A communal catalogue reveals Earth’s multiscale microbial diversity

Luke R. Thompson1,2,3, Jon G. Sanders1, Daniel McDonald1, Amnon Amir1, Joshua Ladu4, Kenneth J. Locey5, Robert J. Prill6, Anupriya Tripathi1,7,8, Sean M. Gibbons9,10, Gail Ackermann1, Jose A. Navas–Molina1,11, Stefan Janssen1, Evgenia Kopylova1, Yoshiki Vázquez–Baeza1,11, Antonio González1, James T. Morton1,11, Slavash Mitrab1, Zhenjiang Zech Xu1, Lingjing Jiang1,13, Mohamed F. Haroon1, William Kanbar1, Qiyun Zhu1, Se Jin Song1, Tomasz Koscielak1, Nicholas A. Bokulich1,12, Joshua Lefser1, Col J. Brislawn1, Gregory Humphrey1, Sarah M. Owens1, Jarrad Hampton–Marcell1,17, Donna Berg–Lyons1,19, Valerie McKenzie1,2, Noah Fierer3,20, Jed A. Fuhrman21, Aaron Clauset19,23, Rick L. Stevens24,25, Ashley Shade26,27,28, Katherine S. Pollard1, Kelly D. Goodwin2, Janet K. Jansson1,29, Jack A. Gilbert1,11,30 & The Earth Microbiome Project Consortium*

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 an 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 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.1–6 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 of 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 have suggested robust general trends in community composition, including the importance of salinity1 and animal association2. 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 that a scalable and accessible analytical framework is needed.

The Earth Microbiome Project (EMP, http://www.earthmicrobiome.org) was founded in 2010 to sample the Earth’s microbial communities at an unprecedented scale in order to advance our understanding of the organizing biogeographic principles that govern microbial community structure2,3. We recognized that open and collaborative science, including scientific crowdsourcing and standardized methods4, would help to reduce technical variation among individual studies, which can overwhelm biological variation and make general trends difficult to detect3. Comprising around 100 studies, over half of which have yielded peer-reviewed publications (Supplementary Table 1), the EMP has now dwarfed by 100-fold the sampling and sequencing depth of earlier meta-analysis efforts1,2; 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 for the exploration of myriad questions and a starting point for the guided acquisition of new data to answer them. As an example of using this

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
tool, we present a 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. We also used a lightweight 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. We indexed the classes in this application ontology (Fig. 1a) for future research.

In situ

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 http://www.earthmicrobiome.org/protocols-and-standards/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).

We surveyed bacterial and archaeal diversity using amplicon sequencing of the 16S rRNA gene, a common taxonomic marker for bacteria and archaea that remains a valuable tool for microbial ecology despite the introduction of whole-genome methods (for example, shotgun metagenomics) that capture gene-level functional diversity. 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 shown to recover sequences from most bacterial taxa and many archaea. We note that these primers may miss newly discovered phyla with alternative ribosomal gene structures, and subsequent modifications not used here have shown improved efficiency with certain clades of Alphaproteobacteria and Archaea. We generated sequence reads of 90–151 base pairs (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. While this approach was useful for certain analyses, for many sample types, especially plant-associated and free-living communities, one-third of reads or more could not be mapped to existing rRNA databases (Extended Data Fig. 2b). We therefore used a reference-free method, Deblur, to remove suspected error sequences and provide single-nucleotide resolution ‘sub-OTUs’, also known as ‘amplicon sequence variants’, 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. Because exact sequences are stable identifiers, unlike OTUs, they can be compared to any 16S rRNA or genomic database now and in the future, thereby promoting reusability.

**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, 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, more than 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, although 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 investigate 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 more than 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 positive, 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). Although contamination remains an issue in microbiome studies, 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 and 45% of SILVA 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 swathes 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% identity OTUs (Extended Data Fig. 2b) supports this observation.

**Patterns of diversity reflect environment**

We used a structured categorical variable of microbial environments, EMPO, to analyse diversity in the EMP catalogue in the context of lessons from previous investigations. 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 previous predicted genome properties (ecological strategy). EMPO, to analyse diversity in the EMP catalogue in the context of environmental surveys. These matches accounted for 10% and 13% 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, although 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 investigate 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 more than 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 positive, 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). Although contamination remains an issue in microbiome studies, 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 and 45% of SILVA 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 swathes 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% identity OTUs (Extended Data Fig. 2b) supports this observation.

