Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Meta-analysis of the agreement and accuracy of functional inference from 16S amplicon studies in arsenic-affected soils and aquifers

Md Rakib Hasan

rakibhhridoy@fermium.systems

University of Dhaka https://orcid.org/0009-0002-4007-7590

Shoumik Zubyer

University of Dhaka https://orcid.org/0009-0006-4085-1493

Mst Anika Khatun Rupa

University of Dhaka https://orcid.org/0009-0008-4347-4288

Mohammad Fazle Alam Rabbi

University of Dhaka

Anwar Zahid

University of Dhaka https://orcid.org/0000-0002-0456-6760

Research Article

Keywords: Functional inference; 16S rRNA gene amplicon; PICRUSt2; Shotgun metagenomics; Meta-analysis; Arsenic cycling; Public sequence archives; Groundwater

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11639131/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Article *Preprint*

Meta-analysis of the agreement and accuracy of functional inference from 16S amplicon studies in arsenic-affected soils and aquifers

Md Rakib Hasan [1,2] ✉ Shoumik Zubyer [1] Mst Anika Khatun Rupa [1] Mohammad Fazle Alam Rabbi [1] Anwar Zahid [1,3]

Background. Almost all archived environmental sequence data are 16S amplicon, so gene content is inferred from lineage rather than measured, yet inferred profiles are routinely read as measurements. Whether the reliability of inference differs between genes, and whether the predictor's own confidence metric identifies the unreliable ones, has not been tested at archive scale. Here, 3,041 archived 16S amplicon runs from 78 studies of rice-paddy soil, arsenic-contaminated soil and arsenic-affected aquifers were re-analysed through one pipeline. Gene content was inferred by a curated marker-taxa bridge and by PICRUSt2, their agreement was quantified per gene and tested against candidate predictors, and both routes were checked against paired shotgun metagenomes. Compartments were compared under depth adjustment and study-level permutation.
Results. Agreement between the routes varied widely by gene, from $\rho = 0.906$ (*mcrA*) to $\rho = -0.118$ (*arsM*), around a weak median of $\rho = 0.286$ across 698 gene-study comparisons in 57 studies. PICRUSt2's confidence metric did not explain that range, whereas agreement tracked how consistently the marker taxa carried each gene in PICRUSt2's reference genomes ($\rho = +0.66$). FAPROTAX, a published lookup, matched PICRUSt2 no better than the bridge, whereas Tax4Fun2, a second genome-based predictor, matched it more closely ($\rho = +0.43$ against $\rho = +0.29$). Sequencing depth lowered inferred abundance in 15 of 20 genes, while its effect on detection was positive for every gene. Against shotgun metagenomes from 267 biosamples in 16 studies, PICRUSt2 tracked measured abundance weakly (median $\rho = +0.21$), Tax4Fun2 no better, and the bridge barely ($\rho = +0.05$). Agreement between the bridge and PICRUSt2 predicted the accuracy of the bridge but not of PICRUSt2. No compartment difference was resolved after adjustment.
Conclusions. Agreement between amplicon-based inference routes differs by gene within a single panel. Neither the predictor's confidence metric nor phylogenetic conservation identifies the discordant genes, while carriage within the marker taxa partly anticipates them. Disagreement between independent routes flags method-dependent genes, but agreement does not certify accuracy. The routes disagree on *arsM*, one of only two arsenic-cycling genes both score. Compartment differences in arsenic-cycling potential cannot be resolved from public amplicon data without shotgun metagenomes.

Keywords: Functional inference; 16S rRNA gene amplicon; PICRUSt2; Shotgun metagenomics; Meta-analysis; Arsenic cycling; Public sequence archives; Groundwater

[1]Department of Soil, Water and Environment, University of Dhaka, Dhaka-1000, Bangladesh. [2]Fermium Systems, Dhaka-1207, Bangladesh. [3]Institute of Water Modelling, Dhaka, Bangladesh. ✉e-mail: rakibhhridoy@fermium.systems

Background

Arsenic in shallow groundwater exposes tens of millions of people across South and Southeast Asia to a known carcinogen, and the mechanism usually held responsible is biological. Microbes respire iron oxyhydroxides, the mineral surfaces that hold arsenic in place, and the arsenic those minerals carry passes into solution as they dissolve [1,2]. Other organisms then methylate, oxidise or respire the arsenic itself, so what reaches a drinking well depends on which of these processes prevails [3]. Each step rests on strong work at individual field sites, and whether the same functional signature appears wherever arsenic mobilises has never been tested on uniformly processed data.

Public sequence archives appear to make that test straightforward. Hundreds of studies of arsenic-affected soils and aquifers are deposited, and processing them through one pipeline with one reference database and one gene panel should show whether the pattern generalises. Re-analysing archived data this way is now ordinary practice in microbial ecology [4–6], and a survey of arsenic-related genes across soil metagenomes has already shown what the approach can deliver where genes are sequenced directly [7].

The obstacle is that almost everything deposited is 16S amplicon, which records which organisms are present rather than what they can do. Gene content must therefore be inferred, and two families of method dominate. A curated lookup maps genera to the functions they are documented to carry. A phylogenetic predictor places sequence variants on a reference tree and infers gene abundances from their neighbours [8]. Other predictors, Tax4Fun2 among them, assign sequence variants to reference genomes differently but share that dependence on them [9], so a curated lookup is the more independent second route. Both are widely used, and once a profile has been computed both are routinely read as though they measured what they only estimate [10]. Each rests on the premise that related organisms do related things [11,12], and functions differ enormously in how far that premise holds.

Where predictions have been checked against shotgun metagenomes from the same samples, the check has covered tens of samples from a few environments, and it has found that accuracy falls sharply outside the human gut and that a correlation between predicted and measured gene profiles is easy to obtain and hard to interpret [8,13]. Accuracy also differs between community types within a single body site [14], and the variation in gene content among related genomes, one cause of the departure of predicted from measured profiles, has been corrected by calibrating against a small set of paired samples [15]. A later evaluation of four predictors against matched metagenomes and simulated communities found them too insensitive to recover functional differences between groups of samples [16], and reviews of the field identify the shortage of reference genomes outside the human gut, in soils especially, as the central limit [17]. The reference behind the most widely used predictor has since been rebuilt from current genomes, but its evaluation rested on simulated communities and on correlation across genes within a sample [18]. The marker gene adds uncertainty of its own through limited taxonomic resolution and variable copy number [19,20]. No such check exists at the scale of an archive, because it needs paired shotgun data, which the archive rarely holds. The archive does allow a test of internal consistency, and that test is answerable with the same corpus assembled to ask the biological one. Two inference routes that share only the sequence-variant table can be run over the same samples and compared gene by gene, and where they disagree, at least one is wrong. This study runs that comparison over 3,041 archived runs from 78 studies of rice-paddy soil, arsenic-contaminated non-paddy soil and arsenic-affected aquifers.

Table 1. Retained amplicon corpus by compartment. Counts are generated directly from the analysis sample sheet.

Compartment	Runs	Studies
Rice-paddy soil	2,210	42
Shallow aquifer (<50 m)	34	2
Deep aquifer (≥50 m)	27	1
Aquifer, depth unspecified	104	7
As-contaminated soil (non-paddy)	666	27
Total	**3,041**	**78**

The study asks four questions of that corpus and of a separate set of paired amplicon and shotgun samples. (i) How closely do a curated marker-taxa lookup, PICRUSt2 and the published lookup FAPROTAX agree, gene by gene, when run over the same sequence variants? (ii) Does the phylogenetic conservation of each gene's marker taxa, the predictor's own confidence metric or the consistency with which the marker taxa carry the gene explain which genes they agree about? (iii) How closely does each route track gene content measured by shotgun sequencing of the same biosamples, and does agreement between the routes indicate which of them is accurate? (iv) Once sequencing depth is accounted for, does inferred arsenic-cycling potential differ between paddy soil, contaminated soil and aquifer? The answers bear on how much weight a functional profile computed from archived amplicon data can carry, for arsenic cycling and beyond.

Methods

The design is outlined in Fig. 1, from the archive search through the three inference routes to the comparisons of route agreement, accuracy against measured gene content and compartment.

Dataset selection and corpus

A structured search of public sequence archives, executed on 20 May 2026, returned 12,114 records that reduced to 7,539 after deduplication. That query is a reproducible basis for dataset selection and not an exhaustive census of the field. Explicit inclusion rules applied to the deduplicated set produced 177 candidate datasets spanning 5,271 sequencing runs. Compartment was resolved from metadata held in the European Nucleotide Archive [21] and not from study titles, since a title records what a paper is about while the metadata record what was actually sequenced. A compartment label therefore records the kind of site sampled and not a measured arsenic concentration. Library layout was read from the deposited file listing rather than the declared field, and single-end Illumina runs, 23.8 % of the final corpus, were denoised without read merging (Supplementary Section S2).

Compartments

Rice-paddy soil and aquifer water were defined at the outset as the two settings in which arsenic mobilisation is most often studied. Arsenic-contaminated non-paddy soil was added as a third arm once the available datasets showed that it, and not the aquifer, was the only compartment with room to expand. It is kept separate from paddy soil so that the gene-prevalence comparison can be run three ways. Table 1 gives the retained corpus.

Exclusion criteria and read depth

Every exclusion is recorded at run level with its cause (Table 2). Platforms whose error profiles the Illumina denoising path cannot model were removed, as were datasets failing the isolation-source and study-design criteria, such as enrichment cultures, arsenic-removal filters and root tissue. Rice rhizosphere soil was retained, because arsenic is mobilised in the rhizosphere of a flooded paddy. Three further categories were found only after retrieval (Supplementary Sections S2 and S14).

A sample-level floor of 10,000 retained reads is applied at consolidation, and depth is carried as a covariate in every compartment model, because 38.8 % of aquifer samples fall below the floor against 11.4 % for paddy soil. Much of that shortfall arises in processing, because primers are removed from the 5′ end only and reads are not truncated (Supplementary Section S2). Aquifer datasets were stratified at 50 m below ground surface where a depth was reported.

Table 2. Run-level exclusion ledger.

Reason for exclusion	Runs
Failed inclusion criterion (isolation source)	211
Constant quality scores: error model not estimable	139
Failed inclusion criterion (study design)	111
Incompatible platform: 454 pyrosequencing	70
Incompatible platform: Ion Torrent	64
Size outlier: shotgun-scale deposit labelled amplicon	30
Failed denoising (no R1/R2 overlap)	27
Incompatible platform: PacBio SMRT	23
Total excluded	**675**

Reference data and software

Taxonomy was assigned against the SILVA release 138.2 genus-level training set [22], with the GTDB r220 fallback [23] disabled for the reported run. The functional panel is a frozen set of 35 genes spanning the arsenic, iron, sulfur, nitrogen and carbon cycles, held at version 1.4 with file checksums recorded. Runs were retrieved with sra-tools [24], primers removed with cutadapt [25], and sequences denoised and assigned with DADA2 [26] and DECIPHER [27] under R [28], scheduled by Snakemake [29]. The arsenic-soil arm used bit-identical versions of these tools, so the three arms rest on the same analytical basis (Supplementary Table S8).

Functional inference and the analysis set

Amplicon data carry taxonomy, not gene content, so panel genes were attributed to samples through a curated marker-taxa bridge (hereafter the bridge) that maps genera to the genes for which they carry documented evidence. Each attribution is weighted by the strength of that evidence, so the per-sample score is an evidence-weighted relative abundance and not a gene copy number. Where a route has no value for a sample the value is left empty and never zero.

