Supplementary material *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]

[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

Contents

S1	Data sources and study inventory	2
S2	Run-level exclusions	8
S3	The marker-taxa lookup	17
S4	Phylogeny of the marker taxa	20
S5	Agreement between the two inference routes	20
S6	Sequencing-depth coefficients	30
S7	Compartment contrasts in full	31
S8	Software and reference data	32
S9	The predictor's own confidence metric	33
S10	Measured gene content from paired shotgun metagenomes	35
S11	FAPROTAX as a third inference route	36
S12	Gene carriage within the marker taxa and study-level moderators	37
S13	Compartment conclusions under each route and the *hydA* ortholog	38
S14	Compartment, composition, soil and corpus analyses in full	39
S15	Accuracy against measured gene content across environments	42
S16	Tax4Fun2 as a fourth inference route	43

S1 Data sources and study inventory

Every sequencing run analysed here was retrieved from a public archive, and no new sequencing was performed. Runs were fetched from the Sequence Read Archive (Leinonen et al., 2010) using accessions resolved against the European Nucleotide Archive (O'Cathail et al., 2024), which holds the metadata from which compartment was assigned.

A structured search executed on 20 May 2026 returned 12,114 records, reducing to 7,539 after deduplication and to 177 candidate datasets under the explicit inclusion rules. Expansion to run level, followed by the exclusions in Section S2, leaves 3,041 amplicon runs across 78 studies. A further 455 shotgun runs from 32 studies were held for later work and are listed here for completeness.

Table S1 names every study in the corpus. Titles are reproduced as deposited, and were retrieved programmatically rather than transcribed. The main text names no individual study, so this table is the record of what the corpus actually consists of. Two studies, both excluded on platform, carried no title in the European Nucleotide Archive and were recovered from the NCBI Sequence Read Archive instead.

Compartment was assigned from archive metadata and not from study titles, since a title records what a paper is about while the metadata record what was sequenced. Where a study contributed no retained amplicon run, its compartment cell is left blank rather than guessed.

A BioProject frequently holds runs this study never had any claim on. 286 runs across the corpus are RNA-seq, whole-chromosome sequencing or ChIP-seq deposited alongside the amplicon or shotgun libraries of interest. These were never candidates and so appear in no exclusion category, but they are counted in Table S1 so that the four count columns sum to every run the BioProject holds. A study whose runs are entirely of this kind shows zeros across the amplicon and shotgun columns.

Table S1. Every study in the corpus, with the title as deposited. Ampl., amplicon runs retained for analysis; Shot., shotgun runs held for later work; Other, runs of a library strategy outside the scope of this study; Excl., runs excluded (reasons in Table S2). Titles were retrieved from the European Nucleotide Archive, and for two studies from the NCBI Sequence Read Archive.

Accession	Study title	Compartment	Geography	Ampl.	Shot.	Other	Excl.
PRJDB10144	Deficiency of alcohol dehydrogenase 2 reduces arsenic in paddy-field rice grains by suppressing silicate transporters during reproductive stage	–	–	0	6	0	0
PRJDB17534	Metatranscriptome of Japanese paddy soil	–	Japan	0	0	18	0
PRJDB28498	The transcriptome mechanism of L-glutamate seed priming enhancing 2-AP formation in fragrant rice by exposing arsenite stress (PRJCA016847)	–	–	0	0	12	0
PRJDB7706	Soil Microbial Community Involved in Reductive Dissolution of Arsenic	–	Japan	0	0	0	36
PRJEB14856	Metagenomic analysis of microbial mat of hot springs	–	India	0	1	0	0
PRJEB26724	Inhibition of Microbial Methylation via arsM in the Rhizosphere investigated via amplicon sequencing in paddy soil from Spain	Paddy soil	Europe	40	0	0	0
PRJEB27237	Soil prokaryotic communities (16S rRNA gene) in the mine waste contaminated Ogosta floodplain.	As-soil	Bulgaria	55	0	0	1
PRJEB28491	16S rRNA metagenomic investigation of arsenotrophic bacteriome of ground water and soil from arsenic-prone area of Bangladesh	–	BengalBasin	0	0	0	2
PRJEB28492	16S rRNA metagenomic investigation of arsenotrophic bacteriome of ground water and soil from arsenic-prone area of Bangladesh	–	BengalBasin	0	0	0	2
PRJEB29426	Microbial community response to environmental changes in a technosol historically contaminated by the burning of chemical ammunitions.	As-soil	France	8	0	0	0
PRJEB45341	16S-soil-ash-Bangladesh, Here we investigated in large scale field trials across two geographic regions and two seasons, if domestic hearth ash can be used as a f...	Paddy soil	BengalBasin	108	0	0	0
PRJEB45605	16S-timepoints-soil, rice paddy pot experiment with 4 fertiliser treatments over the life cycle of rice plants, analysis of 16S rRNA in rhizosphere soil at 4 time...	Paddy soil	China	80	0	0	0

continued

Table S1 continued

Accession	Study title	Compartment	Geography	Ampl.	Shot.	Other	Excl.
PRJEB45610	ITS-timepoints-soil, rice paddy pot experiment with 4 fertiliser treatments over the life cycle of rice plants, analysis of fungal ITS rRNA in rhizosphere soil at...	Paddy soil	China	79	0	0	1
PRJEB45805	16S_Soil_Globaldata_Ricepaddy, Soil microbiome interacts with geochemistry to regulate paddy arsenic and cadmium availability and arsenic transformation	Paddy soil	Mekong	132	0	0	0
PRJEB45806	ITS_Soil_Globaldata_Ricepaddy, Soil microbiome interacts with geochemistry to regulate paddy arsenic and cadmium availability, and arsenic transformation	Paddy soil	Mekong	132	0	0	0
PRJEB8996	Microbial community composition of a household sand filter used for arsenic, iron, and manganese removal from groundwater in Vietnam	–	Mekong	0	0	0	18
PRJNA1004417	Metagenomics sequencing of microbial communities in the rhizosphere soil of arsenic-hyperaccumulator Pteris vittata	–	–	0	0	0	21
PRJNA1013411	Arsenic removal from groundwater for long-term domestic use	–	China	0	0	0	5
PRJNA1026071	16S rRNA gene amplicon sequencing of paddy soil originating from Changsha or Hengnan (China)	Paddy soil	China other	6	0	0	0
PRJNA1039139	Soil arsenic oxidazing microorganisms (aioA)	As-soil	China	91	0	0	0
PRJNA1068274	Paddy soil metagenomes with/out arsenic contamination	–	China	0	168	0	0
PRJNA1068685	Paddy soil metatranscriptomes with/out arsenic contamination	–	China	0	0	45	0
PRJNA1091876	High-throughput sequencing of volatile 16 S rRNA genes from arsenic-antibiotic co-contaminated paddy soil under different molecular weight DOM regulation	Paddy soil	China	9	0	0	0
PRJNA1092548	A new early warning for paddy soil ecological security: Low molecular weight DOM enhances the spread of arsenic-ARGs across the paddy soil-air interface	–	China	0	9	0	0
PRJNA1104782	Arsenic methylation in paddy soil	Paddy soil	China	3	0	0	0
PRJNA1110177	Genetic and molecular basis of organic-arsenic-microbe interactions in arsenic prone aquifers (GOAM)	As-soil	Cambodia	47	0	0	0
PRJNA1118437	Transcriptomics data of iron arsenic affected primed rice seedlings	–	India	0	0	6	0
PRJNA1135448	Midsaeason draining of paddy water suppresses microbial arsenic methylation and alleviates rice straighthead disease	Paddy soil	China	32	0	0	0
PRJNA1136577	well water Raw sequence reads	–	China other	0	0	30	0
PRJNA1142403	Bacterial raw sequence reads in arsenic-contaminated paddy soil under different microplastic treatments	Paddy soil	China	44	0	0	0
PRJNA1142719	Archaeal raw sequence reads in arsenic-contaminated paddy soil under different microplastic treatments	Paddy soil	China	44	0	0	0
PRJNA1166154	Soil humic substances-specific diversity and functional traits of arsenic-reducing bacteria in flooded paddy soil	Paddy soil	China	80	0	0	0
PRJNA1167190	Soil humic substances-specific diversity and functional traits of arsenic-reducing bacteria in flooded paddy soil (metagenomics)	–	China	0	4	0	0
PRJNA1167431	Bacterial community diversity of rhizosphere and non-rhizosphere soils of Hippophae rhamnoides in Arsenic sandstone area of Inner Mongolia	As-soil	China	12	0	0	0
PRJNA1195538	Microbial-driven mechanisms of arsenic methylation during Chinese rice wine fermentation	–	China other	0	18	0	0
PRJNA1254602	Groundwater Raw sequence reads	–	China	0	12	0	0
PRJNA1284197	The environmental pollutants of arsenic and microplastics mediating soil microbiota shift aphid growth in tobacco	As-soil	–	21	0	0	0
PRJNA1346277	Unraveling the non-target effects of molybdate: Inhibition of arsenic methylation in paddy soil	–	China	0	0	6	0

continued

Table S1 continued

Accession	Study title	Compartment	Geography	Ampl.	Shot.	Other	Excl.
PRJNA1363201	Simultaneous arsenate and sulfate reduction in sediment-packed columns to evaluate arsenic removal efficiency and stability of biogenic minerals	–	Mexico	0	0	0	3
PRJNA1367948	Microbial community dynamics in paddy soil of West Bengal, India under varied arsenic concentrations	–	BengalBasin	0	3	0	0
PRJNA211801	Influence of Sulfur on Transcription of Genes Involved in Arsenic Accumulation in Rice Grains	–	–	0	0	3	0
PRJNA253777	Sediment sample Metagenome	–	Taiwan	0	0	0	2
PRJNA270654	Biochar In Co-Contaminated Soil	As-soil	New Zealand	9	0	0	0
PRJNA285433	Microbial communities under arsenic contaminant environment	As-soil	China	10	0	0	0
PRJNA344926	rice paddy metagenome Metagenome	Paddy soil	China	83	0	0	0
PRJNA354810	Heavy metal/metalloid contaminated soil Raw sequence reads	As-soil	China	26	0	0	0
PRJNA355907	rice paddy metagenome Raw sequence reads	Paddy soil	China	91	0	0	0
PRJNA357341	rice paddy metagenome Raw sequence reads	Paddy soil	China	53	0	0	0
PRJNA357668	Nitric Oxide mediated transcriptional Modulation in rice under arsenic stress	–	India	0	0	8	0
PRJNA361507	Anaerobic arsenic enrichment culture microbiome	As-soil	Brazil	6	0	0	0
PRJNA387583	Microbial community analysis from aquifers in three Ohio counties	–	USA	0	0	0	1
PRJNA471565	rice paddy metagenome Raw sequence reads	Paddy soil	China	71	0	0	1
PRJNA478283	Microbial group mediating arsenic methylation and demethylation in paddy soils	Paddy soil	China	18	0	0	0
PRJNA478333	Microbial groups mediating arsenic methylation and demethylation in paddy soils	Paddy soil	China	18	0	0	0
PRJNA492762	arsenic methylation in the enrichment cultures of SRB from paddy soils	–	China	0	0	0	9
PRJNA493661	groundwater metagenome Targeted loci	Aquifer, depth n.s.	Hetao	2	0	0	0
PRJNA493691	Microbial sulfate reduction decreases arsenic mobilization in Fe(II)-rich paddy soils under flooded conditions	Paddy soil	China	18	0	0	0
PRJNA527922	Biotransformations of Arsenic Associated Microbiota in Different Belowground Soil Fauna Guts: a Hidden Hotspot?	–	China	0	0	0	84
PRJNA528209	Dissolved organic matter accelerates arsenic methylation and volatilization of paddy soils	Paddy soil	China	72	0	0	0
PRJNA528471	Microbiomes of groundwater, surface water, and soil within a gold mining area across low arsenic concentrations	Aquifer, depth n.s.	Thailand	12	0	0	0
PRJNA528962	As(V)/sulfate/Fe(III)-reducing bacterial enrichment recuperated from arsenic contaminated sediments	As-soil	Mexico	10	0	0	0
PRJNA531524	Antimony-reducing bacteria identified by DNA-stable isotope probing and their potential metabolic traits revealed by metagenomic-binning Raw sequence reads	Paddy soil	China	32	2	0	0
PRJNA545810	The effect of microbially mediated sulfate reduction on arsenic mobilization in the shallow aquifer from Jianghan Plain, central Yangtze River Basin	–	–	0	60	0	0
PRJNA556952	Bacteriophages from Arsenic Resistant Bacteria Transduced Arsenic Resistance Genes and Transformed Soil Arsenic Profile	As-soil	–	9	0	0	0
PRJNA559049	Arsenite oxidizing bacteria in paddy soil and sediment Raw sequence reads	Paddy soil	China	20	0	0	0
PRJNA562779	Transcriptome and proteome analyses reveal selenium mediated amelioration of arsenic toxicity in rice	–	India	0	0	8	0
PRJNA593718	16S rRNA gene amplicon sequencing of groundwater, sediment and sediment cultures relevant for arsenic mobilization	Shallow aquifer	Viet Nam	23	0	0	0
PRJNA602583	Arsenic biomethylation in the shallow aquifers of Jianghan Plain, central Yangtze River Basin	–	China	0	47	0	0
PRJNA610123	From arsenic-polluted sediments to sulfate/arsenate-reducing bacterial consortia	–	Mexico	0	0	0	8

continued

Table S1 continued

Accession	Study title	Compartment	Geography	Ampl.	Shot.	Other	Excl.
PRJNA628856	16S rRNA amplicon sequencing from groundwater wells in Van Phuc, Hanoi, Vietnam.	Aquifer, depth n.s., Shallow aquifer	Mekong	18	0	0	0
PRJNA632886	soil metagenome Metagenome	–	China	0	0	24	0
PRJNA640466	amplicon and shotgun metagenomic sequencing of nitrate-dependent Sb(III) oxidizers	Paddy soil	China	32	0	0	2
PRJNA642986	Effects of warming on microbial community and diversity in arsenic contaminated paddy soils	–	China	0	0	0	16
PRJNA644740	Effects of arsenic and fish on sediment microorganisms (Bacteria)	–	–	0	0	0	30
PRJNA648294	Combined Transcriptome and Proteome Profiling Reveals Microbial Consortium mediated Growth Promotion and Reduction of Arsenic Burden and Toxicity in Rice (Oryza s...	–	India	0	0	8	0
PRJNA649362	Microcosm based exploration of contaminated groundwater microbial diversity	–	BengalBasin	0	0	0	9
PRJNA649896	Effects of arsenic and fish on sediment microorganisms (fungi)	–	China	0	0	0	24
PRJNA686070	Discerning role of a functional arsenic resistance cassette in evolution and adaptation of a rice pathogen	–	–	0	0	8	0
PRJNA687064	Silicate fertilization effects Metagenome	–	South Korea	0	0	0	4
PRJNA690162	Gene amplicon sequencing from flooded rice paddy mesocosms treated with silicon rich amendments	Paddy soil	USA	124	0	0	31
PRJNA690681	Assessing the diazotrophic community in a vertical soil profile with antimony and arsenic co-contamination	As-soil	–	36	0	0	0
PRJNA694468	bacteria associated with the class Acidobacteriia in metal-contaminated soil revealed by metagenomic-binning Raw sequence reads	As-soil	China	53	19	0	0
PRJNA695517	Synergistic impacts of arsenic and antimony co-contamination on diazotrophic communities	As-soil	–	36	0	0	0
PRJNA714492	Arsenic-methylating anaerobic rice paddy soil-derived microbiomes (Raw DNA sequence reads)	–	Viet Nam	0	12	24	0
PRJNA716501	Microbiomes in arsenic-contaminated soils	As-soil	Thailand	15	0	0	0
PRJNA716505	Diversity of arsM gene in arsenic-contaminated soils	–	Thailand	0	0	0	5
PRJNA719005	16S rRNA gene amplicon sequencing of incubated sediments from an As-contaminated aquifer in Hanoi (Vietnam)	–	Mekong	0	0	0	9
PRJNA754837	Arsenic transformation in high arsenic groundwater	–	China	0	2	0	0
PRJNA766239	antimonate reducing bacteria Raw sequence reads	Paddy soil	China	12	3	0	6
PRJNA768870	Co-occurrence of Antibiotic Resistance and Arsenic Biotransformation Genes in Paddy Soils	–	–	0	0	56	0
PRJNA786456	Impact of different manure sources on the fate of arsenic-antibiotic in co-contaminated paddy soils: dissolved organic matter (DOM)	Paddy soil	China	39	0	0	0
PRJNA795569	Full-length 16S rRNA gene high-throughput sequencing of sediments from an As-contaminated aquifer	–	Hetao	0	0	0	11
PRJNA798572	Arsenic-methylating microorganisms and total bacterial and archaeal communities in the methanogenic paddy soil	Paddy soil	China	108	0	0	0
PRJNA799794	paddy soil enrichment Raw sequence reads	–	China	0	8	0	0
PRJNA820970	predatory bacteria diversity of arsenic contamination soils	As-soil	China	24	0	0	0
PRJNA823829	Active microbial communities in arsenic-contaminated soil during 15-day flooding	As-soil	–	18	0	0	0
PRJNA824136	Metagenomics of arsenic-contaminated soils	–	Thailand	0	0	0	2
PRJNA826636	Incubation from paddy soil Raw sequence reads	–	China	0	0	29	0
PRJNA832082	Influence of manure application on electrokinetic remediation of arsenic- antibiotic co-contaminated paddy soils	Paddy soil	China	36	0	0	0

continued

Table S1 continued

Accession	Study title	Compartment	Geography	Ampl.	Shot.	Other	Excl.
PRJNA858795	16S rRNA gene sequencing of bacterial and archaeal communities living in Italian rice paddies rhizosphere soil and pore water	Paddy soil	Italy	114	0	0	0
PRJNA862747	groundwater metagenome sequencing	–	China	0	3	0	0
PRJNA873282	Fate of antibiotic resistance genes in arsenic and antibiotic co-contaminated paddy soil regulated by composts-derived humic substances	Paddy soil	China	39	0	0	0
PRJNA886312	Genetic diversity of viral arsM in arsenic-contaminated soil during 15-day flooding	As-soil	–	18	0	0	0
PRJNA916093	Staphylococcus aureus, E. coli, Klebsiella spp Metagenome	–	Bangladesh	0	6	0	0
PRJNA922084	16S rRNA gene amplicon sequencing of paddy soil microcosms originating from Vercelli (Italy)	–	Europe	0	0	0	6
PRJNA954693	Effects of different dissolved organic matter inputs on microbial mediated arsenic migration and transformation in groundwater with high arsenic content	–	China	0	0	0	23
PRJNA961897	Effects of different dissolved organic matter inputs on microbial mediated arsenic migration and transformation in groundwater with high arsenic content	–	China	0	0	0	6
PRJNA980983	Soil microbiomes and arsenic-transforming microorganisms	As-soil	Thailand	12	6	0	0
SRP003266	Microbial communities in rice rhizosphere altered by intermittent and continuous flooding in fields with long-term arsenic application	–	–	0	0	0	4
SRP005643	Microbial communities and arsenic concentrations affected by water managements and soil arsenic levels in rice field	–	–	0	0	0	16
SRP039499	Molecular Characterization of the Structural and Functional Microbial Community in Sediments of a Northern Utah, Basin-Fill Aquifer	–	USA	0	0	0	9
SRP043614	Bacterial communities in drinking water-wells from northern Mexico. Metagenome	–	Mexico	0	0	0	9
SRP044066	Arsenic enriched groundwater Metagenome	–	Taiwan	0	0	0	3
SRP046303	Microbially mediated reduction of Fe(III) and As(V) in Cambodian sediments amended with 13C-labelled hexadecane and kerogen	–	Cambodia	0	0	0	3
SRP059905	arsenic contaminated paddy soil Raw sequence reads	Paddy soil	China	5	0	0	0
SRP061694	Arsenic contaminated groundwater Raw sequence reads	–	BengalBasin	0	0	0	2
SRP067270	Cache Valley Sediments and groundwater arrA gene metagenome	–	USA	0	0	0	9
SRP092662	high arsenic groundwater Genome sequencing and assembly	–	China	0	0	0	1
SRP094596	Nitrate-mediated Fe(II) and As(III) oxidation in paddy soil Raw sequence reads	Paddy soil	China	63	4	0	0
SRP154379	Arsenic contaminated groundwater Raw sequence reads	–	India	0	0	0	4
SRP162785	groundwater metagenome Metagenome	–	China	0	1	0	0
SRP166039	arsenic and antimony co-contamination on sedimentary microbial communities in rivers with different pollution gradients Raw sequence reads	As-soil	China	31	0	0	0
SRP173566	Microbial arsenic cycling in Italian contaminated groundwaters	–	Europe	0	0	0	6
SRP186087	Vegetation types impact microbial interaction with Sb contaminants in mining contaminated soil Raw sequence reads	Paddy soil	China	61	0	0	0
SRP246194	Metagenomic Sequencing	–	India	0	2	0	0
SRP252167	Increased arsenic mobilization in the rice rhizosphere is mediated by iron-reducing bacteria	Paddy soil	China	36	0	0	0
SRP254590	Metagenomics and Metatranscriptomics analysis of microbial communities in paddy soils with different arsenic levels	–	China	0	12	0	0
SRP278891	Chromated copper arsenate soil bacterial communities	As-soil	Denmark	39	0	0	0

continued

Table S1 continued

Accession	Study title	Compartment	Geography	Ampl.	Shot.	Other	Excl.
SRP286951	soil microbial community structure and function influenced by Arsenic and antimony co-contamination Raw sequence reads	As-soil	China	23	4	0	0
SRP298835	Microbial diversity and function of arsenic contaminated rice paddy soil sequence reads	–	India	0	2	0	11
SRP313593	Groundwater microbiome study of an arsenic contaminated hotspot of West Bengal, India monitored over dry seasons.	–	India	0	0	0	11
SRP315170	Metagenomic study of aquifer sediments from the Hetao Basin, China	–	China	0	0	0	2
SRP321433	Silicon fertilization influences microbial assemblages in rice roots and decreases arsenic concentration in grain: a five-season in-situ remediation field study	–	China	0	0	0	77
SRP329786	Coupling mechanism of arsenic and C-N-Fe transformation process in paddy soil after straw incorporation	Paddy soil	–	24	0	0	0
SRP339895	Anammox coupled with arsenate reduction	Paddy soil	China	21	0	0	3
SRP354046	Metagenomics of sediments from an As-contaminated aquifer	–	China	0	6	0	0
SRP370321	aesenic sediment sample DNA	As-soil	China	6	2	0	0
SRP376535	microbial community diversities of groundwater	Aquifer, depth n.s.	China	11	0	0	0
SRP376561	arsenic functional genes in groundwater	Aquifer, depth n.s.	China	32	0	0	0
SRP406096	Microbial diversity of a mercury and arsenic-contaminated mining site	As-soil	Spain	22	0	0	0
SRP415393	Cambodian paddy soil metagenomics	–	Cambodia	0	10	0	0
SRP432433	Microbial composition in high arsenic groundwater	Aquifer, depth n.s.	China	9	0	0	0
SRP437741	Metagenome of agricultural soil sample at a gradient of Arsenic contamination	–	India	0	0	0	4
SRP439241	Microbial community of sediment treated by arsenic	–	South Korea	0	0	0	11
SRP452018	Selective pressure of arsenic and antimony co-contamination on microbial community in alkaline sediments	As-soil	China	19	2	0	0
SRP453992	2Continuously Flooded(CF) rep_set	–	China	0	0	1	0
SRP466511	High arsenic groundwater microbial community	Aquifer, depth n.s.	China	31	0	0	0
SRP468581	Supplementary information of evidence for the role of microbes in the Silicon-regulated arsenic concentrations of rice roots in the soil environment	–	China	0	0	0	92
SRP468584	Evidence for the role of microbes in the silicon-regulated arsenic concentrations of rice roots in the soil environment	Paddy soil	China	108	0	0	0
SRP470479	Microbial community in high arsenic geothermal groundwater	–	China	0	0	0	17
SRP500229	Rice Rhizosphere Soil Metagenome	Paddy soil	India	5	0	0	0
SRP507371	Aquifer microbial communities in thermally mobilized arsenic plumes	Deep aquifer	Canada	27	11	0	0
SRP507442	Microbial Communities in Photoelectron-Assisted Arsenate Reduction Systems	As-soil	China	10	0	0	3
SRP580424	Calcium peroxide applications suppress microbial arsenic methylation and straighthead disease in rice	Paddy soil	China	32	0	0	0
SRP641943	Unraveling the non-target effects of molybdate: Inhibition of arsenic methylation in paddy soil	–	China	0	6	0	0
SRP657214	Dose-Dependent Control of Arsenic in Rice by Sulfate-Mediated Iron-Sulfide Formation	Paddy soil	China	6	0	0	0
SRP666293	DOM-Driven Methane Cycling and Arsenic Mobilization in Paddy Soils	Paddy soil	China	80	4	0	0

As-soil, arsenic-contaminated non-paddy soil. The four count columns sum to every run the BioProject holds. Compartment is blank where a study contributed no retained amplicon run, which is why a study can show zeros across the first two columns. Geography lists at most two regions per study.

