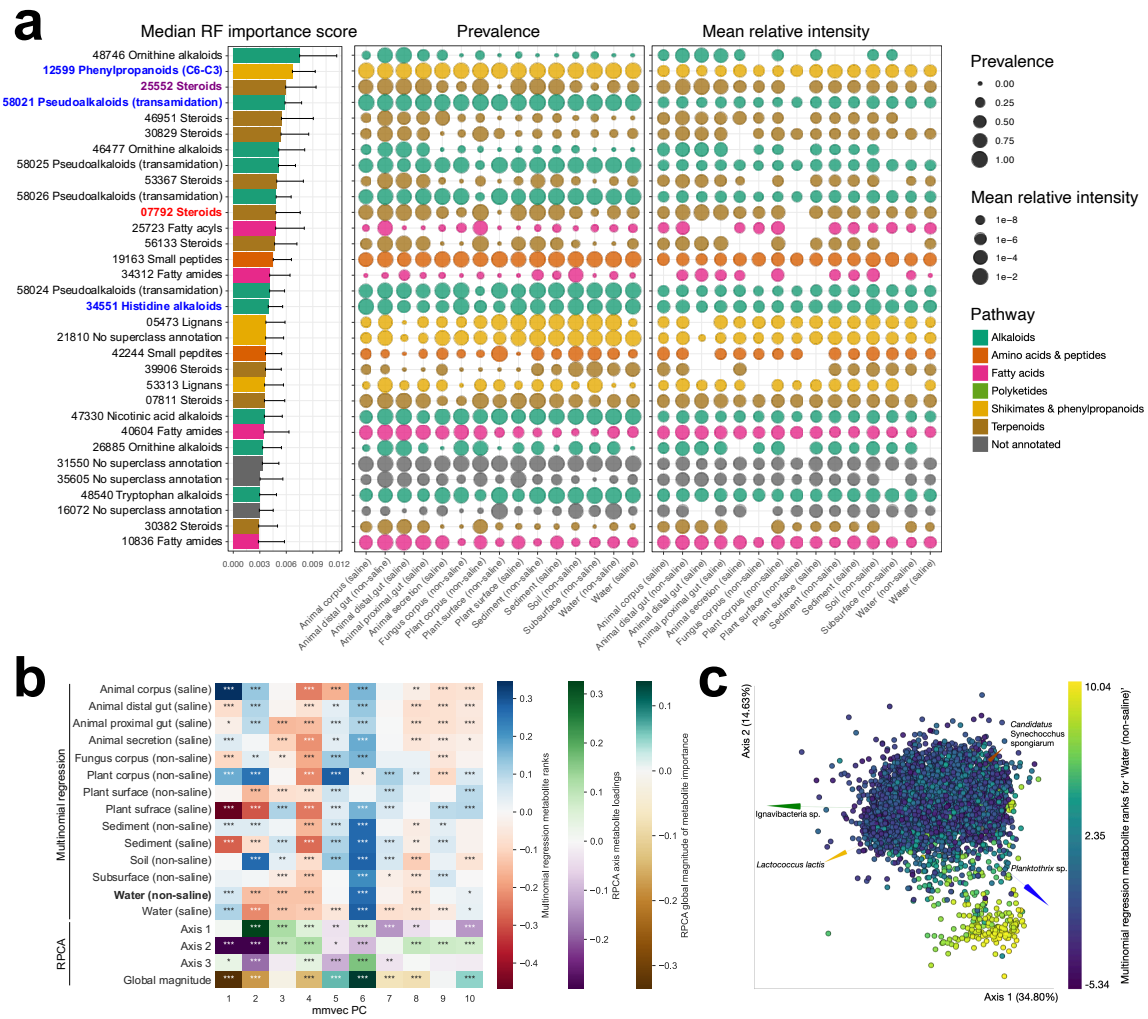


Fig. 3 | Microbially-related metabolites classify environments with a high accuracy and co-occur with specific microbial taxa. a, The random-forest importance score, environment-wide prevalence, and mean relative intensity for the top 32 most important microbially-related metabolites contributing to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

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 an early version of 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), 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. 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 U.S. Office of Biological and Environmental Research and located at PNNL. PNNL is a multiprogram national laboratory operated by Battelle for the Department of Energy (DOE) under contract DE-AC05-76RLO 1830. 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). J.S. was supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Environmental System Science (ESS) Program, and his contribution originates from the River Corridor Scientific Focus Area (SFA) project at PNNL. P.C.D. was supported by the Gordon and Betty Moore Foundation (award GBMF7622) and the NIH (award R01-GM107550). 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.