**Patterns of diversity reflect environment**

We used a structured categorical variable of microbial environments, EMPO, to analyse diversity in the EMP catalogue in the context of lessons from previous investigations. 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 previous findings that host association is a fundamental environmental factor that differentiates microbial communities (Fig. 2c, Extended Data Fig. 2d). We build on this pattern by showing that there is less richness in host-associated communities than in free-living communities (Fig. 2a), with the noted exception of plant rhizosphere samples, which resemble free-living soil communities in both richness (Fig. 2a) and composition (Fig. 2c). Our findings also confirm the major compositional distinction between saline and non-saline communities (Fig. 2c). The 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; these applications will require large curated datasets, such as the EMP. Supervised machine learning demonstrated that samples could be distinguished as being 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–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 the 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, such as 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; this method has been used, for example, to link the taxonomic groups associated with copiotrophic and oligotrophic behaviours in soils to high and low rRNA gene copy numbers, respectively. Approximately half the dataset centred on an ACN of 2.2 (free-living and plant-associated samples) and the other half on an ACN of 3.4 (animal-associated samples) (Fig. 2d). Greater per-genome rRNA operon copy number has been found to be associated with rapid maximum growth rates, which may provide a selective advantage when resources are abundant, such as in animals. 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 that have been 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, but there is substantial variation among studies examining latitudinal trends of microbial diversity. The large EMP dataset analysed here reveals a weak but significant trend towards increasing diversity at lower 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. 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 the 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. For example, richness has been shown to increase...
up to neutral pH$^{36}$ and often to decrease above neutral pH$^{3,35}$ in soil communities. Richness has been shown to increase with temperature up to a limit and then to decrease beyond that limit in seawater (maximum at about 19°C$^{33}$) and to increase with temperature in soil (up to at least around 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 (around 7) and relatively cool temperatures (about 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 that maximum microbial richness occurs within a relatively narrow range of intermediate pH and temperature values. These patterns, while robust in the context of the EMP dataset, necessarily reflect only the subset of samples that assemble nested subsets from more complex communities, or by dispersal abilities$^{41}$ and cascading source–sink colonization processes$^{39}$ and the filtering of complex communities over time$^{42}$, 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 taxa with lower prevalence. Notably, finer taxonomic groupings showed less nestedness (Fig. 3c), indicating that the processes that underlie nested patterns of turnover are likely to reflect conserved aspects of microbial biology, and not to result from 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 us to observe and analyse microbial distribution patterns at finer resolution than is 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 that are 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 towards one or a few environments (low Shannon entropy); by 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 investigated how entropy changes as a function of phylogenetic distance. We calculated entropy for each node of the phylogeny and visualized it as a function of maximum tip-to-tip branch length (Fig. 4c). While entropy decreased gradually 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 the 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 can be queried using the Redbiom software, with command-line (https://github.com/biocore/redbiom) and web-based (http://qiita.microbio.me) interfaces to find samples based on sequences, taxa, or sample metadata, and to export selected sample data and metadata (instructions at https://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, additional tag sequences can be added to this dataset from new studies without reprocessing existing samples. Future combinations 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 used as environmental indicators and representing targets for cultivation, genome sequencing, and laboratory study. In addition, these tools and approaches, although developed for bacteria and archaea, could be expanded to all domains of life. 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 used 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 reveal basic biogeographic patterns of microbial ecology, with resolution and scope rivaling data compilations currently available for ‘macrobial’ ecology. 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, conclusions should be viewed with caution and form starting points for future hypothesis-directed investigations. There is a need to span gradients of geography (for example, latitude and elevation) and chemistry (for example, temperature, pH, and salinity) more evenly—assisted by tools for more comprehensive collection and curation of metadata—and to track environments over time. In addition, biotic factors (for example, animals, fungi, plants, viruses, and eukaryotic microbes) not measured in this study have important roles in determining community structure. The scalable framework introduced here can be expanded 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, the EMP provides one example of a logistically feasible standardization framework to maximize interoperability across diverse and independent studies, in particular using stable identifiers (exact sequences) to enable enduring utility of environmental sequence 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 microorganisms on Earth.

Online Content Methods, along with any additional Extemed Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 March; accepted 10 October 2017.
Published online 1 November 2017.

1. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. Proc. Natl Acad. Sci. USA 104, 11436–11440 (2007).
2. Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. & Gordon, J. I. Worlds within worlds: evolution of the vertebrate gut microbiota. Nat. Rev. Microbiol. 6, 776–788 (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., Lermarceau, P. & van der Putten, W. H. Going back to the roots: the microbial ecology of the microbiota. 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. Nucleic Acids Res. 39, 411–420 (2011).
11. Buttergick, P. L., Morrison, N., Smith, B., Mungall, C. J. & Lewis, S. E. The environment ontology: contextualising biological and biomedical entities. J. Biomed. Semantics 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. https://doi.org/10.3389/fmars.2017.00107 (2016).
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 (Suppl. 1), 4516–4522 (2011).
15. Cavender-Bares, J., Avila, N. N., Woyte, T. & Kyripides, 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. Collister, S. B., McMurry, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker gene data analysis. ISME J. http://dx.doi.org/10.1038/ismej.2017.119 (2017).
18. Huse, S. M., Welch, D. M., Morrison, H. G. & Sogin, M. L. Ironing out the wrinkles in microbial ecology through improved GTU clustering. Environ. Microbiol. 12, 1889–1898 (2010).
19. Saltor, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 12, 87 (2014).
20. 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).
21. Quist, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596 (2013).
22. Nemergut, D. R. et al. Decreases in average bacterial community rRNA operon copy number during succession. ISME J. 10, 1147–1156 (2016).
23. Gibbons, S. M. et al. Invasive plants rapidly reshape soil properties in a natural grassland ecosystem. Environ. Microbiol. 18, 397–407 (2016).
24. Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. RNA operon copy number reflects ecological strategies of bacteria. Appl. Environ. Microbiol. 66, 1328–1333 (2000).
25. Jilska, J. & Harris, S. V. How important are the generalities of the latitudinal diversity gradient. Am. Nat. 163, 192–211 (2004).
26. Fuhrman, J. A. et al. A latitudinal diversity gradient in planktonic marine bacteria. Proc. Natl Acad. Sci. USA 105, 7774–7778 (2008).
27. Lauber, C. L. et al. Global marine bacterial diversity peaks at high latitudes in winter. ISME J. 7, 1669–1677 (2013).
28. Milici, M. et al. Low diversity of planktonic bacteria in the tropical ocean. Sci. Rep. 6, 19054 (2016).
29. Zhu, 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).
30. 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, 400093 (2017).
31. Zhou, J. et al. Temperature mediates continental-scale diversity of microbes in forest soils. Nat. Commun. 7, 12083 (2016).
32. 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, 1757–1767 (2017).
33. Carvalho, J. C., Cardoso, P., Borges, P. & Schmera, D. Measuring fractions of beta diversity and their relationships to neddiness: a theoretical and empirical comparison of new approaches. Ökos 122, 825–834 (2013).
34. Sonnenburg, E. et al. Diet-induced reductions in the gut microbiota compound over generations. Nature 529, 212–215 (2016).
35. Atmar, W. & Patterson, B. D. The measure of order and disorder in the distribution of species in fragmented habitat. Oecologia 96, 373–382 (1993).
36. Lomolino, M. V. Investigating causality of nestedness of insular communities: selective immigrations or extinctions? J. Biogeogr. 23, 699–703 (1996).
37. Gaston, K. & Blackburn, T. Pattern and Process in Macroecology (Wiley-Blackwell, 2000).
38. 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).
39. Donovas, M. et al. Assemblage time series reveal biodiversity change but not systematic loss. Science 344, 296–299 (2014).
40. Amano, T., Lamming, J. L. & Sutherland, W. J. Spatial gaps in global biodiversity information and the role of citizen science. Bioscience 66, 350–360 (2016).
41. Ioannidis, J. P. A. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q. 94, 485–514 (2016).
42. Davies, N. et al. The founding charter of the Genomic Observatories Network. Environ. Microbiol. Rep. 3, 2 (2014).
43. Alvisatos, A. P. et al. MICROBIOME. A unified initiative to harness Earth’s microbiomes. Science 350, 507–508 (2015).

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

Acknowledgements We thank J. DeReus for management of information systems; J. Huntley and K. Jepson for management of sequencing facilities; B. Erickson for administrative assistance; J. Lennon for discussions about macroecological theory; S. Peddada for assistance with effect size calculations; P. L. Buttergick, C. Mungall, and D. Sieggele for assistance with ontologies; A. Rose, A.-S. Roy, A. Bearquiver, B. Cohen, C. Tuscher, C. Feh, D. Winters, E Jones, J. Heidelberg, J. Eisen, K. E. Wormack, K. Docherty, K. Keegan, K. Konstantindis, M. Bailey, M. Sullivan, N. Desai, N. Kyripides, N. Pace, P. Baiaji, R. Gallery, R. Mackelprang, R. D’Oro, R. Ley, T. Vogel, T. Chen, and W. Feng. This work was supported by the John Templeton Foundation (grant ID 44000, Convergent Evolution of the Vertebrate Microbiome), the W. M. Keck Foundation (DT011413), Argonne National Laboratory (US Department of Energy contract DE-AC02-06CH11377), the Australian Research Council, the Tula Foundation, the Samuel Lawrence Foundation, and the Extreme Science and Engineering Discovery Environment (XSEDE, project number OAC150043), which is supported by the National Science Foundation grant number ACI-1053975. 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, Borealis Genomics, Illumina, Roche, and Integrated DNA Technologies for in-kind support at various phases of the project.

