A communal catalogue reveals Earth's multiscale microbial diversity

Luke R. Thompson[1,8,9], Jon G. Sanders[1], Daniel McDonald[1], Amnon Amir[1], Joshua Ladau[10], Kenneth J. Locey[11], Robert J. Prill[12], Anupriya Tripathi[1,2,3], Sean M. Gibbons[13,14], Gail Ackermann[1], Jose A. Navas-Molina[1,4], Stefan Janssen[1], Evguenia Kopylova[1], Yoshiki Vázquez-Baeza[1,4], Antonio González[1], James T. Morton[1,4], Siavash Mirarab[6], Zhenjiang Zech Xu[1], Lingjing Jiang[1,5], Mohamed F. Haroon[15], Jad Kanbar[1], Qiyun Zhu[1], Se Jin Song[1], Tomasz Kosciolek[1], Nicholas A. Bokulich[16], Joshua Lefler[1], Colin J. Brislawn[17], Gregory Humphrey[1], Sarah M. Owens[18], Jarrad Hampton-Marcell[18,20], Donna Berg-Lyons[21], Valerie McKenzie[22], Noah Fierer[23], Jed A. Fuhrman[25], Aaron Clauset[21,24], Rick L. Stevens[19,26], Ashley Shade[28], Katherine S. Pollard[10], Kelly D. Goodwin[9], Janet K. Jansson[17], Jack A. Gilbert[18,27], Rob Knight[1,4,7] & The Earth Microbiome Project Consortium

[1]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA. [2]Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA. [3]Skaggs School of Pharmacy, University of California San Diego, La Jolla, CA, USA. [4]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA. [5]Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA, USA. [6]Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA, USA. [7]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA. [8]Department of Biological Sciences and Northern Gulf Institute, University of Southern Mississippi, Hattiesburg, MS, USA. [9]Ocean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, stationed at Southwest Fisheries Science Center, La Jolla, CA, USA. [10]The Gladstone Institutes and University of California, San Francisco, CA, USA. [11]Department of Biology, Indiana University, Bloomington, IN, USA. [12]Industrial and Applied Genomics, IBM Almaden Research Center, San Jose, CA, USA. [13]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [14]The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [15]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. [16]Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA. [17]Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA. [18]Biosciences Division, Argonne National Laboratory, Argonne, IL, USA. [19]Computing, Environment and Life Sciences, Argonne National Laboratory, Argonne, IL, USA. [20]Department of Biological Sciences, University of Illinois at Chicago, Chicago, IL, USA. [21]BioFrontiers Institute, University of Colorado, Boulder, CO, USA. [22]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA. [23]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA. [24]Department of Computer Science, University of Colorado, Boulder, CO, USA. [25]Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA. [26]Department of Computer Science, University of Chicago, Chicago, IL, USA. [27]Department of Surgery, University of Chicago, Chicago, IL, USA. [28]Department of Microbiology and Molecular Genetics and Program in Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, MI, USA.

Running title: The Earth Microbiome Project (EMP 16S Release 1)

Our growing awareness of the microbial world's importance and diversity contrasts starkly with our limited understanding of its fundamental structure. Despite recent advances in DNA sequencing, a lack of standardized protocols and common analytical frameworks impedes comparisons among studies, hindering the development of global inferences about microbial life on Earth. Here we present a meta-analysis of microbial community samples collected by hundreds of researchers for the Earth Microbiome Project. Coordinated protocols and new analytical methods, particularly the use of exact sequences instead of clustered operational taxonomic units, enable bacterial and archaeal ribosomal RNA gene sequences to be followed across multiple studies and allow us to explore patterns of diversity at unprecedented scale. The result is both a reference database giving global context to DNA sequence data and a framework for incorporating data from future studies, fostering increasingly more complete characterization of Earth's microbial diversity.

A primary aim of microbial ecology is to determine patterns and drivers of community distribution, interaction, and assembly amidst complexity and uncertainty. Microbial community composition has been shown to change across gradients of environment, geographic distance, salinity, temperature, oxygen, nutrients, pH, day length, and biotic factors [1,2,3,4,5,6]. These patterns have been identified mostly by focusing on one sample type and region at a time, with insights extrapolated across environments and geography to produce generalized principles. To assess how microbes are distributed across environments globally—or whether microbial community dynamics follow fundamental ecological 'laws' at a planetary scale—requires either a massive monolithic cross-environment survey or a practical methodology for coordinating many independent surveys. New studies interrogating microbial environments are rapidly accumulating; however, our ability to extract meaningful information from across datasets is outstripped by the rate of data generation. Previous meta-analyses suggested robust general trends in community composition, including the importance of salinity [1] and animal association [2]. These findings, although derived from relatively small and uncontrolled sample sets, support the utility of meta-analysis to reveal basic patterns of microbial diversity and suggest the need for a scalable and accessible analytical framework.

The Earth Microbiome Project (EMP, www.earthmicrobiome.org) was founded in 2010 to sample Earth's microbial communities at unprecedented scale in service of advancing our understanding of the organizing biogeographic principles governing microbial community structure [7,8]. We recognized that open and collaborative science, including scientific crowdsourcing and standardized methods [8], would help reduce technical variation among individual studies, which can overwhelm biological variation and make general trends difficult to detect [9]. Comprising ~100 studies, over two-thirds of which have yielded peer-reviewed publications (Supplementary Table 1), the EMP has now dwarfed by a hundred-fold the sampling and sequencing depth of prior meta-analysis efforts [1,2]; concurrently, powerful analysis tools have been developed, opening a new and larger window into the distribution of microbial diversity on Earth. In establishing a scalable framework to catalogue microbiota globally, we provide both a resource to explore myriad questions and a starting point for the guided acquisition of new data to test them. As an example of using this tool, we present a first meta-analysis of the EMP archive, tracking individual sequences across diverse samples and studies with standardized environmental descriptors, investigating large-scale ecological patterns, and exploring key hypotheses in ecological theory to serve as seeds for future research.

A standardized and scalable approach

The EMP solicited the global scientific community for environmental samples and associated metadata spanning diverse environments and capturing spatial, temporal, and/or physicochemical covariation. The first 27,751 samples from 97 independent studies (Supplementary Table 1) represent diverse environment types (Fig. 1a), geographies (Fig. 1b), and chemistries (Extended Data Fig. 1). The EMP encompasses studies of bacterial, archaeal, and eukaryotic microbial diversity. The analysis here focuses exclusively on the bacterial and archaeal components of the overall database (for concision, use of 'microbial' will hereafter refer to bacteria and archaea only). Associated metadata included environment type, location information, host taxonomy (if relevant), and physicochemical measurements (Supplementary Table 2). Physicochemical measurements were made *in situ* at the time of sampling. Investigators were encouraged to measure temperature and pH at minimum. Salinity, oxygen, and inorganic nutrients were measured when possible, and investigators collected additional metadata pertinent to their particular investigations.

Metadata were required to conform to the Genomic Standards Consortium's MIxS and Environment Ontology (ENVO) standards [10,11]. We additionally used a light-weight application ontology built on top of ENVO: the EMP Ontology (EMPO) of microbial environments. EMPO was tailored to capture two major environmental axes along which microbial beta-diversity has been shown to orient: host association and salinity [1,2]. We applied the classes in this application ontology (Fig. 1a) as levels of a structured categorical variable to classify EMP samples as host-associated or free-living (level 1). Samples were respectively categorized within those classes as animal- vs. plant-associated or saline vs. non-saline (level 2). A finer level (level 3) was

then assigned that satisfied the degree of environment granularity sought for this meta-analysis (e.g., sediment (saline), plant rhizosphere, animal distal gut). We expect EMPO to evolve as new studies and sample types are added to the EMP and as additional patterns of beta-diversity are revealed.

We surveyed bacterial and archaeal diversity using amplicon sequencing of the 16S rRNA gene, a common taxonomic marker for bacteria and archaea[12] that remains a valuable tool for microbial ecology despite the introduction of whole-genome methods (e.g., metagenomics) that capture gene-level functional diversity[13]. DNA was extracted from samples using the MO BIO PowerSoil DNA extraction kit, PCR-amplified, and sequenced on the Illumina platform. Standardized DNA extraction was chosen to minimize the potential bias introduced by different extraction methodologies; however, extraction efficiency may also be subject to interactions between sample type and cell type, and thus extraction effects should be considered as a possible confounding factor in interpreting results. We amplified the 16S rRNA gene (V4 region) using primers[14] shown to recover sequences from most bacterial taxa and many archaea[15]. We note that these primers may miss newly discovered phyla with alternative ribosomal gene structures[16], and subsequent modifications not used here have shown improved efficiency with certain clades of Alphaproteobacteria and Archaea[17,18,19]. We generated sequence reads of 90–151 bp (Extended Data Fig. 2a, Supplementary Table 1), totaling 2.2 billion sequences, an average of 80,000 sequences per sample.

Sequence analysis and taxonomic profiling were done initially using the common approach of assigning sequences to operational taxonomic units (OTUs) clustered by sequence similarity to existing rRNA databases[20,14]. While this approach was useful for certain analyses, for many sample types, especially plant-associated and free-living communities, a third of reads or more could not be mapped to existing rRNA databases (Extended Data Fig. 2b). We therefore employed a reference-free method, Deblur[21], to remove suspected error sequences and provide single-nucleotide resolution 'sub-OTUs', also known as 'amplicon sequence variants'[22], here called 'tag sequences' or simply 'sequences'. Because Deblur tag sequences for a given meta-analysis must be the same length in each sample, and some of the EMP studies have read lengths of 90 bp, we trimmed all sequences to 90 bp for this meta-analysis. We verified that the patterns presented here were not adversely affected by trimming the sequences (Extended Data Fig. 3). As we show, 90-bp sequences were sufficiently long to reveal detailed patterns of community structure. And because exact sequences are stable identifiers, unlike OTUs, they can be compared to any 16S rRNA or genomic database now and going forward, promoting reusability[22].

Microbial ecology without OTU clustering

While earlier large-scale 16S rRNA amplicon studies adopted OTU clustering approaches in part out of concern that erroneous reads would dominate diversity assessments[23], patterns of prevalence (presence–absence) in our results suggest that Deblur error removal produced ecologically informative sequences without clustering. After rarefying to 5,000 sequences per sample, a total of 307,572 unique sequences were contained in the 96 studies and 23,828 samples of the 'QC-filtered' Deblur 90-bp observation table. Among studies, over half (57%) of all obtained sequences were observed in two or more studies, but only 5% were observed in more than ten studies; the most prevalent sequence was found in 88 of 96 studies (Extended Data Fig. 4a). Among samples, while most sequences (86%) were observed in two or more samples, only 7% were observed in more than 100 samples (Extended Data Fig. 4b). As expected, the most prevalent sequences were also the most abundant (Extended Data Fig. 4c).

Our analyses were carried out using a modest sequencing depth of 5,000 observations per sample after Deblur and rarefaction. To examine how prevalence estimates were affected by sequencing depth, we focused on four major environment types for which we had the greatest number of samples with $>$50,000 observations (soil, saltwater, freshwater, and animal distal gut). The relationship between average tag sequence prevalence and sequencing depth differed among these environments (Extended Data Fig. 4d) but was generally increasing, suggesting that our global analysis underestimated true prevalence. Animal-associated microbiomes were a notable exception, with an upper bound on prevalence apparently imposed by host-specificity when all host species were considered (Extended Data Fig. 4e); this bound disappeared when considering only human-derived samples (Extended Data Fig. 4f). While contamination remains an issue in microbiome studies[24], most of the very highly abundant and prevalent sequences here had higher mean relative abundances among samples than among no-template controls (Supplementary Table 3), suggesting that they did not originate from reagents.

Matches between our sequences and existing 16S rRNA gene reference databases highlight the novelty captured by the EMP. Exact matches to 46% of Greengenes[25] and 45% of SILVA[26] rRNA gene databases were found in our dataset, indicating that we 'recaptured' nearly half of the reference sequence diversity with just under 100 environmental surveys. These matches accounted for 10% and 13%, respectively, of the tag sequences in our dataset, indicating that large swaths of microbial community diversity are not yet captured in full-length sequence databases. The failure of many sequences to be mapped in reference-based alignments to Greengenes and SILVA 97% OTUs (Extended Data Fig. 2b) supports this observation.

Patterns of diversity reflect environment

We used a structured categorical variable of microbial environments, EMPO, to analyze diversity in the EMP catalogue in the context of lessons from previous investigations[1,2]. We observed environment-dependent patterns in the number of observed tag sequences (alpha-diversity), turnover and nestedness of taxa (beta-diversity), and predicted genome properties (ecological strategy). Derived from a more standardized methodology, our dataset confirms the prior finding[2] that host association is a fundamental environmental factor differentiating microbial communities (Fig. 2c, Extended Data Fig. 2d). We build on this pattern by showing less richness in host-associated communities (Fig. 2a), with the noted exception of plant rhizosphere samples, which resemble the higher richness (Fig. 2a) and composition (Fig. 2c) of free-living soil communities. Our findings also confirm the major compositional distinction between saline and non-saline communities[1] (Fig. 2c). Effect sizes of environmental factors on alpha- and beta-diversity generally showed large contributions of environment type and (for host-associated samples) host species to both types of diversity (Extended Data Fig. 5a, b).

The ability to identify sample provenance using only a microbial community profile has applications ranging from criminal forensics to mistaken sample identification; this will require large curated datasets like the EMP. Supervised machine learning demonstrated that samples could be distinguished among animal-associated, plant-associated, saline free-living, or non-saline free-living with 91% accuracy based solely on community composition, and to fine-scale environment with 84% accuracy (Extended Data Fig. 5c, d, e). The most commonly misclassified samples were soil, non-saline surface and aerosol, and animal secretion. In many of these cases, misclassification can be attributed to limitations of EMPO. As additional samples are classified, classification can be improved by iteratively and empirically redefining categories using machine learning. Conversely, with continuous factors like salinity, categorical definitions cannot perfectly capture intermediate values. High classification success to environment type was supported by source-tracking analyses (Extended Data Fig. 5f, g), with the exception of plant rhizosphere samples, owing to their similarity to soil samples.

Predicted average community copy number (ACN) of the 16S rRNA gene was another metric found to differentiate microbial communities in both host-associated and free-living communities (Fig. 2d). ACN can be predicted from 16S rRNA amplicon data[27]; this method has been used, for example, to link the taxonomic groups associated with copiotrophic and oligotrophic behaviors in soils to high and low rRNA gene copy numbers, respectively[28]. Approximately half the dataset centered on an ACN of 2.2 (free-living and plant-associated samples) and the other half on 3.4 (animal-associated samples) (Fig. 2d). Greater per-genome rRNA operon copy number has been found to correlate with rapid maximum growth rates[29], which may provide a selective advantage when resources are abundant, such as in animal hosts. While ACN is an estimate rather than a measurement of average rRNA copy number and is subject to potential biases in the underlying reference database, the distributions we observed are consistent with 16S rRNA copy number reflecting differences in ecological strategies among environments.

A resource for theoretical ecology

The coordinated accumulation of data across studies allows investigations of patterns within (alpha-diversity) and among (beta-diversity) microbial communities at scales that vastly exceed what could be measured in any individual study. Patterns of alpha-diversity in meta-analyses have revealed global trends key to the development of major ideas in macroecological theory, but fundamental patterns have been more difficult to discern in microbial ecology. For example, a nearly ubiquitous tendency towards greater diversity in the tropics is evident in macroecology[30], but there is substantial variation among studies examining latitudinal trends of microbial diversity[31,32,33]. The large EMP dataset analyzed here reveals a weak but significant trend towards decreasing diversity at higher latitudes in non-host-associated environments (Extended Data Fig. 5h). An effect of latitude was apparent both within and across studies, consistent with global trends in latitudinal microbial diversity being an emergent function of locally selective environmental heterogeneity[34]. However, substantial study-to-study variation in richness highlights the caveats inherent in meta-analysis; more coordination of sample collections from similar environments across larger gradients is necessary to better address this question.

The EMP has potential to link global patterns of microbial diversity with physicochemical parameters—if appropriate metadata are provided by researchers. Microbial community richness has been found to correlate with environmental factors, including pH and temperature[3,33,35,36]. For example, richness has been shown to increase up to neutral pH[36] and often decrease above neutral pH[3,35] in soil communities. Richness has been shown to increase with temperature up to a limit and then decrease beyond that limit in seawater (maximum at ~19 °C)[33] and to increase with temperature in soil (up to at least ~26 °C)[36]. However, general relationships of richness to temperature and pH remain unresolved[37]. Here, across samples from non-host-associated environments where pH or temperature were measured (mostly freshwater and soil environments), richness was greatest near neutral pH (~7) and relatively cool temperatures (~10 °C) (Fig. 2b). We observed apparent upper bounds on richness with both

temperature and pH that were best fit by two-sided exponential (Laplace) curves. Thus, the present dataset suggests a relatively narrow range of intermediate pH and temperature values at which maximum microbial richness occurs. These patterns, while robust in the context the EMP dataset, necessarily reflect only the subset of sample types for which variables were measured (Supplementary Table 2); they thus should be interpreted cautiously. Understanding universal relationships between richness and environmental factors will require information from more studies with detailed and carefully collected physicochemical metadata.

Beyond measured physical covariates, the breadth of environments in the EMP catalogue allows for detailed exploration of how microbial diversity is distributed across environments. Diversity among communities (beta-diversity) is driven by turnover (replacement of taxa) and nestedness (gain or loss of taxa resulting in differences in richness)[38]. If turnover dominates, then disparate communities will harbour unique taxa. If nestedness dominates, then communities with fewer taxa will be subsets of communities with more taxa. We tested for nestedness using a 2,000-sample subset with even representation across environments and studies. Given the contrasting environments and geographic separation among the many studies in the EMP, we expected different environments to contain unique sets of taxa and to show little nestedness. Surprisingly, we found communities across environments to be significantly nested (Fig. 3a, b; $p < 0.05$) in comparison to null models (Fig. 3c), accounting for the observed patterns of richness. At coarse taxonomic levels, an average of 84% of phyla, 73% of classes, and 58% of orders that occurred in less diverse samples also occurred in more diverse samples. Nestedness was observed even when the most prevalent taxa were removed and was robust across randomly chosen subsets of samples (Extended Data Fig. 6). These patterns could have resulted from several mechanisms, including ordered extinctions[39] and the filtering of complex communities over time[40], differential dispersal abilities[41] and cascading source–sink colonization processes that assemble nested subsets from more complex communities, or by the tendency of larger habitat patches to support more rare species with lower prevalence[42]. Notably, finer taxonomic groupings showed less nestedness (Fig. 3c), indicating that the processes underlying nested patterns of turnover are likely reflective of conserved aspects of microbial biology, and not due to the interplay of diversification and dispersal on short timescales.

These global ecological patterns offer a glimpse of what is possible with coordinated and cumulative sampling—in addition to the specific questions addressed by individual studies, context is built and easily queried across studies. They also necessarily highlight the inherent limitations to decentralized studies, especially regarding the collection of comparable environmental data. Future studies will be able to use the current EMP data as a starting point for more explicit tests of broad ecological principles, both to identify gaps in current knowledge and to more confidently plan large directed studies with sufficient power to fill them.

A more precise and scalable catalogue

An advantage of using exact sequences is that they enable observation and analysis of microbial distribution patterns at finer resolution than possible with traditional OTUs. As an example, we applied a Shannon entropy analysis to tag sequences and higher taxonomic groups to measure biases in the distribution of taxa. Taxa equally likely to be found in any environment will have high entropy and low specificity, whereas taxa found only in a single environment will have low entropy and high specificity (note that we use 'specificity' solely to denote distributional patterns, not to imply adaptation or causality). Tag sequences exhibited high specificity for environment, with distributions skewed toward one or a few environments (low Shannon entropy); in contrast, higher taxonomic levels tended to be more evenly distributed across environments (high Shannon entropy, low specificity) (Fig. 4a). Entropy distributions across all tag sequences at each taxonomic level show that this pattern is general (Fig. 4b). Seeking a more precise measure of the divergence at which a taxon is specific for environments, we next asked how entropy changes as a function of phylogenetic distance. We calculated entropy for each node of the phylogeny and visualized as a function of maximum tip-to-tip branch length (Fig. 4c). While entropy gradually decreased at finer phylogenetic resolution, it dropped sharply at the tips of the tree. We conclude that environment specificity is best captured by individual 16S rRNA sequences, below the typical threshold defining microbial species (97% identity of the 16S rRNA gene).

The EMP dataset provides the ability to track individual sequences across Earth's microbial communities. Using a representative subset of the EMP (Extended Data Fig. 7a), we produced a table of sequence counts and distributions, including among environments (EMPO) and along environmental gradients (pH, temperature, salinity, and oxygen). From this we generated 'EMP Trading Cards', which promote exploration of the dataset and here highlight the distribution patterns of three prevalent or environment-correlated tag sequences (Extended Data Fig. 7b, Supplementary Table 3). The entire EMP catalogue is queryable using the Redbiom software, with command-line (github.com/biocore/redbiom) and web-based (qiita.microbio.me) interfaces to find samples based on sequences, taxa, or sample metadata, and to export selected sample data and metadata (instructions at github.com/biocore/emp). User data generated from the EMP protocols can be readily incorporated into this framework: because

Deblur operates independently on each sample[21], additional tag sequences can be added to this dataset from new studies without reprocessing existing samples. Future combination of datasets targeting the same genomic region but sequenced using different methods may be admissible but would require considerations to account for methodological biases.

