scientific data

OPEN

ANALYSIS

Trait biases in microbial reference genomes

Sage Albright[1] & Stilianos Louca[1,2] ✉

Common culturing techniques and priorities bias our discovery towards specific traits that may not be representative of microbial diversity in nature. So far, these biases have not been systematically examined. To address this gap, here we use 116,884 publicly available metagenome-assembled genomes (MAGs, completeness $\geq$80%) from 203 surveys worldwide as a culture-independent sample of bacterial and archaeal diversity, and compare these MAGs to the popular RefSeq genome database, which heavily relies on cultures. We compare the distribution of 12,454 KEGG gene orthologs (used as trait proxies) in the MAGs and RefSeq genomes, while controlling for environment type (ocean, soil, lake, bioreactor, human, and other animals). Using statistical modeling, we then determine the conditional probabilities that a species is represented in RefSeq depending on its genetic repertoire. We find that the majority of examined genes are significantly biased for or against in RefSeq. Our systematic estimates of gene prevalences across bacteria and archaea in nature and gene-specific biases in reference genomes constitutes a resource for addressing these issues in the future.

Introduction

Culturing remains the golden standard for studying bacterial and archaeal (henceforth "prokaryotic" for brevity) physiology, metabolism and pathogenicity[1], and for obtaining high-quality genome sequences, such as those in the NCBI RefSeq reference sequence database[2]. The thousands of prokaryotic genomes now available in RefSeq, in turn, enable large-scale analyses that yield insight into the processes shaping microbial genome structure and evolution[3–9], and also form the starting pool for curated gene ontologies or databases such as eggNOG[10] and rrnDB[11]. Reference genome databases and culture-based phenotype databases also enable predictions of the likely gene contents and traits of other less studied prokaryotic clades seen in environmental samples based on phylogenetic relationships, e.g., using tools such as PICRUSt[12], Tax4Fun[13] and FAPROTAX[14]. However, to date the vast majority of extant prokaryotic diversity remains uncultured and lacking a whole genome sequence, partly due to the difficulties associated with determining the proper growth conditions for each species. Conventional culturing techniques, the fact that pure cultures must grow in the absence of syntrophic partners, and typical research priorities are thought to bias our discovery towards traits that may not be representative of the broader prokaryotic diversity in nature[1,15]. These biases can distort our vision of prokaryotic diversity, limit our capacity to discover new useful biochemical functions[15], introduce biases in comparative phylogenetic and other evolutionary analyses[16], and most likely bias phylogeny-based predictions of gene content and traits in uncultured organisms[16]. For example, trait biases in RefSeq are expected to cause corresponding prediction biases in PICRUSt and Tax4Fun. Systematically quantifying the true distribution of traits across prokaryotic clades and determining trait biases in reference genome databases (and by extension, in prokaryotic cultures) is required for assessing the extent of these important issues and addressing them in the future. To date such an analysis across a broad range of clades and environments is lacking, one reason being that it was until recently impossible to efficiently recover a large culture-independent set of microbial genomes from natural environments.

Recent advances in genome-resolved metagenomics now enable the recovery of nearly-complete prokaryotic genomes from complex natural microbial communities without the need for culturing[17–20]. Here we use 116,884 previously published prokaryotic metagenome-assembled genomes (MAGs) from around the world to obtain a culture-independent sample of extant prokaryotic diversity in nature. We use this collection of MAGs to estimate the true prevalences of thousands of different genes (considered here as proxies of traits) across prokaryotic clades. We compare these gene prevalences to those in the widely used NCBI RefSeq prokaryotic reference genome database[2], and quantify gene-dependent biases of clades represented in RefSeq. The RefSeq genome

[1]Department of Biology, University of Oregon, Eugene, USA. [2]Institute of Ecology and Evolution, University of Oregon, Eugene, USA. ✉e-mail: louca.research@gmail.com

database was chosen for comparison as it is one of the oldest, most comprehensive and most widely used reference genome databases, and because it is largely based on cultured organisms, although we acknowledge that some cultured organisms have not yet had their genome sequenced and accessioned in RefSeq. We use statistical models to examine to what extent specific genes influence the probability of a random MAG being represented in RefSeq to at least 95% average nucleotide identity (ANI), which is a common modern measure for delineating prokaryotic species[6,21–24]. To account for obvious environmental preferences in both culturing and metagenomic sequencing efforts, we perform our analyses separately for different environment types, including the human microbiome, the microbiomes of non-human animals (henceforth "animals" for brevity), bioreactors, the ocean, soil, and lakes.

Results

A diverse collection of MAGs. Our collection of 116,884 prokaryotic MAGs was obtained from 203 distinct studies, covering the human microbiome (20 studies), other animals (30), bioreactors (including wastewater treatment plants, 35), the ocean (including estuaries and coastal lagoons, 72), soils (23) and lakes (28) (overview in Supplemental Table S2), and covers over 150 different phyla (Supplemental Fig. S2). All MAGs were estimated to be at least 80% complete and exhibit no more than 5% contamination, based on a set of universal single-copy marker genes (details in Methods section, overview in Supplemental Fig. S1). To avoid redundancies in species representation, we clustered MAGs into species genome bins (SGBs) based on an average nucleotide identity (ANI) threshold of 95%[6,21–24]. The probability that a randomly chosen prokaryotic species is represented in RefSeq (henceforth "coverage") was determined based on the fraction of MAG-SGB representatives that could be matched to at least one RefSeq genome at ANI $\geq$95%.

Overall, we found that only a small fraction of MAG-SGBs could be matched to a RefSeq genome, although strong differences existed between environments. Notably, by far the highest coverage was found for human-associated prokaryotes (33%) and the lowest coverages were found for lakes (2.2%) and soil (4.9%) (overview in Supplemental Table S2). This is consistent with a previous study that showed that the cultured fraction of prokaryotes associated with the human gut is substantially above the average across environments[25], and confirms the general expectation that only a small fraction of non-human-associated prokaryotic clades has been cultured. Our coverage estimates are also comparable to the global coverage estimated previously by Zhang *et al.*[26] based on 16S SSU rRNA amplicon sequences (~2.1%).

Gene prevalence estimates. To estimate how the prevalence of various genes differs between MAGs and prokaryotic RefSeq genomes, we searched for KEGG gene orthologs (KO's) in MAGs as well as in RefSeq genomes using Hidden Markov Models (HMMs) from the KOfam database[27] (annotation summaries in Supplemental Table S1). We chose to focus on KEGG because (a) it is widely used in microbial ecology, (b) it provides ready-to-use HMMs for more accurate gene annotation than BLAST-based searches, and (c) its functional focus facilitates the interpretation of the gene-specific biases examined here. In order to avoid Eukaryote-specific genes, only genes found in at least one MAG or prokaryotic RefSeq genome were considered (12,454 genes). To eliminate redundancies in species representation in RefSeq, we clustered RefSeq genomes into species bins based on their provided species-level taxon-IDs (STIBs), and only considered a single representative per STIB. Note that we focus on the true prevalence of each gene in the original populations represented by the MAG-SGBs or RefSeq STIBs (henceforth denoted α), and not merely the detection rate of that gene in the MAG-SGBs or RefSeq STIBs; indeed, mere detection rates are generally lower than true prevalences due to the incompleteness of many MAGs and some RefSeq genomes. To account for the incompleteness of each MAG and RefSeq genome in our gene prevalence estimates, we used an appropriate probabilistic model that we fitted via maximum-likelihood (details in Methods).

Overall, we found that gene detection rates were strongly skewed towards the lower end regardless of environment, with the majority of genes detected in fewer than 5% of MAG-SGBs and RefSeq STIBs (histograms in Supplemental Fig. S3 and Supplemental Fig. S4). As expected, estimated gene prevalences in MAG-SGBs (i.e., accounting for MAG incompleteness) were generally greater than mere gene detection rates (Supplemental Fig. S6), although gene prevalences still exhibited a strong skew towards lower values regardless of environment (Supplemental Fig. S5). MAG-SGB-based gene prevalences exhibited substantial and clearly significant positive correlations between environments, i.e., a gene that was widespread in species from one environment also tended to be widespread in species from other environments (Pearson $r \geq 0.840$ and $P < 0.001$ for all environment pairs, Supplemental Fig. S7). That said, the degree to which gene prevalences correlated between environments varied considerably. For example, the strongest correlation was found between lakes and bioreactors ($r = 0.991$), between humans and other animals ($r = 0.990$) and between lakes and ocean ($r = 0.985$), while the weakest correlations were found between soil and humans ($r = 0.840$) and between soil and other animals ($r = 0.846$). These results suggest that the selective forces determining the prevalences of various genes across species tend to be somewhat similar across environments, and tend to be particularly similar between the human microbiome are other animal microbiomes, and between lakes, bioreactors and the ocean.

Estimated gene prevalences in MAG-SGBs correlated positively with those in RefSeq STIBs, with Pearson correlation coefficients (r) ranging between 0.89 and 0.95 depending on the environment ($P < 0.001$ in all cases, Fig. 1). While these correlations may seem high, we point out that they are generally similar to the correlations observed between environments (discussed in the previous paragraph), in other words differences between MAG-SGBs and RefSeq STIBs (when controlling for environment) are comparable to differences between environments. This compromises the ecological conclusions that one may draw based on gene prevalences in reference databases such as RefSeq. In fact, the vast majority of genes displayed significantly different prevalences between MAG-SGBs and RefSeq STIBs, with a general tendency for gene prevalences to be higher among RefSeq STIBs than among MAG-SGBs. Specifically, the median ratio between MAG-SGB-based prevalences and RefSeq STIB-based prevalences ("median prevalence ratio", or MPR) was substantially above 1 in all environments,

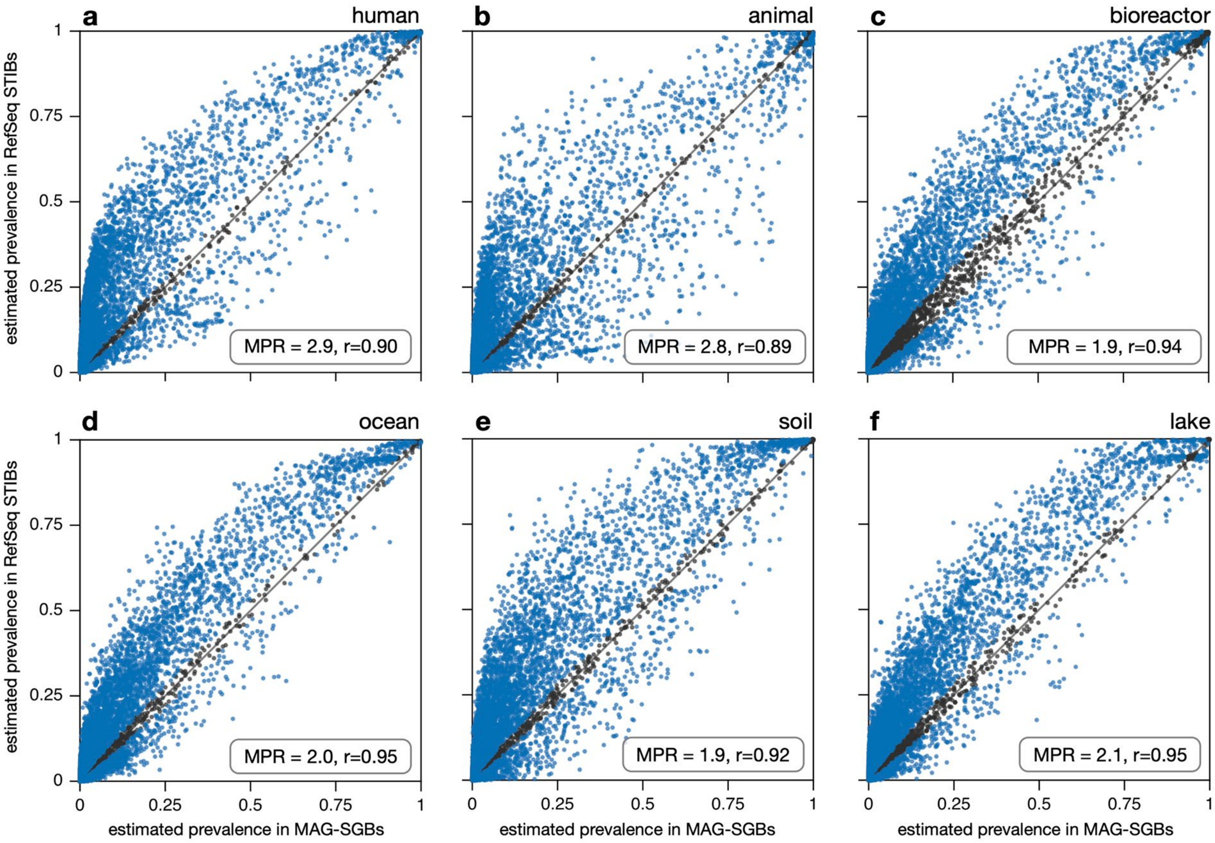

Fig. 1 Gene prevalences (MAG-SGBs vs RefSeq-STIBs). Estimated gene prevalences in MAG-SGBs (horizontal axes) compared to RefSeq STIBs (vertical axes), separately for SGBs/STIBs associated with (**a**) humans, (**b**) other animals, (**c**) bioreactors, (**d**) ocean, (**e**) soil and (**f**) lakes. Every dot represents a distinct gene (KEGG Ortholog). Gene prevalences refer to the populations represented by the MAGs and STIBs, i.e., correcting for genome incompleteness. Blue dots denote genes whose estimated prevalence is significantly different in SGBs compared to STIBs (i.e., the 95% confidence intervals of the two estimates do not overlap), while black dots denote genes whose prevalence is not significantly different. The diagonal is shown for reference. The median prevalence ratio (MPR, prevalence in STIBs divided by the prevalence in SGBs, median taken across genes) and the Pearson correlation coefficient (*r*) are shown in each plot; all correlations were highly significant based on a permutation test ($P < 0.001$). For similar plots showing genome-size adjusted gene prevalences in STIBs see Supplemental Fig. S8. Abbreviations: MAG, metagenome-assembled genome; SGB, species genome bin; STIB, species taxon-ID bin; ANI, average nucleotide identity.