Author Contributions J.A.G., K.J.K., and R.K. conceived the idea for the project and coordinated the project. R.K. coordinated the taxonomic processing and web-based tools. Nucleic Acids Res. 41, D590–D596 (2013).
**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 (http://www.earthmicrobiome.org/protocols-and-standards/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 and not due to inherent biases in the extraction protocol. To avoid known issues that arise when multiple amplicon strategies are combined, we also standardized PCR primers, amplification strategy, and sequencing. More recent studies not included in this meta-analysis adopted additional primer modifications to allow recovery of key taxa in marine and soil samples. Data reporting standards, including the MIX5 (minimal information about any sequence) metadata standard developed by the Genomic Standards Consortium 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 that 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) were represented either in this ontology or in those with which it interoperates. For ease of access, we gathered the contributions into 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 the ENVO release cycle to maximize interoperability.

Metadata curation was automated using Pandas (http://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 (https://qiita.microbio.me), the BIOM format 14, new ‘OTU picking’ methods Deblur 11 and a subsampled open-reference procedure, a scalability improvement of Fast UniFrac phylogenetic inference software, speed improvements to sequence-insertion tree method SEPP 15, and speed and feature improvements to Emperor ordination visualization software (http://biocore.github.io/emporer).

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 (for example, soil, sediment, faeces) and fractionated bulk samples (for example, sponge coral surface tissue, centrifuged turbid water) were taken using microcentrifuge tubes. Swabs (BD SWURE 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 faecal 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. Metadata were collected in compliance with MIMARKS 15, EBI (https://www.ebi.ac.uk/ena), and Qiita (https://qiita.microbio.me). The ENVO team described in the text and the 13,828 data files. The QC-filtered Deblur 90-bp sequences were error-filtered and trimmed to the length of the shortest sequencing run (90 bp) using the Deblur software; 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 tags 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 13.8.23 and SILVA 123 using SortMeRNA 41, and subsampled open-reference OTU picking was done against Greengenes 13.8. These unfILTERED tables and the filtered and subset tables described below are available at ftp://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 provided BIOM tables (ftp://ftp.microbio.me/emp/release1/otu_tables/ and https://zenodo.org/record/890000), the ‘release1’ 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 ≥1,000 observations in each of four observation tables: closed-reference Greengenes, open-reference Greengenes, and Deblur 90-bp controls (emps_1 == ‘Control’) are excluded. Subsets were then generated which give equal (as possible) representation across studies (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 ≥5,000 observations in the Deblur 90-bp observation table and ≥10,000 observations in each of the closed-reference Greengenes, closed-reference SILVA, and open-reference Greengenes observation tables. Note that Deblur removes approximately one-third to one-half of sequences owing to suspected errors, which is consistent with a sequence length of ~90–150 bp and an average error rate of 0.006 per position.

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 the SILVA 123 database using the open-source sequence search tool VSEARCH 21 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 subset contain 10,000, 5,000, and 50,000 sequences per sample. The 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 (samples with sequence 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. © 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
fewer than 30 samples with > 50,000 sequences were discarded. To investigate how mean tag sequence prevalence changes with increasing sequencing depth across environments, we calculated the average mean tag sequence prevalence across three replicate rarefactions. We plotted the average and standard deviation in mean across mean replicate subsamples over a subsampling gradient (that is, 50, 100, 500, 1,000, 5,000, 10,000 and 50,000 sequences per sample).