S2 Run-level exclusions

675 runs were excluded, and Table S2 gives the reason for each one individually. The main text reports only the totals per category.

The eight categories divide into two groups that differ in when they could be detected. Isolation-source, study-design and platform exclusions are resolvable from archive metadata before anything is downloaded. Size outliers, constant quality scores and upstream failures are not, and were found only after the sequences themselves were retrieved and inspected. That division is the point of Figure 2 in the main text.

Deposits labelled as amplicon libraries but carrying between 5 and 63 million reads were removed as size outliers. A 16S amplicon run in this corpus carries between 10^4 and 10^6 reads and the distribution contains no run at all between 10^6 and 5×10^6, so the threshold falls in an empty interval and is not an arbitrary cut. The ENA library_strategy field declares all of them as amplicon, so the screen acts on deposit size instead. 139 runs, in three whole studies and one isolated run, carried an identical quality score at every base and were removed because an error model cannot be estimated from invariant quality and denoising fails outright. Quality was assessed on the archive-normalised files retrieved by the toolkit, since those files and the mirrored FASTQ copies can differ in quality encoding for the same run. One study deposited paired files that differ in length by one to three reads in every sample. Those runs reach the denoiser unchecked, because a sample in which no primer is detected bypasses trimming and is copied through without pairing validation, whereas trimmed samples are checked by the trimming step itself.

Library layout was read from the deposited file listing, because the ENA library_layout field describes library preparation and not the contents of the deposit. Four studies in the initial corpus, and a further 91 runs in the arsenic-soil expansion, were declared paired while depositing a single file, and every one of them would have failed retrieval had the declared field been trusted. Illumina single-end data are supported directly, since denoising skips read merging and the denoised forward reads become the sequence variants.

Much of the depth shortfall that the read floor absorbs is a processing consequence. Primers are removed from the 5′ end only and reads are not truncated to a fixed length, because the corpus spans regions from V4 to V3–V4 and any single truncation length discards one region or destroys read overlap in another. When a read is longer than its amplicon, as with 2×300 sequencing of the 253-bp V4 region, the 3′ tail runs through the opposite primer into adapter sequence at end-of-read quality, and the expected-error filter rejects most reads before denoising, leaving fewer than one read in a hundred in the worst affected runs. Adding 3′ adapter trimming would recover these reads, but only by reprocessing every sample.

Table S2. Run-level exclusion ledger, all 675 excluded runs. The first five reasons are resolvable from archive metadata; the last three became apparent only after the sequences were retrieved and inspected.

Run	Study	Reason	Compartment
SRR25923915	PRJNA1013411	Failed isolation-source criterion	Aquifer, depth n.s.
SRR25923916	PRJNA1013411	Failed isolation-source criterion	Aquifer, depth n.s.
SRR25923917	PRJNA1013411	Failed isolation-source criterion	Aquifer, depth n.s.
SRR25923918	PRJNA1013411	Failed isolation-source criterion	Aquifer, depth n.s.
SRR25923919	PRJNA1013411	Failed isolation-source criterion	Aquifer, depth n.s.
SRR13298376	PRJNA687064	Failed isolation-source criterion	Paddy soil
SRR13298377	PRJNA687064	Failed isolation-source criterion	Paddy soil
SRR13298378	PRJNA687064	Failed isolation-source criterion	Paddy soil
SRR13375690	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375696	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375700	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375703	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375709	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375712	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375717	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375723	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375728	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375736	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375742	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375749	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375752	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375761	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375773	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375815	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR13375816	PRJNA690162	Failed isolation-source criterion	Paddy soil
SRR14660463	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660464	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660465	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660466	SRP321433	Failed isolation-source criterion	Paddy soil

continued

Table S2 continued

Run	Study	Reason	Compartment
SRR14660467	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660468	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660469	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660470	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660471	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660472	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660473	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660474	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660475	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660476	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660477	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660478	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660479	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660480	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660481	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660482	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660483	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660484	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660485	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660486	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660487	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660488	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660489	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660490	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660491	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660492	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660493	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660494	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660495	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660496	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660497	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660498	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660499	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660500	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660501	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660502	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660503	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660504	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660505	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660506	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660507	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660508	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660509	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660510	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660511	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660512	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660513	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660514	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660515	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660516	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660517	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660518	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660519	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660520	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660521	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660522	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660523	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660524	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660525	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660526	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660527	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660528	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660529	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660530	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660531	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660532	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660533	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660534	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660535	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660536	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660537	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660538	SRP321433	Failed isolation-source criterion	Paddy soil
SRR14660539	SRP321433	Failed isolation-source criterion	Paddy soil

continued

Table S2 continued

Run	Study	Reason	Compartment
SRR26535889	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535890	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535891	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535892	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535893	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535894	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535895	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535896	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535897	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535898	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535899	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535900	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535901	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535902	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535903	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535904	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535905	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535906	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535907	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535908	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535909	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535910	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535911	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535912	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535913	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535914	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535915	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535916	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535917	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535918	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535919	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535920	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535921	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535922	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535923	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535924	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535925	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535926	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535927	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535928	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535929	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535930	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535931	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535932	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535933	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535934	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535935	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535936	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535937	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535938	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535939	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535940	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535941	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535942	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535943	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535944	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535945	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535946	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535947	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535948	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535949	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535950	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535951	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535952	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535953	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535954	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535955	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535956	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535957	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535958	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535959	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535960	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535961	SRP468581	Failed isolation-source criterion	Paddy soil

continued

Table S2 continued

Run	Study	Reason	Compartment
SRR26535962	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535963	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535964	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535965	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535966	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535967	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535968	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535969	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535970	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535971	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535972	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535973	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535974	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535975	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535976	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535977	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535978	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535979	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26535980	SRP468581	Failed isolation-source criterion	Paddy soil
SRR26693479	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693480	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693481	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693482	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693483	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693484	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693485	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693486	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693487	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693488	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693489	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693490	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693491	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693492	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693493	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693494	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
SRR26693495	SRP470479	Failed isolation-source criterion	Aquifer, depth n.s.
DRR161114	PRJDB7706	Failed study-design criterion	As-soil
DRR161115	PRJDB7706	Failed study-design criterion	As-soil
DRR161116	PRJDB7706	Failed study-design criterion	As-soil
DRR161117	PRJDB7706	Failed study-design criterion	As-soil
DRR161118	PRJDB7706	Failed study-design criterion	As-soil
DRR161119	PRJDB7706	Failed study-design criterion	As-soil
DRR161120	PRJDB7706	Failed study-design criterion	As-soil
DRR161121	PRJDB7706	Failed study-design criterion	As-soil
DRR161122	PRJDB7706	Failed study-design criterion	As-soil
DRR161123	PRJDB7706	Failed study-design criterion	As-soil
DRR161124	PRJDB7706	Failed study-design criterion	As-soil
DRR161125	PRJDB7706	Failed study-design criterion	As-soil
DRR161126	PRJDB7706	Failed study-design criterion	As-soil
DRR161127	PRJDB7706	Failed study-design criterion	As-soil
DRR161128	PRJDB7706	Failed study-design criterion	As-soil
DRR161129	PRJDB7706	Failed study-design criterion	As-soil
DRR161130	PRJDB7706	Failed study-design criterion	As-soil
DRR161131	PRJDB7706	Failed study-design criterion	As-soil
DRR161132	PRJDB7706	Failed study-design criterion	As-soil
DRR161133	PRJDB7706	Failed study-design criterion	As-soil
DRR161134	PRJDB7706	Failed study-design criterion	As-soil
DRR161135	PRJDB7706	Failed study-design criterion	As-soil
DRR161136	PRJDB7706	Failed study-design criterion	As-soil
DRR161137	PRJDB7706	Failed study-design criterion	As-soil
DRR161138	PRJDB7706	Failed study-design criterion	As-soil
DRR161139	PRJDB7706	Failed study-design criterion	As-soil
DRR161140	PRJDB7706	Failed study-design criterion	As-soil
DRR161141	PRJDB7706	Failed study-design criterion	As-soil
DRR161142	PRJDB7706	Failed study-design criterion	As-soil
DRR161143	PRJDB7706	Failed study-design criterion	As-soil
DRR161144	PRJDB7706	Failed study-design criterion	As-soil
DRR161145	PRJDB7706	Failed study-design criterion	As-soil
DRR161146	PRJDB7706	Failed study-design criterion	As-soil
DRR161147	PRJDB7706	Failed study-design criterion	As-soil
DRR161148	PRJDB7706	Failed study-design criterion	As-soil
DRR161149	PRJDB7706	Failed study-design criterion	As-soil
SRR36010944	PRJNA1363201	Failed study-design criterion	As-soil

continued

Table S2 continued

Run	Study	Reason	Compartment
SRR36010945	PRJNA1363201	Failed study-design criterion	As-soil
SRR36010946	PRJNA1363201	Failed study-design criterion	As-soil
SRR7892495	PRJNA492762	Failed study-design criterion	Paddy soil
SRR7892496	PRJNA492762	Failed study-design criterion	Paddy soil
SRR7892497	PRJNA492762	Failed study-design criterion	Paddy soil
SRR7892498	PRJNA492762	Failed study-design criterion	Paddy soil
SRR7892499	PRJNA492762	Failed study-design criterion	Paddy soil
SRR7892500	PRJNA492762	Failed study-design criterion	Paddy soil
SRR7892501	PRJNA492762	Failed study-design criterion	Paddy soil
SRR7892502	PRJNA492762	Failed study-design criterion	Paddy soil
SRR7892503	PRJNA492762	Failed study-design criterion	Paddy soil
SRR11234098	PRJNA610123	Failed study-design criterion	As-soil
SRR11234099	PRJNA610123	Failed study-design criterion	As-soil
SRR11234100	PRJNA610123	Failed study-design criterion	As-soil
SRR11234101	PRJNA610123	Failed study-design criterion	As-soil
SRR11234102	PRJNA610123	Failed study-design criterion	As-soil
SRR11234103	PRJNA610123	Failed study-design criterion	As-soil
SRR11234104	PRJNA610123	Failed study-design criterion	As-soil
SRR11234105	PRJNA610123	Failed study-design criterion	As-soil
SRR14127365	PRJNA719005	Failed study-design criterion	Shallow aquifer
SRR14127366	PRJNA719005	Failed study-design criterion	Shallow aquifer
SRR14127367	PRJNA719005	Failed study-design criterion	Shallow aquifer
SRR14127368	PRJNA719005	Failed study-design criterion	Shallow aquifer
SRR14127369	PRJNA719005	Failed study-design criterion	Shallow aquifer
SRR14127370	PRJNA719005	Failed study-design criterion	Shallow aquifer
SRR14127371	PRJNA719005	Failed study-design criterion	Shallow aquifer
SRR14127372	PRJNA719005	Failed study-design criterion	Shallow aquifer
SRR14127373	PRJNA719005	Failed study-design criterion	Shallow aquifer
SRR23023961	PRJNA922084	Failed study-design criterion	Paddy soil
SRR23023962	PRJNA922084	Failed study-design criterion	Paddy soil
SRR23023963	PRJNA922084	Failed study-design criterion	Paddy soil
SRR23023964	PRJNA922084	Failed study-design criterion	Paddy soil
SRR23023965	PRJNA922084	Failed study-design criterion	Paddy soil
SRR23023966	PRJNA922084	Failed study-design criterion	Paddy soil
SRR24140687	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140688	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140689	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140690	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140691	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140692	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140693	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140694	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140695	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140696	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140697	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140698	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140699	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140700	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140701	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140702	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140703	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140704	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140705	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140706	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140707	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140708	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24140709	PRJNA954693	Failed study-design criterion	Aquifer, depth n.s.
SRR24303576	PRJNA961897	Failed study-design criterion	Aquifer, depth n.s.
SRR24303577	PRJNA961897	Failed study-design criterion	Aquifer, depth n.s.
SRR24303578	PRJNA961897	Failed study-design criterion	Aquifer, depth n.s.
SRR24303579	PRJNA961897	Failed study-design criterion	Aquifer, depth n.s.
SRR24303580	PRJNA961897	Failed study-design criterion	Aquifer, depth n.s.
SRR24303581	PRJNA961897	Failed study-design criterion	Aquifer, depth n.s.
SRR24717624	SRP439241	Failed study-design criterion	As-soil
SRR24717625	SRP439241	Failed study-design criterion	As-soil
SRR24717626	SRP439241	Failed study-design criterion	As-soil
SRR24717627	SRP439241	Failed study-design criterion	As-soil
SRR24717628	SRP439241	Failed study-design criterion	As-soil
SRR24717629	SRP439241	Failed study-design criterion	As-soil
SRR24717630	SRP439241	Failed study-design criterion	As-soil
SRR24717631	SRP439241	Failed study-design criterion	As-soil
SRR24717632	SRP439241	Failed study-design criterion	As-soil
SRR29636498	SRP439241	Failed study-design criterion	As-soil

continued

Table S2 continued

Run	Study	Reason	Compartment
SRR29636499	SRP439241	Failed study-design criterion	As-soil
ERR6046508	PRJEB45610	Failed during retrieval or denoising	Paddy soil
SRR5660399	PRJNA387583	Failed during retrieval or denoising	Aquifer, depth n.s.
SRR7176811	PRJNA471565	Failed during retrieval or denoising	Paddy soil
SRR12052938	PRJNA640466	Failed during retrieval or denoising	Paddy soil
SRR12052953	PRJNA640466	Failed during retrieval or denoising	Paddy soil
SRR13298375	PRJNA687064	Failed during retrieval or denoising	Paddy soil
SRR13375783	PRJNA690162	Failed during retrieval or denoising	Paddy soil
SRR13375788	PRJNA690162	Failed during retrieval or denoising	Paddy soil
SRR14069154	PRJNA716505	Failed during retrieval or denoising	As-soil
SRR14069155	PRJNA716505	Failed during retrieval or denoising	As-soil
SRR14069156	PRJNA716505	Failed during retrieval or denoising	As-soil
SRR14069157	PRJNA716505	Failed during retrieval or denoising	As-soil
SRR14069158	PRJNA716505	Failed during retrieval or denoising	As-soil
SRR16110010	PRJNA766239	Failed during retrieval or denoising	Paddy soil
SRR16110011	PRJNA766239	Failed during retrieval or denoising	Paddy soil
SRR16110012	PRJNA766239	Failed during retrieval or denoising	Paddy soil
SRR16110013	PRJNA766239	Failed during retrieval or denoising	Paddy soil
SRR16110014	PRJNA766239	Failed during retrieval or denoising	Paddy soil
SRR16110015	PRJNA766239	Failed during retrieval or denoising	Paddy soil
SRR2131194	SRP061694	Failed during retrieval or denoising	Aquifer, depth n.s.
SRR2142328	SRP061694	Failed during retrieval or denoising	Aquifer, depth n.s.
SRR4996298	SRP092662	Failed during retrieval or denoising	Aquifer, depth n.s.
SRR14264037	SRP315170	Failed during retrieval or denoising	Deep aquifer
SRR14264038	SRP315170	Failed during retrieval or denoising	Deep aquifer
SRR16188586	SRP339895	Failed during retrieval or denoising	Paddy soil
SRR16188587	SRP339895	Failed during retrieval or denoising	Paddy soil
SRR16188588	SRP339895	Failed during retrieval or denoising	Paddy soil
ERR2625034	PRJEB27237	Constant quality score at every base	As-soil
SRR8751084	PRJNA527922	Constant quality score at every base	As-soil
SRR8751085	PRJNA527922	Constant quality score at every base	As-soil
SRR8751086	PRJNA527922	Constant quality score at every base	As-soil
SRR8751087	PRJNA527922	Constant quality score at every base	As-soil
SRR8751088	PRJNA527922	Constant quality score at every base	As-soil
SRR8751089	PRJNA527922	Constant quality score at every base	As-soil
SRR8751090	PRJNA527922	Constant quality score at every base	As-soil
SRR8751091	PRJNA527922	Constant quality score at every base	As-soil
SRR8751092	PRJNA527922	Constant quality score at every base	As-soil
SRR8751093	PRJNA527922	Constant quality score at every base	As-soil
SRR8751094	PRJNA527922	Constant quality score at every base	As-soil
SRR8751095	PRJNA527922	Constant quality score at every base	As-soil
SRR8751096	PRJNA527922	Constant quality score at every base	As-soil
SRR8751097	PRJNA527922	Constant quality score at every base	As-soil
SRR8751098	PRJNA527922	Constant quality score at every base	As-soil
SRR8751099	PRJNA527922	Constant quality score at every base	As-soil
SRR8751100	PRJNA527922	Constant quality score at every base	As-soil
SRR8751101	PRJNA527922	Constant quality score at every base	As-soil
SRR8751102	PRJNA527922	Constant quality score at every base	As-soil
SRR8751103	PRJNA527922	Constant quality score at every base	As-soil
SRR8751104	PRJNA527922	Constant quality score at every base	As-soil
SRR8751105	PRJNA527922	Constant quality score at every base	As-soil
SRR8751106	PRJNA527922	Constant quality score at every base	As-soil
SRR8751107	PRJNA527922	Constant quality score at every base	As-soil
SRR8751108	PRJNA527922	Constant quality score at every base	As-soil
SRR8751109	PRJNA527922	Constant quality score at every base	As-soil
SRR8751110	PRJNA527922	Constant quality score at every base	As-soil
SRR8751111	PRJNA527922	Constant quality score at every base	As-soil
SRR8751112	PRJNA527922	Constant quality score at every base	As-soil
SRR8751113	PRJNA527922	Constant quality score at every base	As-soil
SRR8751114	PRJNA527922	Constant quality score at every base	As-soil
SRR8751115	PRJNA527922	Constant quality score at every base	As-soil
SRR8751116	PRJNA527922	Constant quality score at every base	As-soil
SRR8751117	PRJNA527922	Constant quality score at every base	As-soil
SRR8751118	PRJNA527922	Constant quality score at every base	As-soil
SRR8751119	PRJNA527922	Constant quality score at every base	As-soil
SRR8751120	PRJNA527922	Constant quality score at every base	As-soil
SRR8751121	PRJNA527922	Constant quality score at every base	As-soil
SRR8751122	PRJNA527922	Constant quality score at every base	As-soil
SRR8751123	PRJNA527922	Constant quality score at every base	As-soil
SRR8751124	PRJNA527922	Constant quality score at every base	As-soil
SRR8751125	PRJNA527922	Constant quality score at every base	As-soil
SRR8751126	PRJNA527922	Constant quality score at every base	As-soil
SRR8751127	PRJNA527922	Constant quality score at every base	As-soil

continued

Table S2 continued

Run	Study	Reason	Compartment
SRR8751128	PRJNA527922	Constant quality score at every base	As-soil
SRR8751129	PRJNA527922	Constant quality score at every base	As-soil
SRR8751130	PRJNA527922	Constant quality score at every base	As-soil
SRR8751131	PRJNA527922	Constant quality score at every base	As-soil
SRR8751132	PRJNA527922	Constant quality score at every base	As-soil
SRR8751133	PRJNA527922	Constant quality score at every base	As-soil
SRR8751134	PRJNA527922	Constant quality score at every base	As-soil
SRR8751135	PRJNA527922	Constant quality score at every base	As-soil
SRR8751136	PRJNA527922	Constant quality score at every base	As-soil
SRR8751137	PRJNA527922	Constant quality score at every base	As-soil
SRR8751138	PRJNA527922	Constant quality score at every base	As-soil
SRR8751139	PRJNA527922	Constant quality score at every base	As-soil
SRR8751140	PRJNA527922	Constant quality score at every base	As-soil
SRR8751141	PRJNA527922	Constant quality score at every base	As-soil
SRR8751142	PRJNA527922	Constant quality score at every base	As-soil
SRR8751143	PRJNA527922	Constant quality score at every base	As-soil
SRR8751144	PRJNA527922	Constant quality score at every base	As-soil
SRR8751145	PRJNA527922	Constant quality score at every base	As-soil
SRR8751146	PRJNA527922	Constant quality score at every base	As-soil
SRR8751147	PRJNA527922	Constant quality score at every base	As-soil
SRR8751148	PRJNA527922	Constant quality score at every base	As-soil
SRR8751149	PRJNA527922	Constant quality score at every base	As-soil
SRR8751150	PRJNA527922	Constant quality score at every base	As-soil
SRR8751151	PRJNA527922	Constant quality score at every base	As-soil
SRR8751152	PRJNA527922	Constant quality score at every base	As-soil
SRR8751153	PRJNA527922	Constant quality score at every base	As-soil
SRR8751154	PRJNA527922	Constant quality score at every base	As-soil
SRR8751155	PRJNA527922	Constant quality score at every base	As-soil
SRR8751156	PRJNA527922	Constant quality score at every base	As-soil
SRR8751157	PRJNA527922	Constant quality score at every base	As-soil
SRR8751158	PRJNA527922	Constant quality score at every base	As-soil
SRR8751159	PRJNA527922	Constant quality score at every base	As-soil
SRR8751160	PRJNA527922	Constant quality score at every base	As-soil
SRR8751161	PRJNA527922	Constant quality score at every base	As-soil
SRR8751162	PRJNA527922	Constant quality score at every base	As-soil
SRR8751163	PRJNA527922	Constant quality score at every base	As-soil
SRR8751164	PRJNA527922	Constant quality score at every base	As-soil
SRR8751165	PRJNA527922	Constant quality score at every base	As-soil
SRR8751166	PRJNA527922	Constant quality score at every base	As-soil
SRR8751167	PRJNA527922	Constant quality score at every base	As-soil
SRR12176171	PRJNA644740	Constant quality score at every base	As-soil
SRR12176172	PRJNA644740	Constant quality score at every base	As-soil
SRR12176173	PRJNA644740	Constant quality score at every base	As-soil
SRR12176174	PRJNA644740	Constant quality score at every base	As-soil
SRR12176175	PRJNA644740	Constant quality score at every base	As-soil
SRR12176176	PRJNA644740	Constant quality score at every base	As-soil
SRR12176177	PRJNA644740	Constant quality score at every base	As-soil
SRR12176178	PRJNA644740	Constant quality score at every base	As-soil
SRR12176179	PRJNA644740	Constant quality score at every base	As-soil
SRR12176180	PRJNA644740	Constant quality score at every base	As-soil
SRR12176181	PRJNA644740	Constant quality score at every base	As-soil
SRR12176182	PRJNA644740	Constant quality score at every base	As-soil
SRR12176183	PRJNA644740	Constant quality score at every base	As-soil
SRR12176184	PRJNA644740	Constant quality score at every base	As-soil
SRR12176185	PRJNA644740	Constant quality score at every base	As-soil
SRR12176186	PRJNA644740	Constant quality score at every base	As-soil
SRR12176187	PRJNA644740	Constant quality score at every base	As-soil
SRR12176188	PRJNA644740	Constant quality score at every base	As-soil
SRR12176189	PRJNA644740	Constant quality score at every base	As-soil
SRR12176190	PRJNA644740	Constant quality score at every base	As-soil
SRR12176191	PRJNA644740	Constant quality score at every base	As-soil
SRR12176192	PRJNA644740	Constant quality score at every base	As-soil
SRR12176193	PRJNA644740	Constant quality score at every base	As-soil
SRR12176194	PRJNA644740	Constant quality score at every base	As-soil
SRR12176195	PRJNA644740	Constant quality score at every base	As-soil
SRR12176196	PRJNA644740	Constant quality score at every base	As-soil
SRR12176197	PRJNA644740	Constant quality score at every base	As-soil
SRR12176198	PRJNA644740	Constant quality score at every base	As-soil
SRR12176199	PRJNA644740	Constant quality score at every base	As-soil
SRR12176200	PRJNA644740	Constant quality score at every base	As-soil
SRR12374203	PRJNA649896	Constant quality score at every base	As-soil
SRR12374204	PRJNA649896	Constant quality score at every base	As-soil
SRR12374205	PRJNA649896	Constant quality score at every base	As-soil