A second and independent route was computed by PICRUSt2 [8], which places sequence variants on a reference phylogeny and predicts KEGG ortholog abundances from their neighbours. Version 2.6.3 was run with its default reference, PICRUSt2-SC, which holds 27,870 genomes from GTDB release r214 [18], so the predictions are those a current user obtains without changing a setting. The two routes share only the sequence-variant table, so their agreement measures the inference rather than the data. PICRUSt2 output exists for 67 of the 78 retained studies (Supplementary Section S5), so the primary gene matrix rests on the bridge alone and the comparison between routes is restricted to the studies where both exist and the gene maps to a KEGG ortholog. *arsM* and *aioA* enter as AS3MT (K07755) and aoxB (K08356), and *hydA* as the four [FeFe]-hydrogenase orthologs K17997, K00532, K00533 and K18332, giving 698 gene-study comparisons across 57 studies and 13 genes. Agreement is the Spearman correlation between the two per-sample scores, taken within a study so that laboratory and primer effects are held constant, and summarised as a median across studies for each gene. The correlation runs across samples for one gene, because a correlation across genes within a sample is inflated by the abundance structure all genes share and survives permutation of samples [13].

A third route tested whether the behaviour of the bridge is common to curated lookups. FAPROTAX 1.2.12 [11], the published taxon-to-function database most used in environmental surveys,

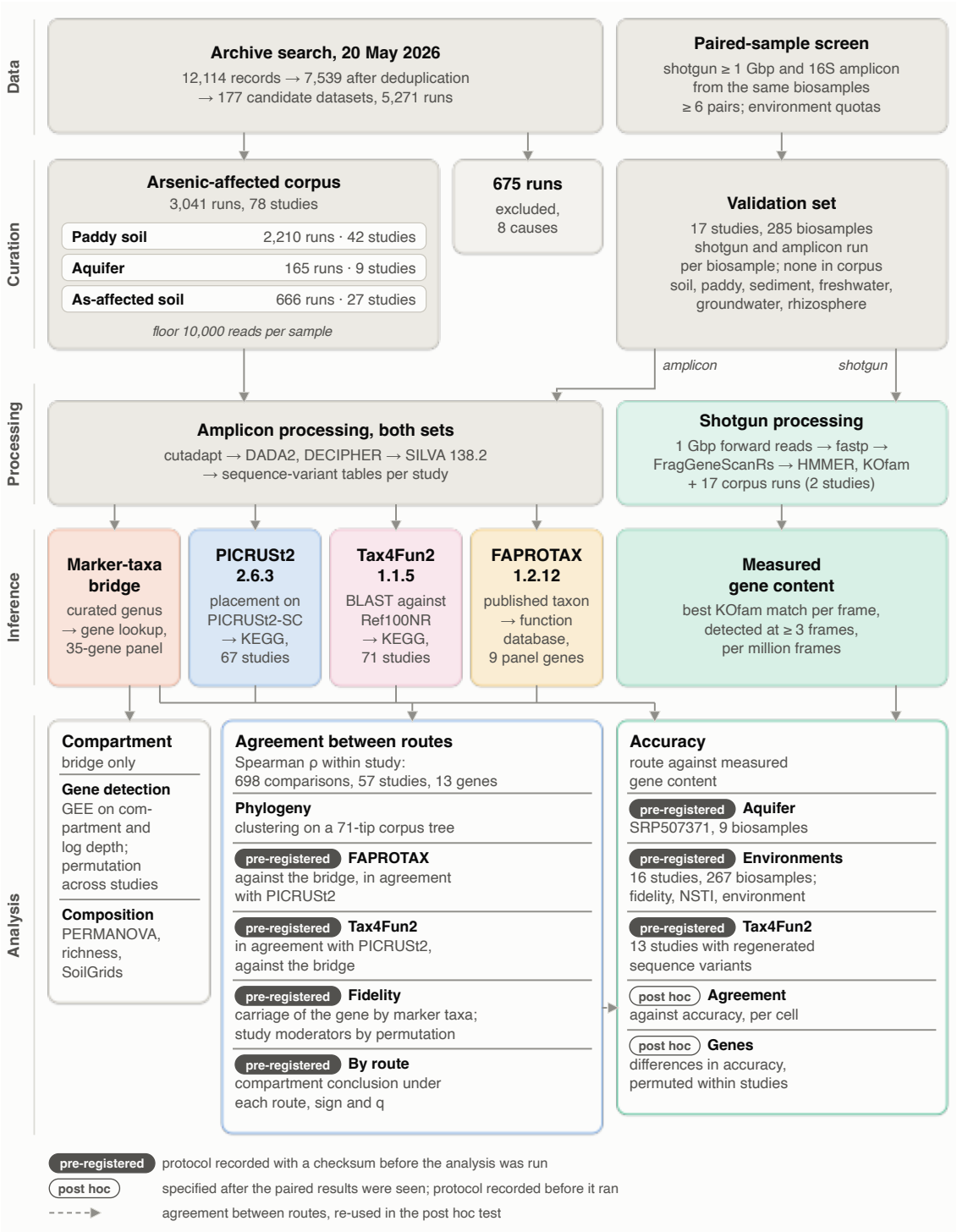


Fig. 1. Study design. The arsenic-affected corpus (left) and a validation set of paired shotgun and amplicon biosamples from other environments (right) pass through the same amplicon processing and the same four inference routes. Gene content measured in shotgun metagenomes enters only the accuracy comparison, and the compartment analysis uses the marker-taxa bridge alone. Filled tags mark analyses fixed in a protocol recorded with a checksum before they were run, and outlined tags the two specified after the paired results were seen. Counts are those of the retained corpus (Table 1) and the validation set (Supplementary Table S15b).

was run unmodified over every taxonomy string in the corpus. A sample's score for a gene is the relative abundance of the sequence variants assigned to the function that gene defines, which exists for 9 panel genes. FAPROTAX and the bridge are compared by the paired difference in their agreement with PICRUSt2, with an interval resampling studies (Supplementary Section S11).

A fourth route tested whether the weak agreement belongs to curated lookups or to inference from 16S data in general. Tax4Fun2 1.1.5[9], the other widely used genome-based predictor, assigns each sequence variant to its nearest reference 16S sequence by BLAST at 97 % identity or more and takes that genome's ortholog profile, rather than placing it on a tree. It was run with its default Ref100NR reference and settings on the sequence variants and counts given to PICRUSt2, for the corpus and for the paired studies of 'Paired amplicon and shotgun data across environments', and a gene's score is the summed share of its KEGG orthologs in the predicted profile, on the scale of the PICRUSt2 score. Agreement with the other routes, accuracy in the deep aquifer and accuracy across environments were computed on the same cells and with the same intervals as for PICRUSt2, under a protocol whose checksum was recorded before any Tax4Fun2 output existed. Because the two predictors draw

on overlapping reference genomes, that protocol fixed in advance that agreement between them would not be read as accuracy. A sample in which no read reaches a reference receives an empty profile and scores zero as registered, and leaving such samples unscored is reported as a sensitivity analysis added after the results were seen (Supplementary Section S16).

Whether agreement depends on how a study was conducted was tested on each study's mean departure from the per-gene median agreement, against aquifer or soil origin, read layout, sample number and median retained reads, each by permutation over studies with Holm correction (Supplementary Section S12). Whether the routes lead to the same compartment conclusion was tested by regressing each study's log median score under each route on compartment and median $\log_{10}$ retained reads, with compartment labels permuted across studies. The routes were then compared on the sign of each contrast and on the contrasts significant under one route only (Supplementary Section S13).

Confidence intervals on headline correlations and medians are 95 % percentile intervals from 10,000 bootstrap resamples with a fixed seed. Each resamples the independent unit of its summary, which is the study for median agreement, the nearest-sequenced-taxon test and the comparison across environments, the gene for the conservation test and the biosample for the comparison within one aquifer. Detection of a gene in a sample is a non-zero marker-taxa score, and its prevalence in a study is the fraction of that study's samples in which it is detected.

Measured gene content from paired shotgun metagenomes

Neither inference route is a measurement, so both were checked against gene content measured directly wherever the archive allowed it. Two corpus studies deposited shotgun metagenomes from the same biosamples as their amplicon libraries, giving 17 shotgun runs from 13 biosamples. In the deep-aquifer study SRP507371, 9 of its 11 pairs have an amplicon run above the read-depth floor and enter the correlation. The second study holds two biosamples, too few for a correlation, and is described in Supplementary Section S10.

Forward reads were filtered with fastp 1.4.0 [30] and gene-called with FragGeneScanRs 1.1.0 [31], a reimplementation of FragGeneScan [32] with the same gene models. Reading frames were searched with HMMER 3.4 [33] against KOfam profiles [34] for the 26 panel genes that map to a KEGG ortholog, *hydA* by K17997 alone. *arsM* and *aioA* were counted in a second pass, with curated HMMs for the arsenic genes [7] as a cross-check. KOfam's adaptive thresholds are calibrated on full-length proteins, and no short read reaches them for 21 of the 26 panel orthologs. Each reading frame was therefore assigned to the panel ortholog it matches best at $E < 10^{-3}$, a gene counts as detected at three or more assigned frames, and abundance is assigned frames per million frames called. That rule was fixed before any per-gene result was examined, and three sensitivity analyses varied it (Supplementary Section S10). Accuracy is the Spearman correlation across biosamples between each route and the measured abundance, gene by gene.

Paired amplicon and shotgun data across environments

The corpus holds no other study with enough paired biosamples, so the comparison was repeated on studies selected for the purpose across environments and not for arsenic, under a protocol recorded with a checksum before any sequence was downloaded. A study was eligible where Illumina shotgun metagenomes and bacterial 16S amplicon libraries came from at least six shared biosamples and its record described no virome, stable-isotope, culture or incubation work. A deterministic draw with quotas by environment gave 285 biosamples from 17 studies of soil, paddy soil, sediment, freshwater, groundwater and rhizosphere, none of them in the corpus, and 267 biosamples in 16 studies remained after one study failed PICRUSt2's alignment check (Supplementary Section S15). Amplicon libraries passed through the same pipeline with PICRUSt2 held at 2.6.3, and shotgun runs were processed as for the aquifer from the first 1 Gbp of forward reads, with every panel ortholog searched.

For each gene and study in which a route and the measured abundance both vary across at least five biosamples, accuracy is their Spearman correlation. Each route is summarised as the median over those cells with a 95 % interval resampling studies, and the routes are compared by the paired difference on the cells both reach. Secondary analyses fixed in the same protocol related accuracy to carriage fidelity ('Carriage of panel genes within the marker taxa') and to the weighted nearest-sequenced-taxon index, and described accuracy by environment. Two further tests were specified after these results were seen, each in a protocol recorded with a checksum before it ran, and are reported as post hoc. The first correlates agreement between the routes with each route's accuracy across cells. The second tests for differences in accuracy between genes with a Kruskal–Wallis statistic against 10,000 permutations of gene labels within studies.

Phylogeny of the marker taxa

Whether agreement follows phylogenetic conservation was tested on a tree built from this corpus, one representative sequence variant per marker genus, aligned with MAFFT [35] and inferred with FastTree [36], with 71 genera entering the tree. Clustering of each gene's marker taxa was measured as their mean pairwise phylogenetic distance standardised against 999 random draws of the same number of tips, and was correlated with per-gene agreement (Supplementary Section S4).

Carriage of panel genes within the marker taxa

For every gene and each of its marker taxa, the share of that taxon's genomes in the PICRUSt2-SC reference that carry the gene's KEGG ortholog was computed, with taxa matched by name in the NCBI taxonomy and, as a sensitivity analysis, in GTDB. Taxa with fewer than three reference genomes were left out, and the mean share across a gene's remaining marker taxa is termed here its carriage fidelity. It was correlated with per-gene median agreement, with a one-sided permutation *p*-value from 100,000 permutations and a bootstrap interval over genes, and within genes against each study's agreement. The quantities and tests were fixed in a protocol recorded with a checksum before any of them was computed (Supplementary Section S12).