ranging from 1.9 for bioreactors and soil up to 2.9 for humans and 2.8 for non-human animals. This tendency of genes to be more prevalent in RefSeq STIBs than MAG-SGBs could be caused by three distinct but not mutually exclusive mechanisms: First, there may be a general bias in RefSeq towards larger genomes. Indeed, prokaryotes with larger genomes tend to be more metabolically versatile and thus likely less dependent on syntrophic partners, which facilitates their culturing[28–30]. Consistent with this interpretation, we found that genome sizes among RefSeq STIBs tended to be larger on average than genome sizes among MAG-SGBs in all considered environments, even after correcting for MAG incompleteness (Fig. 2). This result confirms and extends a previous finding that uncultured human gut bacteria tend to have significantly smaller genomes when compared to cultured ones[31]. To further examine to what extent size biases in RefSeq can explain the generally higher gene prevalences therein, we adjusted the gene prevalence estimates in RefSeq STIBs for the differences in the size distribution between MAG-SGBs and RefSeq STIBs. The adjustment applied was analogous to those commonly done in stratified demographic surveys with disproportionate sampling of strata, where stratum averages are weighted by each stratum's proportion in the overall population in order to obtain an unbiased population average[32] (see Methods for details). We found that this adjustment indeed reduced the discrepancy in gene prevalences between MAG-SGBs and RefSeq STIBs for all environments, with Pearson correlations increasing slightly in all cases and MPRs decreasing substantially towards a value of 1 (Supplemental Fig. S8). Nevertheless, in all environments the MPR remained greater than 1, with the greatest MPR (1.7) found for animal-associated and lake prokaryotes and the smallest MPR (1.3) found for human-associated and ocean prokaryotes. This suggests that size biases partly — but not fully — explain the generally greater gene prevalences estimated for RefSeq STIBs compared to MAG-SGBs.

Second, it is in principle possible that our gene detection approach was less effective in MAGs than RefSeq genomes, for example due to the generally lower quality of MAGs, rather than there being true gene prevalence differences between the two datasets. Further, difficulties still exist in fully reconstructing genomes from

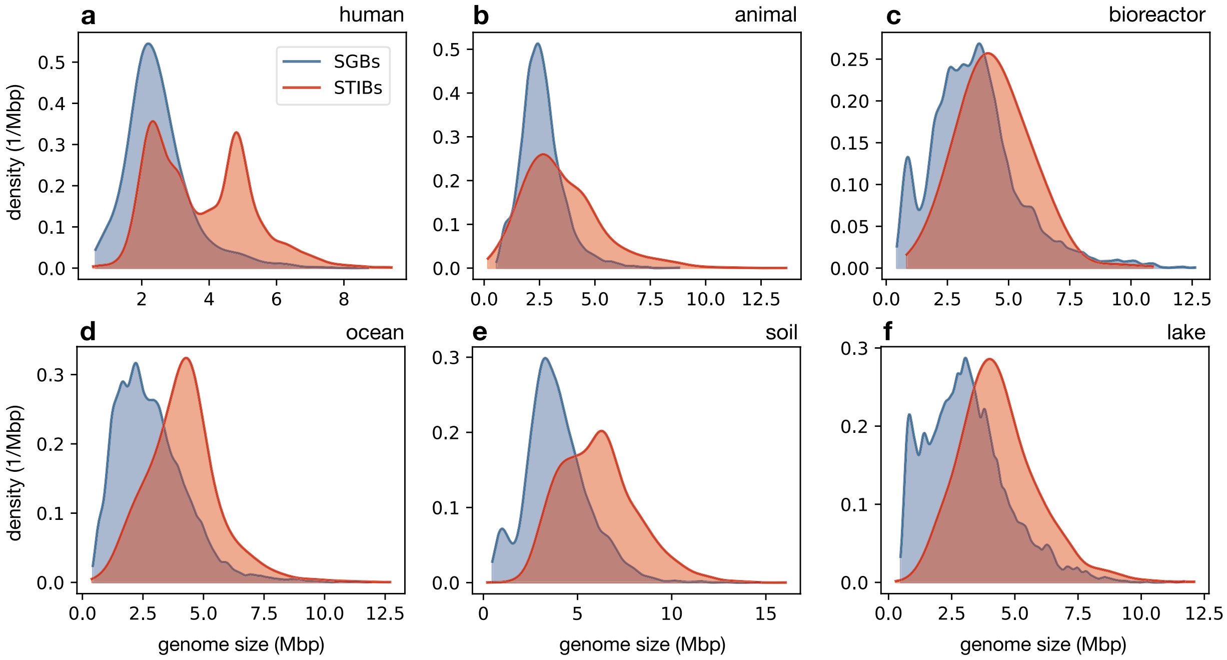

Fig. 2 Genome sizes (MAG-SGBs vs RefSeq-STIBs). Kernel density estimates (KDE) of the genome size distribution among MAG-SGBs (blue curves, correcting for incompleteness) and RefSeq STIBs (red curves), separately for each environment. In all cases, the mean and median genome size among MAG-SGBs was substantially lower than among RefSeq STIBs. For mean genome sizes and KDE bandwidths see Supplemental Table S3.

metagenomes, notably regarding the inclusion of plasmids and genomic islands[33]. These issues could in principle also cause an under-representation of genes in MAGs, compared to genomes from cultures. To examine this possibility, we repeated our gene prevalence estimates for the subset of MAG-SGBs that could be matched to a RefSeq genome, and for the subset of RefSeq genomes matched by a MAG-SGB. By restricting our gene prevalence estimates to these subsets of MAG-SGBs and RefSeq genomes, we eliminated any major differences in species representation between the two datasets, thus focusing on potential differences in gene inclusion/detection efficacy. For these restricted datasets we found a much closer agreement between gene prevalences in MAG-SGBs and RefSeq genomes, with nearly none of the genes exhibiting a statistically significant difference in prevalence and with MPRs being nearly identical to 1 for all environments ($0.97 \leq \mathrm{MPR} \leq 1.03$, Supplemental Fig. S9). We thus conclude that difficulties in including plasmids and genomic islands in MAGs, as well as differences in gene detection efficacy between MAGs and RefSeq genomes, are negligible and, in particular, are not a major source of the gene prevalence differences seen between the full datasets.

Third, RefSeq STIBs may be truly biased towards clades that exhibit the genes examined, to an extent beyond that caused by mere genome size biases. Indeed, KEGG is a highly curated, experimentally informed and functionally focused gene database, and it is possible that genes represented in KEGG tend to be of particular industrial, medical or environmental interests; fewer than half of protein-coding genes predicted in MAGs could be assigned to a KEGG ortholog (overview in Supplemental Table S1). These same interests presumably also guide the majority of culturing and whole sequencing efforts. Similarities between gene characterization biases and culturing/genome sequencing biases will inevitably lead to a tendency for RefSeq STIBs to be rich in genes catalogued in KEGG, even when controlling for genome sizes; we henceforth refer to this mechanism as "intentional" biases. Further, mainstream culturing approaches undoubtedly cause additional unintended trait biases, by favoring fast growers or generalists and disfavoring organisms that depend on syntrophic partners to survive. Genes responsible for (or at least correlating with) traits favored by typical culturing approaches will tend to be more prevalent in cultured species than among prokaryotes in general, and will thus be more likely to be characterized and included in databases such as KEGG. In other words, KEGG could be "unintentionally" biased towards genes that correlate with traits that facilitate culturing. That said, we mention that some genes exhibited lower prevalences among RefSeq STIBs than among MAG-SGBs, suggesting that common culturing approaches may also bias against some genes.

To tease apart intentional from unintentional gene prevalence biases in KEGG, we repeated our analyses with an annotation-independent gene database, the evolutionary gene genealogy of non-supervised orthologous groups (eggNOGs[34]). If biases in KEGG are mostly intentional (as defined above), then one would expect eggNOGs to display a much weaker over-prevalence in RefSeq STIBs than KEGG orthologs do (after adjusting for genome size distributions). In contrast, if biases in KEGG are mostly unintentional (as defined above), then one would expect eggNOGs to display a similar over-prevalence in RefSeq-STIBs as KEGG orthologs do. We found that, when adjusting for genome size distributions, eggNOG MPRs were substantially smaller than KEGG MPRs in 3 out of 6 environments (human, other animals and soil), dropping as low as 0.97 for human-associated

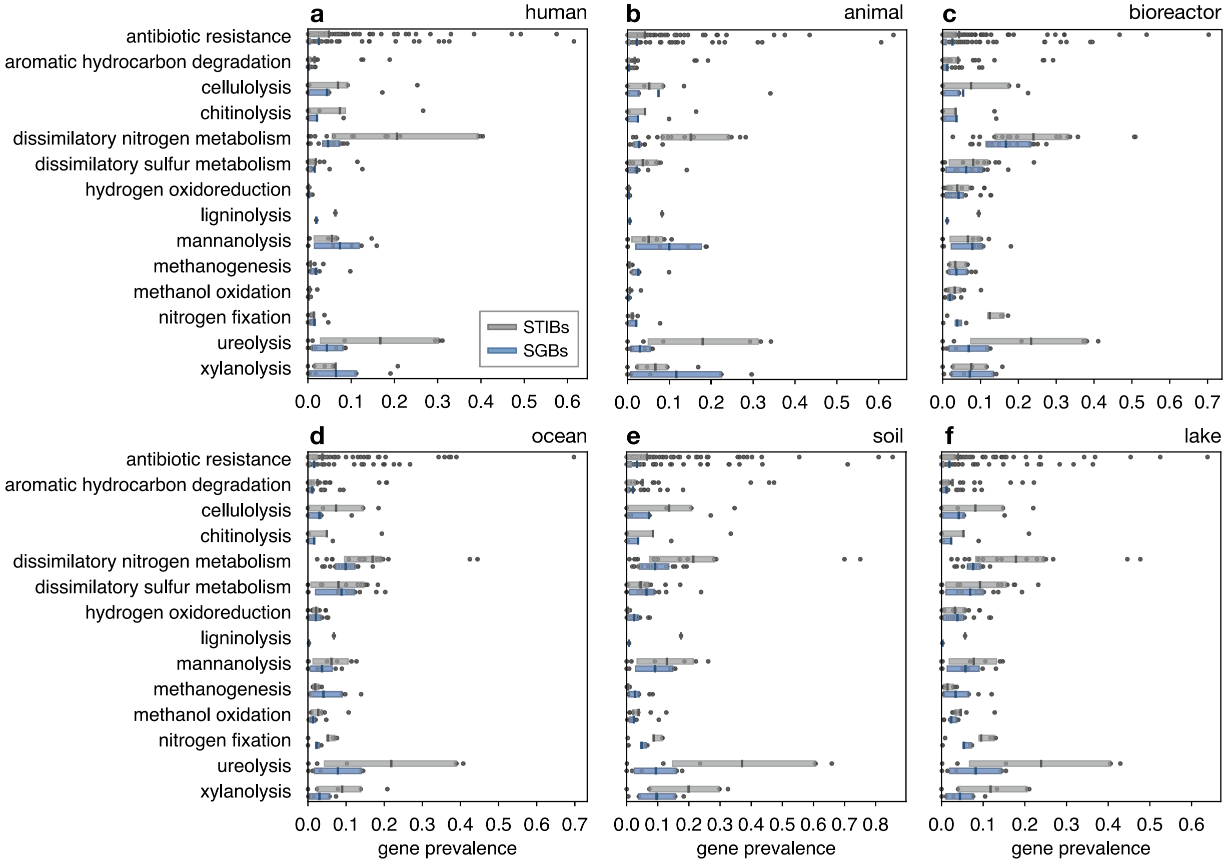

Fig. 3 Gene prevalences for selected metabolic functions (MAG-SGBs vs RefSeq STIBs). Box-plots of estimated gene prevalences based on MAG-SGBs and RefSeq STIBs, for selected functions of particular ecological, industrial or medical interest, separately in each environment. Gene prevalences refer to the populations represented by the MAGs and STIBs, i.e., correcting for genome incompleteness. Each box represents a specific function, each point represents a single gene, vertical bar segments denote mean prevalences (i.e., averaged over all genes associated with a specific function), and boxes span the 2nd and 3rd quartile. In most cases gene prevalences are higher among RefSeq STIBs compared to MAG-SGBs, a notable exception being methanogenesis. For similar figures showing gene prevalences in RefSeq STIBs adjusted for the distribution of genome sizes see Supplemental Fig. S12.