continued

Table S2 continued

Run	Study	Reason	Compartment
SRR12374206	PRJNA649896	Constant quality score at every base	As-soil
SRR12374207	PRJNA649896	Constant quality score at every base	As-soil
SRR12374208	PRJNA649896	Constant quality score at every base	As-soil
SRR12374209	PRJNA649896	Constant quality score at every base	As-soil
SRR12374210	PRJNA649896	Constant quality score at every base	As-soil
SRR12374211	PRJNA649896	Constant quality score at every base	As-soil
SRR12374212	PRJNA649896	Constant quality score at every base	As-soil
SRR12374213	PRJNA649896	Constant quality score at every base	As-soil
SRR12374214	PRJNA649896	Constant quality score at every base	As-soil
SRR12374215	PRJNA649896	Constant quality score at every base	As-soil
SRR12374216	PRJNA649896	Constant quality score at every base	As-soil
SRR12374217	PRJNA649896	Constant quality score at every base	As-soil
SRR12374218	PRJNA649896	Constant quality score at every base	As-soil
SRR12374219	PRJNA649896	Constant quality score at every base	As-soil
SRR12374220	PRJNA649896	Constant quality score at every base	As-soil
SRR12374221	PRJNA649896	Constant quality score at every base	As-soil
SRR12374222	PRJNA649896	Constant quality score at every base	As-soil
SRR12374223	PRJNA649896	Constant quality score at every base	As-soil
SRR12374224	PRJNA649896	Constant quality score at every base	As-soil
SRR12374225	PRJNA649896	Constant quality score at every base	As-soil
SRR12374226	PRJNA649896	Constant quality score at every base	As-soil
ERR2778187	PRJEB28491	Platform: Ion Torrent	Aquifer, depth n.s.
ERR2778188	PRJEB28491	Platform: Ion Torrent	Aquifer, depth n.s.
ERR2778189	PRJEB28492	Platform: Ion Torrent	Aquifer, depth n.s.
ERR2778190	PRJEB28492	Platform: Ion Torrent	Aquifer, depth n.s.
SRR12121545	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121546	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121547	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121548	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121549	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121550	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121551	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121552	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121553	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121554	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121555	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121556	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121557	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121558	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121559	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12121560	PRJNA642986	Platform: Ion Torrent	Paddy soil
SRR12344288	PRJNA649362	Platform: Ion Torrent	Aquifer, depth n.s.
SRR12344289	PRJNA649362	Platform: Ion Torrent	Aquifer, depth n.s.
SRR12344290	PRJNA649362	Platform: Ion Torrent	Aquifer, depth n.s.
SRR12344291	PRJNA649362	Platform: Ion Torrent	Aquifer, depth n.s.
SRR12344642	PRJNA649362	Platform: Ion Torrent	Aquifer, depth n.s.
SRR12344643	PRJNA649362	Platform: Ion Torrent	Aquifer, depth n.s.
SRR12344644	PRJNA649362	Platform: Ion Torrent	Aquifer, depth n.s.
SRR12344645	PRJNA649362	Platform: Ion Torrent	Aquifer, depth n.s.
SRR12344646	PRJNA649362	Platform: Ion Torrent	Aquifer, depth n.s.
SRR2990090	SRP067270	Platform: Ion Torrent	Aquifer, depth n.s.
SRR2990091	SRP067270	Platform: Ion Torrent	Aquifer, depth n.s.
SRR2990092	SRP067270	Platform: Ion Torrent	Aquifer, depth n.s.
SRR2993106	SRP067270	Platform: Ion Torrent	Aquifer, depth n.s.
SRR2993107	SRP067270	Platform: Ion Torrent	Aquifer, depth n.s.
SRR2993108	SRP067270	Platform: Ion Torrent	Aquifer, depth n.s.
SRR2993109	SRP067270	Platform: Ion Torrent	Aquifer, depth n.s.
SRR2993110	SRP067270	Platform: Ion Torrent	Aquifer, depth n.s.
SRR2993111	SRP067270	Platform: Ion Torrent	Aquifer, depth n.s.
SRR11462374	SRP154379	Platform: Ion Torrent	Aquifer, depth n.s.
SRR11463299	SRP154379	Platform: Ion Torrent	Aquifer, depth n.s.
SRR7537707	SRP154379	Platform: Ion Torrent	Aquifer, depth n.s.
SRR8148966	SRP154379	Platform: Ion Torrent	Aquifer, depth n.s.
SRR13285303	SRP298835	Platform: Ion Torrent	Paddy soil
SRR13390651	SRP298835	Platform: Ion Torrent	Paddy soil
SRR13439709	SRP298835	Platform: Ion Torrent	Paddy soil
SRR13439710	SRP298835	Platform: Ion Torrent	Paddy soil
SRR13439711	SRP298835	Platform: Ion Torrent	Paddy soil
SRR13439712	SRP298835	Platform: Ion Torrent	Paddy soil
SRR13439713	SRP298835	Platform: Ion Torrent	Paddy soil
SRR13439714	SRP298835	Platform: Ion Torrent	Paddy soil
SRR13439715	SRP298835	Platform: Ion Torrent	Paddy soil
SRR13439717	SRP298835	Platform: Ion Torrent	Paddy soil

continued

Table S2 continued

Run	Study	Reason	Compartment
SRR13439719	SRP298835	Platform: Ion Torrent	Paddy soil
SRR14148512	SRP313593	Platform: Ion Torrent	Aquifer, depth n.s.
SRR14148513	SRP313593	Platform: Ion Torrent	Aquifer, depth n.s.
SRR14148514	SRP313593	Platform: Ion Torrent	Aquifer, depth n.s.
SRR14148515	SRP313593	Platform: Ion Torrent	Aquifer, depth n.s.
SRR14148516	SRP313593	Platform: Ion Torrent	Aquifer, depth n.s.
SRR14148517	SRP313593	Platform: Ion Torrent	Aquifer, depth n.s.
SRR14148518	SRP313593	Platform: Ion Torrent	Aquifer, depth n.s.
SRR14148519	SRP313593	Platform: Ion Torrent	Aquifer, depth n.s.
SRR14148520	SRP313593	Platform: Ion Torrent	Aquifer, depth n.s.
SRR14148521	SRP313593	Platform: Ion Torrent	Aquifer, depth n.s.
SRR14148522	SRP313593	Platform: Ion Torrent	Aquifer, depth n.s.
ERR841906	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841907	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841908	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841909	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841910	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841911	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841912	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841913	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841914	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841915	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841916	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841917	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841918	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841919	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841920	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841921	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841922	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
ERR841923	PRJEB8996	Platform: 454 pyrosequencing	Aquifer, depth n.s.
SRR1501945	PRJNA253777	Platform: 454 pyrosequencing	Deep aquifer
SRR1503966	PRJNA253777	Platform: 454 pyrosequencing	Deep aquifer
SRR063965	SRP003266	Platform: 454 pyrosequencing	Paddy soil
SRR063966	SRP003266	Platform: 454 pyrosequencing	Paddy soil
SRR063967	SRP003266	Platform: 454 pyrosequencing	Paddy soil
SRR063968	SRP003266	Platform: 454 pyrosequencing	Paddy soil
SRR098973	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR098974	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR098975	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR098976	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR098977	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR098978	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR098979	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR098980	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR098981	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR098998	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR099000	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR099002	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR099016	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR099018	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR099020	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR099022	SRP005643	Platform: 454 pyrosequencing	Paddy soil
SRR1185105	SRP039499	Platform: 454 pyrosequencing	Shallow aquifer
SRR1185110	SRP039499	Platform: 454 pyrosequencing	Shallow aquifer
SRR1185115	SRP039499	Platform: 454 pyrosequencing	Shallow aquifer
SRR1185117	SRP039499	Platform: 454 pyrosequencing	Shallow aquifer
SRR1185215	SRP039499	Platform: 454 pyrosequencing	Shallow aquifer
SRR1185217	SRP039499	Platform: 454 pyrosequencing	Shallow aquifer
SRR1185219	SRP039499	Platform: 454 pyrosequencing	Shallow aquifer
SRR1185220	SRP039499	Platform: 454 pyrosequencing	Shallow aquifer
SRR1185410	SRP039499	Platform: 454 pyrosequencing	Shallow aquifer
SRR1463400	SRP043614	Platform: 454 pyrosequencing	Shallow aquifer
SRR1463401	SRP043614	Platform: 454 pyrosequencing	Shallow aquifer
SRR1463403	SRP043614	Platform: 454 pyrosequencing	Shallow aquifer
SRR1463404	SRP043614	Platform: 454 pyrosequencing	Shallow aquifer
SRR1463405	SRP043614	Platform: 454 pyrosequencing	Shallow aquifer
SRR1463406	SRP043614	Platform: 454 pyrosequencing	Shallow aquifer
SRR1463407	SRP043614	Platform: 454 pyrosequencing	Shallow aquifer
SRR1463408	SRP043614	Platform: 454 pyrosequencing	Shallow aquifer
SRR1463409	SRP043614	Platform: 454 pyrosequencing	Shallow aquifer
SRR1507061	SRP044066	Platform: 454 pyrosequencing	Aquifer, depth n.s.
SRR1507062	SRP044066	Platform: 454 pyrosequencing	Aquifer, depth n.s.
SRR1507063	SRP044066	Platform: 454 pyrosequencing	Aquifer, depth n.s.

continued

Table S2 continued

Run	Study	Reason	Compartment
SRR1569172	SRP046303	Platform: 454 pyrosequencing	Shallow aquifer
SRR1569180	SRP046303	Platform: 454 pyrosequencing	Shallow aquifer
SRR1569181	SRP046303	Platform: 454 pyrosequencing	Shallow aquifer
SRR8327188	SRP173566	Platform: 454 pyrosequencing	Aquifer, depth n.s.
SRR8327189	SRP173566	Platform: 454 pyrosequencing	Aquifer, depth n.s.
SRR8327190	SRP173566	Platform: 454 pyrosequencing	Aquifer, depth n.s.
SRR8327191	SRP173566	Platform: 454 pyrosequencing	Aquifer, depth n.s.
SRR8327192	SRP173566	Platform: 454 pyrosequencing	Aquifer, depth n.s.
SRR8327193	SRP173566	Platform: 454 pyrosequencing	Aquifer, depth n.s.
SRR13381176	PRJNA690162	Platform: PacBio SMRT	Paddy soil
SRR13381177	PRJNA690162	Platform: PacBio SMRT	Paddy soil
SRR13381178	PRJNA690162	Platform: PacBio SMRT	Paddy soil
SRR13381179	PRJNA690162	Platform: PacBio SMRT	Paddy soil
SRR13381180	PRJNA690162	Platform: PacBio SMRT	Paddy soil
SRR13381181	PRJNA690162	Platform: PacBio SMRT	Paddy soil
SRR13381182	PRJNA690162	Platform: PacBio SMRT	Paddy soil
SRR13381183	PRJNA690162	Platform: PacBio SMRT	Paddy soil
SRR13381184	PRJNA690162	Platform: PacBio SMRT	Paddy soil
SRR13381185	PRJNA690162	Platform: PacBio SMRT	Paddy soil
SRR13381186	PRJNA690162	Platform: PacBio SMRT	Paddy soil
SRR13381187	PRJNA690162	Platform: PacBio SMRT	Paddy soil
SRR17497199	PRJNA795569	Platform: PacBio SMRT	Shallow aquifer
SRR17497200	PRJNA795569	Platform: PacBio SMRT	Shallow aquifer
SRR17497201	PRJNA795569	Platform: PacBio SMRT	Shallow aquifer
SRR17497202	PRJNA795569	Platform: PacBio SMRT	Shallow aquifer
SRR17497203	PRJNA795569	Platform: PacBio SMRT	Shallow aquifer
SRR17497204	PRJNA795569	Platform: PacBio SMRT	Shallow aquifer
SRR17497205	PRJNA795569	Platform: PacBio SMRT	Shallow aquifer
SRR17497206	PRJNA795569	Platform: PacBio SMRT	Shallow aquifer
SRR17497207	PRJNA795569	Platform: PacBio SMRT	Shallow aquifer
SRR17497208	PRJNA795569	Platform: PacBio SMRT	Shallow aquifer
SRR17497209	PRJNA795569	Platform: PacBio SMRT	Shallow aquifer
SRR25607570	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607571	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607572	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607573	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607574	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607575	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607576	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607577	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607578	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607579	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607580	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607581	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607582	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607583	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607584	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607585	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607586	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607587	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607588	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607589	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR25607590	PRJNA1004417	Size outlier for an amplicon deposit	As-soil
SRR18710329	PRJNA824136	Size outlier for an amplicon deposit	As-soil
SRR18710330	PRJNA824136	Size outlier for an amplicon deposit	As-soil
SRR24565913	SRP437741	Size outlier for an amplicon deposit	As-soil
SRR24565914	SRP437741	Size outlier for an amplicon deposit	As-soil
SRR24565915	SRP437741	Size outlier for an amplicon deposit	As-soil
SRR24565916	SRP437741	Size outlier for an amplicon deposit	As-soil
SRR29014468	SRP507442	Size outlier for an amplicon deposit	As-soil
SRR29014472	SRP507442	Size outlier for an amplicon deposit	As-soil
SRR29014479	SRP507442	Size outlier for an amplicon deposit	As-soil

Compartment records the arm a run would have joined had it been retained. As-soil, arsenic-contaminated non-paddy soil.

S3 The marker-taxa lookup

Table S3 is the curated lookup in full. It maps genera to the panel genes for which they carry documented evidence, and it is the instrument the entire bridge rests on. A gene is scored present in a sample when any taxon mapped to it is present, so the contents of this table determine what the bridge can detect and, as the main text shows, largely determine each gene's detection rate.

The lookup is frozen at version 1.0. Each attribution carries an evidence tier graded by the curator and at least one key reference. The caveat recorded against each row travels with the machine-readable copy in the code repository.

Two limits follow directly from the table's construction and are stated here because they bound every result computed through the bridge. Genes carried on mobile elements have no stable taxonomic signal and therefore cannot appear, which is why the arsenic detoxification operon is absent from the lookup while PICRUSt2 recovers it. And a genus mapped to several genes contributes to all of them at once, so genes sharing marker taxa cannot vary independently.

Table S3. The curated marker-taxa lookup in full. A gene is scored present in a sample when any taxon mapped to it is present, so this table determines what the bridge can and cannot detect.

Guild	Taxon	Rank	Panel genes	Evidence	Key reference
As methylation	*Anabaena*	genus	arsM	inferred	Yin 2011
As methylation	*Clostridium*	genus	arsM	inferred	Wang 2015
As methylation	*Methanosarcina*	genus	arsM	inferred	Wang 2015
As methylation	*Nostoc*	genus	arsM	inferred	Yin 2011
As methylation	*Rhodopseudomonas*	genus	arsM	inferred	Qin 2006
As methylation	*Streptomyces*	genus	arsM	inferred	Wang 2014
As oxidation aerobic	*Achromobacter*	genus	aioA	moderate	Cai 2009
As oxidation aerobic	*Agrobacterium*	genus	aioA	moderate	Salmassi 2002
As oxidation aerobic	*Alcaligenes*	genus	aioA	moderate	Anderson 1992
As oxidation aerobic	*Herminiimonas*	genus	aioA	strong	Muller 2007
As oxidation aerobic	*Hydrogenophaga*	genus	aioA	moderate	vanden Hoven 2007
As oxidation aerobic	*Pseudomonas*	genus	aioA	inferred	Various
As oxidation aerobic	*Rhizobium*	genus	aioA	moderate	Cai 2009
As oxidation aerobic	*Thiobacillus*	genus	aioA, soxB	moderate	Battaglia Brunet 2002
As oxidation aerobic	*Thiomonas*	genus	aioA	strong	Bryan 2009
As oxidation aerobic	*Variovorax*	genus	aioA	moderate	Wang 2018
As oxidation anaerobic	*Alkalilimnicola*	genus	arxA	strong	Hoeft 2007
As oxidation anaerobic	*Ectothiorhodospira*	genus	arxA	moderate	Zargar 2010
As oxidation anaerobic	*Halorhodospira*	genus	arxA	moderate	Zargar 2012
As respiration	*Alkaliphilus*	genus	arrA	moderate	Fisher 2008
As respiration	*Bacillus*	genus	arrA	inferred	Switzer Blum 1998
As respiration	*Chrysiogenes*	genus	arrA	strong	Macy 1996
As respiration	*Desulfitobacterium*	genus	arrA	moderate	Niggemyer 2001
As respiration	*Desulfosporosinus*	genus	arrA, dsrA, dsrB	moderate	Liu 2004
As respiration	*Geobacter*	genus	arrA	inferred	Stolz 2006
As respiration	*Shewanella*	genus	arrA	inferred	Saltikov 2003
As respiration	*Sulfurospirillum*	genus	arrA	moderate	Stolz 2006
As respiration	*Wolinella*	genus	arrA	moderate	Stolz Oremland 1999
Fe oxidation	*Gallionella*	genus	-	strong	Hallbeck Pedersen 2014
Fe oxidation	*Leptothrix*	genus	-	strong	Emerson Weiss 2004
Fe oxidation	*Mariprofundus*	genus	-	strong	Emerson 2007
Fe oxidation	*Sideroxydans*	genus	-	strong	Emerson 2007
Fe reduction	*Albidiferax*	genus	feoB	moderate	Ramana 2009
Fe reduction	*Anaeromyxobacter*	genus	omcS, feoB	strong	Sanford 2002
Fe reduction	*Desulfuromonas*	genus	mtrA, feoB	moderate	Roden Lovley 1993
Fe reduction	*Geobacter*	genus	mtrA, mtrC, omcS, omcZ, feoB	strong	Lovley 1988; Holmes 2002
Fe reduction	*Geobacteraceae*	family	mtrA, mtrC, omcS, omcZ, feoB	strong	Holmes 2004
Fe reduction	*Geothermobacter*	genus	mtrA, omcS, feoB	strong	Kashefi 2003
Fe reduction	*Geothrix*	genus	mtrA, omcS, feoB	strong	Coates 1999
Fe reduction	*Rhodoferax*	genus	feoB	moderate	Finneran 2003
Fe reduction	*Shewanella*	genus	mtrA, mtrB, mtrC, omcS, feoB	strong	Myers Nealson 1988; Newman Kolter 2000
Fe reduction	*Shewanellaceae*	family	mtrA, mtrB, mtrC, feoB	strong	Hau Gralnick 2007
Fe reduction	*Thermincola*	genus	feoB	moderate	Wrighton 2008

continued

Table S3 continued

Guild	Taxon	Rank	Panel genes	Evidence	Key reference
N fixation	*Anabaena*	genus	nifH	strong	Wolk 1996
N fixation	*Azospirillum*	genus	nifH	strong	Bashan 2010
N fixation	*Azotobacter*	genus	nifH	strong	Dixon Kahn 2004
N fixation	*Bradyrhizobium*	genus	nifH	strong	Young 1985
N fixation	*Clostridium*	genus	nifH	moderate	Dilworth 1966
N fixation	*Frankia*	genus	nifH	strong	Benson Silvester 1993
N fixation	*Geobacter*	genus	nifH	moderate	Coppi 2001
N fixation	*Methanosarcina*	genus	nifH	moderate	Belay 1984
N fixation	*Nostoc*	genus	nifH	strong	Meeks 2001
N fixation	*Paenibacillus*	genus	nifH	moderate	Berge 2002
N fixation	*Rhizobium*	genus	nifH	strong	Young 1985
S oxidation	*Beggiatoa*	genus	soxB	moderate	Mussmann 2007
S oxidation	*Halothiobacillus*	genus	soxB	moderate	Kelly Wood 2000
S oxidation	*Sulfuricurvum*	genus	soxB	strong	Kodama Watanabe 2004
S oxidation	*Sulfurimonas*	genus	soxB	strong	Inagaki 2003
S oxidation	*Thiobacillus*	genus	soxB	strong	Friedrich 2001
S oxidation	*Thiomonas*	genus	aioA, soxB	moderate	Bryan 2009
S oxidation	*Thiothrix*	genus	soxB	moderate	Larkin Strohl 1983
S reduction	*Desulfobacter*	genus	dsrA, dsrB	strong	Widdel Pfennig 1981
S reduction	*Desulfobacteraceae*	family	dsrA, dsrB	strong	–
S reduction	*Desulfobacterium*	genus	dsrA, dsrB	strong	Bak Widdel 1986
S reduction	*Desulfobulbus*	genus	dsrA, dsrB	strong	Widdel Pfennig 1981
S reduction	*Desulfococcus*	genus	dsrA, dsrB	strong	Widdel 1980
S reduction	*Desulfomicrobium*	genus	dsrA, dsrB	strong	Rozanova 1988
S reduction	*Desulforhopalus*	genus	dsrA, dsrB	strong	Isaksen Teske 1996
S reduction	*Desulfosporosinus*	genus	dsrA, dsrB	strong	Stackebrandt 1997
S reduction	*Desulfotomaculum*	genus	dsrA, dsrB	strong	Klemps 1985
S reduction	*Desulfovibrio*	genus	dsrA, dsrB	strong	Wagner 1998
S reduction	*Desulfovibrionaceae*	family	dsrA, dsrB	strong	–
S reduction	*Syntrophobacter*	genus	dsrA, dsrB	moderate	Boone Bryant 1980
S reduction	*Syntrophobacteraceae*	family	dsrA, dsrB	moderate	–
ammonia oxidation	*Candidatus Nitrosotalea*	genus	amoA	strong	Lehtovirta Morley 2011
ammonia oxidation	*Candidatus Nitrosotenuis*	genus	amoA	moderate	Lebedeva 2013
ammonia oxidation	*Nitrosococcus*	genus	amoA	strong	Ward OMullan 2002
ammonia oxidation	*Nitrosomonas*	genus	amoA	strong	Koops 2003
ammonia oxidation	*Nitrosopumilus*	genus	amoA	strong	Konneke 2005
ammonia oxidation	*Nitrososphaera*	genus	amoA	strong	Tourna 2011
ammonia oxidation	*Nitrosospira*	genus	amoA	strong	Stephen 1996
denitrification	*Achromobacter*	genus	nirK	moderate	Coyne 2008
denitrification	*Azoarcus*	genus	nirS, nosZ	moderate	Reinhold Hurek 2000
denitrification	*Bradyrhizobium*	genus	nirK, nosZ	moderate	Sameshima Saito 2006
denitrification	*Denitratisoma*	genus	nirS, nosZ	moderate	Fahrbach 2006
denitrification	*Paracoccus*	genus	nirK, nosZ	moderate	Baker 1998
denitrification	*Pseudomonas*	genus	nirK, nirS, nosZ	moderate	Zumft 1997
denitrification	*Rhodanobacter*	genus	nirS, nosZ	moderate	Green 2010
denitrification	*Thauera*	genus	nirS, nosZ	moderate	Macy 1993
fermentation	*Bacteroides*	genus	pflD	moderate	Macy 1979
fermentation	*Clostridium*	genus	pflD, hydA	moderate	Cato 1986
methanogenesis	*Methanobacterium*	genus	mcrA	strong	Boone 1993
methanogenesis	*Methanobrevibacter*	genus	mcrA	strong	Miller 1986
methanogenesis	*Methanocella*	genus	mcrA	strong	Sakai 2008
methanogenesis	*Methanocellaceae*	family	mcrA	strong	–
methanogenesis	*Methanoculleus*	genus	mcrA	strong	Maestrojuan 1990
methanogenesis	*Methanomassiliicoccus*	genus	mcrA	strong	Dridi 2012
methanogenesis	*Methanomicrobiales*	order	mcrA	strong	–
methanogenesis	*Methanoregula*	genus	mcrA	strong	Brauer 2011
methanogenesis	*Methanosaeta*	genus	mcrA	strong	Smith Ingram Smith 2007
methanogenesis	*Methanosarcina*	genus	mcrA	strong	Boone 1993
methanogenesis	*Methanosarcinaceae*	family	mcrA	strong	–
methanogenesis	*Methanothrix*	genus	mcrA	strong	Patel 1992

Evidence tier is the curator's grading of how well documented the attribution is. The lookup is frozen at v1.0 and its caveats travel with each row in the machine-readable copy. Reference keys are reproduced as deposited, with underscores replaced by spaces; a key naming two surnames may be two authors or one compound surname, and the key does not distinguish them. Two guilds carry a curator caveat, lifted here from rows that name no taxon. *arsM* is a textbook horizontally transferred gene, so any taxon-level inference for it is weak at best and shotgun quantification is canonical. The fermentation markers *pflD* and *hydA* are illustrative only, these functions being broadly distributed across anaerobic bacteria.