Composition, soil covariates and sample attributes

Bray-Curtis dissimilarity [37] between genus profiles was compared across compartments by PERMANOVA [38] with labels permuted across studies, and richness and Shannon diversity were regressed on compartment with $\log_{10}$ retained reads as a covariate. Soil properties were taken from SoilGrids v2.0 [39] at 0–5 cm for geolocated samples and related to community distance within studies by Mantel and partial Mantel correlations. The full attribute record of every run was searched for arsenic concentrations and for laboratory manipulation (Supplementary Section S14).

Statistical analysis

Compartment contrasts were tested at study level, since samples within a study share a laboratory, a primer choice and a sequencing run. For each gene, per-sample detection was regressed on compartment and $\log_{10}$ retained reads by generalized estimating equations with a binomial family, a logit link and an exchangeable working correlation clustered by study. The compartment coefficient is then the log-odds of detection between compartments at equal depth. Cluster-robust errors are anti-conservative with fewer than ten aquifer studies, so inference rests on 2,000 permutations of compartment labels across studies, with odds-ratio intervals from the Mancl–DeRouen bias-reduced covariance. Analyses used Python 3.11 with pandas [40], SciPy [41]

and statsmodels[42], and the code is available as described under Availability of data and materials.

Results

Composition and quality of the archived corpus

The retained corpus holds 3,041 amplicon runs from 78 studies across the three compartments (Table 1). Of the 675 excluded runs, the size outliers, the runs with a constant quality score and the upstream failures were found only by inspecting the sequences (Table 2, Fig. 2).

The largest problem was taxonomic, since 11 of 92 studies, holding 578 otherwise retained runs, assigned under 20 % of their sequence variants to Bacteria or Archaea. These are eukaryotic amplicon data deposited under metadata that declares neither, and failure to detect the expected primer in the reads flags every one of them. Five further studies reported spiking, enrichment culture, reactor growth or laboratory incubation in attributes that a search of isolation source does not reach. Arsenic itself is largely unrecorded, with usable concentrations for 58 of 3,041 runs across 3 studies, so the exposure variable the corpus was built around cannot be used as a covariate. PICRUSt2 predicts more detoxification than metabolism genes in 63 of 63 studies, by a median factor of 8.1, consistent with soil metagenomes[7] but still a prediction.

Agreement between independent inference routes and its relation to phylogenetic conservation and gene carriage

Across 698 gene-study comparisons spanning 57 studies and 13 genes, the two agree at a median Spearman $\rho = 0.286$ (95 % CI +0.212 to +0.348, resampling studies). Agreement is positive in 71 % of comparisons and only one gene has a negative median, so the routes are not generating noise against each other. They exceed $|\rho| = 0.5$ in only 38 % of comparisons, however, so a profile computed by one cannot be read as though the other would have produced it (missing cells in Supplementary Section S5).

Agreement spreads widely between genes, being highest for *mcrA* ($\rho = 0.906$), which marks methanogenesis, and lowest for *arsM* ($\rho = -0.118$), the arsenite methyltransferase, whose median is the only negative one in the panel. Within a gene, studies with more samples agree no more closely (median within-gene $\rho = +0.04$, $p = 0.40$), and neither do studies in which both routes detect the gene more often ($\rho = -0.05$, $p = 0.25$; permutation of studies within genes, Fig. 3b,c). Nor does study design predict how closely a study's routes agree, with no association between a study's mean departure from the per-gene medians and aquifer origin ($p = 0.17$, from only 6 aquifer studies), library layout, sample number or read depth across 57 studies (Supplementary Section S12).

Conservation is the premise of any taxonomy-based route, so the routes might be expected to agree wherever a function is phylogenetically conserved, and a tree of the 71 marker genera tests this directly (Fig. 4). The marker taxa of 7 of 15 genes are more closely related than chance, so the lookup captures real phylogenetic structure, but that structure does not predict agreement (Spearman $\rho = -0.27$, 95 % CI -0.896 to +0.495, $p = 0.42$). With only 11 genes the interval cannot exclude a relationship, but the decisive case runs the wrong way, since *mcrA*, on which the routes agree most closely, has the most widely dispersed marker taxa of any gene tested (standardised effect size +2.35).

PICRUSt2's own measure of extrapolation, the nearest-sequenced-taxon index[43], which is the phylogenetic distance from a sequence variant to the closest reference genome, fares no better. Weighted by abundance and averaged per study, it varies 39.3-fold across the 57 studies carrying both routes (0.044 to 1.713) but does not predict agreement ($\rho = +0.111$, 95 % CI -0.197 to +0.397, $p = 0.41$; Supplementary Section S9). The interval excludes a correlation stronger than +0.397, too weak at most to screen a corpus on, and SRP406096, with the least favourable index of any study, shows the closest agreement between routes (+0.611). A predictor's own confidence therefore does not indicate whether an independent route will reproduce it.

A third property of the marker taxa is associated with agreement. In PICRUSt2's reference genomes the mean share of a gene's marker-taxon genomes that carry it runs from 0.99 for *mcrA* to 0.31 for *arsM* and 0.14 for *aioA* (Supplementary Table S12). Across the 13 genes this carriage fidelity correlates with agreement at $\rho = +0.66$ (95 % CI +0.14 to +0.94, one-sided permutation $p = 0.008$; Fig. 5a), and at +0.81 under GTDB matching, although without *mcrA* it falls to $\rho = +0.57$ (-0.00 to +0.91), so a single gene carries much of the evidence. Fidelity alone is not sufficient, as *feoB* shows, carried by a share of 0.94 of marker-taxon genomes but agreeing poorly because only 1.3 % of its carriers lie inside its marker taxa, and the product of fidelity and that coverage, specified as secondary, correlates with agreement at $\rho = +0.75$ (+0.28 to +0.95; Fig. 5b). Within genes, studies whose marker taxa carry the gene more consistently do not agree measurably better ($\rho = +0.05$, $p = 0.34$), so the relationship belongs to the gene and its lookup and not to the study.

hydA is scored by all four [FeFe]-hydrogenase orthologs in KEGG, because its originally frozen ortholog, K17997, is carried by only 1 of 175 *Clostridium* genomes, the only *hydA* taxon in the bridge, so under it the two routes scored different enzymes. The change was made after the first analysis and recorded in the protocol, and no conclusion depends on it (Supplementary Section S13).

FAPROTAX, a published lookup, behaves like the bridge. It reaches 8 of the 9 genes with a corresponding function and none of the arsenic genes, since it lists its arsenite oxidisers at species level and no sequence variant in the corpus is classified to those species. Across the 274 of 478 cells in 54 studies where its scores vary, it agrees with PICRUSt2 at a median $\rho = +0.35$ (95 % CI +0.27 to +0.48), against $\rho = +0.39$ (+0.29 to +0.47) for the bridge on the same cells, a paired difference of +0.001 (+0.000 to +0.035; Supplementary Table S11). Both medians exceed the full-grid value because these cells exclude *arsM*, *feoB* and every cell without a FAPROTAX taxon. The two lookups share few genera (median Jaccard overlap 0.08) yet agree with each other at $\rho = +0.45$ (+0.34 to +0.54), so the weak agreement with PICRUSt2 is common to curated lookups at genus resolution and not a property of this one.

Tax4Fun2, a second genome-based predictor, separates lookups from predictors. It scores 71 of the 78 retained studies, missing only those whose sequences are not bacterial 16S, although in the median study 75 % of reads reach no reference at 97 % identity (Supplementary Section S16). Across 694 cells in 57 studies it agrees with PICRUSt2 at a median $\rho = +0.43$ (95 % CI +0.36 to +0.49), against +0.29 (+0.21 to +0.34) for the bridge, a paired difference of +0.09 (+0.03 to +0.13), while its agreement with the bridge (+0.28) and with FAPROTAX (+0.31) is as weak as theirs with PICRUSt2. The weak agreement is therefore a difference between genus lookups and genome-based predictors, and because both predictors draw on reference genomes their closer agreement carries no information on which route is accurate.

Only part of the arsenic cycle can be compared. *arxA* and *arrA* are reachable only through the bridge, the first lacking a KEGG ortholog and the second mapping to one that postdates the PICRUSt2 reference set, and the detoxification operon, whose horizontally transferred genes carry no stable taxonomic signal, only through PICRUSt2. *arsM* and *aioA* are reached by both and fall at opposite ends of the comparison, *aioA* agreeing at a median $\rho = +0.22$ across 56 studies and *arsM* at $\rho = -0.12$, positive in only 41 % of studies. That order runs against conservation,

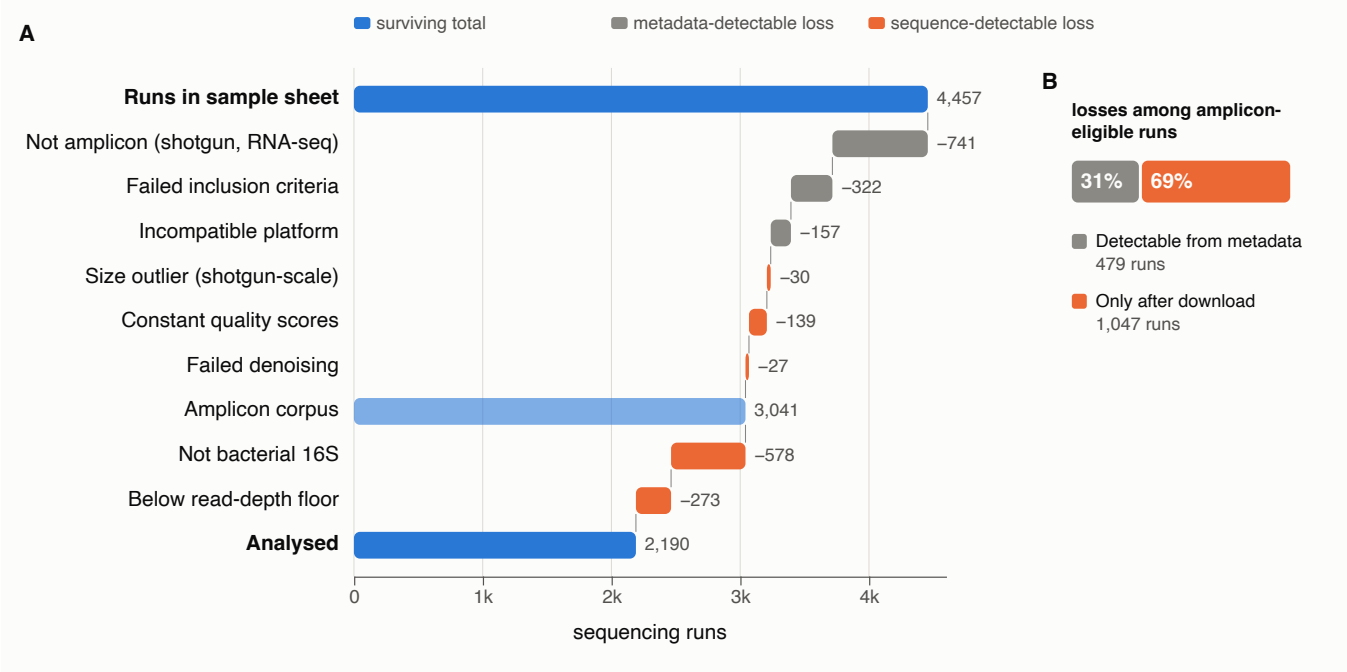


Fig. 2. Corpus attrition. (**a**) Runs surviving each screen, with each loss positioned at the running total it reduces. The read-depth floor counts only samples not already removed as non-bacterial, since 261 fail both. (**b**) Share of losses detectable from metadata and only after download, among runs eligible to be amplicon data. Run-level ledger in Supplementary Table S2.