The growing EMP catalogue is expected to have applications for research and industry, with tag sequences employed as environmental indicators and representing targets for cultivation, genome sequencing, and laboratory study. Additionally, these tools and approaches, while developed for bacteria and archaea, could be expanded to all domains of life[43]. To achieve greater utility for the EMP and similar projects, we must continually improve metadata collection and curation, ontologies, support for multi-omics data, and access to computational resources.

Conclusions and future directions

Here we have employed crowdsourced sample collection and standardized microbiome sequencing and metadata curation to perform a global meta-analysis of bacterial and archaeal communities. Using exact sequences in place of OTUs and a learned structure of microbial environments, we have shown that agglomerative sampling can work to reveal basic biogeographic patterns of microbial ecology, with resolution and scope rivaling data compilations currently available for 'macrobial' ecology[44,45]. Our results point to key organizing principles of microbial communities, with less rich communities nested within richer communities at higher taxonomic levels, and environment specificity becoming much more evident at the level of individual 16S rRNA sequences.

The EMP framework and global synthesis presented here represent value added to the scientific community beyond the substantial contributions of the constituent studies (Supplementary Table 1). However, as with any meta-analysis in which data are gathered primarily in service of separate questions rather than a single theme[46], conclusions should be viewed with caution and form starting points for future hypothesis-directed investigations. There is need to more evenly span gradients of geography (e.g., latitude and elevation) and chemistry (e.g., temperature, pH, and salinity)—assisted by tools for more comprehensive collection and curation of metadata—and to track environments over time. Additionally, biotic factors (e.g., animals, fungi, plants, viruses, and eukaryotic microbes) not measured in this study play important roles in determining community structure[4,5,6]. The scalable framework introduced here allows expansion to address these needs: new studies to fill gaps in physicochemical space, amplicon data for microbial eukaryotes and viruses, and whole-genome and whole-metabolome profiling. At a time when both academic and governmental agencies increasingly recognize the value of communal biodiversity monitoring efforts[47,48], the EMP provides one example of a logistically feasible standardization framework to maximize interoperability across many diverse and independent studies, in particular using stable identifiers (exact sequences) to enable enduring utility of environmental biodiversity data. Given current global sequencing efforts, the use of coordinated protocols and submission to this and other public databases should allow rapid accumulation of new data, providing an ever more diverse reference catalogue of microbes and microbiomes on Earth.

END NOTES

Supplementary Information

Supplementary Information is available in the online version of the paper.

Acknowledgements

We thank Jeff DeReus for management of information systems; Jim Huntley and Kristen Jepsen for management of sequencing facilities; Brent Erickson for administrative assistance; Jay Lennon for discussions about macroecological theory; Shyamal Peddada for assistance with effect size calculations; Pier Luigi Buttigieg, Chris Mungall, and Debby Siegele for assistance with ontologies; Alexandra Rose, Alexandra-Sophie Roy, Angelita Bearquiver, Bob Cohen, Chia L. Tan, Christine Tischer, Claudia Feh, David Winkler, Edwin Jones, Esther Angert, Frederick Blackwolf, Gilles Martin, Harald Schunck, Kelly Hallinger, Lora R. McGuinness, Martin Mühling, Michael Lombardo, Robert Madsen, Saman Bowatte, Sarah Romac, Sylvia Garcia-Houchins, Vanessa Harriman, and Wes James for assistance with sample and/or metadata collection; and the following individuals for supporting the project's founding: Alex Scyzrba, Alice McHardy, Andreas Teske, Andreas Wilke, C. Titus Brown, Chris Brown, Daniel Huson, Dawn Field, Dirk Evers, Doug Wendel, Elizabeth Glass, Eugene Kolke, Fengzhu Sun, Frank Oliver Glöckner, George Kowalchuk, Hans-Peter Klenk, James Tiedje, Jeff Gordon, Jeroen Raes, Jim Knight, Joel Kostka, John Heidelberg,

Jonathan Eisen, K. Eric Wommack, Kathryn Docherty, Kevin Keegan, Konstantinos Konstantindis, Mark Bailey, Matthew Sullivan, Narayan Desai, Nikos Kyprides, Norman Pace, Pavan Balaji, Rachel Gallery, Rachel Mackelprang, Ronald O'Dor, Ruth Ley, Tim Vogel, Ting Chen, and Wu Feng. This work was supported by the John Templeton Foundation (Grant ID 44000, Convergent Evolution of the Vertebrate Microbiome), the W. M. Keck Foundation (DT061413), Argonne National Laboratory (U.S. Dept. of Energy Contract DE-AC02-06CH11357), the Australian Research Council, and the Extreme Science and Engineering Discovery Environment (XSEDE, project number BIO150043), which is supported by National Science Foundation grant number ACI-1053575. Funding for L.R.T. was provided in part by NOAA's Atlantic Oceanographic and Meteorological Laboratory (AOML) and the Mississippi State University/NOAA Northern Gulf Institute. We thank MO BIO Laboratories, Luca Technologies, Eppendorf, Boreal Genomics, Illumina, and Roche 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. L.R.T. coordinated the meta-analysis, performed analysis, and wrote the manuscript. D.M. developed tools, performed analysis, and wrote the manuscript. J.G.S., J.L., K.J.L., R.J.P., S.M.G., A.A., A.T., Z.Z.X., N.A.B., and A.S. performed analysis and wrote the manuscript. Y.V.-B., J.T.M., and S.M. developed tools and performed analysis. A.G. managed the project and performed analysis. J.A.N.-M., S.S., E.K., M.F.H., T.K., S.J., L.J., C.J.B., J.L., Q.Z., J.K., and K.S.P. performed analysis. G.C.H. and G.A. managed the project. S.M.O., J.H.-M., and D.B.-L. managed the project and coordinated DNA sequencing. K.D.G., R.L.S., A.C., J.A.F., and V.M. wrote the manuscript. N.F., J.K.J., J.A.G., and R.K. managed the project and wrote the manuscript.

Author Information

Reprints and permissions information are available at www.nature.com/reprints. The authors declare no competing interests. Correspondence and requests for materials should be addressed to R.K. (robknight@ucsd.edu), J.K.J. (janet.jansson@pnnl.gov), or J.A.G. (gilbertjack@gmail.com).

The Earth Microbiome Project Consortium

Aaron R. Jex[116,87], Alexandra H. Campbell[121], Alexandra M. Linz[139], Alison Berry[98], Allison E. Williams[107], Alyssa Cochran[20], Amy Apprill[152], Andaine Seguin-Orlando[104], Anders Karlsson[62], Andrew Whitehead[98], Andy Rees[73], Anna Forsman[23,101], Anni Moore[60], Anson V. Koehler[116], Antje Gittel[1,95], Antonio M. Martín-Platero[109], Asha Rani[112], Ashish Bhatnagar[53], Aurora MacRae-Crerar[125], Baddr Shakhsheer[102], Bazartseren Boldgiv[64], Beck Wehrle[99], Benjamin B. Crary[139], Benjamin D. Shogan[102], Benjamin L. Turner[80], Bharath Prithiviraj[17], Bonnie Laverock[128], Brenda Casper[125], Brent Stephens[41], Byron Crump[70], Caitlin Potter[8], Carol Robinson[106], Catherine M. Spirito[22], Catherine Pfister[102], Cesar Cardona[102], Chris Freeman[8], Christopher Quince[136], Colleen T. E. Kellogg[96], Congcong Shen[45], Craig Cary[135], D. Lee Taylor[120], Daniel A. Cristol[19], Daniel P. Smith[10], Daniel van der Lelie[37], Daniela Vargas-Robles[126], Danielle C. Claar[133], Danilo Ercolini[119], Dave Shutler[2], David A. Lipson[77], David A. Mills[98], David Armitage[97,123], David Garshelis[59], David Myrold[70], Diogo Jurelevicius[91], Dionysios A. Antonopoulos[5], Donal M. Boyer[151], Donald A. Walker[93], Donglai Gong[146], Douglas C. Woodhams[115], Duoying Cui[11], Elizabeth Pilon-Smits[20], Elliot S. Friedman[23,125], Embriette Hyde[100], Emily M. Landon[102], Eric A. Dubinsky[50,97], Eric Bottos[71], Eric R. Johnston[35], Eske Willerslev[104], Ezequiel M. Marzinelli[121], F. Joseph Pollock[72], Fabian Michelangeli[48], Folker Meyer[5,102], Forest Rohwer[77], Frédéric Delsuc[144], Francis Q. Brearley[54], Gabriela M. Sheets[30], Gary M. King[52], Gavin Collins[63], George W. Kling[117], Giancarlo Galindo[50], Glida Hidalgo[4], Graeme Nicol[143], Gregory D. Mayer[84], Gunnar Gerdts[3], Haiyan Chu[45], Hakdong Shin[79], Hans-Peter Grossart[51,74], Hebe M. Dionisi[15], Helen Findlay[73], Hongxia Zhao[154], Ina Timling[93], Iratxe Zarraonaindia[141], Iris Levin[103], Irma D. Flemming[145], Isaac Gifford[98], J. Gregory Caporaso[66], Jacob Parnell[68], James E. McDonald[8], Jamie M. McDevitt-Irwin[133], Jason Andras[61], Jeff Hooker[16], Jeffrey J. Werner[23,81], Jeffrey Siegel[129], Jenni Hultman[111], Jennifer Defazio[102], Jennifer F. Biddle[105], Jeremiah Minich[100], Jesse Zaneveld[137], Jessica L. Metcalf[20], Jiri Barta[127], John Alverdy[102], JoLynn Carroll[92], Jonathan B. Clayton[118,36], Jonathan E. Hickman[22], Jordan Kueneman[80,103], Jose Agosto Rivera[126], Jose C. Clemente[40], Joseph R. Mendelson III[35,155], Josephine Braun[78], Josh D. Neufeld[138], Jozef I. Nissimov[76], Juan Diego Ibáñez-Álamo[110], Juan J. Soler[27], Juan M. Peralta-Sánchez[109], Julia K. Baum[133], Julie D. Jastrow[5], Julie LaRoche[25], Karen Noyce[59], Karen Tait[73], Karl J. Rockne[112], Kate Ballantine[61], Katherine McMahon[139], Katherine R. Amato[67], Kefeng Niu[32,130], Kelly Lane-deGraaf[107], Kim M. Handley[94], Kim Miller[69], Kirsten S. Hofmockel[71], Krista McGuire[92,23], Kristin West[24], Kristina Guyton[102], L. Margarita Martínez[19], L. Scott Johnson[89], Largus T. Angenent[22,131], Lauren M. Seyler[58], Lee J. Kerkhof[76], Liliana Davalos[82],

Linda A. Whittingham[140], Lisa Al-Moosawi[73], Liza Garcia[126], Lucas Moitinho-Silva[121], Lucie Bittner[142], Lucy Seldin[91], Ludovic Orlando[104], Lukas van Zwieten[121], Luke K. Ursell[13], Mónica Contreras[48], Magda Magris[4], Manuel Lladser[103], Manuel Martín-Vivaldi[109], Manuel Martínez-Bueno[109], Maria Alexandra Garcia-Amado[48], Maria Gloria Dominguez-Bello[65], Mariana Lozada[15], Mark D. Schrenzel[39], Martin Sperling[34], Matthew J. Nolan[116], Matthew Schrenk[58], Maureen L. Coleman[102], Melita A. Stevens[56], Miguel Lentino[18], Miles Richardson[102], Molly K. Gibson[149], Monica Bhatnagar[53], Monica Medina[72], Monika Krezalek[102], Naseer Sangwan[102], Nathalie Fenner[8], Neslihan Tas[50], Nicole M. Scott[13], Nicole Webster[7], Noriko Okamoto[96], Nur A. Hasan[114], Olayinka Osuolale[28], Olivia U. Mason[33], Pamela Weisenhorn[102], Paola Piombino[119], Paola Vitaglione[119], Paul Munroe[121], Peter D. Steinberg[121], Peter Golyshin[8], Peter Larsen[5], Peter O. Dunn[140], Peter Petraitis[125], Pierre Liancourt[44], Qikun Zhang[154], Rachael M. Morgan-Kiss[57], Ravi Ranjan[112], Rebecca Safran[103], Rebecca Vega Thurber[70], Regina Lamendella[49], Rita L. Seger[113], Rita R. Colwell[114], Robert E. Espinoza[14], Robert G. Clark[31], Robin B. Gasser[116,38], Robin Dowell[103], Roger Karlsson[62], Ross Stephen Hall[116], Russell D. Dawson[122], Ryan McMinds[70], Ryan T. Gill[103], Safiyh Taghavi[37], Sara Sjöling[83], Sarah E. Daly[75], Sarah J. Haig[117], Sarah L. O'Brien[5], Selena Marie Rodriguez[126], Seth Kauppinen[97], Shane R. Haydon[56], Shaun Nielsen[121], Shi Wang[50], Simon Creer[8], Simon Lax[102], Sophie Weiss[103], Stefan Hulth[108], Stefano Mocali[23], Stephanie D. Jurburg[148], Stephen A. Wood[153], Stephen B. Pointing[6], Stephen Joseph[121], Steve Simmons[42], Steven J. Hallam[96], Subramanya Rao[6,86], Susan R. Whitehead[147], Suzanne J. Kennedy[12], Tao Wang[116], Tim Urich[134], Tom Weaver[103], Torsten Thomas[121], Trevor Charles[138], Tugrul Girai[126], Ulf Riebesell[34], Vanessa Ezenwa[107], Vanessa Hale[86], Vera Tai[150], Vincenzo Fogliano[119], Virginia Sanz[48], Walter P. MacCormack[90,47], Wayne Roundstone[26], Wenju Liang[43], William A. Walters[55], William Brazelton[132], William van Treuren[103], Wyatt Oswald[29], Yadira Ortiz Castellano[126], Yeqin Yang[88], Yingying Ni[45], Yongqin Liu[46], Yu Shi[45]

Institutions

1. Aarhus University, Aarhus, Denmark
2. Acadia University, Wolfville, NS, Canada
3. Alfred Wegener Institute, Bremerhaven, Germany
4. Amazonic Center for Research and Control of Tropical Diseases (CAICET), Puerto Ayacucho, Amazonas, Venezuela
5. Argonne National Laboratory, Argonne, IL, USA
6. Auckland University of Technology, Auckland, New Zealand
7. Australian Institute for Marine Science, Townsville, Queensland, Australia
8. Bangor University, Bangor, Gwynedd, Wales, UK
9. Barnard College, Columbia University, New York, NY, USA
10. Baylor College of Medicine, Houston, TX, USA
11. Beijing Zoo, Beijing, China
12. Bio-Path Holdings, Inc., Bellaire, TX, USA
13. Biota Technology Inc., San Diego, CA
14. California State University, Northridge, CA, USA
15. Centro para el Estudio de Sistemas Marinos (CESIMAR-CONICET), CCT CENPAT, Puerto Madryn, Chubut, Argentina
16. Chief Dull Knife College, Lame Deer, MT, USA
17. City University of New York, New York, NY, USA
18. Colección Ornitologica W. H. Phelps, Caracas, Venezuela
19. College of William and Mary, Williamsburg, VA, USA
20. Colorado State University, Fort Collins, CO, USA
21. Columbia University, New York, NY, USA
22. Cornell University, Ithaca, NY, USA
23. CREA-AA, Florence, Italy
24. DOCS Global, Research Triangle Park, NC, USA
25. Dalhousie University, Halifax, NS, Canada
26. Department of Environmental Protection and Natural Resources, Northern Cheyenne Tribe, USA
27. EEZA-CSIC, Spain
28. Elizade University, Ilara-Mokin, Ondo State, Nigeria
29. Emerson College, Boston, MA, USA

30. Emory University, Atlanta, GA, USA
31. Environment and Climate Change Canada, Ottawa, Canada
32. Fanjingshan National Nature Reserve Administration, Tongren, China
33. Florida State University, Tallahassee, FL, USA
34. GEOMAR Helmholtz Center for Ocean Research Kiel, Kiel, Germany
35. Georgia Institute of Technology, Atlanta, GA, USA
36. GreenViet Biodiversity Conservation Center, Da Nang, Viet Nam
37. Gusto Global LLC, Charlotte, NC, USA
38. Huazhong Agricultural University, Wuhan, Hubei, China
39. Hybla Valley Veterinary Hospital, Alexandria, VA, USA
40. Icahn School of Medicine at Mount Sinai, New York, NY, USA
41. Illinois Institute of Technology, Chicago, IL, USA
42. Independent Ornithologist, Merced, CA, USA
43. Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang, China
44. Institute of Botany, Czech Academy of Sciences, Dukelská, Trebon, Czech Republic
45. Institute of Soil Science, Chinese Academy of Sciences, Nanjing, China
46. Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China
47. Instituto Antártico Argentino, Buenos Aires, Argentina
48. Instituto Venezolano de Investigaciones Cientificas (IVIC), Venezuela
49. Juniata College, Huntingdon, PA, USA
50. Lawrence Berkeley National Laboratory, Berkeley, CA, USA
51. Leibniz Institute for Freshwater Ecology and Inland Fisheries, Stechlin, Germany
52. Louisiana State University, Baton Rouge, LA, USA
53. Maharshi Dayanand Saraswati University, Ajmer, India
54. Manchester Metropolitan University, Manchester, UK
55. Max Planck Institute for Developmental Biology, Tübingen, Germany
56. Melbourne Water Corporation, Melbourne, Victoria, Australia
57. Miami University, Oxford, OH, USA
58. Michigan State University, East Lansing, MI, USA
59. Minnesota Department of Natural Resources, St. Paul, MN, USA
60. Morningside College, Sioux City, IA, USA
61. Mount Holyoke College, South Hadley, MA, USA
62. Nanoxis Consulting AB, Gothenburg, Sweden
63. National University of Ireland, Galway, Ireland
64. National University of Mongolia, Ulaanbaatar, Mongolia
65. New York University, New York, NY, USA
66. Northern Arizona University, Flagstaff, AZ, USA
67. Northwestern University, Evanston, IL, USA
68. Novozymes North America Inc., Raleigh-Durham, NC, USA
69. Ohio University, Athens, OH, USA
70. Oregon State University, Corvallis, OR, USA
71. Pacific Northwest National Laboratory, Richland, WA, USA
72. Pennsylvania State University, State College, PA, USA
73. Plymouth Marine Laboratory, Plymouth, England, UK
74. Potsdam University, Potsdam, Germany
75. Purdue University, West Lafayette, IN, USA
76. Rutgers University, New Brunswick, NJ, USA
77. San Diego State University, San Diego, CA, USA
78. San Diego Zoo Institute for Conservation Research, Escondido, CA, USA
79. Sejong University, Seoul, South Korea
80. Smithsonian Tropical Research Institute, Panama City, Panama
81. State University of New York, Cortland, NY, USA

82. Stony Brook University, Stony Brook, NY, USA
83. Södertörn University, Huddinge, Sweden
84. Texas Tech University, Lubbock, TX, USA
85. The Hong Kong Polytechnic University, Hong Kong, China
86. The Ohio State University College of Veterinary Medicine, Columbus, OH, USA
87. The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia
88. Tongren University, Tongren, Guizhou, China
89. Towson University, Towson, MD, USA
90. Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina
91. Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
92. Universitetet i Tromsø, The Arctic University of Norway, Tromsø, Norway
93. University of Alaska, Fairbanks, AK, USA
94. University of Auckland, Auckland, New Zealand
95. University of Bergen, Bergen, Norway
96. University of British Columbia, Vancouver, BC, Canada
97. University of California Berkeley, Berkeley, CA, USA
98. University of California Davis, Davis, CA, USA
99. University of California Irvine, Irvine, CA, USA
100. University of California San Diego, La Jolla, CA, USA
101. University of Central Florida, Orlando, FL, USA
102. University of Chicago, Chicago, IL, USA
103. University of Colorado, Boulder, CO, USA
104. University of Copenhagen, Copenhagen, Denmark
105. University of Delaware, Newark, DE, USA
106. University of East Anglia, Norwich, UK
107. University of Georgia, Athens, GA, USA
108. University of Gothenburg, Gothenburg, Sweden
109. University of Granada, Granada, Spain
110. University of Groningen, Groningen, The Netherlands
111. University of Helsinki, Helsinki, Finland
112. University of Illinois, Chicago, IL, USA
113. University of Maine, Orono, ME, USA
114. University of Maryland, College Park, MD, USA
115. University of Massachusetts Boston, Boston, MA, USA
116. University of Melbourne, Melbourne, Victoria, Australia
117. University of Michigan, Ann Arbor, MI, USA
118. University of Minnesota, Saint Paul, MN, USA
119. University of Naples Federico II, Naples, Italy
120. University of New Mexico, Albuquerque, NM, USA
121. University of New South Wales, Sydney, New South Wales, Australia
122. University of Northern British Columbia, Prince George, BC, Canada
123. University of Notre Dame, South Bend, IN, USA
124. University of Oregon, Eugene, OR, USA
125. University of Pennsylvania, Philadelphia, PA, USA
126. University of Puerto Rico, San Juan, Puerto Rico, USA
127. University of South Bohemia, Č?eské Bud?v ejovice, Czech Republic
128. University of Technology, Sydney, New South Wales, Australia
129. University of Toronto, Toronto, ON, Canada
130. University of Turin, Italy
131. University of Tübingen, Tübingen, Germany
132. University of Utah, Salt Lake City, UT, USA
133. University of Victoria, Victoria, BC, Canada

134. University of Vienna, Vienna, Austria
135. University of Waikato, Hamilton, New Zealand
136. University of Warwick, Coventry, England, UK
137. University of Washington Bothell, Bothell, WA, USA
138. University of Waterloo, Waterloo, ON, Canada
139. University of Wisconsin, Madison, WI, USA
140. University of Wisconsin, Milwaukee, WI, USA
141. University of the Basque Country, Bilbao, Spain
142. Université Pierre et Marie Curie, Station Biologique de Roscoff, Roscoff, France
143. Université de Lyon, Lyon, France
144. Université de Montpellier, CNRS, Montpellier, France
145. Vanderbilt University, Nashville, TN, USA
146. Virginia Institute of Marine Science, Gloucester Point, VA, USA
147. Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
148. Wageningen University and Research Centre, Wageningen, Netherlands
149. Washington University, St. Louis, MO, USA
150. Western University, London, ON, Canada
151. Wildlife Conservation Society and Bronx Zoo, New York, NY, USA
152. Woods Hole Oceanographic Institution, Woods Hole, MA, USA
153. Yale University, New Haven, CT, USA
154. Zhejiang Institute of Microbiology, Hangzhou, Zhejiang, China
155. Zoo Atlanta, Atlanta, GA, USA

References

1. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. USA* **104**, 11436–11440 (2007).