S4 Phylogeny of the marker taxa

The tree in Figure 4 was built from this corpus rather than from an external reference, so that the sequences carrying the test are the ones the analysis is based on. One representative 16S sequence was taken per marker genus, choosing the most abundant observed sequence variant assigned to that genus. Representatives were aligned with MAFFT (Katoh and Standley, 2013) and an approximately maximum-likelihood tree inferred with FastTree (Price et al., 2010) under a GTR model, with SH-like local support. 71 of the marker genera yielded a usable sequence and enter the tree; the remainder had no sequence variant assigned to them anywhere in the corpus.

Clustering was measured per gene as the mean pairwise phylogenetic distance between that gene's marker taxa, compared against 999 draws of the same number of tips from the tree at random. The standardised effect size is the observed mean distance minus the null mean divided by the null standard deviation, so negative values indicate marker taxa more closely related than chance.

Table S4 gives the result per gene, with the agreement between routes from Table S5 repeated in the last column. The correlation between those two columns is the test reported in the main text, and it is not significant ($\rho = -0.19$, $p = 0.62$). Six genes carry no agreement value because PICRUSt2 cannot reach them at all, which is itself discussed in the main text.

Table S4. Phylogenetic clustering of each gene's marker taxa, against 999 random draws of the same number of tips from the tree. Negative SES indicates marker taxa closer together than chance.

Gene	Taxa	MPD	Null MPD	SES	p	Agreement ρ
dsrA	12	0.3670	0.6141	-1.68	0.006	0.470
mtrA	6	0.2382	0.6179	-1.68	0.001	–
dsrB	12	0.3670	0.6129	-1.62	0.005	0.470
feoB	9	0.3367	0.6155	-1.60	0.004	0.136
aioA	10	0.3583	0.6140	-1.55	0.002	0.219
omcS	6	0.2916	0.6244	-1.42	0.002	–
mtrC	3	0.2522	0.5976	-1.10	0.017	–
arrA	7	0.4140	0.6226	-1.00	0.089	–
soxB	7	0.4333	0.6197	-0.92	0.137	0.217
nifH	11	0.5045	0.6134	-0.71	0.271	0.265
nirK	4	0.4597	0.5975	-0.52	0.363	0.022
arsM	6	0.6543	0.6080	+0.21	0.727	-0.118
nosZ	7	0.8846	0.6153	+1.34	0.850	0.071
nirS	5	1.0380	0.6276	+1.59	0.869	0.137
mcrA	9	1.0384	0.6168	+2.35	0.978	0.906

MPD, mean pairwise phylogenetic distance. p is the one-sided probability of clustering at least this strong under the null. Agreement ρ is the median agreement from Table S5, repeated here because the correlation between the last two columns is the test reported in the main text.

S5 Agreement between the two inference routes

Table S5 gives every gene-study comparison behind Figure 3. The two routes share only the sequence-variant table and have no reference data, lookup or algorithm in common, so their agreement measures the inference rather than the data.

PICRUSt2 (Douglas et al., 2020) output exists for 67 of the 78 retained studies. Of the remaining 11, 9 deposit libraries that are mostly not bacterial 16S and are removed by the bacterial-fraction gate in any case, 1 failed PICRUSt2's alignment threshold for every sequence variant, and 1, of mixed composition and with over 100,000 variants, was stopped before completion. The primary gene matrix therefore rests on the bridge alone and this comparison is reported separately. A matrix pooling a route present for part of the corpus with one present throughout would be unbalanced by study rather than by biology. Comparisons are further restricted to genes that map to a KEGG ortholog, giving 698 comparisons across 57 studies and 13 genes. Of the 741 study-gene cells behind the median, 15 are missing because the study's PICRUSt2 output lacks the gene's ortholog and 28 because fewer than five samples carry both routes.

Correlations are Spearman, taken within a study so that laboratory and primer effects are held constant, and summarised as a median across studies for each gene. Whether PICRUSt2's own confidence metric anticipates these values is tested in Section S9.

Figure 3b,c tests study size and joint detection within genes, which is the comparison that holds the gene fixed. Between genes the picture differs: across the 13 per-gene medians, agreement is higher for genes scored in larger studies ($\rho = +0.73$, $p = 0.005$) and for genes that the two routes detect together less often ($\rho = -0.61$, $p = 0.03$;

without *mcrA*, $\rho = +0.65$ and -0.53, $p = 0.02$ and 0.08). With thirteen genes these associations cannot be separated from which studies score each gene or from carriage fidelity (Section S12), and they are not tested as a claim. The high agreement on *mcrA* is not an artefact of samples in which neither route detects the gene: in the 11 studies in which both routes detect it in at least 90 % of samples, its median agreement is $\rho = +0.96$.

Table S5. Agreement between the marker-taxa and PICRUSt2 routes, one row per gene and study. Correlations are taken within a study so that laboratory and primer effects are held constant.

Gene	Study	*n*	Spearman ρ	*p*	Both routes
aioA	PRJEB26724	40	0.774	4.7e-09	20
aioA	PRJEB27237	52	-0.019	0.89	5
aioA	PRJEB45341	108	0.055	0.57	108
aioA	PRJEB45605	78	0.233	0.04	76
aioA	PRJEB45805	131	0.588	1.6e-13	104
aioA	PRJNA1026071	6	0.314	0.54	6
aioA	PRJNA1091876	8	0.114	0.79	4
aioA	PRJNA1110177	37	0.048	0.78	27
aioA	PRJNA1135448	32	0.038	0.84	31
aioA	PRJNA1142403	44	-0.565	6.4e-05	44
aioA	PRJNA1142719	44	-0.638	3.2e-06	2
aioA	PRJNA1166154	80	0.340	0.002	2
aioA	PRJNA1284197	8	0.381	0.35	5
aioA	PRJNA270654	9	-0.274	0.48	2
aioA	PRJNA344926	83	0.350	0.0012	6
aioA	PRJNA354810	26	0.185	0.37	22
aioA	PRJNA361507	6	0.500	0.31	6
aioA	PRJNA471565	71	0.352	0.0026	10
aioA	PRJNA478283	18	0.332	0.18	11
aioA	PRJNA478333	18	-0.020	0.94	12
aioA	PRJNA493691	18	0.399	0.1	1
aioA	PRJNA528962	5	0.300	0.62	5
aioA	PRJNA531524	31	-0.239	0.19	29
aioA	PRJNA593718	23	0.205	0.35	23
aioA	PRJNA628856	18	-0.065	0.8	17
aioA	PRJNA640466	32	0.339	0.058	28
aioA	PRJNA690162	45	0.134	0.38	5
aioA	PRJNA694468	40	-0.055	0.73	22
aioA	PRJNA716501	15	-0.375	0.17	15
aioA	PRJNA766239	12	-0.140	0.66	11
aioA	PRJNA786456	32	-0.710	5.2e-06	12
aioA	PRJNA820970	24	-0.179	0.4	24
aioA	PRJNA823829	18	0.292	0.24	7
aioA	PRJNA858795	111	-0.589	1e-11	47
aioA	PRJNA873282	33	-0.100	0.58	6
aioA	PRJNA980983	10	0.755	0.012	6
aioA	SRP094596	63	0.160	0.21	24
aioA	SRP166039	28	0.134	0.5	27
aioA	SRP186087	61	0.243	0.059	26
aioA	SRP252167	22	0.375	0.086	3
aioA	SRP278891	38	-0.135	0.42	7
aioA	SRP286951	23	0.513	0.012	16
aioA	SRP329786	24	-0.113	0.6	4
aioA	SRP339895	18	0.862	4.3e-06	8
aioA	SRP370321	6	-0.393	0.44	1
aioA	SRP376535	9	0.650	0.058	8
aioA	SRP406096	13	0.804	0.00093	3
aioA	SRP432433	8	0.667	0.071	7
aioA	SRP452018	19	0.316	0.19	18
aioA	SRP466511	31	0.179	0.34	31
aioA	SRP468584	108	0.382	4.4e-05	60
aioA	SRP507371	9	0.750	0.02	8
aioA	SRP507442	10	0.709	0.022	10
aioA	SRP580424	32	0.740	1.3e-06	31
aioA	SRP657214	6	0.600	0.21	6
aioA	SRP666293	80	0.197	0.08	39
arsM	PRJEB26724	40	0.792	1.2e-09	20
arsM	PRJEB27237	52	-0.015	0.92	20
arsM	PRJEB29426	8	-0.082	0.85	1
arsM	PRJEB45341	108	-0.260	0.0065	108
arsM	PRJEB45605	78	0.052	0.65	78
arsM	PRJEB45805	131	-0.316	0.00023	131
arsM	PRJNA1026071	6	-0.943	0.0048	6
arsM	PRJNA1091876	8	-0.048	0.91	8

continued

Table S5 continued

Gene	Study	*n*	Spearman ρ	*p*	Both routes
arsM	PRJNA1110177	37	0.538	0.00059	22
arsM	PRJNA1135448	32	-0.579	0.00052	32
arsM	PRJNA1142403	44	-0.816	1.5e-11	44
arsM	PRJNA1142719	44	-0.898	1.3e-16	44
arsM	PRJNA1166154	80	-0.113	0.32	15
arsM	PRJNA1284197	8	-0.381	0.35	8
arsM	PRJNA344926	83	0.906	6.1e-32	6
arsM	PRJNA354810	26	-0.165	0.42	23
arsM	PRJNA361507	6	-0.515	0.3	6
arsM	PRJNA471565	71	-0.044	0.72	66
arsM	PRJNA478283	18	-0.682	0.0018	18
arsM	PRJNA478333	18	-0.348	0.16	18
arsM	PRJNA493691	18	0.120	0.64	16
arsM	PRJNA528962	5	0.400	0.5	5
arsM	PRJNA531524	31	0.328	0.072	29
arsM	PRJNA593718	23	-0.154	0.48	14
arsM	PRJNA628856	18	0.380	0.12	16
arsM	PRJNA640466	32	0.494	0.0041	20
arsM	PRJNA690162	45	0.218	0.15	3
arsM	PRJNA694468	40	-0.123	0.45	26
arsM	PRJNA716501	15	-0.475	0.073	15
arsM	PRJNA766239	12	0.287	0.37	12
arsM	PRJNA786456	32	-0.157	0.39	32
arsM	PRJNA820970	24	-0.269	0.2	24
arsM	PRJNA823829	18	-0.759	0.00026	18
arsM	PRJNA858795	111	0.861	1e-33	110
arsM	PRJNA873282	33	0.349	0.047	33
arsM	PRJNA980983	10	0.715	0.02	6
arsM	SRP094596	63	0.451	0.00021	22
arsM	SRP166039	28	-0.135	0.49	21
arsM	SRP186087	61	-0.350	0.0056	38
arsM	SRP252167	22	-0.494	0.019	22
arsM	SRP278891	38	0.074	0.66	15
arsM	SRP286951	23	-0.405	0.056	20
arsM	SRP329786	24	0.425	0.038	24
arsM	SRP339895	18	0.799	6.9e-05	7
arsM	SRP370321	6	-0.200	0.7	0
arsM	SRP376535	9	-0.270	0.48	5
arsM	SRP406096	13	0.404	0.17	2
arsM	SRP432433	8	-0.216	0.61	4
arsM	SRP452018	19	0.475	0.04	18
arsM	SRP466511	31	0.139	0.45	24
arsM	SRP468584	108	-0.176	0.068	108
arsM	SRP507371	9	-0.356	0.35	6
arsM	SRP507442	10	-0.394	0.26	10
arsM	SRP580424	32	0.635	9.4e-05	32
arsM	SRP657214	6	0.086	0.87	6
arsM	SRP666293	80	-0.141	0.21	52
dsrA	PRJEB26724	40	0.969	1e-24	20
dsrA	PRJEB27237	52	0.212	0.13	20
dsrA	PRJEB29426	8	0.082	0.85	1
dsrA	PRJEB45341	108	0.225	0.019	108
dsrA	PRJEB45605	78	0.467	1.6e-05	78
dsrA	PRJEB45805	131	0.813	4e-32	129
dsrA	PRJNA1026071	6	0.543	0.27	6
dsrA	PRJNA1091876	8	-0.220	0.6	5
dsrA	PRJNA1110177	37	0.454	0.0047	27
dsrA	PRJNA1135448	32	0.694	1.1e-05	32
dsrA	PRJNA1142403	44	0.074	0.64	44
dsrA	PRJNA1142719	44	0.622	6.5e-06	43
dsrA	PRJNA1166154	80	0.202	0.073	8
dsrA	PRJNA344926	83	0.998	1.6e-96	6
dsrA	PRJNA354810	26	0.223	0.27	12
dsrA	PRJNA361507	6	0.515	0.3	6
dsrA	PRJNA471565	71	0.470	3.5e-05	36
dsrA	PRJNA478283	18	-0.164	0.51	15
dsrA	PRJNA478333	18	-0.181	0.47	18
dsrA	PRJNA493691	18	0.669	0.0024	16
dsrA	PRJNA528962	5	1.000	1.4e-24	5
dsrA	PRJNA531524	31	0.609	0.00027	30
dsrA	PRJNA593718	23	-0.233	0.28	15
dsrA	PRJNA628856	18	0.760	0.00025	15
dsrA	PRJNA640466	32	0.722	3.1e-06	27

continued

Table S5 continued

Gene	Study	*n*	Spearman ρ	*p*	Both routes
dsrA	PRJNA690162	45	-0.057	0.71	21
dsrA	PRJNA694468	40	0.176	0.28	16
dsrA	PRJNA716501	15	0.600	0.018	14
dsrA	PRJNA766239	12	0.651	0.022	10
dsrA	PRJNA786456	32	0.561	0.00084	27
dsrA	PRJNA820970	24	0.762	1.5e-05	21
dsrA	PRJNA823829	18	0.164	0.52	18
dsrA	PRJNA858795	111	0.874	5.6e-36	56
dsrA	PRJNA873282	33	0.548	0.00096	31
dsrA	PRJNA980983	10	0.387	0.27	4
dsrA	SRP094596	63	0.182	0.15	20
dsrA	SRP166039	28	0.329	0.087	11
dsrA	SRP186087	61	0.472	0.00012	10
dsrA	SRP252167	22	0.293	0.19	19
dsrA	SRP286951	23	0.777	1.3e-05	15
dsrA	SRP329786	24	0.385	0.063	23
dsrA	SRP339895	18	0.884	1.2e-06	6
dsrA	SRP376535	9	0.667	0.05	9
dsrA	SRP406096	13	0.565	0.044	2
dsrA	SRP432433	8	-0.241	0.57	4
dsrA	SRP452018	19	0.488	0.034	18
dsrA	SRP466511	31	0.506	0.0037	22
dsrA	SRP468584	108	-0.220	0.022	75
dsrA	SRP507371	9	0.383	0.31	9
dsrA	SRP507442	10	0.103	0.78	10
dsrA	SRP580424	32	0.852	6e-10	32
dsrA	SRP657214	6	0.314	0.54	6
dsrA	SRP666293	80	0.126	0.27	45
dsrB	PRJEB26724	40	0.969	1e-24	20
dsrB	PRJEB27237	52	0.212	0.13	20
dsrB	PRJEB29426	8	0.082	0.85	1
dsrB	PRJEB45341	108	0.225	0.019	108
dsrB	PRJEB45605	78	0.468	1.6e-05	78
dsrB	PRJEB45805	131	0.813	4.6e-32	129
dsrB	PRJNA1026071	6	0.543	0.27	6
dsrB	PRJNA1091876	8	-0.220	0.6	5
dsrB	PRJNA1110177	37	0.454	0.0047	27
dsrB	PRJNA1135448	32	0.699	8.4e-06	32
dsrB	PRJNA1142403	44	0.074	0.64	44
dsrB	PRJNA1142719	44	0.622	6.5e-06	43
dsrB	PRJNA1166154	80	0.202	0.073	8
dsrB	PRJNA344926	83	0.998	1.6e-96	6
dsrB	PRJNA354810	26	0.247	0.22	12
dsrB	PRJNA361507	6	0.515	0.3	6
dsrB	PRJNA471565	71	0.470	3.5e-05	36
dsrB	PRJNA478283	18	-0.164	0.51	15
dsrB	PRJNA478333	18	-0.181	0.47	18
dsrB	PRJNA493691	18	0.669	0.0024	16
dsrB	PRJNA528962	5	1.000	1.4e-24	5
dsrB	PRJNA531524	31	0.614	0.00024	30
dsrB	PRJNA593718	23	-0.233	0.28	15
dsrB	PRJNA628856	18	0.717	0.00082	15
dsrB	PRJNA640466	32	0.724	2.8e-06	27
dsrB	PRJNA690162	45	-0.057	0.71	21
dsrB	PRJNA694468	40	0.176	0.28	16
dsrB	PRJNA716501	15	0.600	0.018	14
dsrB	PRJNA766239	12	0.651	0.022	10
dsrB	PRJNA786456	32	0.561	0.00084	27
dsrB	PRJNA820970	24	0.762	1.5e-05	21
dsrB	PRJNA823829	18	0.164	0.52	18
dsrB	PRJNA858795	111	0.874	6.9e-36	56
dsrB	PRJNA873282	33	0.548	0.00096	31
dsrB	PRJNA980983	10	0.387	0.27	4
dsrB	SRP094596	63	0.183	0.15	20
dsrB	SRP166039	28	0.338	0.079	11
dsrB	SRP186087	61	0.480	9.1e-05	10
dsrB	SRP252167	22	0.293	0.19	19
dsrB	SRP286951	23	0.777	1.3e-05	15
dsrB	SRP329786	24	0.385	0.063	23
dsrB	SRP339895	18	0.884	1.2e-06	6
dsrB	SRP376535	9	0.667	0.05	9
dsrB	SRP406096	13	0.565	0.044	2
dsrB	SRP432433	8	-0.241	0.57	4

continued

Table S5 continued

Gene	Study	*n*	Spearman ρ	*p*	Both routes
dsrB	SRP452018	19	0.488	0.034	18
dsrB	SRP466511	31	0.498	0.0043	22
dsrB	SRP468584	108	-0.220	0.022	75
dsrB	SRP507371	9	0.383	0.31	9
dsrB	SRP507442	10	0.103	0.78	10
dsrB	SRP580424	32	0.858	3.4e-10	32
dsrB	SRP657214	6	0.314	0.54	6
dsrB	SRP666293	80	0.126	0.26	45
feoB	PRJEB26724	40	0.848	4.7e-12	20
feoB	PRJEB27237	52	0.308	0.026	35
feoB	PRJEB29426	8	0.071	0.87	8
feoB	PRJEB45341	108	0.053	0.58	108
feoB	PRJEB45605	78	0.119	0.3	78
feoB	PRJEB45805	131	-0.271	0.0018	131
feoB	PRJNA1026071	6	-0.543	0.27	6
feoB	PRJNA1091876	8	0.082	0.85	3
feoB	PRJNA1110177	37	0.597	9.7e-05	32
feoB	PRJNA1135448	32	0.611	0.00021	32
feoB	PRJNA1142403	44	0.436	0.0031	44
feoB	PRJNA1142719	44	-0.687	2.6e-07	44
feoB	PRJNA1166154	80	-0.163	0.15	79
feoB	PRJNA1284197	8	-0.062	0.88	2
feoB	PRJNA270654	9	0.548	0.13	1
feoB	PRJNA344926	83	0.543	1.1e-07	6
feoB	PRJNA354810	26	-0.413	0.036	25
feoB	PRJNA361507	6	1.000	1.8e-32	6
feoB	PRJNA471565	71	0.436	0.00015	70
feoB	PRJNA478283	18	-0.579	0.012	18
feoB	PRJNA478333	18	-0.160	0.53	18
feoB	PRJNA493691	18	0.326	0.19	16
feoB	PRJNA531524	31	0.305	0.095	31
feoB	PRJNA593718	23	-0.130	0.55	22
feoB	PRJNA628856	18	0.373	0.13	18
feoB	PRJNA640466	32	-0.148	0.42	32
feoB	PRJNA690162	45	-0.209	0.17	45
feoB	PRJNA694468	40	-0.121	0.46	37
feoB	PRJNA716501	15	0.050	0.86	15
feoB	PRJNA766239	12	0.336	0.29	12
feoB	PRJNA786456	32	0.202	0.27	32
feoB	PRJNA820970	24	0.525	0.0084	24
feoB	PRJNA823829	18	0.562	0.015	18
feoB	PRJNA858795	111	-0.548	5e-10	67
feoB	PRJNA873282	33	0.543	0.0011	32
feoB	PRJNA980983	10	0.728	0.017	6
feoB	SRP094596	63	0.233	0.066	63
feoB	SRP166039	28	0.120	0.54	26
feoB	SRP186087	61	0.454	0.00023	37
feoB	SRP252167	22	0.091	0.69	22
feoB	SRP278891	38	0.201	0.23	17
feoB	SRP286951	23	0.752	3.5e-05	22
feoB	SRP329786	24	0.242	0.26	24
feoB	SRP339895	18	-0.228	0.36	8
feoB	SRP370321	6	-0.417	0.41	0
feoB	SRP376535	9	-0.100	0.8	8
feoB	SRP406096	13	0.309	0.3	1
feoB	SRP432433	8	-0.390	0.34	5
feoB	SRP452018	19	0.108	0.66	17
feoB	SRP466511	31	-0.437	0.014	31
feoB	SRP468584	108	0.191	0.048	108
feoB	SRP507371	9	0.186	0.63	6
feoB	SRP507442	10	0.152	0.68	10
feoB	SRP580424	32	0.426	0.015	32
feoB	SRP657214	6	-0.086	0.87	6
feoB	SRP666293	80	-0.046	0.68	52
hydA	PRJEB26724	40	0.892	1.2e-14	20
hydA	PRJEB27237	52	0.056	0.69	15
hydA	PRJEB45341	108	0.021	0.83	108
hydA	PRJEB45605	78	0.702	7.8e-13	78
hydA	PRJEB45805	131	0.183	0.037	131
hydA	PRJNA1026071	6	-0.899	0.015	6
hydA	PRJNA1091876	8	0.667	0.071	8
hydA	PRJNA1110177	37	0.625	3.6e-05	15
hydA	PRJNA1135448	32	0.087	0.63	32