species and 1.1 for other animals (Supplemental Fig. S11). This suggests that in these environments the biases in KEGG are to a great extent intentional. In contrast, for bioreactors, ocean and lakes, eggNOG MPRs were greater than KEGG MPRs, suggesting that in these environments the biases in KEGG are largely unintentional and driven by current culturing abilities. That said, further research is needed to better understand the roles of intentional and unintentional biases in the composition of KEGG and other gene databases.

To further illustrate the discovered gene prevalence biases from a functional perspective, we examined genes involved in metabolic functions of particular industrial interest, such as lignin, mannan, xylan and cellulose degradation[35–37], genes involved in functions of particular environmental interest, such as dissimilatory nitrogen and sulfur metabolisms and methanogenesis, as well as genes conferring antibiotic resistance (Fig. 3). For almost all of these functions and regardless of environment, the mean gene prevalences in RefSeq STIBs were considerably higher than in MAG-SGBs. One notable exception was methanogenesis, which was substantially underrepresented in RefSeq STIBs for all environments except bioreactors (where the difference was only minor). This later observation is consistent with the fact that methanogens have been historically difficult to culture[38–40], and suggests than methanogenesis is a much more common trait in prokaryotes in natural environments than one would expect based on reference genomes.

Gene-dependent coverage biases. To more precisely quantify the biases for or against various genes in RefSeq, we estimated the conditional probability that a prokaryotic species is covered by (i.e., represented in) RefSeq given that it either has or does not have a given gene. We henceforth denote these two conditional probabilities by q_1 and q_0, respectively. If a gene is neither biased for nor against, we expect $q_0 = q_1$, while a bias for or against organisms exhibiting the gene would imply $q_1 > q_0$ or $q_1 < q_0$, respectively. We estimated q_1 and q_0 separately for each gene by fitting a probabilistic model that accounts for MAG incompleteness, gene presence/absence across MAG-SGBs and matches between MAG-SGBs and RefSeq genomes at ≥95% ANI. To facilitate comparisons between genes and between environments, we considered a composite variable that we termed

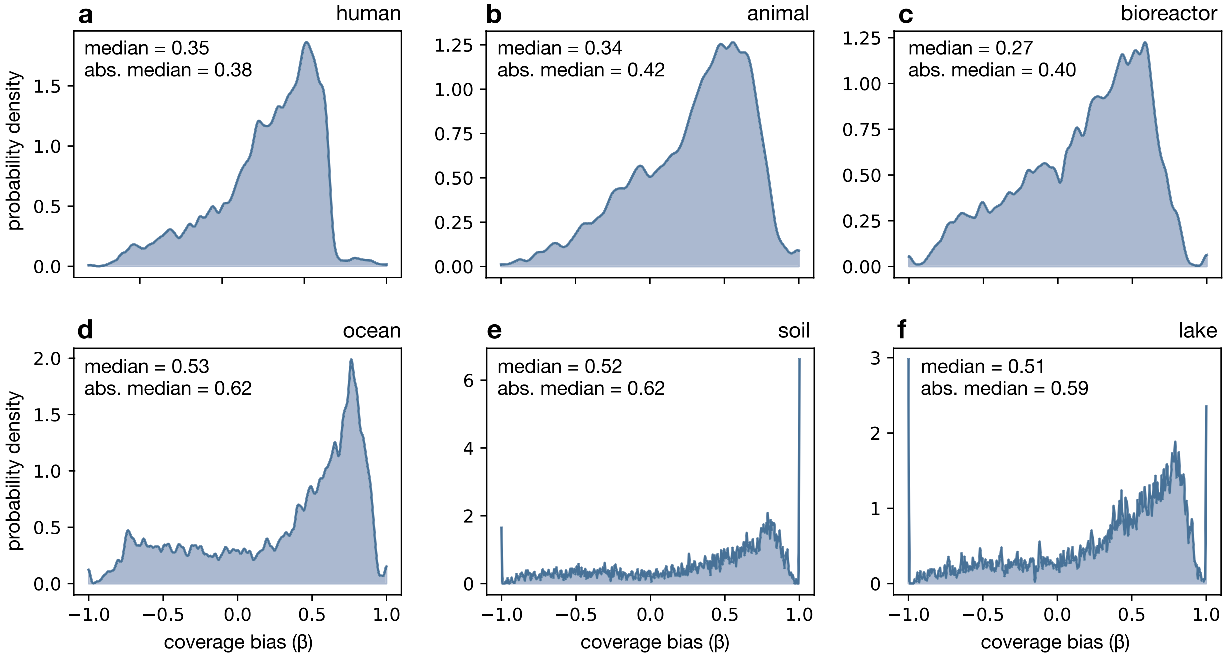

Fig. 4 Distribution of gene-specific coverage biases. (**a**) Distribution of gene-specific coverage biases (β) for human-associated MAG-SGBs, i.e., biases in the probability of matching a RefSeq genome at ≥95% ANI conditioned on the organism having or lacking a specific gene (KEGG ortholog). For any given gene, a positive bias implies that the probability of an SGB matching a RefSeq genome is greater when the gene is present and smaller when the gene is absent (and vice versa for negative biases). In particular, a bias of +1 implies that the probability of matching a RefSeq genome is zero when the gene is absent. Note that gene presence/absence refers to the population represented by a MAG-SGB, i.e., correcting for MAG incompleteness. The distribution was computed using a kernel density estimate over all considered genes. (**b**–**f**) Similar to (**a**), but for alternative environments. The median bias β as well as the median absolute bias (median $|\beta|$), are written in each figure. Observe that in each environment the majority of coverage biases are positive. A summary of coverage biases for each environment, including kernel density bandwidths, is given in Supplemental Table S4.

"coverage bias", denoted β and computed using the estimated q_0, q_1 (see Eq. 4 in Methods for details). The coverage bias is always between −1 and 1, with negative values implying a bias against a gene, positive values implying a bias towards a gene and zero implying no bias ($q_0 = q_1$). A useful property of β is that it only depends on the ratio q_1/q_0 but not on the overall coverage of MAG-SGBs in RefSeq nor on a gene's prevalence (α). We mention that β could not be reliably estimated for all genes, since some genes were too rare ($\alpha \approx 0$) or too prevalent ($\alpha \approx 1$) for estimating the conditional probabilities q_1 or q_0, respectively.

We found that in all environments a substantial fraction (50–82%) of considered genes exhibited a statistically significant coverage bias ($\beta \neq 0$, $P < 0.05$), with the clear majority of significant coverage biases being positive (Fig. 4 and Supplemental Table S4). Accordingly, the median coverage bias across all genes was positive for all environments (0.27–0.53). This pattern is consistent with our previous observation that gene prevalences tend to be inflated in RefSeq STIBs. In fact, for almost all environments (except human) the distribution of coverage biases exhibited a clear spike at a value of $\beta = 1$ (i.e., $q_0 = 0$). For example, among soil-associated prokaryotes 133 out of 4335 considered genes (e.g., *carA*, *gpsA*, *rimP*, *gidB*) exhibited a coverage bias of 1 (details in File `KOfam_gene_prevalences_and_biases.tsv.gz` on Figshare[41]). Hence, the absence of these genes strongly reduces the probability of a species being represented in RefSeq. That said, we point out that we also found many genes with a significant negative coverage bias (e.g., 14% of genes for soil), which means that species exhibiting these genes are underrepresented in RefSeq. When measured in terms of the median absolute coverage bias (median $|\beta|$), we observed the strongest coverage biases for the ocean and soil (median $|\beta| = 0.62$) and the weakest coverage biases for humans (0.38) and bioreactors (0.40). This suggests that existing culturing and genome sequencing pipelines impose weaker trait biases for human- and bioreactor-associated taxa than for ocean- and soil-associated taxa. This observation is perhaps not surprising, given the generally stronger medical and industrial relevance of the first two groups, which probably increases the motivation to culture a broader spectrum of organisms. We also found that coverage biases varied strongly between environments, with the smallest consistency (in terms of the Pearson correlation) seen between humans and soil ($r = 0.41$) and between humans and the ocean ($r = 0.44$, Supplemental Fig. S15).

Strong differences were also observed between different gene categories, as defined by the KEGG KO hierarchy (Fig. 5 and Supplemental Fig. S15). Genes associated with membrane transport, cell motility and nucleotide metabolism generally exhibited the highest median $|\beta|$, although the precise order depended on the environment considered (also note that only a subset of particularly interesting gene categories was included in this comparison). For ocean and soil, which exhibited the highest median $|\beta|$, the two considered gene categories

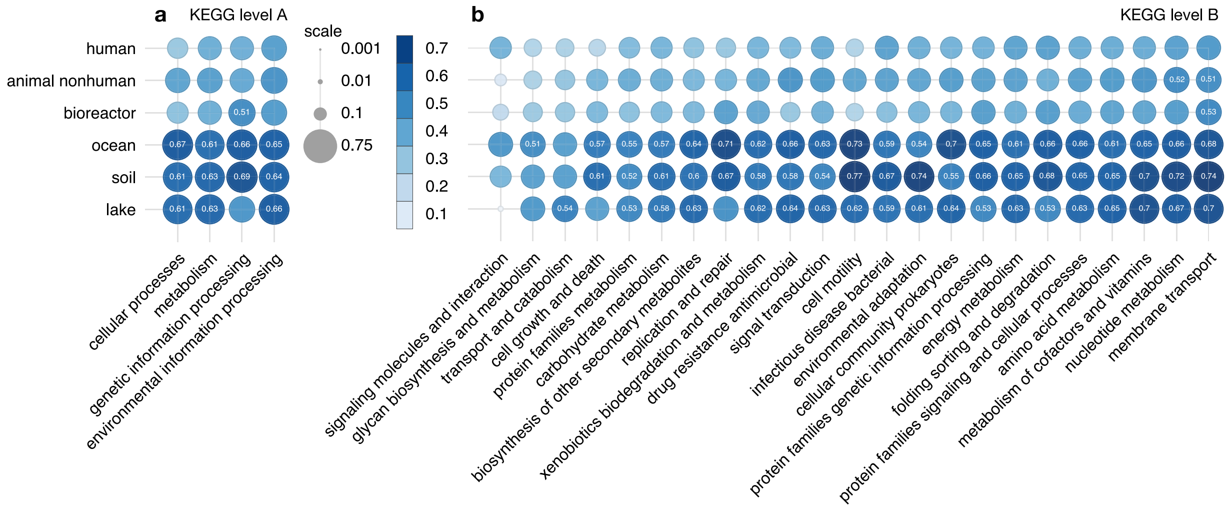

Fig. 5 Median absolute coverage biases by environment and gene category. Circle chart of median absolute coverage bias (median $|\beta|$) for each environment and various gene categories of particular interest (defined according to the KEGG hierarchy, levels A and B). The size and color darkness of each circle are proportional to the median $|\beta|$. Values above 0.5 are inscribed in the circles. Gene categories are sorted by increasing mean value. Ocean and soil generally exhibit the strongest coverage biases. For a similar graphic showing median coverage biases (median β) see Supplemental Fig. S14.

with highest median $|\beta|$ were cell motility and membrane transport (median $|\beta|$ between 0.68 and 0.77). This suggests that traits related to cell motility and membrane transport are particularly strong predictors of culturing success in these two environments.