2. Ley, R., Lozupone, C. A., Hamady, M., Knight, R. & Gordon, J. I. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat. Rev. Microbiol.* (2008).

3. Fierer, N. & Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci. USA* **103**, 626–631 (2006).

4. Steele, J. A. *et al.* Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J.* **5**, 1414–1425 (2011).

5. Philippot, L., Raaijmakers, J. M., Lemanceau, P. & van der Putten, W. H. Going back to the roots: the microbial ecology of the rhizosphere. *Nat. Rev. Microbiol.* **11**, 789–799 (2013).

6. Lima-Mendez, G. *et al.* Ocean plankton. Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).

7. Gilbert, J. A. *et al.* Meeting report: the Terabase Metagenomics Workshop and the vision of an Earth Microbiome Project. *Stand. Genomic Sci.* **3**, 243–248 (2010).

8. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome Project: successes and aspirations. *BMC Biol.* **12**, 69 (2014).

9. Shade, A., Caporaso, J. G., Handelsman, J., Knight, R. & Fierer, N. A meta-analysis of changes in bacterial and archaeal communities with time. *ISME J.* **7**, 1493–1506 (2013).

10. 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).

11. Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J. & Lewis, S. E. The Environment Ontology: contextualising biological and biomedical entities. *J. Biomed. Semant.* **4**, 43 (2013).

12. Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R. & Stahl, D. A. Microbial ecology and evolution: a ribosomal RNA approach. *Annu. Rev. Microbiol.* **40**, 337–365 (1986).

13. Goodwin, K. D. *et al.* DNA sequencing as a tool to monitor marine ecological status. *Front. Mar. Sci.* **4** (2017).

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

15. Eloe-Fadrosh, E. A., Ivanova, N. N., Woyke, T. & Kyrpides, N. C. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat. Microbiol.* **1**, 15032 (2016).

16. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).

17. 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).

18. 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).

19. 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**, e00009–15 (2016).

20. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. USA* **103**, 12115–12120 (2006).

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

22. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* (2017).

23. Huse, S. M., Welch, D. M., Morrison, H. G. & Sogin, M. L. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* **12**, 1889–1898 (2010).

24. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).

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

26. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–596 (2013).

27. Nemergut, D. R. *et al.* Decreases in average bacterial community rRNA operon copy number during succession. *ISME J.* **10**, 1147–1156 (2016).

28. Gibbons, S. M. *et al.* Invasive plants rapidly reshape soil properties in a grassland ecosystem. *mSystems* **2**, e00178–16 (2017).

29. Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* **66**, 1328–1333 (2000).

30. Hillebrand, H. On the generality of the latitudinal diversity gradient. *Am. Nat.* **163**, 192–211 (2004).

31. Fuhrman, J. A. *et al.* A latitudinal diversity gradient in planktonic marine bacteria. *Proc. Natl. Acad. Sci. USA* **105**, 7774–7778 (2008).

32. Ladau, J. *et al.* Global marine bacterial diversity peaks at high latitudes in winter. *ISME J.* **7**, 1669–1677 (2013).

33. Milici, M. *et al.* Low diversity of planktonic bacteria in the tropical ocean. *Sci. Rep.* **6**, 6578 (2016).

34. Chu, H. *et al.* Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environ. Microbiol.* **12**, 2998–3006 (2010).

35. Wu, Y., Zeng, J., Zhu, Q., Zhang, Z. & Lin, X. pH is the primary determinant of the bacterial community structure in agricultural soils impacted by polycyclic aromatic hydrocarbon pollution. *Sci. Rep.* **7**, srep40093 (2017).

36. Zhou, J. *et al.* Temperature mediates continental-scale diversity of microbes in forest soils. *Nat. Comms.* **7**, 12083 (2016).

37. Hendershot, J. N., Read, Q. D., Henning, J. A., Sanders, N. J. & Classen, A. T. Consistently inconsistent drivers of microbial diversity and abundance at macroecological scales. *Ecology* **98**, 1757–1763 (2017).

38. Carvalho, J. C., Cardoso, P., Borges, P. & Schmera, D. Measuring fractions of beta diversity and their relationships to nestedness: a theoretical and empirical comparison of novel approaches. *Oikos* **122**, 825–834 (2013).

39. Sonnenburg, E. D. *et al.* Diet-induced extinctions in the gut microbiota compound over generations. *Nature* **529**, 212–215 (2016).

40. Atmar, W. & Patterson, B. D. The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia* **96**, 373–382 (1993).

41. Lomolino, M. V. Investigating causality of nestedness of insular communities: selective immigrations or extinctions? *J. Biogeogr.* **23**, 699–703 (1996).

42. Gaston, K. & Blackburn, T. *Pattern and Process in Macroecology* (Wiley-Blackwell, 2000).

43. Pointing, S. B., Fierer, N., Smith, G. J. D., Steinberg, P. D. & Wiedmann, M. Quantifying human impact on Earth's microbiome. *Nat. Microbiol.* **1**, 16145 (2016).

44. Dornelas, M. *et al.* Assemblage time series reveal biodiversity change but not systematic loss. *Science* **344**, 296–299 (2014).

45. Amano, T., Lamming, J. D. L. & Sutherland, W. J. Spatial gaps in global biodiversity information and the role of citizen science. *BioScience* **66**, 393–400 (2016).

46. Ioannidis, J. P. A. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Q.* **94**, 485–514 (2016).

47. Davies, N. *et al.* The founding charter of the Genomic Observatories Network. *GigaScience* **3**, 1–5 (2014).

48. Alivisatos, A. P. *et al.* A unified initiative to harness Earth's microbiomes. *Science* **350**, 507–508 (2015).

Figures

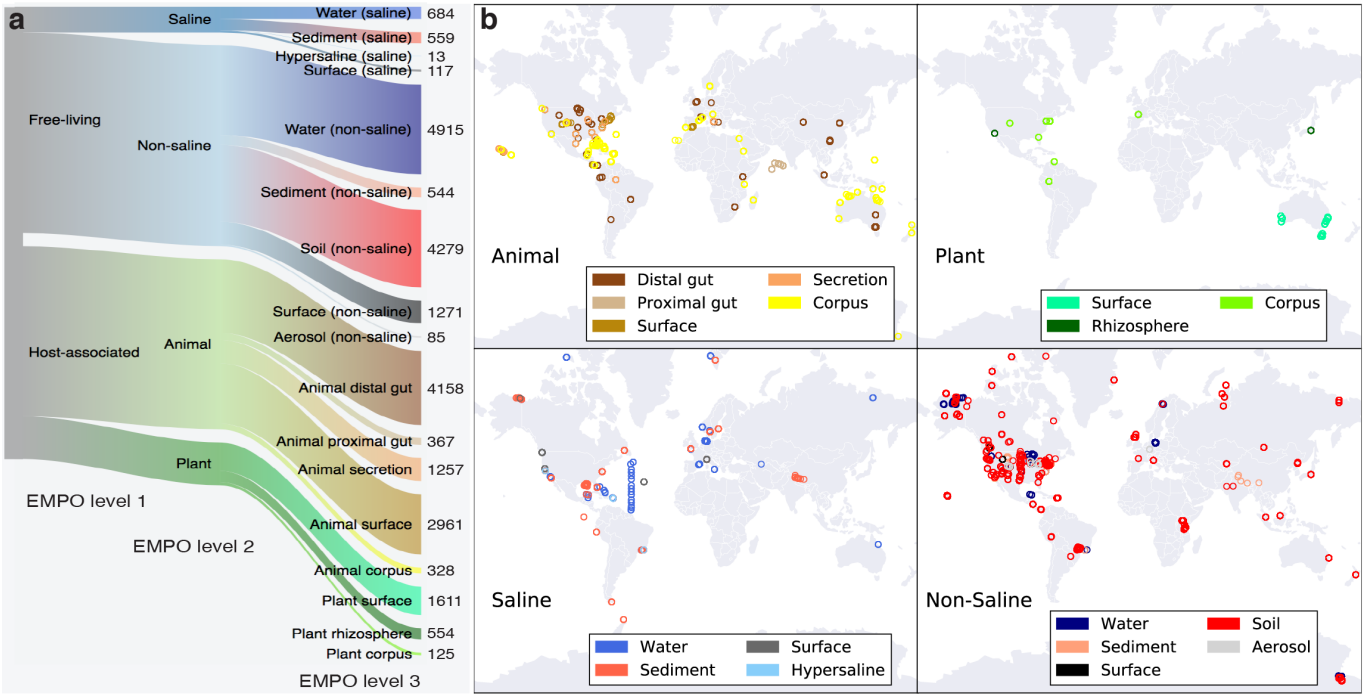


Figure 1 | Environment type and provenance of samples. a, The EMP ontology (EMPO) classifies microbial environments (level 3) as free-living or host-associated (level 1) and saline or non-saline (if free-living) or animal or plant (if host-associated) (level 2). The number out of 23,828 samples in the QC-filtered subset in each environment is provided. EMPO is described with examples at www.earthmicrobiome.org/protocols-and-standards/emp-ontology-empo. **b**, Global scope of sample provenance: samples come from 7 continents, 43 countries, 21 biomes (ENVO), 92 environmental features (ENVO), and 17 environments (EMPO).

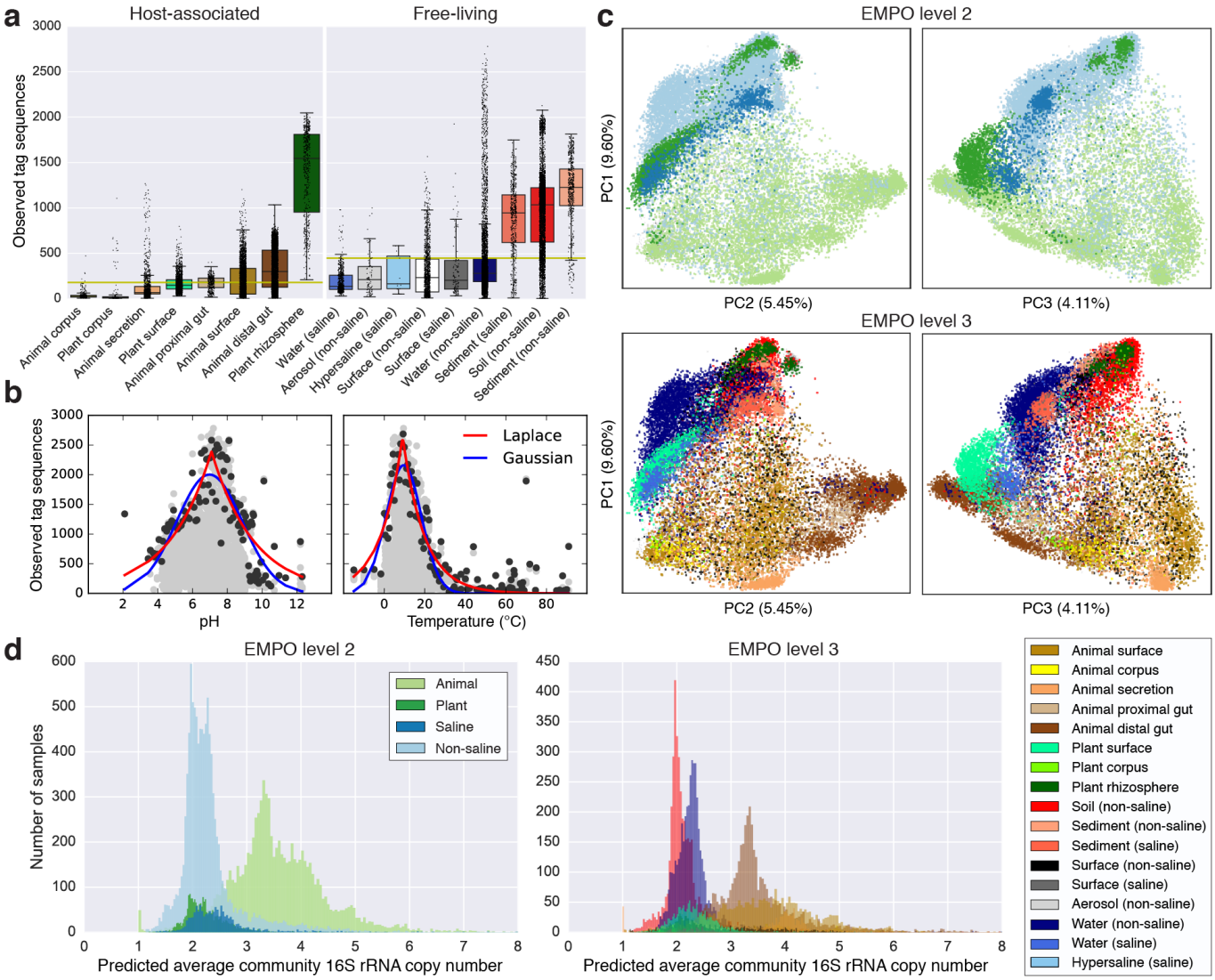


Figure 2 | Alpha-diversity, beta-diversity, and predicted average 16S rRNA gene copy number. a, Within-community (alpha) diversity, measured as number of observed 90-bp tag sequences (richness), in $n = 23,828$ biologically independent samples as a function of environment (per-environment n shown in Fig. 1a), with boxplots showing median, interquartile range (IQR), and 1.5×IQR (with outliers). Tag sequence counts were subsampled to 5,000 observations. Yellow line indicates the median number of observed tag sequences for all samples in that set of boxplots. Free-living communities of most types exhibited greater richness than host-associated communities. **b**, Tag sequence richness (as in **a**) vs. pH and temperature in $n = 3,986$ (pH) and $n = 6,976$ (temperature) biologically independent samples. Black points are the 99th percentiles for richness across binned values of pH and temperature. Laplace (two-sided exponential) curves captured apparent upper bounds on microbial richness and their peaked distributions better than Gaussian curves. Greatest maximal richness occurred at values of pH and temperature that corresponded to modes of the Laplace curves. Maximum richness exponentially decreased away from these apparent optima. **c**, Between-community (beta) diversity among in $n = 23,828$ biologically independent samples: principal coordinates analysis (PCoA) of unweighted UniFrac distance, PC1 vs. PC2 and PC1 vs. PC3, coloured by EMPO levels 2 and 3. Clustering of samples could be explained largely by environment. **d**, 16S rRNA gene average copy number (ACN, abundance-weighted) of EMP communities in $n = 23,228$ biologically independent samples, coloured by environment. EMPO level 2 (left): animal-associated communities had a higher ACN distribution than plant-associated and free-living (both saline and non-saline) communities. Right: soil communities had the lowest ACN distribution, while animal gut and saliva communities had the highest ACN distribution.

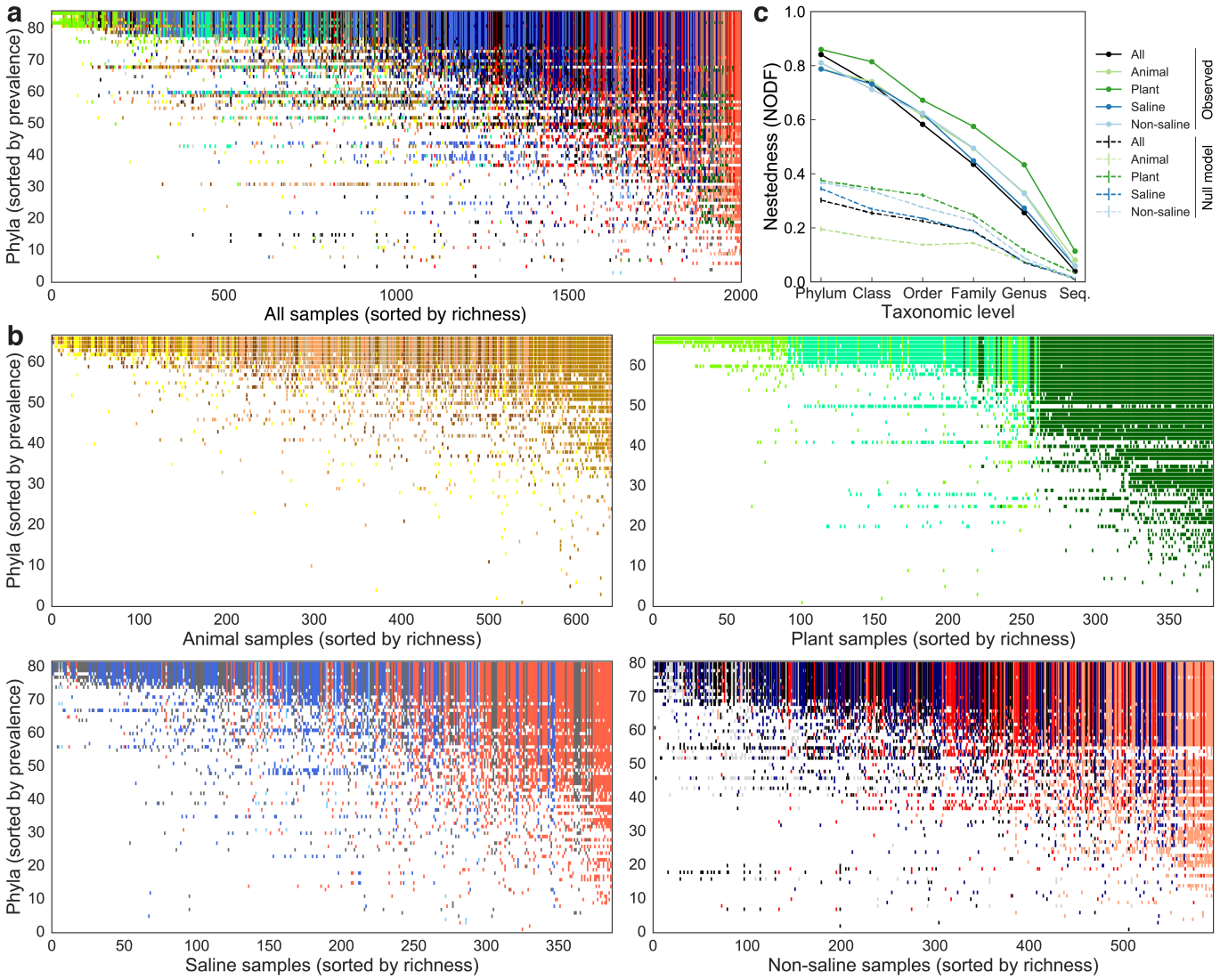


Figure 3 | Nestedness of community composition. a, Presence–absence of phyla across samples, with phyla (rows) sorted by prevalence and samples (columns) sorted by richness. Shown is a subset of the EMP consisting of $n = 2,000$ biologically independent samples with even representation across environments and studies. With increasing sample richness (left to right), phyla tended to be gained but not lost ($p<0.0001$ versus null model; NODF statistic and 95% confidence interval = 0.841 ± 0.018). **b**, As in **a** but separated into non-saline, saline, animal, and plant environments ($p<0.0001$, respective NODF = 0.811 ± 0.013, 0.787 ± 0.015, 0.788 ± 0.018, 0.860 ± 0.021). **c**, Nestedness as a function of taxonomic level, from phylum to tag sequence, across all samples and within environment types. Also shown are median null model NODF scores ($\pm$ standard deviations). NODF measures the average fraction of taxa from less diverse communities that occur in more diverse communities. All environments at all taxonomic levels were more nested than expected randomly, with nestedness higher at higher taxonomic levels (e.g., phyla).