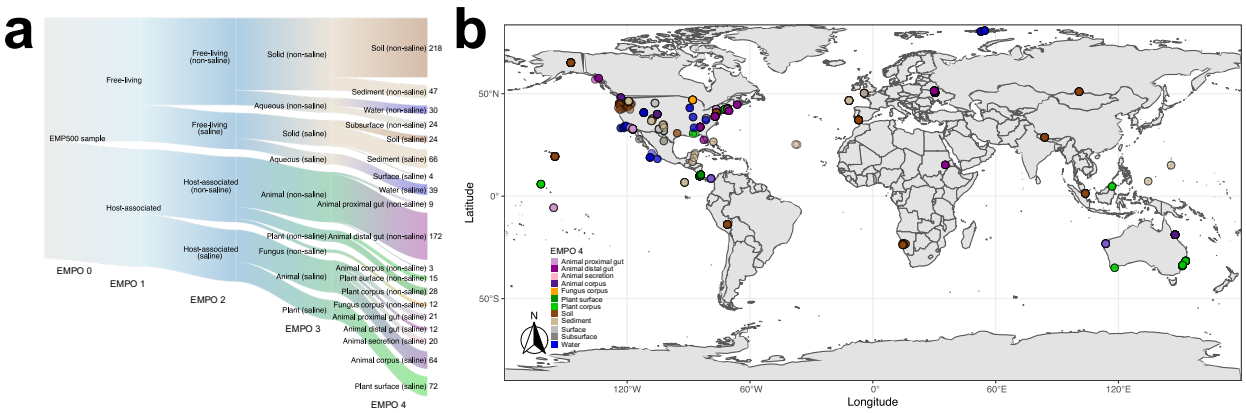
Jennifer L. Bowen[1], Max Chavarría-Vargas[2], Don A. Cowan[3], Jaime Huerta-Cepas[4], Paul Jensen[5], Lingjing Jiang[6], Anton Lavrinienko[7], Thulani P. Makhalanyane[3], Tapio Mappes[7], Ezequiel M. Marzinelli[8], Sou Miyake[9], Timothy A. Mousseau[7], Catalina Murillo-Cruz[2], Adrián A. Pinto-Tomás[2], Dorota L. Porazinska[10], Jean-Baptiste Ramond[3,11], Taniya RoyChowdhury[12,13], Henning Seedorf[9,14], J. Reuben Shipway[15,16], Frank J. Stewart[17], Jana M. U'Ren[18], Phillip C. Watts[7], Nicole S. Webster[19,20], Jesse R. Zaneveld[21], Shan Zhang[22]

[1]Northeastern University, Boston, Massachusetts, USA. [2]University of Costa Rica, San José, Costa Rica. [3]University of Pretoria, Pretoria, South Africa. [4]Universidad Politécnica de Madrid, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Madrid, Spain. [5]University of California San Diego, La Jolla, California, USA. [6]Janssen Research & Development, San Diego, California, USA. [7]University of Jyväskylä, Jyväskylä, Finland. [8]The University of Sydney, Sydney, Australia. [9]Temasek Life Sciences Laboratory, Singapore, Singapore. [10]University of Florida, Gainesville, Florida, USA. [11]Pontificia Universidad Católica de Chile, Santiago, Chile. [12]Pacific Northwest National Laboratory, Richland, Washington, USA. [13]University of Maryland, College Park, Maryland, USA. [14]National University of Singapore, Singapore, Singapore. [15]University of Portsmouth, Portsmouth, United Kingdom. [16]University of Massachusetts Amherst, Amherst, Massachusetts, USA. [17]Montana State University, Bozeman, Montana, USA. [18]University of Arizona, Tucson, Arizona, USA. [19]Australian Institute of Marine Science, Townsville, Qld, Australia. [20]University of Queensland, Brisbane, Australia. [21]University of Washington Bothell, Bothell, Washington, USA. [22]University of New South Wales, Sydney, Australia.