Discussion

We have analyzed the distribution of thousands of genes across a large culture-independent collection of prokaryotic genomes, and have quantified gene-specific biases in RefSeq reference genomes. Given the general close association between prokaryotic culturing and the existence of a reference genome in RefSeq, as well as the close association between gene content and functional traits in prokaryotes, we expect that our conclusions largely translate to trait biases in prokaryotic cultures. We mention that MAG datasets may exhibit their own biases, for example towards more abundant organisms, or against organisms with multiple hard-to-assemble regions such as 16S SSU rRNA genes, and these biases could in principle influence some of the patterns reported here. However, there is little reason to believe at this point that these biases are driven by traits in a manner that is consistent across locations; in other words, we expect that these biases will tend to average out across locations, and thus not be a substantial driver of the trait biases in RefSeq relative to MAGs.

Genome-resolved metagenomics is greatly accelerating the discovery of novel microbial diversity[19,20,42,43]. Notwithstanding these breakthroughs, it should be remembered that culturing and associated whole genome sequencing remain essential tools for understanding the physiology and ecology of prokaryotes, and for decoding their genotype-phenotype mapping[1,44]. Our results suggest that the current culturing and associated whole genome sequencing efforts are heavily biased towards or against a variety of traits, particularly among ocean- and soil-associated prokaryotes. These biases lead to a distorted distribution of annotated genes/traits in reference databases, which in turn compromises the ecological conclusions one may draw from these distributions. These biases will also inevitably influence gene content and trait predictions for novel clades, for example in 16S rRNA amplicon sequencing studies, when these predictions are performed based on reference genome sets (as is common practice, e.g.[13,45]). In fact, the estimated coverage biases of many genes vary strongly between environments (Supplemental Fig. S15), suggesting that site-specific trait databases and statistical corrections may be needed for accurate trait estimation in novel clades. In addition, coverage biases may also delay the discovery of new industrially useful clades, such as novel methanogens for biofuel production, which we showed were severely underrepresented in RefSeq for all natural environments examined. The gene prevalences and coverage biases estimated in this study (provided as file `KOfam_gene_prevalences_and_biases.tsv.gz` on Figshare[41]) could help alleviate these issues in the future. For example, the conditional coverages (q_0 and q_1) estimated here can be used to correct for biases in phylogenetic trait prediction algorithms[16], and help steer future culturing efforts towards under-explored traits.

Methods

MAGs. MAGs from 203 distinct studies were either downloaded from NCBI GenBank or from other locations provided by published studies[46–212]. Only studies in which all MAGs were obtained from a similar environment or in which the environment was specified individually for each MAG were considered. For efficiency in MAG collection, we focused on studies with at least 30 MAGs. Studies focusing on a single clade (e.g., *Thaumarchaeota*) or trait (e.g., only methanogens) were omitted. An overview of included studies, including accession numbers and publication references, is provided in file `project_metadata.tsv` on Figshare[41]. With the exception of one study ("Genomes from Earth's Microbiomes" or GEM[137],), all other studies focused on specific environments. The GEM study itself comprised MAGs recovered from a multitude of environments across the world. Ten GEM MAGs were omitted because of missing taxonomic information. MAG qualities were determined based on the presence or absence of multiple universal single-copy genes using `checkM2` v0.1.3[213]. MAGs with completeness below 80% or contamination above 5% were omitted; thus 116,884 MAGs were kept for our analyses. An overview of completeness and contamination levels of kept MAGs is shown in Supplemental Fig. S1. We mention that these quality criteria do not exactly match commonly suggested conventions for "high" or "medium" quality MAGs[214]. Instead, the chosen thresholds are based on a reasonable balance between dropping too many MAGs and ensuring sufficient quality in the remaining MAG set. For example, a completeness of at least 90% suggested for "high quality" MAGs by[214] would be too stringent for some environments such as soil, while a completeness of at least 50% suggested for "medium quality" MAGs would be too low for reliably estimating gene prevalences.

The full size of the genome represented by a MAG (i.e., correcting for MAG incompleteness) was estimated by dividing the size of the MAG (in base pairs) by the completeness determined using `checkM2`. The taxonomic identities of MAGs were determined using the `GTDB-Tk` v1.4.1 workflow `classify_wf`[215], except for the GEM MAGs for which taxonomic identities were already provided by the original study. All software mentioned above were used with default options unless mentioned otherwise. An overview of represented prokaryotic phyla is shown in Supplemental Fig. S2.

MAGs were associated with various (not necessarily mutually exclusive) environments of interest based on the description in the publication associated with each study (if available), or based on the project description on GenBank, or — in the case of GEM genomes — based on the metadata table provided by the GEM study[137]. MAGs classified as human- and other animal-associated were explicitly excluded from the other environmental categories. Environments associated with each MAG are listed in `MAG_metadata_QF.tsv.gz` on Figshare[41].

To avoid species-level redundancies within the MAG dataset, we clustered MAGs into species genome bins (SGBs) at an ANI cutoff of 95% using a similar approach as described by[216], separately for each environment. Specifically, ANIs between all MAGs from a given environment were calculated using mash v2.3[217], with sketch size 5000 and otherwise default options. The average nucleotide divergence (AND) between any two MAGs was defined as 1-ANI. Bifurcating trees were constructed based on pairwise ANDs and using the hierarchical clustering algorithm implemented in the R package `fastcluster` v1.2.3 (function hclust with average linkage)[218]. For computational efficiency, prior to clustering, MAGs were split into smaller disjoint subsets of moderately to closely related MAGs, based on an AND cutoff threshold of 15%. Hierarchical clustering trees were rooted via the midpoint method[219], using the R package `castor` v1.7.2[220]. Note that each tip in a tree corresponded to a MAG. Next, tips in the hierarchical clustering trees were grouped into SGBs based on a maximum pairwise distance of 5% AND, using the function `collapse_tree_at_resolution` in the R package `castor`[220]. From each SGB, a single representative MAG was kept, chosen to be the MAG with the highest completeness. A total of 29,531 SGBs were thus obtained. An overview of MAGs and SGBs from each environment is given in Supplemental Table S2.

RefSeq genomes. Genomes were downloaded from the NCBI RefSeq database on October 7, 2021. All genomes whose `genome_rep` was "Full", whose `gap_fraction` was below 0.1, and whose `assembly_level` was one of "Complete Genome", "Contig", "Scaffold", "Chromosome", were downloaded. RefSeq genomes were grouped into various (not necessarily mutually exclusive) environments based on the associated biosample's metadata "geo_loc", "biosample_organism_name", "metagenome_source", "env_local_scale", "isolation_source" and "isolation_site", as follows. Genomes whose aforementioned metadata contained any of the words "soil", "rhizosphere", "rhizoplane", "root nodule" or "permafrost" were classified as soil-associated. Genomes whose aforementioned metadata contained any of the words "ocean", "marine" or "estuary" were classified as ocean-associated. Genomes whose aforementioned metadata contained any of the words or phrases "lake", "lakewater", "freshwater sediment", "freshwater mat" or "pond" were classified as lake-associated. Genomes whose aforementioned metadata contained any of the words or phrases "bioreactor", "wastewater treatment", "digester", "digestor", "reactor" or "activated sludge" were classified as bioreactor-associated. Genomes either identified as human-associated using `FAPROTAX` v1.2.4[14] or whose biosample "host" metadata was "homo sapiens", were classified as human-associated. Genomes either identified as animal-associated using `FAPROTAX` v1.2.4[14] or whose biosample "host" metadata was identified as a metazoan (based on metazoan latin names in the Open Tree of Life v13.4[221] and a custom list of animal common names), and not already classified as human-associated, were classified as non-human-animal-associated (in this study simply "animal-associated" for brevity). Note that `FAPROTAX` provides a convenient means to identify human- and animal-associated species, and is used here for the sole reason of increasing the accuracy of the environmental classifications of genomes. Genomes classified as human- and other animal-associated were subsequently excluded from the other environmental categories. A total of 184,131 genomes could be associated with at least one of the above environments. The number of genomes associated with each environment is given in Supplemental Table S2. The environments associated with each RefSeq genome are listed in file `RefSeq_genome_metadata_QF.tsv.gz` on Figshare[41].

To avoid species-level redundancies in some of our subsequent analyses, we clustered RefSeq genomes into species-level bins (STIBs) based on their provided species-level taxon identity (`species_taxid` field). When choosing STIB representatives we prioritized genomes based on their contig-N50 quality metric. An overview of RefSeq genomes and STIBs from each environment is given in Supplemental Table S2.

It is possible that contaminations exist in some RefSeq genomes, with reportedly isolate genomes actually originating from co-cultures, due to the difficulties of growing bacteria (notably Cyanobacteria) axenically[222,223]. Such contaminations, if widespread, could in principle introduce errors in our gene prevalence estimates for

RefSeq, although Cyanobacteria only constitute a very small fraction of the RefSeq genomes and we are unaware of any evidence suggesting that such issues are common across RefSeq. Future similar studies could avoid these issues (as well as assembly/binning issues discussed below) by utilizing single-cell amplified genomes[224].

Gene detection. Protein-coding genes were predicted for each MAG using `prodigal` v2.6.3, and were subsequently annotated (matched to KEGG orthologs) using the `KOfam` Hidden Markov Model database[27] (release 2020-04-02, comprising 21461 genes) and `hmmsearch` v3.3.2[225]. A similar approach was used to detect and annotate genes in prokaryotic RefSeq genomes. To reduce computation time, we only annotated a random subset of 137,726 genomes, however all STIB representatives were explicitly included in this subset. Genes not found in any MAG nor any functionally annotated RefSeq genome were omitted from subsequent analyses, as most of these genes are likely eukaryote specific. Thus, a total of 12,454 genes were kept. The number of MAGs, MAG-SGBs, RefSeq genomes and RefSeq STIBs in which each was gene found is listed in file `KOfam_gene_prevalences_and_biases.tsv.gz` on Figshare[41]. Histograms of the number of MAG-SGBs and RefSeq STIBs in which each gene was detected are shown in Supplemental Figs. S3, S4, respectively. The average number of predicted protein-coding genes per MAG and per genome, as well as the fraction of such genes that could be functionally annotated using KOfam, are listed in Supplemental Table S1.

To examine the role of potential biases in the KEGG database we also matched predicted genes to the eggNOG 5.0 database of orthologous groups[34], as follows. Protein sequences of all eggNOG orthologs were downloaded from the eggNOG website at http://eggnog5.embl.de/download/eggnog_5.0 (file `e5.proteomes.faa.gz`). Amino acid sequences predicted in MAGs or RefSeq genomes were then matched to the downloaded protein sequence database using `diamond` v2.0.15.153[226], and only hits with an e-value below 10^{-10} were kept. Matched sequences were converted to eggNOG ortholog IDs using a lookup table downloaded from the eggNOG website (file `all_members.tsv.gz`). For computational tractability, only 50,000 randomly selected eggNOG orthologs were considered for subsequent analysis. Note that the focus of this article is on KEGG orthologs (KOs), hence unless specified otherwise "gene" refers to a KO rather than an eggNOG ortholog.