Greengenes insertion tree. Deblur tag sequences were inserted into the Greengenes reference tree using SEPP\textsuperscript{55}, which uses a divide-and-conquer technique to place putative phylogenetic placements on a large tree. The SEPP method uses HMMER\textsuperscript{44} internally for aligning each Deblurred sequence to a reference Greengenes alignment (gg.\textsubscript{13.5}\_\textsubscript{s}\_\textsubscript{su}_\textsubscript{align}\_99\_\textsubscript{filtered}\_\textsubscript{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\textsuperscript{49} 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\textsuperscript{52} under the GTRCAT model before the placement. The pipeline used, including the reference trees and alignments can be found at ftp://ftp.microbio.me/emp/release1/otu_info/greengenes_sepp_pipeline, and the bash script is available at https://github.com/biocore/emp/blob/master/code/03-otu-picking-trees/deblur/run_sepp.sh.

Fast UniFrac. Unweighted and weighted UniFrac were computed using the Cythonized\textsuperscript{43} implementation of Fast UniFrac\textsuperscript{44} in scikit-bio\textsuperscript{48}. Fast UniFrac by itself was not scalable for the EMP dataset owing 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 some cases, shrinking N to only the relevant nodes of the phylogeny. This decomposition also allows 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\textsuperscript{69} (https://pypi.python.org/pypi/pyp/low).

Core diversity analyses: alpha- and beta-diversity. Alpha-rarefaction was computed using single_rarefaction.py in QIME 1.1\textsuperscript{11} 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\textsuperscript{70}), chao1 (Chao 1 index\textsuperscript{71}), and faith\_pd (Faith’s phylogenetic diversity\textsuperscript{72}), using the Greengenes insertion tree. Fast UniFrac\textsuperscript{44} was run on the Deblur 90-bp table using the aforementioned approach and the corresponding insertion tree. Principal coordinates were computed using QIME 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, file (metadata) was compiled containing the predictors to be tested: study_id, envo\_biome\_3 (a proxy for biome or environment), empo\_3 (a proxy for sample type) as source samples to train the model, and the remainder as sink samples in the full QC-filtered subset. SourceTracker\textsuperscript{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 (http://github.com/biota/sourcetracker2) to define the degree to which tag sequences are shared among environmental categories from which they came. The R packages caret\textsuperscript{77} and randomForest\textsuperscript{78} were used. Five repeats of tenfold 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 samples in the full QC-filtered subset.

SourceTracker analyses. SourceTracker\textsuperscript{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 (http://github.com/biota/sourcetracker2) to define the degree to which tag sequences are shared among environmental categories from which they came. The R packages caret\textsuperscript{77} and randomForest\textsuperscript{78} were used. Five repeats of tenfold 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 samples in the full QC-filtered subset.

Nestoredness of taxonomic composition. Nestoredness captures the degree to which elements of a large set are contained within progressively smaller sets. We used the NODF statistic\textsuperscript{68} to quantify nestoredness of the sample-by-taxa matrix. The rows of this matrix correspond to specific taxon grouped at particular taxonomic levels (for example, phylum, class, etc.), while the columns correspond to particular environments. A row with the maximum nestoredness was selected and used 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 before feature selection and testing.
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)\(^1\). 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 (\(\geq 2\)); 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

\[
W_E(B) = \frac{\{s \in S: B \in s \land s \in E\}}{\{s \in S: B \in s\}}
\]

(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, for example, Firmicutes), we then calculate cluster distribution vector as

\[
W(T) = (W_E(T), \ldots, W_E(T))
\]

(2)

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

\[
W_E(T) = \text{mean}(W_E(B))
\]

(3)

Clusters of tag sequences were defined in two 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. To calculate entropy of environment distribution as a function of taxonomic level (for example, 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 sample taxonomic group). To calculate 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 versus taxonomy analysis) or the phylogenetic tip placement (for the entropy versus phylogeny analysis). 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 (that is, 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 tagging 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; the taxonomy of the tag sequence from Greengenes; the list of all the samples in which the tag sequence is found. This summary is located at http://ftp.microbio.me/emp/emp/release1/otu.distributions/. Additionally, for each tag sequence with a trading card in Extended Data Fig. 7b or http://www.earthmicrobiome.org/trading-cards, we identified sequences in RDP (http://rdp.cme.msu.edu)\(^2\) matching 100% along the 90-bp region of the 16S rRNA gene. Trading cards at http://www.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 \(\geq 1,000\).

**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 Qita (which includes all of the EMP studies). This service allows users to find samples on the basis of sample details (for example, all soil samples with \(pH < 7\)), to find samples on the basis of features they contain (for example, all samples in which Greengenes ID 131337 exists), to find features on the basis of taxonomy (for example, all samples in which genus Pyrococcus 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 https://pypi.python.org/pypi/redbiom; examples of command-line queries are provided at https://github.com/biocore/redbiom. A graphical user interface for Redbiom is available at https://qiita.microbio.me.