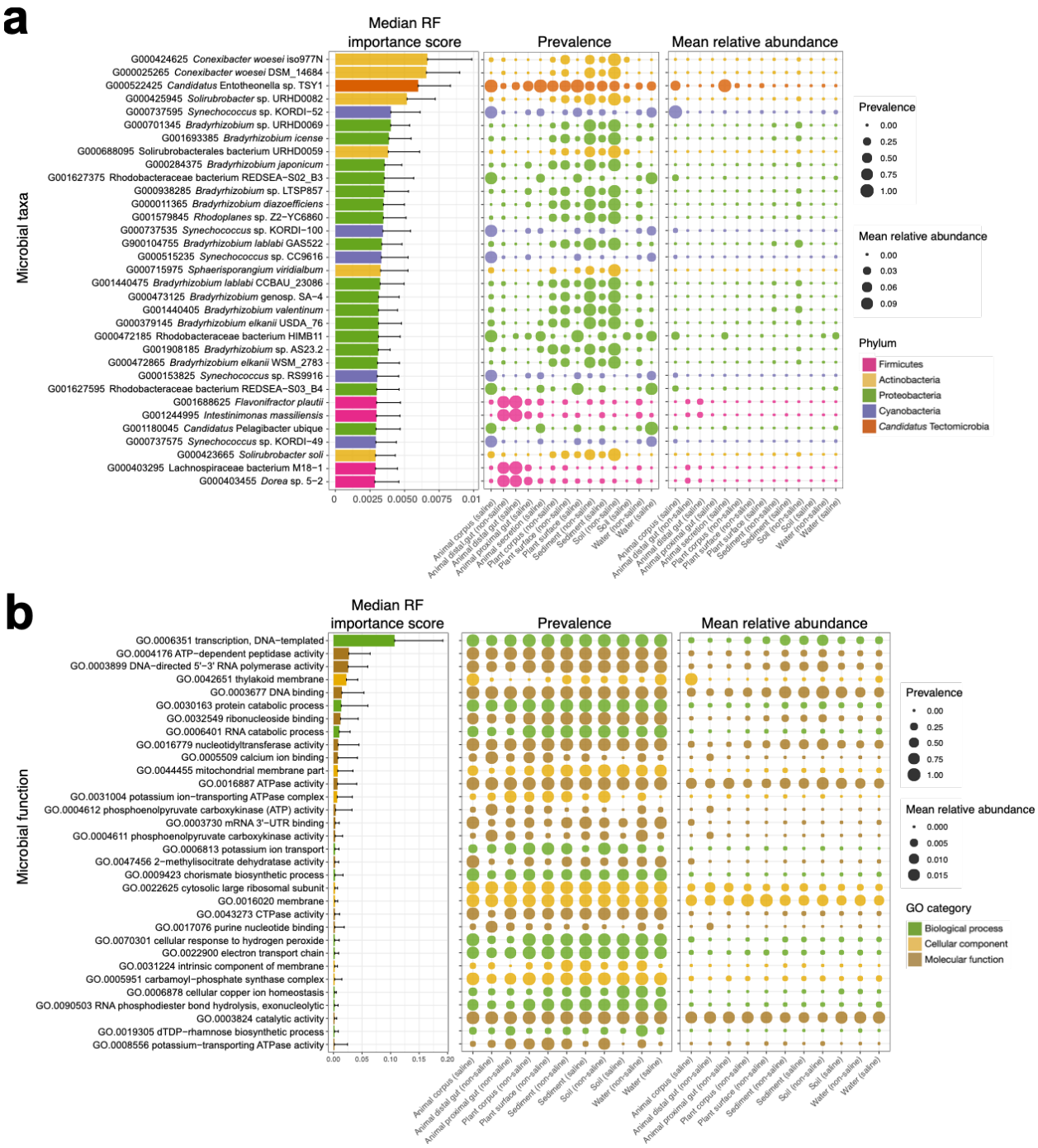
Extended Data.