Estimating gene prevalences in MAG-SGBs. To estimate the prevalence of a given gene (KEGG ortholog or eggNOG) in populations represented by our MAG-SGB set, i.e., the probability α that the population represented by a randomly chosen MAG-SGB exhibits the gene, we proceeded as follows. Throughout the analysis described below, we only considered the single representative MAG of each SGB. We assumed that the probability of a gene being detected in a randomly chosen MAG (using `hmmsearch` as described earlier) is given by the product $\alpha \cdot C$, where α is the true prevalence of the gene across species and C is the completeness of the MAG. We also assumed that the false positive and false negative detection rate of genes in MAGs are negligible. Hence, if M_0 is the set of MAGs in which the gene was not detected, and M_1 the set of MAGs where the gene was detected, the total likelihood of our dataset (for a given α) is given by the product of probabilities:

$$L=\prod_{m\in M_0}(1-\alpha C_m)\cdot\prod_{m\in M_1}\alpha C_m, \tag{1}$$

where C_m is the completeness of the m-th MAG. The maximum-likelihood estimate of α, denoted $\hat{\alpha}$, can be found by demanding that the derivative $\partial L/\partial\alpha$ is zero, which is equivalent to the following equation:

$$|M_1|=\sum_{m\in M_0}\frac{\hat{\alpha}C_m}{1-\hat{\alpha}C_m}, \tag{2}$$

where is the cardinality of $|M_1|$. Note that if all MAGs were complete ($C_m = 1$ for all m), the solution to Eq. (2) would simply be $\hat{\alpha}=|M_1|/(|M_0|+|M_1|)$, however in reality most MAGs were incomplete, making an analytical solution difficult. Equation (2) was thus solved numerically for $\hat{\alpha}$ in python, using the bisection method. Confidence intervals were obtained using parametric bootstrapping, i.e., based on gene presences/absences generated randomly across the MAGs according to the above statistical model and the fitted α. We mention that, like most bioinformatics analyses in this paper, `checkM2` -based completeness estimates may not be fully accurate. Errors in MAG completeness estimates would introduce errors in the gene prevalence estimates, although these errors are suspected to be small based on typical checkM2 errors (mean average error ~3%)[213]. Further, we mention that the MAGs and genomes analyzed were originally generated using a variety of alternative sequencing platforms, assembly and binning tools. This methodological variation could in principle impact contig assembly lengths, which ORFs are included/split on those contigs, the frequency of chimeric contigs, and which contigs get included in each bin, which would by extension impact the recovery of ORFs and the estimation of gene prevalences and coverage biases. That said, our analysis of gene prevalences restricted to MAG-SGBs matched by RefSeq genomes (see main article and Supplemental Fig. S9) indicated that ORF recovery and gene detection efficiency did not noticeably differ between MAG-SGBs and their matched RefSeq genomes.

Estimating gene prevalences in RefSeq STIBs. Prevalences of genes (KEGG orthologs or eggNOGs) across RefSeq STIBs were estimated using a similar approach as for MAG-SGBs, the main difference being that the completeness of a RefSeq genome was computed based on the associated gap fraction listed in RefSeq (that said, we mention that the gap fraction was negligible for the majority of RefSeq genomes considered). To further examine how the gene prevalence estimates across RefSeq STIBs would change in the absence of any genome size biases (relative to the MAG dataset), i.e., correcting for the distribution of genome sizes, we proceeded as follows. We binned MAG-SGBs based on their estimated full genome size (i.e., correcting for MAG incompleteness) into 0.5 Mbp size intervals or “strata” (i.e., 0–0.5 Mbp, 0.5–1 Mbp, 1–1.5 Mbp, …). Similarly, we binned RefSeq STIBs based on their full genome size (attribute `total_length`) into the same size intervals, and independently estimated gene prevalences separately for each size interval, i.e, treating each set of STIBs in a size interval as a separate dataset. For each gene, we then computed the weighted average prevalence across all size intervals, weighting each size interval by the number of MAG-SGBs in that interval. This weighting adjustment is commonly performed in demographic surveys in which different strata are sampled at different proportions[32]. For comparisons of gene prevalence estimates between MAG-SGBs and RefSeq STIBs see Fig. 1, and Supplemental Figs. S8–S11.

Estimating coverage biases. For the following analysis, MAGs have been deduplicated at the SGB level, i.e., we considered only one representative MAG per SGB. We say that a MAG “matches” a RefSeq genome if its ANI to that genome (as calculated using `mash`) was at least 95%. By “coverage” we mean the probability that a randomly chosen MAG matches a RefSeq genome. To investigate the coverages of MAGs, separately for each environment and depending on the presence or absence of specific genes (KEGG orthologs or eggNOGs), we proceeded as follows. For any given environment, let q denote the overall coverage, i.e., the probability that a randomly chosen MAG matches a RefSeq genome. For any given environment and gene, let q_0 and q_1 be the conditional probabilities that a randomly chosen MAG matches a RefSeq genome given that the population represented by the MAG lacked or had the gene, respectively. Note that a gene may be missing from an incomplete MAG even if the represented population had the gene. The two conditional probabilities q_0 and q_1 are a priori unknown, and correspond to the coverage of MAGs in the absence or presence, respectively, of the gene in the represented populations. For example, if $q_0 > q_1$ then this means that RefSeq is biased towards organisms lacking the gene, whereas if $q_0 < q_1$ RefSeq would be biased towards organisms having the gene. Note that by mathematical necessity either $q_0 \leq q \leq q_1$ or $q_0 \geq q \geq q_1$, i.e., q_0 and q_1 cannot be both above or both below the overall coverage q. Our first objective was to estimate the q_0 and q_1 based on our MAG dataset. For any given environment and gene, let M_0^0 be the set of MAGs in which the gene was not detected and which did not match any RefSeq genome, let M_0^1 be the set of MAGs in which the gene was not detected and which did match a RefSeq genome, let M_1^0 be the set of MAGs in which the gene was detected but which did not match any RefSeq genome, and let M_1^1 be the set of MAGs in which the gene was detected and which did match a RefSeq genome. As before, C_m denotes the completeness of the m-th MAG, and α denotes the probability that the population represented by a randomly selected MAG had the gene. Hence, for example, the probability of detecting a gene in a randomly chosen MAG together with that MAG matching a RefSeq genome is given by the product $\alpha C_m q_1$. The probability of not detecting the gene in a randomly chosen MAG and that MAG not matching any RefSeq genome is given by the sum of probabilities of two complementary events: either the represented population did not have the gene and its MAG did not match any RefSeq genome (probability $(1-\alpha)(1-q_0)$), or the population did have the gene but the gene was missing from the MAG (due to incompleteness) and the MAG did not match any RefSeq genome (probability $\alpha(1-C_m)(1-q_1)$). Similar arguments can be made for all other possible scenarios as well, eventually leading to the following expression for the likelihood of our data:

$$\begin{aligned} L = &\prod_{m\in M_0^0} [(1-\alpha)(1-q_0) + \alpha(1-C_m)(1-q_1)] \times \\ &\prod_{m\in M_0^1} [(1-\alpha)q_0 + \alpha(1-C_m)q_1] \\ &\times \prod_{m\in M_1^0} \alpha C_m(1-q_1) \times \prod_{m\in M_1^1} \alpha C_m q_1. \end{aligned} \tag{3}$$

The maximum likelihood estimates $\hat{q}_0$ and $\hat{q}_1$ were obtained by numerically maximizing the log-likelihood $\ln(L)$ in python and using the previously obtained maximum-likelihood estimate $\hat{\alpha}$. To avoid inaccurate estimates of the conditional coverages q_0, q_1, we only considered genes detected in at least 100 MAGs and missing from at least 100 MAGs. Further, optimization of the likelihood failed for a small fraction of genes. Thus, the specific set of genes considered differed somewhat between environments (overview in Supplemental Table S4, details in file `KOfam_gene_prevalences_and_biases.tsv.gz` on Figshare[41]). For an overview of estimated conditional coverages see Supplemental Fig. S13.

To quantify the strength of bias for or against a gene in a way that facilitates comparison between environments, we defined the “coverage bias” as follows:

$$\beta = \mathrm{sign}(q_1 - q_0) \cdot \left[1 - \frac{min(q_0, q_1)}{max(q_0, q_1)}\right]. \tag{4}$$

The coverage bias β can also equivalently be written as follows:

$$\beta = \begin{cases} 1 - \dfrac{q_0}{q1} & : q_1 \geq q_0 \\ \dfrac{q_1}{q_0} - 1 & : q_1 \leq q_0 \end{cases}. \tag{5}$$

Observe that β is always between -1 and 1, and that a positive (or negative) value implies a bias for (or against) the specific gene. A value of $\beta = 0$ implies that $q_0 = q_1 = q$ and hence an absence of any bias related to the gene. A value of $\beta = 1$ implies that $q_0 = 0$, which means that the absence of the gene in an organism makes it improbable that the organism is represented in RefSeq (at ANI $\geq$95%). On the other extreme, a value of $\beta = -1$ implies that $q_1 = 0$, which means that the presence of the gene in an organism makes it improbable that the organism is represented in RefSeq. A useful property of β is that it only depends on the ratio q_1/q_0 but not on the overall coverage q nor on the gene's prevalence α, thus making it suitable for exploring the effects of the environment on gene-specific coverage biases. For example, if MAGs from populations having the gene are 3 times more probable to match a RefSeq genome compared to MAGs from populations lacking the gene (i.e., $q_1 = 3q_0$), then $\beta = 1 - \frac{1}{3} = 2/3$ regardless of whether that environment in and of itself is strongly biased for or against in RefSeq, and regardless of the gene's prevalence in that environment. The two-sided statistical significance of β was determined using parametric bootstrapping under the null model of zero bias ($q_0 = q_1$), i.e., by randomly re-generating gene presences/absences in the MAGs according to the fitted α and accounting for MAG completeness while ignoring MAG coverages. We mention that the above estimates are not adjusted for the differences in the genome size distributions in RefSeq-STIBs versus MAG-SGBs. For example, a q_1 greater than q_0 may be partly due to the fact that the presence of a given gene in a species will tend to correlate positively with the species' genome size (all else being equal), which in turn will correlate positively with the inclusion of the species in RefSeq, thus increasing q_1 relative to q_0. For a summary of coverage biases see Fig. 4, Supplemental Fig. S16 and Supplemental Table S4.

Kernel density estimates of genome size distributions. Gaussian kernel density estimates of the distribution of MAG full genome sizes (i.e., accounting for MAG incompleteness) or RefSeq genome sizes (attribute `total_length`) were computed using the `KernelDensity` function in the python package `scikit-learn` v1.0.2[227]. The optimal KDE bandwidth was determined separately for each environment, and separately for MAGs and RefSeq genomes, via 5-fold cross-validation using the function `GridSearchCV in scikit-learn`. The pool of bandwidths considered ranged from 0.001 up to 10 times the total data range. Optimized KDE bandwidths are listed in Supplemental Table S3.

Data availability

All data have been previously published and are publicly available, as described in the Methods. Supplemental files relating to this analysis are available at Figshare[41]: MAG sources are given in file `project_metadata.tsv`, accession numbers for MAGs (where available) are given in file `MAG_metadata_QF.tsv.gz`, accession numbers for RefSeq genomes are given in file `RefSeq_genome_metadata_QF.tsv.gz`, analysis results for each gene (KEGG ortholog) are given in file `KOfam_gene_prevalences_and_biases.tsv.gz`, a table of all KOs found per MAG is given as file `KOfams_vs_MAGs.tsv.gz` and a table of all KOs found per RefSeq genome is given as file `KOfams_vs_RefSeq_genomes.tsv.gz`.

Code availability

All software used in this paper have been described in the Methods and are freely available online. A copy of our workflow (bash, python, R code files) is also available at Figshare[41].

Received: 23 June 2022; Accepted: 31 January 2023;
Published online: 09 February 2023

References

1. Overmann, J., Abt, B. & Sikorski, J. Present and future of culturing bacteria. *Annual Review of Microbiology* **71**, 711–730 (2017).
2. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733 (2016).
3. Bobay, L. M. & Ochman, H. Biological species are universal across life's domains. *Genome Biology and Evolution* **9**, 491–501 (2017).
4. Magnabosco, C., Moore, K., Wolfe, J. & Fournier, G. Dating phototrophic microbial lineages with reticulate gene histories. *Geobiology* **16**, 179–189 (2018).
5. Louca, S. *et al.* Function and functional redundancy in microbial systems. *Nature Ecology & Evolution* **2**, 936–943 (2018).
6. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90 K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**, 5114 (2018).
7. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. *Nature Communications* **10**, 5477 (2019).
8. Royalty, T.M. & Steen, A.D. Quantitatively partitioning microbial genomic traits among taxonomic ranks across the microbial tree of life. *mSphere* **4** (2019).
9. Murray, C. S., Gao, Y. & Wu, M. Re-evaluating the evidence for a universal genetic boundary among microbial species. *Nature Communications* **12**, 4059 (2021).
10. Powell, S. *et al.* eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research* **42**, D231–D239 (2014).
11. Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. & Schmidt, T. M. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research* **43**, D593–D598 (2014).
12. Douglas, G. M. *et al.* Picrust2 for prediction of metagenome functions. *Nature Biotechnology* **38**, 685–688 (2020).
13. Wemheuer, F. *et al.* Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environmental Microbiome* **15**, 1–12 (2020).
14. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272–1277 (2016).
15. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature* **462**, 1056–1060 (2009).
16. Louca, S. & Pennell, M. W. A general and efficient algorithm for the likelihood of diversification and discrete-trait evolutionary models. *Systematic Biology* **69**, 545–556 (2020).

17. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
18. Sharon, I. & Banfield, J. F. Genomes from metagenomics. *Science* **342**, 1057–1058 (2013).
19. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2**, 1533–1542 (2017).
20. Chen, L. X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Research* **30**, 315–333 (2020).
21. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences* **102**, 2567–2572 (2005).
22. Kim, M., Oh, H. S., Park, S. C. & Chun, J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Journal of Systematic and Evolutionary Microbiology* **64**, 346–351 (2014).
23. Shapiro, B.J. What microbial population genomics has taught us about speciation. In Polz, M.F. & Rajora, O.P. (eds.) *Population Genomics: Microorganisms*, 31–47 (Springer International Publishing, Cham, Switzerland, 2019).
24. Olm, M. R. *et al.* Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. *mSystems* **5**, e00731–19 (2020).
25. Lagkouvardos, I., Overmann, J. & Clavel, T. Cultured microbes represent a substantial fraction of the human and mouse gut microbiota. *Gut Microbes* **8**, 493–503 (2017).
26. Zhang, Z., Wang, J., Wang, J., Wang, J. & Li, Y. Estimate of the sequenced proportion of the global prokaryotic genome. *Microbiome* **8**, 1–9 (2020).
27. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2019).
28. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends in Genetics* **17**, 589–596 (2001).
29. Morris, J. J., Lenski, R. E. & Zinser, E. R. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio* **3**, e00036–12 (2012).
30. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME Journal* **8**, 1553–1565 (2014).
31. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
32. Gary, P.R. Adjusting for nonresponse in surveys. In Smart, J.C. (ed.) *Higher Education: Handbook of Theory and Research*, chap. 8, 411–449 (Springer, Dordrecht, Netherlands, 2007).
33. Maguire, F. *et al.* Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. *Microbial Genomics* **6**, mgen000436 (2020).
34. Huerta-Cepas, J. *et al.* eggnog 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309–D314 (2019).
35. Abdel-Hamid, A.M., Solbiati, J.O., Cann, I.K.O., Sariaslani, S. & Gadd, G.M. *Insights into lignin degradation and its potential industrial applications*, vol. 82, chap. 1, 1–28 (Academic Press, 2013).
36. El-Bondkly, A.M. Sequence analysis of industrially important genes from trichoderma. In *Biotechnology and biology of Trichoderma*, **chap. 28**, 377–392 (Elsevier, 2014).
37. Dawood, A. & Ma, K. Applications of microbial β-mannanases. *Frontiers in Bioengineering and Biotechnology* **8** (2020).
38. Khelaifia, S., Raoult, D. & Drancourt, M. A versatile medium for cultivating methanogenic archaea. *PLOS ONE* **8**, e61563 (2013).
39. Khelaifia, S. *et al.* Aerobic culture of methanogenic archaea without an external source of hydrogen. *European Journal of Clinical Microbiology & Infectious Diseases* **35**, 985–991 (2016).
40. Michał, B. *et al.* Phymet2: a database and toolkit for phylogenetic and metabolic analyses of methanogens. *Environmental Microbiology Reports* **10**, 378–382 (2018).
41. Albright, S. & Louca, S. Trait biases in microbial reference genomes, *figshare.*, https://doi.org/10.6084/m9.figshare.c.6055004.v1 (2022).
42. Castelle, C. J. & Banfield, J. F. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* **172**, 1181–1197 (2018).
43. Murray, A. E. *et al.* Roadmap for naming uncultivated archaea and bacteria. *Nature Microbiology* **5**, 987–994 (2020).
44. Palleroni, N. J. Prokaryotic diversity and the importance of culturing. *Antonie van Leeuwenhoek* **72**, 3–19 (1997).
45. Langille, M. G. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* **31**, 814–821 (2013).
46. Tran, P. Q. *et al.* Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika. *The ISME Journal* **15**, 1971–1986 (2021).
47. Kroeger, M. E. *et al.* New biological insights into how deforestation in amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes. *Frontiers in Microbiology* **9**, 1635 (2018).
48. Nathani, N. M. *et al.* 309 metagenome assembled microbial genomes from deep sediment samples in the Gulfs of Kathiawar Peninsula. *Scientific Data* **8**, 194 (2021).
49. Irazoqui, J. M., Eberhardt, M. F., Adjad, M. M., Amadio, A. F. & Collado, M. C. Identification of key microorganisms in facultative stabilization ponds from dairy industries, using metagenomics. *PeerJ* **10**, e12772 (2022).
50. Hwang, Y. *et al.* Leave no stone unturned: individually adapted xerotolerant Thaumarchaeota sheltered below the boulders of the Atacama Desert hyperarid core. *Microbiome* **9**, 234 (2021).
51. Tully, B., Wheat, C. G., Glazer, B. T. & Huber, J. A dynamic microbial community with high functional redundancy inhabits the cold, oxic subseafloor aquifer. *ISME Journal* **12**, 1–16 (2018).
52. Vanwonterghem, I., Jensen, P. D., Rabaey, K. & Tyson, G. W. Genome-centric resolution of microbial diversity, metabolism and interactions in anaerobic digestion. *Environmental Microbiology* **18**, 3144–3158 (2016).
53. Glasl, B. *et al.* Comparative genome-centric analysis reveals seasonal variation in the function of coral reef microbiomes. *The ISME Journal* **14**, 1435–1450 (2020).
54. Robbins, S. J. *et al.* A genomic view of the reef-building coral *Porites lutea* and its microbial symbionts. *Nature Microbiology* **4**, 2090–2100 (2019).
55. Engelberts, J. P. *et al.* Characterization of a sponge microbiome using an integrative genome-centric approach. *The ISME Journal* **14**, 1100–1110 (2020).
56. Bowerman, K. L. *et al.* Disease-associated gut microbiome and metabolome changes in patients with chronic obstructive pulmonary disease. *Nature Communications* **11**, 5886 (2020).
57. Chen, Y. J. *et al.* Hydrodynamic disturbance controls microbial community assembly and biogeochemical processes in coastal sediments. *The ISME Journal* **16**, 750–763 (2022).
58. Hugerth, L. W. *et al.* Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biology* **16**, 279 (2015).
59. Alneberg, J. *et al.* Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Communications Biology* **3**, 119 (2020).

60. Di Cesare, A. *et al.* Genomic comparison and spatial distribution of different Synechococcus phylotypes in the Black Sea. *Frontiers in Microbiology* **11**, 1979 (2020).
61. van Vliet, D. M. *et al.* The bacterial sulfur cycle in expanding dysoxic and euxinic marine waters. *Environmental Microbiology* **23**, 2834–2857 (2021).
62. Dalcin Martins, P. *et al.* Enrichment of novel Verrucomicrobia, Bacteroidetes, and Krumholzibacteria in an oxygen-limited methane- and iron-fed bioreactor inoculated with Bothnian Sea sediments. *MicrobiologyOpen* **10**, e1175 (2021).
63. Stewart, R. D. *et al.* Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nature Biotechnology* **37**, 953–961 (2019).
64. Segura-Wang, M., Grabner, N., Koestelbauer, A., Klose, V. & Ghanbari, M. Genome-resolved metagenomics of the chicken gut microbiome. *Frontiers in Microbiology* **12**, 726923 (2021).
65. Ruuskanen, M. O. *et al.* Microbial genomes retrieved from High Arctic lake sediments encode for adaptation to cold and oligotrophic environments. *Limnology and Oceanography* **65**, S233–S247 (2020).
66. Haas, S., Desai, D. K., LaRoche, J., Pawlowicz, R. & Wallace, D. W. R. Geomicrobiology of the carbon, nitrogen and sulphur cycles in Powell Lake: a permanently stratified water column containing ancient seawater. *Environmental Microbiology* **21**, 3927–3952 (2019).
67. Spasov, E. *et al.* High functional diversity among Nitrospira populations that dominate rotating biological contactor microbial communities in a municipal wastewater treatment plant. *The ISME Journal* **14**, 1857–1872 (2020).
68. Vigneron, A. *et al.* Genomic evidence for sulfur intermediates as new biogeochemical hubs in a model aquatic microbial ecosystem. *Microbiome* **9**, 46 (2021).
69. Galambos, D., Anderson, R. E., Reveillaud, J. & Huber, J. A. Genome-resolved metagenomics and metatranscriptomics reveal niche differentiation in functionally redundant microbial communities at deep-sea hydrothermal vents. *Environmental Microbiology* **21**, 4395–4410 (2019).
70. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature Communications* **9**, 870 (2018).
71. Xing, P. *et al.* Stratification of microbiomes during the holomictic period of Lake Fuxian, an alpine monomictic lake. *Limnology and Oceanography* **65**, S134–S148 (2020).
72. Zhang, S., Hu, Z. & Wang, H. Metagenomic analysis exhibited the co-metabolism of polycyclic aromatic hydrocarbons by bacterial community from estuarine sediment. *Environment International* **129**, 308–319 (2019).
73. Lin, Y., Wang, L., Xu, K., Li, K. & Ren, H. Revealing taxon-specific heavy metal-resistance mechanisms in denitrifying phosphorus removal sludge using genome-centric metaproteomics. *Microbiome* **9**, 67 (2021).
74. Liu, L. *et al.* High-quality bacterial genomes of a partial-nitritation/anammox system by an iterative hybrid assembly method. *Microbiome* **8**, 155 (2020).
75. Kantor, R. S. *et al.* Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unravelled with genome-resolved metagenomics. *Environmental Microbiology* **17**, 4929–4941 (2015).
76. Zhou, Z. *et al.* Gammaproteobacteria mediating utilization of methyl-, sulfur- and petroleum organic compounds in deep ocean hydrothermal plumes. *The ISME Journal* **14**, 3136–3148 (2020).
77. Reysenbach, A. L. *et al.* Complex subsurface hydrothermal fluid mixing at a submarine arc volcano supports distinct and highly diverse microbial communities. *Proceedings of the National Academy of Sciences* **117**, 32627–32638 (2020).
78. Hou, J. *et al.* Microbial succession during the transition from active to inactive stages of deep-sea hydrothermal vent sulfide chimneys. *Microbiome* **8**, 102 (2020).
79. Campanaro, S. *et al.* Metagenomic analysis and functional characterization of the biogas microbiome using high throughput shotgun sequencing and a novel binning strategy. *Biotechnology for Biofuels* **9**, 26 (2016).
80. Singleton, C. M. *et al.* Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nature Communications* **12**, 2009 (2021).
81. Diamond, S. *et al.* Mediterranean grassland soil C–N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms. *Nature Microbiology* **4**, 1356–1367 (2019).
82. Rasigraf, O. *et al.* Microbial community composition and functional potential in Bothnian Sea sediments is linked to Fe and S dynamics and the quality of organic matter. *Limnology and Oceanography* **65**, S113–S133 (2020).
83. Rissanen, A. J. *et al.* Vertical stratification patterns of methanotrophs and their genetic controllers in water columns of oxygen-stratified boreal lakes. *FEMS Microbiology Ecology* **97**, fiaa252 (2021).
84. Campanaro, S. *et al.* New insights from the biogas microbiome by comprehensive genome-resolved metagenomics of nearly 1600 species originating from multiple anaerobic digesters. *Biotechnology for Biofuels* **13**, 25 (2020).
85. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology* **39**, 105–114 (2021).
86. Zhou, Z. *et al.* Genome- and community-level interaction insights into carbon utilization and element cycling functions of hydrothermarchaeota in hydrothermal sediment. *mSystems* **5** (2020).
87. Pachiadaki, M. G. *et al.* Charting the complexity of the marine microbiome through single-cell genomics. *Cell* **179**, 1623–1635.e11 (2019).
88. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
89. Greenlon, A. *et al.* Global-level population genomics reveals differential effects of geography and phylogeny on horizontal gene transfer in soil bacteria. *Proceedings of the National Academy of Sciences* **116**, 15200–15209 (2019).
90. Hervé, V. *et al.* Phylogenomic analysis of 589 metagenome-assembled genomes encompassing all major prokaryotic lineages from the gut of higher termites. *PeerJ* **8**, e8614 (2020).
91. von Appen, W.J. The expedition PS114 of the research vessel POLARSTERN to the Fram Strait in 2018. Tech. Rep., Alfred Wegener Institute for Polar and Marine Research (2018).
92. Dombrowski, N., Seitz, K. W., Teske, A. P. & Baker, B. J. Genomic insights into potential interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments. *Microbiome* **5**, 106 (2017).
93. Yu, J. *et al.* Dna-stable isotope probing shotgun metagenomics reveals the resilience of active microbial communities to biochar amendment in oxisol soil. *Frontiers in Microbiology* **11**, 587972 (2020).
94. Forster, S. C. *et al.* A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature Biotechnology* **37**, 186–192 (2019).
95. Gharechahi, J. *et al.* Metagenomic analysis reveals a dynamic microbiome with diversified adaptive functions to utilize high lignocellulosic forages in the cattle rumen. *The ISME Journal* **15**, 1108–1120 (2021).
96. Meier, D. V., Imminger, S., Gillor, O. & Woebken, D. Distribution of mixotrophy and desiccation survival mechanisms across microbial genomes in an arid biological soil crust community. *mSystems* **6**, e00786–20 (2021).
97. Haro-Moreno, J. M. *et al.* Dysbiosis in marine aquaculture revealed through microbiome analysis: reverse ecology for environmental sustainability. *FEMS Microbiology Ecology* **96**, fiaa218 (2020).
98. Haro-Moreno, J. M. *et al.* Fine metagenomic profile of the Mediterranean stratified and mixed water columns revealed by assembly and recruitment. *Microbiome* **6**, 128 (2018).