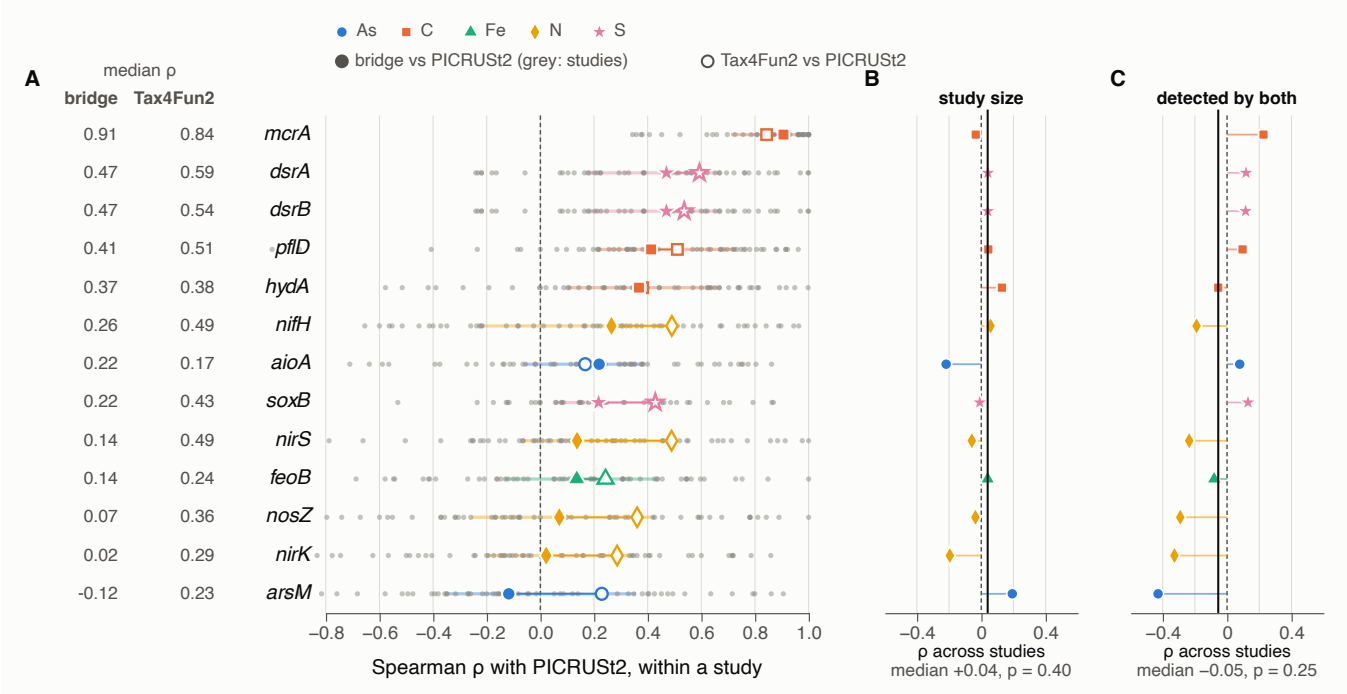


Fig. 3. Agreement between two independent routes from amplicon data to gene content. (**a**) Per-gene Spearman ρ between the marker-taxa and PICRUSt2 routes, taken within a study. Grey points are studies, the filled mark the median, the bar the interquartile range; the open mark is the median agreement of Tax4Fun2 with PICRUSt2 over the same studies where it scores the gene. (**b**, **c**) For each gene, Spearman ρ across studies between agreement and study size (**b**) or the fraction of samples in which both routes detect the gene (**c**); the open mark is the median over genes, tested by permuting studies within genes. Full per-gene values in Supplementary Table S5.

since across soil genomes the presence of *arsM* follows phylogeny while that of *aioA* does not[7], although *arsM* has also been transferred horizontally, repeatedly, in its history[44].

At the level of the contrasts a study would report, the two routes agree more often than their sample-level correlation suggests. Across the 39 gene-contrast cells comparable under both, they give the same direction of difference between compartments in 35 (90 %, binomial $p < 0.001$), and their standardised effects correlate at $\rho = +0.62$ (95 % CI +0.36 to +0.81). The conclusion still depends on the route in 6 cells, 5 significant under PICRUSt2 alone and 1 under the bridge alone, the paddy-aquifer contrast for *aioA* ($q = 0.034$ against $q = 0.32$ under PICRUSt2), and no cell is significant under both routes in opposite directions. These q-values are corrected within the comparable cells only, so they measure how far the choice of route changes a conclusion and are not offered as compartment differences (Supplementary Section S13).

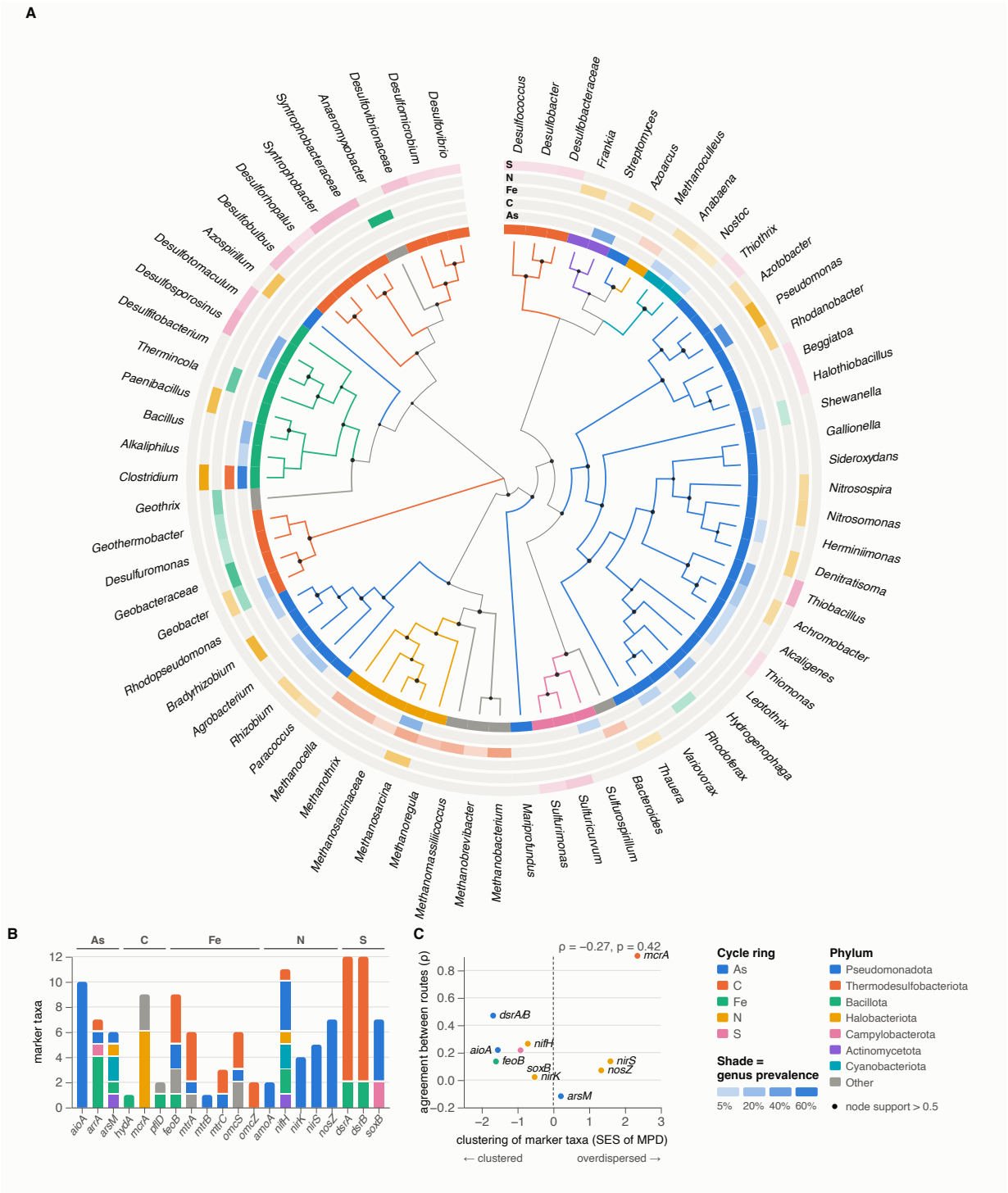


Fig. 4. Phylogeny of the marker taxa the functional bridge relies on. (**a**) Tree of the 71 marker genera. Branch and inner-ring colour denote phylum, node circles local support above 0.5, and the five outer rings the cycles each genus is mapped to, shaded by its prevalence in the corpus. (**b**) Phyla contributing each gene's marker taxa, genes grouped by cycle. (**c**) Standardised effect size of mean pairwise phylogenetic distance against agreement between routes. Tree construction in Supplementary Section S4.

Comparison with gene content measured by shotgun sequencing

In the deep-aquifer study SRP507371 the two routes agree only weakly (median $\rho = +0.28$ across its genes), and shotgun metagenomes from the same biosamples show which of them is closer to the measured gene content (Supplementary Table S10). Across the 9 pairs, PICRUSt2 tracks the measured abundance at a median Spearman $\rho = +0.73$ (95 % CI +0.37 to +0.84, resampling biosamples), positive for 17 of 21 genes and significant for 11, while the bridge reaches +0.00 (−0.54 to +0.56), positive for 6 of 15 genes. Across every sensitivity analysis of the assignment rule and read-depth floor the PICRUSt2 median stays between +0.62 and +0.73 and the bridge median between +0.00 and +0.05. The bridge interval is wide, so the comparison rules out close agreement for the bridge but not a moderate one, and a rank correlation tests whether a route orders the biosamples as measured, not whether it recovers absolute abundance.

Two checks added after the results were seen argue against a property shared by all genes in a sample, such as the two well depths sampled, since PICRUSt2's prediction for one gene does not track the measured abundance of another (median $\rho = +0.00$ across mismatched pairs) and the agreement holds within the