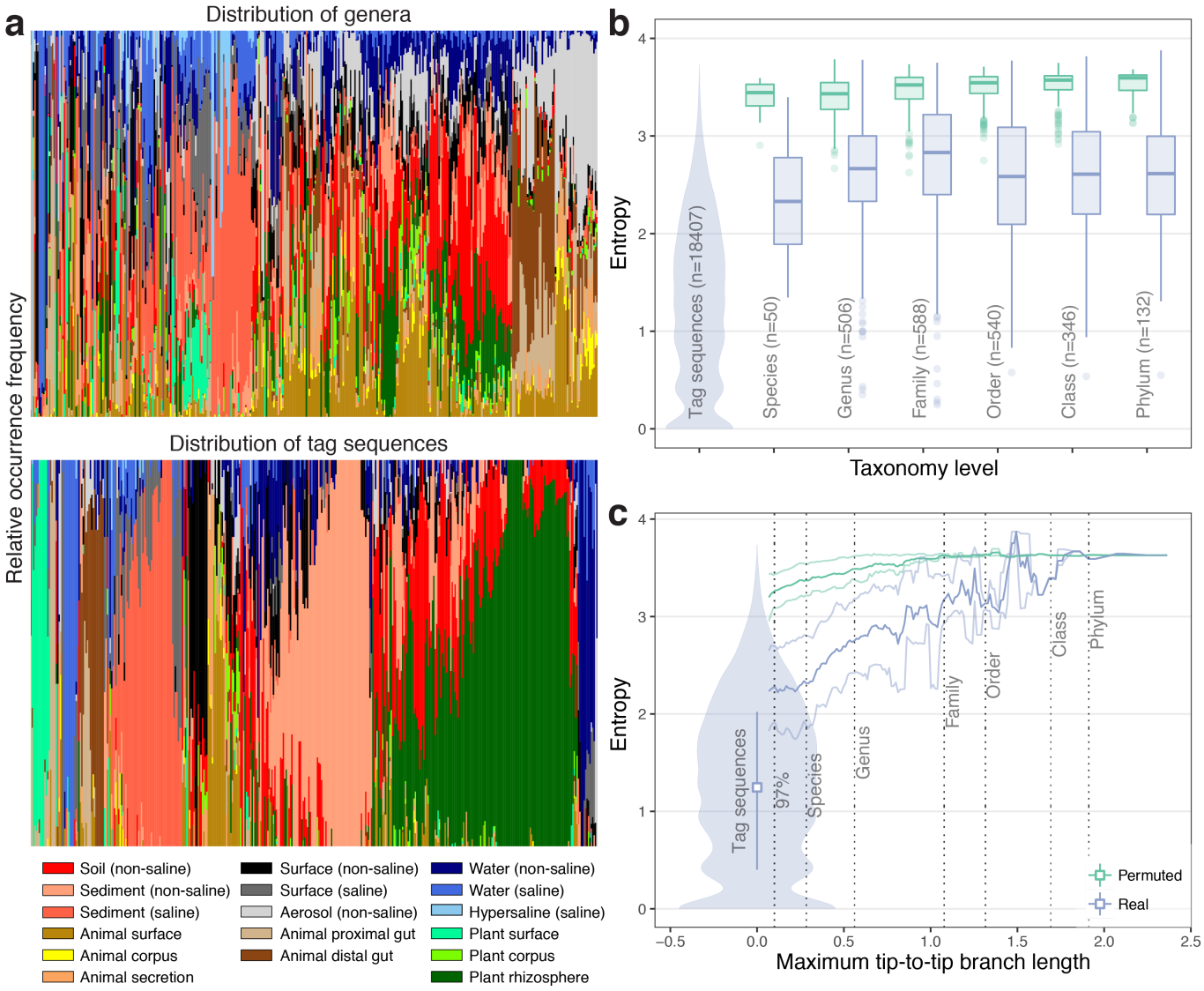


Figure 4 | **Specificity of sequences and higher taxonomic groups for environment.** **a**, Environment distribution in all genera and 400 randomly chosen tag sequences, drawn from $n = 2,000$ biologically independent samples with even representation across environments (EMPO level 3) and studies. Each bar depicts the distribution of environments for a taxon (not relative abundance of taxa): bars composed mostly of one colour (environment) are specific for that environment, as seen with tag sequences; bars composed of many colours are more cosmopolitan, as seen with genera. Tag sequences were more specific for environment than were genera and higher taxonomic levels. **b**, Shannon entropy within each taxonomic group (minimum 20 tag sequences per group) and for the same set of samples with permuted taxonomy labels. Box plots show median, IQR, and 1.5×IQR (with outliers) for each taxonomic level. A violin plot shows the entropy of tag sequences (minimum 10 samples per tag sequence). Specificity for environment occurred predominantly below the genus level. **c**, Shannon entropy within phylogenetic subtrees of tag sequences (minimum 20 tips per subtree) defined by maximal tip-to-tip branch length (substitutions per site) and for the same samples with permuted phylogenetic tree tips. Mean and 20th/80th percentile for a sliding window average of branch length is shown. Violin plot for tag sequences as in **b**. Dotted lines show average tip-to-tip branch length corresponding to 97% sequence identity and taxonomic levels displayed in **b**. The greatest decrease in entropy was between the lowest branch length subtree tested and tag sequences.

Methods

Study design

This effort was possible because of a unified standard workflow that leveraged existing sample and data reporting standards to allow biomass and metadata collection across diverse environments on Earth. After sample collection, all samples were processed following the same protocols. A standard DNA extraction protocol (www.earthmicrobiome.org/emp-standard-protocols/dna-extraction-protocol) was implemented to ensure that trends observed were due either to the biological system or to biases in extraction potential for organisms from different environmental matrices, not due to inherent biases in extraction protocol. To avoid known issues where multiple amplicon strategies are combined[49], we also standardized PCR primers, amplification strategy, and sequencing[50]. More recent studies not included in this meta-analysis adopted additional primer modifications to allow for recovery of key taxa in marine and soil samples[17,18,19]. Data reporting standards, including the MIxS (minimal information about any sequence) metadata standard developed by the Genomic Standards Consortium[10] and the Environment Ontology (ENVO)[11,51], enabled interoperability, data analysis, and interpretation between samples from disparate environments, collected using many different techniques through unconnected programs of investigation.

To transfer our knowledge of microbial environments to the broader community, we engaged with the developers of ENVO to ensure the basic, salient features of microbial environments (host-associated or free-living, and respectively within those, animal- or plant-associated, and saline or non-saline materials) are represented either in this ontology or the others it interoperates with. For ease of application, we gathered these contributions in an application ontology, the EMP Ontology (EMPO) (Fig. 1a). The EMP community will continue to work with ontology engineers to shape ENVO and other ontologies around the EMPO application ontology. EMPO will be maintained as a logical subset of ENVO and integrated into their release cycle to maximize interoperation.

Metadata curation was automated using Pandas (pandas.pydata.org). The size of the dataset also required extensive software development to support analysis at this scale, leading to tools including the data and analysis portal Qiita (qiita.microbio.me), the BIOM format[52], new 'OTU picking' methods Deblur[21] and a subsampled open-reference procedure[53], a scalability improvement of Fast UniFrac phylogenetic inference software[54], speed improvements to sequence-insertion tree method SEPP[55], and speed and feature improvements to Emperor ordination visualization software[56] (biocore.github.io/emperor).

Sample collection

The global community of microbial ecologists was invited to submit samples for microbiome analysis, and samples were accepted for DNA extraction and sequencing provided that scientific justification and high-quality sample metadata were provided before sample submission. Standardized sampling procedures for each sample type were used by contributing investigators. Samples were collected fresh and, where possible, immediately frozen in liquid nitrogen and stored at –80 °C. Detailed sampling protocols are described in publications of the individual studies (Supplementary Table 1). Bulk samples (e.g., soil, sediment, feces) and fractionated bulk samples (e.g., sponge coral surface tissue, centrifuged turbid water) were taken using microcentrifuge tubes. Swabs (BD SWUBE dual cotton swabs or similar) were used for biofilm or surface samples. Filters (Sterivex cartridges, 0.2 μm, Millipore) were used for water samples. Samples were sent to laboratories in the United States for DNA extraction and sequencing: water samples to Argonne National Laboratory, soil samples to Lawrence Berkeley National Laboratory (pre-2014) or Pacific Northwest National Laboratory (2014 onward), and fecal and other samples to the University of Colorado Boulder (pre-2015) or the University of California San Diego (2015 onward).

Metadata curation and EMP Ontology (EMPO)

Metadata were collected in compliance with MIMARKS[10], EBI (www.ebi.ac.uk/ena), and Qiita (qiita.microbio.me) standards, as described in the EMP Metadata Guide (www.earthmicrobiome.org/emp-standard-protocols/metadata-guide). QIIME mapping files (metadata) were downloaded from Qiita, merged, and refined using Python with Pandas, generating quality-controlled mapping files. Mapping file columns are described in Supplementary Table 2. Mapping files for the full EMP dataset and subsets (see below) are available at ftp.microbio.me/emp/release1/mapping_files. The EMP Ontology (EMPO) for microbial environments was devised to facilitate the present analysis while preserving interoperability. Coordinated by the ENVO team, annotations from ENVO[11,51], UBERON (metazoan anatomy)[57], PO (Plant Ontology)[58], FAO (Fungal Anatomy Ontology, purl.obolibrary.org/obo/fao.owl), and OMP (Ontology of Microbial Phenotypes)[59] were mapped to our EMPO levels 2 and 3 (empo_2 and empo_3). Additionally, the free-living or host-associated lifestyles were captured in level 1 categories (empo_1).

Descriptions of empo_3 categories are provided at www.earthmicrobiome.org/protocols-and-standards/emp-ontology-empo. The W3C Web Ontology Language (OWL) document is available at purl.obolibrary.org/obo/envo/subsets/envoEmpo.owl. Map data were derived from the open-source project MatPlotLib package Basemap, which distributes map data from Generic Mapping Tools data (http://gmt.soest.hawaii.edu) released under the GNU Lesser General Public License v3.

DNA extraction, amplicon PCR, sequencing, and sequence pre-processing

DNA extraction and 16S rRNA amplicon sequencing was done using EMP standard protocols (www.earthmicrobiome.org/protocols-and-standards/16s)[14]. Briefly, DNA was extracted using the MO BIO PowerSoil DNA extraction kit (Carlsbad, CA), chosen because of its versatility with diverse sample types (rather than high yields with any given sample type). Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair 515f–806r[50] with Golay error-correcting barcodes on the reverse primer. Although any primer-based method necessarily under-samples diversity, a recent analysis of 16S rRNA genes captured in shotgun metagenomic sequences indicates that this primer pair is among the best available for sampling both bacteria and non-eukaryotic archaea[15]. Amplicons were barcoded and pooled in equal concentrations for sequencing. The amplicon pool was purified with the MoBio UltraClean PCR Clean-up kit and sequenced on the Illumina HiSeq or MiSeq sequencing platform; the same sequencing primers were used with both platforms, and previous work has shown that conclusions drawn from 16S rRNA amplicon data are not dependent on which of these sequencing platforms is used[50]. Sequence data were demultiplexed and minimally quality filtered using the QIIME 1.9.1 script split_libraries_fastq.py[60] with Phred quality threshold of 3 and default parameters to generate per-study FASTA sequence files.

Tag sequence/OTU picking and subsets

Sequence data were error-filtered and trimmed to the length of the shortest sequencing run (90 bp) using the Deblur software[21]; the resulting 90-bp Deblur BIOM table was used for all analyses unless otherwise noted. Deblur tables trimmed to 100 bp and 150 bp were also generated and provided, which contain greater sequence resolution but fewer samples. Deblur observation tables were filtered to keep only tag sequences with at least 25 reads total over all samples. For comparison to existing OTU tables, traditional closed-reference OTU picking was done against 16S rRNA databases Greengenes 13.8[25] and SILVA 123[26] using SortMeRNA[61], and subsampled open-reference OTU picking[53] was done against Greengenes 13.8. These unfiltered tables and the filtered and subset tables described below are available at ftp.microbio.me/emp/release1/otu_tables.

A total of 97 studies and 27,742 samples are included in the present study and in the unfiltered BIOM tables. The QC-filtered subset used in core diversity analyses (Fig. 2) contains 96 studies and 23,828 samples, and it was subset further for some analyses. In the Supporting Data, the 'all_emp' set contains all samples in the 97 studies that have at least one sequence per sample; this set includes controls (blanks and mock communities). The 'qc_filtered' set, from which the subsets are drawn, has samples with $\geq$1000 observations in each of four observation tables: closed-reference Greengenes, closed-reference SILVA, open-reference Greengenes, and Deblur 90-bp; controls (empo_1 == 'Control') are excluded. Subsets were then generated which give equal (as possible) representation across environments (EMPO level 3) and across studies within those environments. The subsets contain 10,000, 5,000, and 2,000 samples (nested subsets). In each subset the samples must have $\geq$5000 observations in the Deblur 90-bp observation table and $\geq$10000 observations in each of closed-reference Greengenes, closed-reference SILVA, and open-reference Greengenes observation tables. Note that Deblur removes approximately one-third to one-half of sequences due to suspected errors, which is consistent with a sequence length of ~90–150 bp and an average error rate of 0.006 per position[62].

Comparison against reference databases

To compare the unique sequence diversity in this study to that in existing databases, sequences from the complete Deblur 90-bp observation table were compared to the set of unique full-length sequences from Greengenes 13.8 and the non-eukaryotic fraction of Silva 128 databases using the open-source sequence search tool VSEARCH[63] in global alignment search mode, requiring 100% similarity across the query sequence and allowing multiple 100% reference matches.

Prevalence as a function of sequencing depth

The QC-filtered Deblur 90-bp observation table was additionally filtered to samples that had at least 50,000 sequences (observations). We chose to focus on four environment types (EMPO level 3) where there were many hundreds of samples with more

than 50,000 sequences: soil ($n = 2,279$), saltwater ($n = 478$), freshwater ($n = 1,508$), and animal distal gut ($n = 695$) environments. For each environment, the observation tables were randomly subsampled to 50, 500, 5,000, and 50,000 sequences per sample. Prevalence of each tag sequence was determined as the number of non-zero occurrences across samples divided by the total number of samples. We then plotted a histogram of tag sequence prevalence at each sampling depth. In order to control for potential study bias, we ran the same analysis on a subset of the observation tables where 30 samples were randomly sampled from each study (studies with fewer than 30 samples with >50000 sequences were discarded). To examine how mean tag sequence prevalence changes with increasing sequencing depth across environments, we calculated the average mean tag sequence prevalence across 3 replicate rarefactions. We plotted the average and standard deviation in mean prevalence across replicate subsamples over a subsampling gradient (i.e., 50, 100, 500, 1,000, 5,000, 10,000, 50,000 sequences per sample).

Greengenes insertion tree

Deblur tag sequences were inserted into the Greengenes reference tree using SEPP[55], which uses a divide-and-conquer technique to enable phylogenetic placement on very large reference trees. The SEPP method uses HMMER[64] internally for aligning each Deblurred sequence to a reference Greengenes alignment (gg_13_5_ssu_align_ 99_pfiltered.fasta) with 99% threshold for clustering (resulting in 203,452 tag sequences) and dividing the reference alignment to subsets with a thousand sequences each. It then uses pplacer[65] to insert the sequences into the reference Greengenes tree (99_otus.tree), dividing it into subsets of size 5,000. The branch lengths on the Greengenes tree were recomputed using RAxML[66] under the GTRCAT model prior to the placement. The pipeline used, including the reference trees and alignments can be found at ftp.microbio.me/emp/latest/otu_info/greengenes_sepp_pipeline, and the bash script is available at github.com/biocore/emp/blob/master/scripts/03-phylogenetic-placement/run_sepp.sh.

Fast UniFrac

Unweighted and weighted UniFrac were computed using the Cythonized[67] implementation of Fast UniFrac[54] in scikit-bio[68]. Fast UniFrac by itself was not scalable for the EMP dataset due to an intermediary data structure required by the algorithm, which scales in space by $O(NM)$, where N is the number of nodes in the phylogeny and M is the number of samples. A workaround was designed and implemented in scikit-bio (skbio.diversity.block_beta_ diversity) which computes partial distance matrices as opposed to all samples pairwise, enabling large reductions within the intermediary data structure by shrinking M and, in tandem, shrinking N to only the relevant nodes of the phylogeny. This decomposition additionally allows for a classic map-reduce parallel approach with low per-process space requirements. Further space and time reductions were obtained through the implementation and use of a balanced parentheses tree representation[69] (pypi.python.org/pypi/iow).

Core diversity analyses: alpha- and beta-diversity

Alpha-rarefaction was computed using single_rarefaction.py in QIIME 1.9.1[60] using as input the Deblur 90-bp BIOM table and rarefaction depths of 1,000, 5,000, 10,000, 30,000, and 100,000. Alpha-diversity was computed using scikit-bio 0.5.0 with the input Deblur 90-bp BIOM table rarefied to 5,000 observations per sample, and alpha-diversity indices were observed_otus (number of unique tag sequences), shannon (Shannon diversity index[70]), chao1 (Chao 1 index[71]), and faith_pd (Faith's phylogenetic diversity[72], using the Greengenes insertion tree). Fast UniFrac[54] was run on the Deblur 90-bp table using the aforementioned approach and the corresponding insertion tree. Principal coordinates were computed using QIIME 1.9.1.

Effect size calculations of alpha- and beta-diversity

A version of the mapping file (metadata) was compiled containing the predictors to be tested: study_id, host_scientific_ name (a proxy for host taxonomy), latitude_deg, longitude_deg, envo_biome_3 (a proxy for biome or environment), empo_3 (a proxy for sample type or environment generally), temperature_deg_c, ph, salinity_psu, and nitrate_umol_per_l (a proxy for nutrient levels generally). Predictors chosen were those expected to be less redundant with other predictors not chosen, with the exception that there was significant overlap between study ID and many of the other predictors—because independent studies typically focused on limited sample types from constrained geographic ranges, it is expected that study ID serves as a proxy for a wide range of other measured and unmeasured environmental variables (see Fig. 5b). Categories for each predictor were chosen as follows: numerical data were first converted to categories using quartiles; then each category was required to be found in at least 0.3% of all samples (corresponding to 75 samples); categories that were less common than this were ignored. Note that some predictors in our data have complex non-linear relationships that multivariate statistical analyses using quartiles may miss, such as the

unimodal upper-constraint-based richness relationships of temperature and pH. We then tested the effect size of each predictor versus the number of observed tag sequences (alpha-diversity) and weighted and unweighted UniFrac distances (beta-diversity). Effect size was calculated using a Python implementation of the mixed-directional false discovery rate (mdFDR)[73,74]. mdFDR reduces the false discovery rates by penalizing the multiple pairwise comparisons within each metadata category and the multiple metadata category comparisons. mdFDR has four steps. First, it performs a pairwise comparison (Mann–Whitney U for alpha-diversity, and PERMANOVA for beta-diversity) of each group within each category. Second, for each category we calculate a pooled p-value based on the p-values of all pairwise comparison for any given category. Third, we apply the Benjamini–Hochberg procedure on the pooled p-values and remove non-significant metadata categories. Finally, we estimate the effect size of those categories found significant in step 3 and that have a pairwise comparison p-value greater than $(R/m * q_i) * \alpha$, where R is the number of categories that were found significant, m is the number of categories that are being compared (the original number of categories in the input mapping file), q_i is the number of pairwise comparisons in each given category, and α is the control level for FDR. The effect size for a given metadata column is calculated as the difference of means of each pairwise comparison divided by pooled standard deviation. To further assess the combined effect size of predictors with non-redundant explanatory power on alpha and beta diversity, the non-redundant predictors were selected by forward stepwise redundancy analysis with the R package *vegan*[75] *ordiR2step* function. This analysis provides an estimate of the relative contribution of each non-redundant predictor to the combined effect size and their independent fraction to the community variation.

Average community 16S rRNA gene copy number

The closed-reference observation table (Greengenes 13.8) was run through the PICRUSt 1.1.0 command normalize_by_copy_number.py script[76], which divides the abundance of each OTU by its inferred 16S rRNA gene copy number (i.e., copy number is inferred from the closest genome representative for a Greengenes 16S rRNA gene reference sequence). Samples with more than 10,000 sequence reads were summed (i.e., OTU abundances were summed within each sample) in both the copy-number normalized and original observation tables. The weighted average community 16S rRNA gene copy number (ACN) for each sample was calculated as the raw sample sum divided by the normalized sample sum.

Covariation of richness with latitude, pH, and temperature

Measurements of alpha diversity were compared to absolute latitude using a linear mixed-effects model incorporating study ID as a random variable and the interaction of environment and latitude as fixed effects; this was performed on a dataset filtered to only include studies comprising samples that spanned at least 10° of absolute latitude. Correlation of richness with pH and temperature were fitted with a Laplace distribution. The Laplace distribution is a continuous probability distribution that simultaneously captures exponential increase and exponential decrease around a modal value (μ). This distribution is also referred to as the double exponential or two-sided exponential because it represents two symmetrical exponential distributions back-to-back. The Laplace is particularly useful for testing the biological hypothesis that a system is under strong selection to take a particular value (μ) and that small deviations from μ produce an exponential decrease, e.g., in diversity. We tested this hypothesis in regards to how tag sequence richness (S) relates to pH and temperature. We used the upper 99th percentile of tag sequence richness across narrow ranges of pH (100 bins) and temperature (120 bins), meaning that our question pertained to the relationship of maximum tag sequence richness (S_{max}) to pH and temperature. We compared our expectations of exponential decrease in maximum S against the fit to a Gaussian curve, which can also predict a steep symmetrical decrease with small deviations from μ.