continued

Table S5 continued

Gene	Study	*n*	Spearman ρ	*p*	Both routes
hydA	PRJNA1142403	44	-0.390	0.0089	44
hydA	PRJNA1142719	44	0.814	1.9e-11	44
hydA	PRJNA1166154	80	0.312	0.0049	11
hydA	PRJNA1284197	8	0.299	0.47	6
hydA	PRJNA344926	83	0.998	4.2e-99	6
hydA	PRJNA354810	26	0.233	0.25	12
hydA	PRJNA361507	6	-0.515	0.3	6
hydA	PRJNA471565	71	-0.192	0.11	65
hydA	PRJNA478283	18	0.668	0.0025	18
hydA	PRJNA478333	18	0.744	0.0004	18
hydA	PRJNA493691	18	0.240	0.34	16
hydA	PRJNA528962	5	0.600	0.28	5
hydA	PRJNA531524	31	0.569	0.00084	28
hydA	PRJNA593718	23	0.311	0.15	14
hydA	PRJNA628856	18	0.443	0.066	16
hydA	PRJNA640466	32	0.379	0.032	19
hydA	PRJNA690162	45	0.107	0.48	3
hydA	PRJNA694468	40	0.515	0.00068	19
hydA	PRJNA716501	15	0.739	0.0016	15
hydA	PRJNA766239	12	0.399	0.2	11
hydA	PRJNA786456	32	0.179	0.33	32
hydA	PRJNA820970	24	0.810	1.7e-06	23
hydA	PRJNA823829	18	-0.104	0.68	17
hydA	PRJNA858795	111	0.444	1.1e-06	62
hydA	PRJNA873282	33	0.631	8.2e-05	33
hydA	PRJNA980983	10	0.948	2.9e-05	6
hydA	SRP094596	63	0.356	0.0042	17
hydA	SRP166039	28	0.737	7.9e-06	18
hydA	SRP186087	61	0.789	4.2e-14	26
hydA	SRP252167	22	0.623	0.002	22
hydA	SRP278891	38	-0.003	0.99	4
hydA	SRP286951	23	0.780	1.1e-05	17
hydA	SRP329786	24	0.102	0.64	24
hydA	SRP339895	18	0.652	0.0034	5
hydA	SRP376535	9	0.000	1	5
hydA	SRP406096	13	0.637	0.019	2
hydA	SRP432433	8	-0.577	0.13	1
hydA	SRP452018	19	0.343	0.15	17
hydA	SRP466511	31	0.285	0.12	20
hydA	SRP468584	108	0.238	0.013	107
hydA	SRP507371	9	-0.288	0.45	6
hydA	SRP507442	10	-0.418	0.23	10
hydA	SRP580424	32	0.916	1.8e-13	32
hydA	SRP657214	6	0.143	0.79	6
hydA	SRP666293	80	0.501	2.2e-06	50
mcrA	PRJEB26724	40	0.988	2.7e-32	20
mcrA	PRJEB27237	52	0.915	2.7e-21	5
mcrA	PRJEB29426	8	0.906	0.0019	3
mcrA	PRJEB45341	108	0.756	3.3e-21	108
mcrA	PRJEB45605	78	0.697	1.4e-12	78
mcrA	PRJEB45805	131	0.912	1.2e-51	116
mcrA	PRJNA1026071	6	0.657	0.16	6
mcrA	PRJNA1110177	37	0.725	3.9e-07	21
mcrA	PRJNA1135448	32	0.970	4.8e-20	31
mcrA	PRJNA1142403	44	0.642	2.6e-06	4
mcrA	PRJNA1142719	44	0.960	7.9e-25	36
mcrA	PRJNA344926	83	0.520	4.6e-07	5
mcrA	PRJNA354810	26	0.343	0.086	1
mcrA	PRJNA361507	6	1.000	0	2
mcrA	PRJNA471565	71	0.807	1.9e-17	6
mcrA	PRJNA478283	18	0.961	2.5e-10	18
mcrA	PRJNA478333	18	0.977	3.3e-12	18
mcrA	PRJNA493691	18	0.831	1.9e-05	16
mcrA	PRJNA531524	31	0.983	9.7e-23	19
mcrA	PRJNA593718	23	0.357	0.094	5
mcrA	PRJNA628856	18	0.808	5e-05	12
mcrA	PRJNA640466	32	0.964	9.3e-19	24
mcrA	PRJNA690162	45	0.921	3.3e-19	11
mcrA	PRJNA694468	40	0.869	3.4e-13	4
mcrA	PRJNA716501	15	0.786	0.00051	4
mcrA	PRJNA766239	12	0.930	1.2e-05	11
mcrA	PRJNA820970	24	0.963	6e-14	14
mcrA	PRJNA858795	111	0.963	6e-64	105

continued

Table S5 continued

Gene	Study	*n*	Spearman ρ	*p*	Both routes
mcrA	PRJNA873282	33	1.000	0	2
mcrA	SRP094596	63	0.621	5.7e-08	21
mcrA	SRP186087	61	1.000	0	4
mcrA	SRP252167	22	0.861	2.6e-07	18
mcrA	SRP278891	38	0.997	2e-41	11
mcrA	SRP286951	23	1.000	0	1
mcrA	SRP329786	24	0.874	2.4e-08	24
mcrA	SRP339895	18	0.375	0.12	1
mcrA	SRP452018	19	0.900	1.6e-07	10
mcrA	SRP466511	31	0.452	0.011	2
mcrA	SRP468584	108	0.979	1.7e-74	52
mcrA	SRP507371	9	0.380	0.31	2
mcrA	SRP580424	32	0.979	2.6e-22	32
mcrA	SRP657214	6	1.000	0	6
mcrA	SRP666293	80	0.407	0.00018	32
nifH	PRJEB26724	40	0.944	5.8e-20	20
nifH	PRJEB27237	52	0.211	0.13	33
nifH	PRJEB29426	8	0.167	0.69	8
nifH	PRJEB45341	108	-0.085	0.38	108
nifH	PRJEB45605	78	0.424	0.00011	78
nifH	PRJEB45805	131	-0.260	0.0027	131
nifH	PRJNA1026071	6	-0.371	0.47	6
nifH	PRJNA1091876	8	0.738	0.037	8
nifH	PRJNA1110177	37	-0.349	0.034	35
nifH	PRJNA1135448	32	-0.563	0.00079	32
nifH	PRJNA1142403	44	-0.553	0.0001	44
nifH	PRJNA1142719	44	0.888	8.9e-16	44
nifH	PRJNA1166154	80	0.070	0.54	69
nifH	PRJNA1284197	8	0.286	0.49	7
nifH	PRJNA270654	9	0.307	0.42	3
nifH	PRJNA344926	83	0.801	1e-19	6
nifH	PRJNA354810	26	-0.082	0.69	26
nifH	PRJNA361507	6	-0.515	0.3	6
nifH	PRJNA471565	71	0.263	0.027	67
nifH	PRJNA478283	18	0.598	0.0088	18
nifH	PRJNA478333	18	0.676	0.0021	18
nifH	PRJNA493691	18	0.760	0.00025	16
nifH	PRJNA528962	5	-0.400	0.5	5
nifH	PRJNA531524	31	0.570	0.00082	30
nifH	PRJNA593718	23	0.336	0.12	21
nifH	PRJNA628856	18	0.162	0.52	18
nifH	PRJNA640466	32	0.321	0.073	27
nifH	PRJNA690162	45	0.223	0.14	9
nifH	PRJNA694468	40	-0.025	0.88	30
nifH	PRJNA716501	15	0.771	0.00076	15
nifH	PRJNA766239	12	-0.217	0.5	12
nifH	PRJNA786456	32	0.012	0.95	32
nifH	PRJNA820970	24	0.510	0.011	24
nifH	PRJNA823829	18	0.670	0.0024	18
nifH	PRJNA858795	111	0.500	2.4e-08	110
nifH	PRJNA873282	33	0.443	0.0099	33
nifH	PRJNA980983	10	0.897	0.00044	6
nifH	SRP094596	63	0.112	0.38	29
nifH	SRP166039	28	0.359	0.061	28
nifH	SRP186087	61	0.454	0.00024	59
nifH	SRP252167	22	0.378	0.083	22
nifH	SRP278891	38	0.301	0.066	26
nifH	SRP286951	23	-0.414	0.05	23
nifH	SRP329786	24	-0.127	0.55	24
nifH	SRP339895	18	0.963	1.4e-10	8
nifH	SRP370321	6	-0.655	0.16	0
nifH	SRP376535	9	-0.226	0.56	5
nifH	SRP406096	13	0.842	0.00031	4
nifH	SRP432433	8	0.482	0.23	4
nifH	SRP452018	19	0.533	0.019	19
nifH	SRP466511	31	-0.452	0.011	29
nifH	SRP468584	108	0.197	0.041	108
nifH	SRP507371	9	-0.220	0.57	6
nifH	SRP507442	10	0.479	0.16	10
nifH	SRP580424	32	-0.238	0.19	32
nifH	SRP657214	6	-0.600	0.21	6
nifH	SRP666293	80	0.265	0.018	52
nirK	PRJEB26724	40	0.786	1.9e-09	20

continued

Table S5 continued

Gene	Study	n	Spearman ρ	p	Both routes
nirK	PRJEB27237	52	-0.091	0.52	21
nirK	PRJEB29426	8	0.024	0.96	8
nirK	PRJEB45341	108	-0.243	0.011	108
nirK	PRJEB45605	78	0.135	0.24	78
nirK	PRJEB45805	131	-0.487	3.6e-09	131
nirK	PRJNA1026071	6	-0.200	0.7	6
nirK	PRJNA1091876	8	-0.500	0.21	8
nirK	PRJNA1110177	37	-0.062	0.72	35
nirK	PRJNA1135448	32	-0.427	0.015	32
nirK	PRJNA1142403	44	-0.776	6e-10	44
nirK	PRJNA1142719	44	-0.832	2.5e-12	36
nirK	PRJNA1166154	80	0.157	0.16	26
nirK	PRJNA1284197	8	0.405	0.32	8
nirK	PRJNA270654	9	-0.567	0.11	9
nirK	PRJNA344926	83	0.718	2.2e-14	5
nirK	PRJNA354810	26	-0.745	1.3e-05	26
nirK	PRJNA361507	6	-0.455	0.37	6
nirK	PRJNA471565	71	-0.136	0.26	70
nirK	PRJNA478283	18	0.391	0.11	18
nirK	PRJNA478333	18	0.360	0.14	18
nirK	PRJNA493691	18	0.341	0.17	8
nirK	PRJNA528962	5	0.718	0.17	3
nirK	PRJNA531524	31	0.445	0.012	31
nirK	PRJNA593718	23	0.409	0.053	22
nirK	PRJNA628856	18	0.222	0.38	18
nirK	PRJNA640466	32	0.015	0.94	30
nirK	PRJNA690162	45	0.022	0.89	8
nirK	PRJNA694468	40	-0.195	0.23	35
nirK	PRJNA716501	15	-0.150	0.59	15
nirK	PRJNA766239	12	0.161	0.62	12
nirK	PRJNA786456	32	0.040	0.83	32
nirK	PRJNA820970	24	-0.337	0.11	24
nirK	PRJNA823829	18	0.331	0.18	18
nirK	PRJNA858795	111	0.235	0.013	40
nirK	PRJNA873282	33	0.083	0.65	24
nirK	PRJNA980983	10	0.600	0.067	6
nirK	SRP094596	63	-0.144	0.26	47
nirK	SRP166039	28	-0.248	0.2	28
nirK	SRP186087	61	-0.104	0.42	56
nirK	SRP252167	22	0.077	0.73	13
nirK	SRP278891	38	-0.174	0.3	27
nirK	SRP286951	23	0.347	0.1	21
nirK	SRP329786	24	0.230	0.28	23
nirK	SRP339895	18	0.859	5e-06	9
nirK	SRP370321	6	-0.131	0.8	1
nirK	SRP376535	9	0.567	0.11	9
nirK	SRP406096	13	0.703	0.0074	6
nirK	SRP432433	8	-0.405	0.32	8
nirK	SRP452018	19	0.553	0.014	19
nirK	SRP466511	31	-0.067	0.72	31
nirK	SRP468584	108	-0.035	0.72	73
nirK	SRP507371	9	0.285	0.46	7
nirK	SRP507442	10	-0.115	0.75	10
nirK	SRP580424	32	-0.625	0.00013	32
nirK	SRP657214	6	0.714	0.11	6
nirK	SRP666293	80	-0.045	0.69	46
nirS	PRJEB26724	40	0.913	2.5e-16	20
nirS	PRJEB27237	52	0.007	0.96	4
nirS	PRJEB29426	8	0.095	0.82	7
nirS	PRJEB45341	108	0.078	0.42	108
nirS	PRJEB45605	78	0.273	0.015	77
nirS	PRJEB45805	131	0.132	0.13	108
nirS	PRJNA1026071	6	-0.371	0.47	6
nirS	PRJNA1091876	8	0.976	3.3e-05	8
nirS	PRJNA1110177	37	-0.020	0.91	33
nirS	PRJNA1135448	32	0.352	0.048	32
nirS	PRJNA1142403	44	-0.788	2.2e-10	44
nirS	PRJNA1142719	44	-0.152	0.33	8
nirS	PRJNA1166154	80	0.137	0.23	80
nirS	PRJNA1284197	8	0.190	0.65	8
nirS	PRJNA270654	9	-0.250	0.52	9
nirS	PRJNA344926	83	0.662	9.6e-12	5
nirS	PRJNA354810	26	0.681	0.00013	25

continued

Table S5 continued

Gene	Study	*n*	Spearman ρ	*p*	Both routes
nirS	PRJNA361507	6	0.455	0.37	6
nirS	PRJNA471565	71	0.371	0.0014	71
nirS	PRJNA478283	18	0.474	0.047	14
nirS	PRJNA478333	18	0.639	0.0043	13
nirS	PRJNA493691	18	0.117	0.64	1
nirS	PRJNA528962	5	1.000	1.4e-24	5
nirS	PRJNA531524	31	-0.661	5.1e-05	31
nirS	PRJNA593718	23	0.519	0.011	21
nirS	PRJNA628856	18	0.110	0.66	18
nirS	PRJNA640466	32	0.459	0.0083	31
nirS	PRJNA690162	45	0.266	0.077	10
nirS	PRJNA694468	40	-0.063	0.7	37
nirS	PRJNA716501	15	-0.504	0.056	15
nirS	PRJNA766239	12	0.301	0.34	12
nirS	PRJNA786456	32	-0.026	0.89	23
nirS	PRJNA820970	24	0.246	0.25	15
nirS	PRJNA823829	18	-0.253	0.31	18
nirS	PRJNA858795	111	0.712	1.9e-18	45
nirS	PRJNA873282	33	-0.006	0.97	4
nirS	PRJNA980983	10	0.858	0.0015	6
nirS	SRP094596	63	0.038	0.77	45
nirS	SRP166039	28	-0.260	0.18	25
nirS	SRP186087	61	0.546	5.4e-06	45
nirS	SRP252167	22	0.388	0.074	17
nirS	SRP278891	38	-0.194	0.24	34
nirS	SRP286951	23	0.071	0.75	14
nirS	SRP329786	24	0.551	0.0053	22
nirS	SRP339895	18	0.772	0.00018	9
nirS	SRP370321	6	-0.131	0.8	1
nirS	SRP376535	9	-0.067	0.86	9
nirS	SRP406096	13	0.885	5.9e-05	6
nirS	SRP432433	8	0.491	0.22	6
nirS	SRP452018	19	0.282	0.24	18
nirS	SRP466511	31	0.080	0.67	31
nirS	SRP468584	108	-0.078	0.42	47
nirS	SRP507371	9	0.317	0.41	8
nirS	SRP507442	10	-0.212	0.56	10
nirS	SRP580424	32	-0.224	0.22	32
nirS	SRP657214	6	0.029	0.96	6
nirS	SRP666293	80	0.213	0.057	40
nosZ	PRJEB26724	40	0.781	2.8e-09	20
nosZ	PRJEB27237	52	-0.077	0.59	21
nosZ	PRJEB29426	8	0.286	0.49	8
nosZ	PRJEB45341	108	-0.083	0.39	108
nosZ	PRJEB45605	78	0.328	0.0034	78
nosZ	PRJEB45805	131	-0.671	1.7e-18	131
nosZ	PRJNA1026071	6	-0.371	0.47	6
nosZ	PRJNA1091876	8	0.524	0.18	8
nosZ	PRJNA1110177	37	-0.142	0.4	36
nosZ	PRJNA1135448	32	-0.023	0.9	32
nosZ	PRJNA1142403	44	-0.741	9e-09	44
nosZ	PRJNA1142719	44	-0.796	1e-10	36
nosZ	PRJNA1166154	80	0.071	0.53	80
nosZ	PRJNA1284197	8	0.000	1	8
nosZ	PRJNA270654	9	-0.600	0.088	9
nosZ	PRJNA344926	83	0.676	2.4e-12	5
nosZ	PRJNA354810	26	0.591	0.0015	26
nosZ	PRJNA361507	6	0.455	0.37	6
nosZ	PRJNA471565	71	0.170	0.16	71
nosZ	PRJNA478283	18	-0.100	0.69	18
nosZ	PRJNA478333	18	-0.370	0.13	18
nosZ	PRJNA493691	18	0.099	0.7	8
nosZ	PRJNA528962	5	1.000	1.4e-24	5
nosZ	PRJNA531524	31	0.842	2.8e-09	31
nosZ	PRJNA593718	23	0.426	0.043	22
nosZ	PRJNA628856	18	0.082	0.75	18
nosZ	PRJNA640466	32	0.115	0.53	31
nosZ	PRJNA690162	45	0.259	0.085	10
nosZ	PRJNA694468	40	0.355	0.025	39
nosZ	PRJNA716501	15	-0.357	0.19	15
nosZ	PRJNA766239	12	0.587	0.045	12
nosZ	PRJNA786456	32	-0.077	0.67	32
nosZ	PRJNA820970	24	-0.305	0.15	24

continued

Table S5 continued

Gene	Study	*n*	Spearman ρ	*p*	Both routes
nosZ	PRJNA823829	18	-0.478	0.045	18
nosZ	PRJNA858795	111	0.782	3.7e-24	51
nosZ	PRJNA873282	33	0.082	0.65	21
nosZ	PRJNA980983	10	0.781	0.0077	6
nosZ	SRP094596	63	0.061	0.63	47
nosZ	SRP166039	28	-0.259	0.18	28
nosZ	SRP186087	61	0.420	0.00075	57
nosZ	SRP252167	22	-0.318	0.15	21
nosZ	SRP278891	38	0.275	0.094	37
nosZ	SRP286951	23	-0.317	0.14	21
nosZ	SRP329786	24	0.600	0.0019	23
nosZ	SRP339895	18	0.889	7.9e-07	9
nosZ	SRP370321	6	-0.131	0.8	1
nosZ	SRP376535	9	-0.167	0.67	9
nosZ	SRP406096	13	0.810	0.00078	6
nosZ	SRP432433	8	-0.286	0.49	8
nosZ	SRP452018	19	0.632	0.0037	19
nosZ	SRP466511	31	-0.181	0.33	31
nosZ	SRP468584	108	-0.111	0.25	61
nosZ	SRP507371	9	0.317	0.41	8
nosZ	SRP507442	10	-0.552	0.098	10
nosZ	SRP580424	32	-0.569	0.00067	32
nosZ	SRP657214	6	0.257	0.62	6
nosZ	SRP666293	80	-0.169	0.13	46
pflD	PRJEB26724	40	0.878	1e-13	20
pflD	PRJEB27237	52	-0.073	0.61	15
pflD	PRJEB45341	108	0.144	0.14	108
pflD	PRJEB45605	78	0.731	2.8e-14	78
pflD	PRJEB45805	131	0.061	0.49	131
pflD	PRJNA1026071	6	-0.899	0.015	6
pflD	PRJNA1091876	8	0.881	0.0039	8
pflD	PRJNA1110177	37	-0.235	0.16	15
pflD	PRJNA1135448	32	0.274	0.13	32
pflD	PRJNA1142403	44	0.824	6.7e-12	44
pflD	PRJNA1142719	44	0.915	3.4e-18	44
pflD	PRJNA1166154	80	0.197	0.08	11
pflD	PRJNA1284197	8	0.635	0.091	6
pflD	PRJNA344926	83	0.914	1.6e-33	6
pflD	PRJNA354810	26	0.440	0.024	22
pflD	PRJNA361507	6	-1.000	1.8e-32	6
pflD	PRJNA471565	71	0.161	0.18	65
pflD	PRJNA478283	18	0.672	0.0023	18
pflD	PRJNA478333	18	0.717	0.00081	18
pflD	PRJNA493691	18	0.256	0.31	16
pflD	PRJNA528962	5	0.600	0.28	5
pflD	PRJNA531524	31	0.696	1.4e-05	30
pflD	PRJNA593718	23	0.626	0.0014	15
pflD	PRJNA628856	18	0.293	0.24	16
pflD	PRJNA640466	32	0.334	0.062	20
pflD	PRJNA690162	45	0.216	0.15	13
pflD	PRJNA694468	40	0.748	3e-08	28
pflD	PRJNA716501	15	0.568	0.027	15
pflD	PRJNA766239	12	0.350	0.27	11
pflD	PRJNA786456	32	0.424	0.016	32
pflD	PRJNA820970	24	0.833	4.4e-07	23
pflD	PRJNA823829	18	0.806	5.4e-05	17
pflD	PRJNA858795	111	0.323	0.00055	62
pflD	PRJNA873282	33	0.682	1.2e-05	33
pflD	PRJNA980983	10	0.961	9.4e-06	6
pflD	SRP094596	63	0.212	0.095	17
pflD	SRP166039	28	0.723	1.4e-05	19
pflD	SRP186087	61	0.672	3.1e-09	26
pflD	SRP252167	22	0.758	4.3e-05	22
pflD	SRP278891	38	0.780	7.5e-09	21
pflD	SRP286951	23	0.322	0.13	19
pflD	SRP329786	24	0.325	0.12	24
pflD	SRP339895	18	0.764	0.00023	6
pflD	SRP376535	9	0.000	1	5
pflD	SRP406096	13	0.585	0.036	2
pflD	SRP432433	8	0.218	0.6	2
pflD	SRP452018	19	0.402	0.088	19
pflD	SRP466511	31	0.097	0.6	22
pflD	SRP468584	108	0.342	0.00029	107

continued

Table S5 continued

Gene	Study	n	Spearman ρ	p	Both routes
pflD	SRP507371	9	-0.407	0.28	6
pflD	SRP507442	10	0.382	0.28	10
pflD	SRP580424	32	0.921	7.9e-14	32
pflD	SRP657214	6	-0.086	0.87	6
pflD	SRP666293	80	0.350	0.0014	50
soxB	PRJEB26724	40	0.869	3.8e-13	20
soxB	PRJEB27237	52	0.180	0.2	3
soxB	PRJEB45341	108	-0.122	0.21	99
soxB	PRJEB45605	78	0.217	0.056	58
soxB	PRJEB45805	131	0.683	2.6e-19	68
soxB	PRJNA1026071	6	0.086	0.87	5
soxB	PRJNA1110177	37	0.407	0.012	20
soxB	PRJNA1135448	32	0.439	0.012	32
soxB	PRJNA1284197	8	0.156	0.71	2
soxB	PRJNA344926	83	0.217	0.049	6
soxB	PRJNA354810	26	0.144	0.48	18
soxB	PRJNA361507	6	0.455	0.37	6
soxB	PRJNA471565	71	0.060	0.62	2
soxB	PRJNA478283	18	0.461	0.054	5
soxB	PRJNA478333	18	0.387	0.11	4
soxB	PRJNA531524	31	0.705	9.4e-06	23
soxB	PRJNA593718	23	0.150	0.5	14
soxB	PRJNA628856	18	0.434	0.072	8
soxB	PRJNA640466	32	-0.018	0.92	25
soxB	PRJNA690162	45	-0.034	0.83	3
soxB	PRJNA694468	40	-0.175	0.28	5
soxB	PRJNA716501	15	0.314	0.25	2
soxB	PRJNA766239	12	0.056	0.86	12
soxB	PRJNA820970	24	0.078	0.72	17
soxB	PRJNA858795	111	0.623	2.8e-13	43
soxB	PRJNA873282	33	0.000	1	1
soxB	PRJNA980983	10	0.792	0.0063	3
soxB	SRP094596	63	0.211	0.097	29
soxB	SRP166039	28	0.080	0.69	23
soxB	SRP186087	61	-0.003	0.98	11
soxB	SRP252167	22	-0.236	0.29	4
soxB	SRP278891	38	0.247	0.13	1
soxB	SRP286951	23	0.240	0.27	12
soxB	SRP339895	18	0.863	4.1e-06	6
soxB	SRP370321	6	-0.531	0.28	0
soxB	SRP376535	9	-0.117	0.77	9
soxB	SRP406096	13	0.508	0.076	2
soxB	SRP432433	8	0.310	0.46	8
soxB	SRP452018	19	0.557	0.013	16
soxB	SRP466511	31	0.548	0.0014	31
soxB	SRP507371	9	0.146	0.71	4
soxB	SRP507442	10	-0.127	0.73	10
soxB	SRP580424	32	0.209	0.25	32
soxB	SRP657214	6	0.486	0.33	6
soxB	SRP666293	80	0.230	0.04	22

Both routes is the count of samples in which both routes returned a non-zero score. Only genes reachable by both routes appear.