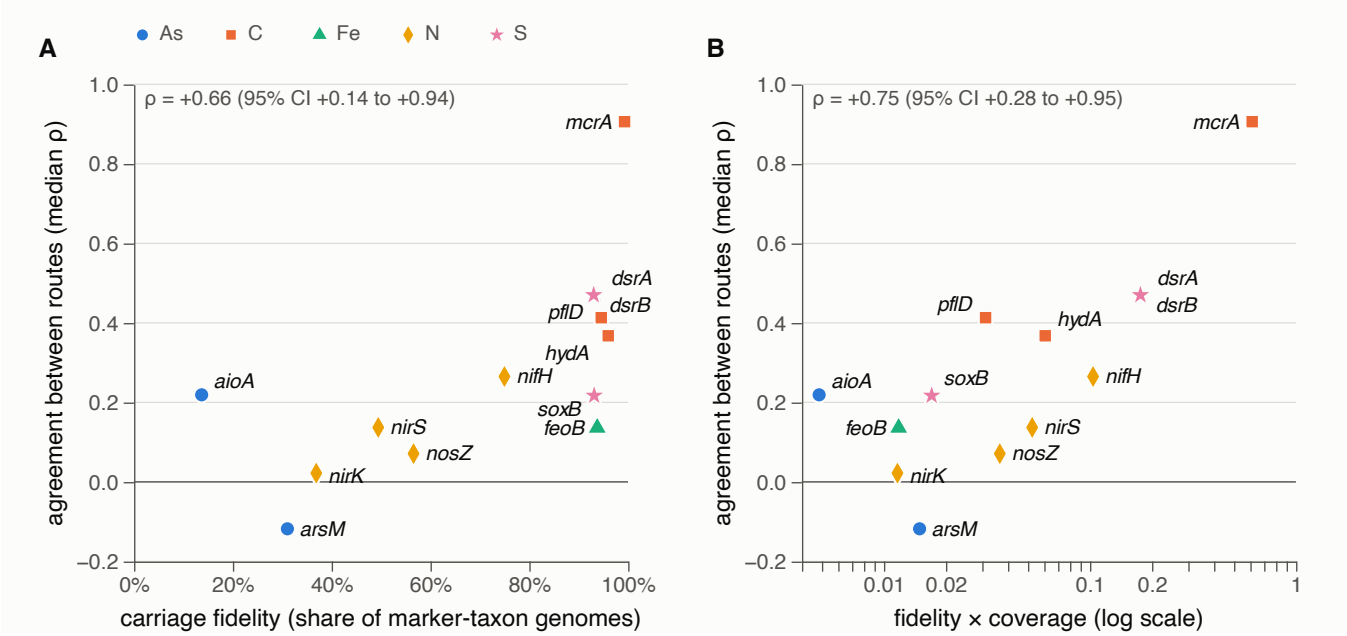


Fig. 5. Agreement between routes follows how consistently the marker taxa carry each gene. (**a**) Median agreement between the bridge and PICRUSt2 against carriage fidelity, the mean share of a marker taxon's PICRUSt2-SC reference genomes that carry the gene's KEGG ortholog, one point per gene. (**b**) The same against fidelity multiplied by coverage, the share of all carrier genomes that lie inside the gene's marker taxa, on a logarithmic axis. Intervals are bootstrap 95 % CIs over genes. Values in Supplementary Table S12.

wells at 160 m ($\rho = +0.70$, $n = 5$) and at 80 m ($\rho = +0.80$, $n = 4$), although groups that small carry little weight.

FAPROTAX varies across the nine biosamples for only 4 genes, which it ranks against measured abundance at a median $\rho = +0.24$ (95 % CI -0.31 to +0.72), between PICRUSt2 (+0.43) and the bridge (-0.02) on the same genes, and four genes cannot separate it from either. Tax4Fun2 shows no detectable accuracy in this aquifer, with a median $\rho = +0.20$ across 24 genes (95 % CI -0.05 to +0.59, resampling biosamples) against +0.73 for PICRUSt2 across its 21 genes, and only *dsrB* ($\rho = +0.73$, $p = 0.025$) and *pflD* ($\rho = +0.73$, $p = 0.025$) reach significance.

For the arsenic genes the results are mixed. PICRUSt2 exceeds $\rho = 0.5$ for 3 of the 6 detoxification genes it reaches, and the bridge runs against the measured abundance of *arrA* ($\rho = -0.54$, and -0.70 under the length-scaled threshold that this molybdopterin-family gene needs; Supplementary Section S10). For *arsM*, on which the routes disagree across the corpus, PICRUSt2 follows the measured abundance at $\rho = +0.50$ ($p = 0.17$, and +0.65 under the stricter cutoff) while the bridge does not ($\rho = -0.03$) and scores *arsM* absent in 3 of the 9 biosamples where the shotgun data detect it. Neither route tracks *aioA* ($\rho = +0.13$ for PICRUSt2, and from -0.42 to +0.48 for the bridge depending on the rule). With 9 biosamples none of these correlations is significant, so they indicate where the *arsM* disagreement lies without establishing it (Supplementary Table S10b), and presence could not be tested because every measurable gene is detected in every biosample.

Accuracy across environments

Across 16 further studies the aquifer result holds in direction but not in size. PICRUSt2 tracks measured gene content at a median $\rho = +0.21$ (95 % CI +0.16 to +0.29, resampling studies) over 375 gene-study cells, positive in 267 and significant in 66, while the bridge reaches $\rho = +0.05$ (-0.01 to +0.15) over 237 cells (Fig. 6a). On the 182 cells both routes reach, PICRUSt2 exceeds the bridge by a median of +0.10 (+0.02 to +0.18). Its median is positive in 14 of the 16 studies and the bridge's in 10, with PICRUSt2's study medians running from -0.05 to +0.70 (Supplementary Table S15b), so the deep aquifer lies at the favourable end of the range and not at its centre. PICRUSt2 reaches +0.39 in the 3 groundwater studies and +0.17 in the 8 soil studies, figures that are descriptive with one to eight studies per environment. Measurement does not limit the comparison, since 15 of the 29 genes are detected in every shotgun run.

Tax4Fun2 tracks measured gene content across environments about as closely as PICRUSt2. Its sequence variants were regenerated by rerunning the unchanged pipeline, and the 13 of 17 studies that reproduced their stored count tables exactly were analysed, the other 4 being set aside as the protocol required. In those studies Tax4Fun2 reaches a median $\rho = +0.17$ (95 % CI +0.09 to +0.25), against +0.23 (+0.15 to +0.29) for PICRUSt2 and +0.01 (-0.07 to +0.10) for the bridge (Fig. 6a). Its paired difference is -0.03 (-0.09 to +0.03) from PICRUSt2 and +0.11 (+0.01 to +0.21) from the bridge, so neither genome-based predictor is detectably closer to measured gene content than the other and both are closer than the genus lookup, a result unchanged when samples in which no read reaches a reference are left unscored (post hoc; Supplementary Section S16).

Neither property examined above as a guide to reliability anticipates accuracy. Carriage fidelity, which tracks agreement between the routes, does not order the 13 genes by the accuracy of PICRUSt2 ($\rho = -0.12$, 95 % CI -0.82 to +0.54) or of the bridge ($\rho = +0.20$, -0.43 to +0.70), and studies with a less favourable nearest-sequenced-taxon index are not detectably less accurate ($\rho = -0.27$, -0.72 to +0.33), with every interval wide enough to admit a moderate relationship in either direction. In a test specified after these results were seen, PICRUSt2's accuracy also does not differ detectably between genes (Kruskal–Wallis $p = 0.36$ under permutation within studies), while under the bridge the difference is marginal ($p = 0.06$), its lowest genes being *nosZ* (-0.21) and *aioA* (-0.12, -0.44 to -0.02; Supplementary Table S15). For *arsM* PICRUSt2 reaches a median of +0.31 against +0.03 for the bridge, the order seen in the aquifer.

Agreement between the routes is no guide to PICRUSt2's accuracy, and in a second post hoc test it runs the other way. Across the 182 cells in which both routes and the measured abundance vary, agreement between the bridge and PICRUSt2 correlates negatively with PICRUSt2's accuracy ($\rho = -0.20$, 95 % CI -0.33 to -0.04) and positively with the bridge's ($\rho = +0.30$, +0.15 to +0.44; Fig. 6b,c). The bridge association holds within genes ($\rho = +0.31$,

+0.17 to +0.44) while the PICRUSt2 association weakens to $\rho = -0.13$ (-0.25 to +0.02), and shuffling the measured values within each cell removes both (two-sided $p = 0.005$ and $p < 0.001$ over 2,000 shuffles), so they do not arise merely because the three correlations share their samples. Agreement therefore marks the cells in which the genus lookup happens to track the measurement, not those in which the predictor does.

Determinants of gene detection by the bridge

The bridge scores a gene present in a sample whenever any of its marker genera is present, so the detection rate of a gene is set by how common those organisms are. Across 20 genes, the prevalence of a gene's single most common marker genus predicts its detection rate at Spearman $\rho = 0.84$ ($p = 4 \times 10^{-6}$, Fig. 7a), as that construction implies, so the result describes the instrument rather than the communities.

The commonest marker genera are ubiquitous soil taxa and not arsenic specialists, *Anaeromyxobacter* occurring in 64 % of samples and *Clostridium*, *Pseudomonas* and *Bradyrhizobium* in roughly half, each mapped to several panel genes (Fig. 7b).

Opposing effects of sequencing depth on detection and abundance

The depth coefficient for gene detection is positive for all 20 genes, as deeper sequencing pushes more marker taxa above the threshold at which they are recorded as present, although only 1 is significant on its own. Depth predicts gene abundance negatively in 15 of 20 genes (positively, and significantly, only for *aioA*), at a median $\beta = -0.399$, consistent with rarer taxa recovered at greater depth diluting the marker fraction, although that cause was not tested directly (Fig. 8). Moving from prevalence to abundance therefore does not escape the confounding, it reverses its sign, and a study reporting one statistic will contradict a study reporting the other from the same data.

Richness shows the same dependence, rising by 242 genera for every tenfold increase in reads ($p < 10^{-90}$). Depth is also unbalanced across compartments, with 38.8 % of aquifer and 32.9 % of arsenic-soil samples below the 10,000-read floor against 11.4 % of paddy samples, so sequencing effort tracks the grouping under test.

Compartment contrasts, composition and soil covariates

Scored by the bridge, with depth controlled and labels permuted across studies, no difference between any pair of compartments is resolved for any gene (Supplementary Table S7 and Figure S14). The null reflects limited resolution rather than equivalence, since the 95 % interval on the odds ratio spans a median 11-fold range across the 60 gene-contrast pairs. The strongest result is *aioA* between paddy soil and aquifer (odds ratio 22.04, $q = 0.154$), abundance in place of detection yields 0 genes after correction, and the leading candidates change identity and direction under the read-depth floor (Supplementary Section S14).

Community composition separates the compartments no better, with study-level PERMANOVA[38] on Bray-Curtis distance[37] giving $p = 0.305$ across the three groups and $p = 0.773$ for paddy soil against aquifer. Within studies, environmental distance from modelled soil properties correlates with community distance (median Mantel $r = +0.119$), but partialling out geographic distance reduces it to $r = +0.002$ ($p = 0.84$), while dissimilarity rises with distance in 18 of 23 studies. These communities are therefore structured in space, and neither the compartment label nor the measured soil environment captures that structure.

Discussion

The central result is a specific one about method. Two unrelated routes for inferring gene content from 16S amplicon data agree closely for one gene of a single panel and not at all for another. Against measured gene content the better of them tracks it only modestly, and their agreement does not show where it does. A functional profile is therefore not one measurement of uniform quality but a set of estimates whose consistency differs from gene to gene, presented in a single table with nothing to distinguish them, and that consistency is not a measure of their accuracy.

Phylogenetic conservation and gene carriage

Phylogenetic conservation, the natural explanation for that spread, does not account for it. Functional traits are conserved to depths that differ systematically by trait[45], and reading a phylogeny for traits rests on that regularity[46,47], so agreement might be expected to follow conservation across a tree of the marker taxa. Most panel genes have marker taxa more closely related than chance, but that clustering does not order the agreement, although with only 11 genes the interval (-0.896 to +0.495) cannot exclude a real relationship. The decisive cases run the wrong way, since the routes agree on *aioA*, which carries no phylogenetic signal across soil genomes, and not on *arsM*, which does[7]. They also agree most closely on the gene with the most dispersed marker taxa of any tested.

A different reading of the spread is supported, if on a small panel. *mcrA* catalyses the final step of methanogenesis and is carried by essentially every methanogen[48], so for this gene a genus-level lookup and a genome-based predictor both reduce to the abundance of methanogens, however widely those are dispersed on the tree. Agreement then depends on how exactly a gene's carriers can be named at the rank the lookup uses, which is a different property from how tightly they cluster. In PICRUSt2's own reference genomes that carriage fidelity runs from near-universal for *mcrA* to under a third for *arsM*, and it correlates with agreement across the panel ($\rho = +0.66$, 95 % CI +0.14 to +0.94). A gene carried by only some members of a marker genus is scored by the lookup wherever the genus occurs and by the predictor only where the placed neighbours carry it, so the two routes diverge by construction. A gene that defines its carriers is scored alike by both. The interval is wide and spans zero once *mcrA* is removed, so the evidence is limited by the panel, but it turns the ordering of the panel into an expectation that can be computed from reference genomes before any sample is analysed. That expectation concerns agreement and not accuracy, because agreement grounded in carriage shows that both routes count the same organisms and not that either counts the gene. The high agreement for *mcrA* should therefore not be read as evidence that inference performs well in general. Measured gene content bears this out, since across 16 studies fidelity does not order the genes by the accuracy of either route ('Accuracy across environments').