99. Dong, X. *et al*. Metabolic potential of uncultured bacteria and archaea associated with petroleum seepage in deep-sea sediments. *Nature Communications* **10**, 1816 (2019).
100. Poghosyan, L. *et al*. Metagenomic profiling of ammonia- and methane-oxidizing microorganisms in two sequential rapid sand filters. *Water Research* **185**, 116288 (2020).
101. Paula, D. M., Jeroen, F., Hugh, M. & Meng, M. L. & J., W.M. Wetland sediments host diverse microbial taxa capable of cycling alcohols. *Applied and Environmental Microbiology* **85**, 00189–19 (2019).
102. Aromokeye, D. A. *et al*. Crystalline iron oxides stimulate methanogenic benzoate degradation in marine sediment-derived enrichment cultures. *The ISME Journal* **15**, 965–980 (2021).
103. Borchert, E. *et al*. Deciphering a marine bone-degrading microbiome reveals a complex community effort. *mSystems* **6**, e01218–20 (2021).
104. Osvatic, J. T. *et al*. Global biogeography of chemosynthetic symbionts reveals both localized and globally distributed symbiont groups. *Proceedings of the National Academy of Sciences* **118**, e2104378118 (2021).
105. Boeuf, D. *et al*. Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean. *Proceedings of the National Academy of Sciences* **116**, 11824–11832 (2019).
106. Woodcroft, B. J. *et al*. Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**, 49–54 (2018).
107. Alqahtani, M. F. *et al*. Enrichment of *Marinobacter sp*. and halophilic homoacetogens at the biocathode of microbial electrosynthesis system inoculated with Red Sea brine pool. *Frontiers in Microbiology* **10**, 2563 (2019).
108. Haroon, M. F., Thompson, L. R., Parks, D. H., Hugenholtz, P. & Stingl, U. A catalogue of 136 microbial draft genomes from Red Sea metagenomes. *Scientific Data* **3**, 160050 (2016).
109. Vavourakis, C. D. *et al*. A metagenomics roadmap to the uncultured genome diversity in hypersaline soda lake sediments. *Microbiome* **6**, 1–18 (2018).
110. Cabello-Yeves, P. J. *et al*. Microbiome of the deep Lake Baikal, a unique oxic bathypelagic habitat. *Limnology and Oceanography* **65**, 1471–1488 (2020).
111. Vavourakis, C. D. *et al*. Metagenomes and metatranscriptomes shed new light on the microbial-mediated sulfur cycle in a siberian soda lake. *BMC Biology* **17**, 69 (2019).
112. Waterworth, S. C., Isemonger, E. W., Rees, E. R., Dorrington, R. A. & Kwan, J. C. Conserved bacterial genomes from two geographically isolated peritidal stromatolite formations shed light on potential functional guilds. *Environmental Microbiology Reports* **13**, 126–137 (2021).
113. Huddy, R. J. *et al*. Thiocyanate and organic carbon inputs drive convergent selection for specific autotrophic Afipia and Thiobacillus strains within complex microbiomes. *Frontiers in Microbiology* **12**, 643368 (2021).
114. Emerson, J. B. *et al*. Diverse sediment microbiota shape methane emission temperature sensitivity in Arctic lakes. *Nature Communications* **12**, 5815 (2021).
115. Chiri, E. *et al*. Termite gas emissions select for hydrogenotrophic microbial communities in termite mounds. *Proceedings of the National Academy of Sciences* **118**, e2102625118 (2021).
116. Gong, G., Zhou, S., Luo, R., Gesang, Z. & Suolang, S. Metagenomic insights into the diversity of carbohydrate-degrading enzymes in the yak fecal microbial community. *BMC Microbiology* **20**, 302 (2020).
117. Zhou, S. *et al*. Characterization of metagenome-assembled genomes and carbohydrate-degrading genes in the gut microbiota of Tibetan pig. *Frontiers in Microbiology* **11**, 595066 (2020).
118. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data* **5**, 170203 (2018).
119. Lavrinienko, A. *et al*. Two hundred and fifty-four metagenome-assembled bacterial genomes from the bank vole gut microbiota. *Scientific Data* **7**, 312 (2020).
120. Peng, X. *et al*. Genomic and functional analyses of fungal and bacterial consortia that enable lignocellulose breakdown in goat gut microbiomes. *Nature Microbiology* **6**, 499–511 (2021).
121. Dudek, N. K. *et al*. Novel microbial diversity and functional potential in the marine mammal oral microbiome. *Current Biology* **27**, 3752–3762.e6 (2017).
122. Pinto, A. J. *et al*. Metagenomic evidence for the presence of comammox nitrospira-like bacteria in a drinking water system. *mSphere* **1**, e00054–15 (2015).
123. Zaremba-Niedzwiedzka, K. *et al*. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
124. Nobu, M. K. *et al*. Catabolism and interactions of uncultured organisms shaped by eco-thermodynamics in methanogenic bioprocesses. *Microbiome* **8**, 111 (2020).
125. Butterfield, C. N. *et al*. Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ* **4**, e2687 (2016).
126. Castelle, C. J. *et al*. Protein family content uncovers lineage relationships and bacterial pathway maintenance mechanisms in DPANN Archaea. *Frontiers in Microbiology* **12**, 660052 (2021).
127. Alteio, L. V. *et al*. Complementary metagenomic approaches improve reconstruction of microbial diversity in a forest soil. *mSystems* **5**, e00768–19 (2020).
128. Shaiber, A. *et al*. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biology* **21**, 292 (2020).
129. Jungbluth, S. P., Amend, J. P. & Rappé, M. S. Metagenome sequencing and 98 microbial genomes from Juan de Fuca Ridge flank subsurface fluids. *Scientific Data* **4**, 170037 (2017).
130. Sheik, C. S. *et al*. Dolichospermum blooms in Lake Superior: DNA-based approach provides insight to the past, present and future of blooms. *Journal of Great Lakes Research* **48**, 1191–1205 (2022).
131. Barnum, T. P. *et al*. Genome-resolved metagenomics identifies genetic mobility, metabolic interactions, and unexpected diversity in perchlorate-reducing communities. *The ISME Journal* **12**, 1568–1581 (2018).
132. Julian, D. *et al*. Coastal ocean metagenomes and curated metagenome-assembled genomes from Marsh Landing, Sapelo Island (Georgia, USA). *Microbiology Resource Announcements* **8**, e00934–19 (2019).
133. Breister, A. M. *et al*. Soil microbiomes mediate degradation of vinyl ester-based polymer composites. *Communications Materials* **1**, 101 (2020).
134. Fu, H., Uchimiya, M., Gore, J. & Moran, M. A. Ecological drivers of bacterial community assembly in synthetic phycospheres. *Proceedings of the National Academy of Sciences* **117**, 3656–3662 (2020).
135. Nobu, M. K. *et al*. Thermodynamically diverse syntrophic aromatic compound catabolism. *Environmental Microbiology* **19**, 4576–4586 (2017).
136. Pasolli, E. *et al*. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662 (2019).
137. Nayfach, S. *et al*. A genomic catalog of Earth's microbiomes. *Nature Biotechnology* **39**, 499–509 (2021).
138. Li, Z. *et al*. Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. *The ISME Journal* **15**, 2366–2378 (2021).
139. Bay, S. K. *et al*. Trace gas oxidizers are widespread and active members of soil microbial communities. *Nature Microbiology* **6**, 246–256 (2021).