S6 Sequencing-depth coefficients

Table S6 gives the depth coefficient for every panel gene under both measurements. Both were fitted on the same corpus so the two columns are directly comparable, which earlier drafts of this analysis did not do and which is necessary for the sign reversal to mean anything.

Detection is modelled as log-odds per $\log_{10}$ read. Abundance is modelled as $\log_{10}$ relative abundance per $\log_{10}$ read, among samples where the gene is detected, so the two are on different scales and only their signs are being compared. The detection coefficient is positive for all 20 genes (significant for 1), and depth significantly lowers abundance in 15 of 20 genes and raises it in 1 (*aioA*), at a median $\beta = -0.399$.

Table S6. Sequencing-depth coefficients for every panel gene, fitted on one corpus so the two columns are directly comparable. Detection is in log-odds per $\log_{10}$ read; abundance is in $\log_{10}$ relative abundance per $\log_{10}$ read, among samples where the gene is detected.

Gene	Detection β	p	Abundance β	p
aioA	+0.803	0.41	+0.227	0.014
amoA	+1.793	0.026	-0.415	0.0001
arrA	+1.155	0.09	-0.558	6.6e-10
arsM	+0.600	0.59	-0.408	8.2e-07
dsrA	+0.499	0.58	+0.133	0.15
dsrB	+0.499	0.58	+0.133	0.15
feoB	+0.601	0.71	-0.391	1.9e-07
hydA	+0.349	0.69	-0.600	1.5e-10
mcrA	+1.136	0.26	+0.033	0.79
mtrA	+0.626	0.48	-0.426	1.9e-07
mtrB	+0.647	0.25	-1.372	4.6e-05
mtrC	+0.495	0.57	-0.534	4.9e-11
nifH	+0.281	0.83	-0.450	3.7e-10
nirK	+0.731	0.42	-0.366	8.1e-07
nirS	+0.879	0.27	-0.230	0.0071
nosZ	+0.422	0.69	-0.335	2.6e-06
omcS	+0.604	0.68	-0.362	1.3e-06
omcZ	+0.379	0.64	-0.469	9.4e-09
pflD	+0.537	0.6	-0.539	8.7e-09
soxB	+1.491	0.05	+0.018	0.89

S7 Compartment contrasts in full

Table S7 gives every depth-adjusted contrast, all genes across all three compartment pairs. The main text reports the strongest contrast and the range of the intervals.

Detection was regressed on compartment and $\log_{10}$ retained reads using generalized estimating equations with a binomial family, a logit link and an exchangeable working correlation, clustered by study. Because cluster-robust standard errors are anti-conservative when clusters are few, and this is a between-cluster comparison with fewer than ten aquifer studies, every gene was also tested by permuting compartment labels across studies rather than across samples, refitting on each of 2,000 permutations. Permutation values are the ones reported for inference. Both are shown here so the gap between them stays visible.

Table S7. Every depth-adjusted compartment contrast, all genes and all three pairs. The main text reports the three strongest per contrast; this is the complete set.

Contrast	Gene	OR	95 % CI	p_{GEE}	q_{GEE}	p_{perm}	q_{perm}
As-soil vs aquifer	*arrA*	3.86	1.56–9.51	0.0006	0.011	0.0350	0.536
As-soil vs aquifer	*soxB*	3.18	0.56–18.12	0.1174	0.391	0.1128	0.536
As-soil vs aquifer	*nirS*	6.04	0.05–720.27	0.3722	0.560	0.1457	0.536
As-soil vs aquifer	*mtrA*	2.38	0.89–6.34	0.0406	0.348	0.1538	0.536
As-soil vs aquifer	*feoB*	2.77	0.70–11.00	0.0695	0.348	0.1808	0.536
As-soil vs aquifer	*dsrB*	2.11	0.46–9.62	0.2276	0.506	0.2048	0.536
As-soil vs aquifer	*dsrA*	2.11	0.46–9.62	0.2276	0.506	0.2052	0.536
As-soil vs aquifer	*aioA*	7.60	0.55–104.53	0.0646	0.348	0.2337	0.536
As-soil vs aquifer	*mtrC*	2.19	0.73–6.57	0.0936	0.374	0.2412	0.536
As-soil vs aquifer	*omcZ*	1.97	0.59–6.54	0.1870	0.506	0.3017	0.603
As-soil vs aquifer	*amoA*	1.75	0.41–7.54	0.3529	0.560	0.4326	0.786
As-soil vs aquifer	*pflD*	0.65	0.22–1.90	0.3514	0.560	0.4940	0.793
As-soil vs aquifer	*arsM*	0.68	0.28–1.66	0.3257	0.560	0.5205	0.793
As-soil vs aquifer	*omcS*	1.45	0.54–3.85	0.3918	0.560	0.5634	0.793
As-soil vs aquifer	*hydA*	0.74	0.23–2.32	0.5350	0.669	0.6330	0.793
As-soil vs aquifer	*nifH*	0.71	0.19–2.56	0.5352	0.669	0.6344	0.793
As-soil vs aquifer	*mtrB*	1.31	0.25–6.84	0.6711	0.789	0.7223	0.850
As-soil vs aquifer	*mcrA*	0.88	0.12–6.32	0.8514	0.896	0.8412	0.935
As-soil vs aquifer	*nirK*	1.20	0.16–8.77	0.8322	0.896	0.8941	0.941
As-soil vs aquifer	*nosZ*	0.95	0.09–10.36	0.9633	0.963	0.9841	0.984
Paddy vs As-soil	*omcZ*	0.31	0.12–0.82	0.0113	0.075	0.0250	0.282
Paddy vs As-soil	*omcS*	0.23	0.05–1.12	0.0518	0.173	0.0431	0.282
Paddy vs As-soil	*mcrA*	0.21	0.09–0.53	0.0002	0.004	0.0438	0.282
Paddy vs As-soil	*feoB*	0.19	0.02–2.06	0.1433	0.260	0.0613	0.282

continued

Table S7 continued

Contrast	Gene	OR	95 % CI	p_{GEE}	q_{GEE}	p_{perm}	q_{perm}
Paddy vs As-soil	*mtrC*	0.38	0.14–1.01	0.0378	0.158	0.0811	0.282
Paddy vs As-soil	*mtrA*	0.39	0.15–1.02	0.0394	0.158	0.0846	0.282
Paddy vs As-soil	*mtrB*	4.50	0.97–20.83	0.0095	0.075	0.1389	0.335
Paddy vs As-soil	*hydA*	0.48	0.18–1.26	0.1099	0.260	0.1648	0.335
Paddy vs As-soil	*dsrB*	0.54	0.22–1.30	0.1418	0.260	0.1715	0.335
Paddy vs As-soil	*dsrA*	0.54	0.22–1.30	0.1418	0.260	0.1988	0.335
Paddy vs As-soil	*pflD*	0.49	0.17–1.45	0.1671	0.260	0.2316	0.335
Paddy vs As-soil	*arsM*	0.45	0.14–1.48	0.1606	0.260	0.2317	0.335
Paddy vs As-soil	*nosZ*	2.14	0.60–7.59	0.2087	0.274	0.2413	0.335
Paddy vs As-soil	*nirS*	1.89	0.70–5.06	0.1753	0.260	0.2436	0.335
Paddy vs As-soil	*aioA*	1.77	0.65–4.88	0.2308	0.274	0.2667	0.335
Paddy vs As-soil	*nirK*	2.04	0.66–6.26	0.1818	0.260	0.2682	0.335
Paddy vs As-soil	*arrA*	0.61	0.26–1.45	0.2330	0.274	0.3377	0.397
Paddy vs As-soil	*amoA*	1.23	0.46–3.31	0.6511	0.723	0.6773	0.753
Paddy vs As-soil	*soxB*	1.12	0.47–2.65	0.7881	0.814	0.7990	0.835
Paddy vs As-soil	*nifH*	0.87	0.26–2.95	0.8142	0.814	0.8347	0.835
Paddy vs aquifer	*aioA*	22.04	2.28–212.81	0.0012	0.008	0.0081	0.154
Paddy vs aquifer	*mtrB*	6.57	1.14–37.79	0.0004	0.008	0.0200	0.154
Paddy vs aquifer	*nosZ*	8.32	1.68–41.25	0.0024	0.012	0.0263	0.154
Paddy vs aquifer	*nirS*	5.95	1.75–20.19	0.0009	0.008	0.0307	0.154
Paddy vs aquifer	*soxB*	3.35	0.74–15.10	0.0603	0.158	0.0540	0.216
Paddy vs aquifer	*nirK*	5.13	0.95–27.78	0.0243	0.081	0.0668	0.223
Paddy vs aquifer	*mcrA*	0.29	0.06–1.45	0.0633	0.158	0.1408	0.402
Paddy vs aquifer	*arrA*	2.58	1.06–6.26	0.0164	0.066	0.1875	0.469
Paddy vs aquifer	*arsM*	0.50	0.21–1.23	0.0937	0.208	0.3286	0.672
Paddy vs aquifer	*amoA*	2.12	0.50–8.92	0.2086	0.373	0.3540	0.672
Paddy vs aquifer	*hydA*	0.51	0.16–1.60	0.1840	0.368	0.3982	0.672
Paddy vs aquifer	*pflD*	0.56	0.19–1.65	0.2240	0.373	0.4326	0.672
Paddy vs aquifer	*dsrB*	1.58	0.49–5.07	0.3755	0.536	0.4645	0.672
Paddy vs aquifer	*dsrA*	1.58	0.49–5.07	0.3755	0.536	0.4705	0.672
Paddy vs aquifer	*omcS*	0.61	0.16–2.36	0.4090	0.545	0.5634	0.736
Paddy vs aquifer	*omcZ*	0.67	0.19–2.28	0.4487	0.561	0.5884	0.736
Paddy vs aquifer	*nifH*	0.70	0.19–2.56	0.5310	0.625	0.6374	0.750
Paddy vs aquifer	*feoB*	0.79	0.15–4.17	0.7480	0.831	0.8531	0.926
Paddy vs aquifer	*mtrC*	0.88	0.28–2.82	0.8074	0.850	0.8801	0.926
Paddy vs aquifer	*mtrA*	1.06	0.36–3.16	0.8983	0.898	0.9331	0.933

OR is the odds of detection in the second compartment relative to the first at equal sequencing depth. The 95 % interval uses the Mancl–DeRouen bias-reduced covariance with a t reference on studies minus two degrees of freedom, is not adjusted for multiplicity, and shows the size of difference each null leaves open. Permutation values are the ones reported for inference, with model-based values shown alongside so the few-cluster inflation is visible.

S8 Software and reference data

Table S8 records the version of every tool and reference dataset used. Package versions were captured as resolved environment specifications at the time of execution, and the arsenic-soil arm was processed with bit-identical versions of the retrieval, trimming and denoising tools to the initial corpus.

Analysis code is available as described in the availability statement of the main text. Every number in the main text and in these tables is generated directly from the sample sheet and the analysis outputs by the scripts in that repository, so no count is transcribed by hand at any point.

Table S8. Software and reference data, with the version actually used. Three Snakemake versions span the run; it schedules rules and does not compute results, so the three do not affect comparability.

Stage	Software or data	Version	Role
Retrieval	sra-tools	3.4.1	FASTQ retrieval from the archive
Primer removal	cutadapt	5.2	5’ primer trimming, per detected primer
Denoising	DADA2	1.38.0	Error model, denoising, merging, chimera removal
Taxonomy	DECIPHER	3.6.0	Genus-level assignment against SILVA
Taxonomy	SILVA	138.2	Reference database, genus-level training set
Taxonomy	GTDB	r220	Fallback taxonomy, disabled for the reported run

continued

Table S8 continued

Stage	Software or data	Version	Role
Functional	PICRUSt2	2.6.3	KEGG ortholog prediction, second inference route, default PICRUSt2-SC reference (GTDB r214, 27,870 genomes)
Functional	Marker-taxa lookup	1.0	Curated genus-to-gene bridge (Table S3)
Functional	FAPROTAX	1.2.12	Published taxon-to-function lookup, third route (Section S11)
Functional	Tax4Fun2	1.1.5	KEGG ortholog prediction by nearest reference 16S sequence, fourth route (Section S16); Zenodo record 10035668
Functional	Tax4Fun2 reference data	v2	Ref100NR reference genomes and 16S sequences
Functional	BLAST+	2.16.0	Assignment of sequence variants for Tax4Fun2, with R 4.4.1
Gene panel	Panel manifest	1.2	35 genes across five cycles, checksummed
Alignment	MAFFT	7	Alignment of marker-genus representatives
Phylogeny	FastTree	2	Approximately maximum-likelihood tree, GTR
Soil covariates	SoilGrids	2.0	0–5 cm properties at 250 m resolution
Workflow	Snakemake	9.22.0–9.25.2	Rule scheduling only; does not compute results
Environment	R	4.5.3	Denoising and taxonomy
Environment	Python	3.11	Analysis, with pandas, SciPy and statsmodels
Shotgun QC	fastp	1.4.0	Read trimming and filtering, R1 only (Section S10)
Gene calling	FragGeneScanRs	1.1.0	Open reading frames on short reads
Homology search	HMMER	3.4	hmmsearch against the KO and Pfam panels
Homology search	KOfam	retrieved 2026-10-16	Profile HMMs and thresholds for the 26 panel KOs, and the 29 of panel v1.3 in the second pass
Homology search	Curated arsenic HMMs	Dunivin et al. 2019	*arsM*, *aioA*, *arrA*, *arxA*; second pass only (Section S10)

S9 The predictor's own confidence metric

Section S4 tested whether phylogenetic conservation of the marker taxa explains which genes the two routes agree about, and it does not. This section tests the other candidate a reader is likely to reach for, which is supplied by the software itself.

PICRUSt2 reports a nearest-sequenced-taxon index (NSTI) for every sequence variant, the phylogenetic distance to the closest genome in its reference set. It is the tool's own statement of how far a prediction has been extrapolated, and a high value is conventionally read as grounds for caution.

Each variant is placed against both the bacterial and the archaeal reference tree, so every variant appears in both output files. The archaeal placement of a bacterial sequence is arbitrarily distant, and in one study of this corpus the archaeal file carries a median index of 3.20 against 0.149 for the bacterial file over the same variants. The index used here is therefore the minimum across the two, which selects the correct domain without requiring the taxonomy. Values were then weighted by variant abundance to give a per-sample figure, as PICRUSt2 defines its own weighted NSTI, and averaged per study. Variants the tool could not place carry no index and are excluded from the weighting rather than counted as zero, which would report a confidence the tool never gave.

Across the 57 studies carrying both routes the index ranges from 0.044 to 1.713, a 39.3-fold spread, while median agreement ranges from −0.118 to +0.859. Both quantities vary substantially, and they are unrelated (Spearman $\rho = +0.111$, $p = 0.41$). Restricting to the 47 studies in which at least 60 % of variants were placed does not recover a relationship ($\rho = +0.195$, $p = 0.19$). The most extreme study runs against the expectation, SRP406096 carrying the least favourable index of any study at 1.713 together with the closest agreement between routes at +0.611.

Two caveats bound this result. A Pearson correlation on the filtered subset reaches +0.724 at $p = 0.002$, but it is driven entirely by SRP406096 sitting at one corner of the scatter and its sign is opposite to what the index means, so the rank statistic is the one reported at this sample size. And with 57 studies the test has roughly 62 % power to detect a correlation of 0.5, so a strong relationship is excluded while a weak one is not.

The practical consequence is that a favourable NSTI does not license trusting a predicted functional profile, and an unfavourable one does not condemn it. Two candidate predictors of reliability have now been tested, one drawn

from trait ecology and one supplied by the software, and neither works.

Table S9. PICRUSt2's nearest-sequenced-taxon index against agreement between the two inference routes, one row per study. NSTI is the phylogenetic distance from a sequence variant to the closest reference genome, abundance-weighted per sample and averaged per study.

Study	Genes	Median ρ	Weighted NSTI	Frac. placed	Variants placed
PRJNA823829	11	+0.164	0.0436	1.00	6844
SRP166039	12	+0.127	0.0486	0.97	7498
PRJNA528962	10	+0.659	0.0581	1.00	234
PRJNA531524	13	+0.570	0.0610	1.00	12016
PRJEB45605	13	+0.328	0.0689	1.00	16948
PRJNA1091876	11	+0.114	0.0704	1.00	1242
SRP432433	12	-0.228	0.0755	1.00	2752
PRJNA344926	13	+0.718	0.0767	0.36	1688
PRJNA873282	13	+0.443	0.0787	1.00	5307
SRP468584	12	+0.078	0.0808	1.00	40219
SRP094596	13	+0.183	0.0842	0.58	5032
PRJNA354810	13	+0.223	0.0885	1.00	14406
PRJNA766239	13	+0.336	0.0894	1.00	3206
SRP507442	12	-0.006	0.0895	1.00	12594
PRJNA786456	11	+0.040	0.0898	1.00	4899
PRJNA471565	13	+0.263	0.0901	0.66	8464
PRJNA270654	6	-0.262	0.0932	0.62	949
PRJNA1284197	10	+0.238	0.0934	1.00	2684
PRJNA361507	13	+0.455	0.0996	1.00	2997
PRJNA1110177	13	+0.407	0.0998	1.00	12882
PRJNA716501	13	+0.314	0.1003	1.00	26411
SRP329786	12	+0.355	0.1071	0.58	8869
SRP278891	11	+0.201	0.1072	1.00	4473
PRJEB29426	9	+0.083	0.1073	1.00	2197
PRJNA640466	13	+0.339	0.1073	0.99	19347
PRJNA593718	13	+0.311	0.1095	1.00	12301
PRJNA1166154	11	+0.157	0.1097	1.00	2920
PRJEB45805	13	+0.133	0.1117	1.00	80110
SRP252167	13	+0.293	0.1124	1.00	2674
SRP466511	13	+0.140	0.1137	1.00	31720
SRP452018	13	+0.488	0.1159	1.00	12690
SRP376535	12	-0.033	0.1175	1.00	3591
SRP580424	13	+0.635	0.1198	1.00	22499
SRP666293	13	+0.197	0.1205	0.43	29633
PRJEB27237	13	+0.056	0.1211	0.99	8884
SRP186087	13	+0.454	0.1240	1.00	18282
PRJNA478333	13	+0.360	0.1272	1.00	10351
PRJNA478283	13	+0.391	0.1273	1.00	11440
PRJNA628856	13	+0.372	0.1274	1.00	17591
PRJNA1135448	13	+0.274	0.1305	1.00	26724
PRJNA694468	13	-0.025	0.1334	1.00	17380
PRJEB26724	13	+0.878	0.1371	0.49	12089
PRJNA493691	12	+0.334	0.1414	0.91	6112
PRJNA1026071	13	-0.371	0.1448	0.98	9983
SRP507371	13	+0.284	0.1458	1.00	3573
SRP286951	13	+0.347	0.1466	1.00	10877
PRJNA690162	13	+0.135	0.1570	1.00	12424
PRJNA820970	13	+0.510	0.1618	1.00	15212
PRJEB45341	13	+0.053	0.1623	1.00	71751
PRJNA1142403	12	-0.471	0.1630	0.56	36418
PRJNA1142719	12	+0.235	0.1867	0.50	10789
SRP657214	13	+0.257	0.1981	0.55	4773
SRP339895	13	+0.859	0.2559	0.99	4299
SRP370321	8	-0.296	0.2671	0.05	102
PRJNA858795	13	+0.623	0.3918	1.00	31797
PRJNA980983	12	+0.768	0.8990	0.25	5219

SRP406096	12	+0.611	1.7128	0.99	3585

Median ρ is the median agreement across the genes reachable by both routes in that study. Frac. placed is the proportion of sequence variants PICRUSt2 placed on its reference tree; variants it could not place carry no index and are excluded from the weighting, which is how PICRUSt2 defines the quantity. Studies are ordered by NSTI, so a relationship with agreement would appear as a trend down the third column. There is none.

S10 Measured gene content from paired shotgun metagenomes

Both inference routes were set against gene content measured directly in the 17 shotgun runs that share a biosample with an amplicon library in this corpus. Processing is described in the main text, and this section records the decisions behind the counts, which matter more than usual because reads are short and the panel is small.

The assignment rule was written down, with a checksum, after the runs had finished and before any per-gene count or correlation was computed. KOfam's adaptive thresholds could not be used as published, because they are calibrated on full-length proteins and for 21 of the 26 panel orthologs no read in any run reaches them, so applying them would have recorded almost every gene as absent. Each reading frame is instead assigned to the one panel ortholog it matches best at $E < 10^{-3}$. Ties on score were to be left unassigned, and one frame in all 17 runs was. The search was run on fixed-size chunks of each run, so the last chunk carries a slightly looser E-value than the rest. Bit-score thresholds do not depend on database size, so the length-scaled rule below is unaffected.

Best-match assignment separates panel genes from one another but cannot separate them from homologues outside the panel, because the search only offers panel profiles to choose from. The length-scaled sensitivity rule addresses that by requiring the KOfam threshold multiplied by the fraction of the profile aligned. It removes most best-match counts for *arrA*, *hydA* and *mtrC*, whose protein families have many members outside the panel, and leaves most other genes almost unchanged (Table S10, third and fourth columns). For those three genes the length-scaled values are the ones to read. Two further sensitivity analyses tightened the cutoff to $E < 10^{-10}$ and admitted the pairs below the read-depth floor, which changes only the PICRUSt2 comparison because bridge scores are not computed below the floor. The lowest *arrA* count in any run falls from 2,230 assigned frames under best-match assignment to 11 under the length-scaled threshold, and *hydA* falls from 3,149 to 0, so most best-match counts for those families probably belong to other members of the family.