Confidence metric, route agreement and accuracy

Agreement is not governed by the predictor's own confidence either, since PICRUSt2's nearest-sequenced-taxon index varies 39.3-fold across these studies and is unrelated to how far an independent route reproduces its predictions ($\rho = +0.111$, 95 % CI -0.197 to +0.397, $p = 0.41$). Across 57 studies spanning a 39.3-fold range of the index, that interval excludes a strong association with cross-route agreement but not a modest one, and a correlation no stronger than +0.397 would be too weak to screen a corpus on. Against measured gene content in 16 studies the index runs in the expected direction without reaching significance ($\rho = -0.27$, -0.72 to +0.33). The relationship on which it was introduced, accuracy against shotgun data falling as the index rises[43], is therefore neither confirmed nor excluded at this number of studies. Neither candidate explanation survives, though for different reasons, since a strong association with the software's own metric is excluded while conservation is left unsupported by a test the gene panel is too small to settle. Carriage fidelity in reference genomes offers a per-gene expectation of agreement, but none of the three offers an expectation of accuracy against measured

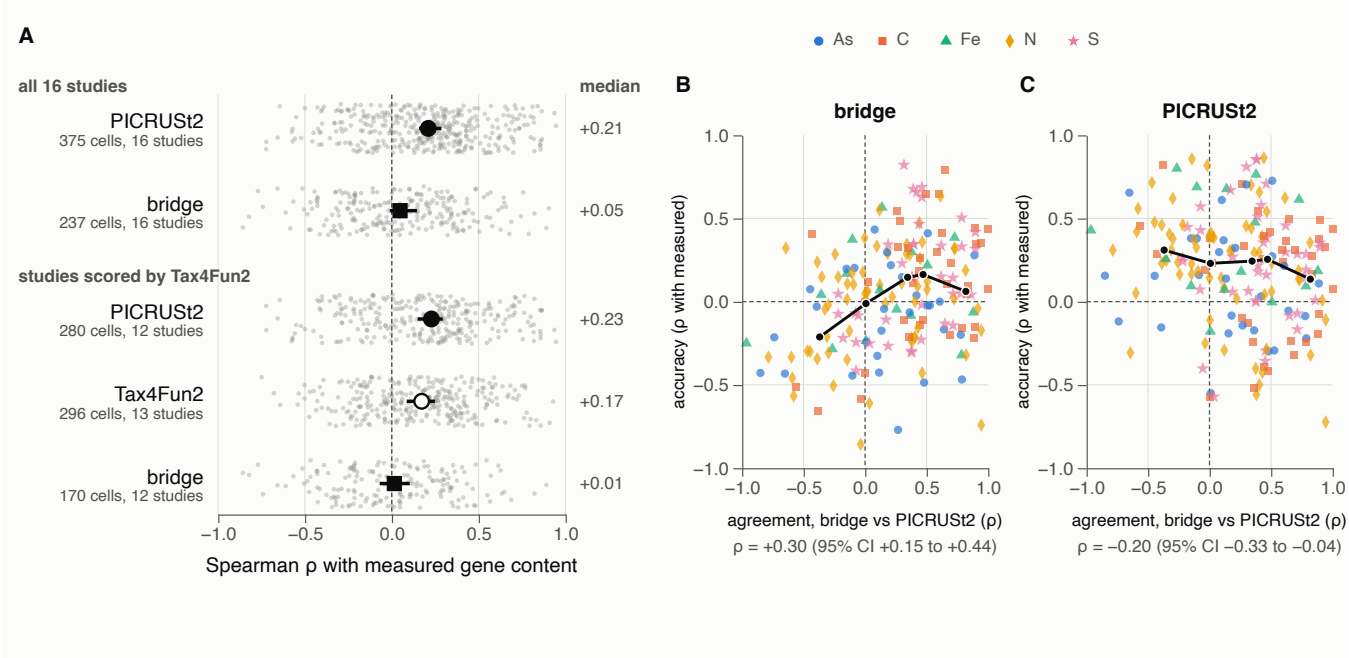


Fig. 6. Agreement between routes does not certify accuracy. (**a**) Spearman ρ between each route and gene content measured in paired shotgun metagenomes, one grey point per gene-study cell, with the median and its bootstrap 95 % CI. Upper block, all 16 studies; lower block, the 13 studies whose regenerated sequence variants Tax4Fun2 scores, on which the three routes are compared directly. (**b**, **c**) Agreement between the bridge and PICRUSt2 against the accuracy of the bridge (**b**) and of PICRUSt2 (**c**), one point per cell, coloured by cycle; the black line joins the medians of quintiles of agreement. Panels b and c were specified after the paired results were seen. Values in Supplementary Tables S15 and S16.

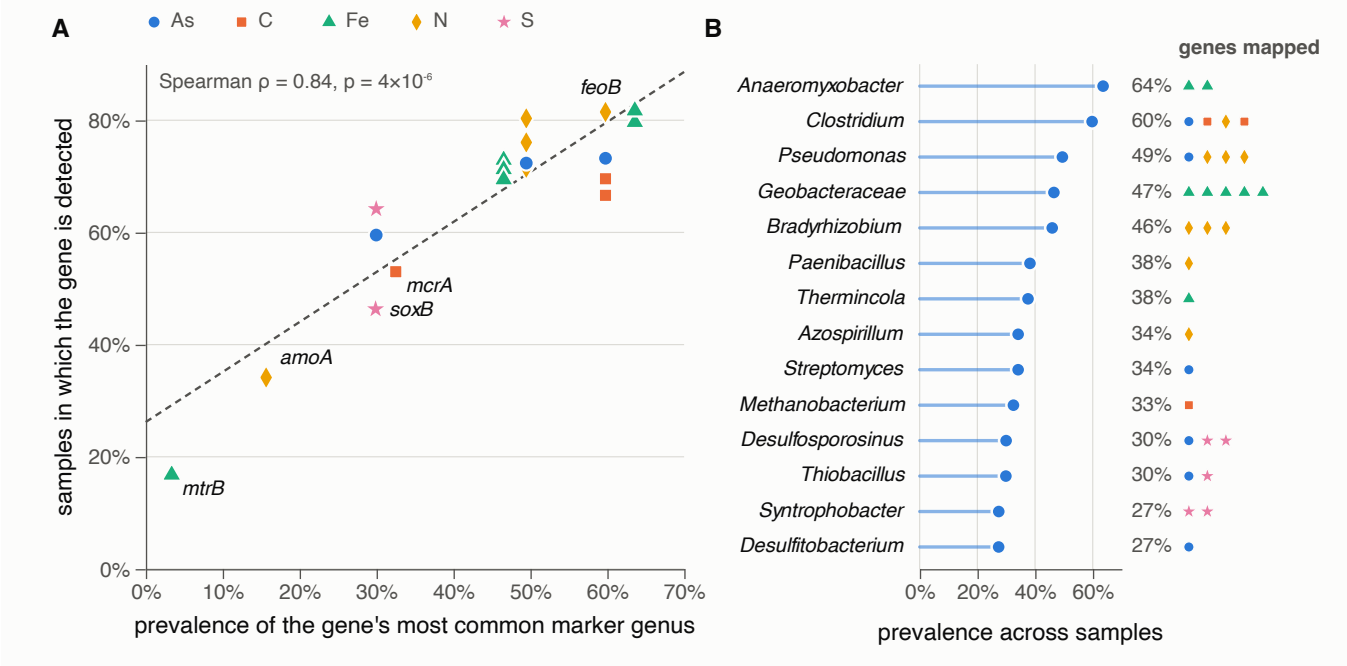


Fig. 7. What the bridge measures. (**a**) Detection rate of each panel gene against the prevalence of its single most common marker genus, one point per gene. (**b**) The most prevalent marker genera in the curated lookup, with one mark per panel gene each is mapped to, coloured and shaped by cycle. The lookup itself is Supplementary Table S3.

gene content. Comparing two independent routes does not supply one either, since both read function from lineage and a bias they share would surface as agreement and not as disagreement. The comparison does supply a one-sided test that any panel and any corpus can run, in which disagreement marks a gene whose inferred profile depends on the method and should not be interpreted without independent evidence. Agreement withholds that verdict without granting the opposite. Measured gene content sharpens the second half of that statement, because agreement predicts the bridge's accuracy and not PICRUSt2's, which is lower where the routes agree than where they part. The routes therefore agree where the genus lookup happens to track the measurement, and agreement is better read as a property of the cell than as evidence about the predictor. Shotgun data supply the missing evidence where they exist, and in the deep aquifer PICRUSt2 ranked biosamples in the measured order (median $\rho = +0.73$) and the bridge did not ($\rho = +0.00$). Across 16 further studies of soil, sediment, freshwater, groundwater and rhizosphere PICRUSt2 kept a smaller but consistent advantage (median +0.21 against +0.05), and Tax4Fun2 matched it within the precision of the comparison (paired difference -0.03, -0.09 to +0.03). Disagreement between the routes therefore points to the bridge more often than to in-

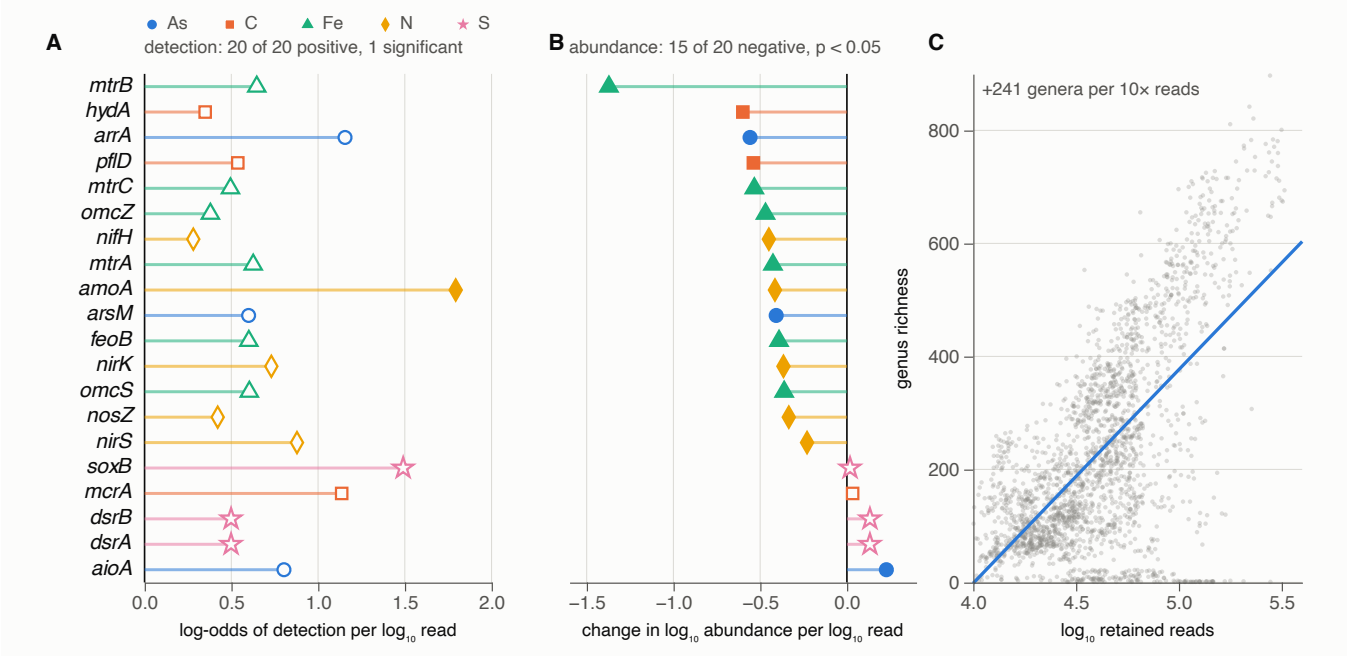


Fig. 8. Sequencing depth biases detection and abundance in opposite directions. (**a**) Depth coefficient for gene detection, in log-odds per $\log_{10}$ read. (**b**) Depth coefficient for gene abundance among samples where the gene is detected, in $\log_{10}$ abundance per $\log_{10}$ read. The two panels share a gene ordering but not a scale. Filled marks indicate $p < 0.05$. (**c**) Genus richness against depth, one point per sample. Per-gene coefficients in Supplementary Table S6.

ference as such, although outside that aquifer PICRUSt2 ranks samples only loosely by the gene content measured in them. The consequence for a reported contrast is narrower than the sample-level figures suggest, since the routes give the same direction for most compartment contrasts. The one arsenic contrast either route would call significant, *aioA* between paddy soil and aquifer, is still called so only by the bridge. The bridge departed from measured gene content where both routes could be checked, and its *aioA* scores ran against measured abundance across the paired studies (median -0.12).