Random forest classification of samples

Random forest classification models were trained on the 2,000-sample subset of the Deblur 90-bp observation table to test classification success of samples into the environmental categories from which they came. The R packages *caret*[77] and *randomForest*[78] were used. Five repeats of 10-fold cross-validation were used to evaluate the classification accuracy. Confusion matrices were computed to measure the agreement between prediction and true observation. The models were then used to classify the other remaining in the full QC-filtered subset.

SourceTracker analyses

SourceTracker[79] uses a Bayesian classification model together with Gibbs sampling to predict the proportion of tag sequences from a given set of source environments that contribute to sink environments. We applied SourceTracker 2.0.1 (github.com/biota/sourcetrack

to define the degree to which tag sequences are shared among environmental samples, using the 2,000-sample subset of the Deblur 90-bp observation table (~20% of each sample type) as source samples to train the model, and the remainder as sink samples to test the model. Additionally, we used leave-one-out cross-validation to predict the sample type of each source sample when that sample type is excluded from the model, in order to evaluate the homogeneity of source samples and independence of each source type. Source and sink samples were rarefied to 1,000 sequences per sample prior to feature selection and testing.

Nestedness of taxonomic composition

Nestedness captures the degree to which elements of a large set are contained within progressively smaller sets. We used the NODF statistic[80] to quantify nestedness of the sample-by-taxa matrix. The rows of this matrix correspond to specific taxa grouped at particular taxonomic levels (e.g., phylum, class, etc.), while the columns correspond to particular samples. After sorting the matrix from greatest-to-least according to row and column sums, we quantified two aspects of the NODF statistic. The first is a 'row' version of NODF that quantifies the degree to which ranges of less prevalent taxa are subsets of the ranges of more prevalent taxa. The second is a 'column' version of NODF that quantifies the degree to which less diverse communities are subsets of more diverse communities. We employed two null models to better interpret the observed values of the NODF statistic. The first is based on a random shuffling of occurrences within each row, holding row sums constant (fixed rows, equiprobable columns), while the second is based on a random shuffling of occurrences within each column, holding column sums constant (equiprobable rows, fixed columns)[81]. The results from both of these null models were qualitatively consistent, so we only report findings using the equiprobable rows, fixed columns model, as it is more consistent with rarefaction of the observation tables. We considered null models at each taxonomic level (phylum, class, order, family, genus), and for all of the samples and each subset of the samples at EMPO level 2. To compute standardized effect scores (SES), we used analytical results based on the hypergeometric distribution to find the expectation and variance of the NODF statistic under both models. SES values were generally very large (>2); we used Wald Tests to compute approximate p-values.

Environment distribution of taxa and Shannon entropy

For each Deblur tag sequence B, sample s in the set of all EMP samples S, and sample type (EMPO level 3) category E, define

$$W_E(B) = \frac{|s \in S : B \in S \wedge s \in E|}{|s \in S : B \in s|} \tag{1}$$

as the fraction of total appearances of tag sequence B in sample type category E (with N possible values). For a given cluster of tag sequences T (phylogenetic subtree or taxonomic group, e.g., Firmicutes), we then calculate cluster distribution vector as

$$W(T) = (W_{E1}(T), \ldots, W_{EN}(T)), \tag{2}$$

where W_E combined for all tag sequences in the sequence cluster is given by

$$W_E(T) = \underset{B \in T}{\text{mean}}(W_E(B)). \tag{3}$$

Clusters of tag sequences were defined in two different ways: first, by partitioning using the taxonomic lineage information for the tag sequences; second, by maximum tip-to-tip branch length for nodes on the phylogenetic tree. For calculating entropy of environment distribution as a function of taxonomic level (e.g., phylum), the mean of Shannon entropies for all taxonomic groups belonging to that taxonomic level was calculated (weighted by the number of tag sequences in each taxonomic group). For calculating the entropy as a function of the phylogenetic subtree group width, cluster Shannon entropy was calculated for all subtrees, as well as the maximum tip-to-tip distance for each subtree. To ascertain whether changes in entropy between taxonomic and phylogenetic levels were expected given the observed distribution of environment entropy among tag sequences, a null model was calculated by randomly permuting the Deblur tag sequence taxonomy associations (for the entropy vs. taxonomy analysis) or the phylogenetic tip placement (for the entropy vs. phylogeny analysis). In order to reduce the effect of discretization on the entropy calculation in both analyses, clusters of tag sequences were included in the analysis only if they had a minimum of 20 tag sequences. For unique tag sequences (i.e., a branch length threshold of 0.0), sequences were required to be found in a minimum of 10 samples. To calculate the approximate branch length corresponding to each taxonomic level, we found the lowest common ancestor for each group and calculated the maximum tip-to-tip distance in that subtree.

EMP trading cards

We started with a BIOM table of 90-bp Deblur tag sequences (16S rRNA gene, V4 region), rarefied to 5,000 observations per sample, containing 2,000 samples evenly distributed across environments and studies (Extended Data Fig. 7a). From this we calculated the following: the number, fraction, and rank of samples in which a tag sequence is found; the abundance, fraction, and rank of observations represented by that tag sequence; taxonomy of the tag sequence from Greengenes; and the list of all the samples the tag sequence is found in. This summary is located at ftp.microbio.me/emp/release1/otu_distributions. Additionally, for each tag sequence sequence with a trading card in Extended Data Fig. 7b or earthmicrobiome.org/trading-cards, we identified sequences in RDP (rdp.cme.msu.edu)[82] matching 100% along the 90-bp region of the 16S rRNA gene. Trading cards at earthmicrobiome.org/trading-cards are those with prevalence or abundance in the top 10 of all tag sequences or the most abundant tag sequence for each environment having a distribution Shannon entropy <1, a proportion of that environment $\geq 25\%$, and total observations ≥ 1000.

Redbiom database service

A metadata and feature search service containing the EMP data is available through Redbiom. Redbiom is a caching layer for BIOM table and sample metadata, where by default it allows users to interact with the public portion of Qiita (which includes all of the EMP studies). This service allows users to find samples based on sample details (e.g., all soil samples with ph < 7), to find samples based on features they contain (e.g., all samples in which Greengenes ID 131337 exists), to find features based on taxonomy (e.g., all samples in which genus *Pyrobaculum* exists), to extract sample data into BIOM tables, and to extract sample metadata. Installation of the command-line client and usage instructions are available at pypi.python.org/pypi/redbiom; examples of command-line queries are provided at github.com/biocore/redbiom. A graphical user interface for Redbiom is available at qiita.microbio.me.

Data Availability

Per-study sequence files and sample metadata are available from EBI (www.ebi.ac.uk/ena) with accession numbers in Supplementary Table 1. Per-study sequence files, sample metadata, and observation tables and information are available from Qiita (qiita.microbio.me/emp) using study IDs in Supplementary Table 1. EMP-wide sample metadata, observation tables and information (trees and taxonomies), alpha- and beta-diversity results, and observation summaries for trading cards are available at ftp.microbio.me/emp/release1; these files plus the Redbiom database at time of publication are archived at zenodo.org with DOI 10.5281/zenodo.890000.

Code Availability

Code for reproducing sequence processing, data analysis, and figure generation is provided at github.com/biocore/emp and is archived at zenodo.org with DOI 10.5281/zenodo.1009693. Redbiom code is available at github.com/biocore/redbiom and is archived at zenodo.org with DOI 10.5281/zenodo.1009150.

References

49. Gohl, D. M. *et al.* Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* **34**, 942–949 (2016).

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

51. Buttigieg, P. L., Pafilis, E. & Lewis, S. E. The Environment Ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semant.* (2016).

52. McDonald, D. *et al.* The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience* **1**, 7 (2012).

53. Rideout, J. R. *et al.* Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* **2**, e545 (2014).

54. Hamady, M., Lozupone, C. & Knight, R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J.* **4**, 17–27 (2010).

55. Mirarab, S., Nguyen, N. & Warnow, T. SEPP: SATé-enabled phylogenetic placement. *Pac. Symp. Biocomput.* 247–258 (2012).

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

57. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5 (2012).

58. Cooper, L. *et al.* The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant Cell Physiol.* **54**, e1–e1 (2013).

59. Chibucos, M. C. *et al.* An ontology for microbial phenotypes. *BMC Microbiol.* **14**, 294–294 (2014).

60. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).

61. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics (Oxford, England)* **28**, 3211–3217 (2012).

62. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLOS ONE* **6**, e27310 (2011).

63. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahe, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

64. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).

65. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinform.* **11**, 538 (2010).

66. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

67. Behnel, S. *et al.* Cython: the best of both worlds. *Comput. Sci. Eng.* **13**, 31–39 (2011).

68. Rideout, J. R. *et al.* scikit-bio: scikit-bio 0.5.0: Python 3 only release (2016).

69. Cordova, J. & Navarro, G. Simple and efficient fully-functional succinct trees. *Theor. Comput. Sci.* **656**, 135–145 (2016).

70. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27** (1948).

71. Chao, A. Nonparametric estimation of the number of classes in a population. *Scand. J. Statist.* (1984).

72. Faith, D. P. Conservation evaluation and phylogenetic diversity (1992).

73. Guo, W., Sarkar, S. K. & Peddada, S. D. Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics* **66**, 485–492 (2010).

74. Grandhi, A., Guo, W. & Peddada, S. D. A multiple testing procedure for multi-dimensional pairwise comparisons with application to gene expression studies. *BMC Bioinform.* **17**, 104 (2016).

75. Oksanen, J. *et al. vegan: Community Ecology Package* (2017). R package version 2.4-3.

76. Langille, M. G. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).

77. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Soft.* **28** (2008).

78. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2-3** (2002).

79. Knights, D. *et al.* Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8**, 761–763 (2011).

80. Almeida-Neto, M., Guimaraes, P., Guimaraes, P. R. J., Loyola, R. D. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos* **117**, 1227–1239 (2008).

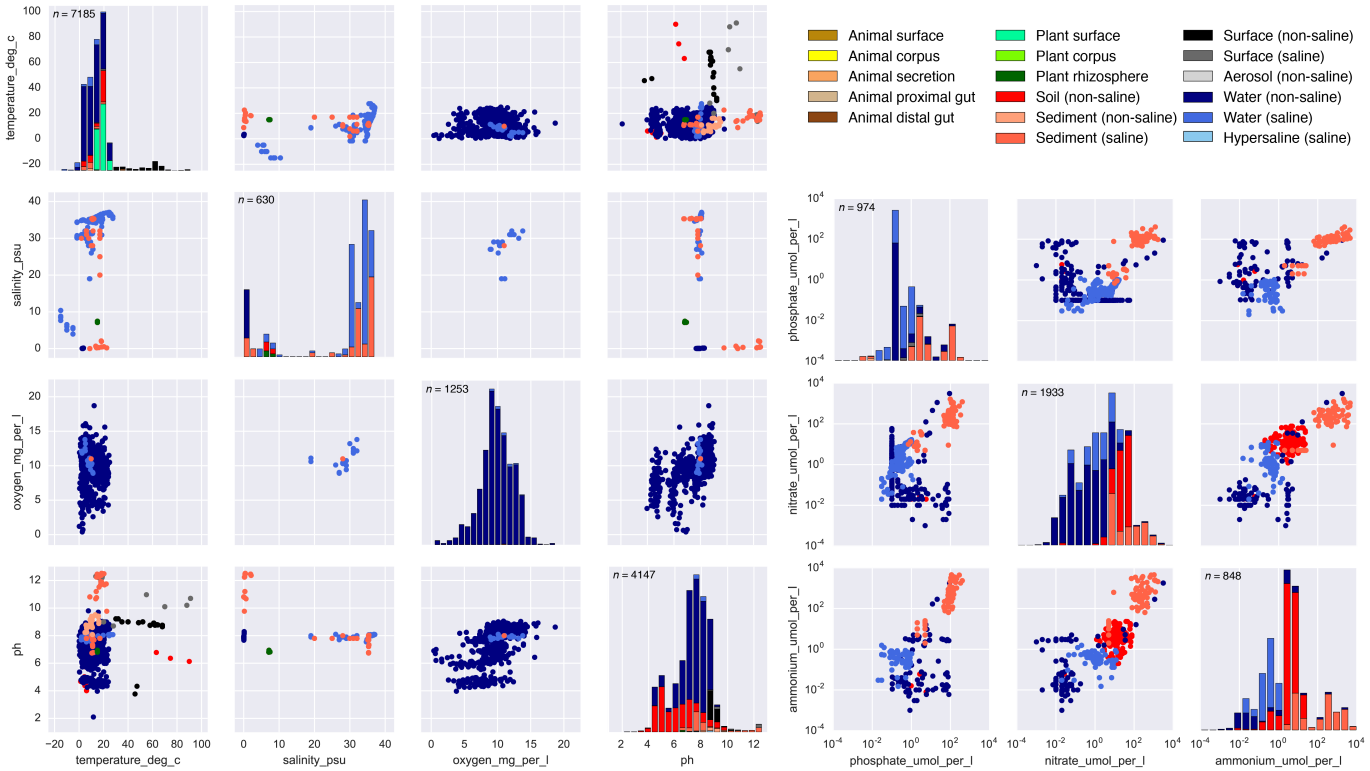
81. Ulrich, W., Almeida-Neto, M. & Gotelli, N. J. A consumer's guide to nestedness analysis. *Oikos* **118**, 3–17 (2009).

82. Cole, J. R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633–D642 (2013).

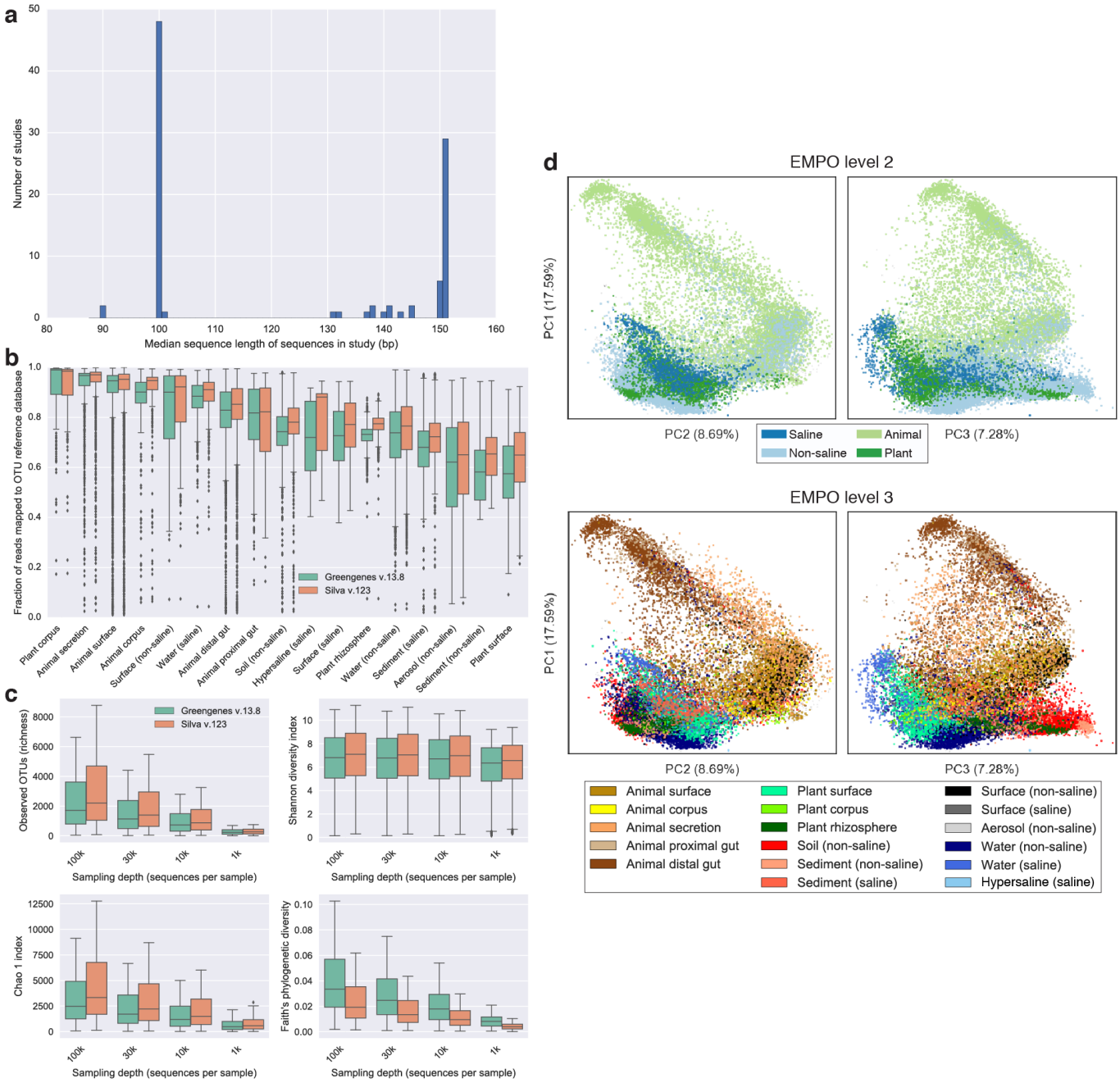
83. Arons, M. S., Fernando, L. & Polayes, I. M. Pasteurella multocida–the major cause of hand infections following domestic animal bites. *J. Hand Surg. Am.* **7**, 47–52 (1982).

84. Ormerod, K. L. *et al.* Genomic characterization of the uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals. *Microbiome* **4**, 36 (2016).