Gene calling used FragGeneScanRs, a reimplementation of FragGeneScan with the same gene models. On a 300,000-read subset of one run, 661 of the 666 reads given an ortholog hit through FragGeneScan were also given one through FragGeneScanRs, which found 670 in total, so the substitution does not change what is counted.

PRJNA980983 contributes two biosamples, each with three shotgun and six amplicon runs, which were pooled per biosample on both sides. Two units cannot support a correlation, so the study is reported descriptively. Every measurable gene is detected in both biosamples, and neither route scores a gene present where the shotgun data lack it, as in SRP507371, so at this sequencing depth detection is saturated and presence carries no information about either route.

Two checks were added after the results were seen and are labelled as such. The first set each gene's PICRUSt2 prediction against the measured abundance of every other gene, to test whether agreement reflects a property shared by all genes in a sample. The second repeated the comparison within each of the two well depths. Both are reported in the main text, and neither changes the result, although neither was planned.

Table S10. Measured against inferred gene content in the deep-aquifer study SRP507371, one row per panel gene with a KEGG ortholog. ρ is the Spearman correlation across paired biosamples between each inferred route and shotgun abundance.

Gene	Cycle	ORFs/M	ORFs/M (S1)	Bridge ρ	Bridge ρ (S1)	PICRUSt2 ρ	PICRUSt2 ρ (S1)
acr3	As	182.95	182.95	–	–	+0.83	+0.83
arrA	As	160.18	9.71	-0.54	-0.70	–	–
arsB	As	25.56	25.56	–	–	-0.10	-0.10
arsC_grx	As	10.68	10.68	–	–	+0.85	+0.85
arsC_myc	As	0.84	0.41	–	–	+0.52	+0.48
arsC_trx	As	76.62	76.62	–	–	+0.40	+0.40
arsR	As	142.84	142.84	–	–	-0.37	-0.37
hydA	C	180.61	0.00	-0.51	-0.56	-0.46	+0.48
mcrA	C	6.30	1.76	-0.05	+0.16	+0.02	+0.33
pflD	C	31.59	31.59	-0.58	-0.58	+0.77	+0.77
feoB	Fe	316.06	316.06	-0.36	-0.36	+0.77	+0.77
mtrA	Fe	40.56	13.55	+0.24	+0.08	–	–
mtrB	Fe	11.33	11.09	–	–	–	–

continued

Table S10 continued

Gene	Cycle	ORFs/M	ORFs/M (S1)	Bridge ρ	Bridge ρ (S1)	PICRUSt2 ρ	PICRUSt2 ρ (S1)
mtrC	Fe	1.42	0.00	+0.15	+0.24	–	–
amoA	N	5.29	3.88	+0.84	+0.95	–	–
nifH	N	82.56	43.86	-0.54	-0.02	+0.08	+0.57
nirK	N	49.51	49.48	-0.01	-0.01	+0.83	+0.83
nirS	N	15.71	7.85	+0.27	+0.27	+0.98	+0.98
nosZ	N	36.10	35.59	+0.10	+0.10	+0.60	+0.60
phoA	P	27.55	27.55	–	–	+0.73	+0.73
phoD	P	31.20	31.20	–	–	+0.98	+0.98
ppk1	P	134.42	134.42	–	–	+0.85	+0.85
pstS	P	140.88	140.88	–	–	+0.33	+0.33
dsrA	S	60.12	57.92	+0.00	+0.00	+0.78	+0.78
dsrB	S	58.41	52.09	+0.00	+0.00	+0.78	+0.78
soxB	S	44.52	44.50	+0.84	+0.84	-0.05	-0.05

ORFs/M is the median across the study's shotgun runs of reads assigned to the gene per million open reading frames. The primary rule assigns each frame to its best-scoring panel KO at $E < 10^{-3}$; S1 additionally requires the KOfam threshold scaled to the aligned length of the profile, which removes reads from homologues outside the panel. A dash marks a gene that route cannot reach (no marker taxa, or a KO absent from the PICRUSt2 reference) or that does not vary across the pairs.

A second pass over the 9 primary SRP507371 runs added the orthologs for *arsM* and *aioA* (K07755 and K08356) and the curated HMMs of Dunivin et al. (2019) for *arsM*, *aioA*, *arrA* and *arxA*. Its quantification was fixed in an addendum to the protocol, with a checksum, before any read was processed, and used the same code and rule as the first pass. The 36 gene-route correlations the two passes share differ by at most 0.10. The curated and KOfam counts rank the biosamples alike for *aioA* ($\rho = +0.93$), *arrA* ($\rho = +0.95$) and *arsM* ($\rho = +0.68$), so the correlations in Table S10b do not depend on which profiles defined the genes.

Table S10b. Arsenic metabolism genes in the second pass over the nine primary SRP507371 pairs. ρ columns are Spearman correlations across the biosamples, between measured abundance and each inferred route.

Gene	KOfam /M	Curated /M	Cur./KOfam ρ	Bridge ρ	Bridge (strict)	PICRUSt2 ρ	PICRUSt2 (strict)
arsM	241	111	+0.68	-0.03	+0.12	+0.50	+0.65
aioA	69	12	+0.93	-0.42	+0.48	+0.13	+0.10
arrA	124	49	+0.95	-0.54	-0.65	–	–
arxA	–	81	–	–	–	–	–

KOfam counts use the primary rule of Table S10 with the v1.3 panel (*arsM* K07755, *aioA* K08356). Curated counts assign each frame hit at $E < 10^{-3}$ by any of the four curated HMMs of Dunivin et al. (2019) to its best-scoring one. *arxA* has no KEGG ortholog, so it has no KOfam count and no PICRUSt2 prediction, and the bridge scores it in none of these samples. Curated counts can include other molybdopterin enzymes and are an upper bound for *arxA*. /M is the median count per million open reading frames; strict is the $E < 10^{-10}$ sensitivity rule. A dash marks a value that cannot be computed.

S11 FAPROTAX as a third inference route

FAPROTAX 1.2.12 (Louca et al., 2016) was run unmodified through its own software over the 4,037 distinct taxonomy strings of the corpus, and each of the nine panel genes with a corresponding FAPROTAX function was scored as the relative abundance of the sequence variants assigned to that function. The gene-to-function mapping, the cells compared and the summaries were fixed in a protocol whose checksum was recorded before any FAPROTAX output existed. *hydA* and *pflD* were mapped loosely to fermentation and are reported separately. A cell enters the comparison when FAPROTAX scores vary across the study's samples, which leaves 274 of the 478 cells of the bridge comparison for these genes. FAPROTAX lists its arsenite oxidisers at species level, and no sequence variant in the corpus is classified to those species, so *aioA* is not reached.

Table S11. FAPROTAX as a third inference route, by gene. F, B and P are FAPROTAX, the marker-taxa bridge and PICRUSt2; their columns are median Spearman ρ across the study-gene cells in which all three can be compared.

Gene	FAPROTAX function	Taxa	Jaccard	Cells	F–P	B–P	F–B	F–shotgun
mcrA	methanogenesis	39	0.36	43	+0.91	+0.91	+1.00	+0.03
nifH	nitrogen fixation	17	0.08	53	+0.28	+0.26	+0.36	-0.15
nirK	nitrite denitrification	3	0.00	28	+0.07	-0.05	+0.06	–
nirS	nitrite denitrification	3	0.00	28	+0.18	+0.12	+0.08	–

continued

Table S11 continued

Gene	FAPROTAX function	Taxa	Jaccard	Cells	F–P	B–P	F–B	F–shotgun
nosZ	nitrous oxide denitrification	1	0.00	3	+0.17	-0.26	+0.37	–
dsrA	sulfate or sulfite respiration	129	0.08	52	+0.52	+0.47	+0.54	+0.45
dsrB	sulfate or sulfite respiration	129	0.08	52	+0.52	+0.47	+0.54	+0.45
soxB	dark thiosulfate oxidation	4	0.22	15	-0.05	+0.21	+0.27	–
aioA	arsenite oxidation (both groups)	0	0.00	0	–	–	–	–
hydA	fermentation (loose)	172	–	53	+0.35	+0.36	+0.26	–
pflD	fermentation (loose)	172	–	53	+0.31	+0.40	+0.36	–

Taxa is the number of corpus taxonomy strings FAPROTAX assigns to the function. Jaccard is the overlap between the genera carrying the gene in the bridge and those FAPROTAX assigns, among genera present in the corpus. F–shotgun is the correlation with measured abundance across the nine SRP507371 biosamples under the primary rule; a dash marks a gene whose FAPROTAX score is constant across them. *arsM* and *feoB* have no FAPROTAX function and are omitted.

S12 Gene carriage within the marker taxa and study-level moderators

The reference is the one that produced the PICRUSt2 route, the trait tables and genome metadata shipped with PICRUSt2 2.6.3 (PICRUSt2-SC, 27,870 genomes), with file checksums recorded. The protocol fixing the quantities, matching rules, tests and interpretation was hashed before the first run and records two deviations. The first is the *hydA* sensitivity analysis, added after the mapping limit described in the main text came to light. The second is the omission of primer region from the study-level moderators, for want of a curated value. Two companions of carriage fidelity were computed, the fraction of all reference genomes carrying the ortholog that fall inside the gene's marker taxa, and the product of the two, which is high only when the marker taxa carry the gene and also account for most of its carriers. Within genes, each study's fidelity was weighted by the abundance of the marker taxa in that study. Fidelity is tested one-sided because the protocol predicts a positive association, and its secondary companions are reported without correction for multiplicity. Table S12 gives the per-gene quantities and Table S12b the moderators.

Table S12. Carriage of each panel gene within its marker taxa in the PICRUSt2-SC reference genomes, against median agreement between routes.

Gene	Taxa	Carriers	Fidelity	Coverage	Product	Fidelity (GTDB)	Agreement ρ
mcrA	11/12	280	0.99	0.618	0.614	0.99	+0.906
dsrA	9/13	842	0.93	0.191	0.179	0.97	+0.470
dsrB	9/13	846	0.93	0.189	0.176	0.96	+0.470
pflD	2/2	6,945	0.95	0.033	0.031	0.98	+0.413
hydA	1/1	2,655	0.96	0.063	0.061	0.96	+0.367
nifH	11/11	2,965	0.75	0.138	0.104	0.75	+0.265
aioA	10/10	904	0.14	0.035	0.005	0.14	+0.219
soxB	6/7	2,727	0.93	0.018	0.017	0.90	+0.217
nirS	4/5	1,347	0.49	0.106	0.053	0.44	+0.137
feoB	9/11	14,692	0.94	0.013	0.012	0.93	+0.136
nosZ	6/7	3,483	0.57	0.065	0.037	0.56	+0.071
nirK	4/4	5,829	0.37	0.032	0.012	0.38	+0.022
arsM	6/6	5,622	0.31	0.048	0.015	0.33	-0.118

Taxa, marker taxa with at least three reference genomes over all marker taxa of the gene. Carriers, reference genomes (of 27,870) carrying the gene's KEGG ortholog. Fidelity, mean share of a marker taxon's genomes carrying it, with taxa matched in the NCBI taxonomy; the GTDB column matches in the GTDB taxonomy with genus suffixes removed. Coverage, share of all carriers inside the gene's marker taxa. Product, fidelity times coverage. Agreement, median of Table S5. *hydA* is scored by its four [FeFe]-hydrogenase orthologs; under K17997 alone the bridge's only *hydA* taxon, *Clostridium*, carries it in 1 of 175 genomes (Section S13).

Table S12b. Study-level moderators of agreement between routes. The response is a study's mean departure from the per-gene median agreement.

Moderator	Effect	Estimate	p	Holm p
Aquifer origin (6 of 57 studies)	rank-biserial	-0.35	0.17	0.70
Median retained reads ($\log_{10}$)	Spearman	-0.12	0.37	1.00
Paired-end layout (43 of 57)	rank-biserial	-0.05	0.77	1.00
Samples per study ($\log_{10}$)	Spearman	+0.02	0.86	1.00

p from 10,000 permutations of the moderator across studies. Rank-biserial is positive when the named group agrees more closely. Primer region was not curated as a study attribute and is not tested. A multivariable model with all four moderators and HC3 errors finds none (all $p > 0.17$).

S13 Compartment conclusions under each route and the *hydA* ortholog

Both analyses follow a second protocol, hashed before they were run. The compartment contrasts of Table S13 use only studies and genes carrying both routes, a median score per study and a regression on compartment and median depth, so that the two routes are compared under one statistic. The primary compartment analysis of the main text remains the detection model of Table S7. The primary model gives the same sign under both routes in 35 of 39 cells and an unadjusted rank-biserial comparison in 30, with the conclusion depending on the route in 6 and 5 cells respectively.

The frozen *hydA* ortholog K17997 is the NAD(P)-linked iron hydrogenase, carried by 64 of the 27,870 PICRUSt2-SC reference genomes and by 1 of 175 *Clostridium* genomes, the only *hydA* taxon in the bridge, against 168 of them for the four [FeFe]-hydrogenase orthologs together. The primary *hydA* map (panel manifest v1.4) therefore scores PICRUSt2 by K17997, K00532, K00533 and K18332 together, with the bridge side unchanged. Under K17997 alone, the map of manifest v1.3, the comparison exists for 41 studies in place of 54, the *hydA* median falls from +0.37 to +0.15, the median over 685 comparisons is 0.256 in place of 0.286 over 698, and the fidelity correlation is +0.58 (+0.03 to +0.90) in place of +0.66. The shotgun comparison scores *hydA* by K17997 under both maps, the only *hydA* ortholog searched in the shotgun run.

Table S13. Compartment contrasts scored by each route on the cells comparable under both. Effects are the compartment coefficient of a regression of the study's log median score on compartment and log median depth.

Contrast	Gene	Studies	Bridge	q	PICRUSt2	q	Same sign
paddy vs As-soil	*aioA*	31/19	+0.48	0.234	+0.35	0.609	yes
paddy vs As-soil	*arsM*	31/19	-0.65	0.210	-0.04	0.866	yes
paddy vs As-soil	*dsrA*	31/16	-0.31	0.440	-0.00	0.992	yes
paddy vs As-soil	*dsrB*	31/16	-0.31	0.440	-0.00	0.992	yes
paddy vs As-soil	*feoB*	31/19	-0.79	0.034	-0.14	0.004	yes
paddy vs As-soil	*hydA*	31/17	-0.69	0.142	-0.56	0.046	yes
paddy vs As-soil	*mcrA*	28/11	-1.52	0.023	-1.27	0.010	yes
paddy vs As-soil	*nifH*	31/20	-0.28	0.428	-0.45	0.044	yes
paddy vs As-soil	*nirK*	31/20	+0.37	0.247	+0.00	0.992	yes
paddy vs As-soil	*nirS*	31/20	+0.35	0.430	+0.03	0.922	yes
paddy vs As-soil	*nosZ*	31/20	+0.30	0.371	+0.03	0.874	yes
paddy vs As-soil	*pflD*	31/17	-0.51	0.247	-0.26	0.245	yes
paddy vs As-soil	*soxB*	23/16	+0.27	0.440	+0.05	0.866	yes
paddy vs aquifer	*aioA*	31/6	+0.91	0.034	+0.71	0.316	yes
paddy vs aquifer	*arsM*	31/6	-1.30	0.116	-0.13	0.651	yes
paddy vs aquifer	*dsrA*	31/6	+0.05	0.957	+0.28	0.316	yes
paddy vs aquifer	*dsrB*	31/6	+0.05	0.957	+0.28	0.316	yes
paddy vs aquifer	*feoB*	31/6	-0.43	0.430	-0.01	0.874	yes
paddy vs aquifer	*hydA*	31/6	-1.28	0.116	-0.54	0.183	yes
paddy vs aquifer	*mcrA*	28/4	-0.71	0.371	-0.38	0.743	yes
paddy vs aquifer	*nifH*	31/6	-0.99	0.142	-0.24	0.418	yes
paddy vs aquifer	*nirK*	31/6	+0.41	0.371	-0.54	0.036	no
paddy vs aquifer	*nirS*	31/6	+0.70	0.371	+0.13	0.866	yes
paddy vs aquifer	*nosZ*	31/6	+0.35	0.430	-0.25	0.183	no
paddy vs aquifer	*pflD*	31/6	-1.26	0.116	-0.25	0.538	yes
paddy vs aquifer	*soxB*	23/6	+0.52	0.415	+0.08	0.866	yes
As-soil vs aquifer	*aioA*	19/6	+0.61	0.371	+0.24	0.866	yes
As-soil vs aquifer	*arsM*	19/6	-0.63	0.430	-0.08	0.866	yes
As-soil vs aquifer	*dsrA*	16/6	+0.36	0.682	+0.30	0.609	yes
As-soil vs aquifer	*dsrB*	16/6	+0.36	0.682	+0.31	0.609	yes
As-soil vs aquifer	*feoB*	19/6	+0.45	0.430	+0.14	0.046	yes
As-soil vs aquifer	*hydA*	17/6	-0.55	0.440	-0.08	0.922	yes
As-soil vs aquifer	*mcrA*	11/4	+0.30	0.540	+0.69	0.176	yes
As-soil vs aquifer	*nifH*	20/6	-0.77	0.242	+0.10	0.874	no
As-soil vs aquifer	*nirK*	20/6	-0.01	0.977	-0.44	0.046	yes
As-soil vs aquifer	*nirS*	20/6	+0.23	0.682	+0.06	0.922	yes
As-soil vs aquifer	*nosZ*	20/6	-0.02	0.957	-0.30	0.176	yes
As-soil vs aquifer	*pflD*	17/6	-0.79	0.268	+0.02	0.984	no
As-soil vs aquifer	*soxB*	16/6	+0.27	0.682	+0.05	0.874	yes

Studies, number in the first and second compartment. Positive effects are higher in the second compartment. q from permutation of compartment labels across studies (10,000) with Benjamini–Hochberg correction across these cells within each route, a smaller family than the panel-wide correction of Table S7, so the q-values compare routes and are not compartment findings. *dsrA* and *dsrB* share their marker taxa and are not independent cells.

S14 Compartment, composition, soil and corpus analyses in full

The main text summarises these analyses because they bear on the compartment question and not on the reliability of inference. The methods and results below are given in full, as they stood before that summary.

Exclusion criteria

Runs were excluded for four kinds of reason, and every exclusion is recorded at run level with its cause. Sequencing platforms whose error profiles the Illumina denoising path cannot model were removed, covering 454 pyrosequencing, Ion Torrent and PacBio SMRT, because forcing them through a single pipeline would produce sequence variants that are not comparable with the rest. Datasets that failed the isolation-source and study-design criteria were removed, including enrichment cultures, a geothermal system, point-of-use arsenic-removal filters and root material, meaning root and rhizoplane tissue and the iron plaque coating rice roots. Rhizosphere soil from rice paddies was retained deliberately, because in a flooded paddy the rhizosphere is where arsenic is mobilised, so the line is drawn between plant tissue and soil adhering to roots. Three further categories were found only after retrieval, namely deposits labelled as amplicon libraries but carrying 5 to 63 million reads, beyond an interval of the read-count distribution that contains no run at all; deposits with a constant quality score at every base, from which no error model can be estimated (three studies and one run); and one study whose paired files differ in length by one to three reads in every sample, which halts denoising. Each is described in Supplementary Section S2, and total exclusions are given in the exclusion ledger of the main text.

Depth stratification and read-depth threshold

Aquifer datasets were stratified at 50 m below ground surface where a depth was reported, with 100 m retained as a sensitivity threshold. Depth is unreported for most aquifer runs, so those runs are retained in the soil-versus-aquifer comparison and excluded from the depth-stratified one, which is reported rather than being pooled into a single aquifer estimate.

A sample-level floor of 10,000 retained reads is applied at consolidation, and it matters more than a threshold of this kind usually would, because retained depth is not balanced across compartments. Aquifer samples fall below the floor at 38.8 % against 11.4 % for paddy soil, and 26.7 % of aquifer samples carry fewer than 1,000 reads against 4.5 % for paddy soil (Mann–Whitney $p = 0.026$). A prevalence-based contrast between compartments of unequal depth would therefore read shallow sequencing as biological absence, which is why sequencing depth is carried as a covariate in the compartment models.

Much of the shortfall is a processing consequence rather than shallow sequencing. Primers are removed from the 5′ end only and reads are not truncated, because the corpus spans regions from V4 to V3–V4, so where a read is longer than its amplicon the 3′ tail runs into adapter sequence and the expected-error filter rejects most reads (Supplementary Section S2). Uniform processing across arms was held above per-sample yield, and the floor and the depth covariate account for the samples lost.

Community composition and diversity

Gene prevalence inferred through a bridge is a weaker signal than the taxonomy it is built from, so composition was also tested directly. Genus relative abundances were computed per sample and Bray-Curtis dissimilarities (Bray and Curtis, 1957) taken between all pairs, on the depth-floored and bacterially gated corpus. Compartments were compared by PERMANOVA (Anderson, 2001) with labels permuted across studies rather than across samples, for the same clustering reason that governs the gene models. Multivariate dispersion was computed per compartment alongside the location test, since PERMANOVA cannot distinguish a difference in centroid from a difference in spread and reporting one without the other is the commonest way the test is over-read.

Genus richness and the Shannon index were regressed on compartment with $\log_{10}$ retained reads as a covariate. Richness is more depth-sensitive than gene detection and depth tracks compartment in this corpus, so an unadjusted comparison of diversity would measure sequencing effort.

Geolocation and soil covariates

A compartment label is coarse, so the measured environment was tested as an alternative. Coordinates were recovered from sample attributes where deposited, giving 20 studies with enough geolocated samples to analyse. Soil pH, clay, sand, silt, organic carbon, cation exchange capacity, bulk density and nitrogen were queried from SoilGrids v2.0 (Poggio et al., 2021) at the 0–5 cm interval, which is the biologically active layer. Values are modelled at 250 m resolution, so they describe the overlying soil at the sampling location rather than the sampled material, which matters most for aquifer samples. Redox potential is not available globally at any resolution and is absent from the covariate set.

Each study with at least eight geolocated samples was tested separately, so that laboratory, primer and depositor effects are constant within a test. Environmental distance was the Euclidean distance between standardised covariates, geographic distance the great-circle distance between coordinates, and community distance the Bray-Curtis dissimilarity already computed. Environment was related to composition by a Mantel correlation and then by a partial Mantel correlation holding geographic distance constant, the difference between the two being the quantity of interest. Distance-decay was tested separately as the correlation between geographic and community distance within each study.

Sample attributes, arsenic and treatment screening

The full attribute record was harvested for every retained run, yielding 47,165 attribute rows across 2,927 samples. Arsenic concentrations were extracted from those rows wherever a field could be parsed to a number, and the units each study reported were recorded alongside. Attributes were also searched for evidence of laboratory manipulation, covering spiking, enrichment culture, reactor growth and incubation, and studies reporting any of these were excluded at run level. Field amendment trials were retained, since treatment applied to soil in place is not the same as material grown in a flask.

Compartment statistics

Compartment contrasts were first tested at study level, since samples within a study share a laboratory, a primer choice and a sequencing run and are not independent. Study-level prevalence was compared between compartments with a Mann–Whitney U test and corrected across genes by the Benjamini–Hochberg procedure. Genes with fewer than three studies in either arm were not tested.

Sensitivity of that comparison to quality control was assessed by repeating it under two gates applied singly and together, the first being the read-depth floor and the second a requirement that at least 20 % of a study's sequence variants be assigned to Bacteria or Archaea. Both gates are applied for the three-compartment analysis, since a study that deposits no bacterial 16S can contribute no signal to a bacterial gene panel.