These predictions come from PICRUSt2-SC, the reference distributed with the software since version 2.6.0. Its developers evaluated it on simulated communities matched to seven datasets, five of them host-associated and none from an aquifer, scoring accuracy as the correlation across genes within each sample[18]. A correlation of that kind survives permutation of samples[13], so the across-sample agreement and the shotgun comparison reported here are, to the authors' knowledge, the first independent tests of the current default reference on environmental samples. The weak median agreement therefore applies to the reference a present-day analysis obtains by default and not to a superseded one.

Genus-level lookups and gene content

The weak agreement belongs to genus-level lookups in general and not to the curated one alone. A lookup assembled by hand from roughly a hundred genera is a thin instrument, but FAPROTAX, built independently and sharing few genera with the bridge, agrees with PICRUSt2 no better than the bridge does on the same cells. Tax4Fun2 agrees with PICRUSt2 more closely than either lookup does, so the divide runs between lookups and genome-based predictors, although their shared reference genomes keep that agreement from counting as accuracy. FAPROTAX reaches none of the arsenic genes, which leaves the bridge as the only curated route to them. The lookup is thin, and the second result shows why that matters beyond this particular lookup. A gene is scored present wherever one of its marker organisms occurs, so its detection rate is largely a restatement of how common those organisms are, and the commonest of them are ordinary soil bacteria found in most samples, not arsenic specialists. Prevalence of a single genus predicts detection of its gene at $\rho = 0.84$, as the construction of the bridge implies, and set against measured gene content in 16 studies the bridge tracked it at a median of only +0.05. Any method that infers function from genus identity inherits this, and a richer lookup would postpone the problem instead of removing it, because a larger marker set makes detection saturate sooner.

Sequencing depth and compartment contrasts

Sequencing depth compounds the problem in a way that has no statistical remedy within this data type. That relative abundances are compositional and that depth cannot be normalised away by rarefying are established[49,50], and the panel adds the sign reversal. Depth raises detection and lowers abundance, so a study that switches statistics to avoid the confounding merely changes its sign, and two analyses of the same data can reach opposite conclusions according to which they report. Depth is also unbalanced across the groups being compared, because low-biomass subsurface samples are harder to sequence deeply for reasons unconnected to arsenic. Its effect on diversity is larger still, at 242 genera per tenfold increase in reads, which exceeds any compartment difference in this corpus by a wide margin.

A second objection is that permuting across studies is too severe a test with eight aquifer studies. The contrast failed to reach significance before any adjustment, its leading genes changed identity and direction under defensible quality gates, and composition and diversity returned the same result independently of the gene panel, so stronger inference would not have rescued it.

Soil properties correlate with composition within studies only until geography is partialled out, while distance-decay remains strong, so these communities are structured in space in a way that neither the compartment label nor standard soil covariates capture. A metagenomic survey of paddy soils along an arsenic gradient reached the same conclusion from measured genes, with geographic distance rather than arsenic concentration structuring the communities and the DNA abundance of arsenic genes unchanged while their transcription responded[51]. A study that attributes composition to pH or texture without a spatial control

risks reporting the same artefact.

Quality of the archived corpus

The corpus also carried more contamination than its metadata revealed, with 11 studies that were not bacterial 16S at all and five more that reported laboratory manipulation in fields a search of isolation source does not reach. None of this is visible before download and all of it is detectable afterwards in minutes, which argues for treating an assembled public corpus as a hypothesis about its own contents. Usable arsenic concentrations appear for only 58 of 3,041 runs, in a corpus assembled to study arsenic cycling, which kept the exposure of interest out of every model.

Limitations and implications

Four caveats apply. The comparison of routes covers 13 genes, because a gene must be reachable by both. *arxA*, *arrA* and the detoxification genes each fall to one route only, so most of the arsenic genes of greatest interest lie beyond it, a property of the reference resources and not of this corpus. The aquifer arm comprises eight studies after quality gating, which is everything the search located. Neither route is a measurement, and no prediction from 16S data can speak to a gene on mobile elements whose distribution departs from lineage, however large the corpus. Finally, accuracy was measured in studies of at most 20 biosamples each, so every per-gene correlation rests on few samples and the per-gene and per-environment values are descriptive. The two tests that bear most on interpretation, of agreement against accuracy and of differences between genes, were specified after the primary result was seen, and shotgun DNA records gene content and not expression. PICRUSt2's accuracy was lower in soil (median +0.17) than in groundwater (+0.39), consistent with the lower accuracy reported for soils[13].

Shotgun sequencing would settle the biological question, since genes are observed rather than predicted there and the arsenic metabolism genes that the amplicon routes reach poorly or not at all become directly countable against curated references[7,52]. This corpus contains 455 shotgun runs from 32 studies, of which only the 17 paired with amplicon libraries from the same biosamples were processed here. Together with 285 paired runs from 17 studies outside the corpus, they served to calibrate the inference and not to answer the biological question, which the remainder would address directly. Until they are processed, the defensible statement is that compartment differences in arsenic-cycling potential are not identifiable from public amplicon data, and reports of such differences should be read alongside the sequencing depth of each group and the phylogenetic breadth of each gene claimed.

Conclusions

Functional profiles inferred from archived 16S amplicon data differ from gene to gene in how far independent routes agree, and nothing in the profile itself shows the difference. Across 57 studies the routes agree closely on *mcrA* and not at all on *arsM*, PICRUSt2's confidence metric does not identify the genes on which they part, and the carriage of each gene within its marker taxa anticipates agreement but not accuracy. Against gene content measured in 16 studies PICRUSt2 tracks it weakly and the bridge barely, so disagreement between routes marks genes whose inferred profiles depend on the method, while agreement is no substitute for measurement. Sequencing depth raises detection and lowers abundance, and belongs in the model as a covariate. Compartment differences in arsenic-cycling potential cannot be resolved from public amplicon data, and the shotgun metagenomes already in the archive are the route to that question.

Abbreviations

CI: confidence interval; ENA: European Nucleotide Archive; FAPROTAX: Functional Annotation of Prokaryotic Taxa; GTDB: Genome Taxonomy Database; GTR: general time-reversible substitution model; HMM: hidden Markov model; ITS: internal transcribed spacer; KEGG: Kyoto Encyclopedia of Genes and Genomes; OR: odds ratio; PERMANOVA: permutational multivariate analysis of variance; PICRUSt2: Phylogenetic Investigation of Communities by Reconstruction of Unobserved States, version 2; SH: Shimodaira–Hasegawa; SMRT: single-molecule real-time; SRA: Sequence Read Archive; V3, V4: hypervariable regions 3 and 4 of the 16S rRNA gene.

Ethics approval and consent to participate

Not applicable. This study analysed only publicly archived sequence data and involved no human participants, no identifiable human data and no animals.

Consent for publication

Not applicable.

Availability of data and materials

No new sequence data were generated. Every run analysed here is publicly archived and is listed individually, with its study accession and study title, in Supplementary Table S1. Runs were retrieved from the Sequence Read Archive using accessions resolved against the European Nucleotide Archive[21,24]. The run-level exclusion ledger, giving the reason each excluded run was removed, is Supplementary Table S2, and the curated marker-taxa lookup on which the bridge rests is Supplementary Table S3. The 285 biosamples of the comparison across environments are listed with their shotgun and amplicon run accessions in the deposit, and their studies in Supplementary Table S15b. Soil covariates were obtained from SoilGrids v2.0[39]. The analysis pipeline, the curated lookup, the gene panel manifest and the scripts that generate every number, table and figure in this article and its supplementary material, together with the per-study inputs those scripts read and the shotgun search tables, are available at the Zenodo deposit, version 1.3 https://doi.org/10.5281/zenodo.23804867, which also holds the per-study Tax4Fun2 results. Files that have not changed are held in the version that first deposited them, the per-run results of the comparison across environments in version 1.2 https://doi.org/10.5281/zenodo.23696363 and the per-study inputs and shotgun search tables of the arsenic-affected corpus in version 1.1 https://doi.org/10.5281/zenodo.23436429. FAPROTAX 1.2.12 was obtained from its authors and is not redistributed in the deposit. Tax4Fun2 1.1.5 and its reference data were obtained from a Zenodo mirror of the original distribution (record 10035668).

Authors' contributions

Md Rakib Hasan: conceptualization, methodology, software, formal analysis, investigation, data curation, writing (original draft), writing (review and editing), visualization. Shoumik Zubyer: data curation, writing (review and editing). Mst Anika Khatun Rupa: investigation, data curation, writing (review and editing). Mohammad Fazle Alam Rabbi: supervision, writing (review and editing). Anwar Zahid: supervision, contribution to study design and hydrogeological interpretation, writing (review and editing). All authors approved the final version.

Competing interests

The authors declare no competing interests.

Funding

This research received no external funding.

Acknowledgements

This work rests entirely on data deposited by other research groups. The authors thank the investigators of the 78 studies whose archived sequence data made this analysis possible, and the staff of the Sequence Read Archive and the European Nucleotide Archive who maintain the infrastructure it depends on.

References

[1] Farhana S. Islam, Andrew G. Gault, Christopher Boothman, David A. Polya, John M. Charnock, Debashis Chatterjee, and Jonathan R. Lloyd. Role of metal-reducing bacteria in arsenic release from Bengal delta sediments. *Nature*, 430 (6995):68–71, July 2004. doi: 10.1038/nature02638.

[2] Scott Fendorf, Holly A. Michael, and Alexander van Geen. Spatial and Temporal Variations of Groundwater Arsenic in South and Southeast Asia. *Science*, 328(5982):1123–1127, May 2010. doi: 10.1126/science.1172974.

[3] Jie Qin, Barry P. Rosen, Yang Zhang, Gejiao Wang, Sylvia Franke, and Christopher Rensing. Arsenic detoxification and evolution of trimethylarsine gas by a microbial arsenite *S*-adenosylmethionine methyltransferase. *Proceedings of the National Academy of Sciences*, 103(7):2075–2080, February 2006. doi: 10.1073/pnas.0506836103.

[4] Claire Duvallet, Sean M. Gibbons, Thomas Gurry, Rafael A. Irizarry, and Eric J. Alm. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, 8(1), December 2017. doi: 10.1038/s41467-017-01973-8.