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

**Data availability.** Per-study sequence files and sample metadata are available from EBI (http://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 Qita (https://qiita.microbio.me) using the 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://ftp.microbio.me/emp/release1, then across the Redbiom database at time of publication are archived at https://zenodo.org with DOI 10.5281/zenodo.890000.
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, 1–26 (2008).
78. Liaw, A. & Wiener, M. Classification and regression by randomForest. R News 2–3, 18–22 (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 (2014).
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 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; the number (n) of samples having data for each factor is provided at the top of each histogram. Samples are coloured by environment, and only QC-filtered samples are included. In sample metadata files, 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 (97% identity OTUs) with closed-reference OTU picking. Boxplots show median, IQR, and $1.5 \times$ 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 100,000, 30,000, 10,000, and 1,000 sequences per sample, displayed as boxplots showing median, IQR, and $1.5 \times$ IQR (with outliers). The sample set for all calculations contained $n=4,667$ biologically independent samples having at least 100,000 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 (PCA) of weighted UniFrac distance. Principal coordinates PC1 versus PC2 and PC1 versus 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 (right: linear scale; left: log scale). b, Per-sample endemism visualized as a histogram of tag sequences binned by the number of samples in which they are observed (right: sample counts up to 92 samples; left: 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 are the average and s.d. 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. The histograms show the distribution moving towards higher prevalences with increasing sequencing depth. Gut data lacked tag sequence prevalences \( > 0.7 \) owing 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 | See next page for caption.
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 because there were 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 matching coloured lines. Samples from studies with narrower latitudinal origins 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 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 before analysis (and rarefying only after this step), it was possible to rule out artefacts 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 to be 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.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
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. 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.
Experimental design

1. Sample size
   Describe how sample size was determined.

   Two subsets of the EMP dataset were used for analyses presented in this paper. For analysis of total diversity across the dataset (alpha- and beta-diversity in Figs. 2-3), we used the full set of 24,910 samples that passed minimal quality controls (QC-filtered) as described in the methods. For Figs. 4-6 and supplementary figures as noted, we used a 2000-sample subset containing samples picked randomly and evenly across 17 habitats and then evenly across studies in each sample type.

2. Data exclusions
   Describe any data exclusions.

   To generate the QC-filtered subset, samples were removed if they had fewer than a predetermined number of observations in the OTU/Deblur tables (see methods). Study no. 1799 was excluded from the QC-filtered subset because of concerns about contamination. For the effect size calculation (ED Fig. 5), categories within each predictor had a minimum of 75 samples per category, and predictors with values for less than half of samples were excluded. For ED Table 3, sequences annotated as chloroplast were excluded before statistics were computed.

3. Replication
   Describe whether the experimental findings were reliably reproduced.

   The experimental findings were reliably reproduced. For the purposes of this meta-analysis, having multiple samples from multiple studies for each habitat type constituted replication. Many studies within the meta-analysis had dedicated biological replicates. Nestedness results were reproduced using 5 additional randomly-selected 2000-sample subsets.

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.

   For creating subsets of samples, samples were drawn randomly, evenly across habitat types and studies. Results were reproduced using 5 additional randomly-selected 2000-sample subsets.

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Investigators were blinded; allocation to groups (subsets) was done entirely computationally and randomly.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.
6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

| n/a | Confirmed |
|-----|-----------|
| ☑   | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) |
| ☑   | A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☑   | A statement indicating how many times each experiment was replicated |
| ☑   | The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) |
| ☑   | A description of any assumptions or corrections, such as an adjustment for multiple comparisons |
| ☑   | The test results (e.g., P values) given as exact values whenever possible and with confidence intervals noted |
| ☑   | A clear description of statistics including central tendency (e.g., median, mean) and variation (e.g., standard deviation, interquartile range) |
| ☑   | Clearly defined error bars |

See the web collection on statistics for biologists for further resources and guidance.

7. Software

Describe the software used to analyze the data in this study.

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.XXXXX. Redbiom code is available at github.com/biocore/redbiom and is archived at zenodo.org with DOI 10.5281/zenodo.XXXXX. (Zenodo DOIs will be provided in proof stage, as discussed with the editor.)

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g., GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e., assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used.
Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

Animal subjects are described in the original studies where animal-associated samples were collected. IACUC protocol numbers can be provided if necessary.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Human subjects are described in the original studies where human-associated samples were collected. IRB protocol numbers can be provided if necessary.