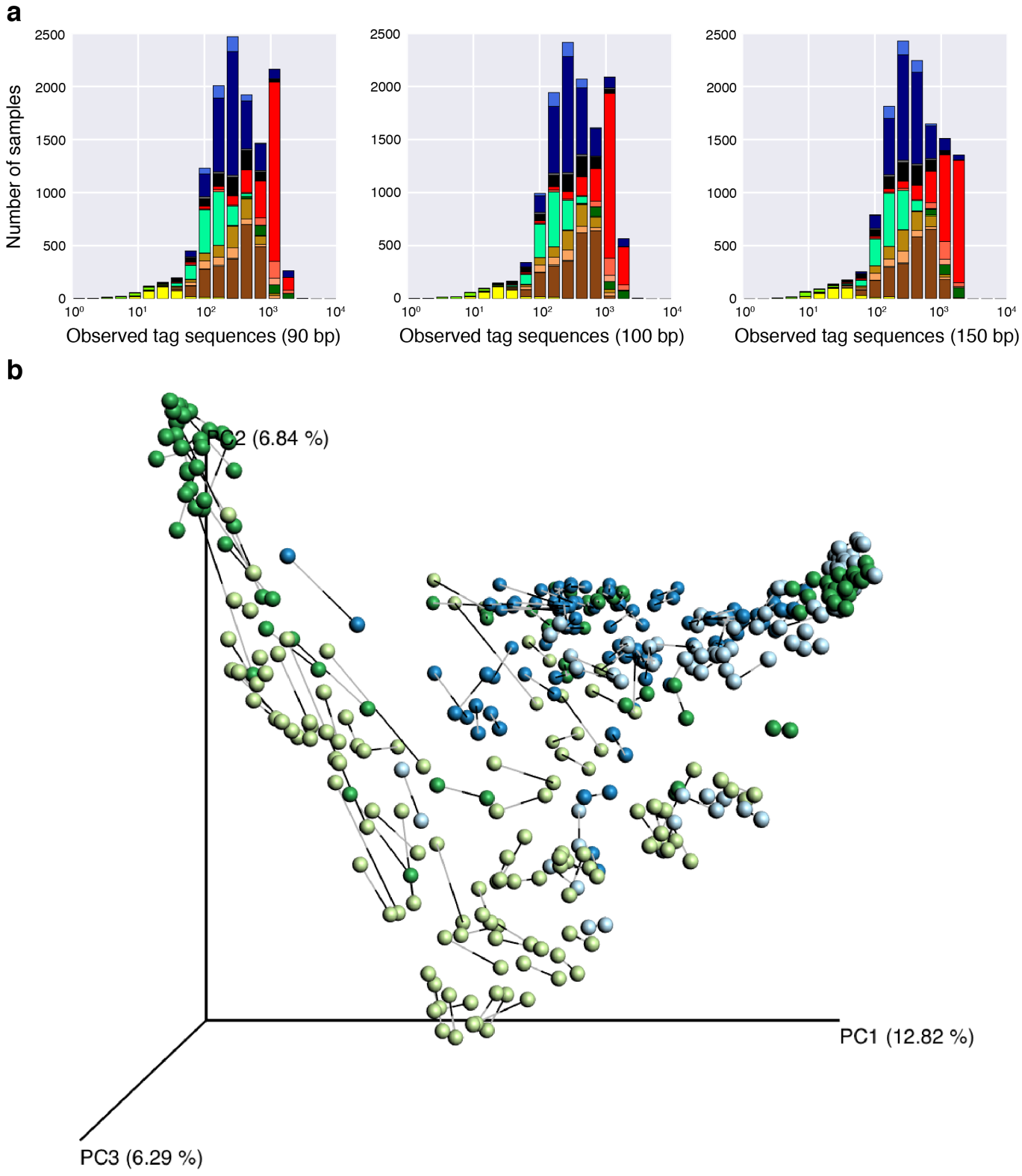
Extended Data Figures



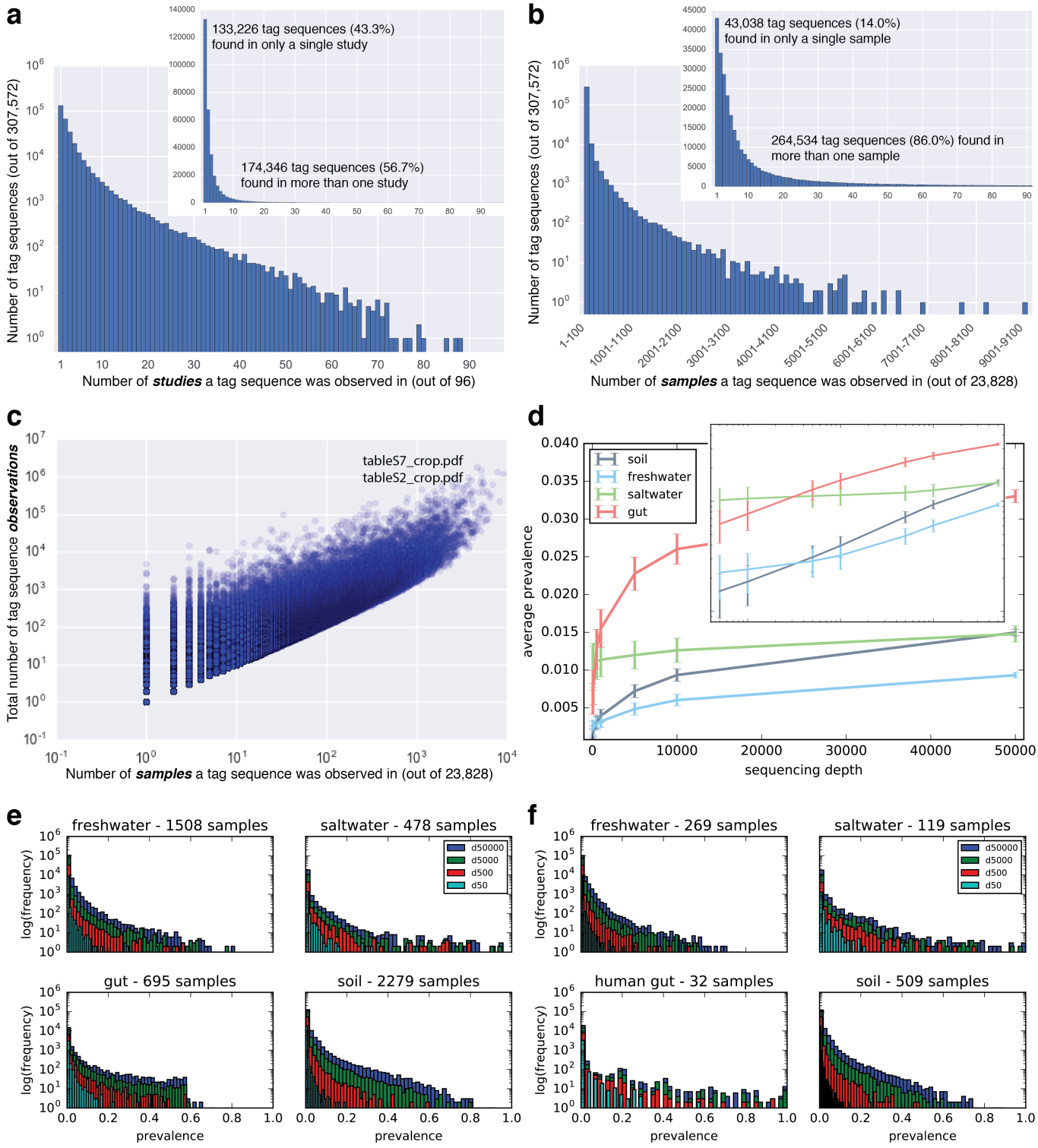
Extended Data Figure 1 | Physicochemical properties of the EMP samples. Pairwise scatter plots of available physicochemical metadata are shown for temperature, salinity, oxygen, and pH, and for phosphate, nitrate, and ammonium. Histograms for each factor are also shown; number (*n*) of samples having data for each factor are provided at the top of each histogram. Samples are coloured by environment, and only QC-filtered samples are included. Environmental factors are named in our recommended format, with analyte name and units combined in the metadata field name.



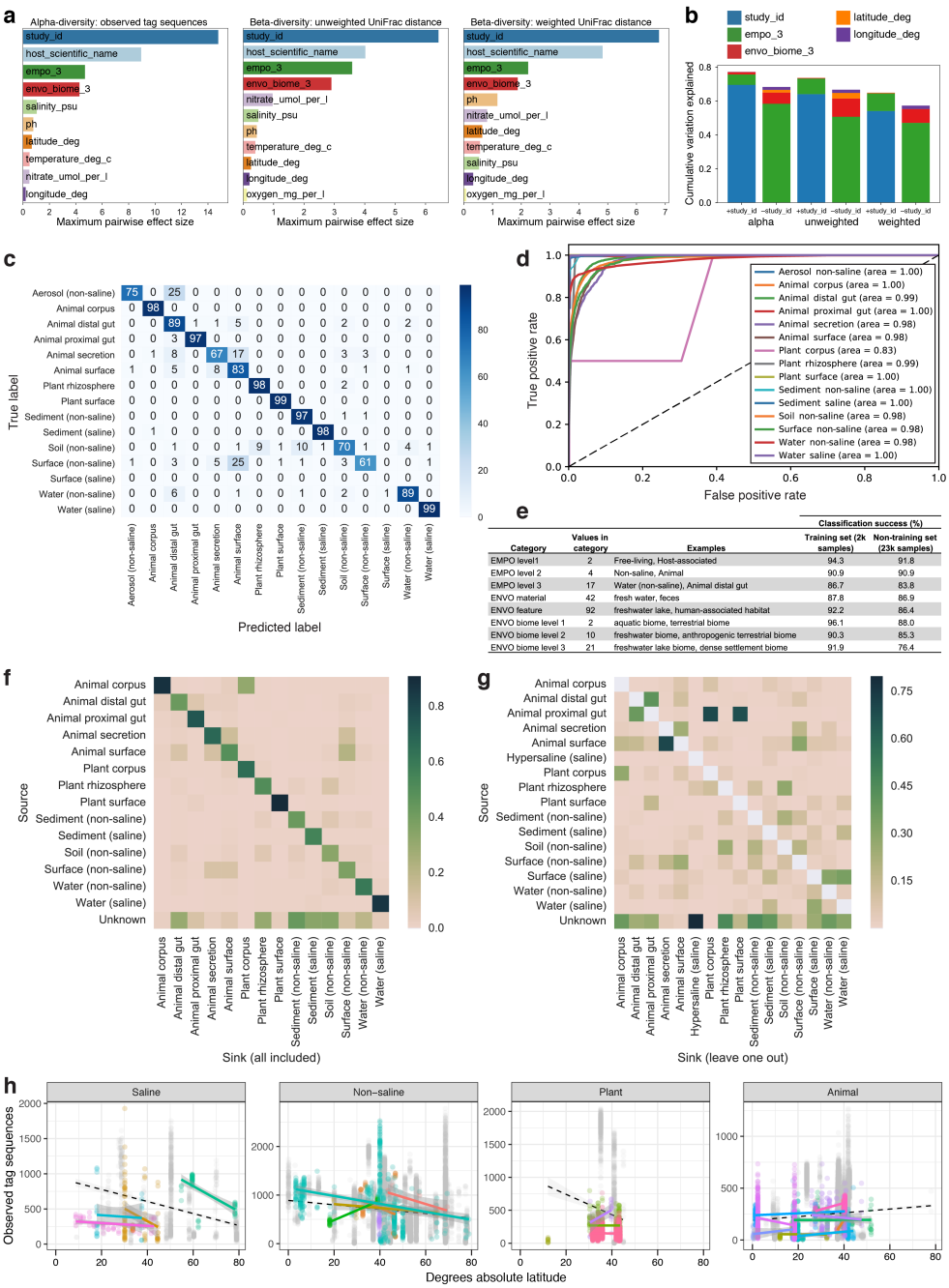
Extended Data Figure 2 | Sequence length, database effects, and beta-diversity patterns. **a**, Median sequence length per study after quality trimming. Original EMP studies used 90-bp reads, which were replaced by 100-bp reads for the majority of studies, and have since been replaced by 150–151-bp reads. For most analyses presented in this manuscript, we used the Deblur algorithm and trimmed tag sequences to 90 bp. This allowed inclusion of older studies with shorter read lengths. **b**, Comparison of Greengenes and SILVA rRNA databases for reference-based OTU picking. Fraction of reads in $n = 23,828$ biologically independent samples—separated by environment (per-environment n shown in Fig. 1a)—mapping to Greengenes 13.8 and SILVA 123 with closed-reference OTU picking. Boxplots show median, IQR, and 1.5×IQR (with outliers). The fraction of reads mapping was similar between Greengenes and SILVA in each environment but slightly higher with SILVA for every environment. **c**, Alpha-diversity in closed-reference OTUs picked against Greengenes 13.8 and SILVA 123, with sequences rarefied to 100k, 30k, 10k, and 1k sequences per sample, displayed as boxplots showing median, IQR, and 1.5×IQR (with outliers). The sample set for all calculations contained $n = 4,667$ biologically independent samples having at least 100k observations in both Greengenes and SILVA OTU tables. Alpha-diversity metrics were higher with SILVA closed-reference OTU picking than with Greengenes. **d**, Beta-diversity among all EMP samples using principal coordinates analysis (PCoA) of weighted UniFrac distance. Principal coordinates PC1 vs. PC2 and PC1 vs. PC3 are shown coloured by EMPO levels 2 and 3. As with unweighted UniFrac distance (Fig. 2c), clustering of samples using weighted UniFrac distance could be explained largely by environment.



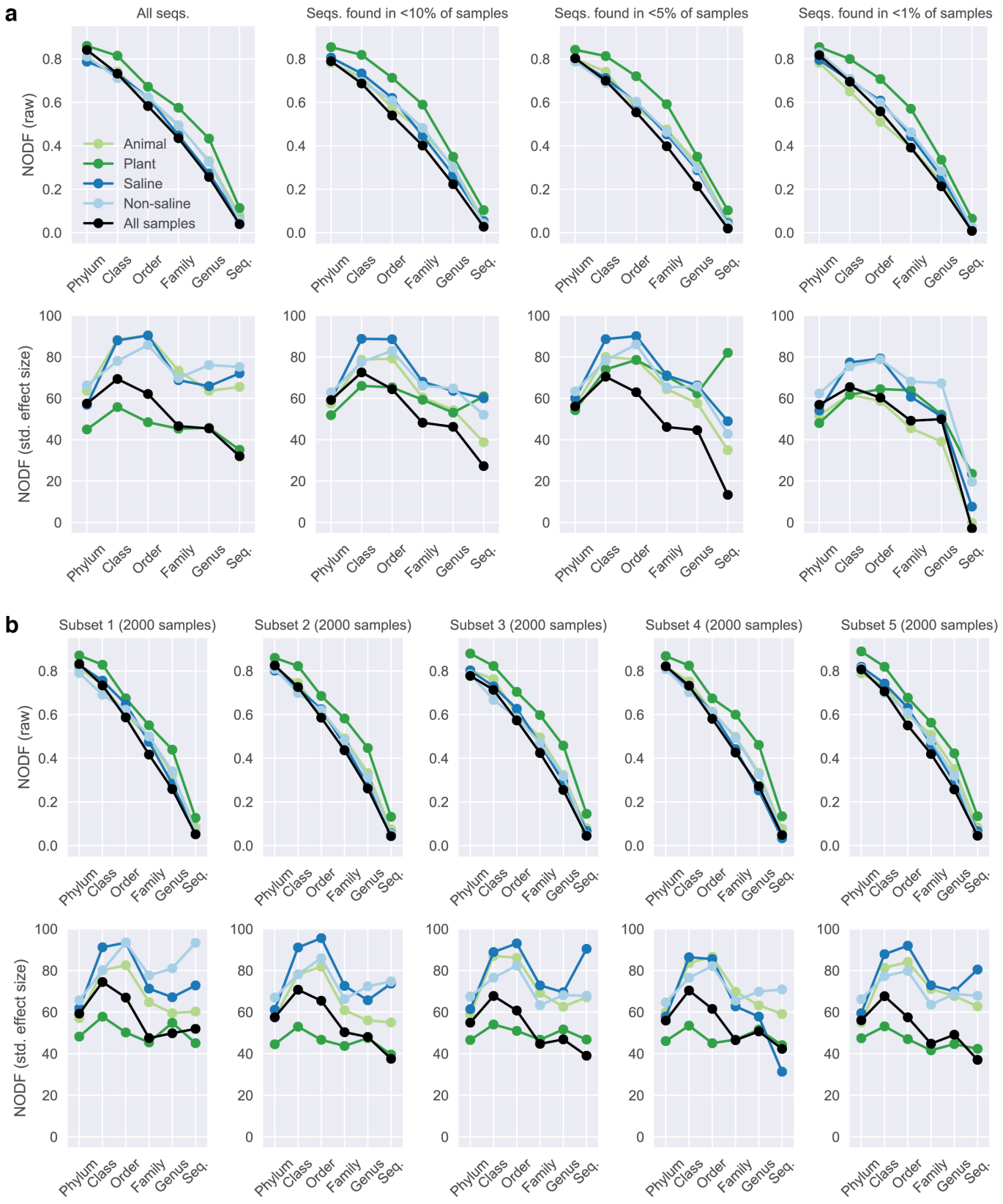
Extended Data Figure 3 | Sequence length effects on observed diversity patterns. The effect of trimming from 150 bp (the approximate starting length of some sequences) to 90 bp (the trimmed length of all sequences in this meta-analysis) was investigated by comparing alpha- and beta-diversity patterns. All samples, at each sequence length, were rarefied to 5,000 sequences per sample. **a**, Alpha-diversity distributions of $n = 12,538$ biologically independent samples displayed as histograms of observed tag sequences coloured by environment (EMPO level 3). Among these samples with sequence length $\geq$150 bp, the distributions are largely preserved when trimming from 150 to 100 to 90 bp. **b**, Procrustes goodness-of-fit between the 90-bp (grey lines) and 150-bp (black lines) Deblur principal coordinates (unweighted UniFrac distance) for $n = 200$ randomly chosen samples coloured by environment (EMPO level 2). Beta-diversity patterns between the two sequence lengths are similar.



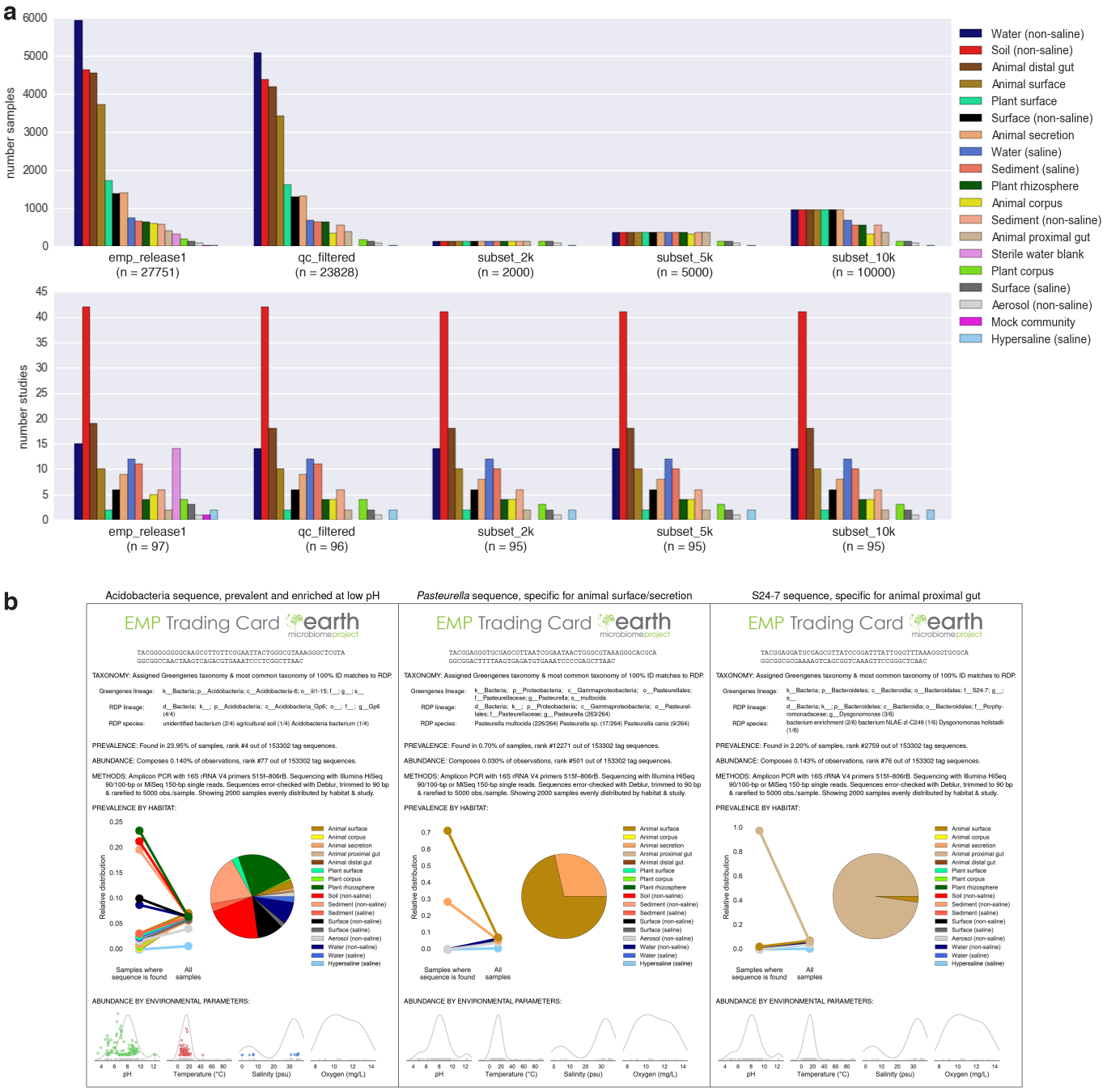
Extended Data Figure 4 | Tag sequence prevalence patterns. Note that for this meta-analysis, the input observation table was filtered to keep only tag sequences with at least 25 observations total over all samples and then rarefied to 5,000 observations per sample. **a**, Per-study endemism visualized as a histogram of tag sequences binned by the number of studies in which they are observed (left: linear scale; right: log scale). **b**, Per-sample endemism visualized as a histogram of tag sequences binned by the number of samples they are observed in (left: sample counts up to 92 samples and the number of tag sequences in linear scale; right: all tag sequences with bin widths of 100 samples and number of tag sequences in log scale). **c**, Abundance (total observations in rarefied table) versus prevalence (number of samples observed in) of $n = 307,572$ tag sequences. Both axes are log scale. The most prevalent tag sequences were also the most abundant. **d**, Prevalence as a function of sequencing depth in $n = 2,279$ soil, $n = 478$ saltwater, $n = 1,508$ freshwater, and $n = 695$ animal distal gut samples having at least 50,000 sequences per sample. Shown is the average and standard deviation of mean prevalence across triplicate rarefied subsamples of 50, 100, 500, 1,000, 5,000, 10,000, and 50,000 sequences per sample. Average prevalence increases with sequencing depth, and the straight-line relationship on the log–log axis is suggestive of a power law. **e**, Histograms of tag sequence prevalences at each sampling depth. The histograms show the distribution moving towards higher prevalences with increasing sequencing depth. Gut data lacked tag sequence prevalences >0.7 due to the inclusion of very different host species; see **f**. **f**, Histograms as in **e** but on a subset of the observation tables where 30 samples were randomly sampled from each study. Restricting to human gut samples only, the full range of prevalences found in the other environments is observed.