Because sequencing depth turned out to differ between compartments, the primary analysis models detection directly. For each gene, per-sample detection was regressed on compartment and $\log_{10}$ retained reads using 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 detecting that gene in one compartment relative to another at equal depth, which is the quantity a prevalence comparison cannot express. Compartments were compared pairwise, giving three contrasts across paddy soil, arsenic-contaminated soil and aquifer.

Cluster-robust standard errors are anti-conservative when clusters are few, and the compartment contrast is a between-cluster comparison with fewer than ten aquifer studies. Every gene was therefore also tested by permutation, shuffling compartment labels across studies rather than across samples, refitting the model on each of 2,000 permutations and comparing the observed coefficient with its permutation null. Permutation p-values are the ones reported for inference, with the model-based values shown alongside so the difference between them is visible. Each odds ratio is reported with a 95 % interval from the Mancl–DeRouen bias-reduced covariance with a t reference on the number of studies minus two, which shows the size of difference each null leaves open.

Compartment contrasts

Scored by the bridge, with depth controlled and labels permuted across studies, no difference between any pair of compartments is resolved for any gene (Table 3 of the main text, Figure S14). That null reflects limited resolution rather than equivalence, since across the 60 gene-contrast pairs the 95 % interval on the odds ratio spans a median 11-fold range and differences of several-fold remain compatible with the data. The 8 nominal intervals that exclude unity do not survive permutation and correction across the panel (lowest $q = 0.15$, Supplementary Table S7). The strongest result anywhere in the panel is *aioA* between paddy soil and aquifer, at an odds ratio of 22.04 and $q = 0.154$. Testing abundance in place of detection raises 5 genes above $p < 0.05$ where none reached it before, although only 0 survive correction across the panel.

Community composition separates the compartments no better. Study-level PERMANOVA (Anderson, 2001) on Bray-Curtis distance (Bray and Curtis, 1957) returns $F = 34.1$ at $p = 0.305$ across the three groups, the pseudo-F being computed over samples while the p-value comes from permuting labels across studies, which is why a large statistic carries a modest p. The paddy-aquifer contrast, which the corpus was assembled to make, reaches only $p = 0.773$, and paddy against arsenic-contaminated soil $p = 0.149$. The one pairwise contrast below 0.05 is arsenic-contaminated soil against aquifer ($F = 25.1$, $p = 0.012$, uncorrected across three pairwise tests), the contrast between the two smallest arms and the one the corpus was not built to make, and since the overall test was not significant it is treated as exploratory. Multivariate dispersion is near-identical across compartments, so none of this is a difference in location masked by unequal spread. Between-study variation is as large as between-compartment variation, and since every study contributes to exactly one compartment the two cannot be sepa-

rated, which applies to the one nominal difference as much as to the absent ones.

The result is also stable under quality control, while the candidates it replaces are not (Table S14). No gene survives correction at baseline, under the read-depth floor, under the bacterial-fraction gate, or under both gates together. Which genes lead does change, since the four strongest candidates at baseline are *arsM*, *mtrB*, *pflD* and *hydA*, all paddy-enriched, while under the read-depth floor they are *aioA*, *mtrB*, *soxB* and *nirK*, all aquifer-enriched, so only *mtrB* appears among the leading four under both. Removing shallow samples raises aquifer prevalence for most genes and leaves paddy prevalence almost unchanged, as expected of a depth artefact. The apparent signal therefore changes direction under a filtering choice a reviewer would accept either way. Model-based inference would have reported otherwise, since cluster-robust standard errors flag several genes in each contrast where the study-level permutation flags none, and with between eight and thirty-seven studies per arm that gap is larger than the effects being tested.

Soil covariates and spatial structure

Repeated against the measured soil environment within studies (see below), the comparison gives the same answer. Environmental distance does correlate with community distance across 20 studies, at a median Mantel $r = +0.119$ ($p = 0.0083$). Geographic distance correlates at least as strongly ($r = +0.133$), and partialling geography out of the environmental correlation collapses it to a median $r = +0.002$, positive in only 10 of 20 studies and indistinguishable from zero ($p = 0.84$). The apparent environmental signal is therefore spatial autocorrelation, in that nearby sites resemble each other both in soil properties and in community composition, and the soil properties add nothing once that is taken into account.

Distance itself behaves as expected, with community dissimilarity rising with geographic separation in 18 of 23 studies (median $\rho = +0.135$, sign test $p = 0.011$), and the relationship reaches significance within 16 studies individually. Whatever structures these communities is therefore organised in space, while neither the compartment label nor the measured soil environment captures it.

Corpus quality

Of the 675 excluded runs, the size outliers (30), the runs with a constant quality score (139) and the upstream failures (27) were removed only after the sequences themselves were inspected, for reasons no metadata field exposes (Methods and Figure 2 of the main text).

Taxonomy exposed a larger problem, in that 11 of 92 studies, holding 578 otherwise retained runs, assigned under 20 % of their sequence variants to Bacteria or Archaea, and several assigned none at all. These are eukaryotic amplicon data deposited under metadata that declares neither, and one of them confirms the diagnosis in its own record by declaring an ITS target elsewhere. Failure to detect the expected primer in the reads flags every one of them and is the cheapest check available.

Five further studies were manipulated instead of mislabelled, reporting spiking, enrichment culture, reactor growth or laboratory incubation in attributes that a search of isolation source alone does not reach, and those runs were removed.

Arsenic itself is largely unrecorded, with usable concentrations for 58 of 3,041 runs across 3 studies, and only one of those states its units, so the exposure variable the corpus was built around cannot be used as a covariate. PICRUSt2 predicts the detoxification operon in roughly nine of every ten samples, and predicted detoxification genes exceed the two metabolism genes it reaches in 63 of 63 studies, by a median factor of 8.1, consistent with soil metagenomes (Dunivin et al., 2019) but still a prediction.

Table S14. Sensitivity of the paddy-aquifer contrast to quality control. Each row applies a different gate before the same study-level test. No gene survives correction under any gate, and which genes lead changes with it.

Quality gate applied	Paddy studies	Aquifer studies	Genes at $q < 0.05$
None (as first analysed)	39	8	0
Read-depth floor	36	8	0
Bacterial fraction	37	8	0
Both	35	8	0

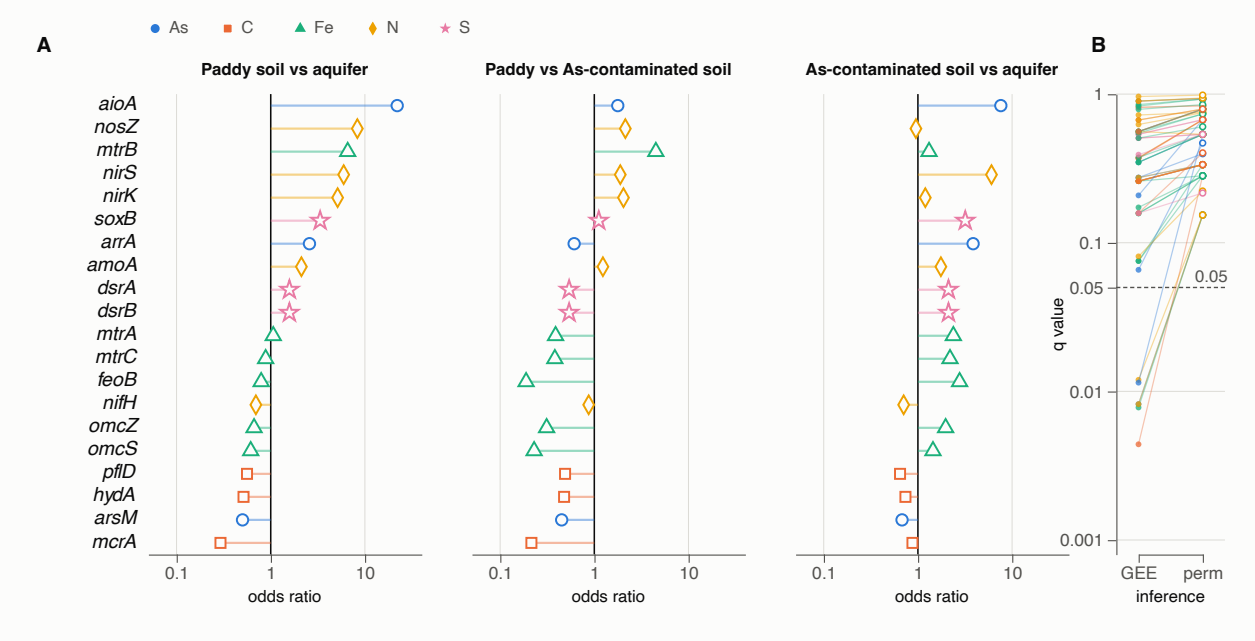


Fig. S14. Compartment contrasts in gene detection at equal sequencing depth. (**a**) Odds ratio for detection in the second compartment relative to the first at equal depth, on a logarithmic axis clipped at 0.05 and 40. A filled mark would indicate $q < 0.05$ by study-level permutation. (**b**) Each line joins one gene's q from cluster-robust standard errors to its q from permutation across studies. All contrasts in Table S7.

S15 Accuracy against measured gene content across environments

The comparison of Section S10 was repeated on 285 biosamples from 17 studies outside the corpus, each carrying one shotgun and one bacterial 16S amplicon run (Table S15b). Selection, processing, the primary outcome and three secondary analyses were fixed in a protocol whose checksum was recorded before any sequence was downloaded. Eligibility required Illumina shotgun metagenomes of at least 1 Gbp and bacterial 16S amplicon libraries of at least 5,000 reads with a mean read length of 250 or more from the same biosamples, at least six such pairs, and no virome, stable-isotope, culture or incubation work in the study record. The deterministic draw took every paddy-soil and groundwater study and at most 20 pairs from any study. The shortest amplicon library carries 17,618 reads, so the read-depth floor removes no pair. The protocol records one deviation, made during the run and before any result was examined. Denoising the amplicon reads of PRJNA843962, at about 2.2 million reads and 700,000 unique sequences per sample, took about seven hours per sample, so its trimmed reads were subsampled to 300,000 pairs with a fixed seed, the same reads being kept in both files of each pair, and the two samples already denoised at full depth were denoised again from the subsample so the study is uniform.

PICRUSt2 was held at version 2.6.3, the version that produced the corpus predictions, with Python 3.12.13 and DendroPy 5.0.6, after an unpinned environment resolved to PICRUSt2 2.6.4 with later Python and DendroPy releases and failed with a recursion error in seven studies. The two studies that had completed under 2.6.4 were rerun under 2.6.3 and their earlier output set aside. Every one of the 8,582 sequence variants of PRJNA788790 failed PICRUSt2's default requirement that 80 % of a sequence align to the reference, so neither route scores that study. Shotgun counts follow the primary rule of Section S10 applied to the first 1 Gbp of forward reads of each run, streamed and therefore not checksum-verified, with the truncation recorded per run.

Two analyses were added after the primary result was seen, each in a protocol recorded with a checksum before it ran, and are post hoc. The first relates agreement between routes to accuracy across cells. Agreement and the two accuracies share each cell's samples, so as a check outside that protocol the measured values were shuffled across biosamples within each cell 2,000 times, which leaves agreement unchanged and removes any association that does not depend on the measurement. The null means were +0.00 for PICRUSt2 and +0.00 for the bridge, against observed values of -0.20 and +0.30. The second tests whether accuracy differs between genes, with gene labels permuted within studies so that a study whose genes are all well or all poorly predicted does not count as a gene effect. Per-gene values in Table S15 are descriptive and no per-gene test is claimed.

Table S15. Accuracy of each route against measured gene content, by gene, across the paired amplicon and shotgun set. Values are medians over studies of the Spearman correlation across biosamples.

Cycle	Gene	Studies	PICRUSt2 ρ	95 % CI	Studies	Bridge ρ	95 % CI
As	*acr3*	16	+0.10	-0.11 to +0.32	–	–	–
As	*aioA*	15	-0.05	-0.19 to +0.16	16	-0.12	-0.44 to -0.02
As	*arrA*	–	–	–	14	+0.08	-0.18 to +0.28
As	*arsB*	16	+0.03	-0.09 to +0.27	–	–	–
As	*arsC_grx*	16	+0.29	+0.06 to +0.53	–	–	–
As	*arsC_myc*	16	+0.20	-0.04 to +0.33	–	–	–
As	*arsC_trx*	16	+0.13	-0.04 to +0.26	–	–	–
As	*arsM*	16	+0.31	+0.12 to +0.61	14	+0.03	-0.21 to +0.22
As	*arsR*	16	+0.10	-0.14 to +0.38	–	–	–
C	*bssA*	14	+0.21	-0.00 to +0.34	–	–	–
C	*hydA*	16	+0.27	-0.16 to +0.32	13	+0.12	+0.05 to +0.32
C	*mcrA*	10	+0.01	-0.13 to +0.38	10	-0.02	-0.16 to +0.38
C	*pflD*	16	+0.25	-0.23 to +0.38	14	+0.15	-0.21 to +0.44
Fe	*feoB*	16	+0.34	+0.09 to +0.61	15	+0.04	-0.08 to +0.30
Fe	*mtrA*	–	–	–	13	+0.11	-0.13 to +0.37
Fe	*mtrB*	–	–	–	4	+0.00	-0.10 to +0.23
Fe	*mtrC*	–	–	–	13	+0.14	-0.18 to +0.45
N	*amoA*	–	–	–	9	+0.22	-0.31 to +0.58
N	*nifH*	16	+0.23	+0.01 to +0.31	16	-0.01	-0.21 to +0.06
N	*nirK*	16	+0.30	+0.16 to +0.41	16	+0.18	-0.08 to +0.29
N	*nirS*	16	+0.29	+0.04 to +0.45	15	+0.10	-0.05 to +0.32
N	*nosZ*	16	+0.27	+0.08 to +0.40	16	-0.21	-0.32 to +0.07
P	*phoA*	16	+0.35	+0.13 to +0.42	–	–	–
P	*phoD*	16	+0.20	+0.10 to +0.58	–	–	–
P	*ppk1*	16	+0.15	-0.02 to +0.36	–	–	–
P	*pstS*	16	+0.21	-0.01 to +0.32	–	–	–
S	*dsrA*	16	+0.29	+0.13 to +0.46	13	+0.16	+0.14 to +0.42
S	*dsrB*	16	+0.24	+0.01 to +0.33	13	+0.15	-0.23 to +0.34
S	*soxB*	16	+0.11	-0.05 to +0.45	13	-0.02	-0.11 to +0.05

Studies, studies in which both the route and the measured abundance vary across at least five biosamples. Intervals resample studies (10,000, seed 42) and are descriptive. Genes with fewer than three studies under a route are omitted for that route. A dash marks a gene the route does not reach or reaches in fewer than three studies.

Table S15b. Accuracy of each route against measured gene content, by study. Values are medians over genes of the Spearman correlation across biosamples.

Study	Environment	Pairs	Genes	PICRUSt2 ρ	Genes	Bridge ρ
PRJNA600216	freshwater	20	24	+0.24	17	+0.01
PRJNA843962	freshwater	20	23	+0.13	15	+0.22
PRJNA1111131	groundwater	11	24	+0.70	17	-0.09
PRJNA699054	groundwater	16	24	+0.39	17	+0.33
PRJNA934733	groundwater	7	24	+0.23	14	-0.18
PRJNA1107715	paddy soil	8	23	+0.29	15	-0.21
PRJNA1121982	paddy soil	18	24	+0.17	16	+0.02
PRJNA1241498	paddy soil	20	24	+0.21	17	-0.09
PRJNA972838	paddy soil	15	24	-0.04	17	-0.04
PRJNA720816	rhizosphere	12	21	+0.19	5	+0.00
PRJNA1457409	sediment	20	24	+0.15	18	+0.06
PRJNA788790	sediment	18	0	–	0	–
PRJNA808857	sediment	20	24	+0.38	17	+0.26
PRJNA1202346	soil	20	24	+0.30	18	+0.13
PRJNA1214840	soil	20	23	+0.34	16	+0.41
PRJNA1233945	soil	20	22	+0.07	6	+0.08
PRJNA317932	soil	20	23	-0.05	12	+0.08

Pairs, biosamples with one shotgun and one amplicon run selected. Genes, genes in which the route and the measured abundance both vary across at least five biosamples. PRJNA788790 enters neither route because every one of its sequence variants failed PICRUSt2's default minimum-alignment check, and the pipeline's bridge step reads the PICRUSt2 table and so did not run for it.

S16 Tax4Fun2 as a fourth inference route

Tax4Fun2 1.1.5 (Wemheuer et al., 2020) and its reference data v2 were obtained from Zenodo record 10035668, a mirror of the original distribution, which is no longer online, and both archives were checked against recorded checksums. The software, reference, settings, score, cells, summaries and interpretation were fixed in a protocol whose checksum was recorded before Tax4Fun2 was run on any sequence. Settings are the package defaults

(Ref100NR, 97 % minimum identity, copy-number normalisation), with BLAST+ 2.16.0 in place of the bundled binary. One departure was specified in advance for tractability. The package asks BLAST for up to 1,000,000 hits per query and keeps the first, and 25 hits were requested instead, after a check on the smallest corpus study gave identical ortholog tables under both settings. Most corpus studies were predicted on a rented server with the same script, package and reference, and one study predicted on both machines differed by at most 3×10^{-16}.

Tax4Fun2 scores 71 of the 78 retained studies. The 7 it does not reach (PRJNA1039139, PRJNA357341, PRJNA559049, PRJNA690681, PRJNA695517, PRJNA886312, SRP376561) are those the bacterial-fraction audit flags as not 16S, and PICRUSt2 reaches none of them either. Across the 63 studies of the agreement comparison, the median share of a study's reads that reach no reference at 97 % identity is 75 %, ranging from 12 to 100 % between studies, so most reads in these environments contribute nothing to the prediction.

The paired-study amplicon outputs retained counts, taxonomy and PICRUSt2 tables but not the variant sequences, so these were regenerated by rerunning the unchanged amplicon pipeline on the same runs, with PRJNA843962 subsampled by the same script and seed as before. The protocol admitted a study only if its regenerated count table equals the stored one entry for entry. 13 of the 17 studies did, and the other 4 differ in 1 to 27 entries, at most 0.007 % of reads, and are excluded (Table S16b). PICRUSt2 and bridge values in the main-text comparison of Tax4Fun2 are those of the accepted studies, which is why they differ slightly from the values over all studies.

The protocol did not anticipate samples in which no read reaches a reference. Tax4Fun2 returns an empty profile for these, and the registered score treats them as zero for every gene. This applies to 138 of the 2,463 samples of the analysis set (5.6 %) and to 22 of the 209 biosample runs of the accepted paired studies. In a sensitivity analysis added after the results were seen, such samples are left unscored. The paired difference in corpus agreement becomes +0.10 (+0.03 to +0.14), the aquifer result is unchanged because every SRP507371 run with an empty profile lies below the read-depth floor, and across the paired studies Tax4Fun2 reaches +0.16 (+0.07 to +0.25), with paired differences of -0.03 (-0.11 to +0.03) from PICRUSt2 and +0.11 (+0.01 to +0.22) from the bridge. No conclusion changes. The 10,000-read floor removes no paired-study sample and so leaves the paired results unchanged. Table S16 gives the registered values by gene, and the protocol, its deviations file and every output are in the code repository.

Table S16. Tax4Fun2 as a fourth inference route, by gene. T, B and P are Tax4Fun2, the marker-taxa bridge and PICRUSt2. Cells to T–B are medians over corpus study-gene cells of the Spearman correlation across samples; T–shotgun is the correlation across the nine SRP507371 biosamples; T accuracy is the median over the accepted paired studies of the correlation with measured abundance.

Gene	Cells	T–P	B–P	T–B	T–shotgun	Studies	T accuracy
acr3	0	–	–	–	+0.05	13	+0.01
aioA	55	+0.17	+0.20	+0.07	+0.23	13	+0.06
arsB	0	–	–	–	+0.17	13	+0.14
arsC_grx	0	–	–	–	+0.22	13	+0.17
arsC_myc	0	–	–	–	+0.07	13	+0.21
arsC_trx	0	–	–	–	-0.03	13	+0.14
arsM	56	+0.23	-0.12	+0.07	-0.10	13	+0.06
arsR	0	–	–	–	-0.02	13	+0.37
bssA	0	–	–	–	+0.46	8	-0.06
dsrA	53	+0.59	+0.47	+0.37	+0.60	10	+0.19
dsrB	53	+0.54	+0.47	+0.37	+0.73	10	+0.21
feoB	56	+0.24	+0.14	+0.14	+0.08	13	+0.01
hydA	54	+0.38	+0.37	+0.44	-0.22	13	+0.27
mcrA	40	+0.84	+0.91	+0.93	+0.05	8	+0.08
nifH	57	+0.49	+0.26	+0.25	+0.18	13	+0.21
nirK	57	+0.29	+0.02	+0.08	+0.33	13	+0.11
nirS	57	+0.49	+0.14	+0.04	+0.15	13	+0.25
nosZ	57	+0.36	+0.07	+0.15	+0.07	13	+0.27
pflD	54	+0.51	+0.41	+0.44	+0.73	13	+0.06
phoA	0	–	–	–	+0.40	13	+0.10
phoD	0	–	–	–	+0.37	13	+0.28
ppk1	0	–	–	–	+0.37	13	+0.35
pstS	0	–	–	–	+0.22	13	+0.25
soxB	45	+0.43	+0.22	+0.44	+0.58	13	+0.11

Cells, corpus cells in which the bridge, PICRUSt2 and Tax4Fun2 all vary across at least five samples. Studies, accepted paired studies in which Tax4Fun2 and the measured abundance both vary across at least five biosamples; the median is shown for genes with three or more. A dash marks a gene a route does not reach or whose score is constant. Values are registered results, with empty profiles scored zero.

Table S16b. Regeneration of the paired-study sequence variants and Tax4Fun2 accuracy by study. T and P are Tax4Fun2 and PICRUSt2; their ρ columns are medians over genes of the Spearman correlation with measured abundance across biosamples.

Study	Accepted	Entries	Differing	Reads (%)	Genes	T ρ	Genes	P ρ
PRJNA1107715	yes	1,302	0	0.000	22	+0.21	23	+0.29
PRJNA1111131	yes	4,716	0	0.000	24	+0.28	24	+0.70
PRJNA1121982	yes	6,043	0	0.000	23	+0.11	24	+0.17
PRJNA1202346	yes	61,868	0	0.000	24	+0.40	24	+0.30
PRJNA1214840	yes	11,258	0	0.000	23	+0.35	23	+0.34
PRJNA1233945	yes	4,158	0	0.000	20	-0.02	22	+0.07
PRJNA1241498	yes	9,668	0	0.000	24	+0.02	24	+0.21
PRJNA1457409	no	57,682	27	0.007	–	–	–	–
PRJNA317932	yes	2,666	0	0.000	20	-0.01	23	-0.05
PRJNA600216	yes	11,342	0	0.000	24	+0.27	24	+0.24
PRJNA699054	no	18,314	1	0.005	–	–	–	–
PRJNA720816	yes	758	0	0.000	20	+0.20	21	+0.19
PRJNA788790	yes	8,868	0	0.000	24	+0.11	–	–
PRJNA808857	no	46,198	7	0.005	–	–	–	–
PRJNA843962	no	54,471	3	0.002	–	–	–	–
PRJNA934733	yes	3,386	0	0.000	24	+0.11	24	+0.23
PRJNA972838	yes	14,933	0	0.000	24	-0.10	24	-0.04

Entries, (variant, sample, count) entries of the stored count table. Differing, entries present in only one of the stored and regenerated tables. Reads, share of reads not reproduced. A study enters the comparison only if its regenerated table equals the stored one. PRJNA788790 has no PICRUSt2 value because every one of its sequence variants failed PICRUSt2's minimum-alignment check, while Tax4Fun2, which assigns by BLAST, scores it.

References

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.