Extended Data Table 1 | Mantel test results comparing data layers generated for the EMP500 samples. Note the strong relationships between the metabolomics data (i.e., LC-MS/MS and GC-MS) and the sequence data from Bacteria and Archaea (i.e., shotgun metagenomics, 16S, and full-length rRNA operon) as compared to between metabolomics data and sequence data from eukaryotes (i.e., 18S and ITS), 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

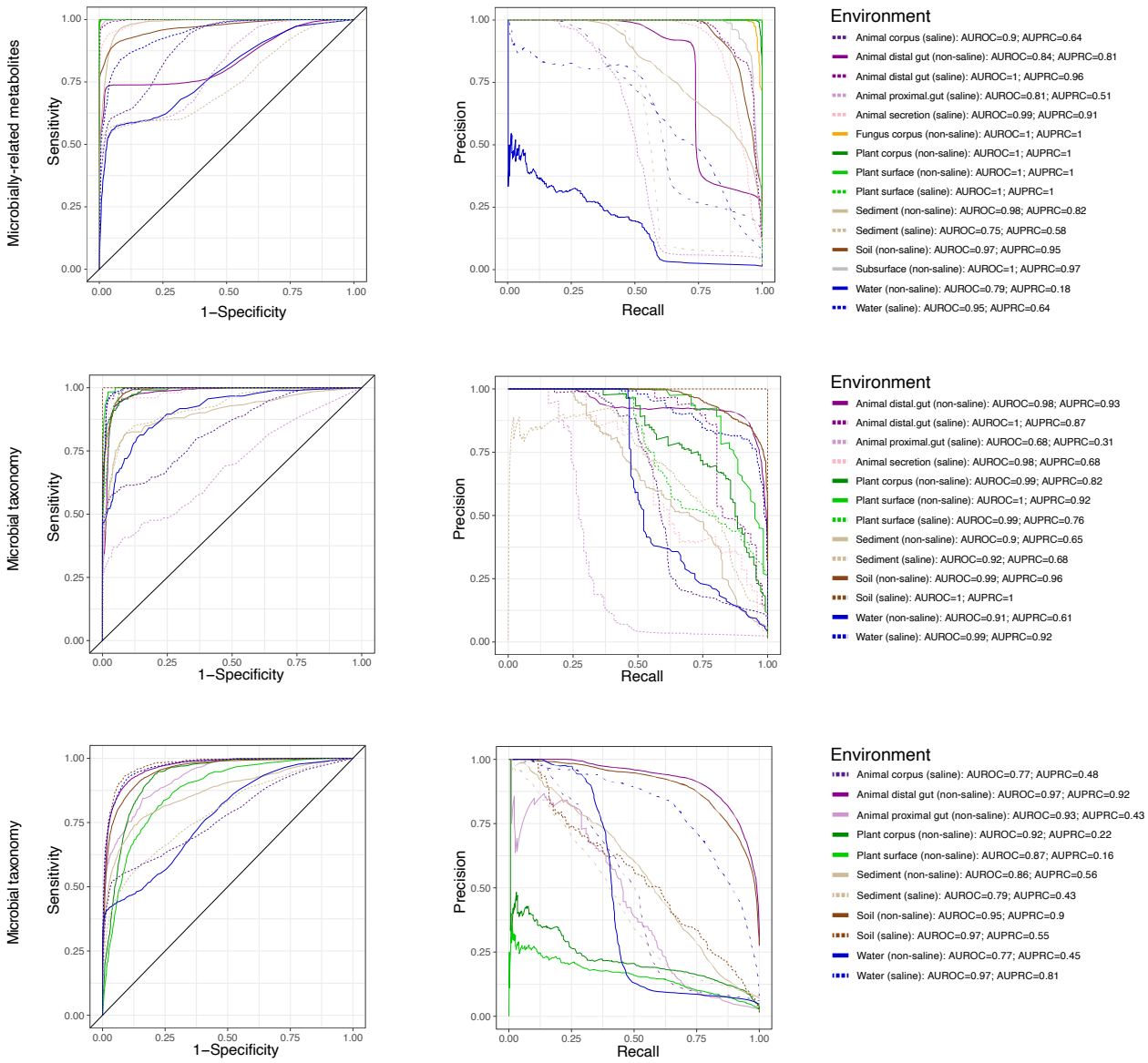


Extended Data Fig. 1 | a, Distribution of samples among the Earth Microbiome Project Ontology (EMPO version 2) categories. **b,** Geographic distribution of samples with points colored by EMPO (version 2, level 4). Extensive information about each sample set contributed is described in Table S1.



Extended Data Fig. 2 | 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.



Extended Data Fig. 3 | 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.