Extended Data Figure 5 | Environmental effect sizes, sample classification, and correlation patterns. **a**, Effect sizes of predictors on alpha- and beta-diversity. Maximum pairwise effect size (difference between means divided by standard deviation) between categories of each predictor plotted for observed tag sequences (alpha-diversity) and unweighted and weighted UniFrac distance (beta-diversity). Response variables (alpha- and beta-diversity) were derived from the QC-filtered subset of the 90-bp Deblur table containing $n = 23,828$ biologically independent samples. Numeric predictor variables were converted to quartiles (categorical predictors). Categories within each predictor had a minimum of 75 samples per category. **b**, Cumulative variation explained by the optimal model of stepwise redundancy analysis (RDA) of predictors: study ID, EMPO level 3, ENVO biome level 3, latitude, and longitude (predictors with values for less than half of samples, including host scientific name, were excluded). Environment (EMPO level 3) and biome (ENVO biome level 3) explained as much variation as study ID when study ID was excluded from the RDA. **c**, Confusion matrix for random forest classifier of samples to environment (EMPO level 3). The classifier was trained on the 2,000-sample subset, which was then tested on the remaining samples (QC-filtered samples minus 2,000-sample subset). Squares are coloured relative to 100 classification attempts for each true label. Overall success rate was 84%, with the most commonly misclassified sample environments being Surface (non-saline), Animal secretion, Soil (non-saline), and Aerosol (non-saline). **d**, Receiver operating characteristic (ROC) curve for classification of samples to environment (EMPO level 3). The AUC (area under curve) indicates the probability that the classifier will rank a randomly chosen sample of the given class higher than a randomly chosen sample of other classes. **e**, Classification success, using a random forest classifier, to EMPO levels 1–3, ENVO material, ENVO feature, and ENVO biome levels 1–3. **f**, Microbial source tracking: mean predicted proportion of tag sequences from each source environment (EMPO level 3) that occurs in each sink environment. The model was trained on a subset of samples (~20% of each environment), and tested to predict tag sequence source composition in all remaining samples. Aerosol (non-saline), Surface (saline), and Hypersaline samples were not included in this analysis due to insufficient sample numbers. **g**, Microbial source tracking: which other environments a sample type most resembles. The model was trained on all source environments except one using a leave-one-out cross-validated model, and then used to classify each sample included in that group. Hence, the predicted classification proportion of environment X to environment X is zero. **h**, Correlation of microbial richness with latitude. Richness of 16S rRNA tag sequences per sample across EMPO level 2 environmental categories as a function of absolute latitude. Samples from studies that span at least 10° latitude are highlighted in colour, with linear fits displayed per-study as matched coloured lines. Samples from studies with more narrow latitudinal origin are shown in grey. The global fit for all samples per category is indicated by a dashed black line.



Extended Data Figure 6 | NODF scores of nestedness across samples by taxonomic level. The NODF statistic (Nestedness metric based on Overlap and Decreasing Fill) represents the mean, across pairs of samples, of the fraction of taxa occurring in less diverse samples that also occur in more diverse samples. A raw NODF of 0.5 would mean that for any pair of samples, on average 50% of the taxa in the less diverse sample would occur in the more diverse sample. **a**, NODF (raw) and NODF standardized effect size in the 2,000-sample subset by taxonomic level. Results are shown first for all tag sequences and then for tag sequences found in <10%, <5%, and <1% of samples. By removing the most prevalent tag sequences prior to analysis (and rarefying only after this step), it was possible to rule out artifacts associated with potential contamination. NODF (raw) is highest at the phylum level and decreases at finer taxonomic levels, and this trend is observed even when the most prevalent tag sequences are removed (removing those occurring in ≥10%, ≥5%, or ≥1% of samples). The decreasing trend is likely partially due to finer taxonomic groups having lower prevalence (and lower matrix fill, among other factors) than coarser taxonomic groups, as standardized effect sizes of the NODF statistic are essentially constant across taxonomic levels. **b**, When five alternate 2,000-sample subsets are randomly drawn (with replacement) from the full (QC-filtered) EMP dataset, the trends in NODF (raw) and NODF standardized effect size remain largely unchanged.



Extended Data Figure 7 | Subsets and EMP Trading Cards. a, Subsets of the EMP dataset with even distribution across samples and studies. Shown are all EMP samples included in this manuscript (Release 1), the QC-filtered subset, and subsets of 10,000, 5,000, and 2,000 samples. The latter three contain progressively more even representation across environments and studies, providing a more representative view of the Earth microbiome and a more lightweight dataset. Top, histograms of samples per environment (EMPO level 3) for each subset. Bottom, histograms of studies per environment (EMPO level 3) for each subset. **b**, EMP Trading Cards: distribution of 16S rRNA tag sequences across the EMP. Trading cards highlight the power of the EMP dataset to help define niche ranges of individual microbial sequence types across the planet's microbial communities. Cards show distribution of 16S rRNA tag sequences in a 2,000-sample subset of the EMP (rarefied to 5,000 observations per sample) having even distribution by environment (EMPO level 3) and study. Taxonomy is from Greengenes 13.8 and Ribosomal Database Project (RDP), with the fraction of exact RDP matches by lineage and species name shown in parentheses. The pie chart and point plot show the relative distribution of environments in which the tag sequence is found (left points) versus the environment distribution of all 2,000 samples (right points). The coloured scatter plots indicate tag sequence relative abundance (normalized to the shared *y*-axis) as a function of metadata values (no points shown indicates that metadata were not provided for that category). For comparison, grey curves with rug plots indicate kernel density estimates of metadata values across all samples in the set of 2,000 (not just samples where the tag sequence was found). Three examples are shown: Left, a prevalent sequence enriched in soil and plant rhizosphere is from the class Acidobacteria, aptly named as this sequence is found at highest relative abundance in low-pH samples. Middle, the sequence most specific for animal surface (also enriched in animal secretion) is annotated as *Pasteurella multocida*, a common cause of zoonotic infections following bites or scratches by domestic animals, such as cats and dogs [83]. Right, the sequence most specific for animal proximal gut belongs to S24-7, a family highly localized to the gastrointestinal tracts of homeothermic animals and predominantly found in herbivores and omnivores, but not in carnivores [84].

Supplementary Tables (next page)

Supplementary Table 1 | Studies within the Earth Microbiome Project included in this meta-analysis. These studies constitute EMP 16S Release 1. Read length is median bp for sequences in that study. DOIs with an asterisk describe that study's samples but did not use EMP sequence data. Sample counts exclude controls.

Study ID	Title	Principal investigator	DOI	EBI accession	Read length (bp)	Number of samples	Number of samples by environment
550	Moving pictures of the human microbiome	Rob Knight	10.1186/gb-2011-12-5-r50	ERP021896	132	1969	Animal distal gut: 467, Animal surface: 993, Animal secretion: 509
632	Canadian MetaMicroBiome Initiative samples	Josh Neufeld	10.4056/sigs.1974654	ERP020023	100	13	Soil (non-saline): 13
638	Protist diversity in a permanently ice-covered Antarctic lake during the polar night transition	Rachael M. Morgan-Kiss	10.1038/ismej.2011.23	ERP020508	100	89	Water (non-saline): 89
659	New Zealand Free Air CO2 Enrichment (FACE) soil samples	Saman Bowatte	10.1016/j.soilbio.2013.03.014*	ERP017166	100	24	Soil (non-saline): 24
662	The role of macrobiota in structuring microbial communities along rocky shores	Catherine Pfister	10.7717/peerj.631	ERP020507	140	47	Animal corpus: 4, Water (saline): 3, Surface (saline): 40
678	Bioturbating shrimp alter the structure and diversity of bacterial communities in coastal marine sediments	Bonnie Laverock	10.1038/ismej.2010.86	ERP017221	143	278	Sediment (saline): 278
713	Diversity of carbonate deposits and basement rocks in continental and marine serpentinite seeps	William Brazelton	10.3389/fmicb.2011.00268	ERP016412	151	51	Sediment (saline): 51
722	Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample (5PRIME)	Rob Knight	10.1073/pnas.1000080107	ERP020884	100	175	Water (non-saline): 35, Animal distal gut: 35, Soil (non-saline): 21, Animal surface: 21, Water (saline): 21, Animal secretion: 21, Sediment (non-saline): 21
723	Catlin Arctic Survey 2010 L3	Helen Findlay	10.3389/fmicb.2014.00490	ERP020022	137	130	Water (saline): 130
755	Replicating the microbial community and water quality performance of full-scale slow sand filters in laboratory-scale filters	Sarah Haig	10.1016/j.watres.2014.05.008	ERP020510	150	1000	Water (non-saline): 77, Soil (non-saline): 923
776	Bioremediation of hydrocarbon-contaminated soils from Maritime Antarctica	Diogo Jurelevicius	–	ERP017438	100	30	Soil (non-saline): 30
804	Biofilms on carbonate chimneys of the Lost City Hydrothermal Field on the Mid-Atlantic Ridge	William Brazelton	10.1073/pnas.0905369107	ERP016395	141	96	Surface (saline): 96
805	Exploring links between pH and bacterial community composition in soils from the Craibstone Experimental Farm	Josh Neufeld	10.1111/1574-6941.12231	ERP020539	100	14	Soil (non-saline): 14
807	Human and environmental impacts on river sediment microbial communities	Jack Gilbert	10.1371/journal.pone.0097435	ERP016468	151	44	Sediment (non-saline): 44
808	NEON: Directions and resources for long-term monitoring in soil microbial ecology	Jacob Parnell	10.1890/ES12-00196.1	ERP020590	100	15	Soil (non-saline): 15
809	Prokaryote populations of extant microbialites along a depth gradient in Pavilion Lake, British Columbia, Canada	Jennifer Biddle	10.1111/gbi.12082	ERP020021	100	21	Surface (non-saline): 21
810	Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific Ocean Margin	Jennifer Biddle	10.1073/pnas.0511033103*	ERP020587	100	7	Sediment (saline): 7
829	Microbial diversity in soil, sand dune and rock substrates of the Thar Monsoon Desert, India	Subramanya Rao	10.1007/s12088-015-0549-1*	ERP020560	100	2	Soil (non-saline): 2
846	Influence of tillage practices on soil microbial diversity and activity in a long-term corn experimental field under continuous maize production	Stefano Mocali	–	ERP020589	100	48	Soil (non-saline): 48
861	Examination of microbial communities through a freshwater/saltwater transition zone in cenotes, Yucatan, Mexico	Anni Moore	–	ERP017176	151	21	Water (non-saline): 7, Water (saline): 12, Sediment (saline): 2
864	Soil bacterial diversity in the semi-arid steppe of Northern Mongolia	Aurora MacRae-Crerar	–	ERP017220	151	230	Soil (non-saline): 230
889	Rees Vulcano Island Mediterranean Sea	Andy Rees	–	ERP017174	151	8	Water (saline): 8
894	Catchment sources of microbes	Robin Gasser	10.1186/s13071-016-1607-1, 10.1016/j.watres.2012.12.027	ERP016405	150	2015	Animal distal gut: 2015
895	Kilauea geothermal soils and biofilms	Gary M. King	–	ERP020591	151	6	Soil (non-saline): 3, Surface (non-saline): 3
905	Anaerobic ammonium-oxidizing bacteria in marine environments: widespread occurrence but low diversity	Stefan Hulth	10.1111/j.1462-2920.2007.01266.x*	ERP021691	151	60	Sediment (saline): 60
910	Biological oxygen demand optode analysis of coral reef-associated microbial communities exposed to algal exudates	Forest Rohwer	10.7717/peerj.107	ERP016416	151	59	Animal corpus: 49, Plant corpus: 10
925	Yellowstone gradients	Greg Caporaso	–	ERP022167	90	464	Water (non-saline): 14, Animal distal gut: 9, Surface (non-saline): 441
933	Latitudinal surveys of algal-associated microorganisms	Torsten Thomas	10.1111/1462-2920.12972	ERP021699	100	339	Plant surface: 339
940	Microbiota of freshwater fish slime and gut from catostomids in Colorado water system	Se Jin Song	–	ERP016495	100	257	Water (non-saline): 3, Animal distal gut: 64, Animal surface: 190
945	Long-term seasonal development in selected lakes of northeast Germany	Hans-Peter Grossart	–	ERP022245	150	1147	Water (non-saline): 1147
958	Saliva from obese individuals suppresses the release of aroma compounds from wine	Danilo Ercolini	10.1371/journal.pone.0085611	ERP016748	151	56	Animal secretion: 56
963	Intergenerational lizard lounges do not explain variation in the gut microbiomes of green iguanas	Beck Wehrle	–	ERP016749	100	100	Animal distal gut: 100
990	Spatial scale drives patterns in soil bacterial diversity	Dionysios Antonopoulos	10.1111/1462-2920.13231	ERP016752	151	708	Soil (non-saline): 708
1001	Understanding cultivar-specificity and soil determinants of the Cannabis microbiome	Suzanne Kennedy	10.1371/journal.pone.0099641	ERP016540	151	27	Soil (non-saline): 18, Plant rhizosphere: 9
1024	The soil microbiome influences grapevine-associated microbiota (MiSeq)	Jack Gilbert	10.1128/mBio.02527-14	ERP006348	151	349	Soil (non-saline): 100, Plant rhizosphere: 79, Plant corpus: 170
1030	Impact of fire on active layer and permafrost microbial communities and metagenomes in an upland Alaskan boreal forest	Janet Jansson	10.1038/ismej.2014.36	ERP016543	100	150	Soil (non-saline): 150
1031	Myrold alder fir	David Myrold	10.1007/s00248-010-9675-9*	ERP016746	100	12	Soil (non-saline): 12
1033	Matagenomic and metaproteomic analysis of hydrocarbon-contaminated Antarctic soils	Diogo Jurelevicius	–	ERP016586	100	15	Soil (non-saline): 15
1034	Distinct microbial communities associated with buried soils in the Siberian tundra	Antje Gittel	10.1038/ismej.2013.219	ERP016735	100	90	Soil (non-saline): 90
1035	The ecological dichotomy of ammonia-oxidizing archaea and bacteria in the hyper-arid soils of the Antarctic Dry Valleys (NZTABS)	Craig Cary	10.3389/fmicb.2014.00515	ERP021864	100	121	Soil (non-saline): 119, Animal corpus: 2
1036	Microbial communities of the deep unfrozen: Do microbes in taliks increase permafrost carbon vulnerability?	Jenni Hultman	10.1038/ismej.2011.163*	ERP016588	100	68	Soil (non-saline): 68
1037	Long-term soil productivity project	Steven Hallam	–	ERP016587	100	24	Soil (non-saline): 24
1038	Myrold Oregon transect	David Myrold	10.1111/j.1758-2229.2011.00290.x*	ERP016539	100	21	Soil (non-saline): 21
1039	Metagenomic analysis of Rio de Janeiro coastline	Diogo Jurelevicius	–	ERP016734	100	25	Water (non-saline): 2, Water (saline): 12, Sediment (saline): 8, Sediment (non-saline): 1, Hypersaline (saline): 2
1041	Great Lake microbiome	Karl J. Rockne	–	ERP016492	151	49	Water (non-saline): 49
1043	Laboratory-directed research and development for biological carbon sequestration	Janet Jansson	–	ERP016592	100	56	Soil (non-saline): 56
1056	Convergence of gut microbiomes in myrmecophagous mammals	Frédéric Delsuc	10.1111/mec.12501	ERP003782	151	93	Animal distal gut: 93
1064	Microbiome of honey bees from Puerto Rico	MG Dominguez-Bello	–	ERP016607	151	391	Animal corpus: 391
1197	Deep sediments following Deepwater Horizon oil spill in Gulf of Mexico	Janet Jansson	10.1038/ismej.2013.254	ERP016581	100	106	Sediment (saline): 106
1198	Polluted polar coastal sediments	Hebe Dionisi	10.1007/s00248-017-1028-5	ERP016557	100	61	Sediment (saline): 61
1222	Ocean acidification shows negligible impacts on high-latitude bacterial community structure in coastal pelagic mesocosms	Jack Gilbert	10.5194/bg-10-555-2013	ERP016464	150	73	Water (saline): 73
1235	Ocean acidification shows negligible impacts on high-latitude bacterial community structure in coastal pelagic mesocosms	Julie LaRoche	10.5194/bg-10-555-2013	ERP016542	100	268	Water (saline): 268
1240	Defining seasonal marine microbial community dynamics	Carol Robinson	10.1038/ismej.2011.107	ERP016541	145	156	Water (saline): 156
1242	A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA	Katherine McMahon	10.1038/ismej.2012.118	ERP016591	100	96	Water (non-saline): 96
1288	Bacterial community spatial and temporal variation in a north temperate bog lake	Katherine McMahon	–	ERP016854	150	1506	Water (non-saline): 1506
1289	Native exotic precipitation study in Temple, TX, USA	Kirsten S. Hofmockel	–	ERP016852	131	65	Soil (non-saline): 65
1453	The gut microbiota distinguishes GI-healthy and -unhealthy captive colobineprimates	Rob Knight	10.1016/j.gecco.2016.06.004	ERP016286	100	318	Animal proximal gut: 318
1481	Whole-grain wheat consumption reduces inflammation in a randomized controlled trial on overweight and obese subjects with unhealthy dietary and lifestyle behaviors: role of polyphenols bound to cereal dietary fiber	Danilo Ercolini	10.3945/ajcn.114.088120	ERP016451	151	96	Animal distal gut: 96
1521	Samples presented at EMP conference June 2011 Shenzhen	Rob Knight	–	ERP023684	90	746	Water (non-saline): 188, Soil (non-saline): 279, Animal secretion: 190, Aerosol (non-saline): 89
1526	Recovery of biological soil crust-like microbial communities in previously submerged soils of Glen Canyon	Greg Caporaso	–	ERP016869	101	95	Soil (non-saline): 95
1578	Changes in microbial communities along redox gradients in polygonized Arctic wet tundra soils	David Lipson	10.1111/1758-2229.12301	ERP010098	100	35	Soil (non-saline): 35
1579	Hawaii Kohala volcanic soils	Eric Dubinsky	–	ERP016879	100	128	Animal distal gut: 1, Soil (non-saline): 127
1580	Halophilic communities as a source for novel lignocellulolytic enzymes	Janet Jansson	–	ERP016883	100	26	Water (saline): 4, Sediment (saline): 11, Hypersaline (saline): 11
1621	Resistance and adaptation to the antibiotic monensin by the anaerobic digestion microbiome	Largus T. Angenent	–	ERP016466	100	184	Animal distal gut: 184
1622	Biodiversity and functional patterns of microbial assemblagesÂ in relation to land-use change in postglacial pond sediment profiles	Alison Berry	–	ERP016496	100	353	Sediment (non-saline): 353
1627	Geographic distance and pH drive bacterial distribution in alkaline lake sediments across Tibetan Plateau	Haiyan Chu	10.1111/j.1462-2920.2012.02799.x	ERP016880	100	18	Sediment (saline): 13, Sediment (non-saline): 5
1632	Bird eggshells from Spain	Juan M. Peralta-Sánchez	10.1111/j.1474-919X.2012.01256.x*	ERP016455	100	608	Animal surface: 608
1642	Microbial community of the bulk soil and rhizosphere of rice plants over its lifecycle	Jose Clemente	–	ERP016900	100	640	Water (non-saline): 10, Soil (non-saline): 320, Plant rhizosphere: 310
1665	Co-diversification of marine mammals and their skin microbiomes	Amy Apprill	10.1371/journal.pone.0090785*	ERP016924	100	161	Animal surface: 161
1673	Mission Bay sediment viromes	Forest Rohwer	–	ERP016923	151	26	Sediment (saline): 26
1674	Urban stress is associated with variation in microbial species composition--but not richness--in Manhattan	Krista McGuire	10.1038/ismej.2015.152	ERP016925	151	152	Soil (non-saline): 152
1692	Friedman Alaska peat soils	Largus T. Angenent	10.3390/min3030318	ERP016927	100	92	Soil (non-saline): 92
1694	Starling eggshells from Spain	Juan M. Peralta-Sánchez	–	ERP016469	100	571	Animal surface: 571
1696	Gut microbiota and health in wild and captive colobine primates	Vanessa Hale	10.1016/j.gecco.2016.06.004,10.1016/j.mimet.2015.03.021	ERP016329	100	160	Animal distal gut: 160
1702	Chu Changbai mountain soil	Haiyan Chu	10.1016/j.soilbio.2012.07.013	ERP016926	100	22	Soil (non-saline): 22
1711	Agricultural intensification and the functional capacity of soil microbes on smallholder African farms (Kakamenga, Kenya)	Krista McGuire	10.1111/1365-2664.12416	ERP021540	151	78	Soil (non-saline): 78
1713	Malaysia Lambir soils	Krista McGuire	–	ERP021541	151	36	Soil (non-saline): 36
1714	Responses of soil fungi to logging and oil palm agriculture in southeast asian tropical forests	Krista McGuire	10.1007/s00248-014-0468-4	ERP021542	151	26	Soil (non-saline): 26
1715	McGuire Nicaragua coffee soil	Krista McGuire	–	ERP021543	151	63	Soil (non-saline): 63
1716	Panama soil precipitation gradient	Krista McGuire	10.1007/s00248-011-9973-x*	ERP021544	151	43	Soil (non-saline): 43
1717	Agricultural intensification and the functional capacity of soil microbes on smallholder African farms (Southwestern Kenya)	Krista McGuire	10.1111/1365-2664.12416	ERP021545	151	56	Soil (non-saline): 56
1721	Comparative analysis of the microbial communities in agricultural soil amended with enhanced biochars or traditional fertilisers	Torsten Thomas	10.1016/j.agee.2014.04.006	ERP016937	138	295	Soil (non-saline): 295
1734	Gut microbiota of phyllostomid bats that span a breadth of diets	Liliana Davalos	–	ERP016131	100	72	Animal distal gut: 72
1736	African buffalo gut microbiome	Vanessa Ezenwa	–	ERP016483	100	642	Animal distal gut: 642
1747	The oral and skin microbiomes of captive Komodo Dragons are significantly shared with their habitat	Rob Knight	10.1128/mSystems.00046-16	ERP016252	151	215	Water (non-saline): 3, Animal distal gut: 49, Soil (non-saline): 4, Animal surface: 64, Animal secretion: 53, Surface (non-saline): 24, Plant corpus: 18
1748	Skin biogeography comparison	Juan M. Peralta-Sánchez	–	ERP022166	100	232	Animal surface: 230, Animal secretion: 2
1773	Characterization of bird gut microbiome - gizzard, upper intestine, lower intestine from birds in Venezuela	Maria Alexandra Garcia-Amado	–	ERP016414	141	124	Animal distal gut: 36, Animal proximal gut: 88
1774	Puerto Rico and Plantanal samples for the western acculturation project	MG Dominguez-Bello	–	ERP016472	100	655	Animal distal gut: 135, Animal surface: 313, Animal secretion: 207
1799	Dominguez sleep-deprived flies	MG Dominguez-Bello	–	ERP023686	151	156	Animal corpus: 156
1883	Microbial diversity in arctic freshwaters is structured by inoculation of microbes from soils	Byron Crump	10.1038/ismej.2012.9	ERP017459	138	3311	Water (non-saline): 2719, Soil (non-saline): 105, Water (saline): 35, Surface (non-saline): 252, Sediment (saline): 47, Sediment (non-saline): 152, Surface (saline): 1
2080	Discerning marine archaeal mixotrophy and heterotrophy in the deep North Atlantic	Lauren Seyler	–	ERP016287	151	55	Water (saline): 55
2182	Gut microbiota and health in wild and captive colobine primates	Vanessa Hale	10.1016/j.gecco.2016.06.004,10.1016/j.mimet.2015.03.021,10.1016/j.mimet.2016.09.017	ERP016285	151	167	Animal distal gut: 167
2192	Longitudinal analysis of microbial interaction between humans and the indoor environment	Jack Gilbert	10.1126/science.1254529	ERP005806	151	1625	Animal surface: 620, Animal secretion: 315, Surface (non-saline): 690
2229	Continental-scale variation in seaweed host-associated bacterial communities is a function of host condition, not geography	Torsten Thomas	10.1111/1462-2920.12972	ERP021895	145	1388	Plant surface: 1388
2300	Gut microbionne of hibernating bears	Rita L. Seger	–	ERP016384	100	88	Animal distal gut: 88
2338	Microbiome of Seba's short-tailed bats, Carollia perspicillata	Susan Whitehead	–	ERP016491	100	157	Animal distal gut: 154, Animal surface: 3
2382	The soil microbiome influences grapevine-associated microbiota (HiSeq)	Jack Gilbert	10.1128/mBio.02527-14	ERP006348	150	401	Soil (non-saline): 64, Plant rhizosphere: 241, Plant corpus: 96