140. Seyler, L. M., Trembath-Reichert, E., Tully, B. J. & Huber, J. A. Time-series transcriptomics from cold, oxic subseafloor crustal fluids reveals a motile, mixotrophic microbial community. *The ISME Journal* **15**, 1192–1206 (2021).
141. Herold, M. *et al*. Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nature Communications* **11**, 5281 (2020).
142. Dong, X. *et al*. Thermogenic hydrocarbon biodegradation by diverse depth-stratified microbial populations at a Scotian Basin cold seep. *Nature Communications* **11**, 5825 (2020).
143. Thompson, L. R. *et al*. Metagenomic covariation along densely sampled environmental gradients in the Red Sea. *The ISME Journal* **11**, 138–151 (2017).
144. Dominik, S., Daniela, Z., Anja, P., Katharina, R. & Rolf, D. Metagenome-assembled genome sequences from different wastewater treatment stages in Germany. *Microbiology Resource Announcements* **10**, e00504–21 (2021).
145. Langwig, M. V. *et al*. Large-scale protein level comparison of Deltaproteobacteria reveals cohesive metabolic groups. *The ISME Journal* **16**, 307–320 (2022).
146. Rezaei Somee, M. *et al*. Distinct microbial community along the chronic oil pollution continuum of the Persian Gulf converge with oil spill accidents. *Scientific Reports* **11**, 11316 (2021).
147. Gilroy, R. *et al*. Metagenomic investigation of the equine faecal microbiome reveals extensive taxonomic diversity. *PeerJ* **10**, e13084 (2022).
148. Bhattarai, B., Bhattacharjee, A. S., Coutinho, F. H. & Goel, R. K. Viruses and their interactions with bacteria and archaea of hypersaline Great Salt Lake. *Frontiers in Microbiology* **12**, 701414 (2021).
149. Liu, L. *et al*. Microbial diversity and adaptive strategies in the Mars-like Qaidam Basin, North Tibetan Plateau, China. *Environmental Microbiology Reports* (2022).
150. Lin, H. *et al*. Mercury methylation by metabolically versatile and cosmopolitan marine bacteria. *The ISME Journal* **15**, 1810–1825 (2021).
151. Martnez-Pérez, C. *et al*. Lifting the lid: nitrifying archaea sustain diverse microbial communities below the Ross Ice Shelf. *SSRN* (2020).
152. Zhang, L. *et al*. Metagenomic insights into the effect of thermal hydrolysis pre-treatment on microbial community of an anaerobic digestion system. *Science of The Total Environment* **791**, 148096 (2021).
153. Starr, E. P. *et al*. Stable-isotope-informed, genome-resolved metagenomics uncovers potential cross-kingdom interactions in rhizosphere soil. *mSphere* **6**, e00085–21 (2021).
154. Matthew, C. *et al*. Archaeal and bacterial metagenome-assembled genome sequences derived from pig feces. *Microbiology Resource Announcements* **11**, 01142–21 (2022).
155. Wang, Y., Zhao, R., Liu, L., Li, B. & Zhang, T. Selective enrichment of comammox from activated sludge using antibiotics. *Water Research* **197**, 117087 (2021).
156. Gilroy, R. *et al*. Extensive microbial diversity within the chicken gut microbiome revealed by metagenomics and culture. *PeerJ* **9**, e10941 (2021).
157. Chen, Y. H. *et al*. Salvaging high-quality genomes of microbial species from a meromictic lake using a hybrid sequencing approach. *Communications Biology* **4**, 996 (2021).
158. Beach, N. K., Myers, K. S., Donohue, T. J. & Noguera, D. R. Metagenomes from 25 low-abundance microbes in a partial nitritation anammox microbiome. *Microbiology Resource Announcements* **11**, 00212–22 (2022).
159. Solanki, V. *et al*. Glycoside hydrolase from the GH76 family indicates that marine Salegentibacter sp. Hel_I_6 consumes alpha-mannan from fungi. *The ISME Journal* **16**, 1818–1830 (2022).
160. Hiraoka, S. *et al*. Diverse DNA modification in marine prokaryotic and viral communities. *Nucleic Acids Research* **50**, 1531–1550 (2022).
161. Haryono, M.A.S. *et al*. Recovery of high quality metagenome-assembled genomes from full-scale activated sludge microbial communities in a tropical climate using longitudinal metagenome sampling. *Frontiers in Microbiology* **13** (2022).
162. Rodrguez-Ramos, J.A. *et al*. Microbial genome-resolved metaproteomic analyses frame intertwined carbon and nitrogen cycles in river hyporheic sediments. *Research Square* (2021).
163. Kim, M., Cho, H. & Lee, W. Y. Distinct gut microbiotas between southern elephant seals and Weddell seals of Antarctica. *Journal of Microbiology* **58**, 1018–1026 (2020).
164. Voorhies, A. A. *et al*. Cyanobacterial life at low O2: community genomics and function reveal metabolic versatility and extremely low diversity in a Great Lakes sinkhole mat. *Geobiology* **10**, 250–267 (2012).
165. McDaniel, E. A. *et al*. Tbasco: trait-based comparative ʻomics identifies ecosystem-level and niche-differentiating adaptations of an engineered microbiome. *ISME Communications* **2**, 111 (2022).
166. Wang, W. *et al*. Contrasting bacterial and archaeal distributions reflecting different geochemical processes in a sediment core from the Pearl River Estuary. *AMB Express* **10**, 16 (2020).
167. Mandakovic, D. *et al*. Genome-scale metabolic models of Microbacterium species isolated from a high altitude desert environment. *Scientific Reports* **10**, 5560 (2020).
168. Wang, Y. *et al*. Seasonal prevalence of ammonia-oxidizing archaea in a full-scale municipal wastewater treatment plant treating saline wastewater revealed by a 6-year time-series analysis. *Environmental Science & Technology* **55**, 2662–2673 (2021).
169. Bulzu, P. A. *et al*. Casting light on Asgardarchaeota metabolism in a sunlit microoxic niche. *Nature Microbiology* **4**, 1129–1137 (2019).
170. Karen, J. *et al*. Hydrogen-oxidizing bacteria are abundant in desert soils and strongly stimulated by hydration. *mSystems* **5**, e01131–20 (2020).
171. Rust, M. *et al*. A multiproducer microbiome generates chemical diversity in the marine sponge *Mycale hentscheli*. *Proceedings of the National Academy of Sciences* **117**, 9508–9518 (2020).
172. Podowski, J. C., Paver, S. F., Newton, R. J. & Coleman, M. L. Genome streamlining, proteorhodopsin, and organic nitrogen metabolism in freshwater nitrifiers. *mBio* **13**, e02379–21 (2022).
173. Coutinho, F. H. *et al*. New viral biogeochemical roles revealed through metagenomic analysis of Lake Baikal. *Microbiome* **8**, 163 (2020).
174. Philippi, M. *et al*. Purple sulfur bacteria fix N2 via molybdenum-nitrogenase in a low molybdenum Proterozoic ocean analogue. *Nature Communications* **12**, 4774 (2021).
175. Katie, S. *et al*. Eight metagenome-assembled genomes provide evidence for microbial adaptation in 20,000- to 1,000,000-year-old Siberian permafrost. *Applied and Environmental Microbiology* **87**, e00972–21 (2021).
176. Mert, K. *et al*. Unexpected abundance and diversity of phototrophs in mats from morphologically variable microbialites in Great Salt Lake, Utah. *Applied and Environmental Microbiology* **86**, e00165–20 (2020).
177. Patin, N. V. *et al*. Gulf of Mexico blue hole harbors high levels of novel microbial lineages. *The ISME Journal* **15**, 2206–2232 (2021).
178. Wang, J., Tang, X., Mo, Z. & Mao, Y. Metagenome-assembled genomes from *Pyropia haitanensis* microbiome provide insights into the potential metabolic functions to the seaweed. *Frontiers in Microbiology* **13**, 857901 (2022).
179. Burgsdorf, I. *et al*. Lineage-specific energy and carbon metabolism of sponge symbionts and contributions to the host carbon pool. *The ISME Journal* **16**, 1163–1175 (2022).
180. Suarez, C. *et al*. Disturbance-based management of ecosystem services and disservices in partial nitritation-anammox biofilms. *npj Biofilms and Microbiomes* **8**, 47 (2022).

181. Kumar, D. *et al*. Textile industry wastewaters from Jetpur, Gujarat, India, are dominated by Shewanellaceae, Bacteroidaceae, and Pseudomonadaceae harboring genes encoding catalytic enzymes for textile dye degradation. *Frontiers in Environmental Science* **9**, 720707 (2021).
182. Seitz, V. A. *et al*. Variation in root exudate composition influences soil microbiome membership and function. *Applied and Environmental Microbiology* **88**, e00226–22 (2022).
183. Lindner, B. G. *et al*. Toward shotgun metagenomic approaches for microbial source tracking sewage spills based on laboratory mesocosms. *Water Research* **210**, 117993 (2022).
184. Yancey, C. E. *et al*. Metagenomic and metatranscriptomic insights into population diversity of microcystis blooms: Spatial and temporal dynamics of mcy genotypes, including a partial operon that can be abundant and expressed. *Applied and Environmental Microbiology* **88**, e02464–21 (2022).
185. Liu, L. *et al*. Charting the complexity of the activated sludge microbiome through a hybrid sequencing strategy. *Microbiome* **9**, 205 (2021).
186. Speth, D. R. *et al*. Microbial communities of Auka hydrothermal sediments shed light on vent biogeography and the evolutionary history of thermophily. *The ISME Journal* **16**, 1750–1764 (2022).
187. Blyton, M. D. J., Soo, R. M., Hugenholtz, P. & Moore, B. D. Maternal inheritance of the koala gut microbiome and its compositional and functional maturation during juvenile development. *Environmental Microbiology* **24**, 475–493 (2022).
188. Nuccio, E. E. *et al*. Niche differentiation is spatially and temporally regulated in the rhizosphere. *The ISME Journal* **14**, 999–1014 (2020).
189. Jaffe, A. L. *et al*. Long-term incubation of lake water enables genomic sampling of consortia involving planctomycetes and candidate phyla radiation bacteria. *mSystems* **7**, e00223–22 (2022).
190. Cabral, L. *et al*. Gut microbiome of the largest living rodent harbors unprecedented enzymatic systems to degrade plant polysaccharides. *Nature Communications* **13**, 629 (2022).
191. Blyton, M. D. J., Soo, R. M., Hugenholtz, P. & Moore, B. D. Characterization of the juvenile koala gut microbiome across wild populations. *Environmental Microbiology* **24**, 4209–4219 (2022).
192. Xu, B. *et al*. A holistic genome dataset of bacteria, archaea and viruses of the Pearl River estuary. *Scientific Data* **9**, 49 (2022).
193. Royo-Llonch, M. *et al*. Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean. *Nature Microbiology* **6**, 1561–1574 (2021).
194. Sun, J., Prabhu, A., Aroney, S. T. N. & Rinke, C. Insights into plastic biodegradation: community composition and functional capabilities of the superworm (*Zophobas morio*) microbiome in styrofoam feeding trials. *Microbial Genomics* **8**, 000842 (2022).
195. Kim, M. *et al*. Higher pathogen load in children from Mozambique vs. USA revealed by comparative fecal microbiome profiling. *ISME Communications* **2**, 74 (2022).
196. Kelly, J. B., Carlson, D. E., Low, J. S. & Thacker, R. W. Novel trends of genome evolution in highly complex tropical sponge microbiomes. *Microbiome* **10**, 164 (2022).
197. Bray, M. S. *et al*. Phylogenetic and structural diversity of aromatically dense pili from environmental metagenomes. *Environmental Microbiology Reports* **12**, 49–57 (2020).
198. Cabello-Yeves, P. J. *et al*. α-cyanobacteria possessing form IA RuBisCO globally dominate aquatic habitats. *The ISME Journal* **16**, 2421–2432 (2022).
199. Berben, T. *et al*. The Polar Fox Lagoon in Siberia harbours a community of Bathyarchaeota possessing the potential for peptide fermentation and acetogenesis. *Antonie van Leeuwenhoek* **115**, 1229–1244 (2022).
200. Tamburini, F. B. *et al*. Short- and long-read metagenomics of urban and rural South African gut microbiomes reveal a transitional composition and undescribed taxa. *Nature Communications* **13**, 926 (2022).
201. Kantor, R. S., Miller, S. E. & Nelson, K. L. The water microbiome through a pilot scale advanced treatment facility for direct potable reuse. *Frontiers in Microbiology* **10**, 993 (2019).
202. Muratore, D. *et al*. Complex marine microbial communities partition metabolism of scarce resources over the diel cycle. *Nature Ecology & Evolution* **6**, 218–229 (2022).
203. Zhou, Y. L., Mara, P., Cui, G. J., Edgcomb, V. P. & Wang, Y. Microbiomes in the challenger deep slope and bottom-axis sediments. *Nature Communications* **13**, 1515 (2022).
204. Zhang, H. *et al*. Metagenome sequencing and 768 microbial genomes from cold seep in South China Sea. *Scientific Data* **9**, 480 (2022).
205. Zhuang, J. L., Zhou, Y. Y., Liu, Y. D. & Li, W. Flocs are the main source of nitrous oxide in a high-rate anammox granular sludge reactor: insights from metagenomics and fed-batch experiments. *Water Research* **186**, e116321 (2020).
206. Shiffman, M. E. *et al*. Gene and genome-centric analyses of koala and wombat fecal microbiomes point to metabolic specialization for eucalyptus digestion. *PeerJ* **5**, 4075 (2017).
207. Murphy, S. M. C., Bautista, M. A., Cramm, M. A. & Hubert, C. R. J. Diesel and crude oil biodegradation by cold-adapted microbial communities in the Labrador Sea. *Applied and Environmental Microbiology* **87**, e00800–21 (2021).
208. Suarez, C. *et al*. Metagenomic evidence of a novel family of anammox bacteria in a subsea environment. *Environmental Microbiology* **24**, 2348–2360 (2022).
209. Dharamshi, J.E. *et al*. Genomic diversity and biosynthetic capabilities of sponge-associated chlamydiae. *The ISME Journal* (2022).
210. Florian, P. O., Hugo, R. & Mathieu, A. Recovery of metagenome-assembled genomes from a human fecal sample with pacific biosciences high-fidelity sequencing. *Microbiology Resource Announcements* **11**, e00250–22 (2022).
211. Bloom, S. M. *et al*. Cysteine dependence of *Lactobacillus iners* is a potential therapeutic target for vaginal microbiota modulation. *Nature Microbiology* **7**, 434–450 (2022).
212. Aylward, F. O. *et al*. Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. *Proceedings of the National Academy of Sciences* **114**, 11446–11451 (2017).
213. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**, 1043–1055 (2015).
214. Bowers, R. M. *et al*. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature biotechnology* **35**, 725 (2017).
215. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
216. Louca, S. The rates of global bacterial and archaeal dispersal. *ISME Journal* **16**, 159–167 (2021).
217. Ondov, B. D. *et al*. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biology* **17**, 132 (2016).
218. Müllner, D. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software* **53**, 1–18 (2013).
219. Kinene, T., Wainaina, J., Maina, S., Boykin, L.M. & Kliman, R.M. *Methods for rooting trees*, vol. 3, 489–493 (Academic Press, Oxford, 2016).
220. Louca, S. & Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* **34**, 1053–1055 (2018).
221. Rees, J. A. & Cranston, K. Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal* **5**, e12581 (2017).
222. Heck, K. *et al*. Evaluating methods for purifying cyanobacterial cultures by qPCR and high-throughput Illumina sequencing. *Journal of Microbiological Methods* **129**, 55–60 (2016).

223. Cornet, L. *et al*. Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLOS ONE* **13**, e0200323 (2018).
224. Alneberg, J. *et al*. Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome* **6**, 173 (2018).
225. Eddy, S. R. Accelerated profile HMM searches. *PLoS Computational Biology* **7**, e1002195 (2011).
226. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60 (2014).
227. Pedregosa, F. *et al*. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

Acknowledgements

S.L. was supported by a startup grant by the University of Oregon and by an Alfred P. Sloan Research Fellowship.

Author contributions

Both authors contributed to the data compilation, analyses and manuscript writing. Correspondence and requests for materials should be addressed to S.L.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-023-01994-7.

Correspondence and requests for materials should be addressed to S.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.