[5] Kelly S. Ramirez, Christopher G. Knight, Mattias de Hollander, Francis Q. Brearley, Bede Constantinides, Anne Cotton, Si Creer, Thomas W. Crowther, John Davison, Manuel Delgado-Baquerizo, Ellen Dorrepaal, David R. Elliott, Graeme Fox, Robert I. Griffiths, Chris Hale, Kyle Hartman, Ashley Houlden, David L. Jones, Eveline J. Krab, Fernando T. Maestre, Krista L. McGuire, Sylvain Monteux, Caroline H. Orr, Wim H. van der Putten, Ian S. Roberts, David A. Robinson, Jennifer D. Rocca, Jennifer Rowntree, Klaus Schlaeppi, Matthew Shepherd, Brajesh K. Singh, Angela L. Straathof, Jennifer M. Bhatnagar, Cécile Thion, Marcel G. A. van der Heijden, and Franciska T. de Vries. Detecting macroecological patterns in bacterial communities across independent studies of global soils. *Nature Microbiology*, 3(2):189–196, November 2017. doi: 10.1038/s41564-017-0062-x.

[6] Mohamed R. Abdelfadil, Sascha Patz, Steffen Kolb, and Silke Ruppel. Unveiling the influence of salinity on bacterial microbiome assembly of halophytes and crops. *Environmental Microbiome*, 19(1):49, 2024. doi: 10.1186/s40793-024-00592-3.

[7] Taylor K. Dunivin, Susanna Y. Yeh, and Ashley Shade. A global survey of arsenic-related genes in soil microbiomes. *BMC Biology*, 17(1):45, May 2019. doi: 10.1186/s12915-019-0661-5.

[8] Gavin M. Douglas, Vincent J. Maffei, Jesse R. Zaneveld, Svetlana N. Yurgel, James R. Brown, Christopher M. Taylor, Curtis Huttenhower, and Morgan G. I. Langille. PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology*, 38(6):685–688, 2020. doi: 10.1038/s41587-020-0548-6.

[9] Franziska Wemheuer, Jessica A. Taylor, Rolf Daniel, Emma Johnston, Peter Meinicke, Torsten Thomas, and Bernd Wemheuer. Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environmental Microbiome*, 15(1), May 2020. doi: 10.1186/s40793-020-00358-7.

[10] Enzong Xiao, Valdis Krumins, Tangfu Xiao, Yiran Dong, Song Tang, Zengping Ning, Zhengyu Huang, and Weimin Sun. Depth-resolved microbial community analyses in two contrasting soil cores contaminated by antimony and arsenic. *Environmental Pollution*, 221:244–255, February 2017. doi: 10.1016/j.envpol.2016.11.071.

[11] Stilianos Louca, Laura Wegener Parfrey, and Michael Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, 2016. doi: 10.1126/science.aaf4507.

[12] Stilianos Louca, Martin F. Polz, Florent Mazel, Michaeline B. N. Albright, Julie A. Huber, Mary I. O'Connor, Martin Ackermann, Aria S. Hahn, Diane S. Srivastava, Sean A. Crowe, Michael Doebeli, and Laura Wegener Parfrey. Function and functional redundancy in microbial systems. *Nature Ecology & Evolution*, 2(6):936–943, April 2018. doi: 10.1038/s41559-018-0519-1.

[13] Shan Sun, Roshonda B. Jones, and Anthony A. Fodor. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome*, 8:46, 2020. doi: 10.1186/s40168-020-00815-y.

[14] Kayla A. Carter, Anthony A. Fodor, Jennifer E. Balkus, Angela Zhang, Myrna G. Serrano, Gregory A. Buck, Stephanie M. Engel, Michael C. Wu, and Shan Sun. Vaginal microbiome metagenome inference accuracy: differential measurement error according to community composition. *mSystems*, 8(2):e01003–22, 2023. doi: 10.1128/msystems.01003-22.

[15] Gongchao Jing, Yufeng Zhang, Wenzhi Cui, Lu Liu, Jian Xu, and Xiaoquan Su. Meta-Apo improves accuracy of 16S-amplicon-based prediction of microbiome function. *BMC Genomics*, 22(1):9, 2021. doi: 10.1186/s12864-020-07307-1.

[16] Monica Steffi Matchado, Malte Rühlemann, Sandra Reitmeier, Tim Kacprowski, Fabian Frost, Dirk Haller, Jan Baumbach, and Markus List. On the limits of 16S rRNA gene-based metagenome prediction and functional profiling. *Microbial Genomics*, 10(2), February 2024. doi: 10.1099/mgen.0.001203.

[17] Christophe Djemiel, Pierre-Alain Maron, Sébastien Terrat, Samuel Dequiedt, Aurélien Cottin, and Lionel Ranjard. Inferring microbiota functions from taxonomic genes: a review. *GigaScience*, 11, 2022. doi: 10.1093/gigascience/giab090.

[18] Robyn J. Wright and Morgan G. I. Langille. PICRUSt2-SC: an update to the reference database used for functional prediction within PICRUSt2. *Bioinformatics*, 41(5):btaf269, 2025. doi: 10.1093/bioinformatics/btaf269.

[19] Vitor Heidrich and Lukas Beule. Are short-read amplicons suitable for the prediction of microbiome functional potential? A critical perspective. *iMeta*, 1(3), 2022. doi: 10.1002/imt2.38.

[20] Stilianos Louca, Michael Doebeli, and Laura Wegener Parfrey. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome*, 6(1), February 2018. doi: 10.1186/s40168-018-0420-9.

[21] Colman O'Cathail, Alisha Ahamed, Josephine Burgin, Carla Cummins, Rajkumar Devaraj, Khadim Gueye, Dipayan Gupta, Vikas Gupta, Muhammad Haseeb, Maira Ihsan, Eugene Ivanov, Suran Jayathilaka, Vishnukumar Kadhirvelu, Manish Kumar, Ankur Lathi, Rasko Leinonen, Jasmine McKinnon, Lili Meszaros, Joana Pauperio, Stephane Pesant, Nadim Rahman, Gabriele Rinck, Sandeep Selvakumar, Swati Suman, Yanisa Sunthornyotin, Marianna Ventouratou, Zahra Waheed, Peter Woollard, David Yuan, Ahmad Zyoud, Tony Burdett, and Guy Cochrane. The European Nucleotide Archive in 2024. *Nucleic Acids Research*, 53(D1):D49–D55, November 2024. doi: 10.1093/nar/gkae975.

[22] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, November 2012. doi: 10.1093/nar/gks1219.

[23] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, September 2021. doi: 10.1093/nar/gkab776.

[24] R. Leinonen, H. Sugawara, M. Shumway, and International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Research*, 39(Database):D19–D21, November 2010. doi: 10.1093/nar/gkq1019.

[25] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10, May 2011. doi: 10.14806/ej.17.1.200.

[26] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7):581–583, May 2016. doi: 10.1038/nmeth.3869.

[27] Erik S. Wright. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *The R Journal*, 8(1):352, 2016. doi: 10.32614/rj-2016-025.

[28] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2026. URL https://www.R-project.org/. Version 4.5.3.

[29] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with Snakemake. *F1000Research*, 10:33, April 2021. doi: 10.12688/f1000research.29032.2.

[30] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018. doi: 10.1093/bioinformatics/bty560.

[31] Felix Van der Jeugt, Peter Dawyndt, and Bart Mesuere. FragGeneScanRs: faster gene prediction for short reads. *BMC Bioinformatics*, 23:198, 2022. doi: 10.1186/s12859-022-04736-5.

[32] Mina Rho, Haixu Tang, and Yuzhen Ye. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research*, 38(20):e191, 2010. doi: 10.1093/nar/gkq747.

[33] Sean R. Eddy. Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10):e1002195, 2011. doi: 10.1371/journal.pcbi.1002195.

[34] Takuya Aramaki, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, 2020. doi: 10.1093/bioinformatics/btz859.

[35] K. Katoh and D. M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780, January 2013. doi: 10.1093/molbev/mst010.

[36] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490, March 2010. doi: 10.1371/journal.pone.0009490.

[37] J. Roger Bray and J. T. Curtis. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, 27(4):325–349, October 1957. doi: 10.2307/1942268.

[38] Marti J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, February 2001. doi: 10.1046/j.1442-9993.2001.01070.x.

[39] Laura Poggio, Luis M. de Sousa, Niels H. Batjes, Gerard B. M. Heuvelink, Bas Kempen, Eloi Ribeiro, and David Rossiter. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7(1):217–240, 2021. doi: 10.5194/soil-7-217-2021.

[40] Wes McKinney. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, SciPy, page 56–61. SciPy, 2010. doi: 10.25080/majora-92bf1922-00a.

[41] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk,

Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, February 2020. doi: 10.1038/s41592-019-0686-2.

[42] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference*, SciPy, page 92–96. SciPy, 2010. doi: 10.25080/majora-92bf1922-011.

[43] Morgan G I Langille, Jesse Zaneveld, J Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A Reyes, Jose C Clemente, Deron E Burkepile, Rebecca L Vega Thurber, Rob Knight, Robert G Beiko, and Curtis Huttenhower. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31(9):814–821, August 2013. doi: 10.1038/nbt.2676.

[44] Song-Can Chen, Guo-Xin Sun, Barry P. Rosen, Si-Yu Zhang, Ye Deng, Bo-Kai Zhu, Christopher Rensing, and Yong-Guan Zhu. Recurrent horizontal transfer of arsenite methyltransferase genes facilitated adaptation of life to arsenic. *Scientific Reports*, 7(1), August 2017. doi: 10.1038/s41598-017-08313-2.

[45] Adam C Martiny, Kathleen Treseder, and Gordon Pusch. Phylogenetic conservatism of functional traits in microorganisms. *The ISME Journal*, 7(4):830–838, December 2012. doi: 10.1038/ismej.2012.160.

[46] Jennifer B. H. Martiny, Stuart E. Jones, Jay T. Lennon, and Adam C. Martiny. Microbiomes in light of traits: A phylogenetic perspective. *Science*, 350(6261), November 2015. doi: 10.1126/science.aac9323.

[47] Marta Goberna and Miguel Verdú. Predicting microbial traits with phylogenies. *The ISME Journal*, 10(4):959–967, 2015. doi: 10.1038/ismej.2015.171.

[48] Peter E. Luton, Jonathan M. Wayne, Richard J. Sharp, and Paul W. Riley. The mcrA gene as an alternative to 16S rRNA in the phylogenetic analysis of methanogen populations in landfill. *Microbiology*, 148(11):3521–3530, 2002. doi: 10.1099/00221287-148-11-3521.

[49] Paul J. McMurdie and Susan Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, 10(4):e1003531, 2014. doi: 10.1371/journal.pcbi.1003531.

[50] Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 8:2224, 2017. doi: 10.3389/fmicb.2017.02224.

[51] Si-Yu Zhang, Xiao Xiao, Song-Can Chen, Yong-Guan Zhu, Guo-Xin Sun, and Konstantinos T. Konstantinidis. High Arsenic Levels Increase Activity Rather than Diversity or Abundance of Arsenic Metabolism Genes in Paddy Soils. *Applied and Environmental Microbiology*, 87(20), 2021. doi: 10.1128/aem.01383-21.

[52] Xinwei Song, Yiqun Li, Erinne Stirling, Kankan Zhao, Binhao Wang, Yongguan Zhu, Yongming Luo, Jianming Xu, and Bin Ma. AsgeneDB: a curated orthology arsenic metabolism gene database and computational tool for metagenome annotation. *NAR Genomics and Bioinformatics*, 4(4), October 2022. doi: 10.1093/nargab/lqac080.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarypreprint.pdf