Supplementary Table 2 | Description of metadata fields in the mapping files. Also shown is data type (Python/Pandas), variable type, and number of the 27,751 samples having metadata for each field.

Section	Field	Description	Data type	Variable type	Number of samples
Sample	#SampleID	unique sample identifier (leading hashtag is required by Qiime)	string	n/a	27751
	BarcodeSequence	barcode sequence	string	n/a	27751
	LinkerPrimerSequence	linker primer sequence	string	n/a	27751
	Description	sample description	string	n/a	27726
	host_subject_id	unique host or subject (can have multiple samples per host_subject_id)	string	categorical: nominal	26303
Study	study_id	study identifier	integer	categorical: nominal	27751
	title	study title	string	n/a	27751
	principal_investigator	PI of the study	string	n/a	27751
	doi	Digital Object Identifier (DOI) or PubMed ID of primary publication	string	n/a	27751
	ebi_accession	European Bioinformatics Institute (EBI) European Nucleotide Archive (ENA) accession number if submitted	string	n/a	27751
Preparation	target_gene	name of gene amplified (e.g., 16S rRNA)	string	categorical: nominal	27751
	target_subfragment	name of subfragment of gene amplified (e.g., V4)	string	categorical: nominal	27751
	pcr_primers	amplicon primer sequences used	string	categorical: nominal	27751
	illumina_technology	model of Illumina sequencer	string	categorical: nominal	27751
	extraction_center	where the sample was physically extracted	string	categorical: nominal	27751
	run_center	where the sample was physically sequenced (CCME=U Colorado Boulder, ANL=Argonne Natl Lab, UCSDMI=UC San Diego, CGS-GL=Wash U, UCD=U Colorado Denver)	string	categorical: nominal	27751
	run_date	date the sample was physically sequenced	datetime	categorical: ordinal	27751
Sequences	read_length_bp	median read length in bp across study after quality filtering	integer	numeric: discrete	27751
	sequences_split_libraries	number of sequences after demultiplexing with split_libraries_fastq.py	integer	numeric: discrete	27751
	observations_closed_ref_greengenes	number of observations in closed-reference Greengenes table	integer	numeric: discrete	27745
	observations_closed_ref_silva	number of observations in closed-reference Silva table	integer	numeric: discrete	27745
	observations_open_ref_greengenes	number of observations in open-reference Greengenes table	integer	numeric: discrete	27745
	observations_deblur_90bp	number of observations in 90-bp Deblur table	integer	numeric: discrete	27745
	observations_deblur_100bp	number of observations in 100-bp Deblur table	integer	numeric: discrete	27745
	observations_deblur_150bp	number of observations in 150-bp Deblur table	integer	numeric: discrete	27745
Subsets	emp_release1	samples with >=1 sequences (split libraries) per sample	boolean	categorical: nominal	27751
	qc_filtered	samples with >=1000 observations in CR-GG & CR-Silva & OR-GG & Deblur-90 but excluding controls (all subsets are in this set)	boolean	categorical: nominal	27751
	subset_10k	10000 samples with >=10000 CR-GG & CR-Silva & OR-GG and >=5000 Deblur-90, randomly selected and evenly distributed across `empo_3` categories and then across studies	boolean	categorical: nominal	27751
	subset_5k	5000 samples with >=10000 CR-GG & CR-Silva & OR-GG and >=5000 Deblur-90, randomly selected and evenly distributed across `empo_3` categories and then across studies	boolean	categorical: nominal	27751
	subset_2k	2000 samples with >=10000 CR-GG & CR-Silva & OR-GG and >=5000 Deblur-90, randomly selected and evenly distributed across `empo_3` categories and then across studies	boolean	categorical: nominal	27751
Taxonomy	sample_taxid	sample NCBI taxonomy ID	integer	categorical: nominal	27751
	sample_scientific_name	sample NCBI scientific name looked up from taxonomy ID	string	categorical: nominal	27751
	host_taxid	host NCBI taxonomy ID	integer	categorical: nominal	13111
	host_common_name_provided	host common name provided in the original mapping file	string	categorical: nominal	12892
	host_common_name	host NCBI common name looked up from taxonomy ID	string	categorical: nominal	10121
	host_scientific_name	host NCBI scientific name looked up from taxonomy ID	string	categorical: nominal	13111
	host_superkingdom	host superkingdom looked up from taxonomy ID	string	categorical: nominal	13111
	host_kingdom	host kingdom looked up from taxonomy ID	string	categorical: nominal	13111
	host_phylum	host phylum looked up from taxonomy ID	string	categorical: nominal	13111
	host_class	host class looked up from taxonomy ID	string	categorical: nominal	13111
	host_order	host order looked up from taxonomy ID	string	categorical: nominal	13111
	host_family	host family looked up from taxonomy ID	string	categorical: nominal	13111
	host_genus	host genus looked up from taxonomy ID	string	categorical: nominal	13111
	host_species	host species looked up from taxonomy ID	string	categorical: nominal	13111
Geography	collection_timestamp	date and time when sample was collected	datetime	categorical: ordinal	26129
	country	country where sample was collected	string	categorical: nominal	27751
	latitude_deg	latitude in degrees	float	numeric: continuous	27738
	longitude_deg	longitude in degrees	float	numeric: continuous	27738
	depth_m	depth in meters of sample below surface (earth surface if soil, sea/lake bottom if sediment, lake surface if lake, sea level if marine)	float	numeric: continuous	26147
	altitude_m	height above surface, usually zero unless the mouse is levitating (either depth or altitude can have a non-zero value but not both)	float	numeric: continuous	27659
	elevation_m	height above sea level in meters (from georeferencing tool)	float	numeric: continuous	27685
Ontology	env_biome	ENVO biome	string	categorical: nominal	27751
	env_feature	ENVO feature	string	categorical: nominal	27751
	env_material	ENVO material	string	categorical: nominal	27751
	envo_biome_0	level 0 ENVO biome looked up from given ENVO biome	string	categorical: nominal	27751
	envo_biome_1	level 1 ENVO biome looked up from given ENVO biome	string	categorical: nominal	27751
	envo_biome_2	level 2 ENVO biome looked up from given ENVO biome	string	categorical: nominal	27219
	envo_biome_3	level 3 ENVO biome looked up from given ENVO biome	string	categorical: nominal	21520
	envo_biome_4	level 4 ENVO biome looked up from given ENVO biome	string	categorical: nominal	16619
	envo_biome_5	level 5 ENVO biome looked up from given ENVO biome	string	categorical: nominal	446
	empo_0	level 0 EMPO habitat from EMPO ontology	string	categorical: nominal	27751
	empo_1	level 1 EMPO habitat from EMPO ontology	string	categorical: nominal	27751
	empo_2	level 2 EMPO habitat from EMPO ontology	string	categorical: nominal	27751
	empo_3	level 3 EMPO habitat from EMPO ontology	string	categorical: nominal	27751
Alpha-diversity	adiv_observed_otus	observed tag sequences of 90-bp Deblur table rarefied to 5000 sequences per sample	integer	numeric: discrete	24049
	adiv_chao1	Chao1 index of 90-bp Deblur table rarefied to 5000 sequences per sample	float	numeric: continuous	24049
	adiv_shannon	Shannon index of 90-bp Deblur table rarefied to 5000 sequences per sample	float	numeric: continuous	24049
	adiv_faith_pd	Faith's phylogenetic diversity of 90-bp Deblur table rarefied to 5000 sequences per sample	float	numeric: continuous	24049
Physicochemical	temperature_deg_c	ambient temperature in degrees Celcius	float	numeric: continuous	8080
	ph	pH value	float	numeric: continuous	4847
	salinity_psu	salinity in practical salinity units (PSU)	float	numeric: continuous	688
	oxygen_mg_per_l	oxygen concentration in mg/L	float	numeric: continuous	1344
	phosphate_umol_per_l	phosphate concentration in µmol/L	float	numeric: continuous	1118
	ammonium_umol_per_l	ammonium concentration in µmol/L	float	numeric: continuous	904
	nitrate_umol_per_l	nitrate concentration in µmol/L	float	numeric: continuous	2247
	sulfate_umol_per_l	sulfate concentration in µmol/L	float	numeric: continuous	407

Supplementary Table 3 | Top tag sequences by prevalence, abundance, and specialization for habitat. The most prevalent (number of samples observed in), abundant (number of total times observed), or specific for a habitat (most abundant tag sequence with prevalence >25% of that habitat and Shannon entropy <1) EMP tag sequences. Tag sequences are 90-bp 16S rRNA gene sequences (V4 region, starting after primer 515f) from the Deblur algorithm. Values are for a subset of the EMP consisting of 2,000 samples with even representation across habitats (EMPO level 3) and studies, except for statistics on blanks, which are for all blank samples in the EMP (values greater in blanks than in the 2000-sample subset are coloured red). Samples were rarefied to 5,000 observations per sample. tag sequences are sorted by prevalence. Sequences annotated as chloroplast were excluded before statistics were computed.

	Prevalence		Abundance		Max. sample abund.		Avg. sample abund.			
16S rRNA tag sequence	fraction of samples	rank	fraction of total observations	rank	subset 2k	blanks	subset 2k	blanks	Taxonomy (Greengenes)	Significance
TACGTAGGTGGCAAGCGTTGTCCGGAATTATTGGGCGTAAAGCGC GCGCAGGCGGTCCTTTAAGTCTGATGTGAAAGCCCACGGCTCAAC	0.3070	1	0.00458	14	0.5482	0.0003	0.00416	0.00001	k__Bacteria; p__Firmicutes; c__Bacilli; o__Bacillales; f__Bacillaceae; g__Bacillus; s__foraminis	Top-10 prevalence
TACGAAGGGGGCTAGCGTTGCTCGGAATCACTGGGCGTAAAGGGT GCGTAGGCGGGTCTTTAAGTCAGGGGTGAAATCCTGGAGCTCAAC	0.2845	2	0.00186	46	0.1182	0.0134	0.00169	0.00136	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhizobiales; f__Bradyrhizobiaceae; g__Bradyrhizobium; s__	Top-10 prevalence
TACGTAGGGCGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGC TCGTAGGCGGTTTGTCGCGTCTGCCGTGAAAGTCCGGGGCTCAAC	0.2695	3	0.01196	1	0.8756	0.0165	0.01085	0.00078	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Micrococcaceae; g__Arthrobacter; s__psychrolactophilus	Top-10 prevalence Top-10 abundance
TACGGGGGGGGCAAGCGTTGTTCGGAATTACTGGGCGTAAAGGGC TCGTAGGCGGCCAACTAAGTCAGACGTGAAATCCCTCGGCTTAAC	0.2395	4	0.00140	77	0.0466	0	0.00127	0	k__Bacteria; p__Acidobacteria; c__Acidobacteria-6; o__iii1-15; f__; g__; s__	Top-10 prevalence
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGC ACGCAGGCGGTCTGTCAAGTCGGATGTGAAATCCCCGGGCTCAAC	0.2225	5	0.00859	4	0.9938	0.6116	0.00779	0.03001	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae	Top-10 prevalence Top-10 abundance
TACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTAAAGCGA GCGCAGGCGGTTAGATAAGTCTGAAGTTAAAGGCTGTGGCTTAAC	0.2185	6	0.01046	2	0.7970	0.2527	0.00948	0.01366	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus	Top-10 prevalence Top-10 abundance
TACAGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGC GCGTAGGTGGTTTGTTAAGTTGGATGTGAAATCCCCGGGCTCAAC	0.2085	7	0.00703	7	0.9964	0.0298	0.00638	0.00172	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas	Top-10 prevalence Top-10 abundance
TACGAAGGGGGCTAGCGTTGCTCGGAATCACTGGGCGTAAAGCGC ACGTAGGCGGACTCTTAAGTCGGTGGTGAAATCCTGGAGCTCAAC	0.2065	8	0.00050	284	0.0172	0.0005	0.00045	0.00002	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhizobiales; f__Hyphomicrobiaceae; g__Rhodoplanes; s__	Top-10 prevalence
TACAGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGC GCGTAGGTGGTTCGTTAAGTTGGATGTGAAATCCCCGGGCTCAAC	0.2025	9	0.00523	11	0.9638	0.7941	0.00474	0.16989	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas; s__viridiflava	Top-10 prevalence
TACGTAGGTGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGCGC GCGTAGGCGGTTTTTTAAGTCTGATGTGAAAGCCCACGGCTCAAC	0.2010	10	0.00838	5	0.9184	0.1169	0.00760	0.00850	k__Bacteria; p__Firmicutes; c__Bacilli; o__Bacillales; f__Staphylococcaceae	Top-10 prevalence Top-10 abundance
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGC ACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAAC	0.1815	14	0.00798	6	0.9934	0.0158	0.00723	0.00134	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae	Top-10 abundance
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGC GCGTAGGTGGTTTGATAAGTTAGATGTGAAAGCCCTGGGCTTAAC	0.1010	118	0.00611	9	0.4228	0	0.00554	0	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria	Top-10 abundance
TACGGAGGGGGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGC ACGTAGGCGGCTTTGTAAGTCAGAGGTGAAAGCCTGGAGCTCAAC	0.0830	206	0.00964	3	0.9016	0.0001	0.00874	0.000003	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Sphingomonadales; f__Sphingomonadaceae; g__Sphingomonas; s__azotifigens	Top-10 abundance
TACGGAGGGTGCGAGCGTTAATCGGAATGACTGGGCGTAAAGGGC ATGTAGGCGGATAATTAAGTTAGGTGTGAAAGCCCTGGGCTCAAC	0.0410	936	0.00703	8	0.9782	0	0.00637	0	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__; g__; s__	Top-10 abundance, Specific for animal corpus
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGT GCGTAGGCGGTTTATTAAGTCAGATGTGAAAGCCCCGGGCTTAAC	0.0350	1286	0.00089	140	0.2076	0	0.00081	0	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria	Specific for saline water
TACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGA GCGCAGACGGTTAATTAAGTCAGATGTGAAATCCCCGAGCTCAAC	0.0350	1302	0.00594	10	0.6638	0	0.00538	0	k__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Neisseriales; f__Neisseriaceae; g__Kingella; s__	Top-10 abundance
TACAGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGC GCGTAGGCGGCTTTGCAAGTCGGGTGTGAAATCCCCAAGCTCAAC	0.0345	1330	0.00198	43	0.2636	0	0.00179	0	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Xanthomonadales; f__Sinobacteraceae; g__Steroidobacter; s__	Specific for plant rhizosphere
TACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGCGTAAAGCGA GCGCAGGCGGAAAAATAAGTCTAATGTGAAAGCCCTCGGCTTAAC	0.0320	1511	0.00035	412	0.0948	0.0060	0.00032	0.00031	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus; s__	Specific for non-saline aerosol
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGT GCGCAGGCGGCGCGAAAAGTCAGCGGTCAAAGTTCCGGGCTCAAC	0.0220	2759	0.00143	76	0.2026	0	0.00130	0	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__S24-7; g__; s__	Specific for animal proximal gut
TACGGAGGGGGTTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGT ACGTAGGCGGATTGGAAAGTAGGGGGTGAAATCCCGGGGCTCAAC	0.0205	3129	0.00061	211	0.2082	0	0.00055	0	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhodobacterales; f__Rhodobacteraceae; g__Loktanella; s__	Specific for plant surface
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGC ACGCAGGCGGTTTGTTAAGCTAGATGTGAAAGCCCCGCGCTCAAC	0.0160	4378	0.00065	202	0.2704	0	0.00059	0	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Alteromonadales; f__Alteromonadaceae; g__Glaciecola; s__	Specific for saline surface
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGC GCGTAGGCGGCCTGATAAGTTGGATGTGAAAGCCCTGGGCTCAAC	0.0120	6236	0.00102	113	0.3154	0.0001	0.00093	0.000005	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Alteromonadales; f__Alteromonadaceae; g__HB2-32-21; s__	Specific for saline sediment
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGA GCGTAGGCGGGATATTAAGTCAGTCGTCAAAGCCTGTGGCTCAAC	0.0115	6560	0.00034	430	0.0800	0.0187	0.00031	0.00175	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__; s__	Specific for animal distal gut
GACAGAGGAGGCGAGCGTTATCCGGAGTCATTGGGCGTAAAGGGT ACGTAGGCGGCTAGTCAAGTCTGAGGTTAAAGACCACAGCTTAAC	0.0110	7235	0.00176	55	0.4336	0	0.00159	0	k__Bacteria; p__Cyanobacteria; c__Gloeobacterophycideae; o__Gloeobacterales; f__Gloeobacteraceae; g__Gloeobacter; s__	Specific for non-saline surface
GACGTAGGATCCAAGCGTTATCCGGAATTACTGGGCGTAAAGGGC GTGTAGGAGGTTGGGCAAGTCGGCCATGAAAGCTCCCGGCTCAAC	0.0100	7840	0.00117	92	0.2698	0	0.00106	0	k__Bacteria; p__Chloroflexi; c__Anaerolineae; o__H39; f__; g__; s__	Specific for non-saline water
TACAGAGGGTGCGAGCGTTGTTCGGAATTATTGGGCGTAAAGGGC GCGTAGGCGGTGTTGTAAGTCACCTGTGAAACCTCTAGGCTTAAC	0.0100	8149	0.00013	1056	0.1906	0	0.00012	0	k__Bacteria; p__Acidobacteria; c__Acidobacteriia; o__Acidobacteriales; f__Koribacteraceae; g__; s__	Specific for soil
TACAGAGGGTGCAAGCGTTAATCGGATTTACTGGGCGTAAAGCGC GCGTAGGCGGTTAGATAAGTCGGATGTGAAAGCCCCGGGCTTAAC	0.0080	10044	0.00042	337	0.0882	0	0.00039	0	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria	Specific for non-saline sediment
TACGGAGGGTGCGAGCGTTAATCGGAATAACTGGGCGTAAAGGGC ACGCAGGCGGACTTTTAAGTGAGATGTGAAATCCCCGAGCTTAAC	0.0070	12271	0.00030	501	0.5018	0	0.00027	0	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae; g__Pasteurella; s__multocida	Specific for animal surface and animal secretion